

Research Article

Concordance Between ChatGPT and Real-World Adjuvant Chemotherapy Decisions in pT3N0 Colon Cancer

 Bengü Dursun,¹  İlgin Akbıyık,²  Bahar Beliz Karaoğlu,³  Sevgi Eryıldız,¹  Dılşa Mızrak Kaya,¹  Güngör Utkan³

¹Department of Medical Oncology, Ankara Etlik City Hospital, Ankara, Türkiye

²Department of Medical Oncology, University of Health Sciences, Dr. Abdurrahman Yurtaslan Ankara Oncology Training and Research Hospital, Ankara, Türkiye

³Department of Medical Oncology, Ankara University Faculty of Medicine, Ankara, Türkiye

Abstract

Objectives: Large language models (LLMs) are increasingly explored as clinical decision-support tools, but their alignment with real-world oncology practice remains uncertain. We evaluated concordance between clinicians and ChatGPT (GPT-5.2) for adjuvant chemotherapy recommendations in pT3N0 colon cancer and examined factors influencing each recommendation pattern.

Methods: This retrospective study included 186 patients with surgically resected pT3N0 colon cancer treated between 2017 and 2024. A standardized clinical vignette was generated for each case and submitted once to ChatGPT to obtain a binary recommendation (chemotherapy vs. observation). Concordance with clinicians' documented decisions was assessed using Cohen's kappa and McNemar's test. Multivariable logistic regression identified factors associated with ChatGPT and clinician recommendations.

Results: Overall concordance was 61.8%, with moderate agreement ($\kappa=0.278$). Discordance was asymmetric, with clinicians recommending chemotherapy more often than ChatGPT (58 clinician-yes/ChatGPT-no vs. 13 ChatGPT-yes/clinician-no; $p<0.001$). ChatGPT recommendations were driven mainly by adverse pathological features, whereas clinician decisions more strongly reflected guideline-modifying variables and surgical quality indicators.

Conclusion: ChatGPT showed only moderate alignment with real-world clinician decisions and appeared more rule-based, supporting its role as an adjunctive decision-support tool rather than an autonomous decision-maker.

Keywords: Adjuvant chemotherapy, ChatGPT, Cohen's kappa, colon cancer, concordance, large language model, McNemar, pT3N0, stage II

Cite This Article: Dursun B, Akbıyık I, Karaoğlu BB, Eryıldız S, Mızrak Kaya D, Utkan G. Concordance Between ChatGPT and Real-World Adjuvant Chemotherapy Decisions in pT3N0 Colon Cancer. EJMI 2026;10(4):508–515.

Colon cancer is the third most common malignancy and the second leading cause of cancer-related mortality worldwide.^[1] Despite advances in molecular profiling and postoperative risk stratification, adjuvant treatment decisions still largely rely on the TNM (tumor–node–metastasis) staging system. Fluoropyrimidine- and oxalip-

tin-based chemotherapy (CTx) is the standard of care for stage III (Node+) colon cancer, reducing recurrence risk by approximately 30%.^[2] In contrast, stage II disease (pT3–4, N0) is clinically heterogeneous, with reported 5-year survival ranging from 68% to 83% after surgery alone.^[3] Consequently, adjuvant CTx is selectively considered in stage II

Address for correspondence: Bengü Dursun, MD. Department of Medical Oncology, Ankara Etlik City Hospital, Ankara, Türkiye

Phone: +90 505 462 76 43 **E-mail:** bengumanti@gmail.com

Submitted Date: June 11, 2026 **Revised Date:** August 25, 2026 **Accepted Date:** September 2, 2026

©Copyright 2026 by Eurasian Journal of Medical Investigation - Available online at www.ejmi.org

OPEN ACCESS This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.



colon cancer based on DNA mismatch repair (MMR) status and the presence of high-risk clinicopathologic features, such as inadequate lymph node sampling, T4 tumors, obstruction or perforation, poor differentiation, lymphovascular or perineural invasion, tumor budding, and elevated preoperative CEA.^[4,5] While pT4N0 is widely regarded as a major high-risk subgroup, the role of adjuvant CTx in pT3N0 disease remains debated, particularly when additional adverse pathological features are present.

Artificial intelligence (AI) systems are increasingly integrated into modern medical workflows, including oncology, where they may support clinicians in synthesizing complex information and standardizing decision processes.^[6,7] Large language models (LLMs) such as ChatGPT can rapidly summarize evidence, map clinical variables to guideline concepts and generate structured recommendations.^[8] This capability is particularly relevant in “gray-zone” scenarios like pT3N0 colon cancer, where decisions often depend on the cumulative weight of multiple risk features, evolving evidence, and nuanced interpretation of guidelines.^[9,10] In such contexts, an LLM-based tool could theoretically function as a real-time decision-support adjunct—helping clinicians check consistency with guideline-based reasoning, reduce omission of key risk factors and facilitate patient–physician discussions by presenting a transparent rationale. However, concerns remain regarding reliability, prompt sensitivity, reproducibility across model versions and the risk of overconfident outputs.^[11] Critically, it is not yet well established how closely LLM-generated adjuvant treatment recommendations align with real-world clinician decisions in preference-sensitive settings, nor whether areas of discordance reflect potential model limitations or historical/center-level practice patterns.

Therefore, this study assessed concordance between adjuvant CTx decisions made by gastrointestinal oncology experts and recommendations generated by ChatGPT for resected pT3N0 colon cancer in a two-center real-world cohort. We aimed to determine whether ChatGPT could serve as a clinically useful decision-support tool in this preference-sensitive setting.

