

Research Article

Evaluating Concordance Between Physicians' Decisions and AI-Generated Treatment Recommendations for Breast Cancer: A Retrospective Pilot Study

 Müge Sönmez,¹  Eser Uyanık²

¹Division of Medical Oncology, Department of Internal Medicine, Ordu University Faculty of Medicine, Ordu State Hospital, Ordu, Türkiye

²Department of Internal Medicine, Ordu State Hospital, Ordu, Türkiye

Abstract

Objectives: This study evaluated concordance between physicians' decisions and recommendations generated by a large language model (LLM) in breast cancer management and explored factors associated with discordance.

Methods: In this study, 101 patients with breast cancer were evaluated. Anonymized clinical data were entered into Chat Generative Pre-trained Transformer (GPT) 5.5 to generate treatment recommendations based on National Comprehensive Cancer Network (NCCN) and/or European Society for Medical Oncology (ESMO) guidelines. Concordance between physician decisions and artificial intelligence-generated (AI-generated) recommendations was classified as full, partial, or non-concordant. Descriptive statistics and exploratory multivariable logistic regression analyses were performed.

Results: The median age was 53 years (IQR, 45–62). Disease stage was Stage I–II in 70 patients (69.3%), Stage III in 20 (19.8%), and metastatic in 11 (10.9%). Full concordance was observed in 83 patients (82.2%), while partial concordance occurred in 18 (17.8%); no cases of complete non-concordance were identified. Partial/non-concordance was more frequent in metastatic disease than in earlier-stage disease (45.5% vs. 12.9–20.0%; $p=0.031$). In multivariable analysis, HER2 (human epidermal growth factor receptor 2) positive disease (adjusted OR, 3.71; 95% CI (confidence interval), 1.09–12.65; $p=0.036$) and the 2024–2025 guideline period (adjusted OR, 4.73; 95% CI, 1.26–17.78; $p=0.022$) were independently associated with partial/non-concordance.

Conclusion: High overall concordance was observed between physician decisions and AI recommendations. Discordance was more common in particularly metastatic disease and periods of evolving treatment guidelines. Further prospective validation is required before routine clinical implementation.

Keywords: Artificial intelligence, breast cancer, ChatGPT, clinical decision support, large language model, treatment concordance

Cite This Article: Sönmez M, Uyanık E. Evaluating Concordance Between Physicians' Decisions and AI-Generated Treatment Recommendations for Breast Cancer: A Retrospective Pilot Study EJMI 2026;10(4):480–490.

Address for correspondence: Müge Sönmez, MD. Division of Medical Oncology, Department of Internal Medicine, Ordu University Faculty of Medicine, Ordu State Hospital, Ordu, Türkiye

Phone: +90 532 749 13 95 **E-mail:** mugemd@yahoo.com

Submitted Date: June 15, 2026 **Revised Date:** August 23, 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.



Breast cancer is one of the most frequently diagnosed malignancies worldwide^[1] and requires individualized treatment planning based on disease stage, tumor biology, patient characteristics, prior therapy, and treatment availability.^[2] International guideline frameworks, such as National Comprehensive Cancer Network (NCCN) recommendations, provide structured, evidence-based pathways for managing breast cancer.^[3] However, real-world oncology decisions often require interpretation beyond algorithmic recommendations because comorbidities, frailty, treatment tolerance, reimbursement restrictions, local drug availability, and patient preferences may influence the final therapeutic plan.^[4,5]

Large language models (LLMs) have recently emerged as potential tools for clinical decision support. In oncology, LLMs may help summarize guideline-based options, structure complex clinical information, and provide second-opinion-style recommendations.^[6] Early evaluations in breast cancer management have demonstrated variable, yet frequently encouraging, concordance between LLM-generated recommendations and multidisciplinary tumor board decisions.^[7] Retrieval-augmented generation (RAG)-enhanced systems have also been explored to improve guideline anchoring and efficacy in tumor board decision support.^[8,9]

Despite this promise, LLM outputs remain prompt-dependent and vulnerable to incomplete clinical context. Recent studies have emphasized that LLMs are not ready for autonomous clinical decision-making, particularly because they may fail to follow instructions, be sensitive to input order and quantity, and not reliably adhere to clinical guidelines in real-world workflows.^[10] Consequently, interpreting concordance necessitates professional caution, as AI-generated recommendations may be inaccurate^[11]; thus, agreement with a physician's decision does not invariably prove optimality, nor does disagreement automatically indicate inappropriate care.^[12]

The present retrospective pilot study evaluated concordance between real-world physician treatment decisions and ChatGPT-generated recommendations for breast cancer. A historical guideline-matching strategy was used to reduce the risk of evaluating earlier treatment decisions against more recent therapeutic standards. In response to the evolving HER2-low treatment paradigm^[13], HER2 status was also reported using detailed Immunohistochemistry/ in situ hybridization (IHC/ISH) categories. Moreover, limitations regarding the separate assessment of HER2-ultralow disease were acknowledged.

Methods

Study design and population

This single-center retrospective pilot study was conducted at the Medical Oncology Clinic of Ordu State Hospital. All

physician treatment decisions included in the analysis represented real-world treatment plans originally documented by the same treating medical oncologist at the study center during routine clinical care. Patients diagnosed with breast cancer and treated with systemic therapy, with treatment planning estimated to span 2021 to 2025, were included. Ethics approval was obtained for retrospective analysis of anonymized clinical data (Ordu University Non-Interventional Clinical Research Ethics Committee, no: 318, date: 26.09.2025) and for AI-assisted evaluation of treatment decision concordance. Notably, the institutional ethics protocol did not specify a particular large language model version for the analysis.

