Interobserver Reproducibility of Her-2/neu Protein Overexpression in Invasive Breast Carcinoma Using the DAKO HercepTest

Chih-Yi Hsu, MD, MHA, Donald Ming-Tak Ho, MD, FRCPC, FCAP, Ching-Fen Yang, MD, Chii-Ru Lai, MD, I-Ting Yu, MD, and Hung Chiang, MD

Key Words: Her2/neu; Breast cancer; Immunohistochemistry; Interobserver reproducibility

Abstract

Although there are criteria for interpretation of the staining results for Her-2/neu in the DAKO HercepTest, the determination of staining intensity and percentage of complete membrane staining is subjective. We studied 46 cases of invasive breast carcinoma to evaluate interobserver reproducibility among 5 pathologists. Complete agreement was achieved in 22 (48%) of 46 cases. Generalized kappa values indicated substantial agreement (0.80). Discrepancies between negative (0, 1+) and positive (2+, 3+) results occurred in 2 cases (kappa = 0.96). One was because of a tangential cut of the basal part of the tumor that mimicked complete membranous staining, and the other was a borderline case that revealed focal (5%-15%) complete membranous staining. Distinguishing weakly (2+) from strongly (3+) positive results showed agreement in only 13 (59%) of 22 positive cases (kappa = 0.38). If more than 50% of tumor cells revealing strong complete membrane staining were regarded as strongly positive, agreement would be improved (kappa = 0.78). While there was a high percentage (70%-80%) of negative cases during routine evaluation, the good interobserver agreement and high negative predictive value made immunohistochemical analysis an effective screening test to exclude negative cases.

Her-2/neu overexpression occurs in 20% to 30% of patients with breast carcinoma and predicts shortened disease-free survival and poor clinical outcome.1-3 The use of trastuzumab (Herceptin) has been shown to result in clinical responses in patients whose metastatic breast cancer overexpresses HER-2/neu.4 Since trastuzumab was approved by the US Food and Drug Administration (FDA) in 1998, the test for HER-2/neu expression has become important.5 Although there is a variety of methods available to assess HER-2/neu status, assessment of protein overexpression using results of immunohistochemical analysis and evaluation of gene amplification using fluorescence in situ hybridization (FISH) are the methods most commonly used.6-8 There has not been an optimal test until now.9 Immunohistochemical analysis is popular and can be performed in nearly every pathology laboratory. However, the positive rate of HER-2/neu using results of immunohistochemical analysis was reported to vary from 2% to 89%.8-12 Standardization of the HER-2/neu test to accurately evaluate HER-2/neu status is necessary.

The first clinical test for HER-2/neu approved by the FDA is the HercepTest (DAKO, Glostrup, Denmark), which has standard reagents, procedures, and reading criteria.5 Although the reading criteria are clearly defined, subjective determination of staining intensity and the percentage of complete membrane staining is still present. For interpreting the results, scores of 0 and 1+ are negative for HER-2/neu overexpression, while scores of 2+ and 3+ are positive. However, the correlation studies showed that only 3+ scores correlated well with the results of the trastuzumab clinical trial and gene amplification. Scores of 2+ did not correlate well, and a score of 2+ is insufficient for trastuzumab treatment.

© American Society for Clinical Pathology
Materials and Methods

A total of 46 specimens from patients with invasive breast carcinoma were used. The specimens were removed by modified radical mastectomy, diagnosed at the Veterans General Hospital-Taipei, Taipei, Taiwan, during 1988 and 1989, and retrieved from the surgical pathology files of the Department of Pathology and Laboratory Medicine. The corresponding H&E-stained slides from all patients were reviewed. Sections containing tumors were selected for study. The adjacent normal breast parenchyma were included when possible.

