The authors declare no financial disclosure or conflict of interest that relate to the research presented in this poster

UK NEQAS
Immunocytochemistry
& In-Situ Hybridisation

# Inter-Rater Agreement of HER2 and HER2-low Scores in Breast Cancer Between the Visiopharm Digital Image Analysis HER2 Application and a Group of Expert Pathologists

USCAP
114th Annual General Meeting
22-27th March 2025. Boston, Massachusetts, USA
Poster ID 485. Program ID: 213

**Andrew Dodson[1], Abeer Shaaban[2], Lila A Zabaglo[1], Suzanne Parry[1]**

1. UK National External Quality Assessment Scheme for Immunocytochemistry and In-Situ Hybridisation, London, UK. 2. University Hospitals Birmingham NHS Foundation Trust Queen Elizabeth Hospital, Birmingham, UK. Corresponding author's email address: adodson@ukneqasiccish.org

## INTRODUCTION

Improvements in the accuracy and reproducibility of HER2 scoring among pathologists working in the breast cancer HER2-low setting remains a largely unmet challenge [1]. The application of artificial intelligence (AI)-powered digital image analysis (DIA) may help to meet these needs by improving the consistency of HER2 expression measurement.

We used published inter-observer HER2 concordance data generated in a study where a group of 16 pathologists with recognised expertise in the area undertook the HER2 status assessment of 'real-world' breast cancer core biopsies, the majority of which had previously been shown to express HER2 in the low range [2].

This data was compared with that produced by DIA of the same sample set using a commercially available application (APP) for HER2 evaluation [3].

The aim of the study was to produce information about the level of agreement between shown between a market-leading DIA APP and a large group of expert raters as a robust measure of the feasibility of implementing DIA in the assessment of HER2 status in the clinical setting.

## MATERIALS and METHODS

The study cases comprised 50 breast cancer samples selected from the routine clinical caseload of a single institution. They were enriched for tumours expressing HER2 in the HER2-low range (0, 1+ and 2+) using the PATHWAY 4B5 assay (Roche, Indianapolis, USA). **Table 1** gives details of the original HER2 scores.

| Original HER2 Score | FISH Status | Count (N) | Proportion (%) | Description of Staining* | HER2 Category |
|---|---|---|---|---|---|
| 0 | N/A | 1 | 2% | No staining | Negative |
|  | N/A | 19 | 38% | Faint/barely perceptible incomplete or complete in <10%, or weak incomplete staining in <10% | Negative |
| 1+ | N/A | 13 | 26% | Weak complete in ≤10%, or weak incomplete in >10%, or faint/barely perceptible in >10% | HER2-low |
| 2+ | FISH -ve | 10 | 20% | Weak to moderate complete in >10%, or moderate to strong complete in ≤10% | HER2-low |
|  | FISH +ve | 2 | 4% |  | Positive |
| 3+ | N/A | 5 | 10% | Strong (intense and uniform) complete in >10% | Positive |

**Table 1. Original clinical HER2 status details.** The descriptions of staining are according to published UK guidelines [4]. N/A = Not Applicable; FISH –ve = Not amplified for *HER2* gene by fluorescent *in-situ* hybridisation; FISH +ve = Amplified for *HER2* gene by fluorescent *in-situ* hybridisation.

Whole slide images were obtained at x40 magnification (Aperio AT2 slide scanner, Leica Biosystems, California, USA). Sixteen expert breast pathologists based in the UK and Republic of Ireland individually assessed them for HER2 expression following UK HER2 guidelines [4] in a study looking at inter-rater concordance [5].

The same set of images were used in this study to compare the performance of a digital image analysis application designed to assess HER2 expression in this clinical setting (VP HER2 APP, #10185), Visiopharm, Hoersholm, Denmark) with that of the expert pathologist group.

*Statistical Analyses*

Inter-rater agreement was assessed using Fleiss' multiple-rater kappa statistic.

## MATERIALS and METHODS (continued)

Cohen's weighted kappa (CW-kappa) coefficient was used to assess the agreement between individual raters' scores (pathologists and VP HER2 APP) and the consensus scores. This was done for the all-cases group, and for the group of cases where the level of agreement (LoA) was at least 0.80 (almost perfect).

