



QuPath-assisted digital scoring of ER, PR, HER2 and Ki-67 compared with manual assessment for surrogate molecular subtyping of invasive breast carcinoma: a diagnostic agreement study from Salah Al-Din, Iraq

Sara Faez Hassan^{1,*}, Shahad Mohammed Mahmood¹, Raghad Hazem Alabbasy²

¹ University of Samarra. College of Applied Sciences. Department of Pathological Analyses, Samarra, Iraq.

² University of Samarra. College of Education for Pure Sciences. Department of Biology, Samarra, Iraq.

*Corresponding author: Sara Faez Hassan.

University of Samarra. College of Applied Sciences.

Department of Pathological Analyses, Samarra, Iraq.

E-mail: sara.f@uosamarra.edu.iq

DOI: <https://doi.org/10.54448/mdnt26402>

Received: 07-11-2026; Revised: 09-19-2026; Accepted: 10-01-2026; Published: 10-02-2026; MedNEXT-id: e26402

Editor: Dr. Idiberto José Zotarelli-Filho, MSc, Ph.D., Post-Doctoral.

Abstract

Introduction: Manual scoring of breast cancer biomarkers is subject to observer variability, and open-source digital image analysis has not been evaluated in Iraqi practice. **Objective:** This study aimed to compare QuPath-assisted digital scoring with manual assessment of ER, PR, HER2 and Ki-67 in invasive breast carcinomas from Salah Al-Din, Iraq, both for the individual measurements and for the surrogate molecular subtypes derived from them. Secondary aims were to characterise the cases in which the two methods disagreed, to examine how QuPath-measured HER2 membrane positivity related to the manual HER2 score, and to explore the associations of the digital readouts with standard histopathological parameters and survival. **Methods:** ER, PR, HER2 and Ki-67 were scored manually and with QuPath in 120 invasive breast carcinomas from Salah Al-Din, Iraq. Agreement was assessed with intraclass correlation coefficients (ICC), Lin's concordance coefficient, Bland-Altman analysis and κ statistics, and surrogate subtypes derived by both methods were compared. **Results:** Agreement was excellent for all continuous measurements (ICC 0.939–0.994). Categorical agreement was 99.2% for ER status, 97.5% for PR status, 97.5% for Ki-67 $\geq 20\%$ and 92.5% for HER2 score (weighted κ 0.940). Subtype agreement was 114/120 (95.0%; κ 0.935). All six discordant cases had ER/PR expression of 1–10%, Ki-67 at the 20% cut-off, or HER2 upgraded from 2+ to 3+ despite non-amplified in situ hybridisation. Triple-negative status was associated with shorter disease-

free survival (QuPath hazard ratio 3.65, 95% CI 1.48–9.00), although events were few. **Conclusion:** QuPath scoring agreed closely with manual assessment in this cohort. Cases near clinical cut-offs and HER2 2+ tumours still require pathologist review and in situ hybridisation. Multicentre validation is needed.

Keywords: Breast neoplasms. Digital pathology. Image analysis. Immunohistochemistry. Ki-67 antigen. Receptor. ErbB-2. QuPath.

Introduction

Breast cancer is the most frequently diagnosed cancer in women worldwide and a leading cause of cancer death, with approximately 2.3 million new cases estimated in 2022 [1]. In Iraq it is the most common cause of cancer-related death among women [2]. Iraqi series from several governorates, including southern Iraq and Erbil, have described the distribution of immunohistochemical (IHC) subtypes [2-5], and all of them depend on conventional manual scoring of the same four biomarkers: oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and Ki-67.

These four markers remain the practical basis of treatment planning in early breast cancer. Combined in the St Gallen surrogate definitions, they separate luminal A-like, luminal B-like, HER2-positive and triple-negative disease, and each category carries different indications for endocrine, anti-HER2 and cytotoxic therapy [6,7]. The interpretive rules are set out in the ASCO/CAP guidelines for ER/PR and HER2 testing and in

the recommendations of the International Ki67 in Breast Cancer Working Group (IKWG) [8–11]. The 2020 ASCO/CAP update added an “ER-low positive” category for tumors with 1–10% ER-positive nuclei, because such tumours often behave more like ER-negative disease [8].

Manual scoring is semi-quantitative, and its weak points lie close to the thresholds used for treatment decisions. Ki-67 is the clearest example: the IKWG considers values between about 5% and 30% unreliable for individual decisions unless the laboratory has demonstrated analytical validity [11], and scores depend on the counting method and the region selected [12]. HER2 has acquired a similar problem. After trastuzumab deruxtecan improved survival in HER2-low metastatic breast cancer (IHC 1+ or 2+ without amplification) [13], the distinction between HER2 0 and 1+ became therapeutically relevant [14,15], yet agreement among pathologists at this boundary is poor [16].

Digital image analysis of scanned slides offers a way to make these measurements more reproducible [17–19]. QuPath is an open-source platform for whole-slide image analysis that supports cell detection, stain separation and marker quantification without licence fees [20,21], which makes it relevant to laboratories with limited budgets. Studies from Europe and China have shown good agreement between digital and manual Ki-67 scoring [22,23] and better reproducibility of HER2 0/1+ calls when pathologists are assisted by image analysis [24].

Two gaps remain. First, most validation work comes from high-income centers, often with commercial software, and we found no published evaluation of digital biomarker scoring from Iraq; Iraqi studies to date report subtype frequencies and clinicopathological associations based on manual assessment alone [2–5,25]. Second, most digital studies examine one marker at a time. In routine practice, however, the clinically relevant output is the surrogate subtype, which combines four measurements, each with its own threshold, and small disagreements in any one of them can change the final category.

The aim of this study was therefore to compare QuPath-assisted digital scoring with manual assessment of ER, PR, HER2 and Ki-67 in invasive breast carcinomas from Salah Al-Din, Iraq, both for the individual measurements and for the surrogate molecular subtypes derived from them. Secondary aims were to characterise the cases in which the two methods disagreed, to examine how QuPath-measured HER2 membrane positivity related to the manual HER2 score, and to explore the associations of the digital readouts with standard histopathological parameters and survival.

Materials and Methods

Study design, setting and reporting

This was an analytical cross-sectional diagnostic agreement study with retrospective collection of follow-up data. Cases were collected from the histopathology services of public hospitals, Salah Al-Din Governorate, Iraq, and from private histopathology clinics in the same city between January 2025 and December 2025. The index method was QuPath-assisted digital scoring and the reference method was manual scoring by a pathologist. Reporting follows the STARD 2015 guideline where applicable [26]. The overall workflow is summarised in Figure 1.

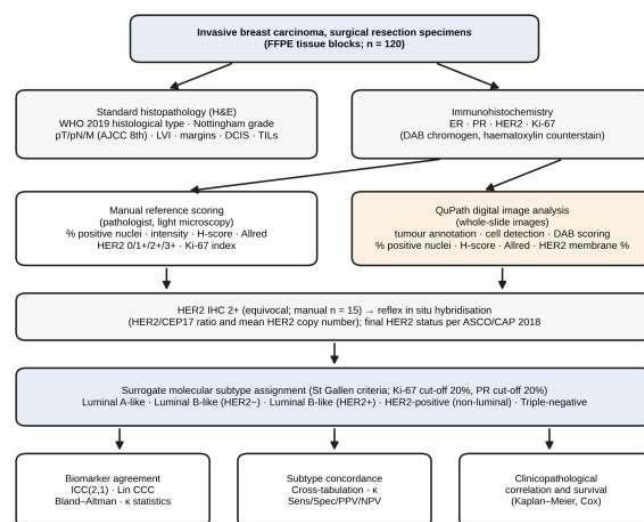


Figure 1. Study design and analytical workflow of the QuPath-assisted diagnostic and subtyping framework. Formalin-fixed paraffin-embedded resection specimens of invasive breast carcinoma (n = 120) underwent standard haematoxylin and eosin (H&E) histopathological evaluation and immunohistochemistry for ER, PR, HER2 and Ki-67. Each immunostain was scored in parallel by manual light microscopy (reference standard) and by QuPath digital image analysis. HER2 IHC 2+ (equivocal) cases underwent reflex in situ hybridisation. Surrogate molecular subtypes were assigned independently from the manual and the digital readouts and compared. Abbreviations: CCC, concordance correlation coefficient; DCIS, ductal carcinoma in situ; FFPE, formalin-fixed paraffin-embedded; ICC, intraclass correlation coefficient; LVI, lymphovascular invasion; TILs, tumour-infiltrating lymphocytes. Source: Own authorship.

Patients and specimens

Women with a histologically confirmed primary invasive breast carcinoma treated by surgical excision with axillary lymph node surgery were eligible if formalin-fixed paraffin-embedded (FFPE) tumour blocks, ER, PR, HER2 and Ki-67 immunostains, complete pathology reports and follow-up information of at least 12 months were available. Exclusion criteria: neoadjuvant systemic therapy, recurrent or bilateral

carcinoma, insufficient invasive tumour in the available block, and slides with extensive artefacts or staining failure. A total of 120 cases were analysed. The cohort was assembled to include 30 cases from each of four receptor-defined groups (luminal A-like, luminal B-like, HER2-positive non-luminal and triple-negative), selected according to the manual reference classification the sampling procedure.

Clinical and laboratory data

Age at diagnosis, menopausal status, residence (urban or rural) and body mass index (weight in kilograms divided by the square of height in meters) were taken from the medical records. Preoperative blood samples were used to record the absolute neutrophil count (ANC) and absolute lymphocyte count (ALC) from the complete blood count, from which the neutrophil-to-lymphocyte ratio (NLR) was calculated as ANC/ALC. Serum cancer antigen 15-3 (CA 15-3) and carcinoembryonic antigen (CEA) were measured by immunoassay, and alanine aminotransferase (ALT), aspartate aminotransferase (AST) and creatinine by routine clinical chemistry. The complete blood count was performed on a Sysmex XN-series automated haematology analyser (Sysmex Corporation, Kobe, Japan). CA 15-3 and CEA were measured by electrochemiluminescence immunoassay on a Cobas e411 analyser (Roche Diagnostics, Mannheim, Germany), and ALT, AST and creatinine were measured on a Cobas c311 chemistry analyser (Roche Diagnostics), all according to the manufacturers' instructions.

