Artificial Intelligence (ChatGPT-4o) in Adjuvant Treatment Decision-Making for Stage II Colon Cancer: A Comparative Analysis with Clinician Recommendations and NCCN/ESMO Guidelines

Fatih KUS[1], Elvin CHALABIYEV[1], Hasan Cagri YILDIRIM[1], Ilgin KOC KUS[1], Firat SIRVAN[2], Omer DIZDAR[1], Suayib YALCIN[1]

[1] Hacettepe University, Faculty of Medicine, Department of Medical Oncology
[2] Hacettepe University, Faculty of Medicine, Department of Internal Medicine

ABSTRACT

The role of artificial intelligence (AI) in oncology decision-making is rapidly expanding, yet its concordance with clinician recommendations and established guidelines remains unclear. This study evaluates the agreement between AI-generated adjuvant therapy recommendations for stage II colon cancer, clinician decisions, and international guidelines (NCCN, ESMO). We conducted a retrospective comparative analysis of 197 stage II colon cancer patients treated at Hacettepe University between 2014 and 2023. AI-generated recommendations (ChatGPT-4o) were compared with clinician decisions and NCCN/ESMO guidelines. Concordance rates were analyzed using Cohen's kappa and McNemar's test. Clinician adherence was highest with NCCN (89.8%) and ESMO (84.8%) guidelines. AI recommendations showed moderate agreement with clinician decisions (65.0%, κ= 0.47). Statistically significant differences were observed between AI and clinical practice ($p< 0.001$), suggesting AI's more conservative approach. While AI demonstrates potential as a clinical decision-support tool, its moderate alignment with real-world decisions highlights the need for further refinement. Future improvements in AI interpretability, real-world validation, and clinician-AI collaboration are essential for its effective integration into oncology practice.

Keywords: Artificial Intelligence, Colon Cancer, Adjuvant treatment

INTRODUCTION

Colon cancer treatment has evolved significantly over the past decades, with advancements in surgical techniques, radiotherapy, and systemic therapies improving patient outcomes.[1] Adjuvant therapy decisions in stage II colon cancer remain a subject of debate, as treatment recommendations depend on multiple patient-specific factors, including tumor characteristics, molecular markers, and high-risk features.[2,3] Clinical decision-making regarding adjuvant treatment is traditionally guided by expert oncologists and evidence-based recommendations from international guidelines such as those from the National Comprehensive Cancer Network (NCCN)[4] and the European Society for Medical Oncology (ESMO).[5] However, treatment decisions are often complex, requiring the integration of multiple patient-specific factors, including tumor stage, histopathological features, molecular markers, and overall patient condition.[6] RAS mutations, particularly KRAS mutations, are common genetic alterations in colorectal cancer (CRC) and significantly influence tumor biology.[7]

Recent studies indicate that activating RAS mutations impair EGFR signaling via the RAS/RAF/MAPK pathway, reducing the effectiveness of EGFR inhibitors. Specifically in microsatellite stable (MSS) tumors, KRAS mutations are linked to shorter recurrence-free and overall survival, highlighting their prognostic and predictive importance in early-stage CRC and emphasizing the need for routine molecular profiling in clinical practice.[7]

The rapid emergence of artificial intelligence in oncology has sparked interest among clinicians regarding its potential role in treatment decisions.[8,9] Clinicians increasingly rely on machine learning and deep learning models to predict treatment outcomes and personalize therapeutic approaches, potentially reducing variability in clinical decision-making.[10] AI systems analyze extensive clinical data, helping clinicians recognize subtle patterns that might otherwise go unnoticed, thereby supporting more personalized treatment recommendations.[11]

Recent studies have specifically assessed the diagnostic capabilities of AI in colon cancer using advanced techniques such as WGCNA and LASSO algorithms, achieving promising accuracy in distinguishing cancerous from healthy tissues.[12]

However, despite promising preliminary results, clinicians continue to debate the real-world applicability and reliability of AI-generated treatment recommendations.[13] Understanding how AI recommendations align or diverge from those made by clinicians and NCCN/ESMO guidelines is essential to evaluating their potential role in routine clinical practice.