Methods

Study Design and Variables

This retrospective study included 186 patients with surgically resected, pathologically confirmed pT3N0 colon cancer who were managed at Ankara University Faculty of Medicine and Ankara Etlik City Hospital between 2017 and 2024. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ankara University Faculty of Medicine Human Research Ethics

Committee (no: 110-928-25; date: 13.11.2025). Collected demographic and clinical variables included age, sex and comorbidities (e.g., diabetes mellitus, hypertension, and atherosclerotic heart disease). Tumor- and surgery-related variables included primary tumor location (right vs. left colon), histologic grade, presentation with obstruction or perforation, lymphovascular invasion (LVI), perineural invasion (PNI), tumor budding, adequacy of lymph node dissection (<12 vs. ≥12 nodes), and preoperative carcinoembryonic antigen (CEA; ≥5 ng/mL). Mismatch repair status (pMMR/MSS vs. dMMR/MSI), when available, was also recorded. Adjuvant CTx decisions were abstracted from contemporaneous medical records and reflected the treating oncologist’s final documented plan. For study labeling, two gastrointestinal medical oncologists independently reviewed each case and any discrepancies were resolved by consensus.

LLM Querying Protocol and Endpoints

For each case, a standardized clinical vignette comprising predefined clinicopathological variables was generated and submitted to ChatGPT (GPT-5.2, OpenAI; accessed on 15 December 2025) using a prespecified prompt to obtain a binary recommendation regarding adjuvant CTx (yes/no). The prompt explicitly instructed the model to provide a binary “Yes” or “No” recommendation regarding whether adjuvant chemotherapy should be given; the first-line binary output (Yes/No) was directly used as the study label, without post-hoc interpretation of free-text. Each vignette was queried once (single-run) without follow-up questions, and the first output was used for the primary analysis. To explore output stability, we performed an exploratory robustness check by repeating the same query five times in ten cases selected to represent a range of clinicopathological risk profiles using identical prompt formatting and settings. Recommendations were concordant across runs in these tested cases (Table S1); however, this limited subsample was not designed or powered to quantify stochastic variability across the full cohort. The full prompt and vignette template are provided in the Supplementary Methods. Model generation parameters (e.g., temperature, top-p, and random seed) were not user-configurable in the ChatGPT web interface and were therefore left at their default values. The primary endpoint was concordance between clinicians’ decisions and ChatGPT recommendations regarding adjuvant CTx. Secondary endpoints were factors associated with (i) ChatGPT recommendations and (ii) clinicians’ adjuvant CTx decisions.

Statistical Analysis

Continuous variables were summarized as median (interquartile range) and categorical variables as counts and per-

centages. Concordance between clinicians and ChatGPT was assessed using paired 2×2 contingency tables. Directional discordance in paired proportions was evaluated with McNemar's test. To identify factors associated with ChatGPT recommendations and clinicians' decisions, binomial logistic regression models were fitted and reported as odds ratios (ORs) with 95% confidence intervals (CIs). Missing data were handled using a complete-case approach; patients with missing values in any covariate included in the multivariable models were excluded from the respective analyses, resulting in model-specific sample sizes. Model discrimination was quantified using the area under the receiver operating characteristic curve (AUC), with classification performance summarized at a probability cut-off of 0.50. To reduce optimism from apparent model performance, internal validation was performed using 10-fold cross-validation and bootstrap resampling (300 iterations), and optimism-corrected performance estimates were reported. Model calibration was assessed using calibration plots and the Brier score; calibration intercept and slope were also calculated. Multicollinearity was assessed using variance inflation factors (VIFs); VIF values were close to 1 (approximately 1.00–1.24 in the ChatGPT model and 1.00–1.23 in the clinician model), indicating no relevant multicollinearity. All statistical analyses were performed using the R package JAMovi version 2.6 (The JAMovi Project, Sydney, Australia) and R version 4.4.0 (R Foundation for Statistical Computing, Vienna, Austria). A two-sided $p < 0.05$ was considered statistically significant.

Results

Baseline Characteristics of Patients

A total of 186 patients with surgically resected and pathologically confirmed T3N0 colon cancer were included in the analysis. The median age at diagnosis was 63.5 years (IQR, 55.7–69.0), and 59.1% of patients were male. Comorbid conditions were present in 41.1% of the cohort, most commonly diabetes mellitus (DM), hypertension (HT), and atherosclerotic heart disease (ASHD). The primary tumor was located in the left colon in 53.0% of cases. Clinical presentation with bowel obstruction or perforation occurred in 10.8% of patients. With respect to adverse pathological features, most tumors were grade 2 (83.2%); lymphovascular invasion (LVI) was identified in 29.7%, perineural invasion (PNI) in 17.8%, and high tumor budding in 21.4% of patients. Deficient mismatch repair (dMMR/MSI) was observed in 12.9% of cases, whereas 87.1% exhibited proficient MMR (pMMR/MSS). Insufficient lymph node dissection (<12 nodes) was documented in 12.5% of patients. In addition, elevated preoperative carcinoembryonic antigen

(CEA ≥ 5 ng/mL) was noted in 28.4% of cases. Table 1 summarizes the clinicopathological characteristics of the study cohort.

Concordance between Clinicians and ChatGPT on Adjuvant Chemotherapy

A total of 186 patients with T3N0 colon cancer were evaluated for concordance between clinician-made decisions and ChatGPT recommendations regarding adjuvant chemotherapy. Overall, 115/186 (61.8%) decisions were concordant: 61 cases were classified as “no chemotherapy” by both, and 54 as “chemotherapy” by both. Discordance was observed in 71/186 (38.2%) cases and was predominantly driven by scenarios in which clinicians recommended chemotherapy but ChatGPT did not (58 cases) rather than

Table 1. Clinicopathological variables of patients with pT3N0 colon cancer

Baseline characteristics	n=186 (%)
Age, year (median) (IQR 25–75)	63.5 (55.7–69.0)
Sex	
Male	110 (59.1)
Female	76 (40.9)
Presence of Comorbidity	76 (41.1)
Primary tumor location	
Left colon	98 (53.0)
Right colon	87 (47.0)
Grade	
1	14 (8.1)
2	144 (83.2)
3	15 (8.7)
Obstruction or Perforation at presentation	20 (10.8)
LVI	
Yes	55 (29.7)
No	130 (70.3)
PNI	
Yes	33 (17.8)
No	152 (82.2)
High Tumor Budding	36 (21.4)
pMMR/MSS	155 (87.1)
dMMR/MSI	23 (12.9)
Adequate Lymph Node Dissection	163 (87.6)
Pre-op CEA ≥ 5 ng/mL	38 (28.4)

