npj | breast cancer

Article

Published in partnership with the Breast Cancer Research Foundation

https://doi.org/10.1038/s41523-025-00872-2

HER2-low and ultralow expression of invasive breast carcinoma: clinicopathological features and immunohistochemical consistency analysis

Yu-han Dong[1], Yi-tong Xu[1], Xi-zi Jiang[1], Lu Xu[2], Rui-Jiao Yu[1], Rui Wang[3] & Qing-fu Zhang[1] ✉

The DESTINY-Breast06 trial demonstrated survival benefits of trastuzumab deruxtecan (T-DXd) in human epidermal growth factor receptor 2 (HER2)-low/ultralow metastatic breast cancer. However, the associated clinicopathological features, HER2 testing consistency, and utility of artificial intelligence (AI)-assisted interpretation remain unclear. We analyzed 1455 HER2-non-amplified cases and conducted a multicenter ring study across 45 laboratories, with 209 tissue sections evaluated by manual and AI-assisted scoring. Compared to HER2-ultralow cases, HER2-null tumors exhibited higher histological grade ($p < 0.001$), lower estrogen receptor (ER) expression, and elevated Ki-67 (all $P < 0.05$). Triple-negative breast cancer (TNBC) patients in the HER2-null subgroup were significantly younger ($P = 0.040$). Significant staining variability was observed between antibodies: the 4B5 clone was associated with more HER2 0 and 1+ scores, while MXR001 yielded higher rates of HER2-low results ($P < 0.05$). Inter-laboratory concordance was moderate (68.89%), being poorest for HER2 1+ (57.58%; Kappa = 0.566). AI-assisted interpretation improved observer agreement (Kappa: 0.703 vs. 0.610 in non-amplified cases) and reduced ambiguous immunohistochemistry (IHC) 2+/0 assignments. These findings delineate distinct clinicopathological characteristics of HER2-ultralow/null tumors, highlight HER2 IHC reproducibility challenges, and validate AI as an effective tool for standardizing scoring to optimize patient selection for T-DXd therapy.

Approximately 15% of breast cancers exhibit human epidermal growth factor receptor 2 (HER2) gene overexpression[1]. In breast cancer molecular classification, HER2 overexpression is recognized as an independent subtype[2], and targeted therapies such as trastuzumab have been developed against this receptor[3]. Current guidelines recommend a combined detection approach using HER2 immunohistochemistry (IHC) with fluorescence in situ hybridization (FISH)/in situ hybridization (ISH) to accurately identify patients eligible for HER2-targeted treatment[4]. In recent years, the novel antibody-drug conjugate trastuzumab deruxtecan (T-DXd) has transformed the treatment landscape by demonstrating efficacy even in HER2-non-amplified breast cancers—a population historically considered unresponsive to conventional HER2-targeted agents like trastuzumab[5]. This breakthrough stems from several key pharmacological advantages of T-DXd, including its unique mechanism to overcome drug resistance, enhanced tumor cell cytotoxicity, and a potent bystander effect[6–8]. Accumulating clinical evidence has confirmed that T-DXd significantly improves survival outcomes in patients with chemotherapy-resistant advanced breast cancer[9] and provides substantial clinical benefit for both HER2-low[10,11] and ultralow-expressing populations[12]. These advances have heightened the clinical relevance of precisely identifying HER2 expression levels, particularly in the low and ultralow ranges.

[1]Department of Pathology, the First Affiliated Hospital and College of Basic Medical Sciences, China Medical University, Shenyang, PR China. [2]Department of Medical Oncology, The First Hospital of China Medical University, Shenyang, China. [3]Department of Pathology, the Fourth Affiliated Hospital of China Medical University, Shenyang, Liaoning Province, PR China. ✉e-mail: qfzhang@cmu.edu.cn

However, accurate HER2 IHC testing—essential for guiding these therapeutic decisions—is hampered by multiple technical challenges. Pre-analytical variables such as variations in fixation time and fixative concentration, analytical inconsistencies in section thickness and staining conditions, and interpreter subjectivity during evaluation all contribute to substantial inter-laboratory variability[13,14]. This lack of reproducibility poses a critical barrier to the reliable identification of HER2-low and ultralow cases, potentially affecting patient access to appropriate targeted therapies.

To address these issues, this study systematically investigates the clinical significance and methodological reliability of HER2 IHC testing through three integrated dimensions: First, we analyze the clinicopathological characteristics associated with different HER2 expression categories—specifically distinguishing HER2-ultralow from HER2-null cases—to uncover potential biological and prognostic implications. Second, we conduct a multicenter reproducibility assessment of HER2 IHC staining and interpretation to identify key variables affecting consistency and to propose standardization strategies. Third, we evaluate the utility of artificial intelligence (AI) tools in improving interpretation agreement and accuracy, thereby providing practical insights into the evolving role of computational pathology in HER2 scoring.

Results

Clinicopathological differences between HER2-low and HER2-ultralow expression

Among 1455 HER2-non-amplified cases, 1379 (94.78%) were invasive breast carcinoma of no special type (NST), while 76 (5.22%) were special subtypes including 7 (0.48%) lobular carcinomas, 30 (2.06%) mucinous carcinomas, 27 (1.86%) micropapillary carcinomas, 4 (0.27%) cribriform carcinomas, 1 (0.07%) adenoid cystic carcinoma, and 7 (0.48%) metaplastic carcinomas. The cohort comprised 1,269 (87.22%) hormone receptor-positive (HR+) cases and 186 (12.78%) triple-negative breast cancers (TNBCs). HER2-low expression was identified in 1,119 cases (76.91%), HER2 0 in 336 (23.09%), HER2-ultralow in 134 (9.21%), and HER2-null in 202 (13.88%). No significant differences were observed among groups regarding patient age, tumor T stage, or N stage. Compared with HER2 0 (including HER2-ultralow and HER2-null) cases, HER2-low tumors showed significantly lower histological grade (all $p < 0.001$), higher estrogen receptor (ER)/ progesterone receptor (PR) expression, and lower Ki-67 indices (HER2-ultralow vs. HER2-low: ER $p < 0.001$, PR $P = 0.006$, Ki-67 $p < 0.001$; HER2-null vs. HER2-low: ER $p < 0.001$, PR $p < 0.001$, Ki-67 $p = 0.004$), while HER2-ultralow and HER2-null groups showed no such significant difference. In addition, HER2-null cases demonstrated significantly higher tumor-infiltrating lymphocytes (TILs) than HER2-low cases ($P = 0.005$) (Table 1).

