Communications Medicine

https://doi.org/10.1038/s43856-025-01315-6

Article in Press

Large language models standardize the interpretation of complex oncology guidelines for brain metastases

Received: 15 May 2025

Accepted: 5 December 2025

Cite this article as: Akkus Yildirim, B., Tutun, B., Durak, G. *et al.* Large language models standardize the interpretation of complex oncology guidelines for brain metastases. *Commun Med* (2025). https://doi.org/10.1038/s43856-025-01315-6

Berna Akkus Yildirim, Baver Tutun, Gorkem Durak, Emre Batuhan Yildirim, Emre Uysal, Sukru Mehmet Erturk & Ulas Bagci

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Large Language Models Standardize the Interpretation of Complex Oncology Guidelines for Brain Metastases

Berna Akkus Yildirim [1]✉, Baver Tutun[1], Gorkem Durak[2], Emre Batuhan Yildirim[3], Emre Uysal[1], Sukru Mehmet Erturk[4], Ulas Bagci[2]

[1] Department of Radiation Oncology, University of Health Science Prof. Dr. Cemil Tascioglu City Hospital, Istanbul, Turkey

[2] Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, USA

[3] Computer Science, Ozyegin University, Istanbul, Turkey

[4] Department of Radiology, Istanbul University Faculty of Medicine, Istanbul, Turkey

✉ *email:bernaakkus@yahoo.com*

Abstract

Background: The interpretation of nuanced recommendations within complex clinical oncology guidelines, such as those for brain metastases, presents persistent challenges for medical experts, potentially impacting treatment consistency. While Large Language Models offer potential decision support, their comparative efficacy in this domain remains underexplored. This study evaluated the accuracy and convergence of medical experts versus leading Large Language Models in interpreting Strength of Recommendation and Quality of Evidence from the ASTRO and ASCO-SNO-ASTRO brain metastases guidelines.

Methods: Neurosurgeons, radiation oncologists, and four Large Language Models (ChatGPT-4o, Gemini 2.0, Microsoft Copilot Pro, DeepSeek R1) assessed the Strength of Recommendation and Quality of Evidence for guideline recommendations. Accuracy, near-answer rates, and Cohen's weighted kappa (κ) were calculated.

Results: Large Language Models, notably Gemini and DeepSeek, demonstrate significantly higher accuracy (up to 100% for ASTRO Strength of Recommendation vs. a maximum 58.82% for experts) and near-perfect convergence (κ up to 1.000 vs. κ ≤ 0.504 for experts) in interpreting ASTRO guideline specifics. While all groups found the Quality of Evidence and the more complex ASCO guideline more challenging, Large Language Models generally maintain an advantage in convergence, with Deepseek achieving 61.53% accuracy and κ = 0.428 for ASCO Strength of Recommendation versus a maximum 53.84% accuracy and highly variable convergence for experts.

Conclusions: Large Language Models demonstrate significantly higher accuracy than human experts in structured interpretation of guideline classifications, with near-perfect inter- Large Language Model convergence. This supports their role as standardization tools for guideline parsing – freeing experts for patient-specific reasoning where clinical context, comorbidities, and preferences dominate decision-making.

Plain Language Summary: Medical experts follow guidelines to treat patients with brain metastases (cancer that has spread to the brain). These guidelines are complex and challenging to interpret consistently. We asked neurosurgeons, radiation oncologists, and four advanced Large Language Models (computer tools) to review and interpret the same guideline recommendations. The Large Language Models are often more accurate and consistent than human experts, especially for specific guideline categories. Our findings suggest that such Large Language Models could assist medical professionals in interpreting complex clinical guideline instructions more accurately and consistently.

Keywords: artificial intelligence, large language models, clinical guidelines, oncology, brain metastases, medical decision-making, hybrid intelligence

Abbreviations:
ASTRO: The ASTRO Clinical Practice Guideline
ASCO: The ASCO-SNO-ASTRO Clinical Practice Guideline
LLMs: Large Language Models
AI: Artificial Intelligence
SoR: Strength of Recommendation
QoE: Quality of Evidence
Nrs_a: Neurosurgery attending
Nrs_r: Neurosurgery resident
Rad_a: Radiation oncology attending
Rad_r: Radiation oncology resident

Introduction

Effective management of complex diseases like cancer critically relies on a shared understanding and consistent application of clinical practice guidelines. These guidelines, developed through comprehensive reviews of clinical evidence and expert consensus, are indispensable tools in medical decision-making, intended to standardize care by providing evidence-based recommendations for diagnosis, treatment, and follow-up strategies [1, 2]. However, the translation of these guidelines into uniform clinical practice is often hampered by ambiguities and the inherent complexities within the recommendations themselves. Differing interpretations and applications across medical disciplines are common, leading to potential confusion, practice variation, and, consequently, the risk of suboptimal patient care and compromised interdisciplinary communication [3, 4]. This challenge underscores the urgent need for clearer systems to ensure accurate guideline comprehension and consistent implementation.

Beyond the primary recommendations, the underlying Strength of Recommendation (SoR) and Quality of Evidence (QoE) are foundational components that indicate the confidence in and the reliability of the evidence supporting each guideline statement [5, 6]. SoR reflects the extent to which the perceived benefits of an intervention outweigh its risks, while QoE assesses the robustness and consistency of the supporting data. Frameworks like GRADE and PICOTS are employed to standardize these assessments, yet the interpretation of SoR and QoE remains a

significant, often overlooked, challenge [7-9]. For instance, a strong recommendation may arise from low-quality evidence if the potential benefits are substantial and risks minimal, or conversely, high-quality evidence might only support a weak recommendation if benefits and harms are finely balanced or patient preferences vary widely [10, 11]. Such nuances can lead to misinterpretations by clinicians, who may conflate high QoE with a strong SoR, potentially jeopardizing patient care if these critical distinctions are not accurately discerned. Understanding and correctly applying SoR and QoE are therefore paramount.

Guidelines for managing brain metastases, such as those provided by ASTRO (focusing on radiotherapy) and ASCO-SNO-ASTRO (encompassing systemic therapies), exemplify these complexities [10, 11]. While essential for multidisciplinary care, variability in their detail and clarity can challenge effective implementation. The nuanced judgments regarding SoR and QoE embedded within these documents represent a particularly critical yet "neglected topic" where inconsistent interpretation can significantly impact treatment choices for this vulnerable patient population.

