Analysis of Large Language Model Decision Making in Hormone Receptor–Positive/Human Epidermal Growth Factor Receptor 2–Negative Early Breast Cancer

Roberto Buonaiuto, MD[1,2,3]; Aldo Caltavituro, MD[2,3]; Rossana Di Rienzo, MD[1,2]; Angela Grieco, MD[3]; Federica P. Mangiacotti, MD[3]; Alessandra Longobardi, MD[3]; Vincenza Cantile, MD[3]; Vittoria Molinaro, MD[3]; Martina Pagliuca, MD[2,4]; Giuseppe Buono, MD[1]; Pietro De Placido, MD[5,6,7]; Erica Pietroluongo, MD[3,8]; Valeria Forestieri, MD[3]; Claudia Martinelli, MD[1,2]; Vincenzo di Lauro, MD[1]; Luigi Leo, MD[9]; Massimiliano D'Aiuto, MD[10]; Giampaolo Bianchini, MD[11,12]; Carmen Criscitiello, MD[13,14]; Roberto Bianco, MD[3]; Lucia Del Mastro, MD[15,16]; Michelino De Laurentiis, MD[1,2]; Grazia Arpino, MD[3]; Carmine De Angelis, MD[1,2]; and Mario Giuliano, MD[3]

DOI https://doi.org/10.1200/CCI-25-00230

ABSTRACT

PURPOSE To assess the ability of GPT-4o in adjuvant treatment decision making in hormone receptor–positive (HR+)/human epidermal growth factor receptor 2–negative (HER2−) early breast cancer by comparing its recommendations with those of clinicians including Oncotype DX data, and to explore its potential as a decision-support tool in routine clinical practice.

METHODS We compared clinician and GPT-4o recommendations in patients tested with Oncotype DX in routine practice at the University of Naples Federico II (n = 607, cohort 1 [C1]) and within the prospective, multicenter PRO BONO study (n = 237, cohort 2 [C2]). Pre– and post–Oncotype DX treatment recommendations were categorized as chemotherapy (CT) + endocrine therapy (ET) or ET alone. Concordance between clinician and GPT-4o recommendations was assessed using agreement rates and Cohen's kappa. The accuracy of Oncotype DX results was evaluated using the AUC metric.

RESULTS The agreement between clinicians and GPT-4o in pretest recommendations was 68% (kappa, 0.381 [95% CI, 0.31 to 0.45], $P < .001$) in C1 and 70% (0.401 [95% CI, 0.29 to 0.52], $P < .001$) in C2. Before Oncotype DX, clinicians recommended CT more frequently than GPT-4o for C1 (58% *v* 38%) and C2 (53% *v* 43%). Post-test agreement increased to 93% (0.814 [95% CI, 0.76 to 0.87], $P < .001$) in C1 and 90% (0.741 [95% CI, 0.64 to 0.84], $P < .001$) in C2. The agreement between pre– and post–Oncotype DX treatment recommendations for clinicians was 56% and 63% versus 68% and 60% for GPT-4o in C1 and C2, respectively. GPT-4o showed higher accuracy in predicting low than high genomic risk in postmenopausal patients (87% *v* 43% in C1; 85% *v* 45% in C2, $P < .001$) and low versus intermediate and high risk in premenopausal patients in both cohorts ($P < .001$).

CONCLUSION The agreement between clinicians and GPT-4o in pretest recommendations was modest but improved post-test, highlighting the importance of multigene testing and the potential of large language models in clinical decision making.

ACCOMPANYING CONTENT

Data Supplement

Accepted January 14, 2026
Published March 6, 2026

JCO Clin Cancer Inform
10:e2500230

BACKGROUND

Artificial intelligence (AI) is progressively transforming health care by corroborating and enhancing physicians' capabilities in complex clinical scenarios, particularly in screening, diagnosis, and treatment.[1,2] Similarly to various medical specialties, AI is rapidly emerging as a pivotal tool in radiology,[3-6] oncology,[7] and pathology.[8-13] Despite this, the integration of AI into treatment decision making remains challenging since this process is inherently multifactorial, necessitating the interplay among physician expertise, tumor biology, and patient characteristics and preferences.

Among AI models, next-generation pretrained large language models (LLMs)[14] represent a practical and user-friendly approach with the ability to generate human-like responses across multiple medical domains.[15] LLMs could assist clinicians in summarizing data, answering clinical

CONTEXT

Key Objective

Could large language models (LLMs), such as GPT-4o, represent potentially useful tools to assist clinical decision making in patients with hormone receptor–positive (HR+)/human epidermal growth factor receptor 2–negative (HER2–) early breast cancer (eBC) undergoing multigene test OncotypeDX?

Knowledge Generated

GPT-4o showed capability to integrate clinicopathologic variables with genomic data in patients with HR+/HER2– eBC. Before multigene testing, the agreement between GPT-4o and clinicians' treatment recommendation was moderate (70%) but increased to more than 90% after OncotypeDX results.

Relevance *(F. Lin)*

This multicenter study provides data that reinforce the importance of exercising discretion when using output from generalist LLMs as decision aids for adjuvant therapy in eBC, particularly when recommendations are derived from clinicopathological variables alone.*

*Relevance section written by *JCO Clinical Cancer Informatics* Deputy Editor Frank Lin, MB ChB, PhD, FRACP, FAIDH.

questions, and providing treatment recommendations supporting medical decision making.[16]

However, their clinical integration presents greater challenges depending on the quality of available patients' and tumor-specific data, raising criticisms about their advantages, limitations, and feasibility in daily practice.

To evaluate the clinical readiness of LLMs and their potential role in supporting medical decision making, we conducted a comparative analysis between LLM-assisted and physician-guided decisions in a commonly encountered clinical scenario: patients with hormone receptor–positive/human epidermal growth factor receptor 2–negative (HR+/HER2–) early breast cancer (eBC), eligible for adjuvant therapy, and undergoing multigene testing. Specifically, we assessed GPT-4o's recommendations regarding multigene testing indications, test selection, and treatment decisions both before and after testing. Additionally, we evaluated the concordance in treatment recommendations between clinicians and GPT-4o pretest, and the agreement between pre- and post-test recommendations provided by both clinicians and GPT-4o. Finally, we assessed GPT-4o's capacity to predict multigene test results.

METHODS

Study Objectives

This study aimed to evaluate the performance of GPT-4o in formulating adjuvant treatment recommendations in patients with HR+/HER2– eBC. The analysis focused on three main areas:

1. Indications for multigene testing and selection of test type by GPT-4o.
2. Pre- and post-test therapy recommendations provided by GPT-4o compared with those proposed by clinicians.
3. Prediction of multigene test results performed by GPT-4o.

