RESEARCH ARTICLE

Editorial Process: Submission:08/02/2025 Acceptance:05/10/2026 Published:05/18/2026

Diagnostic Accuracy and Reproducibility of ChatGPT-4o for HER2 Immunohistochemistry Scoring in Equivocal Breast Cancer Cases

Cheep Charoenlap[1], Pakorn Arunsawat[1], Kittiphan Chienwichai[2], Tanaporn Prateepchaiboon[3], Arunchai Chang[4]*

Abstract

Background: HER2 immunohistochemistry (IHC) plays a central role in therapeutic decision-making for breast cancer. However, interpretation of equivocal (2+) IHC results remains challenging and is subject to interobserver variability, necessitating reflex in situ hybridization testing. This study evaluated the diagnostic performance and reproducibility of ChatGPT-4o, a general-purpose large language model, in scoring HER2 IHC in breast cancer cases initially classified as IHC 2+. **Methods:** We retrospectively analyzed 81 formalin-fixed, paraffin-embedded invasive carcinoma of no special type (NST) cases with prior HER2 IHC 2+ scores and corresponding dual in situ hybridization (DISH) results. Five high-power field images per case were independently analyzed by ChatGPT-4o across three sessions, using a standardized prompt aligned with the ASCO/CAP 2023 guidelines. Cases remaining equivocal after AI-assisted interpretation were excluded from diagnostic performance calculations. HER2 DISH served as the reference standard. **Results:** Fourteen cases (17.3%) remained equivocal following AI interpretation. Among the 67 reclassified cases, ChatGPT-4o demonstrated an overall diagnostic accuracy of 79% (95% CI: 67–88%), a sensitivity of 30%, specificity of 100%, positive predictive value of 100%, and negative predictive value of 77%. Intra-model reproducibility was good (intraclass correlation coefficient = 0.78), whereas agreement with HER2 DISH was fair (Cohen's κ = 0.375). Misclassification predominantly involved false-negative interpretations among HER2-positive cases. **Conclusion:** ChatGPT-4o demonstrated high specificity and reproducibility for identifying HER2 IHC 3+ cases but showed limited sensitivity and only fair concordance with HER2 DISH. These findings indicate that, in its current general-purpose form, ChatGPT-4o is not suitable for independent HER2 assessment and may serve, at best, as an exploratory adjunct to pathologist interpretation.

Keywords: HER2 immunohistochemistry- breast cancer- artificial intelligence- ChatGPT-4o- diagnostic reproducibility

Asian Pac J Cancer Prev, **27 (5)**, 1703-1708

Introduction

Breast cancer remains one of the leading causes of cancer-related morbidity and mortality worldwide, with human epidermal growth factor receptor 2 (*HER2*) status serving as a critical predictive biomarker for targeted therapy. *HER2* overexpression, driven primarily by *ERBB2* gene amplification, identifies a subset of tumors that benefit from anti-*HER2* agents such as trastuzumab and pertuzumab, making accurate *HER2* assessment essential for appropriate treatment selection [1-3].

Current ASCO/CAP guidelines recommend immunohistochemistry (IHC) as the initial method for *HER2* evaluation, with equivocal (2+) results requiring reflex in situ hybridization (ISH) testing. Although this algorithm is well established, interpretation of *HER2* IHC particularly equivocal cases remains challenging. Subjective assessment of membranous staining intensity and completeness, coupled with pre-analytic and technical variability, contributes to interobserver disagreement. These limitations may lead to unnecessary reflex testing, increased costs, delayed turnaround times, and potential misclassification, particularly in tumors with heterogeneous or low-level *HER2* expression [4-9].