Histopathological assessment

Surgical specimens were fixed in 10% neutral-buffered formalin, with fixation times within the 6–72 h range recommended by ASCO/CAP [8], processed routinely and embedded in paraffin. Sections of 4 µm were stained with haematoxylin and eosin. Histological type was assigned according to the fifth edition of the WHO classification of breast tumours [27]. Histological grade was assessed with the Nottingham system, in which tubule formation, nuclear pleomorphism and mitotic count are each scored from 1 to 3 and summed (3–9) to give grade 1 (3–5), grade 2 (6–7) or grade 3 (8–9) [28]. Tumour size was the largest dimension of the invasive component in centimetres. Pathological T and N categories and M status were recorded according to the eighth edition of the AJCC Cancer Staging Manual [29], together with the numbers of lymph nodes examined and involved. Lymphovascular invasion, associated ductal carcinoma in situ (DCIS) and the surgical margin were recorded; margins were classed as positive (ink on tumour), close (<1 mm) or negative,

and the distance to the nearest margin was measured in millimetres. Stromal tumour-infiltrating lymphocytes (TILs) were scored on haematoxylin and eosin sections as the percentage of stromal area within the tumour borders occupied by mononuclear inflammatory cells, following the International TILs Working Group recommendations [30].

Immunohistochemistry and in situ hybridisation

Four-micrometre sections were cut onto charged slides and stained for ER, PR, HER2 and Ki-67 using the following primary antibodies (Agilent/Dako, Glostrup, Denmark): oestrogen receptor α clone 1D5 (catalogue M7047), progesterone receptor clone PgR 636 (M3569), HER2 rabbit polyclonal A0485 (dilution 1:600–1:800) and Ki-67 clone MIB-1 (M7240, dilution 1:75–1:150) on a manual EnVision+ horseradish-peroxidase polymer protocol (anti-mouse K4001 or anti-rabbit K4003, Agilent/Dako), with heat-induced epitope retrieval in a PT Link water bath (Agilent/Dako) for 20 min at 97 °C using high-pH buffer (K8004) for ER and PR and low-pH citrate buffer (K8005, pH 6.1) for Ki-67 and HER2. Diaminobenzidine (DAB) was used as the chromogen with haematoxylin counterstain. Each run included external positive and negative tissue controls; normal ducts and lobules served as internal controls for ER and PR. All tumours scored HER2 2+ (equivocal) by manual assessment underwent reflex dual-colour in situ hybridisation (ISH), and the HER2/CEP17 ratio and mean HER2 copy number per cell were interpreted according to the ASCO/CAP 2018 guideline [9].

Manual scoring (reference method)

Manual scoring was performed by two consultant pathologists with more than ten years of experience in breast pathology, using a conventional light microscope and blinded to the digital results and clinical outcome. For ER and PR, the percentage of positive invasive tumour nuclei (0–100%) and the predominant staining intensity (0, none; 1, weak; 2, moderate; 3, strong) were recorded. The H-score was calculated as (1 × % weakly stained) + (2 × % moderately stained) + (3 × % strongly stained) nuclei, giving a range of 0–300 [31], and the Allred score as the sum of the proportion score (0–5) and intensity score (0–3), giving 0–8 [32]. Tumours with ≥1% positive nuclei were considered positive, and ER expression in 1–10% of nuclei was recorded as ER-low positive [8]. HER2 was scored 0, 1+, 2+ or 3+ according to ASCO/CAP criteria [9,10]; HER2 1+ and 2+ without amplification were considered HER2-low [15]. Ki-67 was recorded as the percentage of positive invasive tumour nuclei following IKWG recommendations [11]. Tumour cellularity was estimated as the percentage of invasive tumour cells

within the scored area.

Digital image acquisition and QuPath analysis (index method)

Immunostained slides were digitised with an Aperio GT 450 whole-slide scanner (Leica Biosystems, Wetzlar, Germany) with a $\times 40$ objective, giving a resolution of approximately $0.25 \mu\text{m}/\text{pixel}$. Images were analysed in QuPath version 0.7.0 [20]. For each image, stain vectors for haematoxylin and DAB were estimated from representative areas, and the invasive tumour region was annotated by the pathologist, excluding DCIS, normal breast tissue, necrosis, crush artefact and edge staining. Nuclei were detected with the positive cell detection command on the haematoxylin channel, and an object classifier trained on annotated examples was used to separate tumour cells from stromal and inflammatory cells. For ER, PR and Ki-67, the mean nuclear DAB optical density of each tumour cell was compared with three intensity thresholds (0.15, 0.35 and 0.55 optical density units for the 1+/2+/3+ boundaries, calibrated on a 20-case training subset) to classify cells as negative, weak (1+), moderate (2+) or strong (3+). The percentage of positive tumour nuclei, the digital H-score and the digital Allred score were calculated from these counts using the same formulas as for manual scoring. For HER2, cell boundaries were approximated by expanding each detected nucleus by $2.5 \mu\text{m}$, and membrane DAB staining was used to calculate the percentage of HER2 membrane-positive tumour cells; these results were converted to a 0/1+/2+/3+ score using thresholds calibrated against the manual scores on the same training subset. Digital tumour cellularity was the proportion of detected cells classified as tumour cells. The operator was blinded to the manual scores. Representative QuPath outputs are shown in Figures 2 and 3.

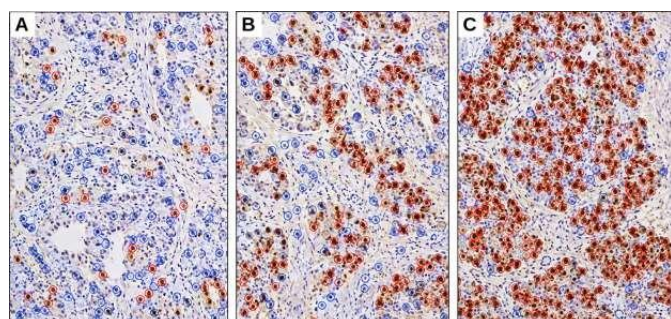


Figure 2. Representative QuPath-assisted quantification of the Ki-67 proliferation index in invasive breast carcinoma of no special type. (A) Low, (B) intermediate and (C) high Ki-67 proliferative activity. Red circles denote tumour nuclei classified as Ki-67-positive and blue circles denote Ki-67-negative tumour nuclei; stromal, endothelial and inflammatory cells were excluded from the tumour-cell denominator. The Ki-67 index was calculated as the

percentage of positive tumour nuclei among all annotated tumour nuclei. Immunoperoxidase staining (DAB chromogen) with haematoxylin counterstain. Original magnification $\times 40$; scale bar = $50 \mu\text{m}$. Source: Own authorship.

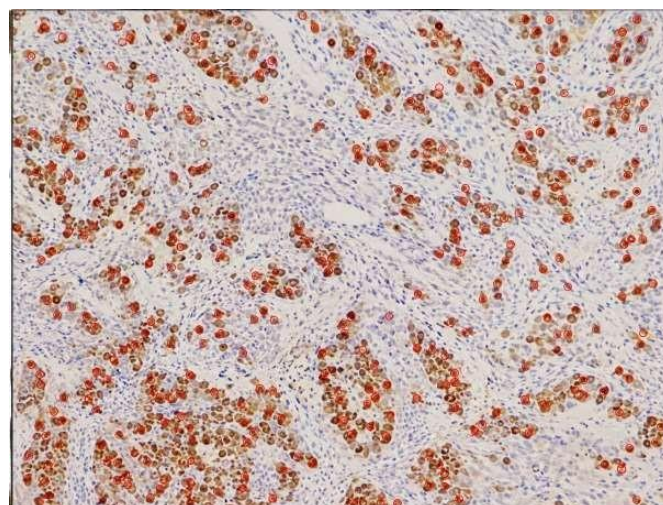


Figure 3. QuPath annotation of positive tumour nuclei in a field of invasive breast carcinoma of no special type. Nests, cords and trabeculae of infiltrating carcinoma cells are separated by cellular desmoplastic stroma. Tumour nuclei showing nuclear DAB immunoreactivity are marked with red circles; spindle-shaped stromal cells and lymphocytes were not scored. Immunoperoxidase staining (DAB chromogen) with haematoxylin counterstain. Original magnification $\times 40$; scale bar = $50 \mu\text{m}$. Source: Own authorship.

Surrogate molecular subtypes

Surrogate subtypes were assigned separately from the manual and the digital readouts using St Gallen definitions [6]. Tumours were hormone receptor-positive if ER or PR was $\geq 1\%$, and HER2-positive if HER2 was 3+ or 2+ with ISH amplification. Luminal A-like tumours were hormone receptor-positive, HER2-negative, with Ki-67 $< 20\%$ and PR $\geq 20\%$; luminal B-like (HER2-) tumours were hormone receptor-positive and HER2-negative with Ki-67 $\geq 20\%$ or PR $< 20\%$; luminal B-like (HER2+) tumours were hormone receptor-positive and HER2-positive; HER2-positive (non-luminal) tumours were hormone receptor-negative and HER2-positive; and triple-negative tumours were negative for ER, PR and HER2. For the digital classification, a QuPath HER2 score of 3+ was taken as positive and a digital 2+ score was resolved by the ISH result.

Follow-up and outcomes

Disease-free survival (DFS) was defined as the time from surgery to the first local, regional or distant recurrence or death from any cause, and overall survival (OS) as the time from surgery to death from any cause. Patients without an event were censored at

the date of last contact.

Statistical analysis

Continuous variables are presented as mean $\pm$ standard deviation (SD) or median (interquartile range, IQR) and categorical variables as numbers and percentages. Characteristics across subtypes were compared with one-way analysis of variance or the Kruskal–Wallis H test, and categorical variables with Pearson's χ^2 test; when more than 20% of cells had expected counts below 5, a Monte Carlo exact χ^2 test with 10,000 samples was used. The Mann–Whitney U test was used for two-group comparisons and Spearman's ρ for monotonic associations. Agreement between manual and digital continuous measurements was assessed with the intraclass correlation coefficient using a two-way random-effects model, absolute agreement and single measures [ICC(2,1)], with 95% confidence intervals (CIs) [33], and interpreted as poor (<0.50), moderate (0.50 – 0.75), good (0.75 – 0.90) or excellent (>0.90) [34]. Lin's concordance correlation coefficient [35] and Bland–Altman analysis, with mean bias and 95% limits of agreement [36], were also calculated, and paired differences were tested with the Wilcoxon signed-rank test. For categorical and ordinal variables, Cohen's κ was calculated, with linear weights for ordinal scales and 95% CIs from 2,000 bootstrap resamples, and interpreted according to Landis and Koch [37]. Directional disagreement was tested with the exact McNemar test for binary variables and Bowker's test of symmetry for multi-category variables.