The CW-kappa coefficient was also calculated for the HER2-low cohort, defined as cases where the consensus HER2 score was 1+ or 2+ (ISH-negative). This again was done for the whole cohort and the group of cases showing an LoA of at least 0.8.

Data were collated using Excel (Office 365, Microsoft, Washington, USA). Statistical analyses were done using SPSS (Version 29.0.2.0 (20), IBM, New York, USA) and Prism (Version 10.3.0, GraphPad, Massachusetts, USA).

## RESULTS

*Pathologists' consensus score versus VP HER2 APP score*

Comparing pathologists' consensus VP HER2 APP scores, 36/49 (73.5%) cases agreed and 13/49 (26.5%) disagreed (one case was excluded because a pathologist' consensus score could not be assigned). Of the thirteen discordant cases, nine (69.2%) occurred between scores that were assigned to cases in the poor or low LoA categories and four (30.8%) to cases in the high LoA category.

*Fleiss' multiple-rater kappa statistic*

The overall agreement between all pathologists rating all cases was 0.433 (moderate agreement), when the scores produced by the VP HER2 APP were included in the analysis, the result was unchanged. **Table 2** gives full details.

| Rating Category | Pathologists |  | Pathologists' + VP HER2 APP |  |
|---|---|---|---|---|
|  | Kappa | Agreement | Kappa | Agreement |
| Overall | **0.433** (0.417 - 0.449) | Moderate | **0.433** (0.418 - 0.448) | Moderate |
| HER2 0 | **0.437** (0.411 - 0.462) | Moderate | **0.444** (0.420 - 0.468) | Moderate |
| HER2 1+ | **0.292** (0.267 - 0.317) | Fair | **0.296** (0.272 - 0.319) | Fair |
| HER2 2+ | **0.431** (0.406 - 0.456) | Moderate | **0.424** (0.400 - 0.447) | Moderate |
| HER2 3+ | **0.803** (0.777 - 0.828) | Almost perfect | **0.808** (0.784 - 0.831) | Almost perfect |

**Table 2. Fleiss' statistic results.** Fleiss' kappa is in bold type, the figures in brackets are the 95% confidence intervals.

Paired results for agreement within individual HER2 categories were closely similar to each other. Paired results for the HER2 0 and the HER2 2+ categories were 0.437:0.444 and 0.431:0.424 respectively (both moderate agreement). However, the paired results for the HER2 1+ category were substantially lower at 0.292:0.296 (fair agreement), and those for scores in the HER2 3+ category substantially higher at 0.803:0.808 (almost perfect).

*Cohen's Weighted Kappa*

When the whole set of 50 cases was considered, the CW-kappa scores for the 16 pathologists together with the VP HER2 APP had a range between 0.412 and 0.854.

The VP HER2 APP was ranked at 12th out of the 17 raters, with a CW-kappa score of 0.638, which is indicative of substantial agreement. See Table 3A for complete data set of this analysis.

## RESULTS (continued)

| Pairing | Kappa | LoA |
|---|---|---|
| C - P11 | 0.854 (0.741 - 0.968) | Almost Perfect |
| C - P14 | 0.824 (0.675 - 0.973) | Almost Perfect |
| C - P16 | 0.815 (0.675 - 0.956) | Almost Perfect |
| C - P04 | 0.757 (0.610 - 0.903) | Substantial |
| C - P05 | 0.749 (0.609 - 0.890) | Substantial |
| C - P07 | 0.748 (0.605 - 0.891) | Substantial |
| C - P02 | 0.742 (0.598 - 0.886) | Substantial |
| C - P01 | 0.710 (0.561 - 0.858) | Substantial |
| C - P09 | 0.709 (0.556 - 0.862) | Substantial |
| C - P06 | 0.680 (0.502 - 0.859) | Substantial |
| C - P12 | 0.651 (0.494 - 0.809) | Substantial |
| **C - HER2 APP** | **0.638 (0.454 - 0.821)** | **Substantial** |
| C - P13 | 0.580 (0.392 - 0.768) | Moderate |
| C - P10 | 0.557 (0.409 - 0.704) | Moderate |
| C - P03 | 0.524 (0.358 - 0.689) | Moderate |
| C - P08 | 0.500 (0.334 - 0.666) | Moderate |
| C - P15 | 0.412 (0.248 - 0.576) | Moderate |