Population

The study included 844 consecutive patients with HR+/HER2– eBC who underwent surgery and subsequent Oncotype DX testing recommended by their physicians at the University of Naples Federico II (cohort 1 [C1]) or within the PRO BONO study[17] (cohort 2 [C2]). Pre- and post-test clinical decisions were prospectively collected for all patients by multidisciplinary oncology teams within multi-institutional health care networks (C1) and academic hospitals (C2). Data on clinicians' pre- and post-test decisions were collected for each patient, including indications for multigene testing, the specific test used, and treatment recommendations. Treatment recommendations were categorized as chemotherapy (CT) followed by endocrine therapy (CT-ET) or ET alone, with additional details on the type of CT and ET prescribed.

Question Format for LLM Treatment Indications and Prediction Assessment

To evaluate the predictive performance of the model, a series of structured, clinically oriented questions were developed. A zero-shot prompting strategy was used. In this approach, the model generated responses based solely on its preexisting knowledge and the information contained within each prompt.[18] The questions were designed to assess the indications for adjuvant therapy as illustrated in Figure 1. This approach was chosen to evaluate how ChatGPT performs, when applied to routine clinical decision-making tasks.

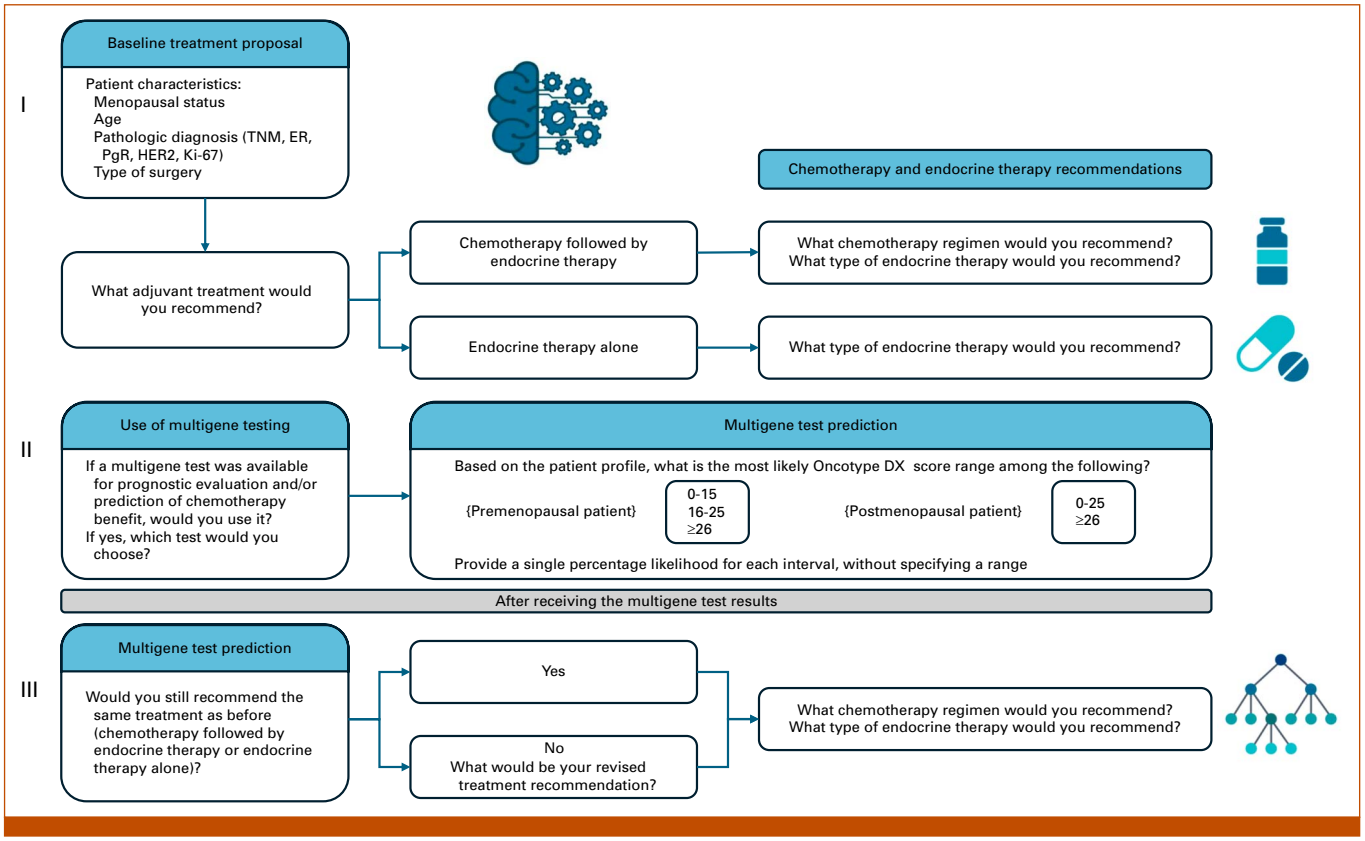


FIG 1. Question format for large language model treatment indications and prediction assessment (zero-shot prompting model). ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PgR, progesterone receptor.

Statistical Analysis

To evaluate the agreement between GPT-4o and clinicians, treatment recommendations were compared in terms of overall strategy (CT-ET *v* ET alone) and the specific types of CT and ET suggested.

Interdecision reliability was assessed using agreement rates and Cohen's kappa coefficient. Separate analyses were performed for C1 and C2 to identify potential differences in the performance of clinical decisions compared with LLM-assisted recommendations within different frameworks.

To further characterize the dynamics of agreement between clinicians and LLM, we evaluated two patterns where only one of the two decision-makers modified their recommendation between the pre- and post-test settings: cases where clinicians changed their recommendation while the LLM's recommendation remained stable (clinician-shift/stable LLM) and vice versa (LLM-shift/stable clinician).

Furthermore, GPT-4o predictions of genomic test results were expressed as percentage probabilities for specific risk categories:

- Postmenopausal patients: High risk (recurrence score [RS] ≥26) versus low risk (RS ≤25).
- Premenopausal patients: High risk (RS ≥26), intermediate risk (RS 16-25), and low risk (RS ≤15).

The accuracy of GPT-4o genomic prediction was evaluated using the AUC metric. The statistical association between the predicted and observed test results was analyzed using the chi-square test.

All statistical analyses were performed with R software (version 4.4.1).