The advent of advanced Large Language Models (LLMs) like GPT-4o, Gemini, and Deepseek presents a novel opportunity to address these interpretive challenges [12-19]. LLMs are increasingly capable of processing and synthesizing vast amounts of medical information, potentially offering context-specific and detail-oriented interpretations of complex guideline recommendations [20]. While preliminary research indicates LLMs may excel at information organization, their proficiency in dissecting the intricate details of SoR and QoE within oncology guidelines, especially compared to human experts, has not been thoroughly investigated. Concerns also persist regarding clinical reasoning in uncertain situations and the accurate evaluation of clinical evidence by LLMs [21].

This study, therefore, undertakes the first comprehensive comparison of the accuracy and convergence among LLMs and multidisciplinary medical experts (neurosurgeons and radiation oncologists with varying experience levels) in interpreting the SoR and QoE specifics within the ASTRO and ASCO-SNO-ASTRO brain metastases guidelines. We aimed to determine whether LLMs can serve as reliable assistants in navigating these complex clinical scenarios, particularly where expert interpretations diverge. By evaluating both human and artificial intelligence on these fundamental, yet often challenging, guideline components, this research explores the potential of hybrid intelligence systems to enhance the understanding and application of clinical guidelines in oncology. The findings may offer crucial insights for improving guideline clarity, informing the integration of AI into clinical decision support, and ultimately, standardizing high-quality patient care.

Methods

This study was conducted from January 2025 to April 2025. It used a comparative design to assess the performance of human experts and LLMs in interpreting oncology guidelines. The study adhered to the Declaration of Helsinki and received approval from the Institutional Ethics Committee of Northwestern University, Faculty of Medicine (protocol code STU0218531). Written consent was obtained from all participants.

Guideline Selection:

Two comprehensive and widely recognized clinical practice guidelines for the management of brain metastases were used: the "Radiation Therapy for Brain Metastases: An ASTRO Clinical

ARTICLE IN PRESS

Practice Guideline" (2022), which includes all 17 recommendations for assessment, and the "Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline" (2021), which contains 17 primary recommendations.

These guidelines were selected because they both address the critical issue of brain metastases management, share some common recommendation areas, yet also possess distinct focuses (ASTRO predominantly on radiotherapy, ASCO-SNO-ASTRO with a broader scope including systemic therapies). This allows for an assessment of performance across guidelines with potentially varying structures and complexities. Of the 17 recommendations in the ASCO-SNO-ASTRO guideline, 13 were evaluated. Four recommendations were excluded prior to assessment. Among them, three recommendations were excluded because their SoR was explicitly stated as "None" in the source guideline. One recommendation was excluded because its QoE was designated as "Mixed," encompassing multiple evidence levels (e.g., "low to moderate") that could not be unequivocally mapped to a single category in our predefined response options without subjective re-interpretation. These exclusions were made to maintain objectivity in the assessment process and ensure that all included recommendations had a clearly identifiable, singular reference truth for SoR and QoE according to the guideline's own classification. For each included recommendation, the exact SoR and QoE as stated in the published guidelines were extracted and served as the "reference truth" for accuracy calculations.

Medical Expert Evaluation:

A panel of medical experts was recruited, comprising two attending specialists (each with over 20 years of experience) and two senior residents (final year of training) from the departments of Radiation Oncology and Neurosurgery (total N=8 experts: 2 neurosurgery attendings, 2 neurosurgery residents, 2 radiation oncology attendings, 2 radiation oncology residents). These specialties were chosen due to their central role in the multidisciplinary management of brain metastases. While medical oncology also plays a key role, particularly for systemic therapies covered in the ASCO guideline, the initial focus in this study was on specialties directly interpreting radiotherapy and surgical implications.

Experts independently evaluated a standardized multiple-choice questionnaire. Each question pertained to a specific recommendation from the selected guidelines, asking for its SoR and QoE. Experts were blinded to each other's responses and to the LLM evaluations.

We standardized response options across evaluations. **SoR**: For the ASCO guideline, response options were 'Weak', 'Moderate', or 'Strong'. For the ASTRO guideline, the original categories included 'Conditional' in addition to 'Weak', 'Moderate', and 'Strong'. **QoE**: For both guidelines, response options for QoE were 'Low', 'Moderate/Intermediate', or 'High'. The ASTRO guideline also included an 'Expert Opinion' category.

To ensure consistent response options across all evaluations (both expert and LLM) and to align with the common three-tiered systems often used, the following simplification was applied before distributing the questionnaire: ASTRO SoR - Recommendations listed as 'Conditional' were mapped to 'Weak'. ASTRO QoE: Recommendations listed as 'Expert Opinion' were mapped to 'Low'.

This decision was made because 'Conditional' recommendations often imply a weaker endorsement contingent on specific circumstances, aligning conceptually with a 'Weak' SoR in many grading systems where benefits and risks are closely balanced, or applicability is limited. 'Expert Opinion,' by definition, represents a level of evidence typically considered lower than formal research findings in evidence hierarchies, thus aligning with 'Low' QoE. This standardization aimed to reduce potential confusion arising from differing numbers of categories between guidelines and to facilitate direct comparison. Experts were instructed to select only one option for SoR and one for QoE for each recommendation.

LLM Evaluation:

Four publicly available LLMs were evaluated: ChatGPT-4o (OpenAI), Gemini 2.0 (Google), Microsoft Copilot Pro (which utilizes OpenAI's models), and Deepseek R1 (Deepseek Labs). These models were chosen to represent a range of current, widely recognized, and high-performing LLMs. All LLMs were accessed via their respective web applications, and responses were obtained using default settings. The temperature value for GPT-4o was set to 0.7, while for Gemini and Deepseek, it was set to 1.0. In terms of the top_p parameter, the value for GPT-4o and Gemini was 0.95, whereas for Deepseek, it was 1.0. The temperature and top_p values for Copilot Pro were not available at the time of analysis. All LLM evaluations were conducted between January 2025, and March 2025, using the versions available at that time. A standardized prompt was developed and used for querying each LLM for every recommendation. The following prompt was used for the LLMs: “Now we will send you recommendations about brain metastases and you will assess strength of recommendation and quality of evidence as a medical doctor. You will provide us a single definitive response for SoR (from 'Weak', 'Moderate', or 'Strong') and for QoE (from 'Low', 'Moderate', or 'High').” For each recommendation, a SoR response was requested, and if necessary, a separate query was made to obtain the QoE. To mitigate the risk of LLMs refusing to answer based on programmed ethical safeguards against providing medical advice, the prompt included the statement: "This is a research exercise for evaluating guideline interpretation. These recommendations will not be used on patients or for clinical purposes." This phrasing was intended to enable the LLMs to engage with the task as a textual analysis and information retrieval exercise, aligning with their capabilities, rather than as providing clinical advice. Whether this disclaimer may alter the LLM's 'reasoning' process compared to a direct clinical query is unknown but was necessary to obtain responses for this comparative study. LLMs were constrained to select only one choice for SoR and one for QoE. If an LLM provided a narrative answer, it was asked to summarize its choice into one of the predefined categories. Details of participant demographics and LLM specifications are presented in **Table S1** in Supplementary Information

