Interobserver variation in the assessment of HER2 low expression in breast cancer: can we improve by adjusting criteria? An international interobserver study.

Ximena Baez Navarro (✉ x.baeznavarro@erasusmc.nl)
Erasmus MC

Mieke Rosalie van Bockstal
Cliniques universitaires Saint-Luc

Diënna Nawawi
diennanawawi@gmail.com

Glenn Broeckx
University Hospital Antwerp https://orcid.org/0000-0002-7543-8061

Cecile Colpaert
AZ Turnhout Campus Sint-Jozef https://orcid.org/0000-0002-9807-0917

Shusma Charlane Doebar
Spaarne Gasthuis

Marieke C.H. Hogenes
Laboratory Pathology East Netherlands

Esther Koop
Gelre Hospital

Kathleen Lambein
Leuven University Hospitals

Dieter J.E. Peeters
General Hospital Sint-Maarten, CellCarta NV

Renata Sinke
Pathan BV

Johannes Bastiaan van Brakel
Skåne University Hospital

José van der Starre-Gaal
Isala Hospitals

Bert Van der Vegt
UMCG https://orcid.org/0000-0002-2613-1506

Koen Van de Vijver
Ghent University Hospital https://orcid.org/0000-0002-2026-9790
Celien P.H. Vreuls
   Utrecht University Medical Center
Willem Vreuls
   Canisius Wilhelmina Hospital
Pieter Westenend
   Laboratory for Pathology Dordrecht
Carolien van Deurzen
   Erasmus Medical Center

**Article**

**Keywords:** HER2 low, interobserver agreement, breast cancer

**Posted Date:** July 14th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1807403/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

The classification of Human Epidermal Growth Factor Receptor 2 (HER2) expression is optimized to detect HER2-amplified breast cancer (BC). However, novel HER2-targeting agents are also effective for BCs with low levels of HER2. This raises the question whether the current guidelines for HER2-testing are sufficiently reproducible to identify HER2 low BC. The aim of this multicenter international study was to assess the interobserver agreement of HER2 low scoring according to the current American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) guidelines. Furthermore, we evaluated whether the agreement improved by redefining immunohistochemistry (IHC) scoring criteria, or by adding fluorescent in situ hybridization (FISH).

We performed a two-round study of 105 non-amplified BCs. During the first assessment, sixteen pathologists used the latest version of the ASCO/CAP guidelines. After a consensus meeting, the same pathologists scored the same digital slides using modified IHC scoring criteria based on the 2007 ASCO/CAP guideline, and an extra ‘ultralow’ category was added.

Overall, the interobserver agreement was limited (4.7% of cases with 100% agreement) in the first round, but this improved by clustering IHC categories. In the second round, the highest reproducibility was seen when comparing IHC 0 versus the grouped cluster of ultralow/1+/2+ (74.3% of cases with 100% agreement). FISH results were not statistically different between HER2 0 and HER2 low cases, regardless of the IHC criteria used.

In conclusion, our study suggests that the modified 2007 ASCO/CAP criteria were more reproducible to distinguish HER2 0 from HER2 low cases as compared to the 2018 ASCO/CAP criteria. However, the reproducibility was still moderate, which was not improved by adding FISH. This could lead to suboptimal selection of patients eligible for novel HER2-targeting agents. There is a need for clearer, more reproducible IHC definitions, training and/or development of more accurate methods to detect HER2 low BC.

Introduction

For more than two decades, overexpression of the Human Epidermal Growth Factor Receptor 2 (HER2) has been recognized as a negative prognostic biomarker and therapeutic target in invasive breast cancer (BC)\(^1\). International guidelines were developed to standardize and optimize HER2-testing, since only those patients with HER2-amplification were likely to respond to HER2-targeted treatment\(^2\,\,3\).