All cases were stratified into HR+ breast cancer and TNBC groups. Among HR+ cases, HER2-ultralow cases showed higher ER expression ($p = 0.018$) compared to HER2-null cases, with ER expression levels showing positive correlation with HER2 IHC scores in HER2-low cases (HER2 2+ vs. HER2 1+, $p = 0.023$). However, no significant difference in ER expression exists between HER2-ultralow and HER2-low cases. Between HER2 IHC scores and patient age, T stage, N stage, or TILs levels, no significant associations were observed. Consistent with the overall analysis,

Table 1 | Comparison of clinicopathological features in HER2-negative breast cancer (n = 1455)

	HER2					*P*-value			
	2+	1+	Low	Ultralow	Null	2+ vs. 1+	Ultralow vs. Low	Ultralow vs. Null	Null vs. Low
Age									
≤50	192	193	385	55	82	0.119	0.128	0.934	0.090
>50	402	332	734	79	120				
T stage									
1–2	582	515	1097	131	199	0.890	1.000*	0.686#	0.856*
3–4	12	10	22	3	3				
N stage									
0	395	345	740	93	138	0.885	0.325	0.966	0.266
1–3	189	162	351	36	54				
Histological grade									
1	32	34	66	2	7	0.327	**<0.001**	0.170	**<0.001**
2	451	378	829	87	110				
3	108	110	218	45	81				
TILs									
<30%	545	477	1022	118	173	0.611	0.142	0.524	**0.005**
≥30%	46	45	91	16	29				
ER									
≤10%	70	77	147	36	67	0.154	**<0.001**	0.220	**<0.001**
>10%	524	448	972	98	135				
PR									
≤20%	172	142	314	53	91	0.478	**0.006**	0.319	**<0.001**
>20%	422	383	805	81	111				
Ki-67									
<30%	297	284	581	46	83	0.161	**<0.001**	0.212	**0.004**
≥30%	297	240	537	88	119				

Bold: p values < 0.05.
P-values with superscript "#" were calculated using continuity correction; *P*-values with superscript "*" were calculated using Fisher's exact test.

Table 2 | Clinicopathological characteristics of HR+/HER2- vs. triple-negative breast cancers by HER2 IHC scores (*n* = 1269 vs. 186)

	HR +									TNBC								
	HER2					*P*-value				HER2					*P*-value			
	2+	1+	L	U	N	2+ vs. 1+	U vs. L	U vs. N	N vs. L	2+	1+	L	U	N	2+ vs. 1 +	U vs. L	U vs. N	N vs. L
Age																		
≤50	182	173	355	41	59	0.378	0.257	0.777	0.352	10	20	30	14	23	**0.013**	0.155	0.748	**0.040**
>50	358	303	661	60	93					44	29	73	19	27				
T stage																		
1–2	528	469	997	98	149	0.377	0.701*	0.685#	1.000*	54	46	100	33	50	0.104#	1.000#	-	0.551#
3–4	12	7	19	3	3					0	3	3	0	0				
N stage																		
0	355	310	665	67	99	0.825	0.789	0.887	0.601	40	35	75	26	39	0.776	0.344	0.633	0.602
1–3	177	150	327	31	44					12	12	24	5	10				
Histological grade																		
1	32	33	65	2	7	0.805	0.194	0.141	**0.011**	0	1	1	0	0	**<0.001#**	0.162#	0.707#	**0.007#**
2	428	373	801	83	106					23	5	28	4	4				
3	77	67	144	16	35					31	43	74	29	46				
TILs																		
<30%	502	442	944	96	140	0.882	0.514	0.360	0.563	43	35	78	22	33	0.425	0.264	0.950	0.172
≥30%	35	32	67	5	12					11	13	24	11	17				
ER																		
≤10%	16	28	44	3	17	**0.023**	0.697*	**0.018**	**<0.001**	-	-	-	-	-	-	-		-
>10%	524	448	972	98	135													
PR																		
≤20%	118	93	211	20	41	0.364	0.819	0.192	0.083	-	-	-	-	-	-	-		-
>20%	422	383	805	81	111													
Ki-67																		
<30%	284	280	564	45	83	**0.042**	**0.034**	0.117	0.824	13	4	17	1	0	**0.030**	0.091*	0.402#	**0.003**
≥30%	256	195	451	56	69					41	45	86	32	50				

Bold: *p* values < 0.05. *L* low, *U* ultralow, *N* Null.
P-values with superscript "#" were calculated using continuity correction; *P*-values with superscript "*" were calculated using Fisher's exact test.

HR+ tumors with HER2-null demonstrated significantly higher histological grade ($p = 0.011$) compared to those with HER2-low, HER2-ultralow cases exhibited higher Ki-67 indices than HER2-low cases ($p = 0.042$). In addition, HER2 1+ cases showed lower Ki-67 indices than HER2 2+ cases ($p = 0.034$).

In TNBC cases, HER2 scores showed no association with T stage, N stage, or TILs. Among TNBC patients, no clinicopathological differences were observed between the HER2-ultralow and HER2-null subgroups. Separately, some variations were noted in other comparisons: the proportion of young patients (≤50 years) was higher in the HER2 1+ group than in the HER2 2+ group ($P = 0.013$), and a higher proportion of young patients was observed in the HER2-null group than in the combined HER2-low group ($P = 0.040$). Regarding histological grade, HER2-null TNBC demonstrated higher grades than HER2-low cases ($p = 0.007$), and HER2 1+ tumors showed higher grades than HER2 2+ tumors ($p < 0.001$). A similar pattern was observed for Ki-67 indices as for histological grade: specifically, both the HER2 1+ group (vs. HER2 2+, $p = 0.030$) and the HER2-null group (vs. the combined HER2-low group, $p = 0.003$) demonstrated significantly higher Ki-67 indices. (Table 2).

HER2 immunohistochemical staining performance

The overall staining concordance rate between all three groups of test slides and reference standards was 72.60%, with Tissue 1(HER2 2+) showing 55.56% concordance (1 case [2.22%] HER2 0 and 19 cases [42.22%] HER2 1+), Tissue 2(HER2 0) demonstrating 66.67% concordance (12 cases [26.67%] HER2 1+ and 3 cases [6.67%] HER2 2+), and Tissue 3(HER2 3+/2+) achieving 95.56% concordance (2 cases [4.44%] HER2 2+). Two laboratories (4.44%) failed external controls (Fig. 1).

Among different antibody clones, the 4B5 group showed higher HER2 1+ rates (78.95%) in HER2 2+, FISH- samples compared to antibody clones refer to laboratory-developed tests (LDTs) (MXR001 vs 4B5 $p = 0.003$; 4B5 vs. LDTs $p < 0.001$). In HER2 0 samples, the 4B5 group had higher HER2 0 rates (84.21%) than MXR001 (MXR001 vs. 4B5 $p = 0.017$), while in HER2 3+, Heterogeneous samples, MXR001 group showed higher HER2 3+ rates (69.23%) than LDTs clones (MXR001 vs. 4B5 $p = 0.011$; MXR001 vs. LDTs $p = 0.015$). Across different platforms, Roche Ventana showed significantly higher HER2 1+ rates (94.12% vs. 10.71%, $p < 0.05$) in HER2 2+, FISH- samples, higher HER2 0 rates (100% vs. 12.5%, $p < 0.05$) in HER2 0 samples compared to Leica BOND MAX/Titan/UltraPATH, and higher HER2 2+ rates (94.12% vs 0%, $p < 0.05$) in HER2 3+, Heterogeneous versus Titan (key findings summarized in Table 3). The full breakdown of the statistical data for all antibody clones and staining platforms can be found in Supplementary Tables S1 and S2.