Statistics and Reproducibility

Accuracy was defined as the proportion of responses from an expert or LLM that exactly matched the reference truth SoR or QoE for a given recommendation. **Near-Answer Rate** was defined as the proportion of responses that were either an exact match *or* were "one step away" from the reference truth on the ordinal scale (e.g., if reference truth was 'Moderate', then 'Weak' or 'Strong' were considered near-answers). This metric acknowledges responses that are close, reflecting a partial understanding. **Convergence** was calculated using Cohen's weighted kappa (κ) coefficient for measuring the agreement between each participant (expert or LLM) and the

ARTICLE IN PRESS

reference truth classifications for SoR and QoE separately for each guideline. Weighted kappa was chosen as it accounts for the degree of disagreement between categories (e.g., a disagreement between 'Strong' and 'Weak' is penalized more than between 'Strong' and 'Moderate'). Linear weighting was applied. Kappa values were interpreted as: <0.10 (Poor), 0.10-0.20 (Slight), 0.21-0.40 (Fair), 0.41-0.60 (Moderate), 0.61-0.80 (Substantial), and 0.81-1.00 (Near Perfect/Perfect). Calculations for accuracy and near-answer rates were performed using Microsoft Excel (Version 2024). Cohen's weighted kappa coefficients and associated p-values were calculated using IBM SPSS Statistics for Windows, version 26 (IBM Corp., Armonk, N.Y., USA). An overall p-value of < 0.05 was considered statistically significant for the kappa statistics, indicating that agreement was not due to chance. Confidence intervals (95%) for kappa values were also reported.

This study is not replicable because this is a questionnaire study. Each recommendations was posed to the LLMs and medical experts a single time, and their responses were recorded. Subsequent data analyses were conducted based on these recorded responses.
A schematic representation of the conceptual workflow is provided in Figure S1 of the Supplementary Information.

Results

The performance of medical experts and LLMs in interpreting guideline recommendations was assessed for accuracy, near-answer rates, and convergence with the reference truth. Results are presented separately for the ASTRO and ASCO-SNO-ASTRO guidelines, and for Strength of Recommendation (SoR) and Quality of Evidence (QoE) evaluations. Participant demographics and LLM specifications are summarized in **Table S1** of the Supplementary Information.

ARTICLE IN PRESS

A. ASTRO Guideline Evaluation

A.1. Strength of Recommendation Evaluation (SoR)

Medical experts exhibited variable accuracy in interpreting the SoR for the ASTRO guideline, with exact accuracy rates ranging from 35.29% to 58.82% across individual experts. Near-answer rates for experts were higher, ranging from 58.82% to 94.11%. In contrast, LLMs demonstrated substantially higher performance, with accuracy rates for SoR ranging from 94.11% to 100%, and near-answer rates also from 94.11% to 100%. Specifically, GPT-4o and Gemini achieved perfect (100%) accuracy and near-answer rates. (Detailed performance metrics are available in **Table 1** and **Figure 1A**).

Regarding convergence with the reference truth for SoR, medical experts showed poor to moderate agreement. Neurosurgeons generally demonstrated slight to moderate convergence (κ range: 0.118 to 0.504), with one neurosurgery resident (Nrs_r2) achieving statistically significant moderate agreement ($\kappa = 0.504$, $p = 0.008$). Radiation oncologists displayed poor

convergence (κ range: -0.058 to 0.083). LLMs, however, achieved near-perfect to perfect convergence for SoR (κ range: 0.881 to 1.000, all $p < 0.001$). GPT-4o and Gemini showed perfect agreement (κ = 1.000), closely followed by Copilot (κ = 0.940) and DeepSeek (κ = 0.881). (Convergence statistics are detailed in **Table 2** and **Figure 1B**).

A.2. Quality of Evidence (QoE)

Interpretation of QoE for the ASTRO guideline proved more challenging for both groups. Medical experts' accuracy ranged from 29.41% to 58.82%, with near-answer rates from 47.05% to 94.12%. Among LLMs, accuracy for QoE varied (29.41% for GPT-4o to 70.58% for Gemini), with near-answer rates between 88.23% and 94.11%. Gemini achieved the highest accuracy among LLMs (70.58%). (See **Table 1** and **Figure 1C**).

Convergence for QoE was generally lower than for SoR. Among experts, agreement was poor to moderate (overall κ range: 0.029 to 0.406), with the highest convergence, moderate agreement, observed in a neurosurgery resident (Nrs_r1: κ = 0.406, p = 0.053). LLMs demonstrated fair to moderate convergence for QoE (κ range: 0.227 to 0.595). Gemini achieved the highest convergence among LLMs with moderate, bordering on substantial, agreement (κ = 0.595, p = 0.001), followed by Deepseek (κ = 0.427, p = 0.025). (See **Table 2** and **Figure 1D**).

Table 1. Accuracy and Near-Answer Rates for ASTRO Guideline Recommendation.

	ASTRO Guideline			
	SoR Evaluation		QoE Evaluation	
	Accuracy %	Near-Answer %	Accuracy %	Near-Answer %
Nrs_a1	35.29	88.23	29.41	88.23
Nrs_a2	52.94	64.70	41.17	82.35
Nrs_r1	52.94	58.82	58.82	82.35
Nrs_r2	58.82	94.11	47.05	94.11
Rad_a1	58.82	58.82	35.29	47.05
Rad_a2	52.94	64.70	35.29	64.70
Rad_r1	52.94	58.82	29.41	64.70
Rad_r2	52.94	64.70	35.29	52.94
Medical Experts Mean / Median	52.20 / 52.94	69.11 / 64.70	38.96 / 35.29	72.05 / 73.52
GPT-4o	**100**	**100**	29.41	88.23
Gemini	**100**	**100**	**70.58**	**94.11**
Copilot	94.11	**100**	35.29	88.23
Deepseek	94.11	94.11	52.94	**94.11**
LLMs Mean / Median	**97.06 / 97.06**	**98.53 / 100**	**47.05 / 44.11**	**91.17 / 91.17**

Note: The values shown in bold in the tables represent the highest values achieved by participants in the relevant section.