LVI: Lymphovascular invasion; PNI: Perineural invasion; pMMR: Proficient mismatch repair; dMMR: Deficient mismatch repair; CEA: Carcinoembryonic Antigen. Data are presented as n available (%), unless otherwise specified. Percentages are calculated using the number of patients with available data for each variable.

the reverse (13 cases). McNemar's test confirmed a significant directional discordance between the two approaches ($\chi^2=28.50$, $p<0.001$), indicating that clinicians recommend adjuvant chemotherapy more frequently than ChatGPT (Fig. 1).

Model Performance

Two multivariable binomial logistic regression models were fitted to evaluate determinants of (i) ChatGPT recommendations for adjuvant CTx and (ii) clinicians' adjuvant CTx decisions. For the ChatGPT outcome, the full ("apparent") model demonstrated very high discrimination (McFadden's $R^2=0.653$; AUC=0.966). In a sensitivity analysis using a reduced covariate set, discrimination remained high (McFadden's $R^2=0.492$; AUC=0.96) (Fig. 2A). Classification performance was 91.2% among cases not recommended CTx and 82.1% among those recommended CTx. Importantly, the

high AUC values for predicting ChatGPT outputs should be interpreted as reflecting the internal consistency and predictability of the model's recommendation policy within a constrained vignette feature space, rather than clinical correctness or superiority of decision-making. In contrast, the clinician decision model showed good but lower discrimination (McFadden's $R^2\approx0.26$; AUC=0.835) (Fig. 2B). Internal validation and calibration analyses are summarized in Table S2.

Factors Associated with ChatGPT Recommendations

For multivariable logistic regression, complete-case analysis was used, resulting in a final analytic sample of $n=147$. In the ChatGPT model, increasing age significantly reduced the likelihood of a chemotherapy recommendation (OR 0.94, $p=0.022$). Among pathological risk factors, three variables emerged as the strongest independent predictors: lymphovascular invasion (LVI) (OR 8.64, $p<0.001$), perineural invasion (PNI) (OR 13.37, $p<0.001$), tumor budding (OR 41.81, $p<0.001$). Conversely, adequate lymph node dissection was inversely associated with chemotherapy recommendation (OR 0.15, $p=0.013$) (Fig. 3). Collinearity diagnostics suggested no meaningful multicollinearity (VIFs ~ 1). The model showed excellent discrimination with an AUC of 0.928.

Factors Associated with Clinician Decisions

Similarly, in the clinical decision model, LVI remained a strong predictor of adjuvant chemotherapy decision (OR 4.12, $p=0.01$). Moreover, microsatellite instability (MSI) was significantly associated with lower chemotherapy use (OR: 0.07, $p=0.01$). Adequate lymph node dissection also predicted reduced likelihood of chemotherapy administration (OR 0.07, $p=0.01$). Other variables, including sex, tumor location, grade were not statistically significant (Fig. 4).

Discussion

In this retrospective real-world cohort of patients with resected pT3N0 colon cancer, we observed moderate concordance between ChatGPT-generated adjuvant CTx recommendations and clinicians' decisions: 61.8% of decisions were concordant, 38.2% were discordant. Clinicians were more likely to recommend adjuvant CTx than ChatGPT, leading to a strongly asymmetric discordance pattern (58 vs. 13; McNemar $p<0.001$) and moderate chance-corrected agreement ($\kappa=0.278$). These findings indicate that, in routine practice, clinicians in our setting were more likely to favor adjuvant CTx for pT3N0 disease than ChatGPT. This is expected in stage II disease, where adjuvant decisions are preference-sensitive and depend on clinical details that may not be fully captured in structured case scenarios.

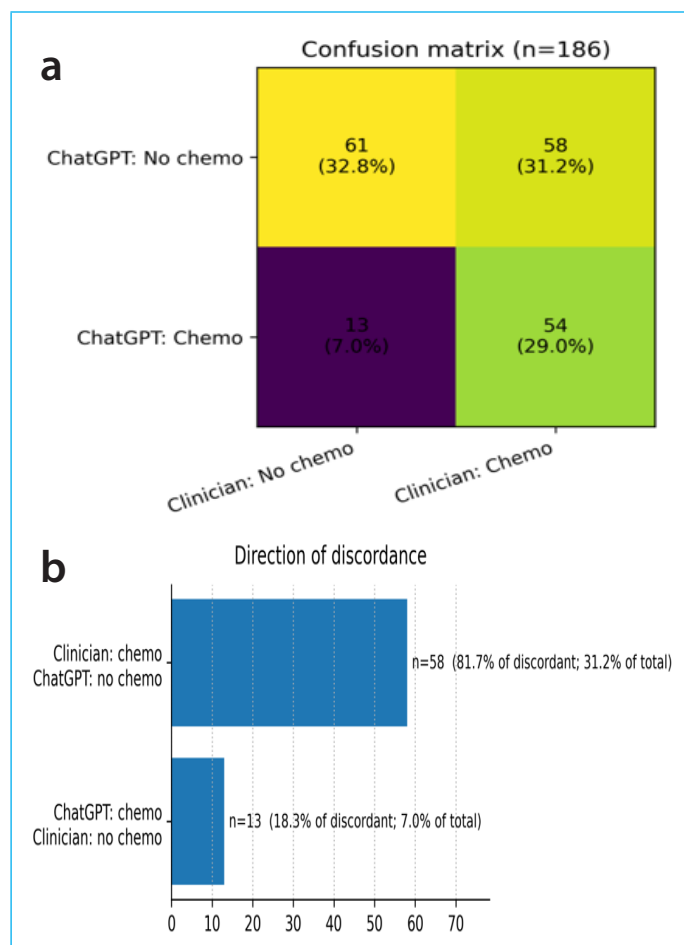


Figure 1. Agreement and direction of discordance between clinicians and ChatGPT for adjuvant chemotherapy ($n=186$). **(a)** Confusion matrix with cell counts (percentages of total cohort). **(b)** Direction of discordant pairs: clinician chemo/ChatGPT no ($n=58$) vs. ChatGPT chemo/clinician no ($n=13$). McNemar's test: $\chi^2=28.50$, $p<0.001$. Cohen's $\kappa=0.278$ (95% CI, 0.158–0.397).