Error Analysis

A human-in-the-loop approach was used[19,20] for error analysis to assess significant discrepancies produced by the LLM model, including hallucinations, inconsistencies, inappropriate results, and reported references. Further information and methods are detailed in the Data Supplement.

Ethics Approval

This study was conducted in accordance with the ethical standards outlined in the Declaration of Helsinki. Ethical approval was obtained from Comitato Etico Campania 3 (reference number, Prot. 331/2024). Written informed consent was obtained from participants before their inclusion in the study.

RESULTS

Baseline Characteristics and Indication for Multigene Testing

A total of 844 patients with HR+/HER2− eBC were included. Among them, 607 (68.7%) patients were treated by local multidisciplinary oncology teams (C1), whereas the remaining 237 patients (31.3%) were enrolled in the PRO BONO study[17] (C2). All patients underwent Oncotype DX testing as recommended by their treating physicians' recommendation. Patient characteristics, including age, menopausal status, nodal involvement, and multigene testing results, are summarized in Table 1.

For all patients (n = 844), GPT-4o prospectively confirmed physicians' indication for genomic testing, with the Oncotype DX test being the only multigene assay selected among all the available assays.

Pretest Treatment Recommendations

In both cohorts, GPT-4o recommended ET alone more frequently than clinicians (Tables 2 and 3). In C1, GPT-4o recommended ET for 375 patients (62%), compared with 255 patients (42%) recommended by clinicians (Table 2). Similarly, in C2, GPT-4o recommended ET for 135 patients (57%), whereas academic clinicians recommended ET for 110 patients (46%; Table 3). Detailed data on specific ET regimens were only available for C1 (Table 2). GPT-4o most frequently selected aromatase inhibitors (ArI; 265 patients, 44%), followed by tamoxifen (TAM; 61 patients, 10%) and TAM plus luteinizing hormone–releasing hormone analog (LHRHa; 48 patients, 8%). In contrast, clinicians recommended ArI for 196 patients (32%), ArI plus LHRHa for 47 patients (8%), and TAM plus LHRHa for nine patients (2%). Conversely, CT-ET was more frequently recommended by clinicians. The most selected CT regimen in C1 was docetaxel plus cyclophosphamide (TC; 186 patients, 31%), followed by doxorubicin plus cyclophosphamide and paclitaxel (AC-T; 167 patients, 27%). In contrast, GPT-4o recommended TC for 42 patients (7%) and AC-T for 168 patients (28%).

Post-Test Treatment Recommendations

After multigene testing, both GPT-4o and clinicians consistently recommended ET in C1 (461 patients, 76%; 460 patients, 76%) and C2 (170 patients, 72%; 178 patients, 75%; Tables 2 and 3). Among the ET regimens (Table 2), ArI monotherapy remained the most frequently selected option, as recommended by both clinicians (332 patients, 55%) and GPT-4o (326 patients, 54%). ArI plus LHRHa was the second choice among clinicians (99 patients, 17%), whereas GPT-4o more frequently recommended TAM (65 patients, 11%) and TAM plus an LHRHa (56 patients, 9%). CT was less frequently recommended by clinicians in both C1 (146 patients, 24%) and C2 (67 patients, 28%; Tables 2 and 3). GPT-4o also had a lower CT recommendation rate, suggesting CT for 146 patients (24%) across both cohorts (Tables 2 and 3). Among CT regimens, epirubicin/doxorubicin plus cyclophosphamide/AC + T was the most frequently selected regimen, favored by both clinicians (82 patients, 13%) and GPT-4o (97 patients, 16%; Table 2).

TABLE 1. Baseline Patients' Characteristics in Cohort 1 and Cohort 2

Characteristic	C1, No. (%)	C2, No. (%)	*P*
Age, years			.01
>50	411 (67.7)	139 (58.6)	
<50	196 (32.3)	98 (41.4)	
Menopausal status			.001
Peri/pre	189 (31)	108 (45.6)	
Post	418 (69)	129 (54.4)	
Histologic subtypes			.98
NST (CDI)	505 (83.2)	198 (83.6)	
CLI	71 (11.7)	28 (11.8)	
Other	31 (5.1)	11 (4.6)	
Tumor size			.93
T1	463 (76.3)	183 (77.2)	
T2	142 (23.4)	53 (22.4)	
T3	2 (0.3)	1 (0.4)	
Nodes			.13
N0	461 (76.0)	176 (74.3)	
N1	133 (21.9)	60 (25.3)	
Nx	13 (2.1)	1 (0.4)	
Ki67, %			.10
<20	168 (27.7)	52 (21.9)	
≥20	439 (72.3)	185 (78.1)	
ER, %			.03
<50	20 (3.3)	1 (0.4)	
≥50	587 (96.7)	236 (99.6)	
Grade			.09
G1	46 (7.6)	9 (3.8)	
G2	327 (53.9)	140 (59.1)	
G3	234 (38.5)	88 (37.1)	
PgR, %			.45
<50	187 (30.8)	80 (33.8)	
≥50	420 (69.2)	157 (66.2)	
Oncotype RS			.45
≤25	497 (81.9)	188 (79.3)	
≥26	110 (18.1)	49 (20.7)	

Abbreviations: C1, cohort 1; C2, cohort 2; ER, estrogen receptor; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; NST, no special type; Nx, unknown nodal status; PgR, progesterone receptor; RS, recurrence score.

Concordance Between LLM and Clinicians' Treatment Recommendations

The agreement between GPT-4o and clinicians' pretest treatment recommendations (CT-ET *v* ET) was 68% (kappa,

TABLE 2. Pre- and Post-Test Treatment Recommendations in Cohort 1

	Pretest		Post-Test	
Treatment	LLM, No. (%)	Clinician, No. (%)	LLM, No. (%)	Clinician, No. (%)
ET	375 (62)	255 (42)	460 (76)	461 (76)
ArI	265 (44)	196 (32)	326 (54)	332 (55)
TAM	61 (10)	1 (0.2)	65 (11)	2 (0.2)
TAM + LHRHa	48 (8)	9 (2)	56 (9)	29 (5)
ArI + LHRHa	2 (0.3)	47 (8)	7 (1)	99 (17)
CT-ET	232 (38)	352 (58)	147 (24)	146 (24)
TC	42 (7)	186 (31)	43 (7)	64 (11)
EC-T	168 (28)	167 (27)	97 (16)	82 (13)
FEC-T	21 (4)	—	7 (1)	—

Abbreviations: ArI, aromatase inhibitor; CT, chemotherapy; EC-T, epirubicin/cyclophosphamide-paclitaxel; ET, endocrine therapy; FEC-T, fluorouracil/epirubicin/cyclophosphamide-docetaxel; LHRHa, luteinizing hormone–releasing hormone analog; LLM, large language model; TAM, tamoxifen; TC, docetaxel-cyclophosphamide.

