npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01483-8

Systematic review and meta-analysis of artificial intelligence in classifying HER2 status in breast cancer immunohistochemistry

Daniel Arruda Navarro Albuquerque[1], Matheus Trotta Vianna[2] ✉, Luana Alencar Fernandes Sampaio[3], Andrei Vasiliu[4] & Eduardo Henrique Cunha Neves Filho[5]

The DESTINY-Breast04 trial has recently demonstrated survival benefits of trastuzumab-deruxtecan (T-DXd) in metastatic breast cancer patients with low Human Epidermal Growth Factor Receptor 2 (HER2) expression. Accurate differentiation of HER2 scores has now become crucial. However, visual immunohistochemistry (IHC) scoring is labour-intensive and prone to high interobserver variability, and artificial intelligence (AI) has emerged as a promising tool in diagnostic medicine. We conducted a diagnostic meta-analysis to evaluate AI's performance in classifying HER2 IHC scores, demonstrating high accuracy in predicting T-DXd eligibility, with a pooled sensitivity of 0.97 [95% CI 0.96–0.98] and specificity of 0.82 [95% CI 0.73–0.88]. Meta-regression revealed better performance with deep learning and patch-based analysis, while performance declined in externally validated and those utilising commercially available algorithms. Our findings indicate that AI holds promising potential in accurately identifying HER2-low patients and excels in distinguishing 2+ and 3+ scores.

Breast cancer is the leading cause of cancer among women worldwide and is expected to result in ~1 million deaths per year by 2040[1]. Between 15% and 20% of cases exhibit overexpression of the Human Epidermal Growth Factor Receptor 2 (HER2)[2], which is associated with poor clinical outcomes. The HER2 is a transmembrane tyrosine kinase receptor often found in breast cancer cells, and its activation promotes cell cycle progression and proliferation[3]. HER2 expression is routinely assessed semi-quantitatively by pathologists using immunohistochemistry (IHC)-stained slides, where HER2 is categorised based on the intensity and completeness of membrane staining into negative (scores 0 and 1+), equivocal (score 2+), or positive (score 3+). Equivocal cases undergo further testing with in situ hybridisation (ISH), and are subsequently classified as either positive or negative depending on HER2 amplification status[2].

Initially, the use of antibody-drug conjugates, such as trastuzumab-deruxtecan (T-DXd), has been recommended for HER2-positive patients with metastatic disease, corresponding to those with IHC scores of 3+, or 2+ with gene amplification confirmed by ISH. Patients with scores of 0, 1+, or 2+ with negative ISH were typically eligible for clinician's choice chemotherapy[4]. However, findings from the DESTINY-Breast04 (DB-04)[5] trial demonstrated that T-DXd significantly improved both progression-free and overall survival in individuals with low expression of HER2, with the latter particularly remarkable in the context of metastatic disease. The eligibility criteria for low HER2 expression in that study required an IHC score of 1+, or 2+ with negative ISH, in patients with metastatic breast cancer. This benefit in survival, observed in a subset of patients categorised as "HER2-low", a designation introduced by the DB-04 trial, has prompted the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) to emphasise the importance of distinguishing between scores 0 and 1+[6], while maintaining the 2018[2] four-class classification system.

Approximately 60% of HER2-negative metastatic breast cancers express low levels of HER2[5]. Accurate differentiation between scores 1+/2+/3+ and 0 has become critical, as it influences clinical decision-making and affects the number of patients eligible for T-DXd.

Nevertheless, the standard IHC visual scoring of HER2 is time-consuming, labour-intensive, and subject to high inter-observer variability[7,8]. Digital pathology has emerged as a promising tool in medical diagnostics, aiding in the diagnosis and grading of cancers such as breast, lung, liver, skin,

[1]School of Medicine, University of Central Lancashire, Preston, UK. [2]Economics Department, School of Social Sciences, The University of Manchester, Manchester, UK. [3]Department of Oncology, Hospital Sírio-Libanês, São Paulo, Brazil. [4] Division of Musculoskeletal & Dermatological Sciences, The University of Manchester, Manchester, UK. [5]Pathus Lab, Fortaleza, Brazil. ✉e-mail: matheus.vianna@manchester.ac.uk

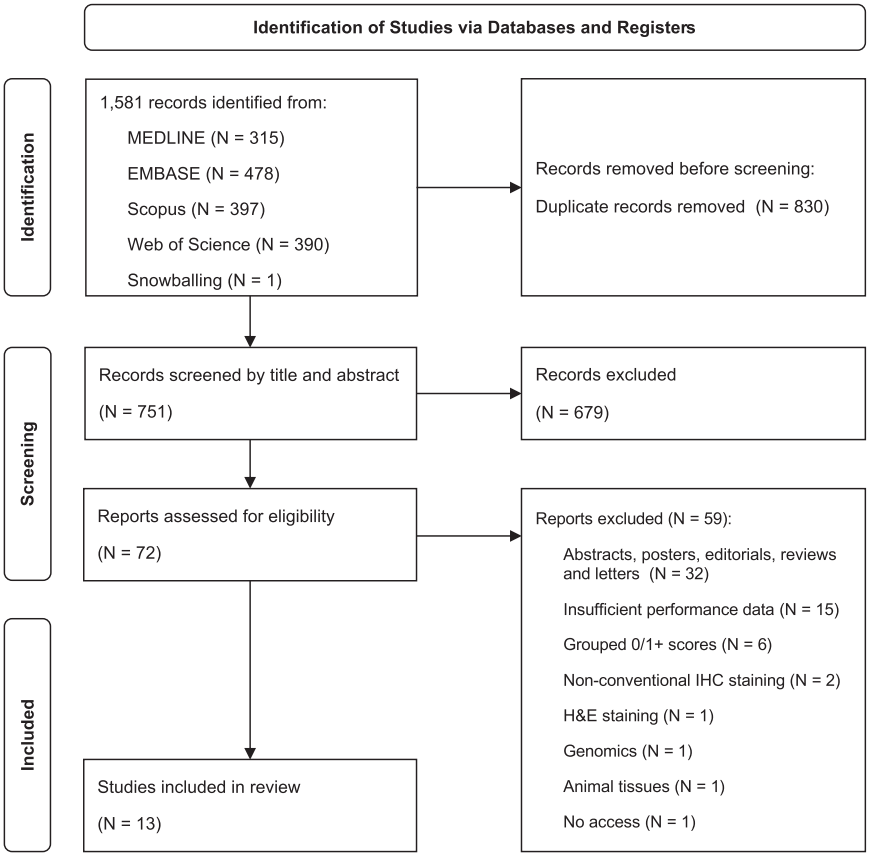

Fig. 1 | PRISMA flow diagram. The flowchart illustrates the process of identifying and screening studies, from database searches to the final selection of studies meeting the eligibility criteria. H&E haematoxylin and eosin, IHC immunohistochemistry.

and pancreatic cancers[9]. To address the subjective nature of IHC visual scoring, artificial intelligence (AI) and deep learning (DL) have become powerful options for analysing histological specimens, particularly for HER2 scoring[10].