With manual assessment as the reference standard, sensitivity, specificity, positive and negative predictive values and accuracy of the digital classification were calculated for each subtype in a one-versus-rest manner, with Wilson score 95% CIs. The ability of QuPath HER2 membrane positivity to identify manual HER2 3+ tumours was examined with a receiver operating characteristic (ROC) curve, and the cut-off was chosen with the Youden index [38]. DFS and OS were estimated by the Kaplan–Meier method and compared with the log-rank test. Hazard ratios (HRs) were estimated by Cox proportional hazards regression. Because only 19 DFS events occurred, the multivariable model was limited to two prespecified covariates to keep about ten events per variable [39]; the proportional hazards assumption was checked with scaled Schoenfeld residuals, and discrimination was summarised with Harrell's C-index [40]. No formal sample size calculation was performed. All tests were two-sided with $p < 0.05$ considered significant. Analyses were performed in Python 3 with the SciPy, scikit-learn and lifelines libraries.

Ethical considerations

The study was approved by the Ethics Committee of the University of Samarra (approval no. 3774) and was conducted in accordance with the Declaration of Helsinki [41]. The requirement for individual informed consent was waived by the committee because archived material and anonymised records were used. Patient identifiers were removed before analysis.

Results

Characteristics of the cohort

The 120 patients had a mean age of 48.6 ± 11.1 years (range 26–78), and 64 (53.3%) were postmenopausal (Table 1). Most tumours were of no special type (112, 93.3%), grade 2 (62, 51.7%) and pT2 (85, 70.8%); the median tumour size was 3.2 cm (IQR 2.2–4.0). Sixty-six patients (55.0%) had nodal metastases and four (3.3%) had distant metastases at diagnosis. By manual assessment, 30 tumours (25.0%) were luminal A-like, 23 (19.2%) luminal B-like (HER2–), 7 (5.8%) luminal B-like (HER2+), 30 (25.0%) HER2-positive (non-luminal) and 30 (25.0%) triple-negative. Grade (P = 0.016), nodal category (p = 0.008) and stromal TILs (P < 0.001) differed between subtypes. Grade 3 was most frequent in triple-negative (53.3%) and HER2-positive (50.0%) tumours, and stromal TILs were highest in triple-negative tumours (median 34.0%, IQR 29.0–41.8). Contrary to what is usually reported, node-negative disease was least frequent in the luminal A-like group (23.3%). Age, body mass index, tumour size, pT category and lymphovascular invasion did not differ between subtypes, and among the haematological and biochemical parameters only AST differed (P = 0.005; Table S2).

Table 1. Baseline demographic and clinicopathological characteristics of the study cohort stratified

Variable	Total (n = 120)	Luminal A-like (n = 30)	Luminal B-like (HER2–) (n = 23)	Luminal B-like (HER2+) (n = 7)	HER2-positive (non-luminal) (n = 30)	Triple-negative (n = 30)	P value
Age at diagnosis, years	48.6 $\pm$ 11.1	47.2 $\pm$ 11.2	50.0 $\pm$ 11.9	47.3 $\pm$ 8.2	50.4 $\pm$ 10.7	47.4 $\pm$ 11.9	0.722 ^a
Menopausal status							0.911 ^a
Premenopausal	56 (46.7)	14 (46.7)	12 (52.2)	3 (42.9)	12 (40.0)	15 (50.0)	
Postmenopausal	64 (53.3)	16 (53.3)	11 (47.8)	4 (57.1)	18 (60.0)	15 (50.0)	
Residence							0.838 ^a
Urban	80 (66.7)	21 (70.0)	13 (56.5)	5 (71.4)	20 (66.7)	21 (70.0)	
Rural	40 (33.3)	9 (30.0)	10 (43.5)	2 (28.6)	10 (33.3)	9 (30.0)	
BMI, kg/m ²	29.6 $\pm$ 4.7	29.6 $\pm$ 4.7	29.7 $\pm$ 4.9	31.3 $\pm$ 5.6	29.6 $\pm$ 4.9	29.4 $\pm$ 4.2	0.905 ^a
Histological type (WHO 2019)							0.897 ^d
NST	112 (93.3)	29 (96.7)	21 (91.3)	6 (85.7)	28 (93.3)	28 (93.3)	
Special types	8 (6.7)	1 (3.3)	2 (8.7)	1 (14.3)	2 (6.7)	2 (6.7)	
Nottingham grade							0.016 ^d
Grade 1	10 (8.3)	7 (23.3)	0 (0.0)	0 (0.0)	2 (6.7)	1 (3.3)	
Grade 2	62 (51.7)	17 (56.7)	15 (65.2)	4 (57.1)	13 (43.3)	13 (43.3)	
Grade 3	48 (40.0)	6 (20.0)	8 (34.8)	3 (42.9)	15 (50.0)	16 (53.3)	
Tumour size, cm	3.2 (2.2–4.0)	2.8 (2.1–3.8)	3.8 (3.0–4.4)	3.2 (2.5–4.0)	3.0 (2.2–3.6)	3.2 (1.9–4.0)	0.320 ^b
pT category (AJCC 8th)							0.540 ^b
T1	26 (21.7)	7 (23.3)	3 (13.0)	1 (14.3)	6 (20.0)	9 (30.0)	
T2	85 (70.8)	22 (73.3)	19 (82.6)	6 (85.7)	21 (70.0)	17 (56.7)	
T3	3 (2.5)	0 (0.0)	1 (4.3)	0 (0.0)	0 (0.0)	2 (6.7)	
T4	6 (5.0)	1 (3.3)	0 (0.0)	0 (0.0)	3 (10.0)	2 (6.7)	
pN category (AJCC 8th)							0.008 ^d
N0	54 (45.0)	7 (23.3)	13 (56.5)	2 (28.6)	18 (60.0)	14 (46.7)	
N1	28 (23.3)	11 (36.7)	5 (21.7)	1 (14.3)	5 (16.7)	6 (20.0)	
N2	19 (15.8)	8 (26.7)	1 (4.3)	4 (57.1)	4 (13.3)	2 (6.7)	
N3	19 (15.8)	4 (13.3)	4 (17.4)	0 (0.0)	3 (10.0)	8 (26.7)	
Distant metastasis							0.297 ^d
M0	116 (96.7)	29 (96.7)	23 (100.0)	6 (85.7)	28 (93.3)	30 (100.0)	
M1	4 (3.3)	1 (3.3)	0 (0.0)	1 (14.3)	2 (6.7)	0 (0.0)	
Lymphovascular invasion							0.267 ^c
Absent	85 (70.8)	20 (66.7)	16 (69.6)	3 (42.9)	25 (83.3)	21 (70.0)	
Present	35 (29.2)	10 (33.3)	7 (30.4)	4 (57.1)	5 (16.7)	9 (30.0)	

Surgical margin status						0.084 ^a
Negative	74 (61.7)	17 (56.7)	17 (73.9)	7 (100.0)	15 (50.0)	18 (60.0)
Close (<1 mm)	23 (19.2)	8 (26.7)	4 (17.4)	0 (0.0)	4 (13.3)	7 (23.3)
Positive (ink on tumor)	23 (19.2)	5 (16.7)	2 (8.7)	0 (0.0)	11 (36.7)	5 (16.7)
Associated DCIS						0.465 ^a
No	64 (53.3)	18 (60.0)	9 (39.1)	5 (71.4)	15 (50.0)	17 (56.7)
Yes	56 (46.7)	12 (40.0)	14 (60.9)	2 (28.6)	15 (50.0)	13 (43.3)
Stromal TILs, %	19.0 (10.0–28.2)	17.0 (8.5–23.0)	20.0 (10.5–25.0)	9.0 (4.0–24.0)	11.0 (7.5–15.8)	34.0 (29.0–41.8)
						<0.001 ^b

Data are presented as n (%), mean $\pm$ standard deviation, or median (interquartile range). Percentages are column percentages within each subtype and may not total 100 because of rounding. Surrogate molecular subtypes were assigned from the manual reference assessment (St Gallen surrogate definitions; Ki-67 cut-off 20%, PR cut-off 20%; HER2 status per ASCO/CAP 2018). Special histological types comprised invasive lobular (n = 4), tubular (n = 3) and mucinous (n = 1) carcinomas; one NST carcinoma with a medullary pattern was classified as NST. a One-way ANOVA. b Kruskal–Wallis H test. c Pearson χ^2 test. d Monte Carlo exact χ^2 test (10,000 samples) because more than 20% of cells had an expected count below 5. P values in bold are statistically significant (P < 0.05). Abbreviations: AJCC, American Joint Committee on Cancer; BMI, body mass index; DCIS, ductal carcinoma in situ; HER2, human epidermal growth factor receptor 2; NST, no special type; TILs, tumour-infiltrating lymphocytes; WHO, World Health Organization. Source: Own authorship.

Agreement for continuous measurements

Agreement between QuPath and manual measurements was excellent for every continuous parameter, with ICC(2,1) values from 0.939 for tumour cellularity to 0.994 for PR-positive nuclei; Lin's coefficients were almost identical (Table 2; Figure 4). Mean differences were small, from –1.40 H-score units for PR to +0.79 for ER. Only PR-positive nuclei showed a statistically significant paired difference (P = 0.027), and its magnitude (mean bias 0.24 percentage points) is unlikely to be clinically relevant. The limits of agreement were narrow for percentages of positive nuclei (for Ki-67, –5.4 to +6.0 percentage points) but wide for the ER H-score (–50.1 to +51.7), indicating that individual H-scores could differ considerably even when overall agreement was high (Figure 5).

Table 2. Agreement between manual and QuPath-derived continuous immunohistochemical measurements (n = 120).

Parameter	Manual Median (IQR)	QuPath Median (IQR)	P value ^a	Spearman ρ	ICC(2,1) (95% CI)	Lin CCC (95% CI)	Bias $\pm$ SD ^b	95% LoA
Tumour cellularity, %	45.0 (34.8–55.0)	45.0 (35.0–57.0)	0.503 ^a	0.928	0.939 (0.914–0.957)	0.939 (0.913–0.957)	–0.21 $\pm$ 5.23	–10.47 to 10.05
ER-positive nuclei, %	2.1 (0.0–66.3)	2.4 (0.2–68.4)	0.663 ^a	0.956	0.993 (0.991–0.995)	0.993 (0.990–0.995)	–0.28 $\pm$ 4.22	–8.56 to 7.99
ER H-score (0–300)	1.5 (0.0–160.5)	4.0 (0.0–158.2)	0.856 ^a	0.979	0.963 (0.947–0.974)	0.962 (0.947–0.974)	0.79 $\pm$ 25.97	–50.11 to 51.69
PR-positive nuclei, %	1.1 (0.0–48.7)	1.6 (0.2–46.8)	0.027^a	0.960	0.994 (0.992–0.996)	0.994 (0.992–0.996)	0.24 $\pm$ 3.07	–5.77 to 6.26
PR H-score (0–300)	1.0 (0.0–104.5)	1.5 (0.0–99.2)	0.863 ^a	0.984	0.982 (0.975–0.988)	0.982 (0.975–0.988)	–1.40 $\pm$ 13.33	–27.53 to 24.73
Ki-67 index, %	33.0 (16.2–44.7)	34.2 (15.9–45.1)	0.274 ^a	0.988	0.988 (0.983–0.992)	0.988 (0.983–0.992)	0.30 $\pm$ 2.89	–5.36 to 5.96

^a Wilcoxon signed-rank test for a systematic difference between paired measurements. Bias = mean difference (QuPath – manual) $\pm$ standard deviation of the differences; 95% limits of agreement (LoA) = bias $\pm$ 1.96 SD (Bland–Altman method). ICC(2,1): two-way random-effects model, absolute agreement, single measures. Agreement was interpreted as poor (<0.50), moderate (0.50–0.75), good (0.75–0.90) or excellent (>0.90) (Koo and Li, 2016). P values

in bold are statistically significant (P < 0.05). Abbreviations: CCC, concordance correlation coefficient; CI, confidence interval; ER, oestrogen receptor; ICC, intraclass correlation coefficient; IQR, interquartile range; PR, progesterone receptor. Source: Own authorship.