The clinical relevance of accurate *HER2* categorization has expanded with the recognition of *HER2*-low and *HER2*-ultralow disease, where subtle differences in IHC interpretation may influence eligibility for antibody–drug conjugates. Consequently, there is growing interest in objective or AI-assisted approaches that may enhance

*[1]Department of Anatomical Pathology, Hatyai Hospital, Songkhla, Thailand. [2]Division of Nephrology, Department of Internal Medicine, Hatyai Hospital, Songkhla, Thailand. [3]Division of Medical Oncology, Department of Internal Medicine, Hatyai Hospital, Songkhla, Thailand. [4]Division of Gastroenterology, Department of Internal Medicine, Hatyai Hospital, Songkhla, Thailand. *For Correspondence: busmdcu58@gmail.com*

consistency in *HER2* IHC assessment. Pathology-specific artificial intelligence systems trained on whole-slide images have shown promising performance; however, such systems require specialized infrastructure, curated datasets, and domain-specific training [6-9].

In parallel, large language models (LLMs), including ChatGPT, have emerged as general-purpose AI tools capable of processing multimodal inputs, including images. Preliminary studies have explored their potential roles in pathology-related tasks such as diagnostic reasoning, image description, and educational support. Nevertheless, these models were not designed for medical diagnostics and lack pathology-specific training and regulatory validation [10-13]. Their diagnostic performance, particularly for visually nuanced tasks such as *HER2* IHC interpretation, remains largely unexplored.

Against this background, the present study investigates the feasibility of using ChatGPT-4o, a general-purpose large language model with image-processing capability, as an exploratory tool for *HER2* IHC scoring in breast cancer cases initially classified as equivocal (2+). Specifically, this study aims to evaluate the diagnostic accuracy of ChatGPT-4o relative to *HER2* DISH results and to assess its intra-model reproducibility across repeated evaluations.

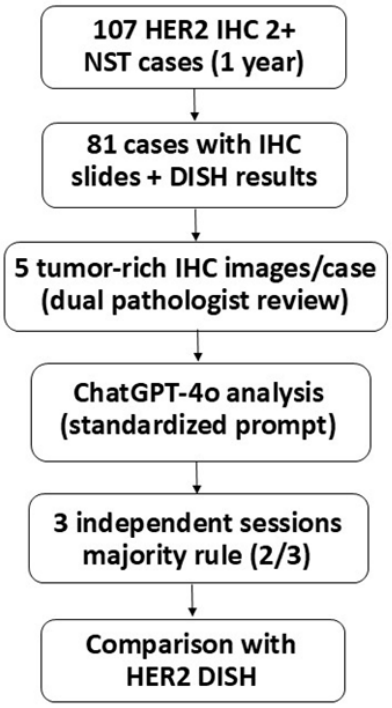

Figure 1. Study Workflow of ChatGPT-4o HER2 IHC Interpretation

Materials and Methods

Study Design and Population

This retrospective diagnostic study was conducted using archival formalin-fixed, paraffin-embedded breast cancer specimens obtained at Hatyai Hospital, Thailand, between January 2024 and July 2025. A total of 107 cases with *HER2* IHC scores of 2+ were initially identified from the pathology database. After histopathologic confirmation and quality screening, 81 cases met the inclusion criteria: (1) diagnosis of invasive carcinoma of no special type (NST); (2) prior *HER2* IHC score of 2+ assigned by board-certified pathologists; and (3) availability of corresponding *HER2* DISH results and intact *HER2*-stained slides. Cases with insufficient tissue, missing DISH data, or inadequate image quality were excluded. The study workflow is illustrated in Figure 1.

Sample Size Considerations

A preliminary pilot assessment was conducted to evaluate feasibility rather than to test a definitive hypothesis. The sample size was informed by estimation of diagnostic sensitivity with an acceptable margin of error, acknowledging the exploratory nature of the study. Accordingly, findings should be interpreted as hypothesis-generating.