Briefly, the workflow of an AI-based HER2 scoring model involves preprocessing digitalised whole slide images (WSI) by fragmenting the images into smaller patches and selecting a region of interest (ROI) likely to contain relevant neoplastic tissue. These patches are then split into a learning and testing subsets to train and evaluate the algorithm's performance, respectively. An internal validation process is subsequently performed to assess the model's accuracy using WSIs from the same training dataset. To assess the algorithm's reproducibility, some models undergo external validation, where performance is tested with an independent external dataset. The final outcome is the overall HER2 score, which is derived from the scoring aggregation of each individual patch[10].

In the context of digital pathology, patches are cropped, fixed-size image sections (typically square or rectangular) extracted from full-size WSIs. These patches facilitate computational analysis of high-resolution WSIs by breaking down the image into manageable units[11]. For the purpose of this review, the term "cases" refers to a set of WSIs obtained from a single patient.

In line with the challenges associated with HER2 scoring and the advancements in AI, we conducted an updated and clinically focused meta-analysis aimed at: (1) evaluating the performance of AI compared to pathologists' visual scoring in accurately identifying patients eligible for T-DXd based on HER2 IHC score; and (2) assessing the accuracy of AI in classifying each individual HER2 score.

Results

Study selection and characteristics

Our search yielded a total of 1581 records published from inception to April 2024, which were reduced to 751 after the removal of 830 duplicates. One relevant study[12] was found through article's bibliography search. The remaining records were then screened for relevance based on their titles and abstracts. Subsequently, 72 records were deemed suitable for inclusion and were assessed in full text. Of these, 59 studies were excluded, with the reasons summarised in Fig. 1. Publications that grouped 0/1+ scores together and those that reported accuracy metrics solely as percentages without providing raw data were excluded. Corresponding authors were contacted in both instances, but failed to provide the necessary data. Finally, a total of 13 studies were included in this review for meta-analysis.

Table 1 presents a summary of the included studies published between 2018 and 2024, encompassing 1285 cases, 168 WSIs, and 24,626 patches collected from 25 contingency tables. Only Palm et al.[13] and Sode et al.[14] specified the use of primary tumours, while the others did not mention the source. All authors reported invasiveness status, except for Fan et al.[15]. The type of tissue extraction (biopsy vs. resection) was reported in only five studies. The IHC assay was described in five studies, and Fan et al.[15] failed to give scanner details. The discrepancies in sample sizes observed in eight studies are attributable to the use of patches as the data unit, as multiple patches can be derived from a single WSI. Eight studies utilised, either wholly or in part, the HER2 Scoring Contest (HER2SC)[16] database, which may have resulted in some overlap of sample sizes. However, quantification of this was not possible, as authors did not specify which cases were used in their analyses. Ten studies employed DL, all of which utilised convolutional neural networks (CNN) algorithms. Only two studies exclusively utilised pathologist-assisted algorithms. In Pham et al.[17], pathologists manually selected an ROI before AI classification, while in Sode et al.[14], the final scoring required a pathologist's review. Palm et al.[13] and Jung et al.[18] tested the same algorithm for both automated and pathologist-assisted modes. One contingency table from Qaiser et al.[16] was excluded, as it used haematoxylin and eosin (H&E)-stained images as inputs for training the algorithm. Oliveira et al.[12] reported performance metrics using patches obtained from IHC-stained WSIs as an input for further

https://doi.org/10.1038/s41746-025-01483-8

Table 1 | Characteristics of the 13 included studies based on 25 contingency tables

First author and year	Participants	Sample size	Data unit	Index test	Reference standard	Deep learning	Transfer learning	Autonomy	Internal validation	External validation	Commercially available	Dataset
Borquez et al. (2023)[40]	86 cases of invasive breast carcinoma	41 2898	WSIs Patches	Bayesian deep learning with Monte Carlo Dropout (CNN)	ASCO/ CAP 2013	Yes	No	Fully automated	5-fold cross-validation	No	No	HER2SC
Fan et al. (2024)[15]	104 cases of breast tumour with overexpressed HER2	4000	Patches	CNN	Average score of three pathologists (NR)	Yes	No	Fully automated	Random split-sample validation	No	No	Nanjing Medical University
Jung et al. (2024)[18]	201 breast cancer patients	201	Cases	CNN: DeepLabv3+ ResNet-34 DeepLabv3 ResNet-101	ASCO/ CAP 2023	Yes	Yes	Fully automated Assisted	Random split-sample validation	Yes	No	Cureline (Brisbane, CA, US) Superbiochips (Seoul, Republic of Korea) Kyung Hee University Hospital (Seoul, Republic of Korea)
Kabir et al. (2024)[41]	86 cases of invasive breast carcinoma	77 5626	WSIs Patches	CNN: DenseNet201 GoogleNet MobileNetV2 Vision Transformer	ASCO/ CAP 2018	Yes	Yes	Fully automated	5-fold cross-validation	No	No	HER2SC
Mirimoghaddam et al. (2024)[42]	86 cases of invasive breast carcinoma 126 cases of invasive ductal breast carcinoma	3200	Patches	CNN: InceptionResNetV2 InceptionV3	ASCO/ CAP 2018	Yes	Yes	Fully automated	5-fold cross-validation	No	No	HER2SC Ghaem Educational Research and Treatment Centre (Mashhad University)
Mukundan (2019)[43]	86 cases of invasive breast carcinoma	1206	Patches	Multi-class logistic regression	ASCO/ CAP 2018	No	No	Fully automated	5-fold cross-validation	No	No	HER2SC
Oliveira et al. (2020)[12]	52 cases of invasive breast cancer	2495	Patches	CNN	ASCO/ CAP 2018	Yes	No	Fully automated	Random split-sample validation	No	No	HER2SC
Palm et al. (2023)[13]	Core needle biopsies of 67 primary invasive breast cancer	67 201	Cases Patches	HER2 4B5® (Roche, Switzerland)	ASCO/ CAP 2018	No	No	Fully automated Assisted	NR	Yes	Yes	Pathologie Institute Enge (Zurich, Switzerland)
Pedraza et al. (2024)[44]	86 cases of invasive breast carcinoma 52 cases of invasive ductal carcinomas 15 cases of invasive and in situ ductal carcinomas	5000	Patches	CNN: ResNet-101 DenseNet-201	ASCO/ CAP 2018	Yes	Yes	Fully automated	Random split-sample validation	Yes	No	HER2SC Servicio de Salud de Castilla-La Mancha Servicio Andaluz de Salud
Pham et al. (2023)[44]	86 cases of invasive breast carcinoma	50	WSIs	CNN: U-Net DenseNet121 ImageNet ResNet18 ImageNet	ASCO/ CAP 2018	Yes	Yes	Assisted	4-fold cross-validation	Yes	No	HER2SC Erasme Hospital