Inclusion criteria were histopathologically confirmed breast cancer, receipt of systemic therapy, availability of hormone receptor and HER2 status, and adequate clinical follow-up data. Exclusion criteria were incomplete essential clinical or molecular data, insufficient follow-up, lack of systemic treatment, or inability to determine the approximate treatment-planning year.

Data extraction

Demographic, clinical, pathological, molecular, and treatment-related variables were collected retrospectively from the hospital information management system, electronic medical records, and archived patient files. Data were extracted regarding age, menopausal status, diagnosis date, stage group, metastatic status, histological subtype, grade, hormone receptor status, HER2 status, Ki-67 index, lymphovascular invasion, and genomic assay results (when available). Systemic treatment components included chemotherapy, endocrine therapy, anti-HER2 therapy, radiotherapy, and surgery.

HER2 classification and HER2-low/HER2-ultralow handling

To address the current HER2-low treatment paradigm, HER2 status was reported using mutually exclusive categories: HER2-positive disease was defined as IHC 3+ or IHC 2+/ISH-positive. HER2-low disease was defined as IHC 1+ or IHC 2+/ISH-negative. HER2-zero disease was defined as IHC 0.^[14]

In the component-level analysis, conventional anti-HER2 treatment concordance was evaluated only among HER2-positive patients. HER2-low and HER2-zero cases were excluded from the denominator for conventional anti-HER2 treatment concordance. For cases in contemporary metastatic treatment periods where HER2-low status could be therapeutically relevant^[15], this was considered during interpretation and discussed as a limitation of historical guideline matching.

AI workflow, prompt architecture, and RAG status

AI analyses were conducted in May 2026 using ChatGPT via the OpenAI web interface. GPT-5.5 was utilized as the available model. Because the ChatGPT web interface did not provide a stable, API-style model identifier or snapshot, an exact endpoint identifier (e.g., a dated API model version) could not be independently verified. No API-based automation, external fine-tuning, local model customization, external plug-in workflow, vector database, or institutional data connection was used.

A standardized zero-shot structured prompt was used across all cases. The standardized AI prompt template used for all cases is provided in Appendix A. The prompt architecture comprised four components: (1) role assignment as a breast cancer oncology decision-support assistant; (2) patient-level clinical input; (3) historical guideline instruction requiring the model to use the NCCN and/or ESMO framework corresponding to the estimated treatment-planning year; and (4) structured output requirements, including a treatment recommendation and a concordance classification. Each patient received one primary AI-generated recommendation, and repeated prompting or iterative correction was avoided to reduce output variability.

RAG was not used. No external NCCN/ESMO guideline documents, uploaded reference libraries, institutional databases, vector databases, or real-time web retrieval systems were connected to the model. This design was chosen to evaluate the performance of a broadly accessible, standardized ChatGPT workflow rather than a customized guideline-retrieval engine. However, the absence of RAG limits traceability to specific guideline passages and does not guarantee complete isolation from contemporary medical knowledge. This limitation was explicitly acknowledged as a methodological constraint.

Concordance definitions and component-level analysis

The primary endpoint was concordance between the real-world physician's treatment decision and the AI-generated recommendation. Full concordance was defined as agreement on the core oncologic treatment strategy and major guideline-based therapeutic components. Partial concordance was defined as agreement on the general treatment principle, with differences in regimen, sequencing, duration, intensity, or component selection. Non-concordance was defined as a major disagreement in treatment strategy or a clear divergence from the historical guideline-based approach.

Notably, partial concordance and non-concordance were not considered clinically inappropriate physician treatment.

Real-world divergence from guideline-based recommendations may be clinically justified by patient age, frailty, comorbidities, toxicity concerns, and patient preference.^[16]

Component-level concordance was evaluated for definitive surgery, chemotherapy, endocrine therapy, anti-HER2 therapy, and radiotherapy. Components not applicable to an individual patient were excluded from the component-specific denominator.

Interobserver reliability and AI confidence score

Although Cohen's kappa analysis was planned to assess interobserver agreement, it could not be performed because the available dataset lacked separate blinded patient-level concordance classifications from two independent reviewers. Therefore, concordance categories were reported based on the available clinical evaluation and consensus-based assessment rather than a formal quantitative interobserver reliability analysis. Cohen's kappa requires independent categorical ratings from exactly two reviewers for each case; the absence of such data is a methodological limitation.^[17]

Given the absence of a prospectively generated model-derived confidence score within the original dataset, a post hoc rule-based proxy AI confidence score was developed for exploratory purposes. The score ranged from 0 to 10 and incorporated completeness of patient-level data, clarity of guideline-based recommendations, and degree of clinical ambiguity. Confidence categories were defined as high (≥ 8.0), intermediate (6.0–7.9), and low (< 6.0).

Statistical analysis

Continuous variables were summarized as mean $\pm$ standard deviation or as median with interquartile range, as appropriate. Categorical variables were presented as frequencies and percentages. Group comparisons were performed using chi-square or Fisher's exact tests for categorical variables and Mann-Whitney U tests or independent samples t-tests for continuous variables, as appropriate. Exploratory multivariable logistic regression was conducted to identify factors associated with partial concordance versus full concordance. Given the pilot design and the limited number of partial-concordance events, the model was restricted to a small set of clinically relevant variables (age, HER2 positivity, metastatic disease, and guideline period 2024–2025). Regression findings were interpreted as hypothesis-generating rather than confirmatory. A two-sided p-value < 0.05 was considered statistically significant. Statistical analyses were performed using Python version 3.13.5 (Python Software Foundation, Wilmington, DE, USA). Data management and descriptive analyses were conducted using pandas version 2.2.3 and NumPy version 2.3.5. Inferential statistical analyses

were performed using SciPy version 1.17.0, while multivariable logistic regression was conducted using statsmodels version 0.14.6. The forest plot displaying the adjusted odds ratios was generated using Matplotlib version 3.10.8.