Concordance of HER2 score interpretation

The overall concordance rate between the tested laboratories' interpretations and the reference scores was 68.89% across all three slide sets. Analyzed for each predefined reference score, the concordance was highest for

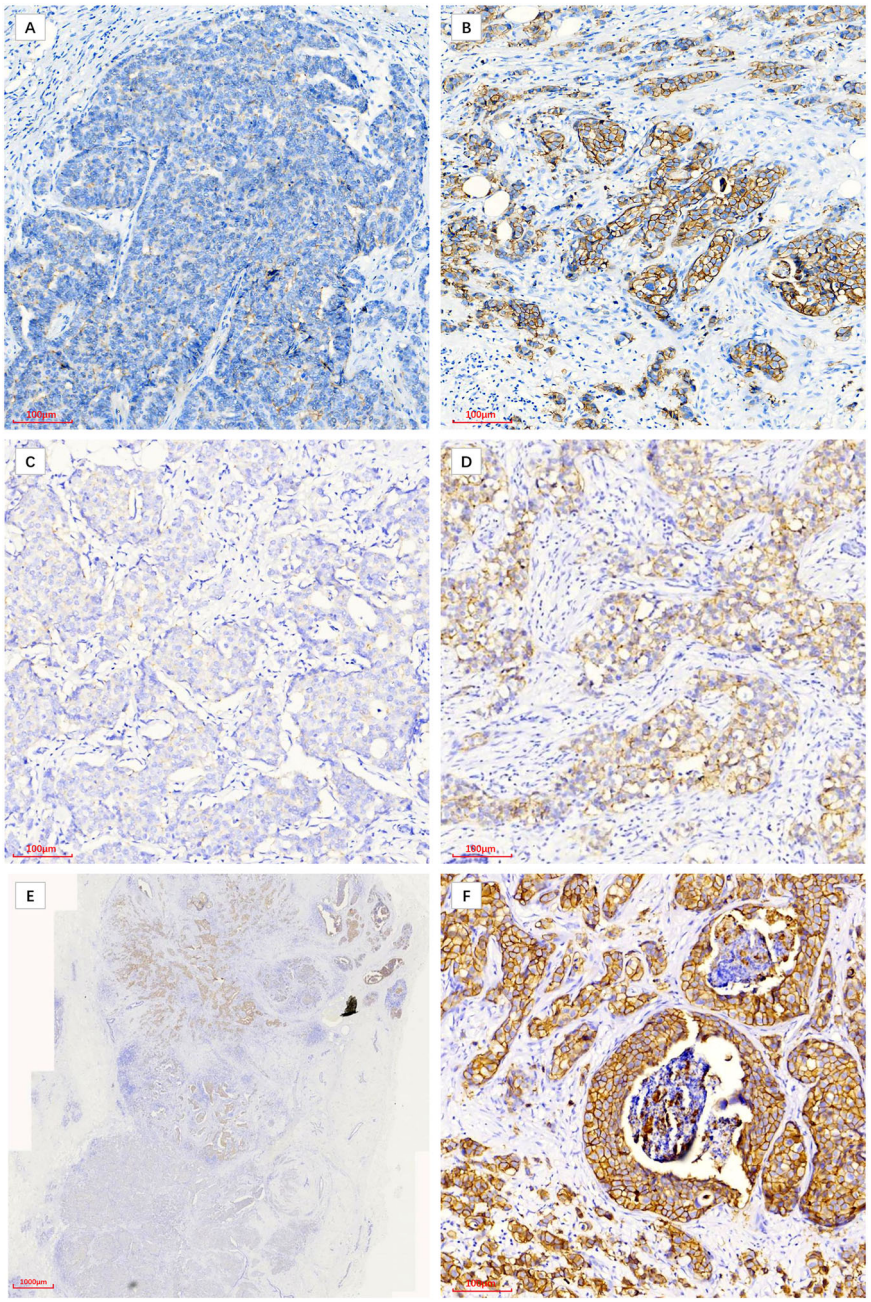

Fig. 1 | Examples of unacceptable HER2 IHC staining in test slides. A A test slide from Tissue 1 showing unacceptable IHC 1+ staining (200×). **B** Weak external control for the test slide from Tissue 1 (200×). **C** IHC 1+ region in a test slide from Tissue 3 (200×). **D** IHC 2+ region in the same test slide from Tissue 3 (200×). **E** Tumor heterogeneity observed in the test slide from Tissue 3 (20×). **F** Weak external control corresponding to the test slide from Tissue 3 (200×).

Table 3 | Analysis of key technical factors influencing HER2 immunohistochemical staining

Technical factor	Comparison group	Key finding[a]	*P*-value
Antibody Clone	MXR001 **vs**. 4B5	In HER2 2+ (FISH-) samples, the 4B5 clone yielded significantly more HER2 1+ scores, while MXR001 yielded significantly more HER2 2+ scores.	0.003
		In HER2 0 samples, the 4B5 clone yielded significantly more HER2 0 scores.	0.017
		In HER2 3+ (Heterogeneous) samples, the MXR001 clone yielded significantly more HER2 3+ scores.	0.011
	MXR001 **vs**. LDTs[c]	In HER2 3+ (Heterogeneous) samples, the MXR001 clone yielded significantly more HER2 3+ scores.	0.015
	4B5 **vs**. LDTs[c]	In HER2 2+ (FISH-) samples, the 4B5 clone yielded significantly more HER2 1+ scores.	<0.001
Staining Platform	Ventana **vs**. Other Platforms[d]	In HER2 2+ (FISH-) samples, the Ventana platform yielded significantly more HER2 1+ scores.	<0.05[b]
		In HER2 0 samples, the Ventana platform yielded significantly more HER2 0 scores.	<0.05[b]
		In HER2 3+ (Heterogeneous) samples, the Ventana platform yielded significantly more HER2 2+ scores (compared to the Titan platform).	<0.05[b]

[a]This table summarizes the key scenarios where significant disparities were identified between different antibody clones and staining platforms.
[b]For staining platforms, significant differences were determined by whether they belonged to the same subset in the χ^2 test at the 0.05 level (Table S2).
[c]LDTs (Antibody Clones): Refers to laboratory-developed tests using non-FDA-approved clones.
[d]"Other Platforms": Refers to a group of automated staining platforms (e.g., Leica BOND MAX, DAKO Link 48) used for pairwise comparisons against Ventana.

slides with a reference score of HER2 3+ (86.67%), followed by HER2 0 (70.97%), and lowest for HER2 1+ (57.58%). The concordance for slides defined as HER2-low was 65.56% (Fig. 2).

Overall interpretation consistency was moderate (kappa = 0.566, $p < 0.001$), with Tissue 2 (HER2 0) showing the highest agreement (kappa = 0.458, $p < 0.001$, moderate strength), while the other two groups demonstrated fair agreement (Tissue 1 [HER2 2+]: kappa = 0.391, $p = 0.001$; Tissue 3 [HER2 2+/3+]: kappa = 0.373, $p = 0.004$) (Table 4).