Table 1 (continued) | Characteristics of the 13 included studies based on 25 contingency tables

First author and year	Participants	Sample size	Data unit	Index test	Reference standard	Deep learning	Transfer learning	Autonomy	Internal validation	External validation	Commercially available	Dataset
Qaiser et al. (2018)[16]	86 cases of invasive breast carcinoma	28	Cases	MUCS-1 team: AlexNet (CNN) VISILAB team: GoogleNet (CNN)	ASCO/ CAP 2018	Yes	Yes	Fully automated	Random split-sample validation	No	No	"Men versus Machine" event of HER2SC
Sode et al. (2023)[14]	727 cases of invasive primary breast carcinoma	761	Cases	HER2-CONNECT® (Visiopharm, Denmark)	ASCO/ CAP 2018	No	No	Assisted	NR	Yes	Yes	Zealand University Hospital
Yao et al. (2022)[45]	228 biopsies of invasive breast carcinoma of no special type	228	Cases	GrayMap + CNN	ASCO/ CAP 2018	Yes	No	Fully automated	5-fold cross-validation	No	No	Peking University Cancer Hospital & Institute

ASCO American Society of Clinical Oncology, *CAP* College of American Pathologists, *CNN* convolutional neural networks, *HER2* Human Epidermal Growth Factor Receptor 2, *HER2SC* HER2 Scoring Contest, *NR* not reported, *WSI* whole slide images.

prediction of HER2 status on H&E-stained WSIs, and as such, it was included in this meta-analysis.

Pooled performance of AI and heterogeneity

To assess the clinical relevance of AI in determining eligibility for T-DXd, we evaluated its performance in distinguishing scores 1+/2+/3+ from score 0 with data obtained from 25 contingency tables. When the threshold of positivity was set at score 1+, 2+, or 3+, with score 0 considered negative, the meta-analysis showed a pooled sensitivity of 0.97 [95% CI 0.96–0.98], a pooled specificity of 0.82 [95% CI 0.73–0.88], and an area under the curve (AUC) of 0.98 [95% CI 0.96–0.99] (Fig. 2).

We also examined the performance of AI in distinguishing scores 1+, 2+, and 3+ individually from their counterparts. In this case, the test was considered positive if the respective score was identified by the models and negative if any other score was detected. Figure 3 and Supplementary Fig. 1 present the corresponding summary receiver operating characteristics (SROC) curves and forest plots, respectively. Notably, the performance improved with higher HER2 scores (2+ and 3+). The pooled sensitivity for score 1+ was 0.69 [95% CI 0.57–0.79], with a specificity of 0.94 [95% CI 0.90–0.96] and an AUC of 0.92 [95% CI 0.90–0.94] (Fig. 3a). For score 2+, AI achieved a pooled sensitivity of 0.89 [95% CI 0.84–0.93], a pooled specificity of 0.96 [95% CI 0.93–0.97], and an AUC of 0.98 [95% CI 0.96–0.99] (Fig. 3b). Finally, for score 3+, the AI demonstrated near-perfect performance, with a sensitivity of 0.97 [95% CI 0.96–0.99], specificity of 0.99 [95% CI 0.97–0.99], and an AUC of 1.00 [95% CI 0.99–1.00] (Fig. 3c).

Our results indicate substantial heterogeneity, with Higgins inconsistency index statistic (I^2) values ranging from 94% to 98% for both sensitivity and specificity (Fig. 2a and Supplementary Fig. 1). In all analyses, no significant threshold effect was identified (Supplementary Table 3).

Figure 4 provides a heatmap with a visual representation of the agreement between AI and pathologists across all HER2 scores. The highest agreement was observed at score 3+, with a concordance of 97% [95% CI 96–98%], while AI performed less accurately at score 1+ (88% [95% CI 86–90%]) (Table 2).

Subgroup analysis and meta-regression

To investigate sources of heterogeneity in the 1+/2+/3+ vs. 0 analysis, we conducted meta-regression utilising potential covariates, including: (1) DL (present vs. not present); (2) transfer learning (present vs. not present); (3) commercially available (CA) algorithms (present vs. not present); (4) autonomy (assisted vs. automated); (5) type of internal validation (random split sample vs. *k*-fold cross-validation); (6) external validation (present vs. not present); (7) sample size (≤761 vs. >761, where 761 is the median across studies); (8) data unit (WSIs/cases vs. patches); and (9) dataset (own vs. HER2SC). The performance estimates corresponding to each covariate and the meta-regression results is outlined in Supplementary Table 4. Two studies[13,14] failed to provide information about the type of internal validation and were not included in this subgroup analysis.

Considering sensitivity alone, we identified statistically significant differences across studies using DL, CA algorithms, and external validation. Studies incorporating DL achieved greater sensitivity (0.98 [95% CI 0.97–0.99] vs. 0.94 [95% CI 0.93–0.95]). In contrast, CA algorithms exhibited lower sensitivity (0.93 [95% CI 0.90–0.95]) relative to experimental-only algorithms (0.98 [95% CI 0.97–0.99]). Similarly, studies employing external validation have also performed worse (Sensitivity of 0.96 [95% CI 0.95–0.97] vs. 0.98 [95% CI 0.96–0.99]) (Supplementary Table 4).

Regarding specificity, only the covariates sample size and data unit demonstrated significant differences. Analyses with sample sizes greater than 761 showed improved specificity (0.88 [95% CI 0.81–0.93]) compared to those with sample sizes ≤761 (0.70 [95% CI 0.55–0.82]). When patches were used as the data unit, specificity was significantly higher (0.87 [95% CI 0.79–0.92]) in contrast to 0.70 [95% CI 0.53–0.83] for WSIs/cases (Supplementary Table 4). The covariates transfer learning, autonomy, type of internal validation, and dataset did not show a statistically significant impact on either sensitivity or specificity.