Image Acquisition

For each case, five representative high-power field (400×) images of *HER2*-stained tumor regions were digitally captured using a standardized microscope-mounted imaging system at a resolution of 2048 × 1536 pixels (TIFF format). Tumor-rich areas were selected to reflect staining heterogeneity across the lesion (as shown in Figure 2). Images with poor focus, artifacts, or suboptimal staining were excluded. To minimize selection bias, images were reviewed independently by two pathologists prior to upload, and investigators were blinded to *HER2* DISH results during image selection.

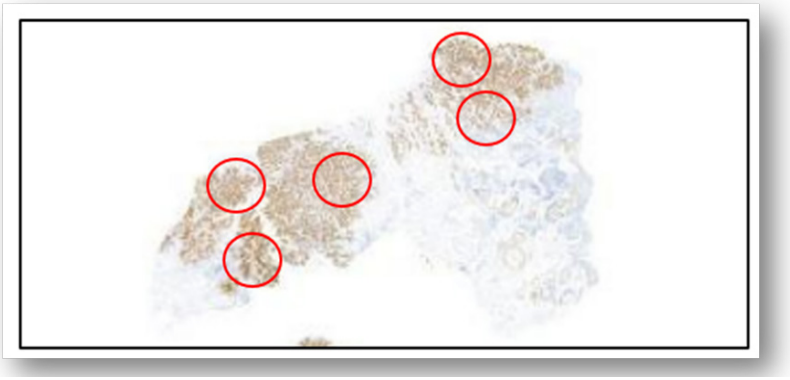

Figure 2. Demonstration of the Selection of Five Representative Fields for HER2 IHC Imaging

ChatGPT-4o Interaction and Scoring Protocol

A pretrained ChatGPT-4o model (OpenAI, May 2025 version) was applied in a zero-shot setting without additional fine-tuning or dataset-specific training. Each set of five images per case was uploaded in a new chat session using a standardized prompt based on ASCO/CAP 2023 criteria:

"Please evaluate these *HER2* immunohistochemistry images based on the 2023 ASCO/CAP guidelines. For each image, consider intensity of membranous staining (weak, moderate, strong), completeness (complete or incomplete), and the proportion of positively stained tumor cells. Provide a final *HER2* score (0, 1+, 2+, or 3+) for the case based on majority criteria."

ChatGPT-4o was blinded to all clinical data and DISH results. A final *HER2* IHC score for each case was determined by majority interpretation across the five images. Representative outputs are shown in Figure 3.

Reproducibility Assessment

To assess intra-model reproducibility, the complete image dataset was reanalyzed in two additional independent ChatGPT-4o sessions using the identical prompt and image input format. Each session was conducted separately to minimize contextual memory effects. Reproducibility was defined as concordant *HER2* IHC scores in at least two out of three evaluation rounds. The intraclass correlation coefficient (ICC) was used to quantify scoring reliability across sessions.

Reference Standard and Concordance Criteria

HER2 dual in situ hybridization (DISH) was used as the reference standard for *HER2* status determination. Interpretation followed the 2023 ASCO/CAP guidelines [14]. Cases were categorized as:

HER2-positive

DISH amplification (Group 1); or IHC 3+; or IHC 2+ with DISH Group 3.

HER2-negative

DISH non-amplification (Group 5); or IHC 0–2+ with DISH Group 2 or 4 that remains in the same group after repeat review/recount by a second observer; or IHC 0–1+ with DISH Group 3.

Cases that remained classified as *HER2* IHC 2+ after ChatGPT-4o interpretation were considered unresolved and were excluded from diagnostic accuracy calculations.

HER2 Immunohistochemistry Procedure

HER2 immunohistochemistry was performed on 4-µm formalin-fixed, paraffin-embedded tissue sections using the Ventana BenchMark ULTRA automated staining platform. The PATHWAY anti-*HER2*/neu antibody (clone 4B5) was used according to the manufacturer's protocol. Antigen retrieval was carried out with CC1 buffer at 95 °C for 36 minutes, followed by ultraView DAB detection and automated counterstaining. Each staining run included validated internal and external quality controls.