The role of AI in HER2 immunohistochemical interpretation

Following AI-assisted interpretation, the most frequent score adjustment across all groups (Pathologist A, Pathologist B, and Consensus Scoring [CS]) occurred in cases reclassified from HER2 0 to HER2 1+ (A: 10.5%; B: 4.3%; CS: 5.3%), resulting in an overall reduction of 3.7% in HER2 0 cases. For HER2 2+ cases, AI-assisted interpretation led to balanced redistributions toward both ends (HER2 1+ and HER2 3+), though manual interpretations exhibited a pronounced tendency to downgrade HER2 2+ to HER2 1+ rather than upgrade to HER2 3+ (A: 4.8% vs. 0.5%; B: 3.3% vs. 0%; CS: 2.9% vs. 0%). Notably, all groups demonstrated a modest yet consistent decrease in HER2 2+ cases (A: 1 case [3.0%]; B: 2 cases [7.1%]; CS: 3 cases [9.4%]), while changes in other score categories showed no discernible pattern (Fig. 3).

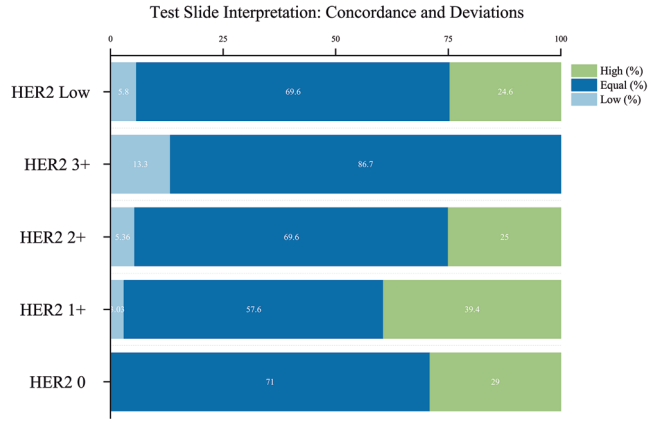

Fig. 2 | Concordance and variability in test slide evaluations.

AI assistance improved both concordance rates and agreement levels between pathologists, with overall interpretation concordance increasing from 79.43% to 85.17% and kappa values improving from 0.711 ($p < 0.001$) to 0.794 ($p < 0.001$). For non-amplified cases specifically, concordance rose from 77.02% to 82.61% with kappa improvement from 0.610 ($p < 0.001$) to 0.713 ($p < 0.001$). When comparing individual pathologists' results with consensus scores, Pathologist A's concordance improved from 86.60% to 91.39% (kappa 0.813–0.881), while Pathologist B's results advanced from 92.82% to 93.78% (kappa 0.899–0.913). Non-amplified cases showed relatively lower but still improved agreement (Pathologist A: kappa 0.751–0.836; Pathologist B: kappa 0.861–0.875; all $p < 0.001$) (Table 5, Fig. 4).

AI interpretation markers for HER2 3+ to 0 cases are demonstrated in Fig. 5. Manual review identified interpretation discrepancies in 35 AI-classified cases, primarily attributable to unrecognized cells (21/35, 60%) and cell type misclassification (9/35, 25.7%), with excessive staining intensity interfering with cell detection (3/35, 8.6%) and nonspecific cytoplasmic staining affecting interpretation accuracy (2/35, 5.7%) constituting remaining challenges.

Discussion

Previous studies have suggested that HER2-ultralow tumors demonstrate higher histological grade and lymph node metastasis rates along with decreased ER/PR expression compared to HER2-low cases, HER2-null tumors exhibit higher histological grade and lower ER/PR expression than HER2-ultralow cases[15]. Our findings confirm that invasive carcinomas of the HER2-null subtype are associated with a higher histological grade than HER2-low cases. Although the HER2-ultralow group also showed a trend towards higher grading compared to the HER2-low group in the overall cohort, this difference was no longer significant after stratifying the cases into HR+ breast cancer and TNBC subgroups for separate analysis. In HR+ tumors, we confirmed higher

Table 4 | Concordance analysis between testing laboratories and review team interpretations of HER2 immunohistochemical staining

Reference slides	Testing group score	Review team score				Kappa value	*P*-value
		HER2 0	HER2 1+	HER2 2+	HER2 3+		
HER2 2+, FISH-	HER2 0	0	0	0	0	0.391	0.001
	HER2 1+	1	9	0	0		
	HER2 2+	0	10	22	0		
	HER2 3+	0	0	3	0		
HER2 0	HER2 0	22	1	0	0	0.458	<0.001
	HER2 1+	8	10	3	0		
	HER2 2+	0	1	0	0		
	HER2 3+	0	0	0	0		
HER2 3+, heterogeneous	HER2 0	0	0	0	0	0.373	0.004
	HER2 1+	0	0	0	0		
	HER2 2+	0	2	17	2		
	HER2 3+	0	0	11	13		
Overall	HER2 0	22	1	0	0	0.566	<0.001
	HER2 1+	9	19	3	0		
	HER2 2+	0	13	39	2		
	HER2 3+	0	0	14	13		

The table shows the agreement between testing laboratories and the reference review team in HER2 immunohistochemical scoring. Kappa values indicate: 0.41–0.60 = moderate agreement; 0.21–0.40 = fair agreement.