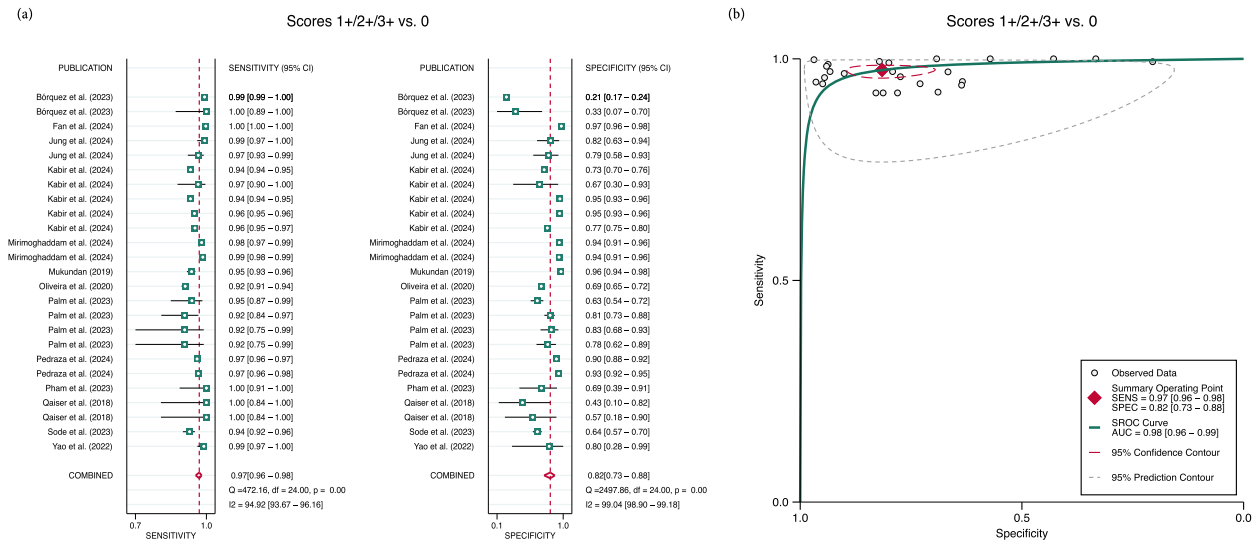

Fig. 2 | Pooled performance of AI in identifying HER2-low individuals from 25 contingency tables. a Forest plot of paired sensitivity and specificity for HER2 scores 1+/2+/3+ vs. 0. **b** SROC curves for HER2 scores 1+/2+3+ vs. 0. The combined performance and summary operating point were calculated using the bivariate random-effects model. Scoring 1+, 2+ or 3+ indicates eligibility for T-DXd therapy. AUC area under the curve, CI confidence interval, SENS sensitivity, SPEC specificity, SROC summary receiver operating characteristics.

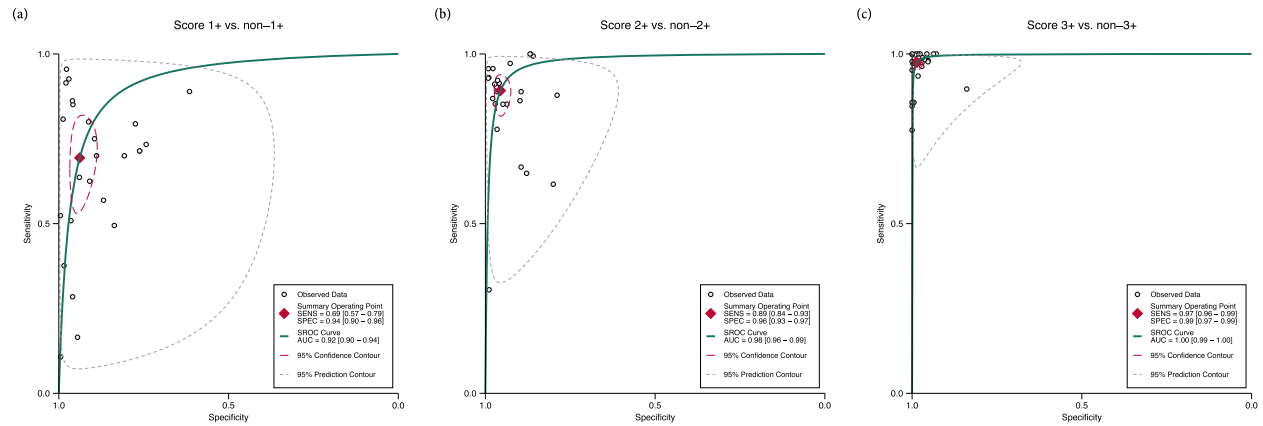

Fig. 3 | Summary receiver operating characteristics curves for individual HER2 scores. a score 1+ vs. non-1+. **b** score 2+ vs. non-2+. **c** score 3+ vs. non-3. Sensitivity and specificity increased with higher HER2 scores, demonstrating near-perfect performance at score 3+. The summary operating points were calculated using the bivariate random-effects model. Heterogeneity is graphically represented by the 95% prediction contour and is more pronounced at lower HER2 scores. AUC area under the curve, SENS sensitivity, SPEC specificity, SROC summary receiver operating characteristics.

Publication bias and sensitivity analysis

To assess publication bias, we performed Deek's funnel plot asymmetry test for all threshold analyses. The studies showed reasonable symmetry around the regression lines, with a non-significant effect, indicating a low likelihood of publication bias (Supplementary Fig. 2). Sensitivity analysis revealed no significant changes in performance estimates for the 1+/2+/3+ vs. 0 meta-analysis (Supplementary Table 5).

Quality assessment

The Quality Assessment Tool for Artificial Intelligence-Centered Diagnostic Test Accuracy Studies (QUADAS-AI) assessment of the included studies is summarised in Supplementary Fig. 3, with a detailed description of the questionnaire outlined in Supplementary Table 2. Briefly, 11 studies demonstrated a high risk of bias in the "patient selection" domain due to unclear reporting of eligibility criteria, such as the modality of tissue extraction (biopsy vs. resection), tumour invasiveness, and failure to specify whether the tumours were primary or metastatic. In the same domain, two studies[13,14] that utilised CA algorithms did not provide a clear rationale neither a detailed breakdown of their training and validation sets. The "index test" domain exhibited high risk of bias in eight studies, as they lacked external validation. Furthermore, eight studies that evaluated AI performance using patches were deemed to have high applicability concerns regarding their index tests, as the use of WSIs/cases instead, would be more representative of real-world practice.