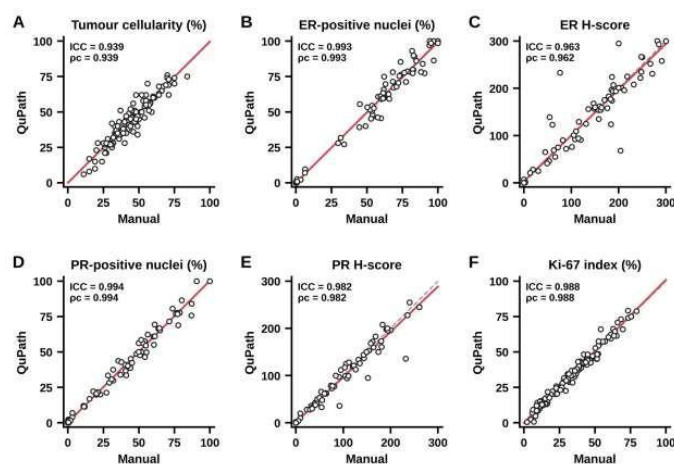


Figure 4. Correlation between manual and QuPath-derived immunohistochemical measurements. Scatter plots of QuPath versus manual values for (A) tumour cellularity, (B) ER-positive nuclei, (C) ER H-score, (D) PR-positive nuclei, (E) PR H-score and (F) Ki-67 index (n = 120). Each circle represents one tumour. The dashed grey line represents the line of identity (y = x) and the solid red line the least-squares regression line. ICC, intraclass correlation coefficient (two-way random effects, absolute agreement, single measures); pc, Lin's concordance correlation coefficient. Source: Own authorship.

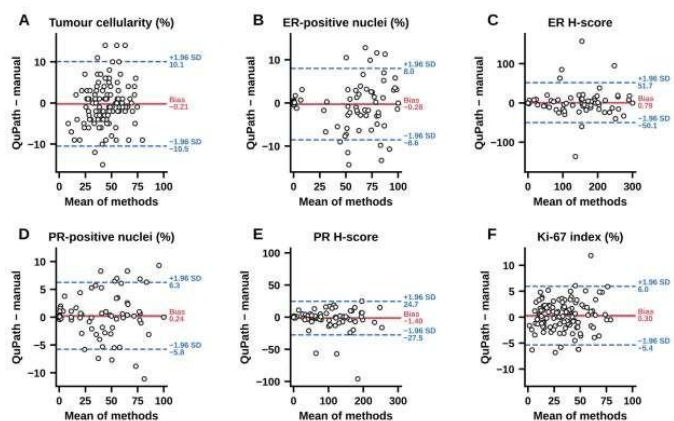


Figure 5. Bland–Altman analysis of agreement between QuPath-derived and manual immunohistochemical measurements. The difference between methods (QuPath – manual) is plotted against their mean for (A) tumour cellularity, (B) ER-positive nuclei, (C) ER H-score, (D) PR-positive nuclei, (E) PR H-score and (F) Ki-67 index (n = 120). The solid red line indicates the mean difference (bias) and the dashed blue lines the 95% limits of agreement (bias $\pm$ 1.96 SD). Source: Own authorship.

Agreement for categorical biomarker status

Binary status showed almost perfect agreement: ER positivity 119/120 (99.2%, κ 0.983), PR positivity 117/120 (97.5%, κ 0.950), PR $\geq$ 20% 118/120 (98.3%, κ 0.966) and Ki-67 $\geq$ 20% 117/120 (97.5%, κ 0.940),

with no significant directional disagreement (Table 3). Exact agreement for the Allred score was lower (76.7% for ER and 74.2% for PR), although weighted κ remained above 0.90. For both receptors the Bowker test showed systematic disagreement (ER $P = 0.025$; PR $P < 0.001$), because QuPath assigned a higher Allred score more often than a lower one (ER, 22 vs 6 cases; PR, 28 vs 3 cases); differences exceeded one point in four cases for each receptor.

Table 3. Categorical concordance of immunohistochemical biomarker status and scores between manual and QuPath-assisted assessment (n = 120).

Biomarker / category	Categories	Exact agreement n/N (%)	Agreement statistic	κ (95% CI) ^a	P value
ER status ($\geq 1\%$)	Negative / Positive	119/120 (99.2)	Cohen κ	0.983 (0.949–1.000)	1.000 ^b
ER category (ASCO/CAP 2020)	Negative / Low / Positive	119/120 (99.2)	Linear-weighted κ	0.992 (0.953–1.000)	0.317 ^c
ER staining intensity	0 / 1 / 2 / 3	113/120 (94.2)	Linear-weighted κ	0.926 (0.865–0.976)	0.675 ^c
ER Allred score	0–8 (ordinal)	92/120 (76.7)	Linear-weighted κ	0.919 (0.851–0.938)	0.025^c
PR status ($\geq 1\%$)	Negative / Positive	117/120 (97.5)	Cohen κ	0.950 (0.884–1.000)	0.250 ^b
PR $\geq 20\%$	<20% / $\geq 20\%$	118/120 (98.3)	Cohen κ	0.966 (0.913–1.000)	1.000 ^b
PR staining intensity	0 / 1 / 2 / 3	114/120 (95.0)	Linear-weighted κ	0.952 (0.910–0.988)	0.199 ^c
PR Allred score	0–8 (ordinal)	89/120 (74.2)	Linear-weighted κ	0.904 (0.856–0.927)	<0.001^c
HER2 IHC score	0 / 1+ / 2+ / 3+	111/120 (92.5)	Linear-weighted κ	0.940 (0.899–0.974)	0.193 ^c
Ki-67 index ($\geq 20\%$)	Low / High	117/120 (97.5)	Cohen κ	0.940 (0.865–1.000)	0.250 ^b

^a 95% confidence intervals were estimated by non-parametric bootstrap resampling (2,000 replicates). Linear weights were applied to ordinal scales. κ values were interpreted as slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) or almost perfect (0.81–1.00) (Landis and Koch, 1977). Exact McNemar test. ^c Bowker test of symmetry. These tests assess directional (systematic) disagreement; $p \geq 0.05$ indicates no systematic bias of the digital method. ER category (ASCO/CAP 2020): negative, <1%; low positive, 1–10%; positive, >10% of tumour nuclei. p-values in bold are statistically significant ($P < 0.05$). Abbreviations: ASCO/CAP, American Society of Clinical Oncology/College of American Pathologists; CI, confidence interval; ER, oestrogen receptor; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; PR, progesterone receptor. Source: Own authorship.

HER2 assessment

HER2 scores agreed in 111/120 cases (92.5%; linear-weighted κ 0.940), and every disagreement involved adjacent categories (Table 4; Figure 6A). Seven disagreements were at the 0/1+ boundary (five manual 1+ scored 0 by QuPath and two manual 0 scored 1+), and two manual 2+ tumours were scored 3+ by QuPath; one of these was ISH-amplified and the other was not (Table S1). When tumours were grouped as HER2-0, HER2-low or HER2-positive, agreement was 112/120 (93.3%; κ 0.900). The QuPath percentage of HER2 membrane-positive cells increased across manual scores, with medians of 3.4%, 18.0%, 42.7% and

83.5% for scores 0, 1+, 2+ and 3+ ($p < 0.001$); Figure 6B). This measure identified manual 3+ tumours with an area under the ROC curve of 0.982 (95% CI 0.956–0.999); at a cut-off of $\geq 62.2\%$, sensitivity was 100% and specificity 95.8% (Figure 6C). Because the cut-off was derived and tested in the same cases, these estimates are likely to be optimistic.

Table 4. Cross-tabulation of HER2 immunohistochemical scores assigned by manual assessment and QuPath digital analysis (n = 120).

Manual HER2 IHC score	QuPath 0	QuPath 1+	QuPath 2+	QuPath 3+	Total
Score 0	40 (95.2)	2 (4.8)	0 (0.0)	0 (0.0)	42 (100.0)
Score 1+	5 (13.2)	33 (86.8)	0 (0.0)	0 (0.0)	38 (100.0)
Score 2+	0 (0.0)	0 (0.0)	13 (86.7)	2 (13.3)	15 (100.0)
Score 3+	0 (0.0)	0 (0.0)	0 (0.0)	25 (100.0)	25 (100.0)
Total	45 (37.5)	35 (29.2)	13 (10.8)	27 (22.5)	120 (100.0)

Data are n (% within manual score). Diagonal (concordant) cells are shown in bold. Overall exact agreement was 111/120 (92.5%); linear-weighted $\kappa = 0.940$ (95% CI 0.899–0.974); Bowker test of symmetry $P = 0.193$. All 15 manually equivocal (2+) cases underwent reflex in situ hybridisation (ISH): 12 amplified and 3 not amplified. Of the two cases upgraded from 2+ to 3+ by QuPath, one was ISH-amplified (BC060) and one was not amplified (BC010). All discordances were confined to adjacent score categories ($0 \leftrightarrow 1+$, $n = 7$; $2+ \rightarrow 3+$, $n = 2$). No manual 0/1+ case was scored 3+ by QuPath, and no manual 3+ case was downgraded. Abbreviations: HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry. Source: Own authorship.