Results

Patient characteristics

A total of 101 patients were included in the analysis. The mean age was 53.4 ± 12.2 years, and the median age was 53 years (IQR, 45–62; range, 26–83). Furthermore, 68 patients (67.3%) were postmenopausal, whereas 33 (32.7%) were premenopausal. Stage I–II disease was observed in 70 patients (69.3%), Stage III in 20 (19.8%), and Stage IV/metastatic disease in 11 (10.9%). Moreover, hormone receptor-positive disease was identified in 73 patients (72.3%) (Table 1).

HER2-ultralow could not be assessed separately because historical pathology reports did not consistently document faint, incomplete staining within the IHC 0 category.^[18]

Detailed HER2 distribution

The detailed HER2 categories were evaluated as mutually exclusive cohorts; moreover, the derived HER2-low row represents a summary category and was excluded from the mutually exclusive groupings.

Table 1. Descriptive statistics of the study cohort

Variable	Value
Total patients	101
Age (years), median	53 (45–62)
Follow-up duration (months)	39 (32–46)
Stage I–II	70 (69.3%)
Stage III	20 (19.8%)
Stage IV/metastatic	11 (10.9%)
Premenopausal	33 (32.7%)
Postmenopausal	68 (67.3%)
Grade 1	13 (12.9%)
Grade 2	56 (55.4%)
Grade 3	31 (30.7%)
Grade unknown	1 (1.0%)
HR-positive	73 (72.3%)
HR-negative	28 (27.7%)
HER2-positive	46 (45.5%)
HER2-low	48 (47.5%)
HER2-zero	7 (6.9%)
Triple-negative breast cancer	20 (19.8%)

HR: Hormone receptor; HER2: Human epidermal growth factor receptor 2.

Overall concordance

Full concordance between physician decisions and AI-generated recommendations was observed in 83 patients ($n=83$; 82.2%; 95% CI: 73.6–88.4), whereas partial concordance was documented in 18 patients ($n=18$; 17.8%; 95% CI: 11.6–26.4). Notably, no cases of complete non-concordance were recorded in the final analytic dataset ($n=0$; 0.0%; 95% CI: 0.0–3.7). Consequently, subsequent binary analyses were restricted to a comparison between full and partial concordance (Table 2).

Component-level concordance

Components not applicable were excluded from the component-specific denominators. In this component-level framework, conventional anti-HER2 therapy was considered applicable only for HER2-positive disease (Table 3).

This component-level discordance rate differs from patient-level concordance analysis, in which HER2-positive patients were classified according to the overall treatment strategy rather than anti-HER2 therapy alone. Notably, no statistically significant difference in concordance with anti-HER2 therapy was identified between IHC 2+/ISH-positive and IHC 3+ disease (Fisher exact test, $p=1.000$) (Table 4).

Factors associated with partial concordance

Furthermore, significant variance in concordance was identified across stage groups ($p=0.031$). Full concordance (FC) was observed to be lower in metastatic disease (54.5%) than in Stage I–II disease (87.1%), and Stage III disease (80.0%). Hormone receptor status was not significantly associated with concordance ($p=0.144$). HER2-positive patients showed a numerically higher partial concordance (PC) rate than HER2-negative/HER2-low/HER2-zero patients, although this difference did not reach statistical significance using Fisher's exact test (26.1% vs. 10.9%; $p=0.067$). Moreover, the distribution of concordance was found to differ by guideline year ($p=0.035$), with a higher partial concordance rate in 2024 (Table 5). Patients with partial concordance were not statistically significantly older than those with full concordance (median age: 56 vs. 52 years; $p=0.201$).

Table 2. Overall concordance distribution and component-level agreement

Concordance category	n	%	95% CI
Full concordance	83	82.2	73.6–88.4
Partial concordance	18	17.8	11.6–26.4
Non-concordance	0	0.0	0.0–3.7

CI: Confidence interval.

Table 3. Concordance of individual treatment components between physician decisions and AI-generated recommendations

Component	Applicable (n)	Matched (n) (%)	Discordant (n) (%)
Definitive surgery	90	86 (95.6%)	4 (4.4%)
Chemotherapy	98	98 (100.0%)	0 (0.0%)
Anti-HER2 therapy	46	39 (84.8%)	7 (15.2%)
Endocrine therapy	73	68 (93.2%)	5 (6.8%)
Radiotherapy	75	75 (100.0%)	0 (0.0%)

HER2: Human epidermal growth factor receptor 2.

Table 4. Concordance of anti-HER2 therapy, stratified by HER2-positive status

HER2-positive subgroup	Applicable n	Matched n (%)	Discordant n (%)
IHC 2+/ISH-positive	9	8 (88.9%)	1 (11.1%)
IHC 3+	37	31 (83.8%)	6 (16.2%)
HER2-positive total	46	39 (84.8%)	7 (15.2%)

IHC: Immunohistochemistry; ISH: In situ hybridization; HER2: Human epidermal growth factor receptor 2.

