CANCER DIAGNOSIS & PROGNOSIS *6*: 548-558 (2026)
doi: 10.21873/cdp.10555

Large Language Models as Decision-support Tools for Adjuvant Therapy Planning in Early-stage Hormone Receptor–positive Breast Cancer

BERKAN KARABUĞA[1], MUSTAFA BÜYÜKKÖR[1], EKIN KONCA KARABUĞA[2], ERGIN AYDEMIR[1], OSMAN BILGE KAYA[1], MEHMET EMIN YILMAZ[1], ENES KAPTAN[3], ÖZTÜRK ATEŞ[1] and FATIH YILDIZ[1]

[1]Department of Medical Oncology, Dr. Abdurrahman Yurtaslan Ankara Oncology Research and Training Hospital, Ankara, Türkiye;
[2]Department of Medical Oncology, Ankara Etlik City Hospital, Ankara, Türkiye;
[3]Ankara Gölbaşı State Hospital, Department of Internal Medicine, Ankara, Türkiye

Abstract

Background/Aim: Adjuvant treatment decisions in hormone receptor–positive (HR), HER2-negative early-stage breast cancer are frequently guided by multigene assays; however, limited access to genomic testing remains a significant challenge, particularly in resource-limited settings. This study aimed to evaluate the concordance between adjuvant treatment recommendations generated by large language models (ChatGPT-4o and ChatGPT-o3) and those of an experienced medical oncologist in HR+/HER2– early-stage breast cancer patients when genomic assay results were unavailable.

Patients and Methods: Clinical and pathological data from 411 patients with HR+/HER2– early-stage breast cancer were provided to ChatGPT-4o and ChatGPT-o3. Both models generated adjuvant treatment recommendations, chemotherapy plus endocrine therapy (CT+ET) or endocrine therapy alone (ET) based on ESMO and NCCN guidelines. These recommendations were compared with those of a medical oncologist. Agreement was assessed using Fleiss's and Cohen's kappa statistics, and differences among evaluators were analyzed using Cochran's Q test.

Results: Overall agreement among the clinician and the two models was substantial (κ=0.67). Moderate agreement was observed between the clinician and ChatGPT-4o (κ=0.60) and between the clinician and ChatGPT-o3 (κ=0.55). Agreement between the two language models was almost perfect (κ=0.88). ChatGPT-4o demonstrated closer alignment with clinical judgment.

Conclusion: Large language models showed substantial concordance with clinician decision-making in adjuvant therapy planning for HR+/HER2– early-stage breast cancer in the absence of genomic testing. These findings suggest that such models may serve as supportive decision-making tools rather than independent decision-makers, particularly in settings with limited access to multigene assays.

Keywords: Artificial intelligence, breast cancer, chemotherapy, ChatGPT, endocrine therapy.

Berkan Karabuğa, Department of Medical Oncology, Dr. Abdurrahman Yurtaslan Ankara Oncology Research and Training Hospital, Ankara, Türkiye. Tel: +90 5378563905, e-mail: berkan.karabuga@saglik.gov.tr

Received January 12, 2026 | Revised February 6, 2026 | Accepted February 11, 2026

Introduction

Breast cancer is the most common type of cancer among women, and numerous scientific studies and clinical guidelines have been developed to guide its management. Thanks to global screening programs and awareness campaigns, early detection and treatment of breast cancer have become possible. Hormone receptor (HR)-positive, HER2-negative luminal-type breast cancer constitutes the majority of these cases. Despite advancements in disease management, there is still no clear consensus regarding adjuvant treatment in early-stage disease (2). When planning adjuvant therapy for these patients, several factors play a critical role, including the tumor (T) stage, patient's age, menopausal status, Ki-67 index, presence of lymphovascular (LVI) and perineural invasion (PNI), tumor grade, and recurrence scores obtained from multigene assays (3). Multigene assays such as Oncotype DX have a broad impact spectrum in breast cancer care, influencing adjuvant treatment decisions in HR–positive, HER2-negative patients as well as the identification of HER2-low disease (4). Unfortunately, access to reliable tools such as multigene assays remains limited in many parts of the world, and even in developing countries, many clinicians are left to make decisions on their own when managing these patients.

In this context, artificial intelligence (AI)-based language models are emerging as potentially valuable and accessible tools in clinical practice (5). ChatGPT, developed by OpenAI, is an AI model capable of providing human-like responses to user queries by accessing online sources. One of the most commonly used versions is ChatGPT-4o(6). Large language models (LLMs) based on the GPT architecture stand out from traditional AI models due to their versatility, human-like language comprehension, and broad knowledge base(7). Although some studies in the literature have compared ChatGPT-4o with the previously high-performing model, ChatGPT3.5, no study to date has compared ChatGPT-4o and ChatGPT-o3 across a large patient population (8). While the role of AI in cancer management has been increasingly explored in the era of personalized oncology, there is limited data on its utility in managing early-stage luminal-type breast cancer. In our study, we aimed to investigate whether ChatGPT-4o and ChatGPT-o3 can be used as decision-support tools in making adjuvant treatment decisions for one of the most clinically challenging subgroups: HR-positive, HER2-negative, early-stage breast cancer patients without the guidance of multigene assay results. We also sought to evaluate the concordance between the decisions of these two models and those of clinicians.