Discussion

In this meta-analysis, we found that AI demonstrated high performance in distinguishing HER2 scores of 1+/2+/3+ from 0, a critical cut-off with significant clinical importance. Considering the 1+, 2+ and 3+ scores individually, the AI models showed increasing sensitivity and specificity as scores rose, with particularly strong performance at higher scores. For score 3+, AI had almost perfect concordance with pathologist's visual scoring. Meta-regression revealed that the use of DL, larger sample sizes, and patches

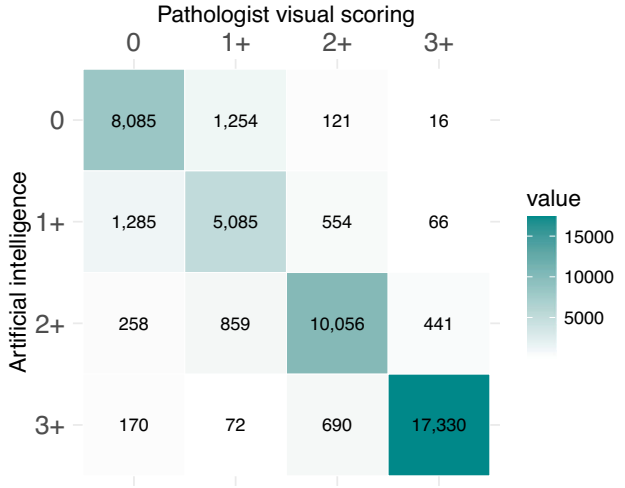

Fig. 4 | Concordance of AI vs. pathologists' visual scoring. The heatmap displays the frequency of accurate and inaccurate AI predictions compared to manual scoring for each HER2 score. Cell values represent the cumulative sample sizes derived from 25 contingency tables corresponding to each prediction. Colour intensity reflects the proportion of correct predictions, with higher intensity indicating greater accuracy. The highest agreement was observed at score 3+, while AI demonstrated lower accuracy at score 1+, with most incorrect predictions skewed towards score 0.

Table 2 | Concordance ratios of correct vs. incorrect predictions across different HER2 scores

HER2 score	Concordance [95% CI]
0	93% [90–94%]
1+	88% [86–90%]
2+	93% [91–94%]
3+	97% [96–98%]
Overall	93% [92–94%]

Pooled concordance ratios were estimated using a random-effects meta-analysis of proportions, applying the DerSimonian and Laird method. *CI* confidence interval, *HER2* Human Epidermal Growth Factor Receptor 2.

positively influenced AI performance, while the use of externally validated and CA models showed less favourable outcomes.

To the best of the authors' knowledge, only one meta-analysis, conducted by Wu et al.[19], has been published on the use of AI for HER2 scoring in breast cancer. However, this review included a limited number of studies, leading to performance metrics with a wide margin of uncertainty. Additionally, sources of heterogeneity were not explored, given the small sample size. These authors assessed AI-assisted HER2 status based on four studies differentiating scores 1+/2+/3+ from 0, finding a pooled sensitivity of 0.93 [95% CI 0.92–0.94], a pooled specificity of 0.80 [95% CI 0.56–0.92], and an AUC of 0.94 [95% CI 0.92–0.96]. It is noteworthy that, for comparison purposes, we swapped the values of sensitivity and specificity, as the original criterion for positivity in their study was set to score 0, whereas in this review the positivity threshold was set to score 1+, 2+, or 3+. Despite the broad 95% CI observed in that review, we found a significantly higher sensitivity (Fig. 2). In contrast, our findings for specificity were similar, and AUCs were borderline comparable (Fig. 2b). The less availability of studies in their analysis likely explains the discrepancy in sensitivity. While they applied similar inclusion criteria and a broader search strategy, yielding four times more studies than this review, five of our included studies were published after their search period concluded. Interestingly, these studies were released after the 2023 ASCO/CAP updated recommendations on HER2-low[6] were made public, indicating a growing demand for more accurate AI models since then.

The classification of scores 0 and 1+ is often challenging for pathologists, leading to significant inter-observer variability. A survey by Fernandez et al.[8], conducted across 1452 laboratories over two years, found that pathologists agreed on the differentiation of scores 0 vs. 1+ in only 70% of cases or fewer. Given the findings of the DB-04 trial, our analysis primarily focused on the distinction between scores 1+/2+/3+ and 0, as patients scoring 1+, 2+ or 3+ are potentially eligible for T-DXd therapy, irrespective of ISH amplification status. Clinically, it is crucial to minimise the number of patients who might miss the opportunity to benefit from T-DXd and those who might be inappropriately treated. The pooled performance metrics found in this work, however, may not be readily intuitive. Put simply, for every 100 patients diagnosed with metastatic breast cancer, AI missed three eligible patients for T-DXd therapy, while eighteen patients would be erroneously treated (Fig. 2), experiencing the drug's adverse effects without a clear therapeutic benefit. This is especially relevant given the potential life-threatening side effects associated with antibody-drug conjugates. For instance, in the T-DXd arm of the DB-04 trial, 45 patients (12.1%) experienced drug-related interstitial lung disease or pneumonitis, leading to the death of three patients (0.8%).

Meta-regression indicated that DL improved significantly the sensitivity in identifying patients eligible for T-DXd (Supplementary Table 4). DL is an AI tool that offers higher accuracy and robustness compared to traditional machine learning (ML) models. It has been widely applied in diagnostic medicine, particularly in histopathology[20], due to its effectiveness in analysing visual data. In this review, all studies that employed DL used CNN, a type of DL model that automatically learns and extracts image features, requiring minimal human intervention. CNN can automatically identify features such as texture, colour intensity, and shape with little need of pathologists' annotations or selection of ROIs[21]. In contrast, traditional ML models are heavily dependent on pathologists' expertise and experience, usually requiring human inputs for maximum effectiveness. The higher sensitivity observed in the DL subgroup is likely explained by its ability to identify image features that may have been overlooked by the human eye. However, DL automation requires large datasets and high-performance computing infra-structure for optimal performance[22], which can be a limitation in practice, as laboratories may not be equipped with such resources.

We also found significantly better specificity in the subgroups with a sample size >761 and those that used patches as the primary unit of analysis. These results are likely interrelated, as both subgroups showed similar disparities in specificity (Supplementary Table 4). Patches are small, localised segments of a larger WSI, designed to facilitate the analysis of WSIs with extremely high resolution[20], therefore studies that reported performance metrics using patches are expected to have larger sample sizes. In HER2 scoring, models typically analyse each patch separately and then aggregate the results into a patient-level overall score, taking into account spatial distribution and the relative weight of scores across individual patches. Although patches can be utilised at some point during AI training, the overall WSI score is the most applicable in practice. However, studies that reported performance outcomes using patches lacked an overall aggregated score. From a practical perspective, the higher specificity observed in the patches subgroup suggests that the real-world performance of these models may be overestimated, as the ability to reliably aggregate individual patch scores into a comprehensive patient-level score was not fully evaluated.

Studies that underwent external validation and utilised CA models showed poorer sensitivity. Since CA algorithms rely on pre-designed and pre-trained models, their testing datasets are inherently external to the training data, leading to similar outcomes across both subgroup analyses (Supplementary Table 4). Indeed, to ensure that models generalise beyond the training set, it is necessary to use samples from an external dataset. This accounts for variations in WSIs across different settings, such as distinct staining protocols, scanner resolutions, and imaging

artefacts. As a result, algorithms tested on the same training dataset were expected to perform significantly better, but not necessarily would be a viable option in practice.

