

AI at the Tumor Board: Benchmarking Three Mainstream Chatbots Against Mirrors of Medicine Uro-Oncology Experts

Anthony Mina, Rami Halabi, Anthony Kanbar, Joey Khoury, Angelo Assaf, Bertrand Tombal, Raghid El Khoury

Introduction:

Artificial intelligence (AI) is increasingly considered for clinical decision support, yet its role in multidisciplinary tumor boards (MDTs) remains unclear. This study assessed the feasibility, reliability, and quality of large language model (LLM)-generated recommendations for uro-oncology MDT cases using a standardized, guideline-based evaluation framework.

Methods:

Ninety MDT recommendations were generated across 30 uro-oncology cases (prostate, bladder, renal) using three anonymized LLMs (Models A, B, C). All models received identical prompts built on EAU guidelines. Two blinded board certified urologists reviewers independently evaluated each output using a validated 17-item rubric covering accuracy, guideline concordance, comprehensiveness, practicality, justification, communication quality, and ethical awareness. Continuous outcomes were compared using Kruskal–Wallis and Mann–Whitney U tests; binary safety outcomes with χ^2 /Fisher's tests. Inter-rater reliability was assessed using ICC and Cohen's κ .

Results:

Across models, mean clinical accuracy and guideline concordance exceeded 4/5, demonstrating strong evidence-based alignment. Composite quality scores ranged from 77.1 ± 14.7 to 85.6 ± 14.5 and remained consistent after uncertainty adjustment (Model B: 80.7 ± 20.6 ; Model A: 78.8 ± 23.4 ; Model C: 72.8 ± 18.4). Significant inter-model differences were identified for guideline concordance, comprehensiveness, justification, communication clarity, ethical awareness, and both composite indices (all $p \leq 0.05$). Models A and B consistently outperformed C, with Model B achieving the highest communication quality ($p < 0.001$). Safety outcomes—unsafe recommendations (13–17%), critical omissions (17–22%), and fabricated citations (<5%)—did not differ significantly between models. Inter-rater agreement was excellent for composite scores (ICC = 0.94–0.95) and substantial for binary items ($\kappa = 0.58$ –0.72).

Conclusion:

LLMs can generate feasible, reliable, and guideline-concordant MDT recommendations in uro-oncology. Models A and B showed superior performance to C, though variability in contextualization persists. Despite low hallucination rates, expert review remains crucial. This study offers a reproducible, SPIRIT-AI-aligned framework for evaluating AI participation in multidisciplinary cancer care