Table 5. Categorical association tests for full versus partial concordance

Variable	Category	FC n (%)	PC n (%)	Total (n)	p
Stage group	Metastatic	6 (54.5%)	5 (45.5%)	11	0.031
	Stage I-II	61 (87.1%)	9 (12.9%)	70	
	Stage III	16 (80.0%)	4 (20.0%)	20	
HR status	Negative	26 (92.9%)	2 (7.1%)	28	0.144
	Positive	57 (78.1%)	16 (21.9%)	73	
HER2 status	HER2-negative/HER2-low/HER2-zero	49 (89.1%)	6 (10.9%)	55	0.067
	HER2-positive	34 (73.9%)	12 (26.1%)	46	
Guideline year	2021	14 (77.8%)	4 (22.2%)	18	0.035
	2022	29 (87.9%)	4 (12.1%)	33	
	2023	28 (90.3%)	3 (9.7%)	31	
	2024	9 (56.2%)	7 (43.8%)	16	
	2025	3 (100.0%)	0 (0.0%)	3	

FC: Full concordance; PC: Partial concordance; HR: Hormone receptor; HER2: Human epidermal growth factor receptor 2.

Exploratory multivariable logistic regression

In the exploratory multivariable logistic regression model, HER2-positive disease showed higher odd ratios (OR) of partial concordance compared with non-HER2-positive disease (adjusted OR, 3.71; 95% CI, 1.09–12.65; $p=0.036$). The 2024–2025 guideline period also showed higher ORs of partial concordance (adjusted OR, 4.73; 95% CI, 1.26–17.78; $p=0.022$). Age and metastatic disease were not statistically significant predictors in the model. Given the limited number of partial-concordance events, these results should be interpreted cautiously and considered hypothesis-gener-

ating (Table 6) (Fig. 1). The dependent variable was partial concordance versus full concordance, and the model included age, HER2 positivity, metastatic disease, and the 2024–2025 guideline period as covariates.

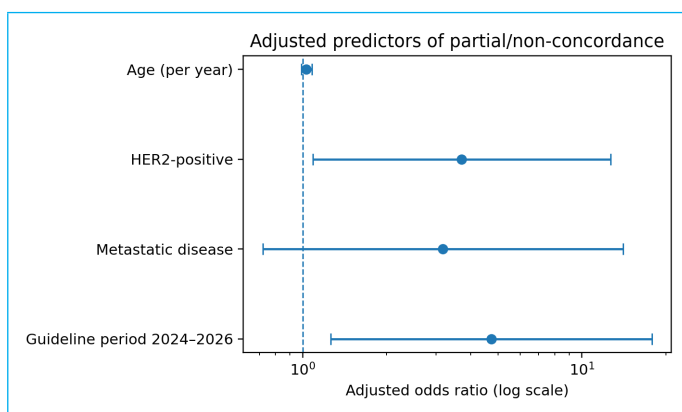
Exploratory rule-based confidence score

The overall median AI confidence score was 9.3 (IQR, 8.5–9.3). Significantly higher AI confidence scores were observed in patients with full concordance, compared to those with partial concordance (median 9.3 [IQR, 8.5–9.3] vs. 8.6 [IQR, 7.8–9.3]; Mann-Whitney U test, $p=0.0238$). High-confidence outputs were identified more frequently

Table 6. Exploratory multivariable logistic regression for partial concordance versus full concordance

Predictor	Adjusted OR	95% CI	p
Age (per year)	1.03	0.99-1.08	0.158
HER2 positivity	3.71	1.09-12.65	0.036
Metastatic disease	3.18	0.72-14.06	0.128
Guideline period 2024-2025	4.73	1.26-17.78	0.022

HER2: Human epidermal growth factor receptor 2.

**Figure 1.** Multivariable analysis of adjusted OR for partial concordance.

within the full concordance group compared to the partial concordance cohort (95.2% vs. 72.2%; Fisher's exact test, $p=0.0084$) (Table 7). The AI confidence score, generated post hoc using a rule-based exploratory framework, should not be interpreted as a prospectively derived model-probability estimate.

Discussion

This retrospective pilot study demonstrated high overall concordance between physician treatment decisions and ChatGPT-generated breast cancer treatment recommendations. Full concordance was observed in 82.2% of patients, partial concordance in 17.8%, and no cases of complete non-concordance. Notably, these findings are consistent with previous evaluations of large language models and clinical decision support systems, which report a wide range of concordance rates.^[19] Strategies involving struc-

tured data, explicit prompt requests for guideline reasoning, orchestrated protocols, and self-refinement have been shown to improve LLM alignment with guideline-oriented treatment logic.^[9] However, as emphasized in methodological analyses of AI in oncology, concordance should not be interpreted as accuracy, superiority, or clinical effectiveness; rather, it remains a measure of agreement within the specific assumptions of the study design rather than proof that the AI-generated recommendation or the physician decision was clinically optimal or led to improved patient outcomes.^[12]

The concordance rate observed in this cohort is broadly consistent with recent literature on breast cancer LLMs. Prior studies have reported that publicly available LLMs may approximate multidisciplinary tumor board decisions in selected breast cancer scenarios, but performance varies by model type.^[7] RAG-enhanced systems have also been evaluated in breast cancer tumor board settings and may improve guideline anchoring and traceability compared with unguided LLM workflows.^[9] In contrast to these systems, the present study utilized a standardized zero-shot ChatGPT workflow without RAG, external guideline documents, or multi-agent orchestration. Consequently, the results should be interpreted as reflecting the feasibility of an accessible prompt-based workflow rather than the performance of a guideline-retrieval engine.

Clinical complexity appeared to influence concordance. Full concordance was lower in metastatic disease than in Stage I-II or Stage III disease, and partial concordance was significantly more frequent in metastatic cases. Notably, this observation is clinically plausible, as the therapeutic management of metastatic breast cancer necessitates the integration of complex clinical features, biomarkers, and patient preferences.^[20] In contrast, many early-stage treatment decisions are more standardized and guided by relatively stable algorithms.^[21] Partial concordance in metastatic disease may therefore reflect the complexity of individualized real-world oncology rather than a failure of physician judgment or AI reasoning.