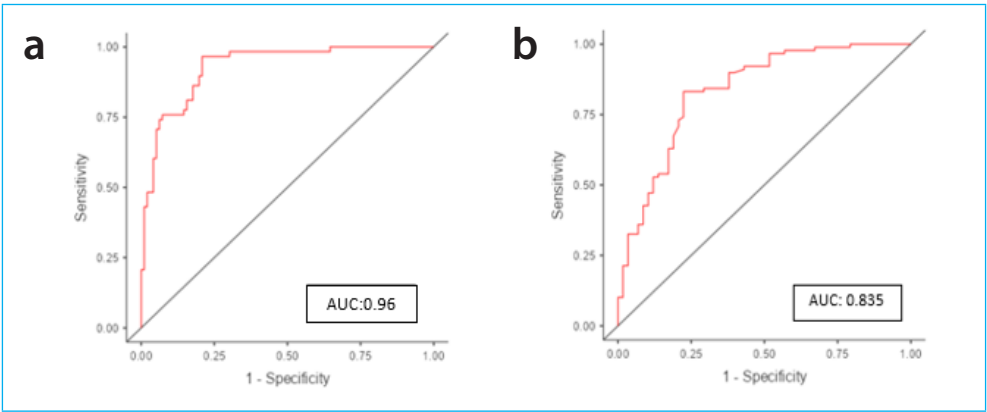


Figure 2. ROC curves for **(a)** the ChatGPT recommendation model and **(b)** the clinician decision model.

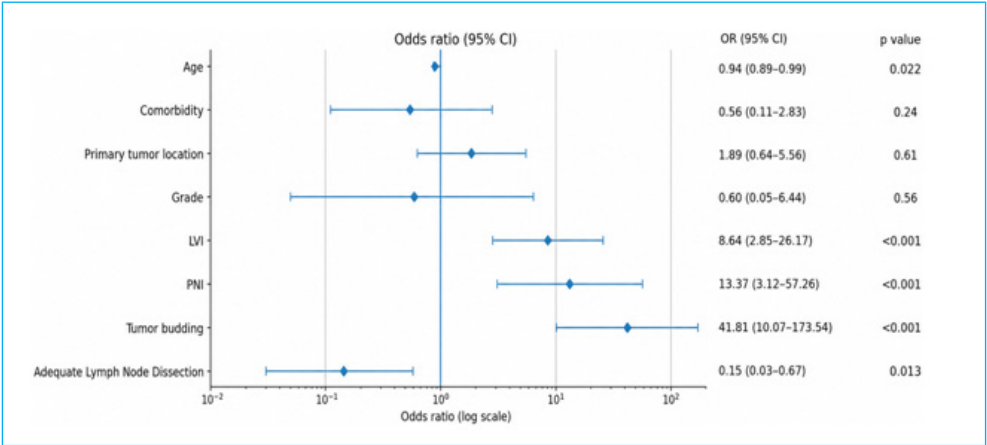


Figure 3. Multivariable analysis of factors associated with ChatGPT recommendations for adjuvant chemotherapy (adjusted OR, 95% CI) (complete-case n=147).

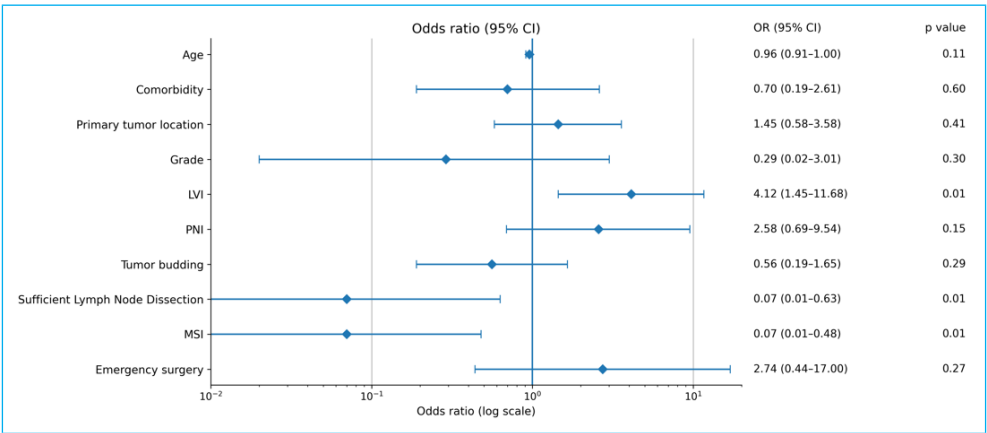


Figure 4. Multivariable logistic regression of factors associated with clinician recommendations for adjuvant chemotherapy (adjusted OR, 95% CI) (complete-case n=147).

