

Introduction

Artificial intelligence (AI) tools, such as **Large Language Models (LLMs)** are increasingly used to support oncology decision-making¹⁻⁴, yet their agreement with expert oncologist recommendations across clinically and genomically diverse NSCLC scenarios remains uncertain.

Objectives

- Compare AI-generated treatment recommendations with expert thoracic oncologist decisions.
- Evaluate concordance across non-driver-mutated, HER2-mutated, and KRAS G12C-mutated advanced NSCLC.
- Compare general-purpose LLMs with evidence-based clinical decision support platforms.

A substantial head-to-head evaluation of AI systems against a human expert standard had not been performed in this setting.

Methods

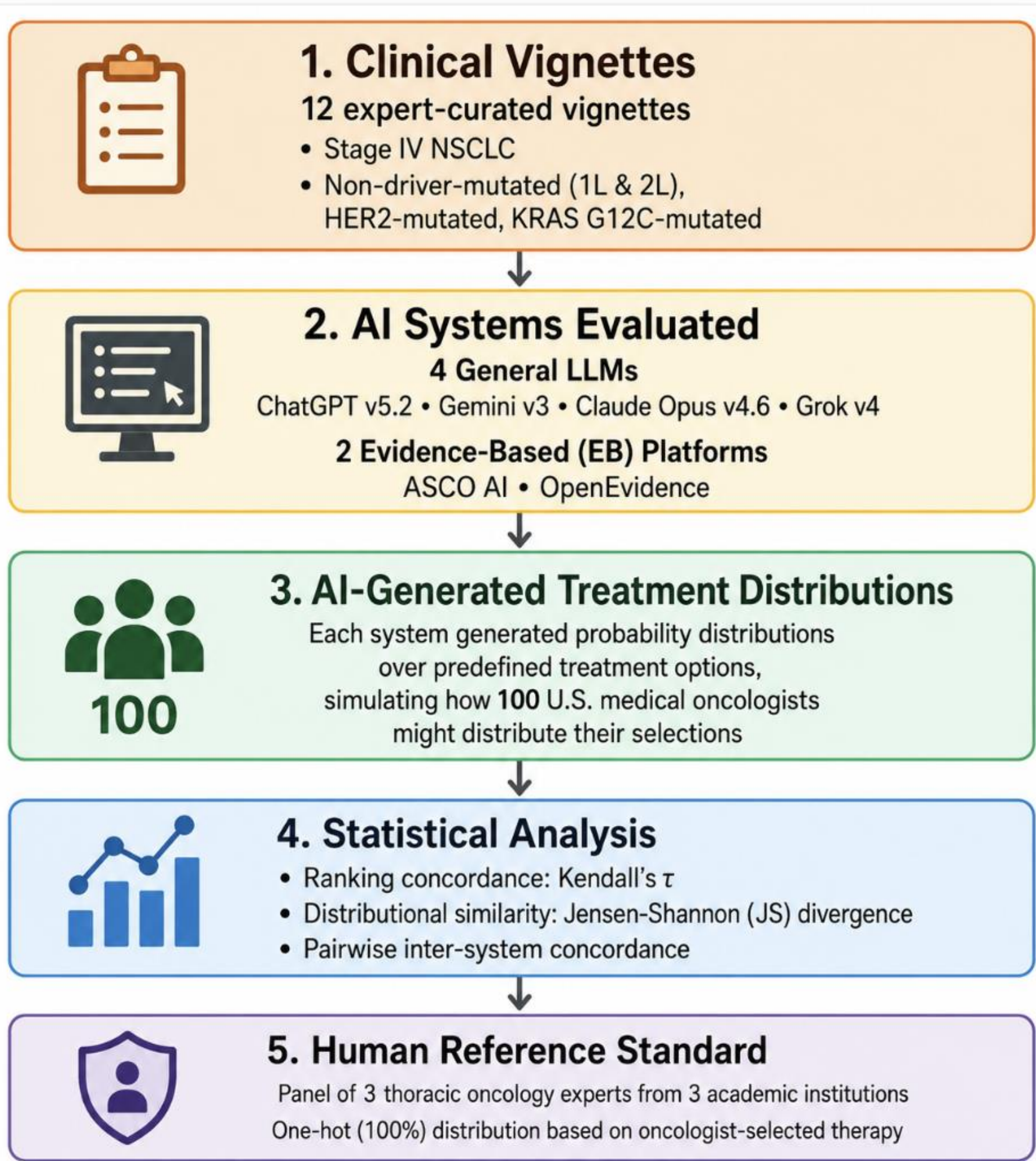


Figure 1. Flowchart of Methodology.