Patients and Methods

Study design and patient selection. This study included 411 patients who were diagnosed with early-stage breast carcinoma (HR-positive, HER2-negative, pT1b-T1c-T2 N0M0) and were treated and followed at Dr. Abdurrahman Yurtaslan Ankara Oncology Training and Research Hospital between 01/01/2020 and 01/01/2024. Only patients with complete clinical data and no history of next-generation genomic testing were included.Inclusion criteria were: availability of complete data on diagnosis, staging, treatment history, and follow-up; age 18 years or older; and absence of any concurrent active malignancy. Patients with incomplete data were excluded from the study. Informed consent was obtained from all participants prior to their inclusion in the study.

Patient assessment form and AI evaluation. A standardized patient assessment form was developed, including parameters such as histopathological subtypes, stage, tumor grade, CerbB2 score, estrogen receptor (ER), progesterone receptor (PR), Ki67 index, lymphovascular invasion (LVI) and perineural invasion (PNI) status, ECOG performance status, menopausal status, and comorbidities. Patients included in this study had previously been evaluated by medical oncology specialists, each with a minimum of 10 years of clinical experience, and had either completed or were still receiving their adjuvant treatment. Information regarding the administered adjuvant therapies was retrospectively obtained from patient records. The histopathological and demographic characteristics of

the patients were then provided to the ChatGPT-o3 and ChatGPT-4o models through a patient assessment form, and the models were asked to recommend an adjuvant treatment option as if making the clinical decision based on the NCCN and ESMO guidelines. Adjuvant treatment recommendations were categorized as either chemotherapy plus endocrine therapy (CT and ET) or endocrine therapy (ET) alone by investigators.

The AI-based language models used in this study were the pro versions of ChatGPT-4o and ChatGPT-o3. To avoid attenuation bias, no prior training was provided to the models, and no prompt engineering techniques were employed. For each patient, ChatGPT was prompted with the following standardized instruction: "Using the current ESMO and NCCN guidelines, and considering the clinical features provided in the patient assessment form, choose the most appropriate adjuvant treatment option for the patient: either chemotherapy plus endocrine therapy or endocrine therapy alone." Although it was recognized that using Turkish commands might affect the models' performance in accessing English-language sources, the prompts were intentionally given in Turkish to evaluate the natural language processing capabilities of the models.The Turkish form of prompt was the following: Güncel ESMO ve NCCN kılavuzlarını kullanarak ve hasta değerlendirme formunda sunulan klinik özellikleri dikkate alarak, hasta için en uygun adjuvan tedavi seçeneğini belirle: kemoterapi + endokrin tedavi veya yalnız endokrin tedavi. Clinician decisions were defined as the adjuvant treatment choices that were actually administered to patients in routine clinical practice, and these decisions were used as the reference standard for comparison with treatment recommendations generated by ChatGPT-4o and ChatGPT-o3.

GPT-4o and o3 models were accessed *via* the OpenAI interface with the following fixed parameters: temperature=0, max tokens=512, and the default system role (general assistant, with no customized instructions). A temperature value of 0 was used to ensure deterministic behavior, meaning identical inputs generated identical outputs across repeated evaluations. This configuration was chosen to maintain consistency and reproducibility of AI-based treatment recommendations.

Statistical analysis. Statistical analyses were performed using SPSS Version 27.0. Descriptive statistics were used to summarize patients' demographic characteristics as well as the pathological and immunohistochemical features of the tumors. The normality of continuous variables was assessed using the Kolmogorov-Smirnov and Shapiro-Wilk tests, along with evaluations of skewness and kurtosis. The concordance between the decisions of the clinician, ChatGPT-4o, and ChatGPT-o3 was assessed using Fleiss's Kappa and Cohen's Kappa (κ) tests, interpreted as follows: κ<0.00=poor agreement, κ=0.00-0.20=slight agreement, κ=0.21-0.40=fair agreement, κ=0.41-0.60=moderate agreement, κ=0.61-0.80=substantial agreement, and κ=0.81-1.00=almost perfect agreement. Statistical differences among the decisions of the clinician, ChatGPT-4o, and ChatGPT-o3 were evaluated using Cochran's Q test.

Logistic regression analysis was performed to compare patients with concordant chemotherapy plus endocrine therapy (CT+ET) recommendations by both the clinician and ChatGPT-4o with those who had concordant ET–only recommendations by both evaluators. The dependent variable was concordant CT+ET recommendation (yes/no), and odds ratios reflect clinicopathologic factors associated with concordant CT+ET *versus* concordant ET-only decisions.A p-Value of <0.05 was considered statistically significant.

Statement of ethics. For this study, single-center ethical approval was obtained from the Non-Interventional Clinical Research Ethics Committee of the University of Health Sciences, Dr. Abdurrahman Yurtaslan Ankara Oncology Training and Research Hospital, under the approval number 2024-09/123.

