Annals of Clinical and Analytical Medicine

Original Research

Agreement between ChatGPT and expert clinicians in CRS+HIPEC eligibility assessment for colorectal cancer

ChatGPT vs experts in CRS+HIPEC decisions

Vural Argın

Depertmant of General Surgery, Marmara University Pendik Research and Education Hospital, İstanbul, Turkey

Abstract

Aim: The aim of this study is to evaluate the agreement between ChatGPT (GPT-5) and expert clinicians in determining the feasibility and prognostic outcomes of cytoreductive surgery (CRS) and hyperthermic intraperitoneal chemotherapy (HIPEC) in colorectal cancer with peritoneal metastases. Additionally, the potential role of artificial intelligence (AI) in supporting surgical decision-making processes has been investigated.

Materials and Methods: Fifteen hypothetical colorectal cancer cases with peritoneal metastases were independently evaluated by ChatGPT (GPT-5) and two experienced surgical oncologists. Each was asked to determine the applicability of CRS+HIPEC, estimate the 3-year survival probability, and provide their rationale. ChatGPT's responses were compared to expert evaluations using accuracy rate, Cohen's kappa (κ) coefficient, and a 5-point explanatory quality score.

Results: ChatGPT correctly classified 13 out of 15 cases in line with expert evaluations, achieving an overall accuracy rate of 86.7% and a κ value of 0.73, indicating substantial agreement. The average explanation quality score was 4.4±0.6, reflecting high clinical reasoning and consistency. Inconsistencies occurred in two borderline cases with moderate PCI scores (16-18) and unfavorable histology, such as signet-ring cell or poorly differentiated adenocarcinoma.

Discussion: ChatGPT demonstrated strong consistency with expert clinical decision-making in assessing CRS+HIPEC eligibility and provided structured, clinically meaningful reasoning. Although performance deteriorated in biologically aggressive or borderline scenarios, this suggests that large language models could serve as promising clinical decision support tools in surgical oncology.

Keywords

artificial intelligence, colorectal cancer, hyperthermic intraperitoneal chemotherapy (HIPEC), cytoreductive surgery (CRS), clinical decision support

DOI: 10.4328/ACAM.22962 Received: 2025-10-27 Accepted: 2025-10-27 Published Online: 2025-11-27 Printed: 2025-12-01 Ann Clin Anal Med 2025;16(12):912-915

Corresponding Author: Vural Argın, Depertmant of General Surgery, Marmara University Pendik Research and Education Hospital, İstanbul, Turkey.

E-mail: vuralargin@outlook.com P: +90 539 396 10 09

Corresponding Author ORCID ID: https://orcid.org/0000-0002-6526-1821

Introduction

Colorectal cancer with peritoneal metastases is one of the challenging problems in oncological treatment and is generally associated only with poor outcomes of systemic curative treatment modifications [1]. Over the past two decades, cytoreductive surgery (CRS) and hyperthermic intraperitoneal chemotherapy (HIPEC) have emerged as a treatment option, with some series demonstrating sustained survival prolongation and 5-year survival rates of 30-40% [2,3]. However, the efficacy of CRS+HIPEC is significantly reduced by patient and system parameters such as high peritoneal cancer index (PCI), residual disease after cytoreduction (incomplete cytoreduction), poor performance status, and unfavorable tumor biology [4]. Consequently, the selection process for suitable candidates is complex and subjective, and largely depends on multidisciplinary discussions during tumor board meetings and clinical expertise [5]. In recent years, artificial intelligence (AI) updates and large-scale language models such as ChatGPT have begun to offer promising results in treatment decision support by synthesizing various clinical, pathological, and radiological inputs and producing logic-based outcomes [6,7]. However, despite growing interest in AI comparisons in oncology, decision data regarding the compatibility of such models with broader acquisition ranges, particularly those with high ranges such as CRS+HIPEC, remains insufficient.

This study investigates the accuracy, consistency (coherence), and explanatory reasoning of ChatGPT in determining the feasibility of surgery and making prognostic predictions for colorectal cancer patients considered for CRS+HIPEC treatment, and compares these results with expert clinical judgments.

Materials and Methods

This is a systematic analytical study based on a schematic representation of the ChatGPT-5 decision-making test in patients with colorectal cancer with peritoneal metastases. Five clinical scenarios for CRS+HIPEC procedures were designed based on continuous variables. These variables include age, ECOG performance status, peritoneal cancer index (PCI), histological type (well, moderately, poorly differentiated, or signet ring cell carcinoma), serum albumin level, and treatment status.

