Published in final edited form as:
Mod Pathol. 2023 January ; 36(1): 100032. doi:10.1016/j.modpat.2022.100032.

Multi-institutional Assessment of Pathologist scoring HER2 Immunohistochemistry

Charles J Robbins[1], **Aileen I Fernandez**[1], **Gang Han**[2], **Serena Wong**[1], **Malini Harigopal**[1], **Mirna Podoll**[3], **Kamaljeet Singh**[4], **Amy Ly**[5], **M. Gabriela Kuba**[6], **Hannah Wen**[6], **Mary Ann Sanders**[7], **Jane Brock**[8], **Shi Wei**[9], **Oluwole Fadare**[10], **Krisztina Hanley**[11], **Julie Jorns**[12], **Olivia L. Snir**[13], **Esther Yoon**[14], **Kim Rabe**[15], **T. Rinda Soong**[16], **Emily S Reisenbichler**[17], **David L Rimm, M.D.- Ph.D.**[1,18]

[1]Department of Pathology, Yale School of Medicine, New Haven, CT, USA

[2]Texas A&M University, College Station, TX, USA.

[3]Dept of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA

[4]Dept of Pathology and Laboratory medicine, Brown University, Providence, RI, USA

[5]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

[6]Dept of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[7]Dept of Pathology, Norton Healthcare, Louisville, KY, USA

[8]Dept of Pathology, Brigham and Women's Hospital, Boston, MA, USA

[9]Dept of Pathology, University of Kansas Medical Center, Kansas City, KS, USA

[10]Dept of Pathology, University of California San Diego, San Diego, CA, USA

[11]Dept of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, USA

[12]Dept of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA

[13]Dept of Pathology, Providence Health & Services, Portland, OR, USA

[14]Dept of Pathology, MD Anderson, Cancer Center, Houston, TX, USA

Corresponding Author: David L. Rimm, M.D.- Ph.D., Professor of Pathology and Medicine (Oncology), Department of Pathology, Yale University School of Medicine, 310 Cedar Street, BML 116, New Haven, CT 06520-8023, Phone: 203-737-4204, david.rimm@yale.edu.

Author Contributions

ESR, DLR performed study concept and design; CJR, AIF, ESR, & DLR performed development of methodology and writing, review, and revision of the paper; CJR, AIF, & GH provided acquisition, analysis and interpretation of data, and statistical analysis; All pathologist authors provided their expertise in breast cancer pathology and HER2 IHC scoring to support this study. All authors read and approved the final paper.

Ethics Approval and Consent to Participate

Written informed consent or waiver of consent was provided by all the patients. This study was approved by Yale Human Investigation Committee protocol ID 9505008219. This study was performed in accordance with the Declaration of Helsinki.

[15]Dept of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA

[16]Dept of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[17]Department of Pathology, SSM Health Saint Louis University Hospital, St. Louis, MO, USA

[18]Department of Medicine (Oncology), Yale School of Medicine, New Haven, CT, USA

Abstract

The HercepTest was approved 20+ years ago as the companion diagnostic test for trastuzumab in HER2 amplified/overexpressing breast cancers. Subsequent HER2 immunohistochemistry (IHC) assays followed, including the now most common Ventana 4B5 assay. While this IHC assay has become the clinical standard, its reliability, reproducibility, and accuracy have largely been approved and accepted based on concordance between small numbers of pathologists without validation in a real-world setting. In this study, we evaluate the concordance and inter-rater reliability of scoring HER2 IHC in 170 breast cancer biopsies by 18 breast cancer-specialized pathologists from 15 institutions. We used the ONEST (Observers Needed to Evaluate Subjective Tests) method to determine the plateau of concordance and the minimum number of pathologists needed to estimate inter-rater agreement values for large numbers of raters, as seen in the real-world setting. We report substantial discordance within the intermediate categories (<1% agreement for 1+ and 3.6% agreement for 2+) in the four-category HER2 IHC scoring system. The discordance within the IHC 0 cases is also substantial with an overall percent agreement (OPA) of only 25% and poor inter-rater reliability metrics (0.49 Fleiss' kappa, 0.55 intraclass correlation coefficient). This discordance can be partially reduced by using a three-category system (28.8% vs. 46.5% OPA for four and three-category scoring systems respectively). ONEST plots suggest that the OPA for the task of determining a HER2 IHC score 0 from not 0 plateaus statistically around 59.4% at 10 raters. Conversely, at the task of scoring HER2 IHC as 3+ or not 3+ pathologists' concordance was much higher with an OPA that plateaus at 87.1% with 6 raters. This suggests that legacy HER2 IHC remains valuable for finding HER2 gene amplified patients, but unacceptably discordant in assigning HER2-low or negative status for emerging HER2-low therapies.

Introduction

Accurate quantification of human epidermal growth factor 2 (HER2) expression levels is critical in the management of breast cancer patients. HER2 expression in breast cancer spans a large dynamic range of 3 logs so immunohistochemistry (IHC) cannot adequately assess HER2 concentrations throughout this dynamic range[1–3]. Since only patients with amplified, over expressed HER2 benefitted from the initial HER2 axis drugs (e.g. trastuzumab)[4–9], subsequent commercial assays were designed to detect high HER2 expression. The current most common companion diagnostic test is the Ventana 4B5 assay and its dynamic range is best in tumors that have over 100,000 molecules of HER2 protein per cell. Even so, the current American Society of Clinical Oncology/College of American Pathologists Clinical Practice (ASCO/CAP) guidelines require reflex gene amplification testing by fluorescence in situ hybridization (FISH) of all IHC 2+ cases so that HER2 amplified tumors are not missed[10,11]. More recently the landscape has changed as there are now new anti-HER2 drugs, for example, trastuzumab deruxtecan (T-DXd), that are effective in this low HER2 expressing subgroup[12–17]. These recent clinical trials for T-DXd have attempted to define this low HER2 expressing subgroup as 1+ or 2+ cases without gene amplification using the 2018/current ASCO/CAP guidelines for the legacy HER2 assays[11], which were originally designed for detecting amplified HER2 expression. This raises new questions about the conventional FDA approved HER2 assays and their performance for both the historical drugs for which the assay was approved and for the new drugs for which it may be used.