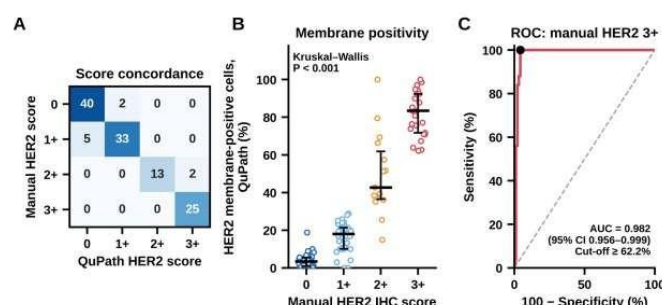


Figure 6. Digital assessment of HER2 immunohistochemistry. (A) Heat map of the cross-tabulation of manual versus QuPath HER2 IHC scores; exact agreement 111/120 (92.5%), linear-weighted $\kappa = 0.940$. (B) Percentage of HER2 membrane-positive tumour cells measured by QuPath according to manual HER2 score; horizontal bars indicate the median and whiskers the interquartile range (Kruskal–Wallis $P < 0.001$). (C) Receiver operating characteristic curve of QuPath membrane positivity for identifying manual HER2 3+ tumours; area under the curve (AUC) = 0.982 (95% CI 0.956–0.999; bootstrap, 2,000 replicates). The black dot marks the optimal cut-off (Youden index) of $\geq 62.2\%$ membrane-positive cells (sensitivity 100.0%, specificity 95.8%). Source: Own authorship.

Concordance of surrogate molecular subtypes

The two methods assigned the same surrogate subtype in 114 of 120 tumours (95.0%; κ 0.935, 95% CI 0.880–0.978) (Table 5; Figure 7). Per-subtype

sensitivity ranged from 86.7% for triple-negative tumours to 100% for the luminal B-like and HER2-positive groups, and specificity from 94.8% for luminal B-like (HER2-) to 100% (Table 6). The lowest positive predictive value (82.1%) was for luminal B-like (HER2-), which received five tumours classified differently by manual assessment. Because the cohort was not sampled consecutively, predictive values should not be extrapolated to routine practice.

Table 5. Cross-tabulation of surrogate molecular subtypes assigned by manual reference assessment and by the QuPath-assisted digital framework (n = 120).

Manual reference subtype	Lum A	Lum B (HER2-)	Lum B (HER2+)	HER2+ (non-lum)	TNBC	Total
Luminal A-like	28	2	0	0	0	30
Luminal B-like (HER2-)	0	23	0	0	0	23
Luminal B-like (HER2+)	0	0	7	0	0	7
HER2-positive (non-luminal)	0	0	0	30	0	30
Triple-negative	0	3	0	1	26	30
Total	28	28	7	31	26	120

Rows: manual reference subtype; columns: QuPath digital subtype. Diagonal (concordant) cells are shown in bold. Overall agreement was 114/120 (95.0%); Cohen κ = 0.935 (95% CI 0.880–0.978; bootstrap, 2,000 replicates), indicating almost perfect agreement. The six discordant cases are detailed in Table 7. Abbreviations: HER2, human epidermal growth factor receptor 2; Lum, luminal; non-lum, nonluminal; TNBC, triple-negative breast cancer. Source: Own authorship

Table 6. Diagnostic performance of the QuPath-assisted framework for each surrogate molecular subtype, using manual assessment as the reference standard (n = 120).

Surrogate subtype	TP/FP/FN/TN	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)	NPV % (95% CI)	Accuracy % (95% CI)	Balanced accuracy
Luminal A-like	28/0/2/90	93.3 (78.7–98.2)	100.0 (95.9–100.0)	100.0 (87.9–100.0)	97.8 (92.4–99.4)	98.3 (94.1–99.5)	0.967
Luminal B-like (HER2-)	23/5/0/92	100.0 (85.7–100.0)	94.8 (88.5–97.8)	82.1 (64.4–92.1)	100.0 (96.0–100.0)	95.8 (90.6–98.2)	0.974
Luminal B-like (HER2+)	7/0/0/113	100.0 (64.6–100.0)	100.0 (96.7–100.0)	100.0 (64.6–100.0)	100.0 (96.9–100.0)	100.0 (96.9–100.0)	1.000
HER2-positive (non-luminal)	30/1/0/89	100.0 (88.6–100.0)	98.9 (94.0–99.8)	96.8 (83.8–99.4)	100.0 (95.9–100.0)	99.2 (95.4–99.9)	0.994
Triple-negative	26/0/4/90	86.7 (70.3–94.7)	100.0 (95.9–100.0)	100.0 (87.1–100.0)	95.7 (89.6–98.3)	96.7 (91.7–98.7)	0.933

Performance was calculated for each subtype in a one-versus-rest manner, with the manual pathologist assessment as the reference standard. 95% confidence intervals were calculated by the Wilson score method. Balanced accuracy = (sensitivity + specificity)/2. Abbreviations: CI, confidence interval; FN, false negative; FP, false positive; HER2, human epidermal growth factor receptor 2; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive. Source: Own authorship.

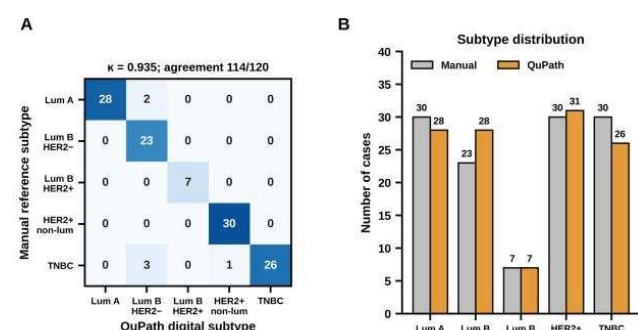


Figure 7. Concordance of surrogate molecular subtype

assignment between manual assessment and the QuPath-assisted framework. (A) Confusion matrix of manual reference versus QuPath digital subtypes; overall agreement 114/120 (95.0%), Cohen κ = 0.935 (95% CI 0.880–0.978). (B) Distribution of cases across surrogate subtypes according to each method; numbers above bars indicate case counts. Abbreviations: Lum, luminal; non-lum, non-luminal; TNBC, triple-negative breast cancer. Source: Own authorship.

Discordant cases

The six discordant cases fell into three groups (Table 7). In three tumours classified as triple-negative by manual assessment, QuPath detected 1–10% positive nuclei (ER 2.5% and PR 2.3% in one tumour, and PR 1.6% and 1.1% with ER <1% in the other two), which moved them into the luminal B-like (HER2-) group. Two luminal A-like tumours with manual Ki-67 of 19.0% had digital values of 21.3% and 20.0% and were reclassified as luminal B-like. The sixth tumour was upgraded from HER2 2+ to 3+ by QuPath despite a non-amplified ISH result, and was therefore digitally classified as HER2-positive. No discordance occurred in a tumour whose measurements lay away from a clinical cut-off.

Table 7. Clinicopathological and immunohistochemical profile of cases with discordant surrogate subtype assignment between manual and QuPath assessment (n = 6).

Case	Manual subtype	QuPath subtype	ER % M/Q	PR % M/Q	HER2 IHC M/Q	ISH	Ki-67 % M/Q	Attributed source of discordance
BC009	Triple-negative	Luminal B-like (HER2-)	0.8 / 2.5	0.3 / 2.3	0 / 0	NP	61.2 / 59.5	ER/PR in 1–10% (ER-low) zone: digital ER 2.5% and PR 2.3% crossed the 1% positivity threshold
BC010	Triple-negative	HER2-positive (non-luminal)	0.0 / 0.2	0.3 / 0.3	2+ / 3+	Not amplified	52.4 / 54.4	HER2 IHC upgraded 2+→3+ by QuPath although reflex ISH was not amplified
BC011	Triple-negative	Luminal B-like (HER2-)	0.2 / 0.1	0.0 / 1.6	2+ / 2+	Not amplified	57.1 / 57.5	Digital PR 1.6% crossed the 1% threshold (ER-negative/PR-low); HER2 2+; ISH not amplified
BC014	Triple-negative	Luminal B-like (HER2-)	0.2 / 0.0	0.0 / 1.1	1+ / 1+	NP	79.3 / 78.8	Digital PR 1.1% crossed the 1% threshold (ER-negative/PR-low)
BC04	Luminal A-like	Luminal B-like (HER2-)	96.9 / 100.0	21.1 / 21.6	1+ / 1+	NP	19.0 / 21.3	Ki-67 near the 20% cut-off (manual 19.0% vs digital 21.3%)
BC06	Luminal A-like	Luminal B-like (HER2-)	72.8 / 75.4	27.1 / 21.3	1+ / 1+	NP	19.0 / 20.0	Ki-67 at the 20% cut-off (manual 19.0% vs digital 20.0%)

Values are shown as manual (M) / QuPath (Q). Discordance arose from three mechanisms: (i) ER/PR expression within the 1–10% low-positive range, where digital detection of $\geq 1\%$ positive nuclei reclassified triple-negative tumours as luminal (n = 3); (ii) Ki-67 values at or immediately adjacent to the 20% cut-off (n = 2); and (iii) digital upgrading of a HER2 2+ case to 3+ despite a non-amplified reflex ISH result (n = 1). Under ASCO/CAP 2020 guidance, tumours with ER 1–10% are reported as “ER-low positive”; BC009 meets this definition by digital assessment. BC016 and BC047 were ER-negative with PR expression of 1–2%, a rare phenotype that is generally managed as triple-negative. For BC010, the ISH result (not amplified) supersedes the IHC score, so the final HER2 status is negative. Abbreviations: ER, oestrogen receptor; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; ISH, in situ hybridisation; NP, not performed (not indicated); PR, progesterone receptor. Source: Own authorship.

Digital biomarkers and clinicopathological parameters

The distributions of the digital readouts across

QuPath subtypes are shown in Figure 8; the differences are expected, because the subtypes are defined by these markers. QuPath Ki-67 was lower in node-positive than in node-negative tumours (median 26.9% vs 38.2%, $P = 0.020$), and ER H-score was higher in node-positive tumours (66 vs 0, $p = 0.020$) (Table 8). These associations probably reflect the high frequency of nodal metastases in the luminal A-like group of this cohort rather than a biological effect of the markers. None of the digital readouts differed significantly by grade, pT category, lymphovascular invasion or age group. Correlations with continuous histological variables were weak ($|p| \leq 0.22$); the strongest were inverse correlations of stromal TILs with ER H-score ($\rho = -0.219$, $p = 0.017$) and HER2 membrane positivity ($\rho = -0.215$, $p = 0.018$).