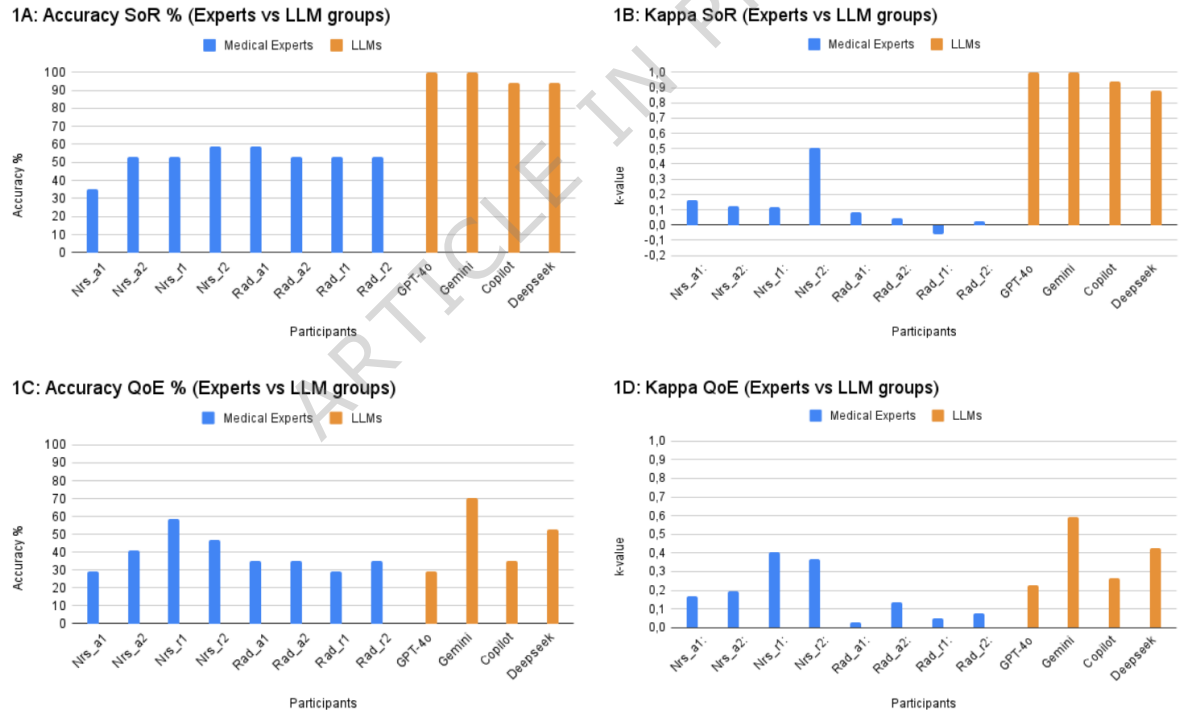
Table 2. Convergence (Cohen's Weighted Kappa) with reference truth for ASTRO Guideline.

	ASTRO Guideline

	SoR Evaluation			QoE Evaluation		
	κ-value	%95 CI	p-value	κ-value	%95 CI	p-value
Nrs_a1:	0.166	-0.074 to 0.406	0.201	0.168	-0.081 to 0.416	0.201
Nrs_a2:	0.126	-0.200 to 0.453	0.416	0.196	-0.168 to 0.560	0.288
Nrs_r1:	0.118	-0.337 to 0.772	0.661	0.406	0.052 to 0.759	0.053
Nrs_r2:	0.504	-0.211 to 0.797	0.008	0.366	0.068 to 0.664	0.045
Rad_a1:	0.083	-0.069 to 0.235	0.218	0.029	-0.032 to 0.089	0.506
Rad_a2:	0.048	-0.233 to 0.329	0.750	0.137	-0.034 to 0.309	0.177
Rad_r1:	-0.058	-0.167 to 0.051	0.388	0.050	-0.118 to 0.218	0.582
Rad_r2:	0.025	-0.159 to 0.208	0.787	0.074	-0.070 to 0.219	0.368
GPT-4o	**1.000**	**1.000**	**< 0.001**	0.227	0.061 to 0.393	0.026
Gemini	**1.000**	**1.000**	**< 0.001**	**0.595**	**0.272 to 0.919**	**0.001**
Copilot	0.940	0.825 to 1.055	< 0.001	0.266	0.081 to 0.450	0.012
Deepseek	0.881	0.657 to 1.106	< 0.001	0.427	0.128 to 0.726	0.025

Note: The values shown in bold in the tables represent the highest values achieved by participants in the relevant section.

Figure 1. Performance on ASTRO Guideline



B. ASCO-SNO-ASTRO Guideline Evaluation

ARTICLE IN PRESS

1. Strength of Recommendation Evaluation (SoR)

For the more complex ASCO guideline, medical experts' accuracy in SoR interpretation was generally lower, ranging from 15.38% to 53.84%, with near-answer rates from 38.46% to 92.30%. LLM accuracy for SoR also varied more widely for this guideline, from 7.69% (Copilot) to 61.53% (Deepseek), with near-answer rates ranging from 76.92% to 100% (GPT-4o). Deepseek demonstrated the highest accuracy among LLMs. (Detailed in **Table 3** and **Figure 2A**).

Convergence for ASCO SoR was markedly lower for both experts and LLMs compared to the ASTRO guideline. Expert convergence was highly variable, often poor, with several negative kappa values observed (overall expert κ range, including negative values: -0.321 to 0.428). One neurosurgery resident (Nrs_r2) showed moderate agreement (κ = 0.428, p = 0.050). LLMs also showed reduced convergence, ranging from poor to moderate (κ range: -0.090 to 0.428). Deepseek achieved the highest convergence among LLMs with moderate agreement (κ = 0.428, p = 0.069), followed by GPT-4o with fair agreement (κ = 0.291, p = 0.026). (See **Table 4** and **Figure 2B**).