During the updates of the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) guidelines, considerable changes were made in the immunohistochemistry (IHC) cut-off points\(^4\,\,5\,\,6\). According to the first version of the guideline, published in 2007, BC was categorized as IHC 0 (no staining), IHC 1+ (weak, incomplete membrane staining in any proportion of tumor cells or weak complete staining in < 10% of cells); IHC 2+ (equivocal), or IHC 3+ (uniform intense membrane staining of the HER2 protein in > 30% of invasive tumor cells). For treatment considerations, HER2 was defined as
positive in case of IHC 2+ with amplification after reflex testing or IHC 3+\textsuperscript{5}. In the updated versions of the guidelines, published in 2013 and 2018, the definitions of the IHC scoring were modified. The cut-off point for IHC 3+ was changed to complete, intense staining of >10% of the tumor cells instead of 30%. The definition of IHC 0 was adapted to either no staining or incomplete membrane staining that is faint/barely perceptible and within <10% of the invasive tumor cells\textsuperscript{4,6}. This change resulted in a substantial increase of IHC 0 cases according to the 2013 version compared to the 2007 guidelines\textsuperscript{7,8}.

In recent years, the development of an emerging group of HER2-targeted drugs, the so-called antibody drug conjugates (ADCs), has led to a different view on this historical HER2 classification system. In ongoing clinical trials, these drugs have demonstrated efficacy and safety against HER2 positive metastatic BC\textsuperscript{9,10}. Moreover, due to its favorable drug to antibody ratio and its bystander killing effect, ADCs such as trastuzumab-deruxtecan (T-DXd) have also proved to have a significant antitumor action in BCs with a low expression level of HER2 (HER2 low), comprising IHC 1+ and IHC 2+ cases without amplification\textsuperscript{9–15}. The Destiny-Breast06 clinical trial is currently evaluating the effect of T-DXd in BC with even lower levels of HER2 expression (IHC >0, <1+), the so-called HER2 ultralow category. These HER2 ultralow cases would have been classified as IHC 1+ according to the ASCO/CAP guideline of 2007, but as IHC 0 according to the 2013 and 2018 editions. Obviously, these novel treatment options raise the question whether our current method of IHC-testing, historically optimized to detect HER2 amplified BC, is sufficiently robust and reproducible to discern HER2 low BC as well.

Previous studies have evaluated the HER2 IHC interobserver reproducibility using the 2013 and 2018 versions of the ASCO/CAP guidelines with inconsistent results\textsuperscript{16–19}. In a recent study by Fernandez et al., data from around 1400 laboratories around the world were collected\textsuperscript{19}. The lowest IHC agreement was found between HER2 0 vs HER2 1+ (less than 70% agreement). An interobserver analysis of 92 cases graded as IHC 0 or IHC 1, resulted in a 90% agreement (17 out of 18 pathologists) in only 26% of the cases (24 out of 92)\textsuperscript{19}.

In a study by Schettini et al. five specialized observers evaluated 100 BC cases using the 2018 guidelines. Overall, thirty-five cases out of a total number of 100 samples were discordant, from which the highest disagreement was found between IHC 1+ vs. IHC 0 (n = 15)\textsuperscript{17}. Interestingly, older studies that used the cut-off points of the 2007 version of the ASCO/CAP guideline, seem to perform better to differentiate IHC 1+ from 0\textsuperscript{20–22}. Umemura et al. reported a good general agreement in 14 out of 20 cases evaluated by seven observers. In this study IHC 2+ and 3+ were the cases with low concordance (55–64%), while cases with HER2 0 and 1+ showed a high concordance (90–100%)\textsuperscript{20}. Additionally, Thomson et al. assessed 127 cases scored by three observers and reported a high interobserver agreement (kappa = 77–95.6) for IHC 0 and 3+ cases, while it was generally poor (kappa = 32.8–59.1) for cases with 1+ and 2+ staining\textsuperscript{21}.

Thus, we hypothesized that the IHC scoring criteria according to the 2007 version of the ASCO/CAP recommendations for HER2 testing are likely to be more reproducible, in particular to distinguish between
IHC 0 and 1+, compared to the criteria from 2013 and 2018. The primary objective of this multicenter international study was to quantify the interobserver agreement of HER2 low scoring according to the current guidelines, and to evaluate whether we could improve this agreement by redefining some of the current IHC scoring criteria or by adding in situ hybridization.