Despite the poorer performance observed in externally validated and CA algorithms, regulatory agencies typically do not require a predefined minimum performance threshold for commercialisation approval, with quality assessments conducted on a case-by-case basis. The assessment of models' reproducibility is a crucial aspect of the regulatory process, as it ensures that their performance generalises beyond the dataset they were trained on. In the US and Europe, for such technologies to be approved for commercial use, it must be demonstrated that their benefits outweigh the risks and that they are effective and safe for their intended purpose[23]. There is no specific regulatory framework for the use of AI in medical applications; such technologies are generally categorised and regulated as medical devices. However, initiatives have been undertaken to address those reproducibility issues. A collaboration between the US Food and Drug Administration (FDA), Health Canada, and the UK Medicines and Healthcare Products Regulatory Agency (MHRA) established guiding principles for Good Machine Learning Practice[24], highlighting the importance of external validation to ensure model reproducibility across different settings. Additionally, an FDA action plan[25] proposed methodologies to mitigate the limitations of biased datasets and encouraged the use of real-world training data to enhance algorithm robustness. The European Union Medical Devices Regulation 2017/745[26] requires software validation and testing in both in-house and simulated or actual user environments prior to final release (Annex II, Section 6.1b). While our subgroup analysis revealed that externally validated CA algorithms may exhibit reduced performance, their certification by regulatory bodies is justified by their superior generalisability and reliability in real-world settings beyond their development environment.

The individual analysis of scores 1+, 2+, and 3+ vs. their counterparts showed that sensitivity, specificity and concordance improved as scores increased (Figs. 3 and 4; Supplementary Fig. 1; Table 2). This trend aligns with the similar meta-analysis by Wu et al.[19]. Unlike the analysis of T-DXd eligibility, where patient management and classification in HER2-low were clinician-driven, AI's performance in classifying individual HER2 scores is more relevant in a histopathology context, as the classic four-class HER2 recommendations remain in effect[6]. The lower sensitivity in score 1+ may reflect AI's limitations in detecting subtle immunohistochemical changes, particularly at the 10% threshold of faint membranes between scores 0 and 1+ as per ASCO/CAP[2,27]. This may suggest that either the detection or counting of faint/barely perceptible membranes is downregulated, as most incorrect predictions were biased towards score 0 (Fig. 4). Conversely, the high specificity for score 2+ indicates that only a small number of false positives (FP)s would undergo unnecessary ISH testing, thereby reducing turnaround time for clinical decision-making.

The intra- and inter-observer concordance rates for IHC manual HER2 scoring vary significantly across the literature. Concordance studies differ in methodology, as pathologists' performance can be assessed either through agreement with peers or with more advanced lab assays. Additionally, factors such as the number of raters, wash-out period, variations in IHC assays, and the accuracy of the reference standard test used as the ground truth introduce further heterogeneity to these studies. A multi-institutional cohort study conducted with 170 breast cancer biopsies across 15 institutions reported an overall four-class (0, 1+, 2+, and 3+) concordance among pathologists as low as 28.82% [95% CI 22.01–35.63%][28]. In contrast, another study found a markedly better figure of 94.2% [95% CI 91.4–96.5%][29]. Studies comparing pathologists' performance in IHC vs. other techniques reported concordance rates of 93.3%[30], 91.7%[31], 82.9%[32] with RT-qPCR, and 82%[33] with ISH. Based on these metrics, the overall concordance rate of 93% [95% CI 92–94%] (Table 2) identified in this review implies that AI may hold the potential to achieve performance comparable to pathologists' visual scoring.

However, the findings of this work have limitations. The exclusion of studies that used grouped 0/1+ scores and those that did not report sufficient performance outcomes may have influenced our pooled metrics. This is likely because, in earlier studies, separating 0/1+ scores was not relevant for T-DXd eligibility at that time, and DL technology had not yet become as widespread and advanced as it currently is. The notably high heterogeneity found in our analyses suggests that AI may lack standardisation, due to inherent differences in models design as well as pre-analytical factors, such as IHC assay, WSI scanner resolution, tumour heterogeneity, and tissue artefacts. The design of a standardised testing protocol is likely to make the various AI models more comparable to each other. Moreover, the high risk of bias and applicability concerns identified in QUADAS-AI (Supplementary Fig. 3) represent an additional limitation. Since some included studies failed to report important sample characteristics, such as tumour invasiveness, method of extraction (biopsy vs. resection), or whether tumours were metastatic or primary, we cannot extrapolate our findings to samples that differ in these aspects. For instance, it cannot be ignored that algorithms may perform differently if non-breast cells are present in metastatic samples or if slides vary in size depending on the method of extraction. It remains essential that upcoming validation studies in AI provide a more detailed description of the samples, taking pre-analytical clinical and factors into account.

In light of the current ASCO/CAP recommendations, the pooled performance findings in this meta-analysis apply to the HER2-low population, though ongoing research may prompt changes in the HER2 classification system. Despite the FDA[34] and the European Medicines Agency[35] approvals of T-DXd for metastatic breast cancer following the DB-04 trial, it is widely agreed among ASCO/CAP and European Society of Medical Oncology (ESMO) that HER2-low does not represent a biologically distinct entity, but rather a heterogeneous group of tumours influenced by HER2 expression[6,36]. We acknowledge that the 2023 ASCO/CAP update reaffirms the 2018 version of the HER2 scoring guideline, with no prognostic value attributed to HER2-low and no changes recommended in reporting HER2 categories. The term HER2-low was used as an eligibility criterion in the DB-04 trial and should guide clinical decision-making only. To date, HER2 classification remains as negative (0 and 1+), equivocal (2+), and positive (3+). This is primarily because the DB-04 trial did not include patients with score 0, and the possibility that some patients in the 0-1+ threshold spectrum might benefit from T-DXd should not be ignored. Addressing this uncertainty, the recently published phase 3 DESTINY-Breast06[37] trial found a survival benefit in HER2-ultralow patients, i.e., a subset of patients within the IHC category 0 who express faint membrane staining in ≤10% of tumour cells. Accordingly, HER2-ultralow hormone receptor-positive patients who had undergone one or more lines of endocrine-based therapy had longer progression-free survival when treated with T-DXd compared to those receiving chemotherapy. These findings align with the interim outcomes of the ongoing phase 2 DAISY trial[38]. The grey zone within the 0-1+ threshold has several implications for this review's outcomes. All AI algorithms included in this analysis used the conventional four-class HER2 scoring system (0, 1+, 2+, 3+), and it is clear that this classification may need to be revised. Subsequent AI algorithms will need to be trained using inputs that account for staining intensity and the proportion of stained cells within the 0–1+ spectrum.