The FDA granted approval for the conventional HER2 IHC assays based on the ability to detect positive or negative cases compared to HER2 gene amplification or the agreement with the original Dako HercepTest (only using a 0/1+, 2+, and 3+ scoring system). Additionally, all of these historical HER2 assays were approved with a relatively low inter-rater agreement requirements as can be seen in the FDA's published SSEDs[18–20]. Specifically, these assays were only required to be evaluated by 2 to 3 pathologists for FDA approval. The decision of whether a case was a score of 0 vs. 1+ (or "low" expressing) was not required to be accurate, reproducible, or concordant based on the FDA summary of safety and effectiveness datasheets (SSEDs) for these assays from over 20 years ago. The distinction between 0 vs. 1+ cases was simply not a meaningful category for FDA approval of these HER2 IHC assays, whereas the ASCO/CAP guidelines for pathologists and assay package insert information have featured the 0 and 1+ categories since the original Dako HercepTest. The general inattention toward reproducible scoring of the 0/1+ categories and subsequent "lumping" of these cases into a negative class presumably did not have significant clinical ramifications as the assays were "fit-for-purpose" to detect amplified cases[21]. However, now it is clinically relevant to distinguish "true negative" from HER2-low cases for these emerging therapies (namely antibody-drug conjugates including T-DXd), and the question is whether the legacy HER2 assays should be used for this task.

Early studies on the performance of these HER2 IHC assays focused on 3+ and 2+ scores, as these cut-points indicated trastuzumab therapy or reflex FISH testing based on the evolving FDA and ASCO/CAP guidelines (note, FISH as evolved to also included chromogenic in situ hybridization (CISH) and thus in the remainder of this work, we simply use ISH). Many independent studies demonstrated that HER2 IHC is a reasonable first test for HER2 overexpression with high negative predictive value as well as an acceptable positive predictive value when paired with reflex ISH testing for IHC 2+ cases[22–26]. However, inter-observer concordance for HER2 IHC scoring is mixed with some studies reporting satisfactory agreement for positive (2+/3+) and negative (0/1+) cases[27–30], whereas others demonstrated significant discordance particularly on 2+ and negative (0/1+) IHC scoring[23,31–34]. Even though HER2 ISH is used as the gold-standard reference assay for gene amplification in these studies, testing HER2 ISH for all breast cancer patients did not replace HER2 IHC testing likely due to the complexity and cost[35]. Past studies have reported performing HER2 ISH testing for all patients including IHC 0 & 1+. While HER2 IHC 0 & 1+ patients with ISH amplification can benefit from anti-HER2 therapies[36,37], the prevalence of HER2 gene amplification with 0/1+ HER2 IHC expression is low (1.5% to 5% of 0/1+ are ISH positive compared to 20% to 30% of 2+ cases)[22,35].

More recent studies report substantial inter-rater discordance and poor reproducibility amongst low vs." true negative" (IHC 0) cases when determining HER2 IHC status in breast cancer[32,34,38,39]. Lambein et al. reported disagreement rates as high as 85% for HER2 0 IHC scores using the Ventana 4B5 assay between their local laboratory and central assessment[38]. In a retrospective study investigating agreement of HER2 IHC classification across 5 breast cancer-specialized pathologists, discordance was mostly driven by 0 vs. 1+ cases (43% of all discordant cases, 15% of total cases)[39]. Results from studies evaluating the evolution of HER2-low expression status in primary to recurrent/advanced breast cancer[40,41] or other associations in HER2-low breast cancer could be confounded by the high inter-rater discordance and poor reproducibility in the low expression range for the historical HER2 IHC assays. Despite these concerning results, these studies were only able to make limited conclusions on the inter-rater reliability of scoring HER2 IHC due to their small number of pathologists or cases. Furthermore, conventional methods of inter-rater reliability for a small number of raters can poorly generalize to the broad population of raters.

In the real world, there are not just 3 pathologist raters, but thousands of pathologists scoring these assays. There is no established method for examination of concordance between large numbers of observers nor is there an established statistical method to determine how many observers are needed to represent real world pathologist performance. Recently we have described a method to examine this issue. The ONEST method (Observers Needed to Evaluate Subjective Tests)[42], allows us to assess the likelihood of this assay showing concordance amongst many pathologist readers. Our goal here is to use this method to better understand the past and future value of the conventional legacy HER2 IHC assays designed for detecting high HER2 expression. The ONEST method is based on calculation of the overall percent agreement within many combinations of pathologists/raters in order to determine if there is a plateau in overall percent agreement. The presence of a plateau strongly suggests that the metric will be stable even if the number of raters/pathologists continues to increase.

Here, we examine the concordance or overall percent agreement at the various cut-points used in the HER2 IHC assay. In this multi-institutional study, we evaluate the inter-rater reliability of scoring HER2 IHC in 170 breast cancer cases by a group of 18 breast cancer-specialized pathologists. We quantify inter-rater reliability with common metrics as well as the recently developed ONEST method to better generalize the performance of the legacy HER2 assay to larger populations of pathologists in routine clinical practice settings.

Materials and methods

Patient biopsies and immunohistochemistry

We retrospectively collected 170 breast biopsies from the archives of the Department of Pathology at Yale School of Medicine. These were all from patients with breast cancer seen in 2018. This set was enriched for HER2 positive cases defined as those with 3+ score by immunohistochemistry (IHC), or 2+ by IHC and HER2 positive by fluorescent in situ hybridization (FISH), as defined by American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) clinical practice guidelines[10,11]. The archival hematoxylin and eosin (H&E) slides and HER2 IHC slides were reviewed, and quality checked by a board-certified pathologist to confirm stain integrity and checked for strong membranous staining of positive controls, that the slide and coverslips were not broken, and that all slides had enough tissue to assess. The slides were scanned using Aperio ScanScope Console (v10.2.0.2352) using bright field Whole Slide Scanning at 20× magnification and sent to eighteen board-certified pathologists, most with over 5 years' experience. Pathologists scored the cases as HER2 0, 1+, 2+, or 3+ according to the current ASCO/CAP criteria[11].