Results



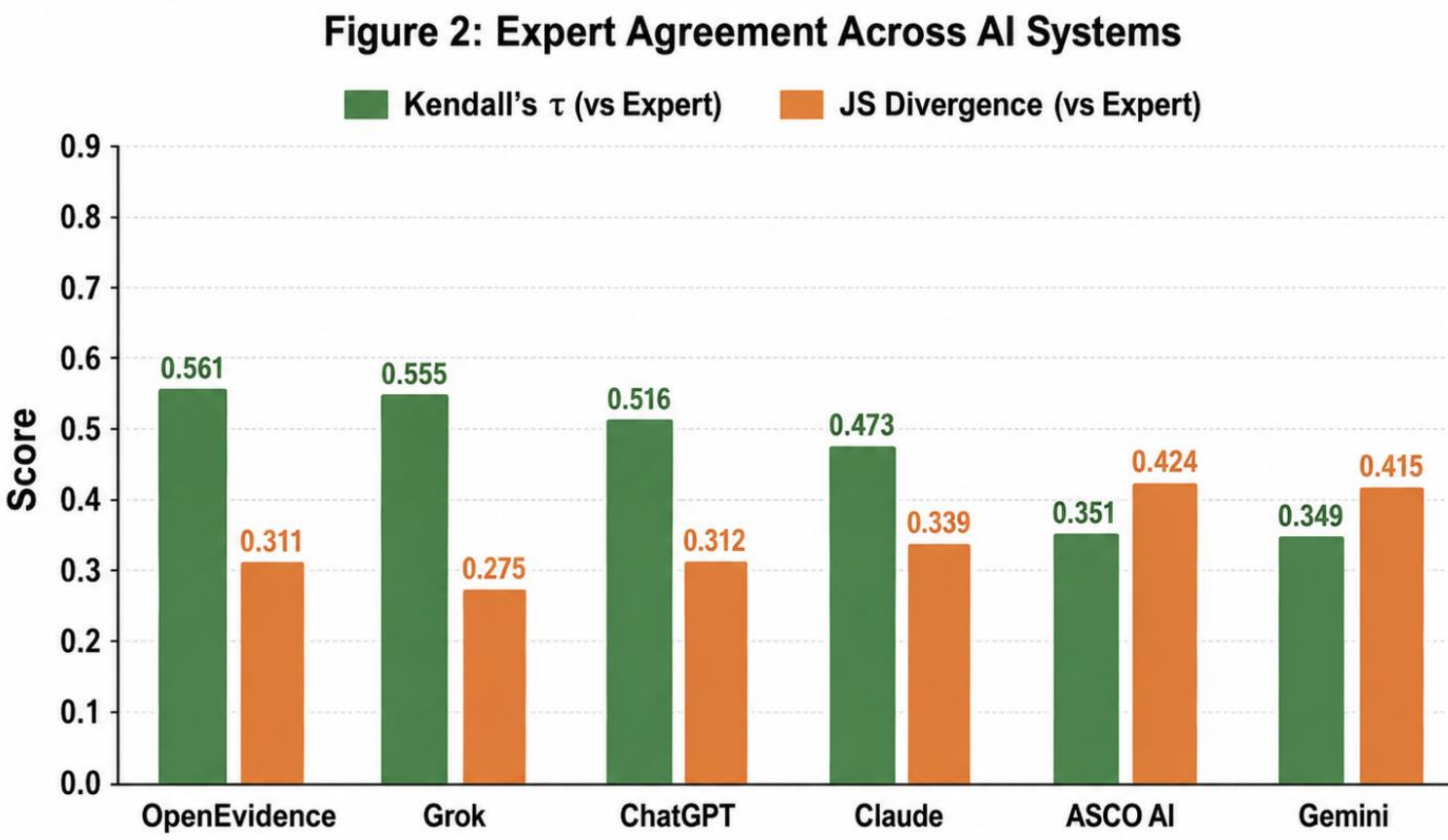
System	Kendall's τ (vs Expert)	JS Divergence (vs Expert)	Highest Pairwise Partner (τ)
OpenEvidence	0.561	0.311	Grok (0.779)
Grok	0.555	0.275	Gemini (0.822)
ChatGPT	0.516	0.312	Claude (0.844)
Claude	0.473	0.339	ChatGPT (0.844)
ASCO AI	0.351	0.424	Gemini (0.876)
Gemini	0.349	0.415	ASCO AI (0.876)