0.381 [95% CI, 0.31 to 0.45], $P < .001$) in C1 and 70% (kappa, 0.401 [95% CI, 0.29 to 0.52], $P < .001$) in C2 (Table 4). After multigene testing, concordance improved significantly to 93% (kappa, 0.814 [95% CI, 0.76 to 0.87], $P < .001$) in C1 and 90% (kappa, 0.741 [95% CI, 0.64 to 0.84], $P < .001$) in C2 (Table 4). In C1, the concordance between clinicians' pretest and post-test recommendations was 56% (kappa, 0.189 [95% CI, 0.13 to 0.25], $P < .001$), whereas GPT-4o showed slightly higher consistency, with a 68% agreement rate (kappa, 0.262 [95% CI, 0.19 to 0.34], $P < .001$; Table 4). In C2, clinicians exhibited 63% agreement (kappa, 0.284 [95% CI, 0.17 to 0.38], $P < .001$) between pre- and post-test recommendations, whereas GPT-4o demonstrated 60% agreement (kappa, 0.14 [95% CI, 0.01 to 0.25], $P < .001$; Table 4). Changes in treatment recommendations in both cohorts are detailed as alluvial plots in the Data Supplement.

Agreement Between Pretest and Post-Test Decisions Provided by LLM

In C1, GPT-4o agreement was higher in N0 patients compared with N+ patients (70% *v* 52%, $P = .001$). Although numerically higher in patients with RS ≤25 compared with those with RS ≥ 26 (60% *v* 60%, $P = .08$), the difference was not statistically significant. GPT-4o agreement was independent of age ($P = .608$). Similarly, in C2, GPT-4o demonstrated significantly higher agreement in treatment recommendations for patients with node-negative (N0) compared with those with node-positive (N+) tumors (69% *v* 40%, $P = .001$). Concordance did not differ between patients with an Oncotype DX RS <25 and those with RS ≥25 ($P = .834$). GPT-4o agreement was also higher in patients age ≤50 years compared with those age >50 years (67% *v* 59%, $P = .001$).

Agreement Between Pretest and Post-Test Decisions Provided by Clinicians

In C1, clinician agreement was higher for patients with N0 than for those with N+ tumors (56% *v* 41%, $P = .002$). In contrast to GPT-4o, clinician agreement was significantly higher for patients with RS ≥26 than for those with RS ≤25 (67% *v* 50%, $P = .001$) but remained independent of age ($P = .741$). In C2, clinician agreement was also significantly higher for patients with RS ≥26 than for those with RS ≤25 (75% *v* 60%, $P = .04$). However, in contrast to the data from C1, clinician agreement in C2 was independent of age and nodal status ($P = .846$).

Determinants of Treatment Decisions

In C1, the disagreement between GPT-4o and clinicians on pretest decisions was independent of nodal status (N0 *v* N+, $P = .407$) and RS (≥26 *v* ≤25, $P = .658$). A nonsignificant trend was observed for age (>50 *v* ≤50 years, 30% *v* 37.6%, $P = .06$). Conversely, in post-test decisions, disagreement remained

TABLE 3. Pre- and Post-Test Treatment Recommendations in Cohort 2

	Pretest		Post-Test	
Treatment	LLM, No. (%)	Clinician, No. (%)	LLM, No. (%)	Clinician, No. (%)
ET	135 (57)	110 (46)	178 (75)	170 (72)
CT-ET	102 (43)	127 (54)	59 (25)	67 (28)

Abbreviations: CT, chemotherapy; ET, endocrine therapy; LLM, large language model.

TABLE 4. Concordance Between LLM and Clinicians' Decision in Cohort 1 and Cohort 2

Metric	Pretest LLM-Clinician Decisions,	Pre- *v* Post-Test Clinician Decisions	Pre- *v* Post-Test LLM Decisions	Post-Test Clinician-LLM Decisions	Pretest LLM-Clinician Decisions	Pre- *v* Post-Test Clinician Decisions	Pre- *v* Post-Test LLM Decisions	Post-Test Clinician-LLM Decisions
Agrmt, %	68	56	68	93	70	63	60	90
Kappa	0.381	0.189	0.262	0.814	0.401	0.284	0.136	0.741
P value	.001	.001	.001	.001	.001	.001	.001	.001

Abbreviations: Agrmt, agreement; C1, cohort 1; C2, cohort 2; LLM, large language model.

independent of age ($P = .348$), nodal status ($P = .351$), and RS ($P = .508$). In C2, pretest disagreement was independent of age ($P = .529$), nodal status ($P = .921$), and RS ($P = .174$). However, for post-test decisions, a higher rate of disagreement was observed for N+ compared with N0 patients (18% *v* 8%, $P = .03$), whereas disagreement remained independent of age ($P = .328$) and RS ($P = .223$). Cases of clinician-shift/stable LLM were significantly more frequent among premenopausal patients in C1 (22.5%) and patients with an RS ≤25 in both cohorts (18% in C1 and 19.3% in C2); conversely, cases of LLM-shift/stable clinician were significantly more frequent among patients with high RS (≥26) in C1 (37.3%) and N+ in C2 (31.7%; Data Supplement, Tables S1 and S2).

LLM Prediction of Oncotype DX Result

GPT-4o prediction of RS result varied by menopausal status and risk category in both cohorts (Fig 2). In C1 (Fig 2A), GPT-4o achieved an AUC of 0.817 (95% CI, 0.790 to 0.846) in postmenopausal patients and 0.700 (95% CI, 0.695 to 0.740) in premenopausal patients, whereas in C2 (Fig 2B), the AUC was 0.803 (95% CI, 0.758 to 0.858) and 0.654 (95% CI, 0.567 to 0.759), respectively. Among postmenopausal patients, GPT-4o demonstrated higher accuracy in predicting low than high genomic risk (C1: 75% *v* 62%; C2: 85% *v* 45%; $P < .001$). In premenopausal patients, GPT-4o demonstrated higher accuracy in predicting low genomic risk compared with both intermediate and high genomic risk, in both C1 (61% *v* 50% *v* 42%, $P < .001$) and C2 (48% *v* 43% *v* 30%, $P = .04$; Fig 2).

Error Analysis

Hallucinations in the initial case summary were identified in 3% of the cases. No factual, logical, or linguistic hallucinations were detected at any of the three decision points. By contrast, at least one reference hallucination was observed in up to 28% of cases. Further information is supplied in the Data Supplement.