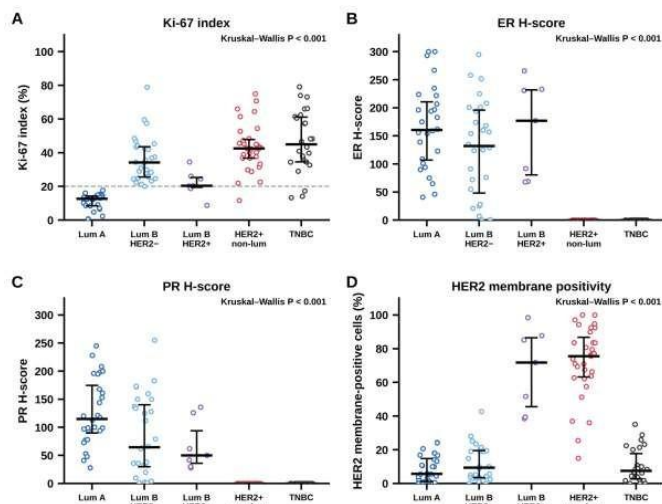


Figure 8. Distribution of QuPath-derived biomarker readouts across digital surrogate molecular subtypes. (A) Ki-67 index, (B) ER H-score, (C) PR H-score and (D) percentage of HER2 membrane-positive tumour cells, grouped by QuPath digital subtype ($n = 120$). Each circle represents one tumour; horizontal bars indicate the median and whiskers the interquartile range. The dashed line in (A) marks the 20% Ki-67 cut-off used for surrogate subtyping. P values were derived from the Kruskal–Wallis H test. Because subtypes are defined by these biomarkers, between-group differences are expected by design; the figure illustrates the spread of continuous digital readouts within each category. Source: Own authorship.

Table 8. Association of QuPath-derived biomarker readouts with standard histopathological and clinical parameters ($n = 120$).

Variable	n	Ki-67 index (%)	ER H-score (0–300)	PR H-score (0–300)	HER2 membrane-positive cells (%)
Nottingham grade					
Grade 1	10	13.2 (9.5–38.8)	92 (19–185)	58 (7–93)	5.7 (1.5–19.1)
Grade 2	62	30.4 (16.4–45.4)	62 (0–157)	29 (0–120)	17.4 (4.5–41.9)
Grade 3	48	34.6 (24.6–44.7)	0 (0–139)	0 (0–66)	20.8 (5.2–65.2)
P value		0.134 ^a	0.090 ^a	0.056 ^a	0.274 ^a

pT category					
T1	26	33.7 (15.3–47.6)	0 (0–158)	0 (0–70)	12.1 (4.5–63.9)
T2–T4	94	34.2 (16.5–43.2)	23 (0–158)	15 (0–109)	18.0 (4.4–51.6)
P value		0.644 ^b	0.592 ^b	0.436 ^b	0.871 ^b
Nodal status					
N0	54	38.2 (24.8–45.2)	0 (0–124)	0 (0–71)	16.8 (4.6–59.6)
N1–N3	66	26.9 (13.9–44.6)	66 (0–194)	37 (0–110)	18.0 (4.4–41.9)
P value		0.020^b	0.020^b	0.089 ^b	0.762 ^b
Lymphovascular invasion					
Absent	85	35.7 (17.3–44.6)	0 (0–158)	0 (0–99)	18.8 (4.3–62.2)
Present	35	29.9 (14.8–46.1)	69 (0–166)	38 (0–108)	10.6 (4.4–32.5)
P value		0.496 ^b	0.296 ^b	0.242 ^b	0.439 ^b
Age group					
<50 years	67	34.4 (16.9–44.4)	1 (0–158)	4 (0–86)	16.7 (4.3–50.1)
≥50 years	53	34.0 (14.7–47.3)	7 (0–162)	1 (0–117)	18.8 (5.5–51.7)
P value		0.949 ^b	0.982 ^b	0.933 ^b	0.979 ^b
Spearman ρ (P value)					
Nottingham total score	120	0.101 (0.273)	–0.169 (0.065)	–0.186 (0.042)	0.072 (0.435)
Tumour size, cm	120	0.018 (0.847)	0.054 (0.556)	0.051 (0.579)	–0.030 (0.745)
Positive lymph nodes, n	120	–0.175 (0.056)	0.171 (0.062)	0.146 (0.113)	0.040 (0.668)
Stromal TILs, %	120	0.179 (0.0499)	–0.219 (0.017)	–0.168 (0.066)	–0.215 (0.018)

Data are median (interquartile range) unless otherwise indicated. The final section shows Spearman rank correlation coefficients (ρ) with P values in parentheses. ^a Kruskal–Wallis H test. ^b Mann–Whitney U test. P values in bold are statistically significant ($P < 0.05$). Abbreviations: ER, oestrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor; TILs, tumour-infiltrating lymphocytes. Source: Own authorship.

Survival

Over a median follow-up of 46 months (IQR 36.8–52.0; range 24–60), 19 DFS events and nine deaths occurred. In univariable Cox analysis, triple-negative status was associated with shorter DFS whether defined manually (HR 2.99, 95% CI 1.21–7.36, $p = 0.017$) or by QuPath (HR 3.65, 95% CI 1.48–9.00, $p = 0.005$) (Table 9). Ki-67 was not associated with DFS by either method, and its discrimination was close to chance (C-index 0.528 manual and 0.531 QuPath). In the two-variable model, neither digital Ki-67 nor nodal status was significant. Differences in DFS across the five subtypes were of borderline significance (log-rank $p = 0.051$), and OS did not differ ($p = 0.431$) (Figure 9). With so few events, these survival analyses should be regarded as exploratory.

Table 9. Cox proportional hazards regression analysis of factors associated with disease-free survival ($n = 120$; 19 events).

Variable	Univariable HR (95% CI)	P value	Multivariable ^a HR (95% CI)	P value
Age (per 10 years)	1.24 (0.84–1.83)	0.279	—	—
Tumour size (per 1 cm)	1.11 (0.85–1.45)	0.439	—	—
Nottingham grade 3 (vs 1–2)	2.20 (0.89–5.48)	0.090	—	—
Nodal status N1–N3 (vs N0)	1.85 (0.70–4.86)	0.214	1.99 (0.74–5.33)	0.170
Lymphovascular invasion (present vs absent)	0.44 (0.13–1.52)	0.197	—	—

Ki-67, manual (per 10%)	1.04 (0.82–1.32)	0.761	—	—
Ki-67, QuPath (per 10%)	1.06 (0.84–1.34)	0.611	1.10 (0.87–1.38)	0.426
ER H-score, QuPath (per 50 units)	1.01 (0.79–1.28)	0.956	—	—
Triple-negative, manual (vs non-TNBC)	2.99 (1.21–7.36)	0.017	—	—
Triple-negative, QuPath (vs non-TNBC)	3.65 (1.48–9.00)	0.005	—	—

^a The multivariable model was restricted to two prespecified covariates (QuPath Ki-67 index and nodal status) to respect the events-per-variable limit (19 disease-free survival events). The proportional hazards assumption was satisfied (global test on scaled Schoenfeld residuals, $p = 0.75$). Triple-negative status was entered in separate univariable models according to manual and QuPath classification and was not combined with other covariates. Harrell's C-index of the continuous Ki-67 index for disease-free survival was 0.528 (manual) and 0.531 (QuPath). P-values in bold are statistically significant ($p < 0.05$). Abbreviations: CI, confidence interval; HR, hazard ratio; TNBC, triple-negative breast cancer. Source: Own authorship.

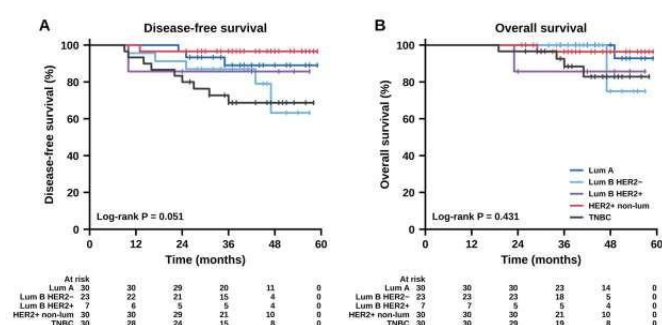


Figure 9. Kaplan–Meier survival curves according to manual surrogate molecular subtype. (A) Disease-free survival (19 events; log-rank $P = 0.051$) and (B) overall survival (9 events; log-rank $p = 0.431$). Vertical tick marks indicate censored observations. Median follow-up was 46 months (interquartile range 36.8–52.0). The numbers of patients at risk at each time point are shown below each panel. Abbreviations: Lum, luminal; non-lum, non-luminal; TNBC, triple-negative breast cancer. Source: Own authorship.

Discussion

In this cohort of 120 invasive breast carcinomas from Samarra, QuPath-assisted scoring of ER, PR, HER2 and Ki-67 agreed closely with manual assessment. ICC values for the continuous measurements exceeded 0.93, binary status agreed in 97–99% of cases, and the surrogate subtype was the same in 95% of tumours. The disagreements that did occur were not scattered: everyone involved a measurement near a treatment-relevant threshold, either ER/PR expression between 1% and 10%, Ki-67 close to 20%, or HER2 at the 0/1+ or 2+/3+ boundaries. To our knowledge this is the first evaluation of digital biomarker scoring in breast cancer from Iraq.

Agreement for continuous measurements

The agreement we observed for Ki-67 (ICC 0.988) is higher than that reported by Rewcastle et al., who found ICC values above 0.78 between manual and

digital scoring in 367 hormone receptorpositive, node-negative tumours [22], and is in keeping with the experience of Catteau et al., who tested a simple QuPath-based workflow for daily practice and found it promising [23]. Several features of our design probably raised the agreement. Both methods were applied in a single laboratory to the same stained sections, and the pathologist defined the tumour region used by QuPath, so differences in region selection, which are a major source of variation in Ki-67 scoring [11,12], were largely removed. Our estimates therefore describe agreement between two ways of reading the same slide and not the inter-laboratory reproducibility that matters most for external validity.

High ICC values did not mean that individual results were interchangeable. The limits of agreement for the ER H-score spanned about ± 50 units, and QuPath tended to give higher Allred scores than the pathologist. Both findings point to intensity rather than proportion as the weaker component. DAB is not a stoichiometric chromogen, and its measured optical density depends on section thickness, fixation, antigen retrieval, stain vectors and scanner settings [17], so fixed intensity thresholds may not match a pathologist's visual judgement of weak, moderate and strong staining. Because treatment decisions depend mainly on the proportion of positive cells, the practical effect of this intensity shift is probably small, but laboratories that report H-scores or Allred scores from digital analysis should calibrate intensity thresholds against local reference cases.