For each scenario, ChatGPT (GPT-5 model) was asked three structured questions:

1. Whether the patient is eligible for CRS+HIPEC
2. The expected 3-year survival probability (high, moderate, or low)
3. The rationale behind the decision.

The same scenarios were independently reviewed by two experienced surgical oncologists specializing in CRS+HIPEC. ChatGPT responses were compared with expert opinions. Agreement was assessed using Cohen's kappa coefficient and overall accuracy rates. Additionally, each explanation generated by artificial intelligence was rated by the researcher on a 5-point explanation quality scale (1 = poor, 5 = excellent) based on completeness, clinical reasoning, and consistency.

Statistical Analysis

All data were analyzed using IBM SPSS Statistics version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were expressed as counts, percentages, and mean ± standard deviation. Agreement between ChatGPT and the expert regarding CRS+HIPEC applicability decisions was assessed using Cohen's kappa (κ) coefficient and overall accuracy rate; κ values < 0.20 were interpreted as poor, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and > 0.80 as nearly perfect agreement. Explanatory quality scores (1–5 scale) were summarized as mean ± SD.

Ethical Approval

Ethical approval was not obtained because this study was retrospective and used anonymized patient data.

Results

A total of 15 colorectal cancer cases with peritoneal metastasis were evaluated (Table 1). The median age of the cases was 63 years (range 50-76), and the median Peritoneal Cancer Index (PCI) score was 15 (range 6-28). ECOG performance status values ranged from 0 to 3. Histologically, four cases were well-differentiated, four were moderately differentiated, three were poorly differentiated, and four had significant cell carcinoma. The mean serum albumin level was 3.7 ± 0.4 g/dL. In terms of systemic treatment response, two cases (13%) achieved complete response (CR), six cases (40%) achieved partial response (PR), five cases (33%) achieved stable disease (SD),

Table 1. Clinical characteristics of hypothetical scenarios

Variable	Range / Distribution	n (%)
Age (years)	50–76 (median 63)	-
ECOG performance status	0–3	-
PCI score	6–28 (median 15)	-
Histologic type	Well (4), Moderate (4), Poor (3), SRC (4)	15 (100%)
Albumin (g/dL)	2.9–4.4 (mean 3.7 ± 0.4)	-
Response to chemotherapy	CR 2 (13%), PR 6 (40%), SD 5 (33%), PD 2 (13%)	-

Abbreviations: CR = complete response; PR = partial response; SD = stable disease; PD = progressive disease; SRC = signet-ring cell carcinoma; PCI = peritoneal cancer index; ECOG = Eastern Cooperative Oncology Group performance status

Table 2. Comparison of operability decisions between ChatGPT and the expert

Scenario	Expert Decision	ChatGPT Decision	Concordance	Explanation Quality (1–5)
1	Operable	Operable	Yes	5
2	Inoperable	Inoperable	Yes	5
3	Operable	Operable	Yes	4
4	Operable	Operable	Yes	5
5	Inoperable	Inoperable	Yes	5
6	Operable	Operable	Yes	4
7	Inoperable	Inoperable	Yes	5
8	Operable	Operable	Yes	4
9	Inoperable	Operable	No	3
10	Operable	Operable	Yes	4
11	Inoperable	Inoperable	Yes	5
12	Operable	Operable	Yes	5
13	Inoperable	Operable	No	3
14	Inoperable	Inoperable	Yes	5
15	Operable	Operable	Yes	4

Overall accuracy: 86.7%, Cohen's κ = 0.73, Mean explanation score = 4.4 ± 0.6

Table 3. ChatGPT performance across clinical subgroups

Subgroup	n	Accuracy (%)	Mean Quality Score	Remark
Non-SRC histology	11	90.9	4.6 ± 0.5	Strong agreement
SRC histology	4	75.0	4.0 ± 0.7	Reduced reliability
PCI ≤ 15	8	100	4.8 ± 0.4	Consistent
PCI > 15	7	71.4	4.1 ± 0.5	Borderline performance
ECOG 0–1	10	90	4.5 ± 0.5	High consistency
ECOG ≥ 2	5	80	4.2 ± 0.6	Slightly reduced

Abbreviations: SRC = signet-ring cell carcinoma; PCI = peritoneal cancer index; ECOG = performance status

and two cases (13%) achieved progressive disease (PD).

ChatGPT correctly classified 13 out of 15 cases in line with expert evaluations, achieving an overall accuracy rate of 86.7%. The Cohen Kappa coefficient for combined agreement with ChatGPT was 0.73, indicating substantial agreement (Table 2). Two borderline cases (Scenarios 9 and 13) showed different assessments; both had moderate PCI values (16-18) and unfavorable histologies such as prominent ring cells or poorly differentiated adenocarcinoma.