Clarifying the areas of agreement and discrepancy among AI, clinician decisions, and guidelines is vital to understand whether AI can reliably support oncologists in daily practice. Therefore, this study aims to comparatively analyze AI-generated adjuvant treatment recommendations for colon cancer against clinician decisions and NCCN/ESMO guidelines. By evaluating concordance rates, potential discrepancies, and the clinical reasoning behind variations, we seek to clarify whether AI can effectively complement clinical decision-making. Additionally, we aim to identify factors influencing differences among decision-making sources, including variations in AI training data, clinician expertise, and guideline interpretation.

In conducting this comparative analysis, we aim to better understand if and how AI recommendations can realistically complement the clinical judgment of oncologists, highlighting both potential advantages and important areas needing further development.

MATERIALS AND METHODS

Study Design and Data Collection

This study was designed as a retrospective comparative analysis evaluating adjuvant treatment recommendations for colon cancer generated by an Artificial Intelligence (AI) model (ChatGPT-4o), clinician decisions, and established international oncology guidelines (NCCN and ESMO). Patient data were retrospectively reviewed from electronic medical records of stage II colon cancer patients treated at Hacettepe University over a six-year period (January 2014 to January 2023). The collected dataset included detailed demographic information, tumor characteristics (including tumor stage, histopathological features, molecular markers such as mismatch repair status), treatment details (including surgical outcomes and administered therapies), and follow-up data.

AI Model and Treatment Recommendation Analysis

The AI model employed in this study was ChatGPT-4o, a large language model equipped with advanced reasoning capabilities and trained on extensive medical literature and clinical datasets. For generating AI-driven treatment recommendations, structured clinical scenarios were systematically created, incorporating predefined clinical parameters such as tumor stage (T-stage), molecular markers (e.g., mismatch repair status), surgical outcomes, and patient demographics (age, gender, presence of comorbidities). These structured scenarios were inputted into ChatGPT-4o, allowing the AI to produce recommendations for or against adjuvant chemotherapy based on current clinical evidence and probabilities derived from its training datasets.

To evaluate the practical validity of AI-generated suggestions, we directly compared these with actual clinical decisions made by oncologists documented in patient medical records. Additionally, the concordance of AI recommendations was evaluated against the latest NCCN and ESMO guidelines, using guideline-specific criteria for adjuvant therapy eligibility.

In situations where AI recommendations differed from clinician decisions or guideline criteria, detailed subgroup analyses were conducted to identify specific reasons, including differences in tumor assessment or clinical judgment.

The study was conducted with strict adherence to ethical considerations and was approved by the ethics committee of Hacettepe University, with decision number: SBA 24/593 and date: 21.05.2024.

Statistical Analysis

Statistical analyses were performed using SPSS software version 27. Cohen's kappa coefficient was calculated to quantify the agreement levels between AI-generated recommendations, clinician decisions, and NCCN/ESMO guideline recommendations. Descriptive statistics (means, standard deviations, frequencies, and percentages) were used to summarize demographic variables and clinical characteristics of the patient population. Subgroup analyses were conducted to examine differences in recommendation concordance based on specific patient characteristics (e.g., tumor stage, lymphovascular or perineural invasion, mismatch repair status). Differences among recommendation sources (AI, clinicians, NCCN, ESMO) were statistically evaluated using McNemar's test, and statistical significance was defined as a p-value <0.05.

RESULTS

We retrospectively analyzed clinical data from 197 stage II colon cancer patients treated at our institution. Patients had a mean age of 61.2 ± 12.9 years, with a male predominance (65.5%). The majority of patients presented with T3 tumors (71.6%), while T4 tumors accounted for 24.9%, and only a small subset (3.6%) were T2. Lymphovascular invasion (LVI) was observed in 14.7% of patients, and perineural invasion (PNI) in 20.8%. Adjuvant therapy was administered to 30.5% of patients, while 69.5% did not receive adjuvant therapy. Additional patient characteristics, including mismatch repair (MMR) status, are detailed in Table 1.