In conclusion, the results of this meta-analysis, encompassing 25 contingency tables across thirteen studies, suggest that AI achieved promising outcomes in accurately identifying HER2-low individuals eligible for T-DXd therapy. Our findings indicate that, to date, AI achieves substantial accuracy, particularly in distinguishing higher HER2 scores. Future validation studies using AI should focus on improving accuracy in the 0–1+ range and provide more detailed reporting of clinical and pre-analytical data. Lastly, this review highlights that the progressive development of DL is steering towards greater automation, and pathologists will need to adapt to effectively integrate this tool into their practice.

Methods

Protocol registration and study design

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis of Diagnostic Test Accuracy (PRISMA-DTA) guidelines (Supplementary Table 1) and was registered in the International Prospective Register of Systematic Reviews (PROSPERO) under the identification number CRD42024540664. This research utilised publicly accessible data, thus ethical approval and informed consent were not applicable.

Eligibility criteria

Inclusion in this meta-analysis was limited to studies that fulfilled all of the following criteria: (1) studies that utilised primary or metastatic breast cancer tissues in the form of digitalised WSIs, from female patients at any age, irrespective of tumour stage, histological subtype and hormone receptor status, employing conventional 3,3'-diaminobenzidine (DAB) staining; (2) studies evaluating the performance of AI algorithms as an index test; (3) studies comparing these algorithms to pathologists' visual scoring as the reference standard; and (4) studies that reported AI performance metrics. Only original research articles published in English in peer-reviewed journals were included. No limitations were imposed on the publication date.

Studies were excluded if they lacked sufficient data to build a 2×2 contingency table of true positives (TP), FP, true negatives (TN), and false negatives (FN) across each HER2 category (0, 1+, 2+, and 3+). Additionally, studies that combined 0/1+ scores were excluded, given the clinical significance of this distinction for the purposes of this review. Studies conducted using animal tissues, those employing H&E staining, multi-omics approaches, as well as reviews, letters, preprints, and conference abstracts, were excluded.

Literature search

A comprehensive literature search was conducted in MEDLINE, EMBASE, Scopus, and Web of Science from inception up to 3 May 2024 using the following terms: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural networks") AND "breast cancer" AND "HER2". We also examined the reference lists of studies to identify relevant articles. Two independent authors screened titles, abstracts, and full texts for eligibility. Disagreements were resolved through consensus with a senior pathologist.

Data extraction

The following data were collected using a pre-designed spreadsheet: study characteristics (author, year, and country), participant details (tumour features, sample size, data unit, and dataset), index test (algorithm specifications, deep learning, transfer learning, autonomy, external validation, type of internal validation, and the use of CA algorithms), reference standard (manual scoring and classification system), and 2×2 contingency tables for TP, FP, FN, and TN for each 0, 1+, 2+, and 3+ score. Supplementary data were requested from corresponding authors when necessary. Performance metrics were calculated using three distinct data units: cases, WSIs, and patches, which varied across the included studies. A second independent reviewer cross-checked all the collected data. If studies provided multiple contingency tables representing various outcomes for a single AI algorithm, only the data reflecting the best performance were collected. In cases where studies presented results evaluating different types of automation (assisted vs. automated) or different data units for the same algorithm, these tables were treated as independent.

Outcomes

The primary outcome was the AI pooled sensitivity, specificity, AUC, and concordance in distinguishing HER2 scores of 1+/2+/3+ from 0. The cut-off for positivity was defined as a score of at least 1+, since the decision to initiate T-DXd treatment is based on this threshold. The secondary outcome was to estimate the same pooled performance metrics for each individual score of 1+, 2+, and 3+ compared to their counterparts.

Each study contributed with at least four contingency tables, corresponding to the performance of each individual score. Importantly, the term "positivity," as used here, applies solely to defining the cut-off for the meta-analysis and is not related to the histological classification of HER2.

Statistical analysis

Meta-analysis was conducted using the bivariate random-effects model to calculate the pooled sensitivity and specificity, as significant heterogeneity was expected. SROC curves with 95% CI and 95% prediction region (PR) were employed for visual comparisons and AUC estimation. Heterogeneity was evaluated using the I^2, with 50% defined as moderate and ≥75% defined as high. Pooled concordance was calculated using a random effects meta-analysis of proportions, following the DerSimonian and Laird method. Potential sources of heterogeneity were explored through subgroup analysis and meta-regression, utilising covariates that were most likely to contribute to heterogeneity. A "leave-one-out" sensitivity analysis was performed by sequentially excluding one study at a time. Publication bias was evaluated using Deek's funnel plot asymmetry test. Statistical analysis was conducted using STATA (version 17.0, StataCorp LLC, Texas, USA) with `midas`, `metandi` and `metaprop` modules. To assess the threshold effect, the Spearman correlation coefficient was calculated using Meta-DiSc (version 1.4, Ramón y Cajal Hospital, Madrid, Spain). A $p < 0.05$ was considered significant for the meta-analysis, meta-regression, and threshold effect evaluation. For the assessment of publication bias, a significance cut-off of $p < 0.10$ was applied.

Quality assessment

The risk of bias and applicability concerns in the included articles were meticulously evaluated independently by two authors using the adapted QUADAS-AI[39] across four domains: patient selection, index test, reference standard, and flow and timing. The detailed questionnaire is provided in Supplementary Table 2.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4o version 1.2024.227 in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

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

Received: 7 November 2024; Accepted: 27 January 2025;
Published online: 06 March 2025

References

1. Arnold, M. et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. *Breast* **66**, 15 (2022).
2. Wolff, A. C. 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.* **36**, 2105–2122 (2018).
3. Schlam, I. & Swain, S. M. Her2-positive breast cancer and tyrosine kinase inhibitors: the time is now. *npj Breast Cancer* **7**, 1–12 (2021).
4. Bradley, R. et al. rastuzumab for early-stage, her2-positive breast cancer: a meta-analysis of 13864 women in seven randomised trials. *Lancet Oncol.* **22**, 1139–1150 (2021).
5. Modi, S. et al. Trastuzumab deruxtecan in previously treated her2-low advanced breast cancer. *N. Engl. J. Med.* **387**, 9–20 (2022).