HER2 Dual In Situ Hybridization (DISH)

HER2 gene amplification was evaluated using the VENTANA INFORM *HER2* Dual ISH DNA Probe Cocktail, targeting the *HER2* gene and chromosome 17 centromere (CEP17). Hybridization and signal detection were performed on the Ventana BenchMark ULTRA platform in accordance with manufacturer-recommended protocols. *HER2* signals were visualized as black signals, and CEP17 signals as red signals.

Statistical Analysis

Diagnostic performance was evaluated using 2 × 2 contingency tables comparing ChatGPT-4o–derived *HER2* IHC classifications with *HER2* DISH results as the reference standard. Overall accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated with corresponding 95% confidence intervals.

Intra-model reproducibility across three independent ChatGPT-4o evaluation rounds was assessed using the intraclass correlation coefficient (ICC), based on a two-way random-effects model with absolute agreement. ICC values were interpreted as follows: < 0.50, poor reliability; 0.50–0.75, moderate reliability; 0.75–0.90,

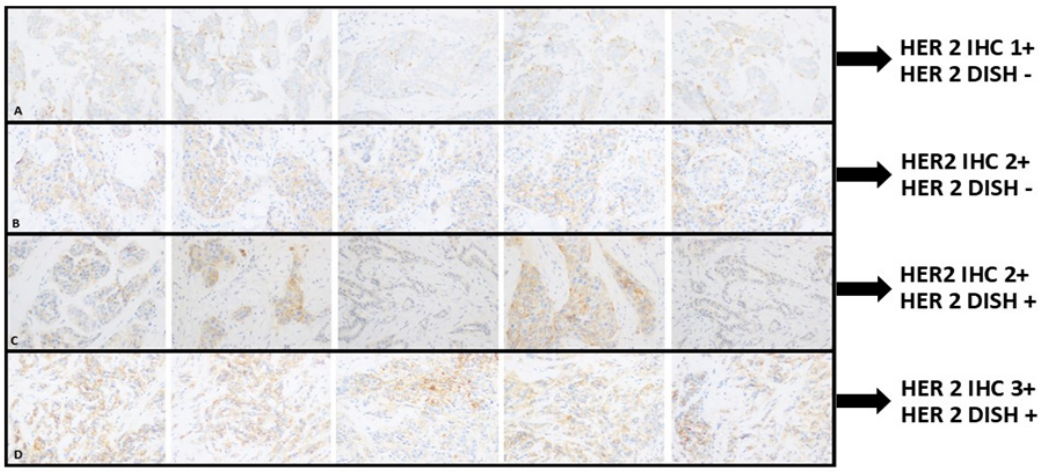

Figure 3. Sample Image Sets with ChatGPT-4o-assigned HER2 IHC Scores (1+, 2+, and 3+) and Corresponding HER2 DISH Results for Initially Equivocal (2+) cases.

good reliability; and > 0.90, excellent reliability.

Agreement between ChatGPT-4o *HER2* IHC classifications and *HER2* DISH results was evaluated using Cohen's kappa (κ) statistic, with agreement strength categorized as slight (< 0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (> 0.80).

All statistical analyses were performed using R software (version 3.3.1; R Foundation for Statistical Computing).

Results

Study Cohort

A total of 107 breast cancer cases with initial *HER2* IHC scores of 2+ were identified during the study period. After exclusion of cases with insufficient tissue, missing *HER2* DISH results, or inadequate image quality, 81 cases were included in the final analysis. All patients were female, with a mean age of 54.9 years (range 31–77 years; SD 10.9). All tumors were classified as invasive carcinoma of no special type (NST).

HER2 DISH analysis demonstrated *ERBB2* gene amplification. *HER2* DISH results were positive in 26 cases (32.1%) and negative in 55 cases (67.9%). Based on ASCO/CAP criteria, the distribution of DISH groups was as follows: Group 1 (amplified), 27 cases (33.3%); Group 2 (equivocal), 8 cases (9.9%); and Group 5 (non-amplified), 46 cases (56.8%). Baseline demographic and *HER2* DISH characteristics are summarized in Table 1.