DISCUSSION

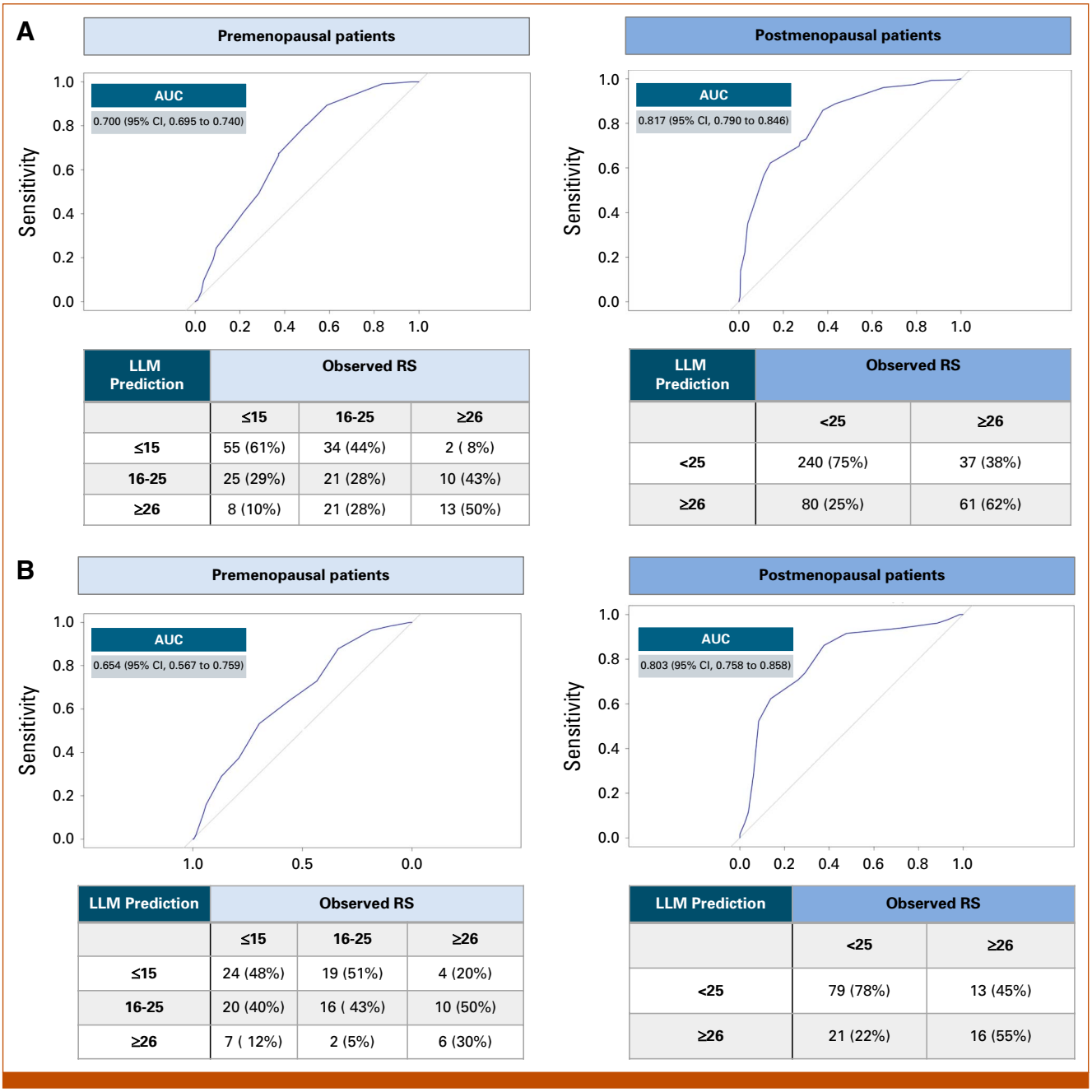
We conducted a comprehensive analysis of LLM-assisted decision making in HR+/HER2− eBC, both in routine clinical practice (C1) and within a clinical study[17] (C2). Our findings underscore key insights into the potential role of LLMs in supporting oncology decision making while also highlighting the challenges that need to be addressed.

Previous studies have investigated the use of ChatGPT in breast cancer management.[21,22] ChatGPT showed to provide appropriate breast cancer screening recommendations[23] and demonstrated substantial accuracy and reliability in answering questions raised by patients with breast cancer about surgical treatment.[24] A recent trial suggested that GPT-4 assistance could effectively improve physician performance in patient care tasks.[25] Finally, an independent analysis evaluating the concordance of treatment recommendations between the Multidisciplinary Tumor Board (MTB) and five publicly available LLMs in 20 complex breast cancer cases showed high agreement of GPT-4o and the MTB (82.4%).[26]

To our knowledge, our study is the first to systematically evaluate the potential of ChatGPT to generate personalized treatment recommendations for a large cohort of patients with eBC.

First, we evaluated the indication for multigene testing and the selection of the test provided by GPT-4o. In both C1 and C2, multigene testing was recommended in all patients. Oncotype DX was the test selected for 100% of patients, in agreement with highest level of evidence (IA), as reported by international guidelines.[27] The rationale for the choice of Oncotype DX was its prognostic and predictive value, longer time of use also supported by extensive real-world evidence. This result seems to be in slight contrast to the findings of a recent study[28] where ChatGPT was used to assess the need for both Oncotype DX in 100 patients with HR+/HER2− eBC at intermediate risk. In contrast to our findings, ChatGPT suggested Oncotype DX in most but not all patients (61%). This discrepancy may be due to methodological differences. In particular, comorbidities were not reported in our study, and this may have limited the capability of GPT-4o to identify patients not eligible to CT and thus to multigene testing.