Notably, the association between HER2-positive disease and partial concordance is clinically significant. HER2-directed treatment has evolved rapidly across neoadjuvant,

Table 7. AI confidence score and physician–AI concordance in the study cohort

AI confidence level	FC n (%)	PC n (%)	Total n (%)
High confidence	79/83 (95.2%)	13/18 (72.2%)	92/101 (91.1%)
Intermediate confidence	4/83 (4.8%)	5/18 (27.8%)	9/101 (8.9%)
Low confidence	0/83 (0.0%)	0/18 (0.0%)	0/101 (0.0%)

AI: Artificial intelligence; FC: Full concordance; PC: Partial concordance.

adjuvant, and metastatic settings, with changes in escalation, de-escalation, sequencing, and ADC strategies.^[22] The emergence of HER2-low disease has further altered the interpretation of HER2 testing. The 2023 American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) update emphasized the clinical relevance of reporting IHC 1+ and IHC 2+/ISH-negative results, while noting that HER2-low is not a new ASCO/CAP interpretive category in the same way as HER2-positive disease.^[23] ESMO expert consensus statements similarly emphasize HER2-low as an emerging therapeutic category, particularly because of ADC in metastatic disease.^[24] In the present study, detailed HER2 reporting was added to distinguish IHC 0, IHC 1+, IHC 2+/ISH-negative, IHC 2+/ISH-positive, and IHC 3+ categories. Nevertheless, the analysis of HER2-ultralow disease was precluded, as historical pathology documentation did not reliably distinguish faint, incomplete staining patterns within the IHC 0 category. This limitation may affect interpretation of contemporary metastatic cases, but it avoids unsupported retrospective reclassification.

The 2024–2025 guideline period was associated with partial concordance. This likely reflects the rapid evolution of breast cancer management in recent years, including expanding indications for cyclin-dependent kinase 4/6 (CDK 4/6) inhibitors, ADC, immunotherapy for selected triple-negative disease, and increasingly biomarker-driven treatment sequencing.^[22] Clinical recommendations have become increasingly sensitive to prior therapies, local access, and patient-specific factors.^[25] Physician treatment decisions were made within the national healthcare context and may have been influenced by national drug approval status, reimbursement conditions, institutional drug availability, patient-specific clinical factors, and patient preferences. These contextual factors were considered when interpreting partial concordance between physician decisions and AI-generated recommendations. While LLM-generated outputs may identify guideline-concordant options, they potentially fail to capture the institutional and socioeconomic context in which real-world clinical decisions are made.

Component-level concordance provides a more nuanced interpretation than overall concordance alone. Chemotherapy and radiotherapy showed strong agreement across applicable cases, suggesting that AI recommendations may align well with established components when indications are clear. Notably, anti-HER2 therapy and endocrine therapy were identified as the primary sources of component-level discordance. These domains represent areas wherein clinical context, tumor biology, menopausal status, treatment sequencing, therapy duration, and drug accessibility strongly influence therapeutic decision-making.^[25]

This finding supports using component-level analysis in future AI concordance studies because it identifies the therapeutic domains where LLM outputs are most vulnerable to ambiguity.

The post hoc confidence analysis further supports the role of clinical ambiguity. High-confidence outputs were more frequent in fully concordant cases, whereas intermediate-confidence outputs were relatively more common in partially concordant cases. Notably, these observations suggest that AI-generated recommendations demonstrate increased concordance with clinician decision-making when guideline indications are clear, clinical evidence is robust, and patient-level data are complete. Conversely, lower or intermediate confidence scores often pinpoint areas where physician judgment and institutional factors become more influential because of inherent clinical uncertainty. Nevertheless, as the confidence score was generated retroactively using a rule-based proxy framework—rather than prospectively—it should be considered exploratory and hypothesis-generating. Future studies should define confidence metrics prospectively and validate them against independent expert assessment and clinical outcomes.

The study also highlights important methodological limitations of LLM evaluation. LLM outputs are sensitive to prompt wording and variations across models. In this study, a standardized prompt and single-output strategy were used to reduce variability, but formal prompt sensitivity testing was not performed. The exact backend model snapshot could not be independently verified because the analysis used the ChatGPT web interface rather than a versioned API endpoint. Moreover, the analysis did not employ any RAG or multi-agent workflow. While this design reflects accessible real-world ChatGPT use, it limits reproducibility and prevents direct source-level verification of guideline passages. Current evidence indicates that RAG may improve grounding of evidence^[26], but RAG systems introduce their own design dependencies, including source selection, retrieval quality, chunking strategy, and validation of retrieved material. The present study should also be interpreted in light of emerging reporting standards for clinical artificial intelligence research. Developmental and Exploratory Clinical Investigation of Decision-support Systems Driven by Artificial Intelligence (DECIDE-AI) was developed for the developmental and exploratory clinical evaluation of AI-driven decision-support systems and provides AI-specific reporting items intended to improve transparency regarding key information items; although the present study was retrospective and did not involve live deployment or real-time clinician interaction, these principles are directly relevant to its methodological interpreta-

tion.^[27] Relevant considerations were identified to encompass the AI workflow, input variables, prompt structure, model access route, historical guideline-matching strategy, the absence of RAG, and limitations related to reproducibility and model updating. By contrast, Standards for Reporting Diagnostic Accuracy Studies–Artificial Intelligence (STARD-AI) was developed primarily for AI-centered diagnostic accuracy studies and is therefore not directly applicable to the present treatment-concordance design. Nevertheless, selected STARD-AI principles remain conceptually useful for strengthening transparency.^[28] Relevant areas for transparency included the dataset, AI system, evaluation framework, risk of bias, applicability, and generalizability. Future prospective studies should be designed according to dedicated AI reporting frameworks to improve reproducibility and facilitate comparison across studies.^[29]