Our real-world discordance rate is plausible and highlights how concordance depends strongly on comparator and evaluation design. Studies using multidisciplinary team (MDT) decisions as the comparator have reported higher agreement. For example, in colorectal cancer MDT settings, ChatGPT-4 showed substantial agreement with MDT recommendations, with complete concordance reported in 72.5% of pre-therapeutic and 82.8% of post-therapeutic

discussions.^[12] Similarly, a prospective mixed-cancer tumor council cohort reported high concordance with MDT decisions ($\kappa=0.764$), with discordance clustering in cases requiring individualized contextual judgment.^[13] Conversely, broader retrospective concordance analyses across diverse solid tumors have reported lower overall concordance (62.3%) with notable variation by tumor type.^[14] Together, these reports support our observation that LLM performance is not uniform across settings and that stage II colon cancer—where benefit is modest and risk–benefit trade-offs are individualized—represents a particularly challenging clinical context. In addition, because clinical decisions were made across 2017–2024, whereas the LLM was queried in December 2025 under “current” guideline framing, temporal mismatch in guideline evolution and practice patterns may also have contributed to systematic differences.

A key insight from our multivariable analyses is that ChatGPT and clinicians appeared to “weight” risk factors differently. ChatGPT recommendations were strongly driven by pathological risk variables—LVI (OR 8.650), PNI (OR 13.380), and especially tumor budding (OR 41.779), while adequate lymph node dissection reduced the likelihood of recommending chemotherapy (OR 0.159). In contrast, clinicians’ decisions were independently associated with LVI (OR 4.117), microsatellite instability (associated with lower chemotherapy use; OR 0.08), and adequate lymph node dissection (OR 0.071), whereas other factors (e.g., grade, location) were not significant. Notably, the clinician model showed good but lower discrimination than the ChatGPT model (AUC ≈ 0.837 vs. 0.966). The markedly higher AUC for predicting ChatGPT outputs (0.966) should not be interpreted as clinical correctness. Rather, it likely reflects the determinism of the model’s recommendation policy within a constrained vignette feature space—i.e., a more rule-like, internally consistent mapping from the provided variables to a binary output. It is likely that clinician decisions integrated contextual elements not fully represented in structured vignettes (e.g., perceived frailty, postoperative recovery, psychosocial factors and nuanced shared decision-making).^[15] These limitations likely contribute both to between-model variability and to within-model variability across repeated queries highlighted by prior guideline scenario studies.^[9,16,17] Improving the clinical relevance of LLM outputs may require explicit guideline grounding via retrieval-augmented generation (RAG) or structured integration of authoritative sources to reduce omission errors and standardize reasoning.^[18] In parallel, stage II colon cancer decision support may benefit from hybrid architectures in which LLMs serve primarily

as a conversational interface for information synthesis and patient-centered communication, while high-stakes recommendations are anchored in auditable guideline logic and complemented by validated, task-specific predictive models that better capture chemotherapy benefit heterogeneity.^[19] Broader reviews of AI for clinical prediction similarly emphasize that such tools should be evaluated against clinically meaningful endpoints and calibrated to real-world workflows.^[20] Several limitations should be emphasized. First, the retrospective design and region-specific practice patterns may limit the external generalizability of our results. Second, ChatGPT recommendations were derived from standardized vignettes with a constrained set of predefined inputs; consequently, key determinants routinely shaping adjuvant decisions (postoperative complications, functional status, patient preference or anticipated toxicity) were not represented and may have underestimated true concordance. Third, ChatGPT responses were obtained using a single-run design for the primary analysis and generation parameters were not user-configurable in the web interface. Although we observed fully concordant outputs in an exploratory repeat-run check of 10 cases, this does not rule out stochastic variation under different runs, prompts, or future model updates; therefore, reproducibility remains a key limitation. Fourth, we assessed concordance with clinician decisions rather than clinical outcomes; thus, our analysis cannot determine which recommendation strategy is “more correct,” only that they differ systematically. Finally, the rapid evolution of LLMs means that results are model- and date-specific, and external validity should be reassessed as models and guideline datasets change. Despite these constraints, our study adds clinically grounded evidence in a setting where decisions are preference-sensitive and where small shifts in risk perception can meaningfully change treatment choices.

Taken together, our findings suggest that current general-purpose LLMs can support information synthesis and decision framing, but their role in adjuvant chemotherapy selection for stage II colon cancer should remain adjunctive, with final recommendations anchored in guideline-based multidisciplinary evaluation and individualized patient assessment.

Conclusion

In this real-world cohort of pT3N0 colon cancer, ChatGPT showed moderate agreement with clinicians and a systematic tendency toward observation compared with clinician practice. The LLM’s recommendations appeared more rule-driven and highly sensitive to how

high-risk pathological features were presented, underscoring that LLM outputs should be used as decision support, checked against contemporary guidelines, and interpreted within a multidisciplinary clinical context. Prospective, multicenter studies incorporating richer patient-level inputs and outcome-based endpoints are needed to define safe and effective use cases in adjuvant decision-making.

Disclosures

Ethics Committee Approval: The study was approved by the Ankara University Faculty of Medicine Human Research Ethics Committee (no: İ10-928-25; date: 13.11.2025).

Informed Consent: Patient consent was waived due to the retrospective study design and the use of anonymized data, as approved by the ethics committee.

Conflicts of Interest: The authors declared no conflict of interest.

Funding: The authors declared that this research did not receive any external funding.

Use of AI for Writing Assistance: The authors declared that in this study, ChatGPT (GPT-5.2, OpenAI) was used as the evaluated large language model to generate binary adjuvant chemotherapy recommendations for standardized clinical vignettes of patients with pT3N0 colon cancer. The model was accessed on 15 December 2025. Each vignette was submitted once using a prespecified prompt, and the first binary output ("Yes" or "No") was used for analysis without post-hoc interpretation of free-text responses. The full prompt and vignette template are provided in the Supplementary Methods. In addition, ChatGPT was used only for limited English language editing and improvement of clarity during manuscript preparation. All AI-assisted outputs were critically reviewed, revised, and approved by the authors. ChatGPT was not used for data collection, statistical analysis, interpretation of results, or generation of scientific conclusions. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Authorship Contributions: Concept – BD; Design – BD; Supervision – GU; Data Collection or Processing: BD, BBK, IA, SE; Analysis or Interpretation: BD; Literature Search: BD, IA, BBK, SE, DMK, GU; Writing: BD; Critical Review: DMK, GU.

