Supplementary Information

Table of Contents

Table S1. Univariate associations between clinical variables and AI–expert concordance

Variable	Entropy correlation r (95% CI)	p-value	AC1 correlation r (95% CI)	p-value
Age	0.037 (–0.077 to 0.150)	0.523	–0.126 (–0.539 to 0.335)	0.596
ECOG	0.000 (–0.113 to 0.114)	0.994	–	–
Stage	0.045 (–0.068 to 0.158)	0.433	–	–
Treatment line	–0.025 (–0.138 to 0.089)	0.669	–0.033 (–0.469 to 0.415)	0.889
Grade	**0.510 (0.073 to 0.783)**	**0.026**	**–0.573 (–0.815 to –0.161)**	**0.010**
Somatostatin receptor status	Kruskal–Wallis	0.195	Kruskal–Wallis	0.734
Primary tumour site	Kruskal–Wallis	0.577	Kruskal–Wallis	0.192
Uncertainty (experts)	0.429 (–0.017 to 0.732)	0.059 †	**–0.595 (–0.821 to –0.207)**	**0.006**

Abbreviations: Pearson correlation for continuous variables, Kruskal–Wallis test for categorical. † = trend ($0.05 \leq p < 0.10$). Bold values = statistically significant ($p<0.05$).

Table S2. Sensitivity analysis: intra-run reproducibility of AI models and inter-expert variability (Bootstrap analysis, 95% CI)

Outcome	Model	Mean AC1 (95% CI)	Mean agreement % (95% CI)
Therapy	GPT	0.843 (0.687–0.961)	80.1 (60–95)
	GPTs	0.766 (0.613–0.919)	69.9 (50–90)
	RAG	1.000 (1.000–1.000)	100.0 (100–100)
	Experts	0.427 (0.294–0.561)	6.0 (0.0–17.6)
Tests	GPT	0.487 (0.270–0.696)	45.2 (25–70)
	GPTs	0.370 (0.163–0.574)	34.5 (15–55)
	RAG	0.946 (0.817–1.000)	95.0 (85–100)
	Experts	0.443 (0.313–0.575)	9.9 (0.0–25.0)
Uncertainty	GPT	0.835 (0.642–0.966)	79.8 (60–95)
	GPTs	1.000 (1.000–1.000)	100.0 (100–100)
	RAG	0.875 (0.670–1.000)	90.3 (75–100)
	Experts	0.239 (0.103–0.394)	5.1 (0.0–15.8)
Completeness	GPT	1.000 (1.000–1.000)	100.0 (100–100)
	GPTs	1.000 (1.000–1.000)	100.0 (100–100)
	RAG	1.000 (1.000–1.000)	100.0 (100–100)
	Experts	0.164 (0.095–0.231)	0.0 (0.0–0.0)
Cost	GPT	0.287 (0.126–0.451)	20.2 (5–40)
	GPTs	0.309 (0.151–0.471)	20.2 (5–40)
	RAG	0.925 (0.811–1.000)	90.3 (75–100)
	Experts	0.323 (0.247–0.403)	0.0 (0.0–0.0)

Abbreviations: Mean AC1 and mean agreement with 95% bootstrap confidence intervals. AI rows report intra-run reproducibility; Expert rows report inter-expert variability.