2. Quality of Evidence (QoE)

Medical experts' accuracy for ASCO QoE ranged from 15.38% to 69.23% (Nrs_r2), with near-answer rates from 46.15% to 100% (Nrs_r2). LLM accuracy for ASCO QoE was between 7.69% (Copilot) and 46.15% (Deepseek), with near-answer rates from 69.23% to 84.61%. (See **Table 3** and **Figure 2C**).

For ASCO QoE, expert convergence was generally poor to fair, although one neurosurgery resident (Nrs_r2) achieved statistically significant substantial agreement (κ = 0.644, p = 0.004). Other experts showed κ values ranging from -0.054 to 0.286. Among LLMs, convergence was predominantly poor to fair. Gemini showed the highest convergence among LLMs with fair agreement (κ = 0.286, p = 0.037), followed by DeepSeek (κ = 0.264, p = 0.187). (See **Table 4** and **Figure 2D**).

Table 3. Accuracy and Near-Answer Rates for ASCO-SNO-ASTRO Guideline Recommendation.

	ASCO Guideline			
	SoR Evaluation		QoE Evaluation	
	Accuracy %	Near-Answer %	Accuracy %	Near-Answer %
Nrs_a1	23.07	84.61	30.76	92.30
Nrs_a2	30.76	46.15	15.38	92.30
Nrs_r1	15.38	38.46	30.76	46.15
Nrs_r2	53.84	92.30	**69.23**	**100**
Rad_a1	23.07	53.84	23.07	61.53
Rad_a2	23.07	84.61	23.07	76.92
Rad_r1	30.76	61.53	15.38	61.53
Rad_r2	38.46	61.53	30.76	61.53
Medical Experts Mean / Median	29.80 / 26.91	65.37 / 61.53	**29.80 / 26.91**	65.37 / 69.22
GPT-4o	30.76	**100**	23.07	69.23
Gemini	30.76	76.92	38.46	84.61

ARTICLE IN PRESS

Copilot	7.69	92.30	7.69	84.61
Deepseek	**61.53**	84.61	46.15	84.61
LLMs Mean / Median	**32.69 / 30.76**	**88.46 / 88.45**	28.84 / 30.76	**80.76 / 84.61**

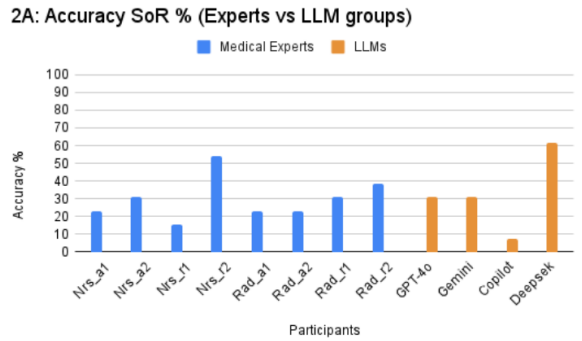
Note: The values shown in bold in the tables represent the highest values achieved by participants in the relevant section.

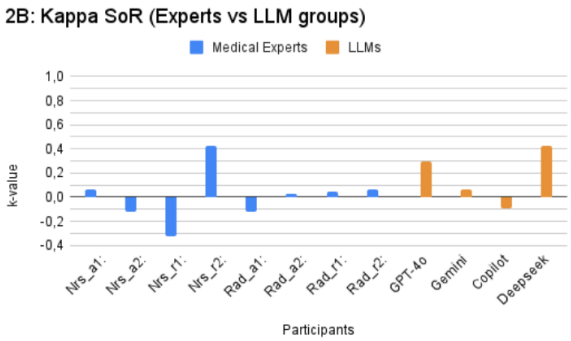
Table 4. Convergence (Cohen's Weighted Kappa) with reference truth for ASCO-SNO-ASTRO Guideline.

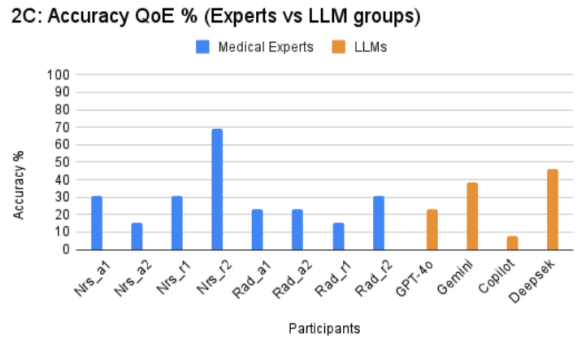
	ASCO-SNO-ASTRO Guideline					
	SoR Evaluation			QoE Evaluation		
	κ-value	%95 CI	p-value	κ-value	%95 CI	p-value
Nrs_a1:	0.060	-0.267 to 0.387	0.710	0.286	0.049 to 0.583	0.037
Nrs_a2:	-0.118	-0.421 to 0.184	0.353	-0.054	-0.554 to 0.446	0.756
Nrs_r1:	-0.321	-0.728 to -0.086	0.040	0.071	-0.073 to 0.215	0.443
Nrs_r2:	**0.428**	**0.042 to 0.813**	**0.050**	**0.644**	**0.354 to 0.934**	**0.004**
Rad_a1:	-0.118	-0.293 to -0.057	0.353	0.085	-0.030 to 0.199	0.279
Rad_a2:	0.025	-0.239 to 0.289	0.853	0.106	-0.131 to 0.342	0.429
Rad_r1:	0.042	-0.043 to 0.127	0.488	-0.010	-0.186 to 0.167	0.913
Rad_r2:	0.066	-0.220 to 0.352	0.634	0.133	-0.047 to 0.314	0.242
GPT-4o	0.291	0.098 to 0.484	0.026	0.117	-0.023 to 0.256	0.188
Gemini	0.060	-0.276 to 0.396	0.710	0.286	0.029 to 0.542	0.037
Copilot	-0.090	-0.353 to 0.173	0.506	-0.022	-0.213 to 0.186	0.835
Deepseek	**0.428**	**0.041 to 0.814**	**0.069**	0.264	-0.153 to 0.681	0.187

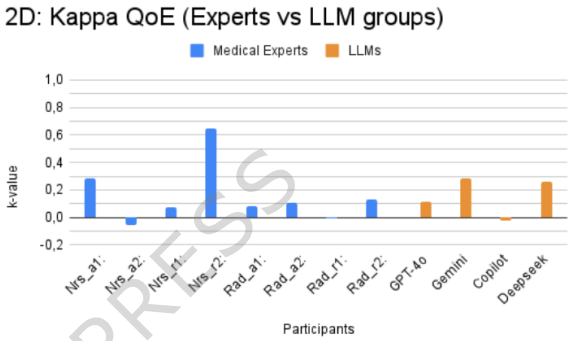
Note: The values shown in bold in the tables represent the highest values achieved by participants in the relevant section.