Results

A total of 411 patients were included in the study, with a mean age of 55.3±11.1 years. The clinical, histopathological,

Table I. *Clinical, histopathological, and demographic characteristics of the patients.*

		Overall		Clinician & GPT-4o: CT and ET Agreement		Clinician & GPT-4o: ET Agreement		*p*-Value
		N	%	N	%	N	%	
Age	<50 years	145	35.3	91	44.2	30	24.0	<0.001*#
	≥50 years	266	64.7	115	55.8	95	76.0	
Menopausal status	Premenopausal	147	35.8	97	47.2	26	20.8	<0.001*#
	Postmenopausal	264	64.2	109	52.9	99	79.2	
Comorbidity	None	188	45.7	110	53.4	47	37.6	
	<2	115	28.0	57	27.7	34	27.2	
	≥2	108	26.3	39	18.9	44	35.2	<0.001*#
Stage	T1c	150	36.5	58	28.2	69	55.2	
	T1b	29	7.1	6	2.9	20	16.0	
	T2	232	56.4	142	68.9	36	28.8	<0.001*#
Histopathological subtype	Idc	347	84.4	180	87.4	100	80.0	
	Ilc	46	11.2	15	7.3	21	16.8	
	Mix type	6	1.5	5	2.4	0	0.0	
	Mucinous	6	1.5	2	1.0	2	1.6	
	Neuroendocrine	2	0.5	2	1.0	0	0.0	
	Papillary	4	1.0	2	1.0	2	1.6	
Grade	Grade 1	55	13.4	5	2.4	39	31.2	
	Grade 2	218	53.0	79	38.3	81	64.8	
	Grade 3	138	33.6	122	59.2	5	4.0	<0.001*#
Cerb-B2 score	Score 0	250	60.8	111	53.9	89	71.2	
	Score 1	88	21.4	49	23.8	26	20.8	
	Score 2	73	17.8	46	22.3	10	8.0	<0.001*#
LVI status	Negative	289	70.3	128	62.1	100	80.0	
	Positive	58	14.1	43	20.9	5	4.0	<0.001*#
	Unknown	64	15.6	35	17.0	20	16.0	
PNI status	Negative	264	64.2	131	63.6	87	69.6	
	Positive	83	20.2	40	19.4	18	14.4	0.244#
	Unknown	64	15.6	35	17.0	20	16.0	
ECOG PS	PS 0	235	57.2	115	55.8	70	56.0	
	PS 1	175	42.6	91	44.2	54	43.2	
	PS 2	1	0.2	0	0	1	0.8	
Ki67 index	≤25	228	55.5	46	22.3	117	93.6	
	>25	183	44.5	160	77.7	8	6.4	<0.001*#

#Chi Square Test, *p<0.05. Idc: Invasive ductal carcinoma; Ilc: invasive lobular carcinoma; LVI: lymphovascular invasion; PNI: perineural invasion.

and demographic characteristics of the overall cohort, as well as those of patients for whom both the clinician and ChatGPT-4o recommended CT+ET or ET alone, are summarized in Table I.

Patients for whom CT+ET was recommended concordantly by both the clinician and ChatGPT-4o demonstrated significantly different clinicopathological features compared with those concordantly recommended ET alone. Specifically, patients in the CT+ET group were more likely to be younger than 50 years, premenopausal, have fewer than two comorbidities, T2-stage disease, grade 3 tumors, a CerbB2 score of 2, positive lymphovascular invasion (LVI), and a Ki-67 index greater than 25% (all p<0.001).

In univariate regression analyses, LVI positivity, Ki-67 status, age, menopausal status, tumor stage, comorbidity burden, tumor grade, and CerbB2 status were all significantly associated with the recommendation of CT+ET. In multivariate analysis, age under 50 years lost statistical significance, while other variables remained

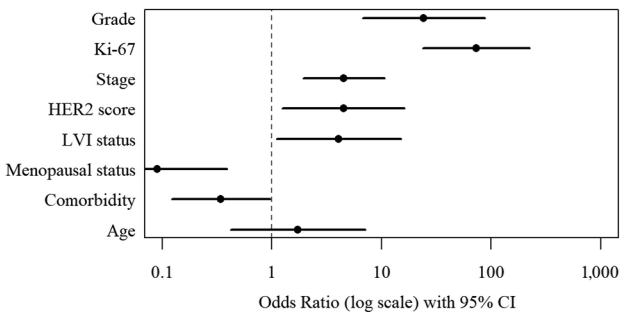

Figure 1. *Forest plot showing the odds ratio (OR) and 95% confidence interval (CI) for each clinicopathological variable. The vertical dashed line indicates the reference value (OR=1).*

independently associated with treatment decisions (Figure 1, Table II).

Among all evaluated factors, a Ki-67 index greater than 25% exerted the strongest influence on the recommendation of CT (OR=73.2, $p<0.001$, 95% CI=24.0-223.6), followed by high tumor grade (OR=24.3, $p<0.001$, 95% CI=6.8-87.1) and premenopausal status (OR=11.1, $p=0.001$, 95% CI=2.6-47.6). The clinical and pathological characteristics of patients for whom CT+ET was recommended by the clinician but ET alone by ChatGPT-4o are presented in Table III.

Most of these patients were over 50 years of age (73.1%) and postmenopausal (71.6%), with predominantly grade 2 tumors (74.6%), a CerbB2 score of 0 (64.2%), low rates of LVI positivity (11.9%), and a Ki-67 index ≤25% in the majority of cases (88.1%). Overall agreement among the clinician, ChatGPT-4o, and ChatGPT-o3 was substantial, with a Fleiss's κ value of 0.67 ($p<0.001$; 95% CI=0.62-0.73). Concordant treatment recommendations among all three evaluators were observed in 85.2% of CT+ET decisions and 81.7% of ET decisions ($p<0.001$). Cochran's Q test demonstrated a statistically significant difference among the three decision sources ($p<0.001$); however, no significant difference was observed between the treatment recommendations generated by ChatGPT-4o and ChatGPT-o3 ($p=0.058$) (Table IV).