In the context of breast cancer management, the role of AI systems is best characterized as supportive rather than substitutive. AI systems may assist multidisciplinary tumor boards by organizing clinical information, identifying missing variables, summarizing guideline-based options, and highlighting acceptable alternative strategies.^[30] However, final decisions should remain the responsibility of expert clinicians because AI systems cannot fully account for nuanced clinical examination, patient values, treatment tolerance, local reimbursement policies, or evolving real-world constraints.^[31,32] Notably, previous studies indicate that discordance between AI recommendations and clinical practice is frequently driven by factors including drug availability, cost, and patient preferences, which are not uniformly captured in standardized guideline datasets.^[32] The absence of complete concordance in this cohort should not be interpreted as validation for autonomous AI use. Rather, it supports further prospective evaluation of AI-assisted decision support under controlled, transparent, and outcome-linked conditions.

Limitations

This study has several limitations. Initially, this retrospective pilot study was characterized by a single-center design, a restricted sample size, and a low frequency of partial-concordance events. As noted in similar clinical decision support evaluations, single-center studies may reflect local institutional practice patterns and specific reimbursement environments, limiting broader generalizability.^[33] Although exploratory multivariable logistic regression was performed, the model was intentionally restricted to a small number of clinically relevant covariates and should be interpreted as hypothesis-generat-

ing rather than confirmatory. This approach is consistent with the preliminary nature of pilot AI evaluations, which prioritize identifying potential trends and associations to inform future, larger-scale prospective validations.^[34]

Second, the historical NCCN/ESMO matching strategy was based on prompt-level instructions and the estimated guideline year. Notably, integration with archived guideline repositories via RAG was not established; this methodology represents a limitation, as RAG-enhanced frameworks have been demonstrated to significantly improve NCCN-concordant breast cancer treatment recommendations compared to standard prompting. Because the model relied on its internal clinical knowledge encoding rather than a multi-agent workflow or an external guideline repository, it cannot be guaranteed to remain isolated from contemporary medical knowledge (or “knowledge drift”), as models frequently exhibit shifts in behavior and internal knowledge over time.^[35]

Third, formal interobserver reliability analysis using Cohen’s kappa could not be performed because separate blinded patient-level classifications from two independent reviewers were unavailable. In clinical research, the kappa statistic is the standard metric for assessing how much observed agreement among raters exceeds the level expected by chance alone. Therefore, in the absence of such quantitative validation, the concordance categories in this study should be interpreted as consensus-based rather than independently verified. This is a recognized limitation in early-phase AI evaluations, as blinded, multi-expert assessment is increasingly recommended to minimize subjective bias and ensure the robustness of AI-physician agreement metrics.^[36]

Fourth, HER2-low status was defined using IHC 1+ and IHC 2+/*ISH*-negative criteria, but HER2-ultralow disease could not be systematically evaluated, as historical pathology reports lacked the requisite granularity for IHC 0 cases. This may be particularly relevant to contemporary metastatic treatment decisions, as the distinction between HER2-negative (IHC 0/null) and HER2-ultralow/low has become a key driver of eligibility for novel antibody-drug conjugates (ADC) like trastuzumab deruxtecan in advanced settings. The rapid evolution of these pathological categories means that retrospective datasets often lack the granularity required to align with the most recent international guidelines, such as the 2023 ASCO/CAP updates, which now incorporate HER2-low and ultralow as distinct therapeutic targets.^[37] Studies indicate that up to 65.8% of cases originally scored as IHC 0 may harbor clinically actionable HER2 expression when re-evaluated for ultralow or low status, highlight-

ing the limitations of historical reporting in the era of personalized medicine.

Fifth, radiology, pathology, genomic testing, Eastern Cooperative Oncology Group (ECOG) status, comorbidities, frailty, treatment access, reimbursement restrictions, and patient preference may have been incompletely captured in the retrospective dataset.^[38] Notably, research has shown that discordance between AI recommendations and clinical practice is frequently driven by factors such as drug availability, cost, regional insurance systems, and patient-specific values, which are not always fully represented in standardized guideline-based datasets. Finally, this analysis did not evaluate survival, toxicity, quality of life, patient satisfaction, cost-effectiveness, or treatment efficacy.^[39] Therefore, concordance should not be interpreted as a measure of accuracy, superiority, or clinical effectiveness.

Conclusion

In this retrospective pilot cohort, physician treatment decisions and AI-generated breast cancer treatment recommendations showed substantial concordance. However, these findings should be interpreted cautiously because of the single-center retrospective design, which may limit broader generalizability and reflect local institutional biases^[33], the limited number of partial-concordance events, the absence of independent interobserver validation using metrics such as Cohen's kappa to adjust for chance agreement^[17], and the lack of outcome-based evaluation. Notably, partial concordance was more frequently observed in clinically complex cases, particularly metastatic disease and contemporary guideline periods^[19], underscoring the influence of significant intertumoral heterogeneity and evolving treatment standards on AI-generated recommendations.^[20] AI-generated recommendations may support guideline-oriented discussion as a second-opinion tool, but they require expert oncologist interpretation, prospective validation, and a focus on human-AI collaboration before routine clinical implementation to ensure they augment rather than replace multidisciplinary human judgment.^[40]

Disclosures

Ethics Committee Approval: The study was approved by the Ordu University Non-Interventional Clinical Research Ethics Committee (no: 318, date: 26.09.2025).