Peer-review: Externally peer-reviewed.

References

1. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. *CA Cancer J Clin* 2023;73(3):233–54. [\[CrossRef\]](#)
2. Gunderson LL, Jessup JM, Sargent DJ, Greene FL, Stewart AK. Revised TN categorization for colon cancer based on national survival outcomes data. *J Clin Oncol* 2010;28(2):264–71. [\[CrossRef\]](#)
3. Argilés G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, et al; ESMO Guidelines Committee. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2020;31(10):1291–305. [\[CrossRef\]](#)
4. Roth AD, Delorenzi M, Tejpar S, Yan P, Klingbiel D, Fiocca R, et al. Integrated analysis of molecular and clinical prognostic factors in stage II/III colon cancer. *J Natl Cancer Inst* 2012;104(21):1635–46. Erratum in: *J Natl Cancer Inst* 2013;105(2):151–2. [\[CrossRef\]](#)
5. Baxter NN, Kennedy EB, Bergsland E, Berlin J, George TJ, Gill S, et al. Adjuvant therapy for stage II colon cancer: ASCO Guideline update. *J Clin Oncol* 2022;40(8):892–910. [\[CrossRef\]](#)
6. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2(4):230–43. [\[CrossRef\]](#)
7. Ferdush J, Begum M, Hossain ST. ChatGPT and clinical decision support: scope, application, and limitations. *Ann Biomed Eng* 2024;52(5):1119–24. [\[CrossRef\]](#)
8. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak* 2025;25(1):117. [\[CrossRef\]](#)
9. Tsai CY, Cheng PY, Deng JH, Jaw FS, Yii SC. ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance. *Digit Health* 2024;10:20552076241269538. [\[CrossRef\]](#)
10. Umihanic S, Osmanovic H, Selak N, Kopric D, Huseinbasic A, Sehic-Kozica E, et al. Evaluating the concordance between ChatGPT and multidisciplinary teams in breast cancer treatment planning: a study from Bosnia and Herzegovina. *J Clin Med* 2025;14(18):6460. [\[CrossRef\]](#)
11. Anh-Hoang D, Tran V, Nguyen LM. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Front Artif Intell* 2025;8:1622292. [\[CrossRef\]](#)
12. Chatziisaak D, Burri P, Sparrn M, Hahnloser D, Steffen T, Bischofberger S. Concordance of ChatGPT artificial intelligence decision-making in colorectal cancer multidisciplinary meetings: retrospective study. *BJS Open* 2025;9(3):zraf040. [\[CrossRef\]](#)
13. Dogan I, Bartin MK, Sonmez E, Seyran E, Bozkurt HA, Yuksek M, et al. Chat GPT performance in multi-disciplinary boards-should AI be a member of cancer boards? *Healthcare (Basel)* 2025;13(18):2254. [\[CrossRef\]](#)
14. Rushanyan M, Aghabekyan T, Tamamyan G, Zohrabyan D, Safaryan L, Bedirian K, et al. Clinical decision making by ChatGPT vs medical oncologists: A retrospective concordance study. *J Clin Oncol* 2024;42(16 Suppl):e13634. [\[CrossRef\]](#)
15. Kaiser KN, Hughes AJ, Yang AD, Turk AA, Mohanty S, Gonzalez AA, et al. Accuracy and consistency of publicly available Large Language Models as clinical decision support tools for the management of colon cancer. *J Surg Oncol* 2024;130(5):1104–10. [\[CrossRef\]](#)
16. Carl N, Schramm F, Haggemüller S, Kather JN, Hetz MJ, Wies C, et al. Large language model use in clinical oncology. *NPJ Precis Oncol* 2024;8(1):240. [\[CrossRef\]](#)
17. Marchi F, Bellini E, Iandelli A, Sampieri C, Peretti G. Exploring the landscape of AI-assisted decision-making in head and neck cancer treatment: a comparative analysis of NCCN guidelines and ChatGPT responses. *Eur Arch Otorhinolaryngol* 2024;281(4):2123–36. [\[CrossRef\]](#)
18. Goh SSN, Mariappan R, Soo Woon Tan G, Yao J, Hew FM, Yeo Y, et al. Augmenting Large Language Models with National Comprehensive Cancer Network Guidelines for improved and stan-

-
- standardized adjuvant therapy recommendations in postoperative breast cancer cases. *JCO Clin Cancer Inform* 2025;9:e2400243. [\[CrossRef\]](#)
19. Xie C, Ning Z, Guo T, Yao L, Chen X, Huang W, et al. Multimodal data integration for biologically-relevant artificial intelligence to guide adjuvant chemotherapy in stage II colorectal cancer. *EBio-Medicine* 2025;117:105789. [\[CrossRef\]](#)
20. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: exploring key domains and essential functions. *Comp Meth Prog Biomed Update* 2024;5:100148. [\[CrossRef\]](#)

Supplementary Methods

LLM querying protocol, vignette template, output adjudication and robustness assessment

1) Standardized clinical vignette template

For each patient, a standardized clinical vignette was constructed using predefined clinicopathological variables and a fixed ordering to minimize formatting-related variation. Each vignette included:

- Date of operation
- Age at diagnosis
- Comorbidities
- Primary tumor location (RIGHT vs LEFT colon)
- Pathological stage (pT3N0; Stage IIA)
- Presentation with bowel obstruction/perforation (YES/NO)
- Preoperative CEA (ng/mL)
- Histologic grade (GRADE 1/2/3)
- Lymphovascular invasion (LVI: PRESENT/ABSENT)
- Perineural invasion (PNI: PRESENT/ABSENT)
- Tumor budding (Bd3/high-grade: PRESENT/ABSENT/UNKNOWN)
- Microsatellite/MMR status (MSI/dMMR vs MSS/pMMR)
- Surgical margins (NEGATIVE/POSITIVE)
- Number of lymph nodes examined

2) Prompt used for LLM querying

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [YEAR], would you recommend adjuvant chemotherapy for this patient?