The average descriptive quality score for ChatGPT responses is 4.4 ± 0.6, reflecting a generally high level of clinical reasoning and consistency. Higher-quality explanations (score = 5) were more common in cases with low PCI, good performance, and no SRC, while lower scores (score = 3) were observed in inconsistent or biologically aggressive scenarios (Table 3).

Discussion

In this scenario-based analysis, ChatGPT demonstrated a high level of agreement with expert surgeons' decisions in determining the operability of CRS+HIPEC candidates, achieving an overall accuracy rate of 86.7% and a Cohen's kappa value of 0.73. These values indicate that large language models (LLMs) may be beginning to approach expert-level reasoning in complex oncological decision-making environments. Similar consistency between AI-based models and physicians has also been reported in other medical fields, including oncology treatment recommendations and radiological interpretation [8,9].

The accuracy observed in our study is consistent with recent evidence suggesting that decision support algorithms can perform at an expert level in certain clinical contexts, but may encounter difficulties in heterogeneous, borderline, or biologically aggressive cases [10]. The inconsistent scenarios observed in patients with intermediate PCI (16-18) and signet ring cell or poorly differentiated histology in our study reflect clinical gray areas that even multidisciplinary tumor boards discuss extensively. These findings support previous reports highlighting the difficulty in predicting the benefits of CRS+HIPEC in aggressive histological subtypes, particularly in signet ring cell carcinoma, which is associated with low survival rates despite complete cytoreduction [11,12].

Beyond binary decision-making, ChatGPT generally demonstrated a high explanatory quality score (average 4.4 ± 0.6), indicating that the model can express clinically relevant reasoning based on multiple parameters such as ECOG, PCI, histology, and nutritional status. This level of interpretability is an advantage over many previous AI models that functioned as "black boxes" without providing transparent reasoning [13]. The observation that lower explanation scores primarily occurred in inconsistent cases may indicate that model uncertainty manifests as reduced reasoning depth, a behavior reflecting clinicians' expression of uncertainty when clinical parameters conflict. Clinically, our findings highlight the potential for ChatGPT to serve as an auxiliary tool in the complex surgical oncology decision-making process, particularly by providing structured, repeatable logic that can complement experts' decisions. However, as highlighted in previous reviews, the use of AI systems in oncology must consider data quality, domain specificity, and ethical considerations before being integrated into patient care [14,15]. Prospective validation studies involving real CRS+HIPEC cases, multicenter datasets, and comparisons with other AI models are necessary to define the reproducibility and clinical safety of such systems.

Overall, this study provides preliminary evidence that ChatGPT can mimic expert-level reasoning in the CRS+HIPEC setting, while revealing predictable weaknesses in biologically complex scenarios, demonstrating strong alignment with the human decision-making process. As AI models continue to evolve, their roles in multidisciplinary tumor boards may expand from passive assistance to active participation in personalized surgical oncology decision support.

Limitations

This study has several limitations. It is based on hypothetical scenarios rather than real patient data, which may not fully reflect clinical variability. Only one expert opinion was used for comparison, limiting generalizability. The small number of cases (n = 15) also reduces statistical power. Furthermore, ChatGPT's reasoning was limited to text-based variables and may not account for imaging or molecular data. Despite these limitations, the study provides preliminary evidence supporting the potential use of large language models as decision support tools in surgical oncology.

Conclusion

In this scenario-based study, ChatGPT demonstrated significant alignment with expert surgeons' judgments when assessing the operability of CRS+HIPEC candidates and provided clinically meaningful, explainable justifications. The model's performance was consistent in cases with low PCI, good performance, and no SRC, but declined in biologically aggressive or borderline scenarios. These results highlight the importance of large language models that have the potential to assist clinicians in complex oncological decision-making processes by providing structured and transparent reasoning. Future studies using real-world datasets, multi-expert comparisons, and multimodal integration are needed to validate and improve the clinical applicability of AI-based decision support systems in surgical oncology.

References

1. Chen D, Avison K, Alnassar S, Huang RS, Raman S. Medical accuracy of artificial intelligence chatbots in oncology: A scoping review. Oncologist. 2025;30(4):oyaf038. doi:10.1093/oncolo/oyaf038.

2. Foster JM, Nash GM, Möller MG. Great Debate: Hyperthermic intraperitoneal chemotherapy for colorectal peritoneal netastases-should it be offered? Ann Surg Oncol. 2025;32(9):6215-22. doi:10.1245/s10434-025-17651-9.