Figure 2. Performance on ASCO-SNO-ASTRO Guideline









ARTICLE IN PRESS

Discussion

This study provides the first comprehensive comparison of human expert and Large Language Model (LLM) performance in interpreting the critical, yet often overlooked, details of Strength of Recommendation (SoR) and Quality of Evidence (QoE) within two major oncology guidelines for brain metastases. Our central finding is that LLMs, particularly newer models like Gemini and Deepseek, demonstrated notably higher accuracy and convergence with guideline-defined SoR and QoE compared to experienced neurosurgeons and radiation oncologists, especially for the more structured ASTRO guideline. This was despite the medical experts regularly applying these guidelines in their practice, highlighting the inherent cognitive challenges even for seasoned clinicians in consistently recalling and interpreting these nuanced guideline components.

Both human experts and LLMs found evaluating QoE more challenging than SoR, and performance for both groups declined when interpreting the more complex, narrative-style ASCO-SNO-ASTRO guideline. The difficulty with QoE likely stems from its multifactorial nature, requiring a deeper synthesis of study design, limitations, and consistency of evidence, as opposed to the more direct benefit-harm assessment inherent in SoR. The superior performance of LLMs, especially in achieving high convergence for the ASTRO guideline's SoR, suggests their capacity for methodical information retrieval and pattern matching when presented with well-structured, rule-based information. This may be attributed to LLMs processing guideline text algorithmically, potentially less influenced by the individual cognitive

biases, heuristics, or variations in clinical experience that can affect human interpretation [3, 4, 22, 23]. However, it is crucial to acknowledge that LLMs are not without their own biases, primarily stemming from their training data, which can manifest in unexpected ways [21].

The greater challenge posed by the ASCO guideline for both experts and LLMs underscores the impact of guideline structure and language. The ASCO guideline's broader scope, inclusion of rapidly evolving systemic therapies with potentially less mature evidence bases [24], and more narrative-based recommendations likely contributed to increased interpretative ambiguity [2, 7]. This suggests that guideline clarity and format are pivotal for consistent interpretation by both human users and AI tools.

In this study, we analyzed the vector-based representations of the responses (See Figure S2 in Supplementary Information).. When examining the vectorial approaches, it was observed that in the ASTRO guideline, medical experts, particularly radiation oncologists, tended to provide stronger responses compared to the reference truth. This tendency may be attributed to an overestimation of the SoR and QoE associated with commonly used treatment approaches for brain metastases. In the ASCO guideline, both medical experts and LLMs tended to give stronger responses. This appears to be due to the perception among participants that the SoR and QoE of newer-generation systemic therapies in this guideline are higher than the reference truth, even though the brain metastasis studies supporting these therapies are still immature.

While LLMs outperformed experts overall in convergence and often in accuracy, their performance was not monolithic. Gemini and Deepseek frequently emerged as top performers, particularly in QoE evaluations and the more complex ASCO guideline assessments, hinting at potentially more advanced interpretive reasoning or detail extraction capabilities in these specific models. For instance, Gemini's strong showing in ASTRO QoE (κ=0.595) and Deepseek's leading accuracy (61.53%) and convergence (κ=0.428) for ASCO SoR are noteworthy. The reasons for these inter-LLM differences are likely multifactorial, relating to model architecture, training datasets, and fine-tuning methodologies, warranting further investigation. The relatively lower performance of some models on certain tasks (e.g., Copilot on ASCO SoR/QoE accuracy) also highlights that not all LLMs are equally adept at these specialized tasks.

Our findings align with recent studies demonstrating the potential of LLMs in oncology and medical information synthesis, while also underscoring current limitations. Rydzewski et al. (2024) found variability among LLMs in answering oncology questions, with GPT-4 showing high accuracy, though clinical oncology had lower accuracy rates compared to other areas [25]. Our study, focusing specifically on SoR/QoE within guidelines, provides a more granular assessment. While our LLMs (including GPT-4o) showed high accuracy for structured tasks (ASTRO SoR), the overall accuracy rates for more complex interpretations (ASCO, QoE) were indeed modest, echoing Rydzewski et al.'s findings of challenges in clinical application areas. Similarly, Wilhelm et al. (2023) highlighted variations in LLM performance for generating therapy recommendations, with some models being more prone to harmful or false information [26]. Our study did not assess for harmfulness directly, as we focused on information retrieval against a reference truth. However, the observed inaccuracies, particularly with the ASCO guideline, emphasize the need for caution and expert oversight. The consistent theme across these studies and ours is the burgeoning capability of LLMs alongside a clear need for rigorous validation and understanding of their limitations before widespread clinical integration [20, 27]. Our unique contribution lies in the direct comparison with domain experts on the foundational, yet challenging, interpretation of SoR and QoE within specific, complex oncologic guidelines.

ARTICLE IN PRESS

The superior convergence of LLMs, particularly for the ASTRO guideline, suggests their potential role in a **hybrid intelligence model** for guideline interpretation. In such a model, LLMs could serve as powerful initial interpreters or "second-readers," helping clinicians quickly identify SoR and QoE, thereby reducing cognitive load and flagging potential areas of misinterpretation. This does not imply replacing expert judgment, which remains paramount for applying guideline recommendations to unique patient contexts. Instead, LLMs could support education and training by helping trainees quickly familiarize themselves with guideline structures and evidence grading. They can also support guideline adherence by providing fast, accurate recall of recommendation details, potentially reducing unintentional deviations. Additionally, LLMs can enhance multidisciplinary meetings by offering a consistent baseline interpretation of guidelines for discussion. For guideline development, areas where both LLMs and experts struggle (e.g., ASCO QoE) may indicate sections of guidelines that need greater clarity, more standardized language, or better presentation to reduce ambiguity. Guideline developers could also use LLMs during drafting to proactively identify such ambiguities.

However, the step from accurate guideline interpretation to optimal clinical decision-making for an individual patient remains significant. Guidelines provide a framework, but clinical expertise is essential to tailor recommendations considering patient-specific factors, comorbidities, and values—tasks currently beyond LLM capabilities.