3) LLM model and query settings

Model: GPT-5.2 (OpenAI)

Access date: 15 December 2025

System message: You are a medical oncologist following NCCN, ESMO, and ASCO guidelines. Provide only a Yes or No answer.

Temperature: default (not user-configurable in the web interface)

Top-p: default (not user-configurable in the web interface)

Max tokens: default

Seed: not available

Follow-up prompts: None

Output format: Binary (Yes/No)

4) Output adjudication rule

“Yes” = model recommends or considers chemotherapy.

“No” = model recommends observation/no chemotherapy.

5) Primary querying approach

Each vignette was submitted once (single-run) and the first response was used for analysis.

6) Robustness assessment

We repeated the LLM query five times in ten representative cases using identical prompts and settings. All runs produced fully concordant outputs.

7) Case example prompt

1. Case (operated on [20 May 2023]):

At diagnosis, the patient was [51] years old with comorbidities including [hypertension and atherosclerotic heart disease]. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [1.2] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [15] lymph nodes were examined.

Question: Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

Answer: NO

ChatGPT 5.2 ▾

"2. Case (operated on [20 May 2023]): At diagnosis, the patient was [51] years old with comorbidities including [hypertension and atherosclerotic heart disease]. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [1.2] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [15] lymph nodes were examined. Question: Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?"

No

2. Case (operated on [09.07.2020]):

At diagnosis, the patient was [68] years old with comorbidities including [hypertension]. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [2.59] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [16] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [YEAR], would you recommend adjuvant chemotherapy for this patient?

Answer: NO

ChatGPT 5.2 ▾

2. Case (operated on [09.07.2020]):
At diagnosis, the patient was [68] years old with comorbidities including [hypertension]. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [2.59] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [16] lymph nodes were examined.
Question:
Considering the NCCN, ESMO, and ASCO guidelines applicable in [2020], would you recommend adjuvant chemotherapy for this patient?

No

3. Case (operated on [09.06.2023]):

At diagnosis, the patient was [71] years old with NO comorbidities. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [1.83] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI PRESENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [10] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

Answer: YES

ChatGPT 5.2

3. Case (operated on [09.06.2023]):

At diagnosis, the patient was [71] years old with NO comorbidities. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [1.83] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI PRESENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [10] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

Yes

4. Case (operated on [09.07.2021]):

At diagnosis, the patient was [51] years old with comorbidities including [hypertension]. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [1.22] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [12] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2021], would you recommend adjuvant chemotherapy for this patient?

Answer: NO

ChatGPT 5.2

4. Case (operated on [09.07.2021]):

At diagnosis, the patient was [51] years old with comorbidities including [hypertension]. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [1.22] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [12] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2021], would you recommend adjuvant chemotherapy for this patient?

No

5. Case (operated on [27.03.2018]):

At diagnosis, the patient was [64] years old with comorbidities including (hypertension and diabetes mellitus). The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [6.73] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI PRESENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [16] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2018], would you recommend adjuvant chemotherapy for this patient?

Answer: YES

ChatGPT 5.2

5. Case (operated on [27.03.2018]):

At diagnosis, the patient was [64] years old with comorbidities including (hypertension and diabetes mellitus). The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [6.73] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI PRESENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [16] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2018], would you recommend adjuvant chemotherapy for this patient?

Yes.

6. Case (operated on [08.07.2020]):

At diagnosis, the patient was [67] years old with comorbidities including (hypertension, diabetes mellitus and atherosclerotic heart disease). The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [2.62] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI PRESENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [20] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2020], would you recommend adjuvant chemotherapy for this patient?

Answer: YES

ChatGPT 5.2

6. Case (operated on [08.07.2020]):

At diagnosis, the patient was [67] years old with comorbidities including (hypertension, diabetes mellitus and atherosclerotic heart disease). The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [2.62] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI PRESENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [20] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2020], would you recommend adjuvant chemotherapy for this patient?

Yes.

7. Case (operated on [23.09.2024]):

At diagnosis, the patient was [62] years old with NO comorbidities. The tumor was a [RIGHT]-sided colon cancer staged as pT3N0 (Stage IIA). There was presentation with bowel obstruction or perforation. Preoperative CEA was [21.9] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [15] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2024], would you recommend adjuvant chemotherapy for this patient?

Answer: YES

7. Case (operated on [23.09.2024]):

At diagnosis, the patient was [62] years old with NO comorbidities. The tumor was a [RIGHT]-sided colon cancer staged as pT3N0 (Stage IIA). There was presentation with bowel obstruction or perforation. Preoperative CEA was [21.9] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [15] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2024], would you recommend adjuvant chemotherapy for this patient?

Yes.

8. Case (operated on [31.05.2023]):

At diagnosis, the patient was [75] years old with comorbidities including (hypertension and diabetes mellitus). The tumor was a [RIGHT]-sided colon cancer staged as pT3N0 (Stage IIA). There was presentation with bowel obstruction or perforation. Preoperative CEA was [1.56] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [UNKNOWN]. Microsatellite status was [MSI]. Surgical margins were [MARGIN NEGATIVE]. A total of [16] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

Answer: NO

ChatGPT 5.2

8. Case (operated on [31.05.2023]):

At diagnosis, the patient was [75] years old with comorbidities including (hypertension and diabetes mellitus). The tumor was a [RIGHT]-sided colon cancer staged as pT3N0 (Stage IIA). There was presentation with bowel obstruction or perforation. Preoperative CEA was [1.56] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [UNKNOWN]. Microsatellite status was [MSI]. Surgical margins were [MARGIN NEGATIVE]. A total of [16] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

No.