Statistical analysis and Observers Needed to Evaluate a Subjective Test (ONEST)

All statistical analyses were performed in R 4.1.0[43]. The ONEST package[44] was used to model and visualize the change in overall percent agreement (OPA) as a function of the number of pathologists scoring the cases. This method is described in detail in Han et al[42]. Briefly, for each subset of cases, we randomly select combinations of pathologists and calculate the OPA for each group (from sizes 2 to 18 in this study of 18 pathologists). 100 curves (of 100 combinations for each group size) were generated for plotting, and 1000 curves were used to estimate the mean and 95% confidence interval of the ONEST plot. The resulting ONEST plots descend and can reach a non-zero plateau that can be validated by estimating the parameters of the statistical model described in (Han G. et al., 2021)[42]. The ONEST model can also be used to estimate the number of raters needed to reach the plateau by calculating when the OPA difference between successive groups becomes clinically insignificant (less than 0.5%). As shown previously[42,45], if a test is easy to interpret and has high concordance, then the plateau will occur at a large OPA value with a small number of raters. Conversely, when there is low concordance for a test, the plateau occurs at low OPA values or may drop to 0 with a large number of raters. The raters package[46] and irr package[47] were used to calculate Fleiss' Kappa and the intraclass correlation coefficient (ICC). ICC was calculated using a two-way random-effects model with absolute agreement as the relationship type and single rater as the measurement unit. The ggplot2 package[48] was used for plotting and data visualizations in this study.

Results

To assess HER2 IHC inter-rater reliability in this study, whole tissue sections of 170 independent breast cancer biopsy cases were evaluated for HER2 IHC by 18 pathologists from 15 institutions. Figure 1 displays stacked bar plots of the HER2 IHC score given for each case (Fig. 1A) or by each of these pathologists (Fig. 1B). For the 170 cases, 121 cases had disagreement on the IHC score amongst the 18 pathologists. The overall percent agreement (OPA), which is the percent of cases where all pathologists/raters in the group agree (in this case all 18 pathologists), for the 170 HER2 IHC cases was 28.8%. Discordance was observed for each cut-point, with the 0 vs. 1+ cut-point displaying the largest number of discordant cases. Of the 170 cases, 92 were read as 0 by at least one pathologist. 23 of these 92 IHC 0 cases were concordant (which corresponds to an OPA of 25%)(Fig. 1A; Table 1). 44 cases were read as 3+ by at least one pathologist. Again, only 22 of the 44 IHC 3+ cases were concordant amongst all readers (Fig. 1A; Table 1). When evaluating how each pathologist scored the 170 cases, the largest discrepancy is in the percent of cases scored as HER2 negative (IHC 0) vs. low (1+ or 2+); some pathologists scored 40% of the cases as IHC 0 whereas others scored 20% of the cases as IHC 0 (Fig. 1B).

Next, we wanted to explore whether these metrics of OPA across the 18 pathologists in this study were generalizable to a larger population of pathologists performing HER2 IHC in breast cancer. To do this, we used the ONEST technique[42] of plotting OPA within different combinations of groups of pathologists to determine if there is a point where OPA plateaus. The ONEST method was developed to not only determine the number of observers needed for evaluation of a subjective test, but also to predict how the test/biomarker would perform in the real world with thousands of pathologist readers. The presence of a plateau strongly suggests that the metric will be stable even if you continue to increase number of raters/pathologists (since there are thousands of pathologists in practice reading IHC, this method has the potential to predict how the biomarker will perform with thousands of raters). Additionally, the point where the metric plateaus indicates the number of pathologists that are required to provide realistic concordance estimates for when the assay is broadly used.

The ONEST plots in Figure 2 of OPA show a decrease in OPA as the number of raters in the group increases, with each plot reaching a plateau between 6 to 12 raters. When considering the concordance amongst pathologists using a four-category score (0, 1+, 2+, 3+) compared to a three-category score (0, Low*, 3+), a three-category score yielded a higher OPA (28.8% OPA for four-category compared to 46.5% OPA for three-category)(Fig. 2A; Table 1). This can also be seen in Figure 1A, as combining 1+ and 2+ categories removes 30 discordant cases. Similarly, the Fleiss' kappa increases using a three-category score compared to a four-category score for reading HER2 IHC (Fleiss' kappa of 0.65 compared to 0.74 for four- and three-category scores respectively)(Fig. S1; Table 1). This suggests that there is substantial discordance when assessing whether a case is 1+ or 2+, and that combining 1+ and 2+ cases into a HER2-low category can result in increased concordance.

We wanted to determine the OPA using the ONEST method for cases that were scored as 0 and cases that were scored as 1+ since this is the cut-point for determining whether a case is HER2-low and hence a candidate for HER2-low therapies including T-DXd. Other studies have reported that the bulk of discordance in HER2 IHC is driven by cases that were discordant between 0 vs 1+[39]. Similarly, concerning disagreement rates of HER2 0 scores between local and central assessment (85%) have been reported[38]. Figure 2B and 2C show the OPA ONEST plots when cases were scored as 0 and 1+ respectively by at least one of the 18 pathologists. The OPA for the cases that were scored as 0 plateaus at 25%, indicating that the pathologists disagreed in 75% of the cases that were scored as 0 by at least one pathologist. The pathologists' discordance of the 0 cases was mainly between scores of 0 vs. 1+ (785/1656 of total ratings within 0 cases, 69/92 of 0 cases read as 1+ by another pathologist) and to a much lesser extent between scores 0 vs. 2+ (85/1656 ratings, 29/92 cases)(Table S1; Table S2). The OPA for the cases that were scored as 1+ reaches less than 1%. This is due to there being only 1 case that all 18 pathologists agreed that was 1+ out of the 102 cases that were scored as 1+ by at least one pathologist (Fig. 1A). The 2+ cases also had a very low OPA that reached 3.6% (Fig. S2-C,H; Table 1). Upon combing the 1+ and 2+ categories into a HER2 low category, the OPA for these low cases increases and plateaus at 27.2% (Fig. S2-D,I; Table 1). Also surprising, only 50% of the HER2 IHC 3+ cases were agreed upon by all of the 18 pathologists (Fig. S2-E,J; Table 1).