**Table 3A. All cases (N = 50)**

| Pairing | Kappa | LoA |
|---|---|---|
| C - P02 | 1.000 (1.000 – 1.000) | Perfect |
| C - P11 | 1.000 (1.000 – 1.000) | Perfect |
| C - P14 | 1.000 (1.000 – 1.000) | Perfect |
| C - P06 | 0.958 (0.876 - 1.040) | Almost Perfect |
| C - P07 | 0.958 (0.876 - 1.040) | Almost Perfect |
| C - P04 | 0.919 (0.808 - 1.029) | Almost Perfect |
| C - P05 | 0.916 (0.800 - 1.033) | Almost Perfect |
| **C - HER2 APP** | **0.916 (0.800 - 1.033)** | **Almost Perfect** |
| C - P01 | 0.885 (0.766 - 1.005) | Almost Perfect |
| C - P16 | 0.885 (0.764 - 1.005) | Almost Perfect |
| C - P09 | 0.851 (0.721 - 0.982) | Almost Perfect |
| C - P10 | 0.795 (0.645 - 0.945) | Substantial |
| C - P12 | 0.789 (0.647 - 0.931) | Substantial |
| C - P13 | 0.742 (0.518 - 0.965) | Substantial |
| C - P03 | 0.741 (0.547 - 0.935) | Substantial |
| C - P08 | 0.683 (0.506 - 0.861) | Substantial |
| C - P15 | 0.664 (0.454 - 0.874) | Substantial |

**Table 3B. High Agreement cases (N = 24)**

| Pairing | Kappa | LoA |
|---|---|---|
| C - P11 | 0.823 (0.462 - 0.796) | Almost Perfect |
| C - P14 | 0.813 (0.541 - 0.873) | Almost Perfect |
| C - P16 | 0.748 (0.251 - 0.609) | Substantial |
| C - P05 | 0.721 (0.480 - 0.840) | Substantial |
| C - P02 | 0.707 (0.563 - 0.879) | Substantial |
| C - P07 | 0.678 (0.337 - 0.760) | Substantial |
| C - P04 | 0.660 (0.505 - 0.850) | Substantial |
| C - P01 | 0.629 (0.194 - 0.538) | Substantial |
| C - P09 | 0.618 (0.444 - 0.792) | Substantial |
| C - P06 | 0.548 (0.281 - 0.569) | Moderate |
| C - P12 | 0.543 (0.692 - 0.954) | Moderate |
| **C - HER2 APP** | **0.535 (0.368 - 0.719)** | **Moderate** |
| C - P13 | 0.512 (0.301 - 0.723) | Moderate |
| C - P03 | 0.430 (0.652 - 0.974) | Moderate |
| C - P10 | 0.425 (0.136 - 0.453) | Moderate |
| C - P08 | 0.366 (0.573 - 0.924) | Fair |
| C - P15 | 0.295 (0.318 - 0.753) | Fair |

**3C. All HER2-low cases (N = 44)**

| Pairing | Kappa | LoA |
|---|---|---|
| C - P02 | 1.000 (1.000 – 1.000) | Perfect |
| C - P11 | 1.000 (1.000 – 1.000) | Perfect |
| C - P14 | 1.000 (1.000 – 1.000) | Perfect |
| C - P07 | 0.933 (0.802 - 1.064) | Almost Perfect |
| C - P06 | 0.931 (0.797 - 1.065) | Almost Perfect |
| C - P04 | 0.871 (0.708 - 1.034) | Almost Perfect |
| C - P05 | 0.868 (0.687 - 1.048) | Almost Perfect |
| **C - HER2 APP** | **0.860 (0.671 - 1.049)** | **Almost Perfect** |
| C - P16 | 0.828 (0.668 - 0.988) | Almost Perfect |
| C - P01 | 0.821 (0.643 – 1.00) | Almost Perfect |
| C - P09 | 0.771 (0.582 - 0.961) | Substantial |
| C - P13 | 0.712 (0.453 - 0.971) | Substantial |
| C - P10 | 0.693 (0.518 - 0.869) | Substantial |
| C - P12 | 0.683 (0.488 - 0.878) | Substantial |
| C - P03 | 0.600 (0.324 - 0.876) | Moderate |
| C - P08 | 0.519 (0.286 - 0.751) | Moderate |
| C - P15 | 0.506 (0.234 - 0.779) | Moderate |

**3D. HER2-low High Agreement cases (N = 20)**

**Table 3 (A-D). CW-kappa scores.** P01 to P16 represent each of the 16 pathologists, C = Consensus score. HER2 APP = VP HER2 APP (truncated to conserve space).