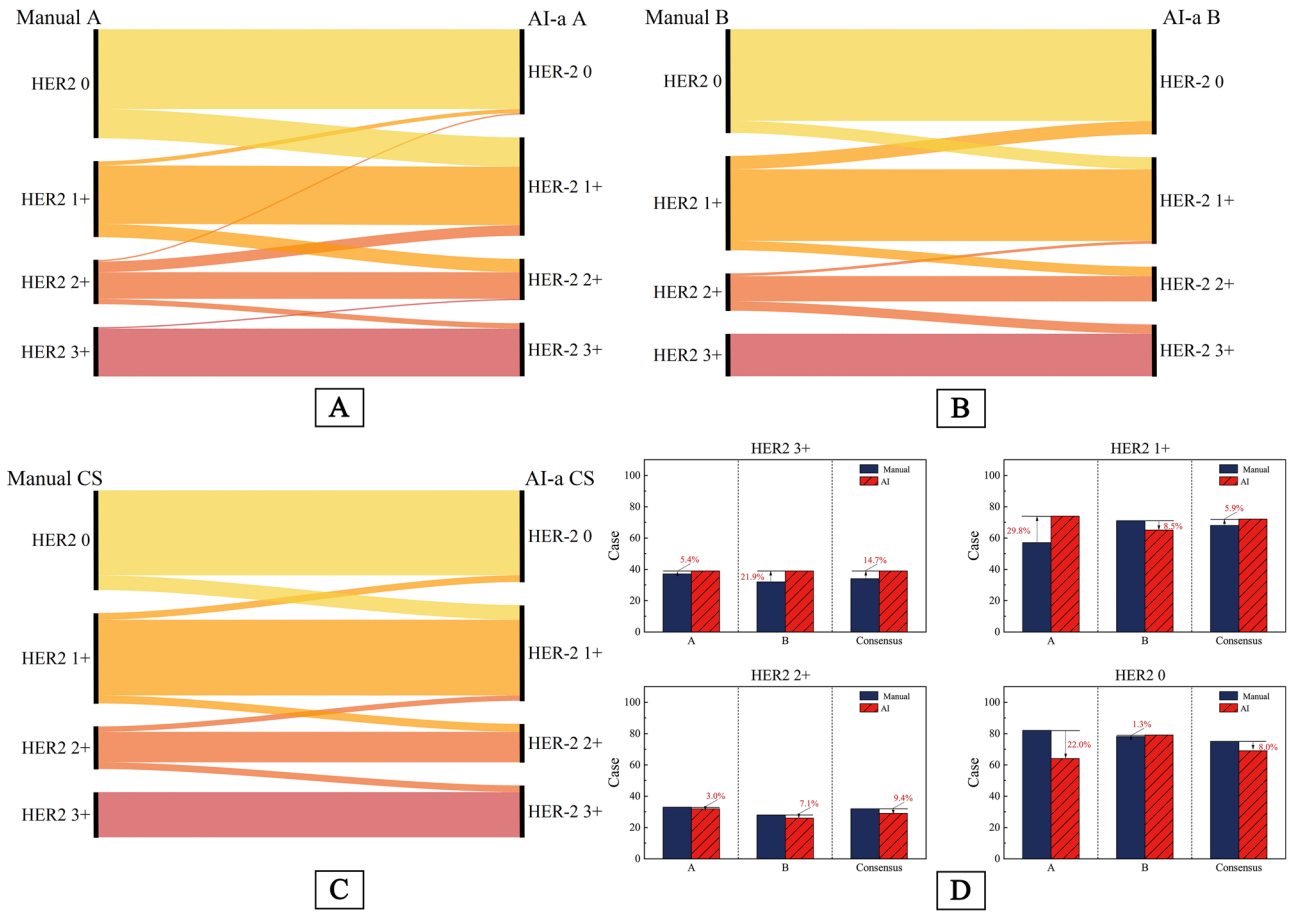

Fig. 3 | Distribution of manual and AI-assisted scoring results. A–C The changes in HER2 IHC scores across different groups following AI interpretation. **D** The quantitative changes in each HER2 IHC score category before and after AI analysis. AI-a AI-assisted, CS consensus scoring.

ER expression in HER2-ultralow versus HER2-null cases, suggesting that tumors with HER2-ultralow expression are more likely to benefit from endocrine therapy than those with HER2-null expression. Based on the combined characteristics above, we propose that HR+ breast cancers with HER2-ultralow more closely resemble tumors with HER2-low rather than HER2-null in their clinicopathological features. While in TNBCs, tumors with HER2-ultralow expression showed a trend toward closer resemblance to those with HER2-null expression, although this difference was not statistically significant. Although tumor-infiltrating lymphocytes (TILs) were significantly more abundant in HER2-null versus HER2-low tumors overall, this difference disappeared when analyzing HR+ and TNBC subgroups separately. This discovery may put forward more specific requirements for the stratification of treatment. Limited research exists on TILs in luminal/HER2-negative breast cancer, though some evidence suggests high TILs may correlate with poorer overall survival (but not disease-free survival) after neoadjuvant chemotherapy[16], contrasting with a new study showing a favorable prognosis associated with high TILs in young patients treated with adjuvant chemotherapy or endocrine therapy[17].

HER2 immunohistochemistry is the most cost-effective and efficient method for patient stratification. Quality control within and between laboratories is particularly important. Therefore, we conducted a multicenter staining and interpretation double assessment. Given the inter-laboratory variability in IHC staining and interpretation, we do not require further subcategorization of HER2 0 into HER2-ultralow and HER2-null in our scoring system. Our multicenter evaluation of HER2 IHC staining and interpretation consistency revealed that HER2 3+ staining showed superior reproducibility compared to HER2 0 and non-amplified HER2 2+ cases, with the latter demonstrating the lowest consistency. Ventana platforms showed distinct staining patterns for non-amplified HER2 2+ cases (higher IHC 1+ rates) and HER2 0 cases (more frequent HER2 0 calls) compared to other platforms. Antibody-specific analysis demonstrated that 4B5 clones generated significantly lower membrane positivity rates in non-amplified specimens, whereas MXR001 tended toward higher membrane staining in amplified cases, aligning with prior reports of 4B5's superior specificity versus HercepTest and MXR001[18]. These variations may directly impact treatment decisions for patients with HER2-low tumors, which underscores how staining reliability depends on multiple technical factors, with FDA-approved platforms/antibodies and rigorous internal quality control being essential.

Analysis of 45 laboratories revealed moderate overall interpretation consistency (kappa = 0.566), with HER2 3+ showing the highest agreement (86.67%), followed by HER2 0 (70.97%), while HER2 1+ demonstrated the poorest reproducibility (57.58%). The major discordance occurred between HER2 0 and 1+, which reflects the declining sensitivity of the 10% cutoff threshold for faint membrane staining[19,20]. Emerging requirements to distinguish HER2-ultralow from HER2-null cases will likely exacerbate these inconsistencies.

Nowadays, AI is gradually being put to use in pathological diagnosis. Therefore, whether AI can assist in the interpretation of IHC is also something we want to know. In our research, it was found that AI-assisted interpretation significantly improved inter-observer agreement while reducing ambiguous HER2 2+ calls and unnecessary FISH tests. However, current AI tools cannot reliably differentiate HER2-ultralow from HER2-null due to persistent challenges in recognizing nonspecific staining, with

Table 5 | Concordance rates and agreement between manual and AI-assisted HER2 scoring

Manual A	Manual B							HER2-Non-amplified							
								Manual A	Manual B						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 0	68	14	0	0	79.43%	0.711	<0.001	HER2 0	68	14	0	0	77.02%	0.61	<0.001
HER2 1+	10	45	2	0				HER2 1+	10	45	1	0			
HER2 2+	0	12	21	0				HER2 2+	0	12	11	0			
HER2 3+	0	0	5	32				HER2 3+	0	0	0	0			
AI-a A	AI-a B							AI-a A	AI-a B						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 0	62	2	0	0	85.17%	0.794	<0.001	HER2 0	62	2	0	0	82.61%	0.713	<0.001
HER2 1+	17	55	2	0				HER2 1+	17	55	2	0			
HER2 2+	0	8	23	1				HER2 2+	0	7	16	0			
HER2 3+	0	0	1	38				HER2 3+	0	0	0	0			
Manual CS	Manual A							HER2-Non-amplified							
								Manual CS	Manual A						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 0	70	5	0	0	86.60%	0.813	<0.001	HER2 0	70	5	0	0	85.09%	0.751	<0.001
HER2 1+	12	50	6	0				HER2 1+	12	50	6	0			
HER2 2+	0	2	27	3				HER2 2+	0	1	17	0			
HER2 3+	0	0	0	34				HER2 3+	0	0	0	0			
AI-a CS	AI-a A							AI-a CS	AI-a A						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 0	62	7	0	0	91.39%	0.881	<0.001	HER2 0	62	7	0	0	90.06%	0.836	<0.001
HER2 1+	2	65	5	0				HER2 1+	2	65	5	0			
HER2 2+	0	2	26	1				HER2 2+	0	2	18	0			
HER2 3+	0	0	1	38				HER2 3+	0	0	0	0			
Manual CS	Manual B							HER2-Non-amplified							
								Manual CS	Manual B						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 0	73	2	0	0	92.82%	0.899	<0.001	HER2 0	73	2	0	0	91.93%	0.861	<0.001
HER2 1+	5	63	0	0				HER2 1+	5	63	0	0			