Before multigene testing, GPT-4o predominantly recommended ET alone as the most appropriate adjuvant therapy (62% of cases in C1 and 57% in C2). By contrast, clinicians recommended ET alone in 42% and 46% of



LLM Prediction	Observed RS		
	≤15	16-25	≥26
≤15	55 (61%)	34 (44%)	2 (8%)
16-25	25 (29%)	21 (28%)	10 (43%)
≥26	8 (10%)	21 (28%)	13 (50%)

LLM Prediction	Observed RS	
	<25	≥26
<25	240 (75%)	37 (38%)
≥26	80 (25%)	61 (62%)

LLM Prediction	Observed RS		
	≤15	16-25	≥26
≤15	24 (48%)	19 (51%)	4 (20%)
16-25	20 (40%)	16 (43%)	10 (50%)
≥26	7 (12%)	2 (5%)	6 (30%)

LLM Prediction	Observed RS	
	<25	≥26
<25	79 (78%)	13 (45%)
≥26	21 (22%)	16 (55%)

FIG 2. AI prediction of Oncotype DX results in cohort 1 and cohort 2. The ROC curves (micro-averaging) show the LLM prediction of Oncotype DX results in cohort 1 (A) and cohort 2 (B). Detailed predictions according to menopausal status are reported in the tables above. AI, artificial intelligence; LLM, large language model; ROC,receiver operating characteristic; RS, recurrence score.

patients in C1 and C2, respectively, more frequently favoring CT-ET than GPT-4o. This discrepancy may reflect the tendency of clinicians to adopt a more aggressive treatment approach, influenced by their clinical experience and the goal of minimizing the risk of undertreatment. Accordingly, the *clinician-shift/stable LLM* pattern was more frequent among premenopausal patients in C1, consistent with a clinician inclination toward CT to reduce the perceived risk of suboptimal therapy in younger patients. Conversely, GPT-4o may exhibit a more conservative approach in cases of uncertainty, at least in part due to intrinsic biases in its information-processing mechanisms or to decision heuristics inherent in its training approach.

Regarding the selection of CT regimens, GPT-4o recommended only standard CT regimens, mostly matching those commonly used by clinicians (Table 3); only rarely (9% in the pretest and 4.7% in the post-test settings), GPT-4o chose less commonly employed regimens (eg, fluorouracil/epirubicin/cyclophosphamide-docetaxel), which, although standard, has become progressively less used.

Despite the relatively low concordance in treatment recommendations observed between clinicians and GPT-4o before receiving Oncotype DX results, the agreement increased significantly after Oncotype DX testing, reaching 93% in the real-world cohort and 90% in the clinical study cohort. These findings highlight the critical role of genomic testing in refining treatment strategies and standardizing clinical decisions, by reducing variability among health care providers.[29]

Although the overall results appear to be consistent between C1 and C2, the agreement between pre- and post-test clinician decisions was numerically higher in C2 than in C1. These imbalances likely stem from differences between real-world practice and clinical study conditions, not from patient factors. The structured, multidisciplinary design of the PRO BONO study, with expert oncologists from high-volume hospitals, likely promoted greater consistency between pre- and post-test decisions.

The LLM predicted Oncotype DX RS values more accurately in postmenopausal than in premenopausal patients, likely due to different classification frameworks (ie, two risk categories for postmenopausal and three for premenopausal patients), making predictions more variable in the latter group. The increased granularity of classification in the premenopausal group may also have introduced greater complexity in accurately identifying the corresponding RS range, potentially contributing to the increased variability in prediction.

Furthermore, GPT-4o demonstrated higher accuracy in predicting low versus high genomic risk both in pre- and postmenopausal patients.

In the error analysis, hallucinations were uncommon in case summaries (3%) and were not observed at the decision points (Data Supplement). By contrast, citation issues were frequent (25%-28% across time points), predominantly topic-related yet context-inappropriate. These findings are consistent with previous evidence[30] and underscore that citation generation and clinical recommendation generation are partially independent behaviors of LLMs. Consequently, references supplied by the model may not reliably substantiate otherwise reasonable recommendations and should be appraised separately in clinical validation workflows.

Several limitations of our study should be acknowledged.

First, our study did not involve any model training or fine-tuning process. Of note, our aim was not to develop a predictive model but rather was to investigate how ChatGPT performs when integrated into routine clinical practice.

Moreover, the use of a single LLM model (GPT-4o) may limit the generalizability of our findings. Furthermore, clinician decisions for both C1 and C2 were collected prospectively, whereas LLM-generated decisions were obtained retrospectively.

Finally, our analyses did not account for long-term patient outcomes, which are critical for validating the clinical utility of LLM-assisted decision making.

Our findings suggest that LLMs, such as GPT-4o, can effectively integrate standardized clinicopathologic variables and genomic data to align with clinician-recommended treatment decisions, particularly following multigene testing. However, caution is warranted when relying solely on LLM for decision making, especially in cases requiring nuanced clinical judgment. Rather than replacing multidisciplinary expertise, LLMs should be viewed as a complementary tool that enhances oncologic decision making by fostering a more standardized and data-driven approach to patient care.

AFFILIATIONS

[1]Department of Breast and Thoracic Oncology, Istituto Nazionale Tumori IRCCS "Fondazione G. Pascale," Naples, Italy
[2]Clinical and Translational Oncology, Scuola Superiore Meridionale (SSM), Naples, Italy
[3]Oncology Unit, Department of Clinical Medicine and Surgery, University of Naples Federico II, Naples, Italy
[4]Université Paris-Saclay, Gustave Roussy, INSERM U981, Molecular Predictors and New Targets in Oncology, Villejuif, France
[5]Department of Advanced Biomedical Sciences, University Federico II, Naples, Italy
[6]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA
[7]Harvard Medical School, Boston, MA
[8]Department of Medicine, Section of Hematology Oncology, Thoracic Oncology Program, The University of Chicago Medicine, Chicago, IL
[9]Oncology Unit, ASL Napoli 3 Sud, Naples, Italy
[10]Breast Surgery Unit, ASL Napoli 3 Sud, Naples, Italy
[11]Department of Medical Oncology, IRCCS Ospedale San Raffaele, Milan, Italy
[12]University Vita-Salute San Raffaele, Milan, Italy
[13]Division of Early Drug Development for Innovative Therapies, IEO, European Institute of Oncology IRCCS, Milan, Italy
[14]Department of Oncology and Haemato-Oncology, University of Milano, Milan, Italy
[15]Department of Medical Oncology, Clinica di Oncologia Medica, IRCCS Ospedale Policlinico San Martino, Genova, Italy
[16]Department of Internal Medicine and Medical Specialties, School of Medicine, University of Genova, Genova, Italy

CORRESPONDING AUTHOR

Grazia Arpino, MD; e-mail: grazia.arpino@unina.it.

EQUAL CONTRIBUTION

R.B. and A.C. contributed equally to this work. C.D.A. and M.G. contributed equally to this work.

PRIOR PRESENTATION

Presented at the San Antonio Breast Cancer Symposium 2024, San Antonio, TX, December 11, 2024; and ESMO Breast Cancer 2025 Congress, Berlin, Germany, May 15, 2025.