Pairwise agreement analyses revealed a moderate and statistically significant concordance between clinician decisions and those generated by ChatGPT-4o (κ=0.60, $p<0.001$) and ChatGPT-o3 (κ=0.55, $p<0.001$). The distribution of treatment recommendations across evaluators is illustrated in Figure 2, demonstrating that the clinician more frequently favored CT+ET compared with both AI models, while ChatGPT-o3 showed a greater tendency toward ET alone. Among the two LLMs, ChatGPT-4o demonstrated closer alignment with clinician decision-making patterns (Figure 3).

Confusion matrices including absolute numbers and corresponding percentages were constructed to compare adjuvant treatment recommendations between the clinician and each language model (Table V). For ChatGPT-4o, concordant CT+ET and ET decisions accounted for 50.1% and 30.4% of the cohort, respectively, while discordant recommendations were observed in the remaining cases. A similar distribution was observed for ChatGPT-o3. In addition, a three-way concordance table summarizing agreement patterns among the clinician, ChatGPT-4o, and ChatGPT-o3 demonstrated complete agreement in 75.9% of patients (Table VI). Detailed distributions of concordant and discordant decision patterns are presented in the respective tables.

Discussion

Although multigene assay methods are considered the most reliable tools for guiding adjuvant treatment decisions in patients with HR-positive luminal-type breast cancer, access to these tests remains limited in many countries. As a result, clinicians are often required to make adjuvant treatment decisions for early-stage HR-positive, HER2-negative breast cancer patients by integrating tumor- and patient-specific factors without the support of genomic tools. In daily clinical practice, there is therefore a clear need for fast, affordable, and reliable decision-support tools. In this context, the present study aimed to investigate whether ChatGPT, an easily accessible AI-based language model, could serve as a supportive tool in adjuvant treatment planning for this patient population.

Table II. *Univariate and multivariate logistic regression models.*

	Univariate logistic regression			Multivariate logistic regression		
	OR	*p*-Value	CI 95%	OR	*p*-Value	95% CI
LVI status	6.3	<0.001*	2.4-16.4	4.0	0.035*	1.1-14.9
Ki67	50.9	<0.001*	23.1-111.8	73.2	<0.001*	24.0-223.6
Age	2.5	<0.001*	1.5-4.1	0.58	0.45	0.14-2.3
Menopausal status	3.4	<0.001*	2.0-5.6	11.1	0.001*	2.6-47.6
Stage	5.4	<0.001*	3.3-8.9	4.5	<0.001*	1.9-10.6
Comorbidity	2.3	0.001*	1.4-3.9	2.9	0.04*	1.0-8.2
Grade	34.9	<0.001*	13.7-89.0	24.3	<0.001*	6.8-87.1
CerbB2 status	3.3	0.001*	1.6-6.8	4.5	0.02*	1.3-16.1

*$p<0.05$. OR: Odds ratio; CI: confidence interval;LVI: lymphovascular invasion.

In our analysis, although a substantial concordance was observed between the adjuvant treatment decisions of a medical oncology specialist and those generated by ChatGPT-4o and ChatGPT-o3, statistically significant differences persisted between clinician and AI-based recommendations. In contrast, no significant difference was observed between the decisions of ChatGPT-4o and ChatGPT-o3, and their agreement was almost perfect. Notably, ChatGPT-4o demonstrated closer alignment with clinician decisions, suggesting that more advanced model architectures may better approximate clinical reasoning patterns.

The AI models employed in the study. In this study, we evaluated two of the most commonly usedversions of ChatGPT, namely ChatGPT-o3 and ChatGPT-4o, in order to compare the most commonly usedmodel with one previously reported to perform well in scientific reasoning. In a study by Rao *et al.* focusing on breast cancer, ChatGPT-4o was shown to outperform ChatGPT-3.5 in radiological breast cancer diagnosis (9). Similar findings have been reported in other tumor types. Studies involving sarcoma and renal cancer patients demonstrated that ChatGPT-4 provided more consistent and reliable treatment recommendations than ChatGPT-3.5 (8, 10). Alsaudi *et al.* demonstrated that ChatGPT-o3 outperformed GPT-3.5 in clinical accuracy and decision support performance, while Naliyatthaliyazchayil *et al.* reported superior clinical reasoning and risk stratification performance of ChatGPT-o3 compared with GPT-4o (11, 12). However, direct head-to-head comparisons between GPT-4o and GPT-o3 under identical clinical scenarios remain scarce. Consequently, conclusions regarding the most robust model rely mainly on indirect inference. Large-scale, standardized, and population-based comparative studies are needed to reliably identify the optimal model for clinical decision support.