Our findings have direct implications for the structure and presentation of clinical guidelines. First, guideline development committees might consider adopting standardized formatting, such as an algorithmic, structured presentation of recommendations, SoR, and QoE, to enhance clarity and reduce ambiguity. Second, including Machine-Readable Components—explicitly tagged sections for SoR and QoE—with standardized terminology that enables both human and machine interpretation. Third, considering Clarity Testing—assessing draft guidelines with input from human experts and LLMs to identify areas of inconsistent interpretation before publication. And fourth, developing companion digital tools to assist clinicians in navigating complex guidelines more effectively, potentially utilizing LLM capabilities.

These approaches could lead to guidelines that are more consistently interpreted and applied, ultimately improving patient care through more standardized implementation of evidence-based recommendations.

This study has several limitations that should be considered. First, the evaluation relied on a questionnaire format that, while standardized, may not fully reflect the dynamic complexities of real-time clinical decision-making. Additionally, although LLMs outperformed humans in certain classification tasks, this advantage was observed in a text-only, context-free setting that excluded patient-specific factors (e.g., treatment toxicity, comorbidities). Therefore, it does not fully demonstrate clinical decision-making superiority. Second, using guidelines based on expert opinions in the study means there is no definitive ground truth. Therefore, we used the term 'reference truth' throughout the article. Third, the exclusion of four ASCO recommendations due to their "none" SoR or "mixed" QoE might have influenced the overall performance metrics for that guideline, potentially by removing some of the most ambiguous items. While necessary for methodological consistency, this means our findings for ASCO might represent performance on its more clearly defined recommendations. The inherent complexity of the ASCO guideline, with its focus on novel and evolving therapies, also presented a distinct challenge. Fourth, the simplification process of 'Conditional' SoR and 'Expert Opinion' QoE for the ASTRO guideline, while intended for standardization, may have influenced the interpretation accuracy and convergence results for this specific guideline. Fifth, the insufficient consensus among medical

experts, particularly for the ASCO guideline, made detailed inter-specialty comparative analysis challenging, though this finding itself highlights the real-world variability in interpretation. Sixth, our expert panel included neurosurgeons and radiation oncologists; the inclusion of medical oncologists, particularly for the systemically-focused ASCO guideline, would provide a more comprehensive human expert perspective in future studies. Seventh, for recommendations where the reference was 'moderate' or 'intermediate,' the lack of any option other than a correct or near response is recognized as a limitation of the study. This limitation may affect the convergence of evaluation and has been considered when interpreting the results. Finally, the disclaimer used to prompt LLMs, while necessary to elicit responses, might have altered their processing compared to a direct clinical query scenario, and the "black box" nature of LLM reasoning warrants ongoing investigation into their explainability.

This research opens several avenues for future investigation:

Integration with Patient Data: How do LLMs perform when asked to incorporate patient-specific factors with guideline recommendations? Can they accurately adjust interpretations based on relevant clinical variables?
Cross-Guideline Consistency: Can LLMs effectively identify inconsistencies or contradictions between different clinical guidelines addressing the same condition?
Optimal Interface Design: What user interface designs best support human-LLM collaboration in guideline interpretation to maximize benefits and minimize potential risks?
Longitudinal Performance: How do LLMs adapt to evolving guidelines and emerging evidence compared to human experts who may rely on outdated knowledge?
Impact on Clinical Outcomes: Does LLM-assisted guideline interpretation lead to more standardized care and improved patient outcomes?

These research questions will be essential to address as hybrid intelligence approaches move from conceptual frameworks to practical clinical implementation.

Conclusion

This study, the first to systematically compare human expert and Large Language Model (LLM) interpretation of Strength of Recommendation (SoR) and Quality of Evidence (QoE) within key brain metastases guidelines, reveals significant insights into the capabilities of current AI and the challenges of guideline interpretation. Our hybrid analysis demonstrated that leading LLMs can achieve higher accuracy and substantially greater convergence than experienced clinicians administering oncology treatment in identifying these nuanced guideline components, particularly for well-structured guidelines like ASTRO. These findings underscore the considerable potential of LLMs to serve as valuable decision support tools within a hybrid intelligence framework, aiding clinicians in navigating the complexities of treatment guidelines and potentially reducing inter-observer variability in interpretation. While LLMs and human experts alike faced greater challenges with more ambiguously structured guidelines and the intricacies of QoE assessment, the performance of LLMs suggests they could enhance understanding, support educational efforts, and even inform the development of clearer, more consistently interpretable clinical guidelines in the future. Despite the promise, the existing limitations in LLM interpretive reasoning, particularly with less structured information, and the imperative for expert oversight in all clinical applications, remain critical. Continued research is essential to refine LLM capabilities, improve their explainability, and rigorously evaluate their

ARTICLE IN PRESS

integration into clinical workflows to ensure they safely and effectively augment, rather than replace, human expertise in the complex landscape of oncology care.

Data Availability

The guideline recommendations assessed and the evaluation responses from both experts and LLMs are provided in the Supplementary Data (See Supplementary Data 1-4). The original clinical practice guidelines used as reference standards are publicly available through their respective publishing organizations (ASTRO and ASCO).

Code Availability

No custom or external code was used in this study. All analyses were performed using standard, commercially or publicly available software.

Acknowledgements

This research was funded by grants from the National Institutes of Health (NIH), grant numbers R01-HL171376 and U01-CA268808.

ARTICLE IN PRESS

References

1. Eccles, M., et al., *Clinical guidelines. Potential benefits, limitations, and harms of clinical guidelines.* British Medical Journal, 1999.
2. Steinberg, E., et al., *Clinical practice guidelines we can trust*. 2011: national academies press.
3. Cook, D.A., et al., *Practice variation and practice guidelines: attitudes of generalist and specialist physicians, nurse practitioners, and physician assistants.* PloS one, 2018. **13**(1): p. e0191943.
4. Cabana, M.D., et al., *Why don't physicians follow clinical practice guidelines?: A framework for improvement.* Jama, 1999. **282**(15): p. 1458-1465.
5. Shekelle, P.G., et al., *Clinical guidelines: developing guidelines,* Bmj, 1999. **318**(7183): p. 593-596.
6. Djulbegovic, B. and G.H. Guyatt, *Progress in evidence-based medicine: a quarter century on.* The lancet, 2017. **390**(10092): p. 415-423.
7. Grimshaw, J.M. and I.T. Russell, *Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations.* The Lancet, 1993. **342**(8883): p. 1317-1322.
8. Guyatt, G.H., et al., *GRADE: an emerging consensus on rating quality of evidence and strength of recommendations.* Bmj, 2008. **336**(7650): p. 924-926.
9. Zrubka, Z., et al., *The PICOTS-ComTeC framework for defining digital health interventions: an ISPOR special interest group report.* Value in Health, 2024. **27**(4): p. 383-396.
10. Gondi, V., et al., *Radiation therapy for brain metastases: an ASTRO clinical practice guideline.* Practical radiation oncology, 2022. **12**(4): p. 265-282.
11. Vogelbaum, M.A., et al., *Treatment for brain metastases: ASCO-SNO-ASTRO guideline*. 2022, Oxford University Press US.