Materials And Methods

Study design and HER2 IHC scoring

We performed a multi-institutional study with two rounds of scoring, including 105 needle biopsies with invasive BC that were scored by sixteen pathologists. These cases were a consecutive series of archived BC cases diagnosed in 2019 with a negative HER2 status according to the original pathology report using the 2018 ASCO/CAP guidelines. Tissue sections were immunostained with the 4B5 HER-2/neu antibody using an automatic immunostainer (Ventana BenchMark Ultra, Roche, Indianapolis, USA). All slides were scanned with the Nanozoomer 2.0-HT (Hamamatsu Photonics, Shizuoka, Japan) which enabled Z-stacking, and they were uploaded to Slide Score B.V. (version 1.2-2022-05-24T15:37:11 (Netherlands Cancer Institute, Amsterdam, The Netherlands), which allowed zooming to a high magnification (objective × 40). This program blinded the case numbers and randomized the slides for the participants. The use of coded leftover patient material is in accordance of the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands.

In the first round of scoring, sixteen pathologists scored a total number of 105 slides according to the ASCO/CAP guidelines of 2018 as either IHC 0, 1+, 2+, 3+, or ‘too few tumor cells’. Once the results from the first round were complete and analyzed, all pathologists participated in an online consensus meeting. This meeting included the following topics: presentation of the IHC criteria of the ASCO/CAP guidelines of 2007, 2013 and 2018, presentation and discussion of slides with good and poor agreement, and proposal of the criteria to use in the second round of scoring. Figure 1 provides an overview of our study design. For the second round of scoring, all pathologists scored the same slides again after randomization of the slide order in SlideScore. The IHC scoring criteria used in the second round of scoring were modified in accordance with the 2007 ASCO/CAP guidelines, supplemented with a separate category of HER2-ultralow, as described in Fig. 1. This resulted in the following categories: IHC 0, ultralow, 1+, 2+, 3 + or ‘too few tumor cells’. For this second round, a document describing the new criteria and examples of each IHC category was sent to the pathologists.

HER2 ISH

To complement this study, fluorescence in situ hybridization (FISH) was performed on all 105 cases using the BenchMark Ultra (Roche, Indianapolis, USA). For detection of HER2, the ZytoLight SPEC ERBB2/CEN 17 Dual Color Probe (Zytovision, Bremerhaven, Germany) was used according to the manufacturer's protocol. Signal numbers for the chromosomal region 17q12-q21.1 harboring the HER2 gene (labeled with SPEC ERBB2, ZytoLight) and the alpha satellite centromeric region of chromosome 17 (CEP17)
(labeled with CEN 17, ZytoLight) were counted in at least 30 invasive tumor cells, and the ratio of the HER2/CEP17 signal numbers was calculated. The analysis of the FISH tests was performed by one observer. A second observer checked and approved the tests’ interpretation.

**Statistical analysis**

The scoring option of ‘too few tumor cells’ was considered as missing data and excluded from the statistical analysis. Krippendorff’s alpha test was used to estimate the interobserver agreement in round one and two\(^24\). The cut off points of interobserver reliability for this test were: \(\alpha\) value < 0.67 (low), between 0.67 and 0.8 (moderate), and > 0.8 (high).

To analyze the correlation between the IHC scores and the FISH results (HER2/CEP17 ratio and the average number of HER2 copies/cell), we used the IHC score that was the most frequently scored by the pathologists in the first and second round. We calculated the mean, median and range for continuous variables. The Shapiro-Wilk test was used to check normal distribution. Kruskal-Wallis and Mann-Whitney tests were used to study the association between the FISH results and the IHC score, as these data were not normally distributed. A \(p\)-value < 0.05 was considered significant, except for the post hoc Mann Whitney tests after the Kruskal-Wallis tests, where a \(p\)-value of < 0.016 was used (i.e. Bonferroni correction for multiple testing).

Statistical analyses were performed in SPSS (IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 28.0. Armonk, New York, USA). Additionally, a macro was downloaded from http://afhayes.com/spss-sas-and-r-macros-and-code.html to perform the Krippendorff’s alpha test in SPSS.