Assessments of AI in different tumor types. Several studies have explored the applicability of AI-based models in oncology across different tumor types. Kuş *et al.* evaluated the consistency of ChatGPT-4o with clinician decisions and NCCN/ESMO guidelines in stage II colon cancer and reported moderate agreement between ChatGPT-4o and clinicians ($\kappa=0.47$, $p<0.001$) (13). In that study, the AI model was pre-trained before evaluation. In contrast, no pre-training was applied in our study in order to avoid attenuation bias and to reflect real-world, non-optimized clinical use. Despite this methodological difference, statistically significant discrepancies between clinician and AI decisions were still observed, although the level of agreement in our study was numerically higher. This difference may be attributable to tumor type–specific decision complexity and the larger patient cohort included in our analysis. In another study by Lechien *et al.* on head and neck cancers, ChatGPT-4o was reported to

Table III. *Clinical, histopathological, and demographic characteristics of patients recommended chemotherapy (CT) and endocrine therapy (ET) by the clinician but ET by ChatGPT-4o.*

		Clinician: CT and ET/ChatGPT-4o: ET	
		N	%
Age	<50 years	18	26.9
	≥50 years	49	73.1
Menopausal status	Premenopausal	19	28.4
	Postmenopausal	48	71.6
Comorbidity	None	26	38.8
	<2	21	31.3
	≥2	20	29.9
Stage	T1c	15	22.4
	T2	52	77.6
Histopathological subtype	Idc	57	85.1
	Ilc	8	11.9
	Mix type	1	1.5
	Mucinous	1	1.5
Grade	Grade 1	10	14.9
	Grade 2	50	74.6
	Grade 3	7	10.4
Cerb-B2 score	Score 0	43	64.2
	Score 1	8	11.9
	Score 2	16	23.9
LVI status	Negative	53	79.1
	Positive	8	11.9
	Unknown	6	9.0
PNI status	Negative	39	58.2
	Positive	22	32.8
	Unknown	6	9.0
ECOG PS	PS 0	42	62.7
	PS 1	25	37.3
Ki67 percentage	≤25	59	88.1
	>25	8	11.9

Idc: Invasive ductal carcinoma; Ilc: Invasive lobular carcinoma; LVI: lymphovascular invasion; PNI: perineural invasion.

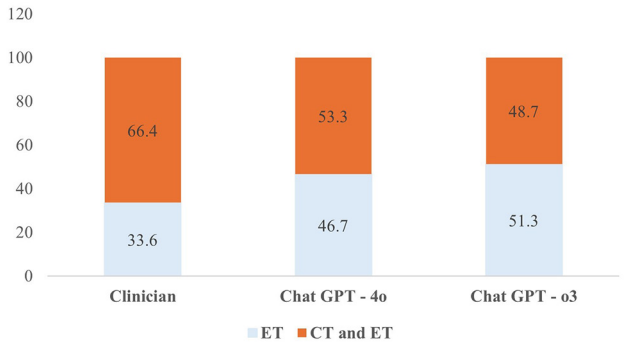

Figure 2. *Comparative distribution of adjuvant treatment recommendations among the clinician and large language models.*

Table IV. *Comparison of adjuvant treatment recommendations among the clinician, ChatGPT-4o, and ChatGPT-o3 using Cochran's Q test and post hocpairwise analyses.*

	p-Value (adjusted)
Clinician – ChatGPT-4o – ChatGPT-o3	<0.001*
ChatGPT-4o – ChatGPT-o3	0.058
Clinician – ChatGPT-o3	<0.001*
Clinician – ChatGPT-4o	<0.001*

*$p<0.05$; Adjusted *p*-values were calculated using the Bonferroni correction for multiple pairwise comparisons.

Table V. *Confusion matrix of adjuvant treatment recommendations between clinicians and ChatGPT-4o.*

Clinician/ChatGPT-4o	CT+ET	ET
CT+ET	206 (50.1%)	67 (16.3%)
ET	13 (3.2%)	125 (30.4%)

CT: Chemotherapy; ET: endocrine therapy.

have limited ability in making critical decisions compared to the tumor board (14). Similarly, in a study by Zabaleta *et al.* on non-small cell lung cancer, AI was considered a useful supportive tool for tumor boards but not suitable as a stand-alone decision-making system (15). These findings across different malignancies are consistent with our results and reinforce the supportive rather than autonomous role of AI in oncology.

Assessments of AI in breast cancer. The importance ofAI in the diagnosis and treatment of breast cancer is increasing steadily, with applications ranging from initial diagnosis to surgical decision-making and adjuvant treatment management (16, 17). Several studies have specifically evaluated AI-based tools in breast cancer management. Nabieva *et al.* compared ChatGPT-4o recommendations with the 18th St. Gallen International Consensus Conference and reported high consistency in some domains, but lower agreement in adjuvant endocrine therapy decisions where consensus is limited