Informed Consent: Informed consent was waived due to the retrospective nature of the study and the use of anonymized patient data.

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

Funding: The authors declared no funding.

Use of AI for Writing Assistance: The authors declared that they used ChatGPT 5.5 (OpenAI, San Francisco, CA, USA) for retrospective AI-assisted comparative analysis and language-editing support.

Authorship Contributions: Concept – EU; Design – MS; Supervision – MS; Fundings – MS; Materials – MS; Data Collection or Processing: MS; Analysis or Interpretation: MS; Literature Search: MS; Writing: EU; Critical Review: EU.

Peer-review: Externally peer-reviewed.

Acknowledgments: The authors thank the medical oncology clinic staff of Ordu State Hospital for their assistance in the organization of retrospective clinical data.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74(3):229–63. [\[CrossRef\]](#)
2. Ellsworth RE, Decewicz DJ, Shriver CD, Ellsworth DL. Breast cancer in the personal genomics era. *Curr Genomics* 2010;11(3):146–61. [\[CrossRef\]](#)
3. Gradishar WJ, Moran MS, Abraham J, Abramson V, Aft R, Agnese D, et al. Breast Cancer, Version 3.2024, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2024;22(5):331–57. [\[CrossRef\]](#)
4. Borstnar S, Bozovic-Spasojevic I, Cvetanovic A, Plavetic ND, Konsoulova A, Matos E, et al. Advancing HER2-low breast cancer management: enhancing diagnosis and treatment strategies. *Radiol Oncol* 2024;58(2):258–67. [\[CrossRef\]](#)
5. Otte I, Salloch S, Reinacher-Schick A, Vollmann J. Treatment recommendations within the leeway of clinical guidelines: A qualitative interview study on oncologists' clinical deliberation. *BMC Cancer* 2017;17(1):780. [\[CrossRef\]](#)
6. Hao Y, Qiu Z, Holmes J, Löckenhoff CE, Liu W, Ghassemi M, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *NPJ Digit Med* 2025;8(1):450. [\[CrossRef\]](#)
7. Griewing S, Knitza J, Boekhoff J, Hillen C, Lechner F, Wagner U, et al. Evolution of publicly available large language models for complex decision-making in breast cancer care. *Arch Gynecol Obstet* 2024;310(1):537–50. [\[CrossRef\]](#)
8. Ferber D, Wiest IC, Wölflein G, Ebert MP, Beutel G, Eckardt JN, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI* 2024;1(6):Alcs2300235. [\[CrossRef\]](#)
9. Liu Q, Hu Z, Huang T, Niu Y, Zhang X, Ma S, et al. EvoMDT: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *NPJ Digit Med* 2026;9(1):124. [\[CrossRef\]](#)
10. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30(9):2613–22. [\[CrossRef\]](#)

11. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in senology-an assessment of concordance with breast cancer tumor board decision making. *J Pers Med* 2023;13(10):1502. [\[CrossRef\]](#)
12. Tupasela A, Di Nucci E. Concordance as evidence in the Watson for Oncology decision-support system. *AI Soc* 2020;35(4):811–8. [\[CrossRef\]](#)
13. Tarantino P, Curigliano G, Tolaney SM. Navigating the HER2-low paradigm in breast oncology: new standards, future horizons. *Cancer Discov* 2022;12(9):2026–30. [\[CrossRef\]](#)
14. Abbasvandi F, Bayat M, Akbari A, Shojaeian F, Zandi A, Rahmani J, et al. Tumor characteristics and survival rate of HER2-low breast cancer patients: a retrospective cohort study. *Sci Rep* 2023;13(1):16719. [\[CrossRef\]](#)
15. Singareeka Raghavendra A, Liu DD, Damodaran S, Pasyar S, Shen Y, Mouabbi JA, et al. Prevalence and outcomes of HER2-low versus HER2-0 status in patients with metastatic breast cancer. *Cancers (Basel)* 2026;18(2):253. [\[CrossRef\]](#)
16. Bierbaum M, Rapport F, Arnolda G, Delaney GP, Liauw W, Oliver I, et al. Clinical practice guideline adherence in oncology: A qualitative study of insights from clinicians in Australia. *PLoS One* 2022;17(12):e0279116. [\[CrossRef\]](#)
17. Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer* 2023;23(1):799. [\[CrossRef\]](#)
18. Lv H, Yue J, Zhang Q, Xu F, Gao P, Yang H, et al. Prevalence and concordance of HER2-low and HER2-ultralow status between historical and rescored results in a multicentre study of breast cancer patients in China. *Breast Cancer Res* 2025;27(1):45. [\[CrossRef\]](#)
19. Oehring R, Ramasetti N, Ng S, Roller R, Thomas P, Winter A, et al. Use and accuracy of decision support systems using artificial intelligence for tumor diseases: a systematic review and meta-analysis. *Front Oncol* 2023;13:1224347. [\[CrossRef\]](#)
20. Singareeka Raghavendra A, Damodaran S, Barcenas CH, Fuqua SA, Layman RM, Tripathy D. Personalizing therapies over the course of hormone receptor-positive/HER2-negative metastatic breast cancer. *CA Cancer J Clin* 2026;76(1):e70055. [\[CrossRef\]](#)
21. Ayala de la Peña F, Antolín Novoa S, Gavilá Gregori J, González Cortijo L, Henao Carrasco F, Martínez Martínez MT, et al. SEOM-GEICAM-SOLTI clinical guidelines for early-stage breast cancer (2022). *Clin Transl Oncol* 2023;25(9):2647–64. Erratum in: *Clin Transl Oncol* 2024;26(2):555–6. [\[CrossRef\]](#)
22. Corti C, Batra-Sharma H, Kelsten M, Shatsky RA, Garrido-Castro AC, Gradishar WJ. Systemic therapy in breast cancer. *Am Soc Clin Oncol Educ Book* 2024;44(3):e432442. [\[CrossRef\]](#)
23. Wolff AC, Somerfield MR, Dowsett M, Hammond MEH, Hayes DF, McShane LM, et al. Human epidermal growth factor receptor 2 testing in breast cancer: ASCO-College of American Pathologists Guideline update. *J Clin Oncol* 2023;41(22):3867–72. [\[CrossRef\]](#)
24. Tarantino P, Viale G, Press MF, Hu X, Penault-Llorca F, Bardia A, et al. ESMO expert consensus statements (ECS) on the definition, diagnosis, and management of HER2-low breast cancer. *Ann Oncol* 2023;34(8):645–59. [\[CrossRef\]](#)
25. Dawood S, Konstantionva M, Dent R, Perazzo F, Kim SB, Villarreal-Garza C, et al. Optimizing treatment selection, and sequencing decisions for Management of HR-Positive, HER2-Negative advanced breast cancer - Proceedings from breast cancer expert group meeting. *BMC Proc* 2021;15(Suppl 10):15. [\[CrossRef\]](#)
26. Managwu C, Shittu L, Obi-Nwankpa D. A hybrid retrieval-augmented generation and language model framework for evidence-grounded review systems. *Int J Adv Res Comp Sci* 2025;16(5):32–6. [\[CrossRef\]](#)
27. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27(2):186–7. [\[CrossRef\]](#)
28. Klontzas ME, Gatti AA, Tejani AS, Kahn CE Jr. AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell* 2023;5(3):e230055. [\[CrossRef\]](#)
29. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med (Lond)* 2024;4(1):71. [\[CrossRef\]](#)
30. Wang X, Wang Q, Ding G, Wang J, Tang Y, Feng Y. Artificial intelligence in multidisciplinary tumor boards enhancing decision making and clinical outcomes in oncology. *iScience* 2025;28(12):114082. [\[CrossRef\]](#)
31. Hindocha S, Badea C. Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare. *AI Ethics* 2022;2(1):167–75. [\[CrossRef\]](#)
32. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *NPJ Digit Med* 2022;5(1):162. [\[CrossRef\]](#)
33. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. [\[CrossRef\]](#)
34. Idlahcen F, Idri A, Goceri E. Exploring data mining and machine learning in gynecologic oncology. *Artif Intell Rev* 2024;57(2):20. [\[CrossRef\]](#)
35. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *Harv Data Sci Rev* 2024;6(2):99608F92.5317da47. [\[CrossRef\]](#)
36. Chen PC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health* 2021;3(11):e693–5. [\[CrossRef\]](#)
37. Rakha EA, Tan PH, Van Bockstal MR, Allison KH, Brogi E, Callagy G, et al; UK National Coordinating Committee of Breast Pathology, the Association of Breast Pathology, the European