Since the emergence of lower levels of HER2 as a target for therapy, determining when a case is 0 vs. not 0 is an important clinical decision threshold for prescription of HER2-low therapies. This new threshold is added on to the existing threshold where trastuzumab is prescribed in HER2 amplified cases, defined as 3+ or 2+ and ISH+. Thus, we next evaluated the pathologists' ability to make clinically impactful/significant reads at both clinical thresholds. First, for the task of determining cases with a 3+ score vs. not 3+ and then for cases with a 0 score vs. not 0. To do this, we grouped the HER2 IHC scores as 3+ or not 3+ and 0 or not 0 for analysis. In the HER2 IHC ONEST plot of the scores grouped as 3+ or not 3+ (Fig. 2D), there was an OPA of 87.1% that plateaued around 6 raters. This OPA demonstrates that this group of pathologists has high agreement for the task of determining 3+ cases from not 3+ cases. Correspondingly, in the ONEST plot of the scores grouped as 0 or not 0 (Fig. S3; Table 1) pathologists in the study had an OPA of 59.4% that plateaued around 10 raters. This observation agrees with previous studies that suggest that pathologists cannot agree on cases with a HER2 0 score, as there is up to a 40.6% disagreement for what cases are IHC 0 or not 0.

Discussion

This multi-institutional study assessing the inter-rater reliability of HER2 IHC scoring demonstrates several findings and offers generalizable concordance estimates for scoring HER2 IHC. The first finding is that the intermediate categories (1+ and 2+) in the four-category HER2 IHC scoring system are a large source of discordance (<1% agreement for 1+ and 3.6% agreement for 2+), and this discordance can be partially reduced by using a three-category system (28.8% vs. 46.5% OPA for four and three-category scoring systems respectively). Intermediate categories being less reproducible than the extreme categories is a trend that has been found in several other studies for different multi-category assays[42,45,49–51]. The low agreement of 2+ cases in this study is due to discordance in scores of 2+ vs. 3+ (22/84 of 2+ cases read as 3+ by another pathologist), 1+ vs. 2+ (61/84), as well as a non-negligible number of discordant cases for scores of 0 vs. 2+ (29/84)(Table S2). The disagreement of cases at the 0 or 1+ vs. 2+ boundaries is concerning as these cases would not receive reflexive FISH testing (under the current ASCO/CAP guidelines) to check for HER2 gene amplification status if they were marked as 0 or 1+. Although the prevalence of HER2 gene amplification in IHC 0/1+ cases is much lower than IHC 2+ cases, past studies have demonstrated that these patients can have pathologic complete response to HER2 amplified therapy regimens[36,37].

We also found that there is a low concordance amongst pathologists in this cohort when evaluating breast cancer cases with HER2 IHC score of 0 and at the task of determining a score of 0 or not 0. This is a critical cut-point for the new HER2 antibody-drug conjugates. Other studies have also reported discordance for scoring HER2, particularly around the 0 to 1+ or 2+ cut-points[34,38,39]. Despite knowing that there was potentially a high level of discordance, these studies did not show generalizability of their results to a large population of pathologists. The ONEST plots in this study suggests that the OPA for the task of determining a HER2 IHC score 0 from not 0 plateaus statistically around 59.4%. The agreement for assigning a HER2 IHC score of 0 vs. not 0 is only slightly better than a coin flip amongst these 18 pathologists in this multi-institutional study. Conversely, at the task of scoring HER2 IHC as 3+ or not 3+ pathologists' concordance was much higher with an OPA that plateaus at 87.1%. These results indicate that the legacy HER2 IHC assay is largely valuable as is to find HER2 gene amplified patients for conventional HER2 targeted therapies, but unacceptably discordant for assigning HER2-low status for emerging HER2-low therapies.

This study has a number of considerations and limitations. One potential limitation is that the 18 pathologists that scored the biopsy cohort were not told that the 0 vs. 1+ concordance level would be assessed. In retrospect, many said they would have examined the low expressing cases more closely. Also, the set of breast cancer biopsies was enriched in 2+ and 3+ cases but most of the pathologists did not know this prior to reading the slides. This could have contributed to pathologists assigning HER2 negative and 1+ scores more frequently as these cases are more common in clinical practice. However, these considerations can be argued as strengths for this study, as compelling pathologists to provide additional scrutiny of the 0/1+ cases or informing them about the cohort composition beforehand would not have provided an accurate reflection of how pathologists really score these cases. Another limitation of this study is that we do not have quantitative molecular measurements of HER2 in the examined core biopsies. However, the absence of a criterion standard is not unusual in pathologist concordance studies.

Although the legacy HER2 IHC assay combined with ISH is the companion diagnostic for amplified HER2 therapies including trastuzumab, the legacy HER2 IHC assay's high discordance and poor inter-rater reliability amongst pathologists for scoring IHC 0, 1+, and 2+ cases demonstrated in this study suggest that this assay will be problematic for the emerging HER2-low treatments. Finally, in agreement with other studies performed with other methods, the high level of discordance for pathologists scoring HER2 IHC 0 vs. not 0 (40.6% disagreement reported by ONEST in this study), suggests that the legacy HER2 IHC assay will likely be inaccurate and arguably insufficient for clinical decision making when prescribing HER2-low specific treatments (e.g. trastuzumab-deruxtecan).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This research was supported by the Breast Cancer Research Foundation (DLR). HYW and MGK are supported in part by a National Institutes of Health/National Cancer Institute Cancer Center Support Grant (P30CA008748)

Conflicts of interest

DLR has served as an advisor for Astra Zeneca, Agendia, Amgen, BMS, Cell Signaling Technology, Cepheid, Daiichi Sankyo, Danaher, Genoptix/Novartis, GSK, Konica Minolta, Merck, NanoString, PAIGE.AI, Perkin Elmer, Roche, Sanofi, Ventana and Ultivue. Astra Zeneca, Cepheid, NavigateBP, NextCure, Nanostring, and Lilly, have funded or currently fund research in DLR's lab. HYW serves as an advisor for Astra Zeneca. SW has served as a breast pathology consultant for Astra Zeneca. OLS has served as a pathologist in a project for Diaceutics. All other authors have nothing to disclose.