Table 1. Patients characteristics

Characteristic	n (%)
Age — yr	
Mean±SD	61.22±12.85
Sex — no (%)	
Male	129 (65.5)
Female	68 (34.5)
T Stage — no (%)	
T2	7 (3.6)
T3	141 (71.6)
T4	49 (24.9)
PNI — no (%)	
Absent	173 (87.8)
Present	24 (12.2)
LVI — no (%)	
Absent	168 (85.3)
Present	29 (14.7)
MMR Status — no (%)	
Unknown	77 (39.1)
Deficient	15 (7.6)
Proficient	105 (53.3)
Adjuvant Treatment — no. (%)	
No	137 (69.9)
Yes	59 (30.1)

Clinicians demonstrated high adherence to established guidelines, with the highest concordance observed with NCCN (89.8%) and slightly lower adherence to ESMO guidelines (84.8%). Recommendations provided by the AI model showed moderate alignment (65.0%) with clinical decisions, highlighting some differences in clinical judgment.

Statistical agreement measured by Cohen's kappa was substantial between clinician decisions and NCCN (κ= 0.73) and ESMO (κ= 0.62) guidelines. In contrast, agreement with AI recommendations was moderate (κ= 0.47), highlighting differences in clinical interpretation (Table 2).

Table 2. Agreement between Real-World Adjuvant Therapy and AI, NCCN, and ESMO Guidelines

Comparison	Agreement Rate (%)	Cohen's Kappa (κ)	McNemar's Test (p)
NCCN vs Real-World Treatment	89.8	0.73	<0.001
ESMO vs Real-World Treatment	84.8	0.62	<0.001
AI vs Real-World Treatment	65.0	0.47	<0.001
AI vs NCCN			<0.001
NCCN vs ESMO			0.04

Using McNemar's test, we found statistically significant differences among recommendation sources, with AI-generated suggestions notably differing from clinician decisions ($p< 0.001$). Similarly, even the well-established NCCN and ESMO guidelines showed statistically significant differences from clinical practice (both $p< 0.001$). Notably, a smaller yet statistically significant variation ($p= 0.04$) between NCCN and ESMO guideline recommendations was also identified, reflecting subtle differences in their clinical risk assessments. ($p= 0.04$).

DISCUSSION

In this study, we compared AI-generated (ChatGPT-4o) adjuvant therapy recommendations with decisions made by clinicians and established guidelines (NCCN and ESMO) for stage II colon cancer. We found that NCCN guidelines had the highest agreement with real-world clinical decisions (89.8%), followed by ESMO (84.8%). In contrast, AI recommendations aligned only moderately (65.0%) with clinician choices, highlighting key differences in decision-making. These findings emphasize the difficulties of integrating AI into oncology practice, where clinical expertise and individualized patient assessment remain essential.

Recent literature highlights the increasing use of artificial intelligence (AI)-based models as clinical decision support systems in oncology. Carl et al. performed a systematic review and meta-analysis demonstrating substantial methodological heterogeneity and performance variability among large language models (LLMs) utilized for clinical oncology questions.[14] They emphasize the urgent need for developing standardized methodological approaches to ensure reliable integration of AI into routine clinical oncology practice.

Previous studies have consistently demonstrated the potential benefits of integrating AI into oncology, including improvements in diagnostic accuracy, personalized treatment planning, and clinical decision-making efficiency. Hassan et al. emphasized the transformative impact of AI in chemotherapy development and cancer treatment, particularly highlighting the enhanced predictive power of deep learning models in personalizing cancer therapies and improving clinical outcomes.[9] Similarly, Khalifa and Albadawy identified significant contributions of AI in various key domains, notably diagnosis, prognosis, risk assessment, and treatment response, underscoring AI's potential to optimize clinical prediction and patient outcomes10.