9. Case (operated on [25.11.2023]):

At diagnosis, the patient was [65] years old with comorbidities including (diabetes mellitus). The tumor was a [RIGHT]-sided colon cancer staged as pT3N0 (Stage IIA). There was presentation with bowel obstruction or perforation. Preoperative CEA was [UNKNOWN] ng/mL. Histologic grade was [GRADE 1]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [12] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

Answer: YES

ChatGPT 5.2

9. Case (operated on [25.11.2023]):

At diagnosis, the patient was [65] years old with comorbidities including (diabetes mellitus). The tumor was a [RIGHT]-sided colon cancer staged as pT3N0 (Stage IIA). There was presentation with bowel obstruction or perforation. Preoperative CEA was [UNKNOWN] ng/mL. Histologic grade was [GRADE 1]. Lymphovascular invasion was [LVI PRESENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [PRESENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [12] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2023], would you recommend adjuvant chemotherapy for this patient?

Yes.

10. Case (operated on [27.10.2022]):

At diagnosis, the patient was [66] years old with NO comorbidities. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [4.85] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [10] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2022], would you recommend adjuvant chemotherapy for this patient?

Answer: YES

10. Case (operated on [27.10.2022]):

At diagnosis, the patient was [66] years old with NO comorbidities. The tumor was a [LEFT]-sided colon cancer staged as pT3N0 (Stage IIA). There was NO presentation with bowel obstruction or perforation. Preoperative CEA was [4.85] ng/mL. Histologic grade was [GRADE 2]. Lymphovascular invasion was [LVI ABSENT]. Perineural invasion was [PNI ABSENT]. Tumor budding (Bd3/high-grade) was [ABSENT]. Microsatellite status was [MSS]. Surgical margins were [MARGIN NEGATIVE]. A total of [10] lymph nodes were examined.

Question:

Considering the NCCN, ESMO, and ASCO guidelines applicable in [2022], would you recommend adjuvant chemotherapy for this patient?

Yes.

Table S1. Output stability assessment in repeated LLM runs (n=10 cases, 5 runs each)

Case ID	Run1	Run2	Run3	Run4	Run5	Stability
Case-01	No	No	No	No	No	100%
Case-02	No	No	No	No	No	100%
Case-03	Yes	Yes	Yes	Yes	Yes	100%
Case-04	No	No	No	No	No	100%
Case-05	Yes	Yes	Yes	Yes	Yes	100%
Case-06	Yes	Yes	Yes	Yes	Yes	100%
Case-07	Yes	Yes	Yes	Yes	Yes	100%
Case-08	No	No	No	No	No	100%
Case-09	Yes	Yes	Yes	Yes	Yes	100%
Case-10	Yes	Yes	Yes	Yes	Yes	100%

The 10 cases selected for repeated-run analysis were chosen to represent a spectrum of clinical risk profiles (low-risk vs high-risk stage II disease), including cases with and without adverse pathological features (LVI, PNI, high tumor budding, MSI status)

Table S2. Internal validation and calibration metrics

Model	n	AUC (apparent)	AUC (cross-validated)	Optimism	AUC optimism (corrected)	Brier	Calibration intercept	Calibration slope
ChatGPT recommendation	147	0.966	0.910	0.043	0.923	0.074	0.000	1.000
Clinician decision model	147	0.837	0.724	0.083	0.754	0.162	0.000	1.000

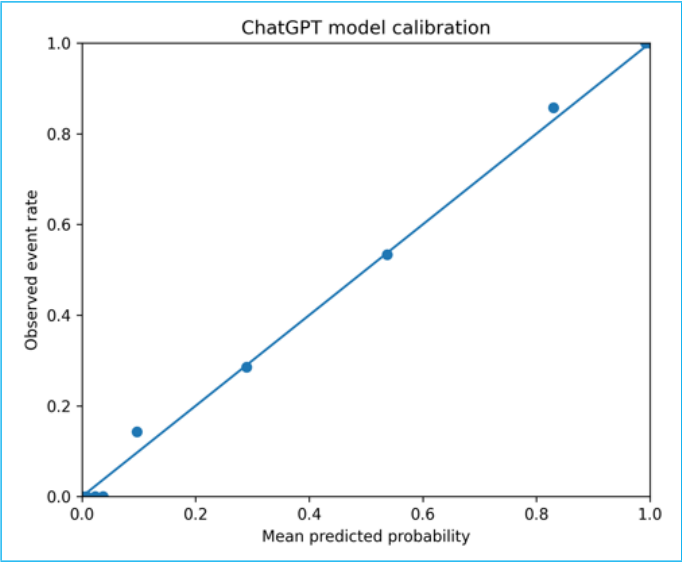


Figure S1. Calibration plot of the ChatGPT model.

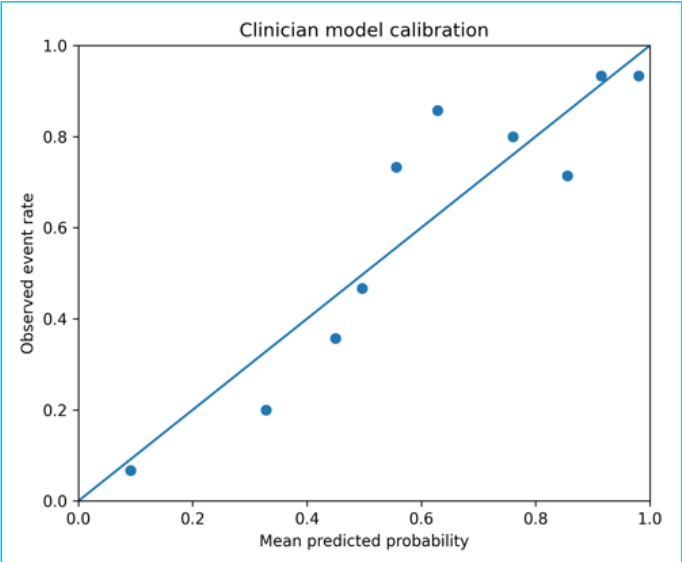


Figure S2. Calibration plot of the clinician model.