ChatGPT-4o HER2 IHC Classification

Following AI-assisted evaluation, 14 of the 81 cases (17.3%) remained classified as *HER2* IHC 2+ and were considered unresolved. These cases were excluded from diagnostic performance calculations in accordance with guideline-based definitions of equivocal results. The remaining 67 cases were reclassified by ChatGPT-4o as either *HER2* IHC 0/1+ or 3+ and were included in subsequent analyses.

Diagnostic Performance

Among the 67 evaluable cases, ChatGPT-4o achieved an overall diagnostic accuracy of 79% (95% CI: 67–88%) when compared with *HER2* DISH results. Sensitivity for detecting *HER2*-positive disease was 30% (95% CI: 12–54%), while specificity reached 100% (95% CI: 92–100%). The positive predictive value was 100% (95% CI: 54–100%), indicating complete concordance among cases classified as IHC 3+. The negative predictive value was 77% (95% CI: 65–87%). Detailed diagnostic performance metrics are presented in Table 2.

Table 1. Demographic Characteristics and HER2 Dual in Situ Hybridization (DISH) Results of the Study Cohort (N = 81)

Characteristic	Value
Age (years), mean ± SD	54.9 ± 10.9
HER2 DISH positive, n (%)	26 (32.1)
HER2 DISH negative, n (%)	55 (67.9)
DISH Group 1 (amplified), n (%)	27 (33.3)
DISH Group 2 (equivocal), n (%)	8 (9.9)
DISH Group 5 (non-amplified), n (%)	46 (56.8)

SD, standard deviation; DISH, dual in situ hybridization.

Misclassification Patterns

Misclassification was predominantly driven by false-negative interpretations. Fourteen of 61 cases (23%) classified as *HER2*-negative (IHC 0/1+) by ChatGPT-4o were *HER2*-positive by DISH. Among these false-negative cases, 10 (71.4%) were consistently interpreted as IHC 1+ across all three evaluation rounds, suggesting systematic underestimation of *HER2* expression, particularly in cases with intratumoral heterogeneity characterized by foci of weak or incomplete membranous staining.

A heatmap illustrating ChatGPT-4o *HER2* IHC interpretations across three independent scoring rounds and the final consensus result, in comparison with *HER2* DISH status, is shown in Figure 4. This visualization highlights inter-round consistency, unresolved equivocal

Table 2. Diagnostic Performance of ChatGPT-4o HER2 Immunohistochemistry Interpretation Compared with HER2 DISH in Reclassified Non-Equivocal Cases (N = 67)

ChatGPT-4o HER2 IHC classification	HER2 DISH positive	HER2 DISH negative	Total
IHC 3+ (positive)	6	0	6
IHC 0–1+ (negative)	14	47	61
Total	20	47	67

Cases remaining equivocal (IHC 2+) after ChatGPT-4o interpretation were excluded from analysis.