DATA SHARING STATEMENT

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Conception and design: Roberto Buonaiuto, Aldo Caltavituro, Grazia Arpino, Carmine De Angelis, Mario Giuliano
Provision of study materials or patients: Luigi Leo, Massimiliano D'Aiuto, Roberto Bianco, Michelino De Laurentiis, Grazia Arpino, Mario Giuliano
Collection and assembly of data: Roberto Buonaiuto, Aldo Caltavituro, Rossana Di Rienzo, Angela Grieco, Federica P. Mangiacotti, Alessandra Longobardi, Vincenza Cantile, Vittoria Molinaro, Martina Pagliuca, Giuseppe Buono, Pietro De Placido, Erica Pietroluongo, Valeria Forestieri, Claudia Martinelli, Vincenzo di Lauro, Luigi Leo, Massimiliano D'Aiuto, Michelino De Laurentiis
Data analysis and interpretation: Roberto Buonaiuto, Aldo Caltavituro, Giampaolo Bianchini, Carmen Criscitiello, Roberto Bianco, Lucia Del Mastro, Michelino De Laurentiis, Grazia Arpino, Carmine De Angelis, Mario Giuliano
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Aldo Caltavituro
Honoraria: Eli Lilly
Travel, Accommodations, Expenses: Lilly

Martina Pagliuca
Research Funding: Gilead Sciences (Inst)
Travel, Accommodations, Expenses: Ipsen

Giuseppe Buono
Consulting or Advisory Role: AstraZeneca, Daiichi Sankyo/AstraZeneca, Novartis, Genetic Spa, Wave Pharma
Speakers' Bureau: Novartis, Lilly, Pfizer, AstraZeneca, Daiichi Sankyo/AstraZeneca, Exact Sciences, Gilead Sciences, MSD Oncology, Pierre Fabre, wave pharma, Genetic Spa
Research Funding: Gilead Sciences (Inst), AstraZeneca (Inst)
Travel, Accommodations, Expenses: Roche, Lilly, Novartis, Daiichi Sankyo/AstraZeneca, AstraZeneca

Claudia Martinelli
Speakers' Bureau: Accademia Nazionale Di Medicina (ACCMED), ANDROMEDA E20 SRL
Travel, Accommodations, Expenses: Roche, Gilead Sciences, Lilly, Novartis, Menarini

Vincenzo di Lauro
Consulting or Advisory Role: Lilly, Amgen, Veracyte, Genetic Spa, Wavepharma, Menarini, Seagen, Accord Healthcare
Travel, Accommodations, Expenses: Pfizer, Novartis, Roche, Gilead Sciences

Giampaolo Bianchini
Honoraria: Roche, MSD, Pfizer, Novartis, Exact Sciences, AstraZeneca/Daiichi Sankyo, Gilead Sciences, Seagen, Menarini, Takeda, Lilly
Consulting or Advisory Role: Roche, AstraZeneca/Daiichi Sankyo, Pfizer, Lilly, Novartis, Gilead Sciences, Exact Sciences, MSD, Seagen, AstraZeneca, Daiichi Sankyo Europe GmbH, Menarini, Helsinn Therapeutics
Speakers' Bureau: Novartis, AstraZeneca, Lilly, MSD
Research Funding: Gilead Sciences (Inst)
Travel, Accommodations, Expenses: Novartis, Roche, Pfizer, AstraZeneca/Daiichi Sankyo, Gilead Sciences, Takeda, Menarini

Carmen Criscitiello
Consulting or Advisory Role: MSD, AstraZeneca, Daiichi Sankyo/AstraZeneca, Seagen, Pfizer, Lilly
Speakers' Bureau: Lilly, Gilead Sciences, Novartis, Pfizer, Roche

Roberto Bianco
Consulting or Advisory Role: Boehringer Ingelheim, AstraZeneca, MSD, Pfizer, Lilly, Roche, Roche

Lucia Del Mastro
Honoraria: Roche, Novartis, Lilly, MSD Oncology, Gilead Sciences
Consulting or Advisory Role: Roche, Novartis, MSD, Pfizer, Ipsen, AstraZeneca, Lilly, Eisai, Pierre Fabre, Daiichi Sankyo, Gilead Sciences, Exact Sciences, Seagen, GlaxoSmithKline, Agendia, Olema Oncology, Menarini
Research Funding: Roche (Inst), Novartis (Inst), Pfizer (Inst), AstraZeneca (Inst), Lilly (Inst), Pierre Fabre (Inst), Daiichi Sankyo/Astra Zeneca (Inst), Seagen (Inst), Gilead Sciences (Inst), Olema Oncology
Travel, Accommodations, Expenses: Roche, Pfizer, Daiichi Sankyo/AstraZeneca, Gilead Sciences, AstraZeneca, Menarini Group

Michelino De Laurentiis
Stock and Other Ownership Interests: Arvinas
Honoraria: Roche, Novartis, Pfizer, Lilly, Pierre Fabre, AstraZeneca, MSD, Seagen, Gilead Sciences, Ipsen, Exact Sciences, TOMA Biosciences, Daiichi Sankyo Europe GmbH, Veracyte
Consulting or Advisory Role: Roche, Novartis, Pfizer, Lilly, AstraZeneca, MSD, Pierre Fabre, Seagen, Gilead Sciences, Ipsen, Daiichi Sankyo Europe GmbH
Speakers' Bureau: Novartis
Research Funding: Novartis (Inst), Roche (Inst), Lilly, Pfizer (Inst), Daiichi Sankyo (Inst), MSD (Inst), Bristol Myers Squibb (Inst), Genzyme (Inst), AstraZeneca (Inst)

Grazia Arpino
Honoraria: Roche, Lilly, Novartis, Novartis (Inst), Pfizer, Pfizer (Inst), AstraZeneca
Consulting or Advisory Role: Roche, Eisai, Lilly, Pfizer, Novartis, Daiichi Sankyo/AstraZeneca, MSD Oncology
Speakers' Bureau: Roche, Pfizer, Lilly, Novartis, Helsinn Therapeutics/QED Therapeutics
Research Funding: Novartis (Inst), pfizer (Inst), eisai (Inst), Roche (Inst)
Travel, Accommodations, Expenses: Pfizer, Roche, Lilly, Daiichi Sankyo/AstraZeneca

Carmine De Angelis
Consulting or Advisory Role: Lilly, Novartis, Gilead Sciences, Daiichi Sankyo/AstraZeneca, GlaxoSmithKline
Research Funding: Daiichi Sankyo/UCB Japan (Inst), Gilead Sciences (Inst)
Travel, Accommodations, Expenses: Daiichi Sankyo/AstraZeneca, Gilead Sciences