Similarly, Nabieva et al. evaluated ChatGPT's alignment with expert recommendations for early breast cancer treatment from the St. Gallen International Consensus.[15] They observed moderate agreement between AI-generated recommendations and expert panel decisions, with notably higher agreement in specific clinical scenarios. Their findings suggest potential for AI integration into clinical guidelines but indicate the necessity for significant refinements before routine clinical adoption.

The potential utility of AI for clinical decision-making in head and neck oncology has also been explored. Lorenzi et al. compared ChatGPT-4 and Gemini Advanced models, finding both capable of providing guideline-concordant recommendations; however, notable inconsistencies occurred in critical decisions such as induction chemotherapy and surgical management.[16] This suggests that AI models, while promising, require further refinement to achieve clinical consistency.

In the context of acromegaly management, Koroglu et al. demonstrated that ChatGPT could provide accurate and reliable patient education, yet its role in managing clinical cases independently was limited.[17] Their study concluded that AI can serve effectively as a supportive tool for clinicians but should not replace clinical judgment

Finally, Fountzilas et al. discussed the evolving role of AI and machine learning in precision oncology, highlighting their capacity to analyze multidisciplinary data for enhanced precision in treatment decisions.[18] Nevertheless, challenges related to data quality, algorithm transparency, and clinical integration remain significant hurdles to overcome

However, our study indicates a moderate alignment between AI-generated decisions and real-world clinical practice (65% agreement rate), suggesting that AI recommendations tend to be more conservative than those outlined by NCCN and ESMO guidelines. This discrepancy likely stems from the fact that clinicians incorporate a wide range of patient-specific factors and contextual details that AI models may not fully account for. Indeed, Han et al. (2023) demonstrated that clinical decision support systems (CDSS) based on AI can significantly standardize oncology treatments and reduce regional and individual variation among physicians, yet these systems still require substantial enhancement in their ability to interpret nuanced clinical contexts and patient-specific variables.[13]

The high concordance between NCCN guidelines and real-world decisions suggests that clinicians generally adhere to well-established, evidence-based recommendations when determining adjuvant therapyNevertheless, the slight discrepancy between NCCN and ESMO adherence (~5%) highlights differences in guideline perspectives and risk stratification approaches. Consistent with our findings, Kann et al. (2021) noted that despite extensive AI-driven advancements, real-world oncology decisions frequently deviate due to individualized clinical judgment and contextual patient factors.[19] The moderate AI-real-world treatment concordance further supports Kann et al.'s assertion that narrow-task AI models tailored for specific clinical scenarios could yield higher predictive validity and greater practical applicability compared to broad, generalized AI systems.

The lower Cohen's kappa values for NCCN (κ= 0.73) and ESMO (κ= 0.62), despite their high agreement with real-world decisions, reflect their broad recommendations for adjuvant therapy, leading to less variability in classification. In contrast, the AI model showed a lower agreement rate (κ= 0.47), indicating moderate alignment with clinical decisions. This suggests that while AI can provide useful recommendations, it still lacks the consistency seen in established guidelines and requires further refinement before it can reliably support oncologists in decision-making.

Statistically significant findings via McNemar's test ($p< 0.001$) further validate that AI recommendations notably differ from standard guideline recommendations. Alowais et al. (2023) suggested that addressing data privacy, biases, and the inherent requirement for human judgment are essential for AI's broader clinical implementation.[20] Thus, the disparity observed in our study underscores the critical need to enhance AI models' ability to integrate detailed patient-specific clinical insights.