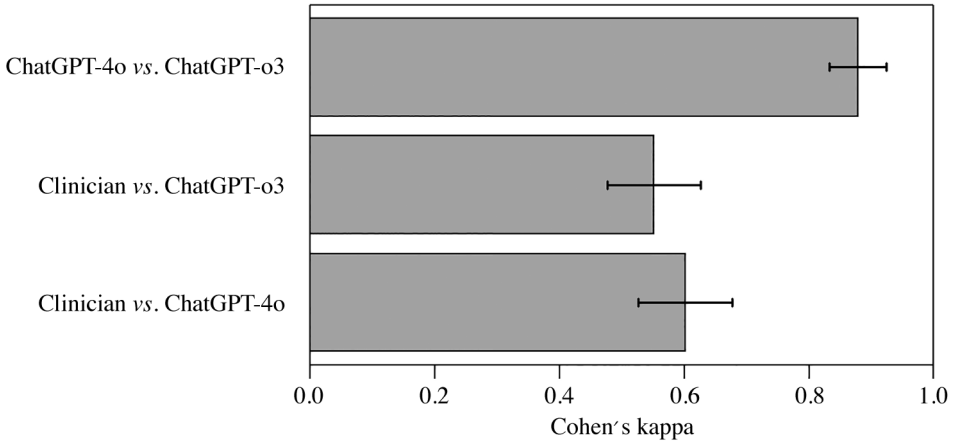

Figure 3. *Level of agreement between groups in Cohen's kappa analysis.*

(18). To the best of our knowledge, there are no published studies evaluating GPT-o3 as a clinical decision-support tool for breast cancer treatment, with most prior work focusing on GPT-3.5. Lukac *et al.* demonstrated that ChatGPT-3.5 could serve as a supportive tool when compared with multidisciplinary tumor board decisions in early-stage breast cancer (19). Similarly, Stalp *et al.* reported overall consistency between ChatGPT-3.5 and tumor board decisions despite minor discrepancies in CT and ET recommendations (20). Sorin *et al.* observed concordant decisions in 7 of 10 early-stage breast cancer cases when comparing ChatGPT-3.5 with tumor board recommendations (21). Considering the available literature and our findings, our study appears to be the first to directly assess the performance of GPT-4o and OpenAI o3 as clinical support mechanisms in breast cancer. Taken together, these results support the inference that GPT-based models may serve as reliable clinical decision-support tools in this setting.

Evaluation of subgroup analyses. Subgroup analyses demonstrated that both clinicians and ChatGPT models prioritized well-established high-risk clinicopathological features when recommending chemotherapy. CT+ET recommendations were more frequent in patients with higher tumor grade, elevated Ki-67 index, lymphovascular invasion positivity, premenopausal status, higher tumor stage, and lower comorbidity burden. Among these factors, Ki-67 >25%, high tumor grade, and premenopausal status emerged as the most influential determinants of chemotherapy recommendation.

Table VI. *Concordance of adjuvant treatment recommendations among clinicians, ChatGPT-4o, and ChatGPT-o3.*

Clinician	ChatGPT-4o	ChatGPT-o3	N (%)
CT+ET	CT+ET	CT+ET	188 (45.7)
ET	ET	ET	124 (30.2)
CT+ET	ET	ET	65 (15.8)
ET	CT+ET	CT+ET	9 (2.2)
CT+ET	CT+ET	ET	2 (0.5)
ET	CT+ET	ET	4 (1.0)
ET	ET	CT+ET	1 (0.2)
CT+ET	ET	CT+ET	18 (4.4)

CT: Chemotherapy; ET: endocrine therapy.

These findings are consistent with previous literature demonstrating the prognostic and predictive value of these parameters in early-stage breast cancer. Lymphovascular invasion has been associated with poorer overall and disease-free survival (22). A study reported that LVI positivity influences adjuvant treatment decisions alongside stage, age, and tumor histopathology (23). In a study involving triple-negative and HER2-positive early-stage breast cancer, patients with a lower comorbidity burden were significantly more likely to receive adjuvant CT (24). Additionally, patients with higher CerbB2 scores

and premenopausal status have been shown to derive greater benefit from chemotherapy in selected settings (25, 26). Although the prognostic role of Ki-67 remains debated, multiple studies support its value as a marker of recurrence risk in HR-positive/HER2-negative disease (27-29). Tumor grade and T stage are also well-established risk indicators guiding adjuvant treatment decisions (30).

Discordance between clinician and ChatGPT-4o recommendations was mainly observed in patients with intermediate-risk profiles, including postmenopausal status, intermediate tumor grade, and lower Ki-67 values. This suggests that ChatGPT relied more strictly on objective risk factors, whereas clinicians may have incorporated additional contextual considerations such as clinical experience, patient preference, and precautionary reasoning. Improving the interpretability and transparency of AI-based decision-support systems through explainable outputs may help bridge this gap and enhance clinician trust.

Study imitations. First, being a retrospective and single-center study may limit the generalizability of the findings. Second, the prompts entered into ChatGPT were written in Turkish to evaluate the model's natural language understanding in a non-English context. However, since ChatGPT primarily relies on English-language sources, this language difference may have slightly affected its performance. Third, due to the immaturity of the current dataset and the insufficient follow-up duration, survival analyses could not be performed. Nevertheless, once the data reach maturity, our results will be updated and shared again. Another limitation is that ChatGPT's recommendations were based on the most recent versions of the ESMO and NCCN guidelines, whereas clinician decisions reflected the guideline updates and treatment standards valid at the time of patient management (2020-2024). Finally, ethical, privacy, and liability considerations related to the use of large language models in oncology decision-making should be acknowledged, emphasizing the need for expert oversight and strict data protection measures when integrating AI tools into clinical workflows.

Conclusion

In our study, a substantial and statistically significant level of concordance was observed between the treatment decisions of the clinician and those of ChatGPT-4o and ChatGPT-o3 in HR-positive/HER2-negative, early-stage breast cancer patients, for whom there is no clear consensus regarding adjuvant therapy in the absence of genomic profiling. Despite this overall consistency, statistically significant differences were also identified between the clinician's decisions and those of both ChatGPT models. Pairwise analyses revealed no significant difference between the decisions of ChatGPT-4o and ChatGPT-o3, and ChatGPT-4o was found to be more numerically aligned with the clinician's decisions.

Our findings suggest that while AI may not yet be suitable as an independent decision-maker in oncology, it shows potential as a decision-support tool under expert supervision. Further multicenter, prospective studies with longer follow-up periods are needed to evaluate the safe and effective integration of AI into oncology practice.