Sorted by Kendall's τ . Higher τ = better ranking agreement; lower JS = less distributional divergence.

Table 1: LLM and Resulting Concordance and Divergence compared to Expert Thoracic Opinion

EXPERT ALIGNMENT

Performance varied substantially across systems. OpenEvidence demonstrated the strongest ranking concordance with expert-selected treatments ($\tau=0.561$), followed closely by Grok ($\tau=0.555$) and ChatGPT ($\tau=0.516$). Claude showed moderate alignment ($\tau=0.473$). ASCO AI and Gemini demonstrated the lowest expert concordance ($\tau=0.351$ and 0.349, respectively).



OpenEvidence demonstrated the strongest ranking concordance with expert-selected treatments, while Grok showed the lowest divergence from expert probability distributions. ASCO AI and Gemini showed the highest inter-system concordance with each other, but not the strongest alignment with oncologist decisions.

PAIRWISE SYSTEM BEHAVIOR

Gemini and ASCO AI showed the highest mutual pair concordance ($\tau=0.876$, JS=0.095), consistent with ASCO AI's Gemini-based architecture; however, this shared behavior did not translate to stronger oncologist alignment. The two evidence-based platforms showed only moderate agreement with each other ($\tau=0.656$, JS=0.114), reflecting divergent architectures. No single system was consistently superior across all vignettes.

Conclusions and Clinical Implications

- Performance varies by platform, not category.**
Across diverse advanced NSCLC scenarios, evidence-based platforms did not consistently outperform general-purpose LLMs.
- Best expert alignment was seen with OpenEvidence and Grok.**
OpenEvidence achieved the highest ranking concordance ($\tau=0.561$). Grok had the lowest divergence (JS=0.275). ChatGPT also demonstrated strong alignment.
- ASCO AI and Gemini showed the weakest oncologist concordance.**
ASCO AI had the highest distributional divergence (JS=0.424), and both ASCO AI and Gemini had the lowest expert concordance overall.

- High agreement between AI systems does not necessarily indicate clinical accuracy**, underscoring the importance of benchmarking against expert-derived reference standards.
- No single AI tool was consistently superior across all clinical scenarios**, limiting reliance on any one platform for treatment selection.
- AI tools should augment—not replace—oncologist judgment**, particularly in complex or biomarker-defined NSCLC.
- Disease- and scenario-specific validation is needed before clinical integration**, with evaluation incorporating both treatment ranking and distributional agreement.

References

- Planchard D, Jänne PA, Cheng Y, Yang JC, Yanagitani N, Kim SW, Sugawara S, et al; FLAURA2 Investigators. Osimertinib with or without chemotherapy in EGFR-mutated advanced NSCLC. *N Engl J Med*. 2023;389(21):1935-1948. doi:10.1056/NEJMoa2306434.
- Felip E, Cho BC, Gutiérrez V, et al. Amivantamab plus lazertinib versus osimertinib in first-line EGFR-mutant advanced non-small-cell lung cancer with biomarkers of high-risk disease: a secondary analysis from MARIPOSA. *Ann Oncol*. 2024;35(9):805-816. doi:10.1016/j.annonc.2024.05.541
- Chen D, Parsa R, Swanson K, et al. Large language models in oncology: a review. *BMJ Oncol*. 2025;4(1):e000759. Published 2025 May 15. doi:10.1136/bmjonc-2025-000759
- Zhou S, Xie W, Li J, et al. Automating expert-level medical reasoning evaluation of large language models. *NPJ Digit Med*. 2025;9(1):34. Published 2025 Dec 6. doi:10.1038/s41746-025-02208-7