Data Availability Statement

The datasets used and/or analyzed during the current study are available within the supplemental materials. The scripts used in the analysis are available at https://github.com/crobbins327/HER2-IHC-ONEST-Multi-Institutional-IRR. Additional datasets can be made available from the corresponding author on request.

References

1. McCabe A, Dolled-Filhart M, Camp RL & Rimm DL Automated Quantitative Analysis (AQUA) of In Situ Protein Expression, Antibody Concentration, and Prognosis. J Natl Cancer Inst 97, 1808–1815 (2005). [PubMed: 16368942]
2. Onsum MD, Geretti E, Paragas V, Kudla AJ, Moulis SP, Luus L et al. Single-Cell Quantitative HER2 Measurement Identifies Heterogeneity and Distinct Subgroups within Traditionally Defined HER2-Positive Patients. Am J Pathol 183, 1446–1460 (2013). [PubMed: 24035511]
3. Rimm DL What brown cannot do for you. Nat Biotechnol 24, 914–916 (2006). [PubMed: 16900128]
4. Chen HL, Chen Q & Deng YC Pathologic complete response to neoadjuvant anti-HER2 therapy is associated with HER2 immunohistochemistry score in HER2-positive early breast cancer. Medicine (United States) 100 (2021).
5. Ennis S Lamon DJ, Rian Eyland -j Ones B. L., Teven Hak SS, Ank Uchs HF, Irginia Aton VP, Harm PD et al. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. N Engl J Med 344, 783–792 (2009).
6. Kaufman B, Mackey JR, Clemens MR, Bapsy PP, Vaid A, Wardley A et al. Trastuzumab plus anastrozole versus anastrozole alone for the treatment of postmenopausal women with human epidermal growth factor receptor 2-positive, hormone receptor-positive metastatic breast cancer: Results from the randomized phase III TAnDEM study. J Clin Oncol 27, 5529–5537 (2009). [PubMed: 19786670]
7. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Davidson NE et al. Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer. N Engl J Med 353, 1673–1684 (2005). [PubMed: 16236738]
8. Sawaki M, Ito Y, Tada K, Mizunuma N, Takahashi S, Horikoshi N et al. Efficacy and safety of trastuzumab as a single agent in heavily pretreated patients with HER-2/neu-overexpressing metastatic breast cancer. Tumori 90, 40–43 (2004). [PubMed: 15143970]
9. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 20, 719–726 (2002). [PubMed: 11821453]
10. Wolff AC, Hammond MEH, Schwartz JN, Hagerty KL, Allred DC, Cote RJ et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 25, 118–145 (2007). [PubMed: 17159189]
11. Wolff AC, McShane LM, Hammond MEH, Allison KH, Fitzgibbons P, Press MF 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. Arch Pathol Lab Med 142, 1364–1382 (2018). [PubMed: 29846104]
12. Dekker TJA HER2-Targeted Therapies in HER2-Low–Expressing Breast Cancer. J Clin Oncol 38, 3350–3351 (2020). [PubMed: 32658628]
13. Eiger D, Agostinetto E, Saúde-Conde R & De Azambuja E The Exciting New Field of HER2-Low Breast Cancer Treatment. Cancers 13, 1015 (2021). [PubMed: 33804398]
14. Modi S, Ohtani S, Lee CC, Wang K, Saxena K & Cameron DA A phase III, multicenter, randomized, open label trial of [fam-] trastuzumab deruxtecan (DS-8201a) versus investigator’s choice in HER2-low breast cancer. J Clin Oncol 37, TPS1102–TPS1102 (2019).

15. Modi S, Park H, Murthy RK, Iwata H, Tamura K, Tsurutani J et al. Antitumor Activity and Safety of Trastuzumab Deruxtecan in Patients With HER2-Low-Expressing Advanced Breast Cancer: Results From a Phase Ib Study. J Clin Oncol 38, 1887–1896 (2020). [PubMed: 32058843]

16. Modi S, Saura C, Yamashita T, Park YH, Kim S-B, Tamura K et al. Trastuzumab Deruxtecan in Previously Treated HER2-Positive Breast Cancer. N Engl J Med 382, 610–621 (2020). [PubMed: 31825192]

17. 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 38, 1951–1962 (2020). [PubMed: 32330069]

18. Administration USFaD. Summary of Safety and Effectiveness Data (SSED) PMA P090015, https://www.accessdata.fda.gov/cdrh_docs/pdf9/P090015b.pdf (2012).

19. Administration USFaD. Summary of Safety and Effectiveness Data (SSED) PMA P980018, https://www.accessdata.fda.gov/cdrh_docs/pdf/P980018.pdf (1998).

20. Administration USFaD. Summary of Safety and Effectiveness Data (SSED) PMA P990081, https://www.accessdata.fda.gov/cdrh_docs/pdf/P990081B.pdf (2000).

21. Torlakovic EE Fit-for-Purpose Immunohistochemical Biomarkers. Endocr Pathol 29, 199–205 (2018). [PubMed: 29696583]

22. Cuadros M & Villegas R Systematic review of HER2 breast cancer testing. Appl Immunohistochem Mol Morphol 17, 1–7 (2009). [PubMed: 18685491]

23. Thomson TA, Hayes MM, Spinelli JJ, Hilland E, Sawrenko C, Phillips D et al. HER-2/neu in breast cancer: Interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. Mod Pathol 14, 1079–1086 (2001). [PubMed: 11706067]

24. Varga Z, Noske A, Ramach C, Padberg B & Moch H Assessment of HER2 status in breast cancer: overall positivity rate and accuracy by fluorescence in situ hybridization and immunohistochemistry in a single institution over 12 years: a quality control study. BMC Cancer 13 (2013).