Mario Giuliano
Stock and Other Ownership Interests: Aldeyra Therapeutics, Karyopharm Therapeutics, BioAtla, Cardiff Oncology, Clearside Biomedical, Mersana, ADC Therapeutics

Honoraria: AstraZeneca, Daiichi Sankyo/AstraZeneca, Lilly, MSD, Novartis, Menarini Group, Pfizer, Exact Sciences, Gilead Sciences
Consulting or Advisory Role: AstraZeneca/Daiichi Sankyo, Lilly, MSD, Novartis, Pfizer, Roche, Gilead Sciences
Speakers' Bureau: AstraZeneca/Daiichi Sankyo, Exact Sciences, Lilly, MSD, Gilead Sciences, Menarini Group, Novartis, Pfizer, Roche
Travel, Accommodations, Expenses: AstraZeneca/Daiichi Sankyo, Lilly, Novartis, Roche, Pfizer, Gilead Sciences

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We acknowledge Genny Paradiso for her support in managing the reporting of the tests, which provided an additional source to support the confirmation of our pre- and post-test decision data.
We acknowledge Domenico Ciuonzo from the Department of Electrical Engineering and Information Technologies (DIETI) of the University of Naples Federico II for the support in conducting the error analysis.

REFERENCES

1. Al Kuwaiti A, Nazer K, Al-Reedy A, et al: A review of the role of artificial intelligence in healthcare. J Pers Med 13:951, 2023
2. Pandav K, Almahfouz Nasser S, Kimball KH, et al: Opportunities for artificial intelligence in oncology: From the lens of Clinicians and patients. JCO Oncol Pract 21:1257-1264, 2025
3. McKinney SM, Sieniek M, Godbole V, et al: International evaluation of an AI system for breast cancer screening. Nature 577:89-94, 2020
4. Dada EG, Oyewola DO, Misra S: Computer-aided diagnosis of breast cancer from mammogram images using deep learning algorithms. J Electr Syst Inf Technol 11:38, 2024
5. Yoon JH, Han K, Suh HJ, et al: Artificial intelligence-based computer-assisted detection/diagnosis (AI-CAD) for screening mammography: Outcomes of AI-CAD in the mammographic interpretation workflow. Eur J Radiol Open 11:100509, 2023
6. Kohli A, Jha S: Why CAD failed in mammography. J Am Coll Radiol 15:535-537, 2018
7. Ahn JS, Shin S, Yang S-A, et al: Artificial intelligence in breast cancer diagnosis and personalized medicine. J Breast Cancer 26:405, 2023
8. Soliman A, Li Z, Parwani AV: Artificial intelligence's impact on breast cancer pathology: A literature review. Diagn Pathol 19:38, 2024
9. Polónia A, Campelos S, Ribeiro A, et al: Artificial intelligence improves the accuracy in histologic classification of breast lesions. Am J Clin Pathol 155:527-536, 2021
10. Sandbank J, Bataillon G, Nudelman A, et al: Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. NPJ Breast Cancer 8:129, 2022
11. Wang R, Gu Y, Zhang T, et al: Fast cancer metastasis location based on dual magnification hard example mining network in whole-slide images. Comput Biol Med 158:106880, 2023
12. Abele N, Tiemann K, Krech T, et al: Noninferiority of artificial intelligence–assisted analysis of Ki-67 and estrogen/progesterone receptor in breast cancer routine diagnostics. Mod Pathol 36: 100033, 2023
13. Hartage R, Li AC, Hammond S, et al: A validation study of human epidermal growth factor receptor 2 immunohistochemistry digital imaging analysis and its correlation with human epidermal growth factor receptor 2 fluorescence in situ hybridization results in breast carcinoma. J Pathol Inform 11:2, 2020
14. Sorin V, Barash Y, Konen E, et al: Large language models for oncological applications. J Cancer Res Clin Oncol 149:9505-9508, 2023
15. Dave T, Athaluri SA, Singh S: ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 6:1169595, 2023
16. Rushanyan M, Aghabekyan T, Tamamyan G, et al: Clinical decision making by ChatGPT vs medical oncologists: A retrospective concordance study. J Clin Oncol 42:e13634, 2024
17. De Placido P, Di Rienzo R, Pietroluongo E, et al: Insights on the association of anthropometric and metabolic variables with tumor features and genomic risk in luminal early breast cancer: Results of a multicentric prospective study. Eur J Cancer 221:115409, 2025
18. Liu J, Liu F, Wang C, et al: Prompt engineering in clinical practice: Tutorial for clinicians. J Med Internet Res 27:e72644, 2025
19. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, et al: Human-in-the-loop machine learning: A state of the art. Artif Intell Rev 56:3005-3054, 2023
20. Wu X, Xiao L, Sun Y, et al: A survey of human-in-the-loop for machine learning. Future Generation Comput Syst 135:364-381, 2022
21. Lukac S, Dayan D, Fink V, et al: Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 308:1831-1844, 2023
22. Sorin V, Klang E, Sklair-Levy M, et al: Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 9:44, 2023
23. Braithwaite D, Karanth SD, Divaker J, et al: Evaluating ChatGPT's accuracy in providing screening mammography recommendations among older women: Artificial intelligence and cancer communication. J Am Geriatr Soc 72:2237-2240, 2024
24. Roldan-Vasquez E, Mitri S, Bhasin S, et al: Reliability of artificial intelligence chatbot responses to frequently asked questions in breast surgical oncology. J Surg Oncol 130:188-203, 2024
25. Goh E, Gallo RJ, Strong E, et al: GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. Nat Med 31:1233-1238, 2025
26. Griewing S, Knitza J, Boekhoff J, et al: Evolution of publicly available large language models for complex decision-making in breast cancer care. Arch Gynecol Obstet 310:537-550, 2024
27. Andre F, Ismaila N, Allison KH, et al: Biomarkers for adjuvant endocrine and chemotherapy in early-stage breast cancer: ASCO guideline update. J Clin Oncol 40:1816-1837, 2022
28. Patanè F, Palumbo F, Lorenzini G, et al: 3P evaluating conversational generative pre-trained transformer (ChatGPT) as a tool in early breast cancer (eBC) cases. ESMO Open 9:102258, 2024
29. Licata L, Viale G, Giuliano M, et al: Oncotype DX results increase concordance in adjuvant chemotherapy recommendations for early-stage breast cancer. NPJ Breast Cancer 9:51, 2023
30. Ferber D, El Nahhas OSM, Wölflein G, et al: Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat Cancer 6:1337-1349, 2025