The DAKO HercepTest was performed according to the instructions of the manufacturer. The staining results were interpreted independently by 5 pathologists (C.-Y.H., D.M.-T.H., C.-F.Y., C.-R.L., I.-T.Y.) using the criteria set by DAKO. No staining observed or membrane staining of fewer than 10% of tumor cells was scored as 0. A faint or barely perceptible incomplete membrane staining detected in more than 10% of tumor cells was scored as 1+. A weak to moderately complete membrane staining observed in more than 10% of the tumor cells was scored as 2+. A strong complete membrane staining observed in more than 10% of the tumor cells was scored as 3+. Scores of 0 and 1+ were regarded as negative, 2+ as weakly positive, and 3+ as strongly positive. In addition to the scoring, we recorded the details of membrane staining patterns, such as presence, completeness, intensity, and percentage. The presence and completeness were coded as yes or no. The intensity was coded as strong or not strong. The percentage was recorded as fewer than 10%, 10% to 50%, and more than 50%. Any discrepancies of the scoring results were reviewed by the 5 pathologists using a multiheaded microscope to obtain consensus. The consensus opinion was viewed as a conclusive result.

The level of interobserver agreement was quantitated using the generalized kappa and pairwise kappa statistics. The pairwise kappa statistics were the proportion of cases in which the 2 observers agreed, adjusted for the level of agreement that would be expected to occur solely by chance. The generalized kappa was the summary of the agreement across all observers. The kappa values were interpreted following the guidelines described by Landis and Koch. In brief, the greater of the kappa values reflected stronger agreement between the raters. If the kappa value ranged from 0.81 to 1.00, the strength of agreement was almost perfect. When the kappa value ranged from 0.61 to 0.80, it showed substantial agreement. When the kappa value ranged from 0.41 to 0.60, 0.21 to 0.40, or 0.00 to 0.20, the strength of agreement was moderate, fair, or slight, respectively. Clinical data were reviewed and correlated with the pathologic findings.

Results

The age of the patients at diagnosis ranged from 31 to 67 years (median, 50 years). Tumor size ranged from 2 to 8 cm (mean, 4.1 cm). Twenty-eight cases had axillary lymph node metastases. Disease stages varied from II to III. Twenty-three patients died of disease. The median follow-up time was 96 months (range, 3-168 months). The histologic grade varied from 1 to 3. Four cases were grade 1, 28 cases were grade 2, and 14 cases were grade 3.

The consensus scores for the material studied are given in Table I. There were 18 cases (39%) that were scored 0, 6 cases (13%) scored 1+, 5 cases (11%) scored 2+, and 17 cases (37%) scored 3+ by consensus. Numbers of cases for which the pathologists made the same scores as a consensus also are shown. Complete agreement of the scoring results among the 5 observers was achieved in 22 (48%) of 46 cases. Nine cases were scored 0, and the others were scored 3+. In the 18 consensus cases scored 0, 3 pathologists also scored 0 in 5 cases. Four pathologists rated the same score in 4 cases, and 5 pathologists made the same scores in the remaining nine cases. There were 90 individual scores that corresponded to the 18 cases that scored 0. The distributions of individual scores were 0 for 76 times and 1+ for 14 times.

Of 6 cases scored 1+ by consensus, 2 were scored 2+ for 3 times by 2 of the 5 pathologists. While reviewing the consensus results, we found that in 1 case there was a tangential cut of the basal part of the tumor that mimicked complete membranous staining; in another case, a borderline case, focal (5%-15%) complete membranous staining was revealed.

All 5 cases scored 2+ by consensus were scored individually as either 2+ or 3+. No complete agreement was achieved in this category. The cases were scored 3+ for 8 times by 2 of 5 pathologists. The percentages of complete membrane staining in 3 of these 5 cases scored 2+ were recorded as 10% to 50% by all 5 pathologists.

Of 17 cases, 13 (76%) scored 3+ by consensus, which was complete agreement. The remaining 4 cases not achieving complete agreement were scored 2+ for 5 times by 2 pathologists. Except for 1 case, the percentages of
complete membrane staining of all cases that scored 3+ were recorded as more than 50% by all pathologists.