**Results**

**Interobserver agreement first round**

In both scoring rounds, each of the sixteen pathologists evaluated 105 BC cases. The results of the first round are presented in supplementary table 1. A total number of 6 cases were scored as “too few tumor cells” and were considered as missing data, resulting in 1674 scores. Overall, 22.8% (383 of 1674) of cases were scored as IHC 0, 44.6% (749 of 1674) were scored as 1+, 31.3% (526 of 1674) were scored as 2+ and 1% (16 of 1674) were scored as IHC 3. The Krippendorff's alpha to estimate the interobserver agreement from this first scoring round showed low agreement for the categories IHC 0, 1+, 2+ and 3+ (\(\alpha=0.63\)). A consistently low agreement (\(\alpha=0.56\)) was found when we grouped the categories 1+ and 2+ together.

Furthermore, the percentage of complete agreement, in which all sixteen pathologists grouped the tumors in the same IHC category (0, 1+, 2+, 3+), was achieved in only 4.7% (5 of 105) of the cases. If we considered an agreement of 87.5% as an acceptable consensus (14 out of 16 pathologists), 30.4% (32 of 105) of the cases were grouped in the same category. Table 1 presents the level of agreement using...
different combinations of IHC-clusters. After clustering IHC 1+ and IHC 2+ together, the percentage of agreement increases to 33.3% (35 of 105) for a complete agreement (all pathologists) and 76.2% (80 of 105) for an agreement between 14 out of 16 pathologists, which highlights that the distinction of 1+ cases from 2+ cases is problematic.

**Consensus meeting**

During the online consensus meeting after the first round of scoring, a representative subset of cases with good and poor interobserver agreement were discussed. Figures 2A and 2B present two examples of cases with low agreement. Several issues caused doubt when distinguishing between IHC 0 and 1+ including difficulty in discriminating non-specific staining from membranous staining. In addition, the application of the term ‘barely perceptible’ was considered highly subjective by most participants. Another difficulty involved the correct estimation of 10% of tumor cells. The use and the resolution of digital slides was also considered as an obstacle by some pathologists. Besides, several pathologists were not used to evaluate the 4B5 HER2 antibody, since they use another antibody in their laboratory. Finally, a small subset of cases was scored as 3+ in this first round, since some pathologists did not realize that only HER2 negative cases were included. Considering all the aforementioned difficulties, we performed a second round using modified criteria based on the 2007 ASCO/CAP guideline (figure 1), as described in detail in the Materials & Methods section.

**Interobserver variation second round**

The results of the second round are presented in supplementary table 2. Three cases were scored as “too few tumor cells”, resulting in 1677 scores. Overall, 16.4% (275 of 1677) were scored as IHC 0, 15.4% (259 of 1677) as ultralow, 42.6% (715 of 1677) as 1+, and 25.5% (428 of 1677) as 2+. No cases were scored as IHC 3+. By analyzing all IHC categories separately, the Krippendorff’s alpha showed poor agreement ($\alpha = 0.32$). The agreement stayed moderate after clustering other IHC groups together ($\alpha = 0.65-0.73$), and the percentage of agreement remained poor to moderate (20.9- 80%), as presented in table 1. Noticeably, the highest percentage of agreement was achieved when we clustered ultralow, 1+ and 2+ together and compared it to IHC 0; the complete agreement (all pathologists) was 74.3% (78 of 105) and 80% (84 of 105) for an agreement among 14 out of 16 pathologists. Figures 3 presents two examples of cases with low agreement in the second round, likely due to difficulty in the distinction between non-specific staining versus true membrane staining (A) and/or the use of the 10% cut off (B).

**FISH analysis and difference between IHC categories**

The FISH assay was performed for all 105 BC cases. The $\text{HER2}/\text{CEP17}$ ratios and average number of $\text{HER2}$ copies/cell were non-normally distributed. Overall, the median $\text{HER2}/\text{CEP17}$ ratio and mean $\text{HER2}$ copy number were 1.15 (range 0.43-4.19) and 1.79 (range 1.07-4.23) respectively.

The FISH results according to the IHC consensus of round 2 are presented in figure 4. For the analysis of the mean $\text{HER2}$ copy number according to the IHC scores, an overall statistically significant difference
was found (p <0.001) when HER2 0 and ultralow scores were grouped together versus HER2 1+ and 2+ scores (Figure 4A). In the post hoc analyses, no significant difference was found in HER2 copy number between HER2 0 and ultralow vs 1+ (p 0.141), while significant differences were found between HER2 0 and ultralow vs HER2 2+ (p <0.001), and between HER2 1+ vs 2+ (p <0.001).