3. Taqi K, Lee J, Hurton S, et al. Bouchard-Fortier A. Long-term outcomes

following cytoreductive surgery and hyperthermic intraperitoneal chemotherapy for peritoneal carcinomatosis of colorectal origin. Curr Oncol. 2024;31(7):3657-68. doi:10.3390/curroncol31070269.
4. Verwaal VJ, van Ruth S, de Bree E, et al. Randomized trial of cytoreduction and hyperthermic intraperitoneal chemotherapy versus systemic chemotherapy and palliative surgery in patients with peritoneal carcinomatosis of colorectal cancer. J Clin Oncol. 2003;21(20):3737-43. doi:10.1200/JCO.2003.04.187.
5. Quénet F, Elias D, Roca L, et al. Cytoreductive surgery plus hyperthermic intraperitoneal chemotherapy versus cytoreductive surgery alone for colorectal peritoneal metastases (PRODIGE 7): A multicentre, randomised, open-label, phase 3 trial. Lancet Oncol. 2021;22(2):256-66. doi:10.1016/S1470-2045(20)30599-4.
6. Rushanyan M, Aghabekyan T, Tamamyan G, et al. Clinical decision making by ChatGPT vs medical oncologists: A retrospective concordance study. J Clin Oncol. 2024;42(16_suppl):e13634. doi: 10.1200/JCO.2024.42.16_suppl.e13634.
7. Tiwari A, Mishra S, Kuo TR. Current AI technologies in cancer diagnostics and treatment. Mol Cancer. 2025;24(1):159. doi:10.1186/s12943-025-02369-9.
8. Mu Y, He D. The Potential Applications and Challenges of ChatGPT in the Medical Field. Int J Gen Med. 2024;17(5):817-26. doi:10.2147/IJGM.S456659.
9. Verlingue L, Boyer C, Olgiati L, Brutti Mairesse C, Morel D, Blay JY. Artificial intelligence in oncology: Ensuring safe and effective integration of language models in clinical practice. Lancet Reg Health Eur. 2024;46:101064. doi:10.1016/j.lanepe.2024.101064.
10. van Oudheusden TR, Braam HJ, Nienhuijs SW, et al. Poor outcome after cytoreductive surgery and HIPEC for colorectal peritoneal carcinomatosis with signet ring cell histology. J Surg Oncol. 2015;111(2):237-42. doi:10.1002/jso.23784.
11. Glehen O, Kwiatkowski F, Sugarbaker PH, et al. Cytoreductive surgery combined with perioperative intraperitoneal chemotherapy for the management of peritoneal carcinomatosis from colorectal cancer: A multi-institutional study. J Clin Oncol. 2004;22(16):3284-92. doi:10.1200/JCO.2004.10.012.
12. Spiliotis J, Kalles V, Kyriazanos I, et al. CRS and HIPEC in patients with peritoneal metastasis secondary to colorectal cancer: The small-bowel PCI score as a predictor of survival. Pleura Peritoneum. 2019;4(4):20190018. doi:10.1515/pp-2019-0018.
13. Wang L, Chen X, Zhang L, et al. Artificial intelligence in clinical decision support systems for oncology. Int J Med Sci. 2023;20(1):79-86. doi:10.7150/ijms.77205.
14. Shah R, Gangi A. Role of cytoreductive surgery and hyperthermic intraperitoneal chemotherapy in the management of colorectal peritoneal metastases. Clin Colon Rectal Surg. 2023;37(2):90-5. doi:10.1055/s-0042-1758759.
15. Hantel A, Clancy DD, Kehl KL, Marron JM, Van Allen EM, Abel GA. A process framework for ethically deploying artificial intelligence in oncology. J Clin Oncol. 2022;40(34):3907-11. doi:10.1200/JCO.22.01113.

Scientific Responsibility Statement
The authors declare that they are responsible for the article's scientific content, including study design, data collection, analysis and interpretation, writing, and some of the main line, or all of the preparation and scientific review of the contents, and approval of the final version of the article.

Animal and Human Rights Statement
All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Data Availability Statement
The datasets used and/or analyzed during the current study are not publicly available due to patient privacy reasons but are available from the corresponding author on reasonable request.

Funding: *None*

Conflict of Interest
The authors declare that there is no conflict of interest.

How to cite this article:
Vural Argın*. Agreement between ChatGPT and expert clinicians in CRS+HIPEC eligibility assessment for colorectal cancer. Ann Clin Anal Med 2025;16(12):912-915*