The low-expression zone for ER and PR

Three of the six subtype discordances arose when QuPath counted 1–10% positive nuclei in tumours that the pathologist had classified as receptor-negative. This zone matters clinically. Tumours with ER expression of 1–10% share many biological and clinical features with triple-negative cancer; in the registry study by Yoder et al., HER2-negative tumours with low ER/PR expression had recurrence-free and overall survival similar to triple-negative tumours [42], and a meta-analysis by Paakkola et al. found that ER-low tumours behaved more like ER-negative than ERpositive disease [43]. ASCO/CAP therefore asks laboratories to report ER-low results with a comment and to review them carefully [8]. Two of our three cases were ER-negative with PR expression of only 1–2%, a pattern that is rare and seldom used on its own to direct endocrine therapy. A small number of positive nuclei can also come from entrapped normal epithelium, edge artefact or misclassified non-tumour cells, and digital detection is particularly sensitive to such events when the true positive fraction is very low. We suggest that digital pipelines should flag tumours

with 1–10% ER or PR positivity for pathologist review instead of assigning a subtype automatically.

Ki-67 near the 20% cut-off

The two luminal A-like tumours reclassified as luminal B-like had manual Ki-67 values of 19.0% and digital values of 20.0% and 21.3%. These differences are well within the limits of agreement we observed (about ± 6 percentage points) and illustrate why the IKGW discourages treatment decisions based on Ki-67 values in the intermediate range [11]. The 2021 St Gallen panel supported Ki-67 mainly at the low ($\leq 5\%$) and high ($\geq 30\%$) extremes [6]. A digital score of 20% is no more certain than a manual one; its advantage is reproducibility, not freedom from measurement error, so results close to the cut-off should be interpreted together with grade, PR level and, where available, genomic tests.

HER2 scoring and HER2-low disease

Agreement for the four-tier HER2 score (92.5%, weighted κ 0.940) was high, but seven of the nine disagreements were at the 0/1+ boundary. Before trastuzumab deruxtecan was shown to benefit patients with HER2-low metastatic disease [13], this boundary had little consequence; it now determines eligibility for treatment [14,15]. Pathologists agree poorly at this level [16], and image analysis has been proposed as a way of improving consistency. Wu et al. found that artificial intelligence assistance improved pathologists' accuracy in distinguishing HER2 0 from 1+ [24]. Our data cannot show whether QuPath or the pathologist was correct in the seven boundary cases, because the manual score itself served as the reference; they show only that the boundary remains the main area of disagreement.

The tumour upgraded from 2+ to 3+ by QuPath despite a non-amplified ISH result is an important reminder of the limits of IHC-based quantification. Under ASCO/CAP rules, ISH determines the final HER2 status of equivocal tumours, and unusual IHC–ISH combinations require review [9,10]. A digital 3+ score should not override a negative ISH result, and the digital classification rule we used, in which QuPath 3+ was accepted without ISH, created this discordance. The strong relation between QuPath membrane positivity and the manual score (AUC 0.982) is encouraging, but the threshold of 62.2% was derived and tested in the same cases, the manual score was the reference, and our membrane measure is not identical to the ASCO/CAP definition of complete, intense circumferential staining. The value should be regarded as specific to this workflow and requires external validation.

Comparison with Iraqi studies

Because our cohort was assembled to include similar numbers of each receptor-defined group, the subtype frequencies we report cannot be compared with population prevalence. Jumaah et al., in 1,227 Iraqi tumours, found luminal A-like tumours in 61.7%, luminal B-like in 20.0%, HER2-enriched in 9.5% and triple-negative tumours in 8.7% of cases [2]. Shubbar reported that hormone receptor-positive, HER2-negative tumours with low Ki-67 were the most common group (48.5%) among 511 women from southern Iraq [4], and Chawsheen et al. likewise found luminal A to be the dominant subtype in 261 cases from Erbil [5]. Earlier national data also described subtype distribution on the basis of ER, PR and HER2 status [3]. Other characteristics of our cohort resemble those of these series. The mean age of 48.6 years is close to the 47.3 years reported by Dosh et al. in Najaf [25] and the 50.2 years reported by Jumaah et al. [2]; grade 2 tumours made up 51.7% of our cohort and 51.8% of the southern series; and 70.8% of our tumours were pT2, similar to the 72.8% of tumours measuring 2–5 cm reported by Shubbar [4], which is consistent with relatively late presentation. All of these Iraqi studies relied on manual scoring, and none assessed the reproducibility of the measurements on which their subtype frequencies depend. Our results suggest that such frequencies are robust for most tumours but may shift by a few percentage points depending on how borderline cases are scored.

One finding differs from the Iraqi and international literature. Nodal metastases were most frequent in the luminal A-like group (76.7%), whereas Dosh et al. found that luminal A tumours tended to be smaller [25], and luminal A disease is generally associated with a lower stage [7]. We think this reflects the way the cohort was assembled rather than the biology of luminal A tumours, but it also explains two apparently paradoxical associations: lower digital Ki-67 and higher ER H-scores in node-positive tumours. These associations should not be interpreted as prognostic effects of the markers.

Clinicopathological correlates and survival

Stromal TILs were highest in triple-negative tumours, in keeping with the recognised immunogenicity of this subtype and the rationale for standardised TIL assessment [7,30]. Triple-negative status was associated with an approximately threefold higher hazard of recurrence, with similar HRs for the manual and digital classifications. This suggests that the digital classification preserved the prognostic separation of the manual one, but the CIs were wide, treatment data were not available for adjustment, and the analysis

included only 19 events. Ki-67 was not prognostic in our cohort, which is not surprising given the small number of events and the mixture of subtypes and treatments; the study was not designed to test the prognostic value of Ki-67.

Implications for practice

For Iraqi laboratories, the main attraction of QuPath is that it is free and runs on ordinary workstations [21]. Our results suggest that, once a slide has been scanned and the tumour region annotated, the software reproduces the manual reading closely for most tumours and could serve as a second read, a quality-assurance tool or a way of documenting how a score was obtained. The same results define where it should not be used alone: ER/PR expression of 1–10%, Ki-67 values near the local cut-off, HER2 scores at the 0/1+ boundary, and any HER2 2+ or 3+ result that conflicts with ISH. A practical approach would be to accept concordant results and route these predefined categories to pathologist review. Implementation would also require access to slide scanners, local validation, standard operating procedures and training [18,19], which are not yet widely available in Iraq.

Strengths and limitations

The study has several strengths. Both methods were applied to the same cases, and agreement was evaluated with measures appropriate to agreement (ICC, Lin's coefficient and Bland–Altman analysis) rather than correlation alone. The comparison extended from individual markers to the surrogate subtype, every discordant case was examined individually, all equivocal HER2 cases had ISH, and follow-up of up to 60 months was available. There are also important limitations. First, the reference standard was manual scoring, which is itself imperfect; no consensus panel or molecular assay was available, so we could measure agreement but not accuracy. Second, the cohort was not consecutive, which limits the generalisability of the subtype frequencies and of the predictive values in Table 6. Third, the study involved one region and one digital platform, and the number of scoring pathologists was two; inter-observer and inter-laboratory variability were not assessed. Fourth, QuPath settings and the HER2 membrane cut-off were established within the same cohort without an independent test set, so performance may be overestimated. Fifth, the survival analyses were limited by 19 DFS events and nine deaths and by the absence of treatment data. Finally, several associations in this cohort, particularly the high nodal involvement in luminal A-like tumours, differ from published patterns and may reflect case selection.

Future directions

A multicentre Iraqi study using consecutive cases, several pathologists and an independent validation set would be the natural next step. Such a study should examine staining variation between laboratories, calibrate intensity thresholds, prospectively test rules for flagging borderline cases, and measure turnaround time and cost. Where feasible, comparison with ISH across all HER2-low tumours and with gene-expression profiling would allow accuracy, rather than agreement, to be assessed.

Conclusion

In this cohort, QuPath-assisted scoring of ER, PR, HER2 and Ki-67 agreed closely with manual assessment and reproduced the manual surrogate subtype in 95% of invasive breast carcinomas. Disagreement was confined to tumours near clinical thresholds: ER/PR expression of 1–10%, Ki-67 around 20%, and HER2 at the 0/1+ and 2+/3+ boundaries. QuPath may therefore be useful as a supporting and quality-assurance tool in Iraqi pathology laboratories, provided that borderline results are reviewed by a pathologist and HER2 equivocal cases are resolved by *in situ* hybridisation. These findings need confirmation in larger multicentre studies with independent validation before routine use.

CRedit

Author contributions: Sara Faez Hassan: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Supervision, and Correspondence. Shahad Mohammed Mahmood: Investigation, Data curation, Formal analysis, Software, Validation, Visualization, and Writing – review & editing. Raghad Hazem Alabbasy: Investigation, Histopathological interpretation, Validation, and Writing – review & editing. All authors reviewed and approved the final manuscript.

Acknowledgment

The authors thank the staff of the histopathology departments of the participating public hospitals and private clinics in Salah Al-Din for their assistance with case retrieval and slide preparation.

Ethical Approval

The study was approved by the Ethics Committee of the University of Samarra (approval no. 3774) and was conducted in accordance with the Declaration of Helsinki. The requirement for informed consent was waived by the Ethics Committee because archived material and anonymised records were used.

Informed Consent

Not applicable. The manuscript contains no identifiable patient information or images.

Funding

This research received no specific grant from any funding agency in the public, commercial or notfor-profit sectors.

Data Sharing Statement

The anonymised dataset analysed in this study is available from the corresponding author on reasonable request.

Publication Statement

All authors have agreed to the publication of this article.

Conflict of Interest

The authors declare no conflict of interest.

Similarity Check

It was applied by Ithenticate@.

Application of Artificial Intelligence (AI)

Not applicable.

Peer Review Process

It was performed.