Table 2 shows the proportion of cases assigned to each category of immunohistochemical score stratified by the observers. Disagreement between scores of 0 and 1+ had no clinical importance because both were considered negative. Therefore, we grouped them into the negative category for the following agreement studies. Interobserver variability was reduced when the scores of 0 and 1+ were combined. Good agreement was seen in the negative immunohistochemical category, with assigned percentages ranging from 48% (22/46) to 52% (24/46). The weakly positive category (2+), consisting of 10.9% (5/46) of total cases by consensus, reflected significant differences in the scoring by the observers, ranging from 4% (2/46) to 20% (9/46). The strongly positive category (3+) showed some differences, ranging from 28% (13/46) to 48% (22/46). Generalized kappa values indicated substantial agreement (0.80). The pairwise kappa values for agreement among observers ranged from substantial to almost perfect (0.61 to 0.96).

Table 3. The generalized kappa values for the details of the membrane staining were 0.82 for presence, 0.96 for completeness, 0.80 for strong intensity, and 0.70 for percentage.

Distinguishing negative (0, 1+) from positive (2+, 3+) showed agreement in 44 (96%) of 46 cases (generalized kappa = 0.96). Distinguishing weakly positive (2+) from strongly positive (3+) showed agreement in only 13 (59%) of 22 positive cases (generalized kappa = 0.38). The pairwise kappa value for the interobserver agreement for weakly positive (2+) vs strongly positive (3+) ranged from slight to almost perfect (0.00 to 0.88).

Table 4. As shown previously, most cases that scored 3+ revealed complete membrane staining in more than 50% of the tumor cells, and most cases that scored 2+ revealed staining in 10% to 50%. If we redefined strongly positive as more than 50% of tumor cells revealing strong, complete membrane staining and the other positive cases were regarded as weakly positive, the agreement would be improved (generalized kappa = 0.78). The pairwise kappa value for the interobserver agreement for weakly positive vs strongly positive by this new criterion revealed some improvement (0.60 to 1.00) (Table 4).
We found frequent discrepancies between the scores of 0 and 1+. The differences between 0 and 1+, by definition, were perceptible partial membrane staining detected in fewer than or more than 10% of tumor cells. The agreement values for presence and percentage of membrane staining were almost perfect (0.82) and substantial (0.70), respectively. It was difficult to determine a 10% area with membrane staining, especially when the staining was faint and the areas were discontinuous and small. However, the differences between the scores of 0 and 1+ had no clinical significance. They were both negative for HER-2 protein overexpression; thus, patients were not candidates for trastuzumab treatment.

It was important and not difficult to distinguish the negative (scores 0 and 1+) from positive (scores 2+ and 3+) when following the criteria for the HercepTest. According to the criteria, the main differences in negative (0, 1+) and positive (2+, 3+) staining results depended on whether the membrane staining was complete and whether the complete membrane staining cells constituted more than 10% or fewer than 10%. The interobserver agreement was almost perfect. Of 46 cases, 2 (4%) did not have full agreement in this part. One was because of a tangential cut of the basal part of the tumor, and the other was a borderline case. About the tangential cut, we found that the basal part of the normal duct may reveal the same staining pattern and should not be regarded as positive staining. Some authors suggest that the final scoring should be based on subtracting the score of the normal ductal epithelium from the tumor score. However, the latter scoring method was not used in the HercepTest. The second case revealed focal complete membranous staining in some of the tumor cells over the peripheral area. This area measured 5% to 15% according to different observers. Applying an arbitrary criterion of 10% in a borderline case was not easy for pathologists. We knew that scoring should be made according to the representative areas. Areas that were not well preserved, such as edges, crushes, and thermal artifacts, should be omitted. Estimating the percentage of positive staining area in a case with a heterogeneous pattern was subjective. Fortunately, except for 1 borderline case, the area that revealed complete membrane staining in most of the positive cases was far greater than 10%.