6. Wolff, A. C. et al. Human epidermal growth factor receptor 2 testing in breast cancer: ASCO-College of American Pathologists guideline update. *J. Clin. Oncol.* **41**, 3867–3872 (2023).
7. DiPeri, T. P. et al. Discordance of her2 expression and/or amplification on repeat testing. *Mol. Cancer Ther.* **22**, 976–984 (2023).
8. Fernandez, A. I. et al. Examination of low erbb2 protein expression in breast cancer tissue. *JAMA Oncol.* **8**, 1–4 (2022).
9. McGenity, C. et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *npj Digit. Med.* **7**, 1–19 (2024).
10. Chan, R. C. et al. Artificial intelligence in breast cancer histopathology. *Histopathology* **82**, 198–210 (2023).
11. Marini, N. et al. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *npj Digit. Med.* **5**, 1–18 (2022).
12. Oliveira, S. P. et al. Weakly-supervised classification of her2 expression in breast cancer haematoxylin and eosin stained slides. *Appl. Sci.* **10**, 4728 (2020).
13. Palm, C. et al. Determining her2 status by artificial intelligence: An investigation of primary, metastatic, and her2 low breast tumors. *Diagnostics* **13**, 168 (2023).
14. Sode, M., Thagaard, J., Eriksen, J. O. & Laenkholm, A.-V. Digital image analysis and assisted reading of the her2 score display reduced concordance: pitfalls in the categorisation of her2-low breast cancer. *Histopathology* **82**, 912–924 (2023).
15. Fan, L., Liu, J., Ju, B., Lou, D. & Tian, Y. A deep learning-based holistic diagnosis system for immunohistochemistry interpretation and molecular subtyping. *Neoplasia* **50**, 100976 (2024).
16. Qaiser, T. et al. Her2 challenge contest: a detailed assessment of automated her2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology* **72**, 227–238 (2018).
17. Pham, M. D. et al. Interpretable her2 scoring by evaluating clinical guidelines through a weakly supervised, constrained deep learning approach. *Comput. Med. Imaging Graph.* **108**, 102261 (2023).
18. Jung, M. et al. Augmented interpretation of her2, er, and pr in breast cancer by artificial intelligence analyzer: enhancing interobserver agreement through a reader study of 201 cases. *Breast Cancer Res.* **26**, 31 (2024).
19. Wu, S. et al. Artificial intelligence for assisted her2 immunohistochemistry evaluation of breast cancer: a systematic review and meta-analysis. *Pathol. Res. Pract.* **260**, 155472 (2024).
20. Yousif, M. et al. Artificial intelligence applied to breast pathology. *Virchows Arch.* **480**, 191–209 (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. Ahmed, A. A., Abouzid, M. & Kaczmarek, E. Deep learning approaches in histopathology. *Cancers* **14**, 5264 (2022).
23. Pantanowitz, L. et al. Regulatory aspects of artificial intelligence and machine learning. *Mod. Pathol.* **37**, 100609 (2024).
24. Good machine learning practice for medical device development: Guiding principles. Food and Drug Administration. Center for Devices and Radiological Health. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (Accessed: 28 December 2024) (2021).
25. Artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd) action plan. *U.S Food and Drug Administration. Center for Devices and Radiological Health*. Available at: https://www.fda.gov/media/145022/download (Accessed: 28 December 2024) (2021).
26. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. EUR-Lex: European Union Law Available at: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng (Accessed: 28 December 2024) (2024).
27. Wolff, A. C. et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin. Oncol.* **31**, 3997–4013 (2013).
28. Robbins, C. J. et al. Multi-institutional assessment of pathologist scoring HER2 immunohistochemistry. *Mod. Pathol.* **36**, 100032 (2023).
29. Garrido, C. et al. Analytical and clinical validation of pathway anti-her-2/neu (4b5) antibody to assess her2-low status for trastuzumab deruxtecan treatment in breast cancer. *Virchows Arch.* **484**, 1005–1014 (2024).
30. Wu, N. C. et al. Comparison of central laboratory assessments of er, pr, her2, and ki67 by ihc/fish and the corresponding mRNAs (esr1, pgr, erbb2, and mki67) by rt-qpcr on an automated, broadly deployed diagnostic platform. *Breast Cancer Res. Treat.* **172**, 327–338 (2018).
31. Chen, L. et al. Comparison of immunohistochemistry and rt-qpcr for assessing er, pr, her2, and ki67 and evaluating subtypes in patients with breast cancer. *Breast Cancer Res. Treat.* **194**, 517–529 (2022).
32. Oma, D. et al. Immunohistochemistry versus pcr technology for molecular subtyping of breast cancer: multicentered experiences from Addis Ababa, Ethiopia. *J. Cancer Prev.* **28**, 64–74 (2023).
33. Sui, W. et al. Comparison of immunohistochemistry (ihc) and fluorescence in situ hybridization (fish) assessment for her-2 status in breast cancer. *World J. Surg. Oncol.* **7**, 83 (2009).
34. FDA approves fam-trastuzumab deruxtecan-nxki for her2-low breast cancer. *U.S Food and Drug Administration* Available at: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-fam-trastuzumab-deruxtecan-nxki-her2-low-breast-cancer (Accessed: 28 December 2024). (2022).
35. Trastuzumab-deruxtecan (enhertu). *European Medicines Agency (EMA)* Available at: https://www.ema.europa.eu/en/medicines/human/EPAR/enhertu (Accessed: 28 December 2024) (2024).
36. 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).
37. Bardia, A. et al. Trastuzumab deruxtecan after endocrine therapy in metastatic breast cancer. *N. Engl. J. Med.* **391**, 2110–2122 (2024).
38. Mosele, F. et al. Trastuzumab deruxtecan in metastatic breast cancer with variable her2 expression: the phase 2 daisy trial. *Nat. Med.* **29**, 2110–2120 (2023).
39. Sounderajah, V. et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: quadas-AI. *Nat. Med.* **27**, 1663–1665 (2021).
40. Borquez, S., Pezoa, R., Salinas, L. & Torres, C. E. Uncertainty estimation in the classification of histopathological images with her2 overexpression using Monte Carlo dropout. *Biomed. Signal Process. Control* **85**, 104864 (2023).
41. Kabir, S. et al. The utility of a deep learning-based approach in her-2/neu assessment in breast cancer. *Expert Syst. Appl.* **238**, 122051 (2024).
42. Mirimoghaddam, M. M. et al. Her2gan: Overcome the scarcity of her2 breast cancer dataset based on transfer learning and gan model. *Clin. Breast Cancer* **24**, 53–64 (2024).
43. Mukundan, R. Analysis of image feature characteristics for automated scoring of her2 in histology slides. *J. Imaging* **5**, 35 (2019).
44. Pedraza, A., Gonzalez, L., Deniz, O. & Bueno, G. Deep neural networks for her2 grading of whole slide images with subclasses levels. *Algorithms* **17**, 97 (2024).
45. Yao, Q. et al. Using whole slide gray value map to predict her2 expression and fish status in breast cancer. *Cancers* **14**, 6233 (2022).

Author contributions

D.A.N.A. developed the study concept and design, conducted the electronic searches, performed data extraction, and wrote the first draft of this manuscript. M.T.V. contributed to data analysis and interpretation, statistical review, and critical revision of the manuscript. A.V. assisted with electronic

searches, data extraction, and risk of bias assessment. L.A.F.S. and E.H.C.N.F. contributed to the conceptualisation, critical review, and supervision. All authors have approved the final version of this manuscript.

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/s41746-025-01483-8.

Correspondence and requests for materials should be addressed to Matheus Trotta Vianna.