The variability observed between guideline adherence and actual clinical practice highlights the need for flexible yet standardized decision-making tools. Our results align with those by Lotter et al. (2024), who described that despite AI's significant potential for enhancing clinical decision-making accuracy, the transition from model development to practical implementation remains challenging, primarily due to differences in data availability, clinical interpretation of nuanced patient conditions, and region-specific practice patterns.[21]

In comparison, Han et al. (2023), who evaluated AI decision quality among oncologists treating breast cancer across different regions, found that AI systems achieved a higher standardized treatment level and lower variability compared to physician decisions alone.[13] Their findings indicated that AI-supported decision-making significantly improved guideline adherence and reduced internal variation among physician recommendations, particularly among less experienced clinicians. Our study's moderate agreement rate for AI highlights both similar potential benefits and the existing gaps that need to be addressed for optimal clinical utility.

Clinical Implications

Our findings suggest that AI holds significant promise as a decision-support tool in clinical oncology, although several steps remain essential for its effective integration into routine practice. Prospective clinical validation studies are necessary to rigorously evaluate the real-world impact of AI-generated recommendations on patient outcomes. Such validation efforts may prove particularly valuable in complex clinical scenarios or in patient populations with higher prognostic uncertainty, where AI's decision-support capabilities could provide substantial clinical benefit. Furthermore, improving data quality and standardization is critical; facilitating interoperability between electronic health records and AI platforms, as well as incorporating continuous clinician feedback into AI model updates, could significantly enhance recommendation accuracy and practical applicability. Finally, addressing ethical concerns—such as ensuring data privacy, algorithm transparency, and clinician accountability—will be essential to foster trust among healthcare providers and support broader acceptance of AI tools in oncology practice.

Limitations

This study has several limitations. Its retrospective design increases the risk of selection bias, potentially affecting concordance rates between AI recommendations, clinician decisions, and guidelines. AI recommendations relied on predefined parameters, limiting their ability to capture real-time clinical nuances. Additionally, data from a single institution may reduce generalizability. The study also lacked long-term follow-up on survival and recurrence outcomes. Prospective, multicenter studies with patient-reported outcomes and extended follow-up are needed to validate AI's clinical impact.

Future Directions

To enhance AI's clinical applicability, future efforts should focus on improving its ability to incorporate real-world patient data and dynamic clinical factors.. First, ensuring comprehensive prospective validation of AI models through multi-institutional studies will be crucial to confirm their clinical effectiveness and reliability across diverse patient populations. Prospective clinical trials specifically designed to assess the impact of AI recommendations on long-term patient outcomes, such as survival and recurrence rates, will provide stronger evidence of clinical utility. Second, enhancing data infrastructure by standardizing data collection processes and improving interoperability between AI systems and electronic health records could significantly improve the quality and reliability of AI-generated recommendations. Third, interdisciplinary collaboration between clinicians, data scientists, ethicists, and healthcare administrators will be essential to develop clear guidelines addressing ethical concerns such as patient privacy, AI transparency, bias minimization, and clinician accountability. Finally, ongoing education and training programs should be implemented for clinicians to facilitate informed use and acceptance of AI tools, enabling smoother adoption and more effective clinical utilization in oncology practice.

Conclusion

Our study showed that AI-generated adjuvant therapy recommendations moderately aligned with clinical decisions but were less consistent than NCCN and ESMO guidelines. While AI holds promise as a decision-support tool, its current limitations—such as lack of real-time clinical context and individualized patient considerations—must be addressed before routine use in oncology.

Future efforts should focus on refining AI models, improving data quality, and ensuring transparency to enhance reliability and clinical trust. Rather than replacing oncologists, AI should be integrated as a complementary tool to support decision-making where guideline recommendations are ambiguous. Continued collaboration between clinicians and AI developers will be key to optimizing its role in cancer treatment.

REFERENCES

1. Gaetani RS, Ladin K, Abelson JS. Journey through the Decades: The evolution in treatment and shared decision making for locally advanced rectal cancer. Cancers (Basel) 16: 2807, 2024.

2. André T, Boni C, Mounedji-Boudiaf L, et al. Oxaliplatin, Fluorouracil, and Leucovorin as adjuvant treatment for colon cancer. N Eng J Med 350: 2343-2351, 2004.