Table 3B shows results for the 24 case sub-set in which at least 13 out of the 17 raters agreed on the consensus score (high agreement). The CW-kappa score range was 1.000 to 0.664 and the VP HER2 APP scored 0.916 indicating almost perfect agreement (ranked 8th out of 17).

When only HER2-low cases were included (N = 44 cases), the CW-kappa score range was 0.823 to 0.295. The VP HER2 APP score was 0.535 indicating moderate agreement, with a ranking of 12th out of 17. See Table 3C.

Restricting the analysed set to HER2-low cases where there was high agreement on the consensus HER2 score (N = 20 cases), the CW-kappa score range was 1.000-0.506. The VP HER2 APP score was 0.860 indicating almost perfect agreement (ranked 8th out of 17 raters). See Table 7D.

## DISCUSSION

A notable strength of the work presented here lies in the robust way in which the ground-truth HER2 scores were derived. By taking a consensus of scores produced by a large number (16) of clinically active specialist breast pathologists we can be confident of the validity of the data. Moreover, it makes it possible to assess the degree to which that confidence should be ascribed by taking into account the agreement level between pathologists on each case individually.

Thus, we have been able to show that the scores derived by the VP HER2 APP are extremely well aligned to those of the pathologists where there is good agreement amongst the pathologist as to what those scores are. And that this is true in the 'standard' HER2 assessment setting and for HER2-low cases.

The case-set was deliberately enriched for challenging cases i.e. those showing heterogeneous HER2 expression, cytoplasmic staining and HER2 expression closely spanning the 10% cut-point; all of which have been well-documented in the literature as leading to poor agreement. This is clearly reflected in the moderate agreement score of 0.433 returned by the Fleiss' kappa analysis. Two of these features, namely heterogeneity and scores close to cut-points are especially confounding to scoring methodologies that rely on estimation, which is the method used universally by pathologists in this setting. Here counting methods as used by DIA applications have been shown to be consistently more accurate and reproducible. Thus, the mid-range performance achieved by the VP HER2 APP when cases with poor levels of agreement were included do not indicate a failing of the software but more probably of the result it is being measured against.

## CONCLUSIONS

☑ **When measured against well-characterised cases DIA produces HER2 scores which are better aligned to the consensus than the majority of highly qualified breast pathologists.**

☑ **This finding applies equally in HER2 'standard' assessment and in the newly introduced setting of HER2-low assessment.**

## REFERENCES

1. Tarantino, P., *et al.*, ESMO expert consensus statements (ECS) on the definition, diagnosis, and management of HER2-low breast cancer. Ann Oncol, 2023. 34(8): p. 645-659.
2. Zaakouk, M., *et al.*, Concordance of HER2-low scoring in breast carcinoma among expert pathologists in the United Kingdom and the republic of Ireland - on behalf of the UK national coordinating committee for breast pathology. Breast, 2023. 70: p. 82-91.
3. Visiopharm. HER2 APP, Breast Cancer (#10185) [cited 2024 27/08/2024]; Available from: https:// visiopharm.com/app-center/app/her2-app-breast-cancer-ruo.
4. Rakha, E.A., et al., UK recommendations for HER2 assessment in breast cancer: an update. J Clin Pathol, 2023. 76(4): p. 217-227.

## ACKNOWLEDGEMENTS

The authors acknowledge: the staff at Visiopharm for their technical help with the set up of the VP HER2 APP; Mohamed Zaakouk (Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK and Cancer Pathology, National Cancer Institute, Cairo University, Cairo, Egypt) for his advice on the statistical analyses used in the original paper; the UK national coordinating committee for breast pathology for permission to use the data contained in the original paper.