25. Vincent-Salomon A, MacGrogan G, Couturier J, Arnould L, Denoux Y, Fiche M et al. Calibration of immunohistochemistry for assessment of HER2 in breast cancer: results of the French multicentre GEFPICS study. Histopathology 42, 337–347 (2003). [PubMed: 12653945]

26. B Z., Y W & H X Concordance of immunohistochemistry and fluorescence in situ hybridization for assessment of HER2 status in breast cancer patients in Xinjiang autonomous region, China. Int J Clin Exp Pathol 10 (2017).

27. Md Pauzi SH, Masir N, Yahaya A, Mohammed F, Tizen Laim NMS, Mustangin M et al. HER2 testing by immunohistochemistry in breast cancer: A multicenter proficiency ring study. Indian J Pathol Microbiol 64, 677–682 (2021). [PubMed: 34673585]

28. Paradiso A, Marubini E, Verderio P, Cortese ME, De Paola F, Silvestrini R et al. Interobserver reproducibility of immunohistochemical HER-2/neu evaluation in human breast cancer: The real-world experience. Int J Biol Markers 19, 147–154 (2004). [PubMed: 15255548]

29. Pfitzner BM, Lederer B, Lindner J, Solbach C, Engels K, Rezai M et al. Clinical relevance and concordance of HER2 status in local and central testing—an analysis of 1581 HER2-positive breast carcinomas over 12 years. Mod Pathol 31, 607–615 (2017). [PubMed: 29271415]

30. Umemura S, Osamura RY, Akiyama F, Honma K, Kurosumi M, Sasano H et al. What Causes Discrepancies in HER2 Testing for Breast Cancer?A Japanese Ring Study in Conjunction With the Global Standard. Am J Clin Pathol 130, 883–891 (2008). [PubMed: 19019764]

31. Bartlett JMS, Going JJ, Mallon EA, Watters AD, Reeves JR, Stanton P et al. Evaluating HER2 amplification and overexpression in breast cancer. J Pathol 195, 422–428 (2001). [PubMed: 11745673]

32. Casterá C & Bernet L HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH. Ann Diagn Pathol 45 (2020).

33. Griggs JJ, Hamilton AS, Schwartz KL, Zhao W, Abrahamse PH, Thomas DG et al. Discordance between original and central laboratories in ER and HER2 results in a diverse, population-based sample. Breast Cancer Res Treat 161, 375–384 (2017). [PubMed: 27900490]

34. Kaufman PA, Bloom KJ, Burris H, Gralow JR, Mayer M, Pegram M et al. Assessing the discordance rate between local and central HER2 testing in women with locally determined HER2-negative breast cancer. Cancer 120, 2657–2664 (2014). [PubMed: 24930388]

35. Dendukuri N, Khetani K, McIsaac M & Brophy J Testing for HER2-positive breast cancer: a systematic review and cost-effectiveness analysis. Can Med Assoc J 176, 1429 (2007). [PubMed: 17485695]

36. Gibbons-Fideler IS, Nitta H, Murillo A, Tozbikian G, Banks P, Parwani AV et al. Identification of HER2 Immunohistochemistry-Negative, FISH-Amplified Breast Cancers and Their Response to Anti-HER2 Neoadjuvant Chemotherapy. Am J Clin Pathol 151, 176–184 (2019). [PubMed: 30339245]

37. Krystel-Whittemore M, Xu J, Brogi E, Ventura K, Patil S, Ross DS et al. Pathologic complete response rate according to HER2 detection methods in HER2 positive breast cancer treated with neoadjuvant systemic therapy. Breast Cancer Res Treat 177, 61 (2019). [PubMed: 31144151]

38. Lambein K, Van Bockstal M, Vandemaele L, Geenen S, Rottiers I, Nuyts A et al. Distinguishing Score 0 From Score 1+ in HER2 Immunohistochemistry-Negative Breast CancerClinical and Pathobiological Relevance. Am J Clin Pathol 140, 561–566 (2013). [PubMed: 24045554]

39. Schettini F, Chic N, Brasó-Maristany F, Paré L, Pascual T, Conte B et al. Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer. npj Breast Cancer 7, 1–13 (2021). [PubMed: 33397968]

40. Miglietta F, Griguolo G, Bottosso M, Giarratano T, Lo Mele M, Fassan M et al. Evolution of HER2-low expression from primary to recurrent breast cancer. npj Breast Cancer 7, 1–8 (2021). [PubMed: 33397968]

41. Tarantino P, Gandini S, Nicolò E, Trillo P, Giugliano F, Zagami P et al. Evolution of low HER2 expression between early and advanced-stage breast cancer. Eur J Cancer 163, 35–43 (2022). [PubMed: 35032815]

42. Han G, Schell MJ, Reisenbichler ES, Guo B & Rimm DL Determination of the number of observers needed to evaluate a subjective test and its application in two PD-L1 studies. Stat Med 41, 1361–1375 (2022). [PubMed: 34897773]

43. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

44. Han Gang and Guo Baihong (2021). ONEST: Observers Needed to Evaluate Subjective Tests. R package version 0.1.0 https://CRAN.R-project.org/package=ONEST.

45. Reisenbichler ES, Han G, Bellizzi A, Bossuyt V, Brock J, Cole K et al. Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer. Mod Pathol 33, 1746–1752 (2020). [PubMed: 32300181]

46. Quatto Piero and Ripamonti Enrico (2014). raters: A Modification of Fleiss' Kappa in Case of Nominal and Ordinal Variables. R package version 2.0.1 https://CRAN.R-project.org/package=raters.