In the second analysis we grouped HER2 ultralow and 1+ together (Figure 4B). We also found a significant difference between the ranks of HER2 copy number (p <0.001). In the post hoc analyses, no significant difference in mean HER2 copy number was found between HER2 0 vs ultralow and 1+ (p 0.149). Significant differences in mean HER2 copy number were observed between HER2 ultralow and 1+ vs 2+ (p <0.001), and between HER2 0 vs 2+ (p <0.001). Regarding the HER2/CEP17 ratios, no significant differences were found between the different IHC categories when clustering 0 and ultralow together (p=0.315) nor when grouping ultralow and 1+ together (p=0.7). These results were consistent when performing the same analyses using the results of the consensus IHC scores of the first scoring round (data not shown). Therefore, FISH results seem to have limited additional value to differentiate between the HER2 0 and HER2 (ultra) low cases.

**Discussion**

The precision of HER2 IHC scoring is essential to select BC patients for existing and novel HER2-targeting agents. Since T-DXd showed effectiveness in patients with HER2 low tumors, the necessity to distinguish between no expression versus low HER2 expression are likely to become clinically relevant. In this study, there was a decrease in the number of cases scored as HER2 0 from the first (22.8%) to the second (16.3%) round, which is in line with literature. Noticeably, the proportion of IHC 2 + cases was also higher (31.3%) with the current criteria compared to the 2007 criteria (25.5%), as described previously.

The present study suggests that the interobserver agreement for IHC interpretation could be improved by adapting the criteria for immunohistochemical assessment. The best (or rather, least poor) reproducibility was seen in the second scoring-round when comparing IHC 0 versus the cluster of ultralow/1 + and 2+. This supports the hypothesis that identifying HER2 0 cases according to the 2007 guideline (complete lack of HER2 expression, or the “all or nothing” principle) is easier compared to the use of a 10% cut point according to 2018 guidelines. In addition, the use of “barely perceptible”, as described in the 2018 guidelines, is perceived as subjective by most pathologists participating in this study. Difficulties in distinguishing non-specific background or cytoplasmic staining and true membrane staining could explain the discrepancies between IHC 0 and ultralow categories. These results are in line with previous studies evaluating the interobserver agreement for each of these ASCO/CAP guidelines. These results support the idea that the current ASCO/CAP IHC criteria, which are optimized to select HER2 amplified cases, are not very robust to distinguish cases without from those with low levels of HER2 expression. Furthermore, in this study we demonstrated that the reproducibility between HER2 ultralow and 1 + is moderate, which also illustrates the difficulty using the 10% cut-off. The clinical relevance of HER2 ultralow tumors is uncertain. A recent study from Diéras et al. demonstrated that T-DXd was effective even in HER2 0 tumors, but it is unclear whether these cases were completely negative or
ultralow. Future research should focus on finding novel methods to distinguish the subtle differences between HER2 0 and HER2 low BC. Some studies have already developed artificial intelligence-algorithms and advanced techniques of targeted mass spectrometry for this purpose, but the studies in this area are still limited and require validation in larger, independent cohorts. Additionally, continuous training of practicing pathologists seems important, as highlighted by the high degree of variability in our study.

An additional aim of our study was to assess whether FISH data, and mean HER2 copy numbers in particular, could improve the discrimination between HER2 0 and HER2 (ultra)low tumors. However, we did not observe any statistical differences in the mean HER2 copy number between the IHC groups 0 and 1+ regardless of where the ultralow group was included. Therefore, using FISH as a companion test for IHC does not seem to have additional value to separate HER2 0 and HER2 low tumors. These data are in contrast with previous studies of Lambein et al, who reported that tumors with an IHC score of 0 had a lower HER2 copy number compared to the IHC 1+ group. In our study, we did not observe this effect. However, although the p value was not significant, the difference in mean rank (Md) was higher in the ultralow and 1+ group (Md = 41.45, p = 0.149) compared to ultralow with 0 (Md = 34.31, p = 0.141). The lack of statistical significance between the IHC 0 and 1+ groups in this study could be due to the smaller sample size compared to other studies, or due to tumor heterogeneity.