Table 5 (continued) | Concordance rates and agreement between manual and AI-assisted HER2 scoring

Manual CS	Manual B							HER2-Non-amplified							
								Manual CS	Manual B						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 2+	0	6	26	0				HER2 2+	0	6	12	0			
HER2 3+	0	0	2	32				HER2 3+	0	0	0	0			
AI-a CS	AI-a B							AI-a CS	AI-a B						
	HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P		HER2 0	HER2 1+	HER2 2+	HER2 3+	CR	κ	P
HER2 0	69	0	0	0	93.78%	0.913	<0.001	HER2 0	69	0	0	0	92.55%	0.875	<0.001
HER2 1+	10	62	0	0				HER2 1+	10	62	0	0			
HER2 2+	0	3	26	0				HER2 2+	0	2	18	0			
HER2 3+	0	0	0	39				HER2 3+	0	0	0	0			

CR concordance rate, *κ* Kappa, *AI-a* AI-assisted.

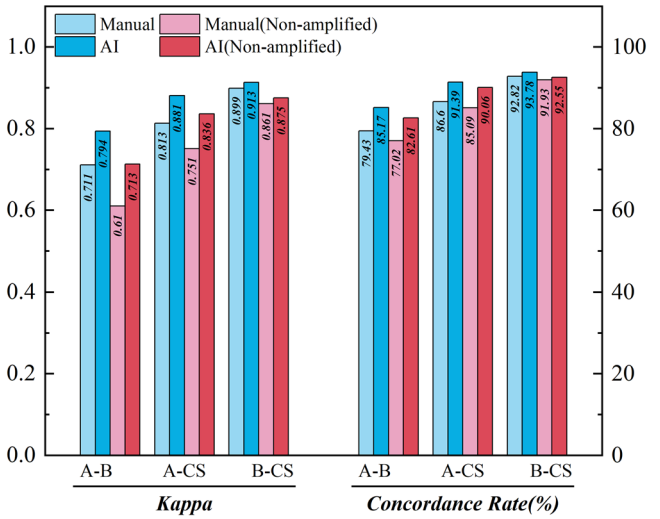

Fig. 4 | Concordance analysis between manual and AI-assisted HER2 scoring. CS consensus scoring.

limitations including inaccurate invasive carcinoma detection and cytoplasmic staining misinterpretation. While conventional machine learning algorithms struggle with image logic interpretation—an area where advanced deep learning shows promise for pathological diagnosis[21]—our results confirm AI's clinical utility in expanding patient eligibility for novel targeted therapies, albeit still requiring manual verification for optimal accuracy. The ultimate clinical value of HER2 IHC stratification warrants further investigation.

Methods

Clinicopathological characteristics of HER2-low and HER2-ultralow expression

This retrospective study identified and included a total of 1455 treatment-naïve cases of HER2-non-amplified invasive breast cancer from the pathological archives of the First Affiliated Hospital of China Medical University. The cases, consecutively accessioned between July 2022 and July 2024, were selected based on the availability of complete clinical, pathological, and immunohistochemical data. The study protocol was reviewed and approved by the Ethics Committee of the First Affiliated Hospital of China Medical University (Approval Number: 2025-79). This study was conducted in accordance with the Declaration of Helsinki. The committee granted a waiver for informed consent due to the retrospective nature of the analysis. Tumor-infiltrating lymphocytes (TILs) were assessed according to the 2024 Chinese Society of Clinical Oncology (CSCO) Breast Cancer Diagnosis Guidelines[22]. HER2 status was evaluated following the Guidelines for HER2 Testing in Breast Cancer (2024 Edition)[23].

Reproducibility of HER2 immunohistochemical testing

Forty-five laboratories were evaluated for staining performance and HER2 interpretation. Three formalin-fixed, paraffin-embedded (FFPE) tissue blocks with distinct HER2 scores were used to prepare control and test slides. Control slides were processed using the Roche Ventana platform with the 4B5 antibody clone. The HER2 scores for each tissue were as follows: Tissue 1: 2+ (IHC 2+ and FISH-negative); Tissue 2: Ultralow expression (<10% of tumor cells showing faint, incomplete membrane staining); Tissue 3: 3+ or 2+ (demonstrating heterogeneity with 50% 3+ and 50% 2+ staining areas confirmed as FISH-amplified) (Fig. 6).

To account for inter-laboratory variability, the interpretation criteria for staining performance are as follows: Tissue 1: 2+, Tissue 2: 0 (<10% of tumor cells showing faint, incomplete membrane staining), Tissue 3: 3+ or 2+ (with more than 50% 2+ staining areas). Interpretative consistency is allowed within a reasonable range. The acceptable reference score ranges for test slides were defined as: Tissue 1: 1+ or 2+; Tissue 2: 1+ or 0;