Conflicts of Interest

The Authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Authors' Contributions

B.K: Writing, Statistical analysis; M.B: Conceptualization; E.K.K: Data curation; E.A: Figures, Data curation; O.B.K: Tables; M.E.Y: Review; E.K: Visualization; İ.D.O: Review; Ö.A: Editing; F.Y: Supervision.

Funding

This study did not receive any financial support or funding.

Artificial Intelligence (AI) Disclosure

Large language models (ChatGPT-4o and ChatGPT-o3) were used as decision-support tools to generate adjuvant treatment recommendations based on predefined clinical and pathological data, in accordance with current ESMO and NCCN guidelines. The AI models did not have access to patient-identifiable information and did not participate in data collection, data interpretation, or final clinical decision-making. All treatment decisions used for comparison were independently made by an experienced medical oncologist, who retained full responsibility for clinical judgment and patient care.

References

1 Stabellini N, Cao L, Towe CW, Amin AL, Montero AJ: Estimating the overall survival benefit of adjuvant chemo-endocrine therapy in women over age 50 with pT1-2N0 early stage breast cancer and 21-gene recurrence score ≥26: A National Cancer Database analysis. Cancer Med 12(19): 19607-19616, 2023. DOI: 10.1002/cam4.6584

2 Liu D, Chang L, Hao Q, Ren X, Liu P, Liu X, Wei Y, Wang M, Wu H, Kang H, Lin S: Is neoadjuvant chemotherapy necessary for T2N0-1M0 hormone receptor-positive/HER2-negative breast cancer patients undergoing breast-conserving surgery? J Cancer Res Clin Oncol 150(5): 285, 2024. DOI: 10.1007/s00432-024-05810-6

3 Sparano JA, Gray RJ, Makower DF, Albain KS, Saphner TJ, Badve SS, Wagner LI, Kaklamani VG, Keane MM, Gomez HL, Reddy PS, Goggins TF, Mayer IA, Toppmeyer DL, Brufsky AM, Goetz MP, Berenberg JL, Mahalcioiu C, Desbiens C, Hayes DF, Dees EC, Geyer CE Jr, Olson JA Jr, Wood WC, Lively T, Paik S, Ellis MJ, Abrams J, Sledge GW Jr: Clinical outcomes in early breast cancer with a high 21-gene recurrence score of 26 to 100 assigned to adjuvant chemotherapy plus endocrine therapy: a secondary analysis of the TAILORx randomized clinical trial. JAMA Oncol 6(3): 367-374, 2020. DOI: 10.1001/jamaoncol.2019.4794

4 Douganiotis G, Kontovinis L, Zarampoukas T, Natsiopoulos I, Papazisis K: Association of oncotype-DX HER2 single gene score with HER2 expression assessed by immunohistochemistry in HER2-low breast cancer. Cancer Diagn Progn 4(5): 605-610, 2024. DOI: 10.21873/cdp.10370

5 Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4): 1234-1240, 2020. DOI: 10.1093/bioinformatics/btz682

6 Haemmerli J, Sveikata L, Nouri A, May A, Egervari K, Freyschlag C, Lobrinus JA, Migliorini D, Momjian S, Sanda N, Schaller K, Tran S, Yeung J, Bijlenga P: ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Inform 30(1): e100775, 2023. DOI: 10.1136/bmjhci-2023-100775

7 Nori H, King N, McKinney SM, Carignan D, Horvitz E: Capabilities of gpt-4 on medical challenge problems. arXiv: 230313375, 2023. DOI: 10.48550/arXiv.2303.13375

8 Liang R, Zhao A, Peng L, Xu X, Zhong J, Wu F, Yi F, Zhang S, Wu S, Hou J: Enhanced artificial intelligence strategies in renal oncology: iterative optimization and comparative analysis of GPT 3.5 *versus* 4.0. Ann Surg Oncol 31(6): 3887-3893, 2024. DOI: 10.1245/s10434-024-15107-0

9 Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, Succi MD: Evaluating GPT as an adjunct for radiologic decision making: GPT-4 *versus* GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 20(10): 990-997, 2023. DOI: 10.1016/j.jacr.2023.05.003

10 Li CP, Jakob J, Menge F, Reißfelder C, Hohenberger P, Yang C: Comparing ChatGPT-3.5 and ChatGPT-4's alignments with the German evidence-based S3 guideline for adult soft tissue sarcoma. iScience 27(12): 111493, 2024. DOI: 10.1016/j.isci.2024.111493

11 Alsaudi EM, Shilbayeh SA, Abu-Farha RK: Benchmarking ChatGPT-3.5 and OpenAI o3 against clinical pharmacists: preliminary insights into clinical accuracy, sensitivity, and specificity in pharmacy MCQs. Healthcare (Basel) 13(14): 1751, 2025. DOI: 10.3390/healthcare13141751

12 Naliyatthaliyazchayil P, Muthyala R, Gichoya JW, Purkayastha S: Evaluating the reasoning capabilities of large language models for medical coding and hospital readmission risk stratification: zero-shot prompting approach. J Med Internet Res 27: e74142, 2025. DOI: 10.2196/74142