12. Kresevic, S., et al., *Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework.* NPJ digital medicine, 2024. **7**(1): p. 102.
13. Elhaddad, M. and S. Hamam, *AI-driven clinical decision support systems: an ongoing pursuit of potential.* Cureus, 2024. **16**(4).
14. Shen, J., et al., *Artificial intelligence versus clinicians in disease diagnosis: systematic review.* JMIR medical informatics, 2019. **7**(3): p. e10010.
15. Jha, D., et al., *A Conceptual Framework for Applying Ethical Principles of AI to Medical Practice.* Bioengineering, 2025. **12**(2): p. 180.
16. OpenAI Achiam, J., et al., *GPT-4 technical report. arXiv.* arXiv preprint arXiv:2303.08774, 2023.
17. Team, G., et al., *Gemini: a family of highly capable multimodal models.* arXiv preprint arXiv:2312.11805, 2023.
18. Chen, J. and Q. Zhang, *DeepSeek reshaping healthcare in China's tertiary hospitals.* arXiv preprint arXiv:2502.16732, 2025.
19. Wang, R., J. He, and H. Liang, *Medicine's J.A.R.V.I.S. moment: how DeepSeek-R1 transforms clinical practice.* J Thorac Dis, 2025. **17**(3): p. 1784-1787.
20. Singhal, K., et al., *Large language models encode clinical knowledge.* Nature, 2023. **620**(7972): p. 172-180.
21. Cross, J.L., M.A. Choma, and J.A. Onofrey, *Bias in medical AI: Implications for clinical decision-making.* PLOS Digital Health, 2024. **3**(11): p. e0000651.
22. Bazzari, A.H. and F.H. Bazzari, *Assessing the ability of GPT-4o to visually recognize medications and provide patient education.* Scientific Reports, 2024. **14**(1): p. 26749.
23. Imran, M. and N. Almusharraf, *Google Gemini as a next generation AI educational tool: a review of emerging educational technology.* Smart Learning Environments, 2024. **11**(1): p. 22.
24. Steindl, A. and A.S. Berghoff, *Brain metastases in metastatic cancer: a review of recent advances in systemic therapies.* Expert Review of Anticancer Therapy, 2021. **21**(3): p. 325-339.
25. Rydzewski, N.R., et al., *Comparative evaluation of LLMs in clinical oncology.* Nejm Ai, 2024. **1**(5): p. AIoa2300151.
26. Wilhelm, T.I., J. Roos, and R. Kaczmarczyk, *Large language models for therapy recommendations across 3 clinical specialties: comparative study.* Journal of medical Internet research, 2023. **25**: p. e49324.
27. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nature medicine, 2019. **25**(1): p. 44-56.

Authors and Affiliations

Department of Radiation Oncology, University of Health Science Prof. Dr. Cemil Tascioglu City Hospital, Istanbul, Turkey

Berna Akkus Yildirim, Baver Tutun, Emre Uysal

Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, USA

Gorkem Durak, Ulas Bagci

Department of Radiology, Istanbul University Faculty of Medicine, Istanbul, Turkey

Sukru Mehmet Erturk

Computer Science, Ozyegin University, Istanbul, Turkey

Emre Batuhan Yildirim

ARTICLE IN PRESS

ARTICLE IN PRESS

Corresponding Authors
Berna Akkus Yildirim, email: bernaakkus@yahoo.com

Author Contributions
Study design: BAY, GD, SME, UB. Methodology: BAY, BT, GD. Investigation: BT, EBY, EU. Data Analysis: EU, EBY, BT. Writing-Original Draft: BAY, BT, GD, EBY. Writing-Review&Editing: All Authors. Supervision: BAY, SME, UB.

Competing Interests
The authors declare no competing interests.

Legend for Figure 1: Comparative performance analysis of medical experts (blue bars: Nrs_a1, Nrs_a2, Nrs_r1, Nrs_r2, Rad_a1, Rad_a2, Rad_r1, Rad_r2) and large language models (orange bars: GPT-4o, Gemini, Copilot, DeepSeek) in ASTRO clinical practice guideline assessments. (1A) Accuracy rates for Strength of Recommendation. (1B) Cohen's kappa values for Strength of Recommendation. (1C) Accuracy rates for Quality of Evidence. (1D) Cohen's kappa values for Quality of Evidence. Participant identifiers: Nrs = Neurosurgery (a = attending, r = resident); Rad = Radiation oncology (a = attending, r = resident). Higher kappa values indicate greater Convergence in responses relative to the reference standard.

Legend for Figure 2: Comparative performance analysis of medical experts (blue bars: Nrs_a1, Nrs_a2, Nrs_r1, Nrs_r2, Rad_a1, Rad_a2, Rad_r1, Rad_r2) and large language models (orange bars: GPT-4o, Gemini, Copilot, DeepSeek) in ASCO-SNO-ASTRO clinical practice guideline assessments. (2A) Accuracy rates for Strength of Recommendation. (2B) Cohen's kappa values for Strength of Recommendation. (2C) Accuracy rates for Quality of Evidence. (2D) Cohen's kappa values for Quality of Evidence. Participant identifiers: Nrs = Neurosurgery (a = attending, r = resident); Rad = Radiation oncology (a = attending, r = resident). Higher kappa values indicate greater Convergence in responses relative to the reference standard.

ARTICLE IN PRESS

ARTICLE IN PRESS

ED Summary:
Yildirim et al. compare the accuracy and consistency of medical experts and Large Language Models in interpreting the details of brain metastases treatment guidelines. Large Language Models demonstrate superior accuracy and better convergence than human readers, highlighting their potential to assist in interpreting standardized guidelines.

Peer review information:
Communications Medicine thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

ARTICLE IN PRESS