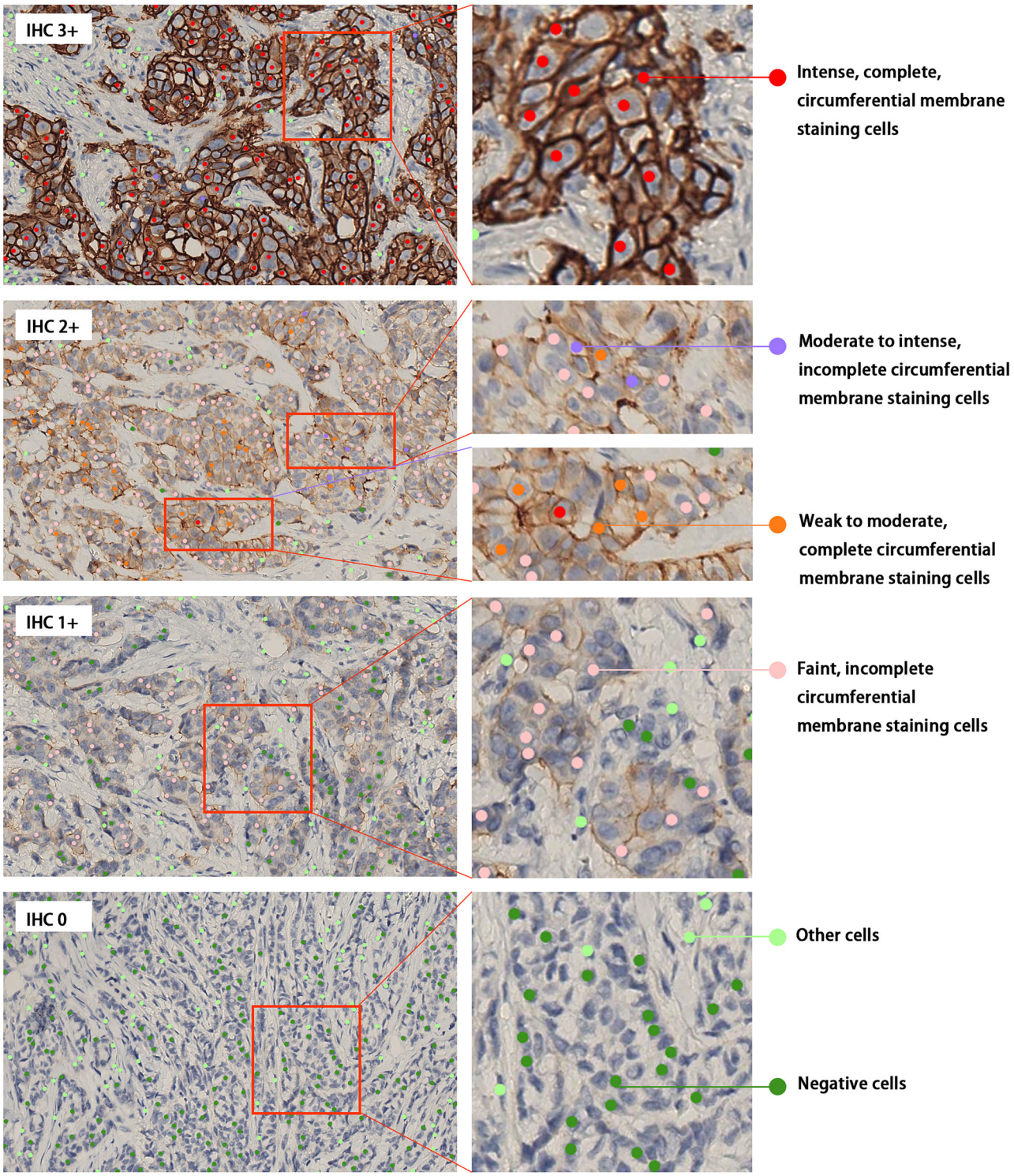

Fig. 5 | AI interpretation examples.

Tissue 3: 2+ or 3+. Test slides were deemed non-conforming if they met any of the following criteria: Failure in internal or external controls; Presence of nonspecific heterogeneity in invasive carcinoma regions.

Significance of AI tools in HER2 immunohistochemical assessment

This study conducted a comparative analysis of 209 HER2 immunohistochemical (IHC) slides from treatment-naïve postoperative invasive breast cancer cases at the First Affiliated Hospital of China Medical University between July 22, 2019, and July 22, 2022. The study evaluated the inter-pathologist concordance in IHC interpretation, the agreement between individual pathologists and consensus scoring (CS), as well as the differences in interpretation accuracy before and after AI-assisted evaluation. The AI tool employed in this study was the D-Path AI platform developed by Dpath Technology Co., Ltd.

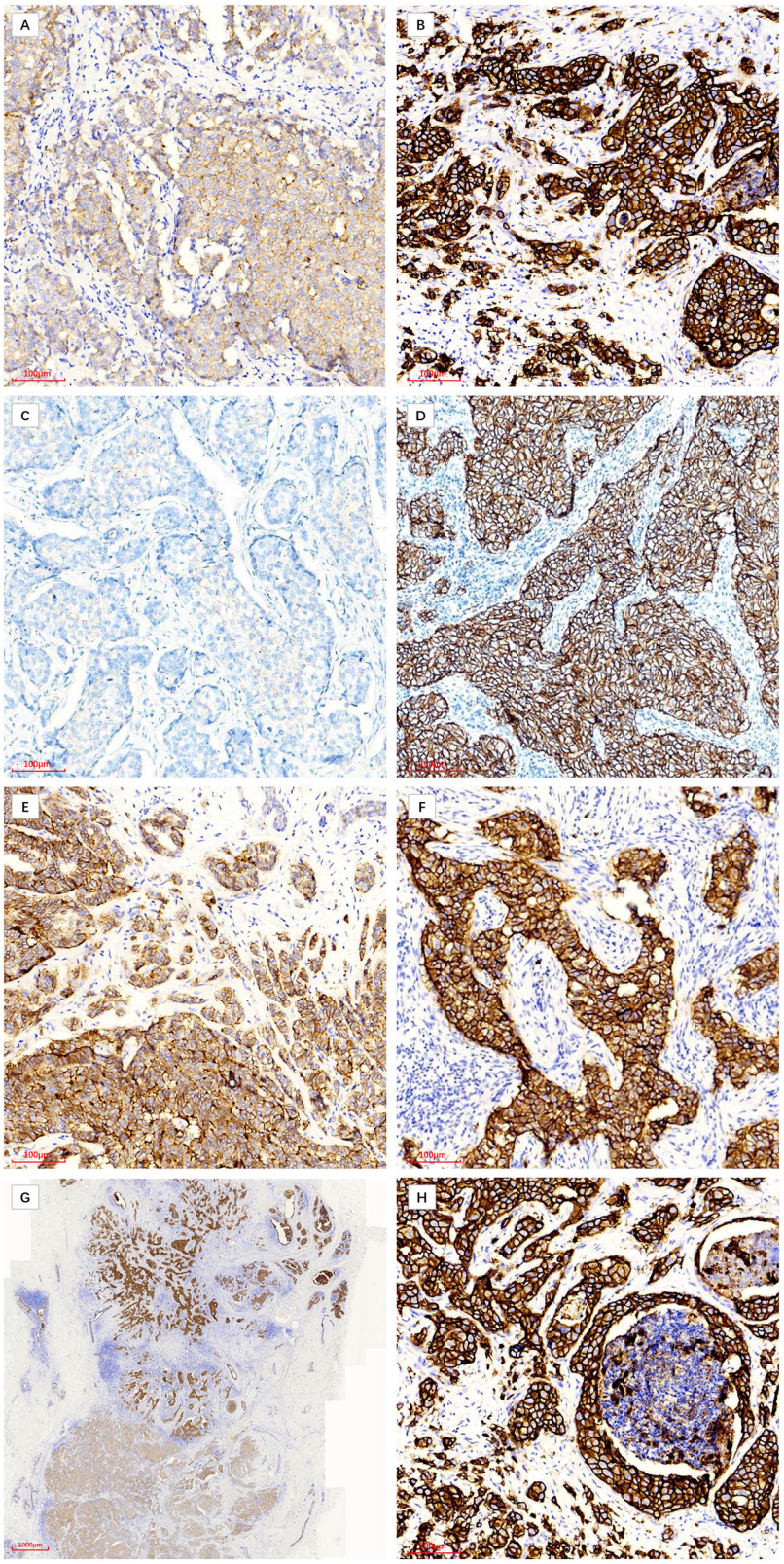

Fig. 6 | HER2 IHC staining of reference tissues and their external controls. **A** Tissue 1: HER2 2+ (200×). **B** External control for Tissue 1 (200×). **C** Tissue 2: HER2 0 (<10% of tumor cells with incomplete/faint membranous staining) (200×). **D** External control for Tissue 2 (200×). **E** IHC 2+ region (FISH-positive) of Tissue 3 (200×). **F** IHC 3+ region of Tissue 3 (200×). **G** Tumor heterogeneity in Tissue 3 (approximately 50% area 3+, 50% area 2+) (20×). **H** Positive control for Tissue 3 (200×).