13 Kus F, Chalabiyev E, Yildirim HC, Koc Kus I, Sirvan F, Dizdar O, Yalcin S: Artificial intelligence (ChatGPT-4o) in adjuvant treatment decision-making for stage II colon cancer: a comparative analysis with clinician recommendations and NCCN/ESMO guidelines. Int J Hematol Oncol 35(1): 68-74, 2025. DOI: 10.4999/uhod.258149

14 Lechien JR, Chiesa-Estomba C, Baudouin R, Hans S: Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. Eur Arch Otorhinolaryngol 281(4): 2105-2114, 2024. DOI: 10.1007/s00405-023-08326-w

15 Zabaleta J, Aguinagalde B, Lopez I, Fernandez-Monge A, Lizarbe JA, Mainer M, Ferrer-Bonsoms JA, de Assas M: Utility of artificial intelligence for decision making in thoracic multidisciplinary tumor boards. J Clin Med 14(2): 399, 2025. DOI: 10.3390/jcm14020399

16 Cossu M, Cuniolo L, Diaz R, Oliva M, Cornacchia C, Murelli F, Depaoli F, Gipponi M, Margarino C, Boccardo C, Franchelli S, Pesce M, Allievi R, D'Agraives AC, Abdallah S, Fregatti P: BREAST AI-PLAN: Prompt-driven AI assistance for breast

surgery planning - a retrospective single-center study. In Vivo 39(6): 3271-3277, 2025. DOI: 10.21873/invivo.14126
17 Park JH, Lim JH, Kim S, Heo J: A multi-label artificial intelligence approach for improving breast cancer detection with mammographic image analysis. In Vivo 38(6): 2864-2872, 2024. DOI: 10.21873/invivo.13767
18 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(24): 4163, 2024. DOI: 10.3390/cancers16244163
19 Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, Janni W, Rack B, Pfister K, Heitmeir B, Ebner F: Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 308(6): 1831-1844, 2023. DOI: 10.1007/s00404-023-07130-5
20 Stalp JL, Denecke A, Jentschke M, Hillemanns P, Klapdor R: Quality of ChatGPT-generated therapy recommendations for breast cancer treatment in gynecology. Curr Oncol 31(7): 3845-3854, 2024. DOI: 10.3390/curroncol31070284
21 Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, Konen E, Barash Y: Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 9(1): 44, 2023. DOI: 10.1038/s41523-023-00557-8
22 Houvenaeghel G, Cohen M, Classe JM, Reyal F, Mazouni C, Chopin N, Martinez A, Daraï E, Coutant C, Colombo PE, Gimbergues P, Chauvet MP, Azuar AS, Rouzier R, Tunon de Lara C, Muracciole X, Agostini A, Bannier M, Charaffe Jauffret E, De Nonneville A, Goncalves A: Lymphovascular invasion has a significant prognostic impact in patients with early breast cancer, results from a large, national, multicenter, retrospective cohort study. ESMO Open 6(6): 100316, 2021. DOI: 10.1016/j.esmoop.2021.100316
23 Kuhn E, Gambini D, Despini L, Asnaghi D, Runza L, Ferrero S: Updates on lymphovascular invasion in breast cancer. Biomedicines 11(3): 968, 2023. DOI: 10.3390/biomedicines11030968
24 Tamirisa N, Dong W, Shen Y, Lin H, Shaitelman SF, Babiera G, Bedrosian I: Sequence of therapy impact on older women with comorbidities and triple-negative or HER2-positive breast cancer. NPJ Breast Cancer 11(1): 21, 2025. DOI: 10.1038/s41523-025-00732-z
25 Roy AM, Jiang C, Perimbeti S, Deng L, Shapiro CL, Gandhi S: Oncotype Dx score, HER2 low expression, and clinical outcomes in early-stage breast cancer: a National Cancer Database analysis. Cancers (Basel) 15(17): 4264, 2023. DOI: 10.3390/cancers15174264
26 Jacobson A: Benefits of adjuvant chemotherapy differ by menopausal status in women with HR+/HER2- early breast cancer, 1-3 positive nodes, and a low recurrence score. Oncologist 27(Suppl 1): S15-S16, 2022. DOI: 10.1093/oncolo/oyac012
27 Criscitiello C, Disalvatore D, De Laurentiis M, Gelao L, Fumagalli L, Locatelli M, Bagnardi V, Rotmensz N, Esposito A, Minchella I, De Placido S, Santangelo M, Viale G, Goldhirsch A, Curigliano G: High Ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in Luminal B HER2 negative and node-positive breast cancer. Breast 23(1): 69-75, 2014. DOI: 10.1016/j.breast.2013.11.007
28 Patel R, Hovstadius M, Kier MW, Moshier EL, Zimmerman BS, Cascetta K, Jaffer S, Sparano JA, Tiersten A: Correlation of the Ki67 Working Group prognostic risk categories with the Oncotype DX Recurrence Score in early breast cancer. Cancer 128(20): 3602-3609, 2022. DOI: 10.1002/cncr.34426
29 Louis DM, Nair LM, Vallonthaiel AG, Narmadha MP, Vijaykumar DK: Ki 67: a promising prognostic marker in early breast cancer-a review article. Indian J Surg Oncol 14(1): 122-127, 2023. DOI: 10.1007/s13193-022-01631-6
30 Walsh EM, Smith KL, Stearns V: Management of hormone receptor-positive, HER2-negative early breast cancer. Semin Oncol 47(4): 187-200, 2020. DOI: 10.1053/j.seminoncol.2020.05.010