3. Quezada-Diaz FF, Smith JJ. Neoadjuvant therapy for rectal cancer. Surg Oncol Clin N Am 31: 279-291, 2022.

4. NCCN. NCCN Clinical Practice Guidelines in Oncology. 2025. https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf

5. Oncology) EESfM. ESMO – European Society for Medical Oncology. 2025. https://www.esmo.org/guidelines/guidelines-by-topic/esmo-clinical-practice-guidelines-gastrointestinal-cancers/localised-colon-cancer

6. Bregni G, Akin Telli T, Camera S, et al. Adjuvant chemotherapy for rectal cancer: Current evidence and recommendations for clinical practice. Cancer Treat Rev 83: 101948, 2020.

7. Yildirim HC, Gunenc D, Almuradova E, et al. A narrative review of RAS mutations in early-stage colorectal cancer: Mechanisms and clinical implications. Medicina 61: 408, 2025.

8. Pandav K, Nasser SA, Kimball KH, et al. Opportunities for artificial intelligence in oncology: From the lens of clinicians and patients. JCO Oncol Pract OP-24-00797. doi: 10.1200/OP-24-00797.

9. Abdul Rasool Hassan B, Mohammed AH, Hallit S et al. Exploring the role of artificial intelligence in chemotherapy development, cancer diagnosis, and treatment: present achievements and future outlook. Front Oncol 15:1475893, 2025.

10. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: Exploring key domains and essential functions. Computer Methods and Programs in Biomedicine Update 5: 100148, 2024.

11. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 23: 689, 2023.

12. Su Y, Tian X, Gao R, et al. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Comput Biol Med 145: 105409, 2022.

13. Han C, Pan Y, Liu C, et al. Assessing the decision quality of artificial intelligence and oncologists of different experience in different regions in breast cancer treatment. Front Oncol 13: 1152013, 2023.

14. Carl N, Schramm F, Haggenmüller S, et al. Large language model use in clinical oncology. npj Precision Oncology 8: 240, 2024.

15. Nabieva N, Brucker SY, Gmeiner B. ChatGPT's agreement with the recommendations from the 18th St. Gallen International Consensus Conference on the Treatment of Early Breast Cancer. Cancers (Basel) 16: 4163, 2024.

16. Lorenzi A, Pugliese G, Maniaci A et al. Reliability of large language models for advanced head and neck malignancies management: a comparison between ChatGPT 4 and Gemini Advanced. Eur Arch Otorhinolaryngol 281: 5001-5006, 2024.

17. Koroglu EY, Ersoy R, Sacikara M et al. Evaluation of the impact Of ChatGPT support on acromegaly management and patient education. Endocrine 87: 1141-1149, 2025.

18. Fountzilas E, Pearce T, Baysal MA et al. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digital Medicine 8: 75, 2025.

19. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. Cancer Cell 39: 916-927, 2021.

20. Alowais SA, Alghamdi SS, Alsuhebany N et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education 23: 689, 2023.

21. Lotter W, Hassett MJ, Schultz N et al. Artificial intelligence in oncology: Current landscape, challenges, and future directions. Cancer Discov 14: 711-726, 2024.

Correspondence:
Dr. Fatih KUS
Hacettepe Universitesi
Onkoloji Enstitusu
Mehmet Akif Ersoy Sokak
No: 19/A 06230 Altindag
ANKARA / TURKIYE

Tel: (+90-552) 210 72 76
e-mail: fatihkush@hotmail.com

ORCIDs:

Fatih Kus	0000-0003-1650-154X
Elvin Chalabiyev	0000-0001-6470-6043
Hasan Cagri Yildirim	0000-0003-3060-377X
Ilgin Koc Kus	0000-0001-9797-8037
Firat Sirvan	0009-0004-8210-0837
Omer Dizdar	0000-0003-0911-9078
Suayib Yalcin	0000-0001-7850-6798