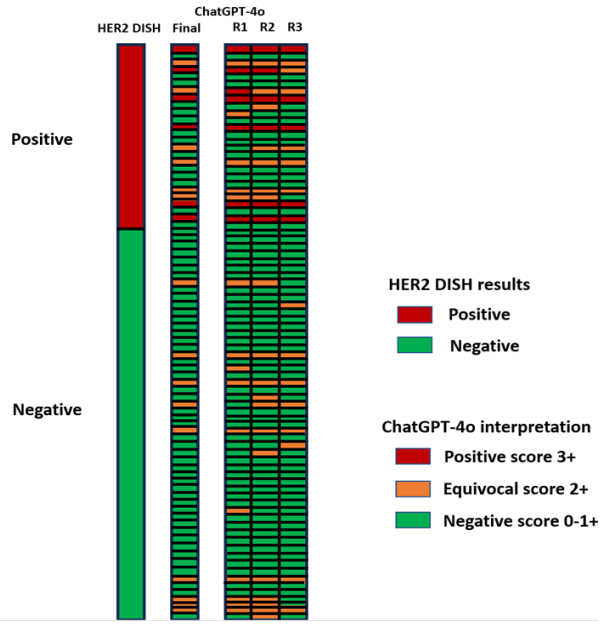

Figure 4. Heatmap Showed ChatGPT-4o HER2 IHC Interpretations (Final and R1–R3) across 81 Cases, Ordered by HER2 DISH Status. Colors represent HER2 IHC categories (0–1+, 2+, 3+). The figure demonstrates inter-round consistency, equivocal cases, and agreement with HER2 DISH.

Table 3. Reproducibility and Agreement of ChatGPT-4o HER2 Immunohistochemistry Assessment

Assessment	Comparison	Statistical measure	Interpretation
Intra-model reproducibility	ChatGPT-4o scoring across three independent evaluation rounds	Intraclass correlation coefficient (ICC) = 0.78	Good reproducibility
Inter-method agreement	ChatGPT-4o HER2 IHC classification vs. HER2 DISH status	Cohen's kappa (κ) = 0.375	Fair agreement

ICC calculated using a two-way random-effects model with absolute agreement.

cases, and patterns of concordance and discordance.

Reproducibility and Agreement

Intra-model reproducibility across three independent ChatGPT-4o scoring rounds was good, with an intraclass correlation coefficient (ICC) of 0.78 (95% CI: 0.70–0.84; $p < 0.001$). Agreement between ChatGPT-4o *HER2* IHC classification and *HER2* DISH status was fair, with a Cohen's kappa value of 0.375 ($p < 0.001$). Reproducibility and agreement metrics are summarized in Table 3.

Discussion

In this exploratory study, we evaluated the diagnostic performance and reproducibility of ChatGPT-4o, a general-purpose large language model, for *HER2* immunohistochemistry interpretation in breast cancer cases initially classified as equivocal (IHC 2+). While ChatGPT-4o demonstrated excellent specificity and perfect positive predictive value for identifying *HER2* IHC 3+ cases, its low sensitivity and only fair agreement with *HER2* DISH substantially limit its clinical utility.

The most notable finding was the consistent underestimation of *HER2* expression in a subset of *HER2*-DISH positive tumors, resulting in false-negative classifications. These errors were not sporadic; rather, many *HER2*-DISH positive cases were repeatedly scored as IHC 1+ across multiple evaluation rounds. This pattern suggests difficulty in recognizing tumor heterogeneity characterized by foci of weak or incomplete membranous staining, a challenge that is particularly relevant in the current era of *HER2*-low and *HER2*-ultralow disease classification [5]. From a clinical perspective, such false-negative interpretations could delay or preclude access to *HER2*-targeted therapies, with direct implications for patient outcomes [15].

Although ChatGPT-4o showed good intra-model reproducibility, reproducibility alone does not equate to diagnostic validity. The fair concordance with *HER2* DISH underscores that consistent outputs may still be consistently incorrect. Importantly, pathologist interpretation remains the reference standard for *HER2* IHC assessment, and all equivocal cases must continue to undergo confirmatory ISH testing in accordance with established guidelines, regardless of any AI-generated interpretation.

When contextualized against pathology-specific AI platforms, the performance of ChatGPT-4o is modest. Deep learning systems trained on annotated whole-slide images have reported higher sensitivity, stronger agreement with reference standards, and superior discrimination across *HER2* categories. These systems benefit from domain-specific training, access to whole-slide context, and explicit modeling of tumor–stroma relationships features absent in general-purpose LLMs. The contrast highlights that model architecture and training paradigm are critical determinants of performance in visually nuanced diagnostic tasks [8, 9, 16-19].