47. Gamer Matthias, Lemon Jim and Ian Fellows Puspendra Singh <puspendra.pusp22@gmail.com> (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1 https://CRAN.R-project.org/package=irr.

48. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

49. Vörös A, Csörgö E, Nyári T & Cserni G An intra- and interobserver reproducibility analysis of the Ki-67 proliferation marker assessment on core biopsies of breast cancer patients and its potential clinical implications. Pathobiology 80, 111–118 (2013). [PubMed: 23258384]

50. Wells CA, Sloane JP, Coleman D, Munt C, Amendoeira I, Apostolikas N et al. Consistency of staining and reporting of oestrogen receptor immunocytochemistry within the European Union--an inter-laboratory study. Virchows Arch 445, 119–128 (2004). [PubMed: 15221370]

51. Pu T, Shui R, Shi J, Liang Z, Yang W, Bu H et al. External quality assessment (EQA) program for the immunohistochemical detection of ER, PR and Ki-67 in breast cancer: results of an interlaboratory reproducibility ring study in China. BMC Cancer 19 (2019).

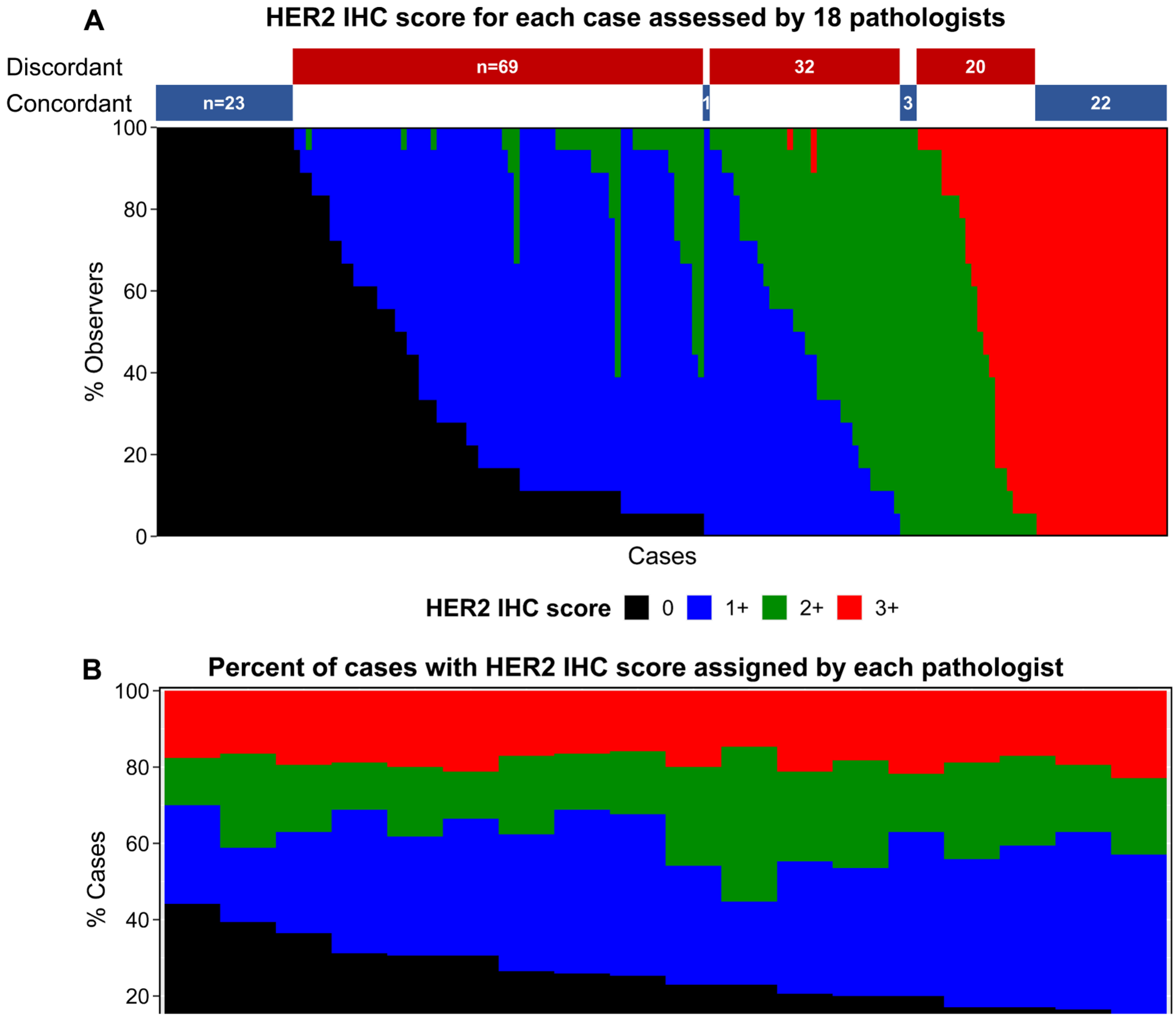

Figure 1: HER2 IHC scores for 170 cases read by 18 pathologists.

HER2 IHC score of the whole tissue sections. **A)** Each case on the x-axis is shown as the percent of observers that called the case HER2 0, 1+, 2+, or 3+. Concordant (100% agreement) and discordant cases are indicated as bars above the plot. **B)** Percent of cases with HER2 IHC score assigned by each of the 18 pathologists.