Statistical analysis

All statistical analyses were performed using SPSS 24.0 software. This study employed both comparative analysis and agreement analysis, with the former conducted using χ^2 tests, Monte Carlo simulations, and Fisher's exact probability tests. The kappa coefficient was used to evaluate inter-rater agreement, with the following interpretation guidelines: kappa ≤ 0.20 indicating poor agreement, 0.20 < kappa ≤ 0.40 indicating fair agreement, 0.40 < kappa ≤ 0.60 indicating moderate agreement, 0.60 < kappa ≤ 0.80 indicating good agreement, and 0.80 < kappa ≤ 1.00 indicating excellent agreement.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy concerns, but are available from the corresponding author on reasonable request.

Code availability

Not applicable.

Received: 18 August 2025; Accepted: 21 November 2025;
Published online: 09 December 2025

References

1. Ge, S. et al. Review: Radionuclide molecular imaging targeting HER2 in breast cancer with a focus on molecular probes into clinical trials and small peptides. *Molecules* https://doi.org/10.3390/molecules26216482 (2021).
2. Angerilli, V. et al. The role of the pathologist in the next-generation era of tumor molecular characterization. *Diagnostics* https://doi.org/10.3390/diagnostics11020339 (2021).
3. Riccardi, F., Dal Bo, M., Macor, P. & Toffoli, G. A comprehensive overview on antibody-drug conjugates: from the conceptualization to cancer therapy. *Front. Pharmacol.* **14**, 1274088 (2023).
4. Marchiò, C. et al. Think "HER2" different: integrative diagnostic approaches for HER2-low breast cancer. *Pathologica* **115**, 292–301 (2023).
5. Fehrenbacher, L. et al. NSABP B-47/NRG oncology phase III randomized trial comparing adjuvant chemotherapy with or without trastuzumab in high-risk invasive breast cancer negative for HER2 by FISH and with IHC 1+ or 2. *J. Clin. Oncol.* **38**, 444–453 (2020).
6. Xu, Z. et al. Novel HER2-targeting antibody-drug conjugates of trastuzumab beyond T-DM1 in breast cancer: trastuzumab deruxtecan(DS-8201a) and (Vic-)trastuzumab duocarmazine (SYD985). *Eur. J. Med. Chem.* **183**, 111682 (2019).
7. Nakada, T., Sugihara, K., Jikoh, T., Abe, Y. & Agatsuma, T. The latest research and development into the antibody-drug conjugate, [fam-] trastuzumab deruxtecan (DS-8201a), for HER2 cancer therapy. *Chem. Pharm. Bull.* **67**, 173–185, https://doi.org/10.1248/cpb.c18-00744 (2019).
8. Meric-Bernstam, F. et al. Efficacy and safety of trastuzumab deruxtecan in patients with HER2-expressing solid tumors: primary results from the DESTINY-PanTumor02 phase II trial. *J. Clin. Oncol.* **42**, 47–58 (2024).
9. Cortés, J. et al. Trastuzumab deruxtecan versus trastuzumab emtansine in HER2-positive metastatic breast cancer: long-term survival analysis of the DESTINY-Breast03 trial. *Nat. Med.* **30**, 2208–2215 (2024).
10. Modi, S. et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N. Engl. J. Med.* **387**, 9–20 (2022).
11. Ivanova, M. et al. Standardized pathology report for HER2 testing in compliance with 2023 ASCO/CAP updates and 2023 ESMO consensus statements on HER2-low breast cancer. *Virchows Arch.* **484**, 3–14 (2024).
12. Bardia, A. et al. Trastuzumab deruxtecan after endocrine therapy in metastatic breast cancer. *N. Engl. J. Med.* **391**, 2110–2122 (2024).
13. Rhodes, A., Jasani, B., Anderson, E., Dodson, A. R. & Balaton, A. J. Evaluation of HER-2/neu immunohistochemical assay sensitivity and scoring on formalin-fixed and paraffin-processed cell lines and breast tumors: a comparative study involving results from laboratories in 21 countries. *Am. J. Clin. Pathol.* **118**, 408–417 (2002).
14. Roche, P. C. et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. *J. Natl Cancer Inst.* **94**, 855–857 (2002).
15. Hatami, E., Fanaei, S. M., Hasteh, F., Fadare, O. & Zare, S. 162 HER2-ultra low and HER2-null Breast Cancers, Incidence and Correlation with Clinicopathologic Parameters. *Lab. Invest.* **105**, 102386 (2025).
16. Denkert, C. et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **19**, 40–50 (2018).
17. Tesch, M. E. et al. Association of tumor-infiltrating lymphocyte subtypes with clinical characteristics and prognosis in young women with hormone receptor-positive breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research, https://doi.org/10.1158/1078-0432.Ccr-25-0948 (2025).
18. Lucas, E. et al. Comparison of Dako HercepTest and Ventana PATHWAY anti-HER2 (4B5) tests and their correlation with fluorescent in situ hybridization in breast carcinoma. *Appl. Immunohistochem. Mol. Morphol.* **27**, 403–409 (2019).
19. Tarantino, P. et al. ESMO expert consensus statements (ECS) on the definition, diagnosis, and management of HER2-low breast cancer. *Ann. Oncol.* **34**, 645–659 (2023).
20. Fernandez, A. I. et al. Examination of low ERBB2 protein expression in breast cancer tissue. *JAMA Oncol.* **8**, 1–4 (2022).
21. Liu, Y., Han, D., Parwani, A. V. & Li, Z. Applications of artificial intelligence in breast pathology. *Arch. Pathol. Lab. Med.* **147**, 1003–1013 (2023).
22. Li, J. et al. Chinese Society of Clinical Oncology (CSCO) breast cancer guidelines 2024. *Transl. Breast Cancer Res.* **5**, 18 (2024).
23. Chinese Society of Pathology, C. M. A. Guideline for HER2 testing in breast cancer (2024 version). *Zhonghua bing li xue za zhi = Chinese journal of pathology* **53**, 1192−1202. https://doi.org/10.3760/cma.j.cn112151-20241009-00664 (2024).

Acknowledgements

This work was partially supported by: 1. the National Natural Science Foundation of China (Grant No. 82404068); 2. Natural Science Foundation of Liaoning Province, China (No. 2024-MS-065); 3. The Basic Scientific Research Project of Liaoning Provincial Department of Education (No. LJ212410159098).

Author contributions

Yu-han Dong: methodology, validation, formal analysis, investigation, resources, data curation, writing – original draft, visualization. Yi-tong Xu: validation, data curation, writing – review & editing. Xi-zi Jiang: writing – review & editing. Lu Xu: funding acquisition, writing – review & editing. Rui-Jiao Yu: writing – review & editing, data curation. Rui Wang: funding acquisition. Qing-fu Zhang: conceptualization, methodology, formal analysis, investigation, resources, writing – review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41523-025-00872-2.

Correspondence and requests for materials should be addressed to Qing-fu Zhang.