- Working Group for Breast Screening Pathology and the International Society of Breast Pathology. International expert consensus recommendations for HER2 reporting in breast cancer: focus on HER2-low and ultralow categories. *Mod Pathol* 2026;39(1):100925. [\[CrossRef\]](#)
38. Chang WH, Lai AG. Pan-cancer analyses of the associations between 109 pre-existing conditions and cancer treatment patterns across 19 adult cancers. *Sci Rep* 2024;14(1):464. [\[CrossRef\]](#)
39. Xu W, Wang X, Yang L, Meng M, Sun C, Li W, et al. Consistency of CSCO AI with multidisciplinary clinical decision-making teams in breast cancer: a retrospective study. *Breast Cancer* (Dove Med Press) 2024;16:413–22. [\[CrossRef\]](#)
40. Chua IS, Gaziel-Yablowitz M, Korach ZT, Kehl KL, Levitan NA, Arriaga YE, et al. Artificial intelligence in oncology: Path to implementation. *Cancer Med.* 2021;10(12):4138–49. [\[CrossRef\]](#)

Appendix A. Standardized AI prompt template

You are an oncology clinical decision-support assistant specializing in breast cancer treatment planning. Analyze the anonymized patient data below and generate a treatment recommendation based on the NCCN and/or ESMO breast cancer guideline framework applicable to the estimated year of the original treatment decision. Do not retrospectively apply treatment standards that were not available or guideline-supported at that time.

Patient-level variables: year of diagnosis or treatment decision; age; menopausal status; ECOG performance status, if available; comorbidities, if available; histology; T stage; N stage; M stage; overall stage; tumor grade; ER status; PR status; HER2 IHC score (0, 1+, 2+, or 3+); ISH result when IHC is 2+; HER2-low or HER2-ultralow status, if explicitly reported; Ki-67; lymphovascular invasion; genomic assay result, if available; surgery; chemotherapy; anti-HER2 therapy; endocrine therapy; radiotherapy; and other systemic treatments.

The output requirements were defined as follows: (1) specify the estimated guideline year; (2) provide a structured treatment recommendation that includes surgery, chemotherapy, endocrine therapy, anti-HER2 therapy, radiotherapy, and other relevant systemic treatment components; (3) compare the AI-generated recommendation with the actual physician treatment; and (4) classify the relationship as full concordance, partial concordance, or non-concordance, with a brief justification. If divergence is clinically reasonable, discuss possible explanations, such as age, comorbidity, performance status, toxicity concerns, patient preference, reimbursement restrictions, treatment availability, or real-world practice constraints.

RAG status: no external documents, databases, real-time web retrieval, uploaded guideline library, or vector database is connected to the model. Recommendations are generated solely from the structured prompt and the model's internal knowledge.