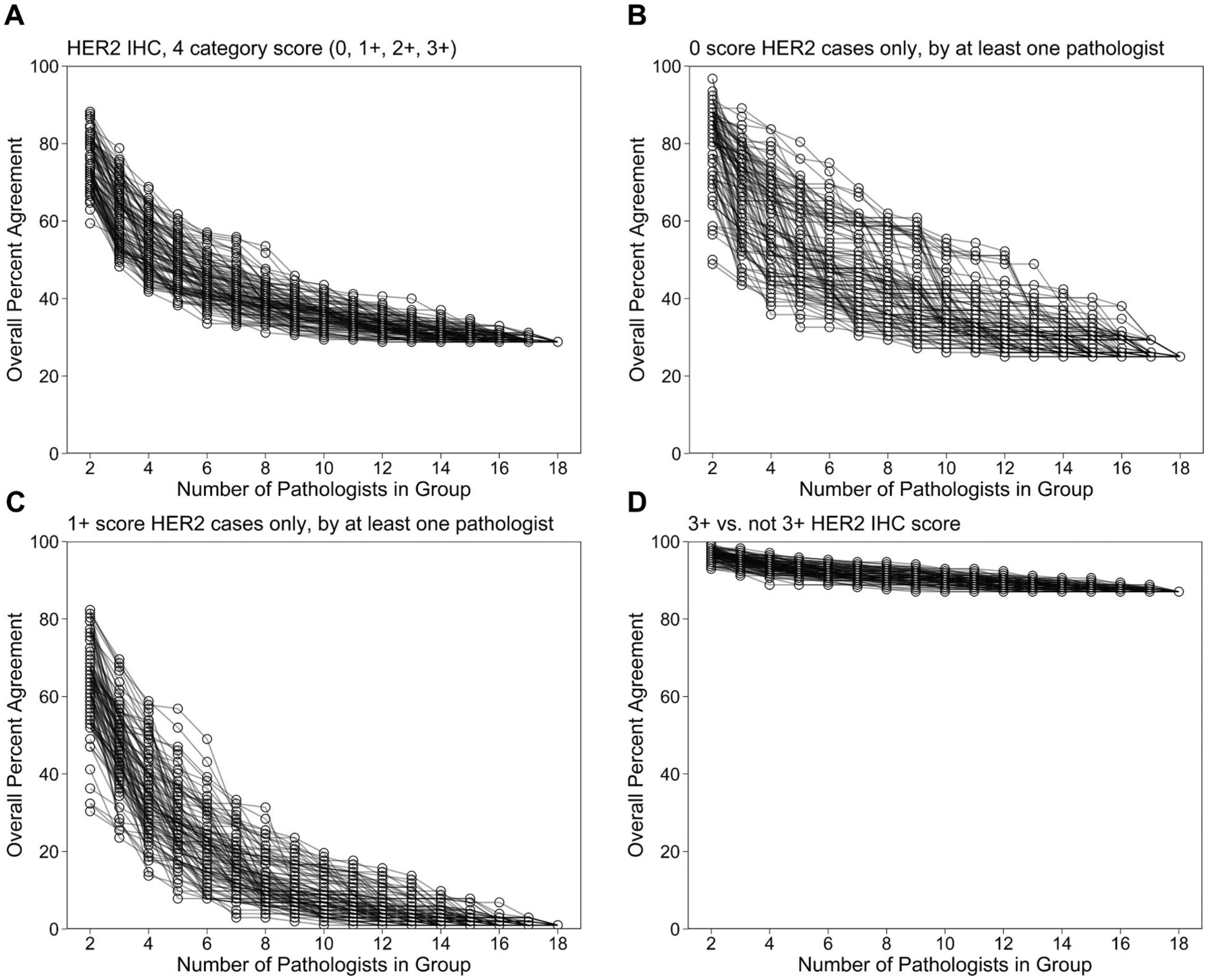

Figure 2: ONEST plots of overall percent agreement for different HER2 IHC groupings.

ONEST plot of HER2 IHC overall percent agreement (OPA) in a four category score (0, 1+, 2+, 3+) **(a)**. OPA ONEST plots for the subset of cases that were read as HER2 IHC 0 **(b)** or 1+ **(c)** by at least one of the 18 pathologist raters. OPA ONEST plots for the task of determining HER2 IHC score of 3+ vs. not 3+ **(d)**. One hundred curves were randomly generated from all possible combinations of pathologists for each HER2 IHC grouping. 0.5%).

Table 1:

Summary of inter-rater reliability metrics for different HER2 IHC groups amongst 18 pathologists in 170 cases of breast cancer

HER2 IHC group	Overall Percent Agreement (95% CI)	Fleiss' Kappa (95% CI)	ICC (95% CI)
4 category (0, 1+, 2+, 3+)	28.82 (22.01, 35.63)	0.65 (0.64, 0.66)	0.88 (0.85, 0.90)
3 category (0, Low[*], 3+)	46.47 (38.97, 53.97)	0.74 (0.73, 0.75)	0.83 (0.79, 0.86)[†]
Only including cases with this score by at least one pathologist			
0 only	25 (16.15, 33.85)	0.49 (0.47, 0.50)	0.55 (0.47, 0.64)
1+ only	0.98 (0, 2.89)	0.35 (0.34, 0.36)	0.52 (0.44, 0.60)
2+ only	3.57 (0, 7.54)	0.46 (0.45, 0.47)	0.67 (0.59, 0.74)
3+ only	50 (35.23, 64.77)	0.63 (0.61, 0.65)	0.69 (0.59, 0.78)
Low[*] only	27.2 (19.4, 35)	0.47 (0.46, 0.48)	0.56 (0.49, 0.63)[†]
0 vs. not 0	59.41 (52.03, 66.79)	0.69 (0.68, 0.70)	0.69 (0.64, 0.74)[‡]
Low[*] vs. not Low[*]	46.47 (38.97, 53.97)	0.69 (0.68, 0.70)	--
3+ vs. not 3+	87.06 (82.01, 92.1)	0.89 (0.88, 0.90)	0.89 (0.87, 0.91)[‡]
< 2+ vs. ≥ 2+	64.12 (56.91, 71.33)	0.77 (0.76, 0.78)	0.77 (0.73, 0.81)[‡]

[*] The Low category is the result of combining the 1+ and 2+ categories.

[†] To calculate ICC for the 3 category score and Low only cases, the IHC scores were converted to 0, 1, or 2 to represent HER2 negative, Low, and 3+ cases respectively.

[‡] To calculate the ICC for these groupings, scores for cases were converted to ordinal scores of 0 or 1 based on increasing IHC score (e.g. IHC 0 or not 0 was converted to 0 and 1 respectively).