Recent interest in applying large language models to pathology has primarily focused on educational support, diagnostic reasoning assistance, and structured reporting, rather than direct image-based diagnosis. While experimental studies suggest that LLMs can synthesize morphological descriptions and assist interpretive reasoning, their role in primary diagnostic decision-making remains unproven. The findings of the present study reinforce this distinction, demonstrating that a general-purpose LLM, even with image-processing capability, is insufficient for reliable biomarker assessment without pathology-specific training and validation [10-13].

This study has several important limitations. First, analysis was restricted to preselected high-power field images rather than whole-slide images, introducing potential selection bias and limiting assessment of architectural context and staining heterogeneity. Second, ChatGPT-4o was applied in a zero-shot setting without pathology-specific training, constraining its ability to recognize subtle, domain-specific visual features. Third, the study was conducted at a single center with a relatively modest sample size and should be interpreted as exploratory. Prospective multi-center validation to improve diagnostic performance is needed. Finally, the exclusion of unresolved equivocal cases from performance calculations may influence generalizability of the reported metrics.

In summary, ChatGPT-4o demonstrated high specificity and good reproducibility in identifying *HER2* IHC 3+ cases among initially equivocal breast cancer samples. However, its low sensitivity, modest negative predictive value, and only fair agreement with *HER2* DISH indicate that it cannot reliably detect all *HER2*-positive tumors. In its current general-purpose form, ChatGPT-4o is not suitable for independent *HER2* assessment and should not be integrated into routine diagnostic workflows. Future advances in pathology-trained, slide-level AI systems may offer more clinically meaningful performance and warrant further investigation.

Author Contribution Statement

CC and AC conceived and designed the study. CC collected data, performed statistical analyses, and drafted the manuscript. CC and PA reviewed images and contributed to data interpretation. KC, TP, and AC

critically revised the manuscript. All authors approved the final version.

Acknowledgements

The authors thank the pathology and clinical staff at Hatyai Hospital for their assistance with data acquisition and slide digitization, as well as the administrative teams who facilitated access to archival materials.

Ethics Approval and Consent to Participate

This study was approved by the Institutional Review Board of Hatyai Hospital (HYH EC 052-68-01) and conducted in accordance with the Declaration of Helsinki. Informed consent was waived due to the retrospective design and use of anonymized data.

Availability of Data and Materials

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request, subject to institutional approval.