To our knowledge this is the first HER2 interobserver study focusing on the novel subgroup of HER2 low BC that estimated interobserver agreement before and after a consensus meeting and adjusting IHC-criteria. In addition, we also assessed the association between IHC and FISH data. In this study, we used the 4B5 HER-2/neu clone on the Ventana platform (Roche), which is also used in the DESTINY-Breast trials. Another strength of this study involved the large number of specialized pathologists participating in both rounds. Furthermore, by including only HER2 negative cases we prevented a false improvement in the agreement, since previous reports already demonstrated that assessment of IHC 3+ cases shows a better concordance. Our study also has some limitations. First, not all pathologists were used to score digital slides in daily practice, which could have affected our results. Second, we used needle biopsies of the primary tumor, while in the clinical setting, HER2 status of the metastases is also relevant. However, since the scoring criteria of metastases do not differ from primary tumors, this is less likely to have affected our results.

**Conclusion**

The present study suggests that the 2007 ASCO/CAP criteria were more reproducible to distinguish HER2 0 cases compared to the 2018 ASCO/CAP criteria. However, the reproducibility is still only moderate and performing FISH does not provide additional support to discriminate between very low levels of HER2 expression. This could potentially lead to suboptimal selection of patients that could benefit from novel treatment options like T-DXd. Our results reinforce the need to develop clearer, more reproducible definitions for IHC scoring and training for pathologists to diagnose HER2 low BC, adapted to the results of ongoing clinical trials. Future research in this field should also focus on the development of novel and more accurate methods to quantify the level of HER2 expression.
Declarations

Acknowledgements

The authors thank H. Sleddens and L. Uytdewilligen for performing and scoring the FISH assays. We thank J. Hudeck for his help with using Slidescore.

Conflict of interests

C.D. was involved in an advisory board of Astra Zeneca/ Daiichi Sankyo and received research funding from Roche and AstraZeneca.

B.V.: honoraria received by UMCG for expertise or scientific advisory board/consultancy (on request): Visiopharm, Philips, MSD/Merck, Daiichi-Sankyo/AstraZeneca; speaker’s fee from Visiopharm, Diaceutics, MSD/Merck. All unrelated to the current work.

Ethics Approval and Consent to Participate

According to the Code of Conduct of the Federation of Medical Sciences in the Netherlands, no informed consent or ethical approval was needed for this study. This study was performed in accordance with the Declaration of Helsinki.

Author Contributions

C.D. performed study concept and design. X.B.N. performed the manuscript writing, C.D., X.B.N., M.R.B., and D.N. performed development of methodology, review and revision of the paper. G.B., C.C., S.C.D., M.C.H.H., E.K., K.L., D.J.E.P., R.H.J.A.S., M.R.B., J.B.B., J.S.G., B.V., K.V., C.P.H.V., W.V., P.J.W. performed the slides score and participated in the consensus meeting. X.B.N. performed the data interpretation and statistical analysis. C.D. provided the tissue material. All authors read and approved the final paper.

Funding

This work was funded by Astrazeneca. The funder had no role in the design of the study, the interpretation of the results or the writing of the article.

Data Availability Statement

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

References


**Table**

Table 1 is available in the Supplementary Files section.

**Figures**

**Figure 1**

Study design and IHC scoring criteria used for the first and the second scoring round.
Figure 2

Examples of cases with low agreement in the first round (amplification x80). A) Seven pathologists scored IHC 0 and nine scored IHC 1+. B) Nine pathologists scored IHC 0 and seven scored IHC 1+. 
Figure 3

Examples of cases with low agreement in the second round (amplification x80). A) Seven pathologists scored IHC 0 and nine scored IHC ultralow. B) Six pathologists scored IHC ultralow and ten scored IHC 1+. 
Figure 4

Boxplot of FISH results according to the consensus IHC categories of round 2, by clustering HER2 ultralow with IHC 0 (A) or by clustering HER2 ultralow with IHC 1+ (B).

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- Table1.xlsx