About The License©

The author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229-63.
2. Jumaah AS, Shaker RA, Al-Ali Z, Altoriah KMJ, Al-Quzweni A, Jumaah SS, et al. Immunohistochemical breast cancer profiling among Iraqi women: molecular subtype classification, clinicopathology associations, and treatment-decision making tools: a crosssectional study. *Health Sci Rep.* 2025;8(5):e70553. doi:10.1002/hsr2.70553
3. Alwan NAS, Tawfeeq FN, Muallah FH. Breast cancer subtypes among Iraqi patients: identified by their ER, PR and HER2 status. *J Fac Med Baghdad.* 2017;59(4):303-7. doi:10.32007/jfacmedbagdad.59469
4. Shubbar YAMH. Correlation between different clinicopathological parameters and molecular subtypes of female breast carcinoma in south region of Iraq. *Wiad Lek.* 2023;76(1):97-107. doi:10.36740/WLek202301114
5. Chawsheen MA, Al-Naqshbandi AA, Ishaac RH, Kaka Sur FA. Epidemiological trends and molecular characterization of breast carcinoma in Erbil, Kurdistan Region of Iraq. *ARO Sci J Koya Univ.* 2025;13(1):86-93. doi:10.14500/aro.11860
6. Burstein HJ, Curigiano G, Thürlimann B, Weber WP, Poortmans P, Regan MM, et al. Customizing local and systemic therapies for women with early breast cancer: the St. Gallen International Consensus Guidelines for treatment of early breast cancer 2021. *Ann Oncol.* 2021;32(10):1216-35.
7. Loibl S, Poortmans P, Morrow M, Denkert C, Curigiano G. Breast cancer. *Lancet.* 2021;397(10286):1750-69.
8. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *J Clin Oncol.* 2020;38(12):1346-66.
9. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *J Clin Oncol.* 2018;36(20):2105-22.
10. Wolff AC, Somerfield MR, Dowsett M, Hammond MEH, Hayes DF, McShane LM, et al. Human epidermal growth factor receptor 2 testing in breast cancer: ASCO-College of American Pathologists guideline update. *J Clin Oncol.* 2023;41(22):3867-72.
11. Nielsen TO, Leung SCY, Rimm DL, Dodson A, Acs B, Badve S, et al. Assessment of Ki67 in breast cancer: updated recommendations from the International Ki67 in Breast Cancer Working Group. *J Natl Cancer Inst.* 2021;113(7):808-19.
12. Robertson S, Acs B, Lippert M, Hartman J. Prognostic potential of automated Ki67 evaluation in breast cancer: different hot spot definitions versus true global score. *Breast Cancer Res Treat.* 2020;183(1):161-75.
13. Modi S, Jacot W, Yamashita T, Sohn J, Vidal M, Tokunaga E, et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N Engl J Med.* 2022;387(1):9-20.

14. Tarantino P, Hamilton E, Tolaney SM, Cortes J, Morganti S, Ferraro E, et al. HER2-low breast cancer: pathological and clinical landscape. *J Clin Oncol*. 2020;38(17):1951-62.
15. Tarantino P, Viale G, Press MF, Hu X, Penault-Llorca F, Bardia A, et al. ESMO expert consensus statements (ECS) on the definition, diagnosis, and management of HER2-low breast cancer. *Ann Oncol*. 2023;34(8):645-59.
16. Fernandez AI, Liu M, Bellizzi A, Brock J, Fadare O, Hanley K, et al. Examination of low ERBB2 protein expression in breast cancer tissue. *JAMA Oncol*. 2022;8(4):607-10.
17. Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, et al. Introduction to digital image analysis in whole-slide imaging: a white paper from the Digital Pathology Association. *J Pathol Inform*. 2019;10:9.
18. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. *J Intern Med*. 2020;288(1):62-81.
19. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775-84.
20. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):16878.
21. Humphries MP, Maxwell P, Salto-Tellez M. QuPath: the global impact of an open source digital pathology system. *Comput Struct Biotechnol J*. 2021;19:852-9.
22. Rewcastle E, Skaland I, Gudlaugsson E, Fykse SK, Baak JPA, Janssen EAM. The Ki67 dilemma: investigating prognostic cut-offs and reproducibility for automated Ki67 scoring in breast cancer. *Breast Cancer Res Treat*. 2024;207(1):167-78.
23. Catteau X, Zindy E, Bouri S, Noël JC, Salmon I, Decaestecker C. Comparison between manual and automated assessment of Ki-67 in breast carcinoma: test of a simple method in daily practice. *Technol Cancer Res Treat*. 2023;22:15330338231169603.
24. Wu S, Yue M, Zhang J, Li X, Li Z, Zhang H, et al. The role of artificial intelligence in accurate interpretation of HER2 immunohistochemical scores 0 and 1+ in breast cancer. *Mod Pathol*. 2023;36(3):100054.
25. Dosh ZN, Muslim LM, Hasan MMSM, Al Janabi A. Correlation between primary tumor sizes with prognostic markers in breast carcinoma in Iraqi women: immunohistochemical study. *Wiad Lek*. 2022;75(11 pt 2):2771-8. doi:10.36740/WLek202211210
26. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527.
27. WHO Classification of Tumours Editorial Board. Breast tumours. 5th ed. Lyon: International Agency for Research on Cancer; 2019. (WHO classification of tumours series; vol. 2).
28. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*. 1991;19(5):403-10.
29. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, et al., editors. *AJCC cancer staging manual*. 8th ed. New York: Springer; 2017.
30. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann Oncol*. 2015;26(2):259-71.
31. McCarty KS Jr, Miller LS, Cox EB, Konrath J, McCarty KS Sr. Estrogen receptor analyses. Correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Arch Pathol Lab Med*. 1985;109(8):716-21.
32. Allred DC, Harvey JM, Berardo M, Clark GM. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Mod Pathol*. 1998;11(2):155-68.
33. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30-46.
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-63.
35. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45(1):255-68.
36. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-10.
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-74.
38. Youden WJ. Index for rating diagnostic tests.

- Cancer. 1950;3(1):32-5.
39. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503-10.
40. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
41. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191-4.
42. Yoder R, Kimler BF, Staley JM, Schwensen K, Wang YY, Finke K, et al. Impact of low versus negative estrogen/progesterone receptor status on clinico-pathologic characteristics and survival outcomes in HER2-negative breast cancer. NPJ Breast Cancer. 2022;8(1):80.
43. Paakkola NM, Karakatsanis A, Mauri D, Foukakis T, Valachis A. The prognostic and predictive impact of low estrogen receptor expression in early breast cancer: a systematic review and metaanalysis. ESMO Open. 2021;6(6):100289Discussion

Supplementary Material

Table S1. HER2 immunohistochemical and in situ hybridisation findings in cases scored as equivocal (2+) by manual assessment (n = 15).

Case	Manual subtype	HER2 IHC (manual)	HER2 IHC (QuPath)	Membrane-positive cells, QuPath (%)	ISH result	HER2/CEP17 ratio	Mean HER2 copy number	Final HER2 status
BC010	Triple-negative	2+	3+	79.5	Not amplified	1.67	3.2	Negative
BC012	HER2-positive (non-luminal)	2+	2+	66.3	Amplified	4.00	6.1	Positive
BC013	Luminal B-like (HER2+)	2+	2+	51.7	Amplified	4.18	7.4	Positive
BC016	Triple-negative	2+	2+	42.7	Not amplified	1.10	2.1	Negative
BC024	HER2-positive (non-luminal)	2+	2+	15.0	Amplified	4.85	10.8	Positive
BC034	HER2-positive (non-luminal)	2+	2+	57.5	Amplified	2.81	4.6	Positive
BC035	HER2-positive (non-luminal)	2+	2+	37.0	Amplified	5.44	9.2	Positive
BC036	Luminal B-like (HER2+)	2+	2+	38.5	Amplified	2.55	10.8	Positive
BC060	HER2-positive (non-luminal)	2+	3+	100.0	Amplified	4.06	11.7	Positive
BC068	HER2-positive (non-luminal)	2+	2+	51.2	Amplified	2.95	12.0	Positive
BC093	HER2-positive (non-luminal)	2+	2+	25.5	Amplified	2.19	7.1	Positive
BC094	Triple-negative	2+	2+	35.1	Not amplified	1.39	1.8	Negative
BC096	Luminal B-like (HER2+)	2+	2+	39.4	Amplified	2.74	7.6	Positive
BC104	HER2-positive (non-luminal)	2+	2+	36.1	Amplified	3.83	8.6	Positive
BC119	HER2-positive (non-luminal)	2+	2+	69.4	Amplified	3.04	7.0	Positive

*Final HER2 status was assigned according to the ASCO/CAP 2018 guideline, in which the ISH result determines the final status of IHC 2+ tumours. Abbreviations: CEP17, chromosome enumeration probe 17; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; ISH, *in situ* hybridisation. Source: Own authorship.

Table S2. Pre-operative haematological and serum biochemical parameters stratified by manual surrogate molecular subtype (n = 120).

Parameter	Total (n = 120)	Luminal A-like (n = 30)	Luminal B-like (HER2-) (n = 23)	Luminal B-like (HER2+) (n = 7)	HER2-positive (non-luminal) (n = 30)	Triple-negative (n = 30)	P value
ANC, $\times 10^9/L$	3.88 (2.77–5.47)	3.82 (3.09–4.67)	4.39 (3.00–5.70)	2.79 (2.52–5.94)	3.06 (2.39–4.54)	4.55 (2.90–6.20)	0.144 ^a
ALC, $\times 10^9/L$	1.94 (1.62–2.33)	1.96 (1.65–2.33)	1.83 (1.62–2.17)	2.22 (1.54–2.38)	1.92 (1.60–2.25)	2.05 (1.69–2.46)	0.713 ^a
NLR	1.96 (1.51–2.79)	1.94 (1.63–2.60)	2.31 (1.75–3.04)	2.01 (1.64–2.50)	1.60 (1.32–2.10)	2.48 (1.52–3.01)	0.095 ^a
CA 15-3, U/mL	29.4 (20.9–39.2)	33.2 (23.4–46.7)	27.5 (20.6–36.3)	45.7 (27.4–53.5)	26.9 (18.4–35.8)	30.4 (21.2–36.1)	0.173 ^a
CEA, ng/mL	2.65 (1.90–3.90)	2.75 (1.95–3.90)	2.50 (1.80–3.30)	2.70 (2.45–5.25)	3.05 (2.30–4.33)	2.30 (1.52–3.60)	0.434 ^a
ALT, U/L	23.0 (16.8–27.0)	21.0 (17.5–28.2)	24.0 (20.5–26.5)	14.0 (13.0–25.5)	24.5 (19.2–28.5)	20.0 (14.2–24.8)	0.231 ^a
AST, U/L	20.0 (15.0–26.0)	15.5 (12.0–22.8)	22.0 (18.0–30.5)	26.0 (23.0–33.0)	19.0 (16.0–27.5)	19.5 (15.0–23.0)	0.005^a
Creatinine, mg/dL	0.79 (0.68–0.88)	0.77 (0.67–0.86)	0.80 (0.62–0.90)	0.81 (0.65–0.85)	0.81 (0.72–0.85)	0.79 (0.72–0.90)	0.982 ^a

Data are median (interquartile range). ^a Kruskal–Wallis H test. P values in bold are statistically significant (P < 0.05). Abbreviations: ALC, absolute lymphocyte count; ALT, alanine aminotransferase; ANC, absolute neutrophil count; AST, aspartate aminotransferase; CA 15-3, cancer antigen 15-3; CEA, carcinoembryonic antigen; NLR, neutrophil-to-lymphocyte ratio. Source: Own authorship.