References

1. International agency for research on cancer. Global cancer observatory: Cancer today [internet]. Lyon (france): International agency for research on cancer; 2024 [cited 2025 jan 10]. Available from: Https://gco.Iarc.Who.Int/today.
2. Bhushan A, Gonsalves A, Menon JU. Current state of breast cancer diagnosis, treatment, and theranostics. Pharmaceutics. 2021;13(5). https://doi.org/10.3390/pharmaceutics13050723.
3. Swain SM, Shastry M, Hamilton E. Targeting *HER2*-positive breast cancer: Advances and future directions. Nat Rev Drug Discov. 2023;22(2):101-26. https://doi.org/10.1038/s41573-022-00579-0.
4. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline focused update. J Clin Oncol. 2018;36(20):2105-22. https://doi.org/10.1200/jco.2018.77.8738.
5. Ferrari P, Macedo M, Cunha I, Soares F. Abstract pr-06: Analysis of pathologists' intraobserver, interobserver and ai agreement in breast cancer *HER2* scoring: Ai-assessed intra-sample tumor heterogeneity relates to lower agreement among pathologists and with ai. Clin Cancer Res. 2025;31:Pr-06. https://doi.org/10.1158/1557-3265.Aimachine-pr-06.
6. Turashvili G, Gao Y, Ai DA, Ewaz AM, Gjeorgjievski SG, Wang Q, et al. Low interobserver agreement among subspecialised breast pathologists in evaluating *HER2*-low breast cancer. J Clin Pathol. 2024;77(12):815-21. https://doi.org/10.1136/jcp-2023-209055.
7. Joaquim AC, Bazzaneze LK, Percicote AP, Sebastião APM. Interobserver variability in *HER2* immunohistochemistry analysis and its clinical implications with the advent of the *HER2*-low category. Surgical and Experimental Pathology. 2025;8(1):24. https://doi.org/10.1186/s42047-025-00199-z.
8. Gavrielides MA, Gallas BD, Lenz P, Badano A, Hewitt SM. Observer variability in the interpretation of *HER2*/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch Pathol Lab Med. 2011;135(2):233-42. https://doi.org/10.5858/135.2.233.
9. Wu S, Yue M, Zhang J, Li X, Li Z, Zhang H, et al. The role of artificial intelligence in accurate interpretation of *HER2* immunohistochemical scores 0 and 1+ in breast cancer. Mod Pathol. 2023;36(3):100054. https://doi.org/10.1016/j.modpat.2022.100054.
10. Laohawetwanit T, Namboonlue C, Apornvirat S. Accuracy of gpt-4 in histopathological image detection and classification of colorectal adenomas. J Clin Pathol. 2025;78(3):202-7. https://doi.org/10.1136/jcp-2023-209304.
11. Sa-ngiamwibool P, Laohawetwanit T. Comparative analysis of chatgpt and human expertise in diagnosing primary liver carcinoma: A focus on gross morphology. Siriraj Med J. 2025;77(2):119-29. https://doi.org/10.33192/smj.v77i2.271596.
12. Laohawetwanit T, Pinto DG, Bychkov A. A survey analysis of the adoption of large language models among pathologists. Am J Clin Pathol. 2025;163(1):52-9. https://doi.org/10.1093/ajcp/aqae093.
13. Laohawetwanit T, Apornvirat S, Asaturova A, Li H, Lami K, Bychkov A. Evaluation of general-purpose large language models as diagnostic support tools in cervical cytology. Pathol Res Pract. 2025;274:156159. https://doi.org/10.1016/j.prp.2025.156159.
14. College of american pathologists. Protocol for the examination of specimens from patients with invasive carcinoma of the breast (version 1.6.0.0) [internet]. Northfield (il): College of american pathologists; 2025 [cited 2025 jun 10]. Available from: Https://www.Cap.org/protocols-and-guidelines.
15. Garrison LP, Jr., Babigumira JB, Masaquel A, Wang BC, Lalla D, Brammer M. The lifetime economic burden of inaccurate *HER2* testing: Estimating the costs of false-positive and false-negative *HER2* test results in us patients with early-stage breast cancer. Value Health. 2015;18(4):541-6. https://doi.org/10.1016/j.jval.2015.01.012.
16. Vandenberghe ME, Scott ML, Scorer PW, Söderberg M, Balcerzak D, Barker C. Relevance of deep learning to facilitate the diagnosis of *HER2* status in breast cancer. Sci Rep. 2017;7:45938. https://doi.org/10.1038/srep45938.
17. Liao CC, Bakoglu N, Cesmecioglu E, Hanna M, Pareja F, Wen HY, et al. Semi-automated analysis of *HER2* immunohistochemistry in invasive breast carcinoma using whole slide images: Utility for interpretation in clinical practice. Pathol Oncol Res. 2024;30:1611826. https://doi.org/10.3389/pore.2024.1611826.
18. Kabir S, Vranic S, Mahmood Al Saady R, Salman Khan M, Sarmun R, Alqahtani A, et al. The utility of a deep learning-based approach in her-2/neu assessment in breast cancer. Expert Systems with Applications. 2024;238:122051. https://doi.org/10.1016/j.eswa.2023.122051.
19. Xiong Z, Liu K, Liu S, Feng J, Wang J, Feng Z, et al. Precision *HER2*: A comprehensive ai system for accurate and consistent evaluation of *HER2* expression in invasive breast cancer. BMC Cancer. 2024;24(1):1204. https://doi.org/10.1186/s12885-024-12980-6.