Another major discrepancy was in the discrimination between the scores of 2+ and 3+. The agreement was fair (kappa = 0.38). All of the consensus for positive (scores 2+ and 3+) cases were scored individually, either 2+ or 3+ according to the 5 pathologists. Although the scores of 2+ and 3+ were both reported as positive for HER-2/neu protein overexpression according to the HercepTest, only patients with a score of 3+ were eligible for trastuzumab treatment. Thus, the distinction between 2+ and 3+ was very important. Complete agreement in the strongly positive (3+) category was achieved in only 76% (13/17) of cases. Complete agreement was not achieved in the weakly positive (2+) category. The individual cases were scored 3+ by some pathologists. Only 17% to 37% of weakly positive (2+) cases had gene amplification in previous studies. Preliminary data also suggested that among cases that were scored 3+, only the cases that also showed gene amplification by FISH responded to trastuzumab. When using the criteria for the HercepTest, the only difference between 2+ and 3+ was the intensity of membrane staining. The agreement concerning strong intensity was not perfect among pathologists. The staining intensity of immunostains

### Table 4

<table>
<thead>
<tr>
<th>Criteria/Osver</th>
<th>Observer</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.51</td>
</tr>
<tr>
<td>2</td>
<td>—</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criteria/Osver</th>
<th>Observer</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.88</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>—</td>
<td>0.88</td>
</tr>
<tr>
<td>3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
was affected by many factors. Although positive and negative external controls using cell lines were provided in the HercepTest kit, only controls for scores of 0, 1+, and 3+ were provided. There was no 2+ control available. Furthermore, no suitable internal control was attainable. Pretaining factors that affected staining results, such as latent time from tumor excision until fixation and tissue processing, could not be tested using external control slides. The only possible negative internal control was normal ductal epithelium, most of which stained as 0 or 1+. However, normal ductal epithelium was not included in every specimen. We did not find any positive internal control in normal breast tissue, let alone an internal control for scores of 2+ and 3+.

The quantity of positive cells was not considered when differentiating scores of 2+ and 3+ in the HercepTest. Results of previous studies showed that the consistency of determining the presence of membranous staining was high, but the consistency of determining staining intensity was poor.22 We found that the presence, intensity, and percentage of membrane staining achieved substantial to almost perfect agreement. However, all 5 pathologists recorded the percentage of the positive area as 10% to 50% in 3 of 5 weakly positive (2+) cases and more than 50% in 16 of 17 strongly positive (3+) cases. Some scoring systems had different views in this part.15,21,22 They considered intensity and quantity of the positive cells to distinguish strongly and weakly positive cases. In this way, strongly positive (3+) cases were defined as more than 50% of tumor cells revealing strong complete membrane staining. Tumors showing 10% to 50% or weaker complete membrane staining were rated as weakly positive (2+). Combining the quantity and quality of positive cells to differentiate 2+ and 3+ would improve the consistency. Three of 5 weakly positive (2+) and 3 of 4 strongly positive (3+) cases that had discrepancies showed complete agreement in our study. Interobserver agreement also improved. In a previous study using this type of criteria, the correlation with FISH was good.21 Some computer assistance or image analysis systems to improve scoring accuracy have been designed.23-25 Better results were achieved; however, those systems are not available in every laboratory and are not suitable for daily practice.

Immunohistochemical analysis is popular and can be performed in nearly every pathology laboratory. Compared with FISH, immunohistochemical analysis is cheaper, less technically challenging, and less time-consuming. Furthermore, it seems more logical to determine HER-2/neu protein overexpression using immunohistochemical analysis than gene amplification using FISH when trastuzumab treatment is targeted toward the HER-2/neu protein on the cell surface. However, the variability among different antibodies and staining methods and the subjective scoring systems make FISH more reliable. Thus, comprehensive standardization of methods and the inclusion of validated controls to be applied in immunohistochemical analysis are mandatory. A stable and reliable internal control, such as that provided in the FDA-approved PathVysion HER-2 DNA Probe Kit (Vysis, Downers Grove, IL), is not available in immunohistochemical stains for HER-2/neu. The DNA probe in FISH, chromosome enumeration probe 17, is less affected by tissue fixation and processing. A standardized external control for immunohistochemical analysis such as a formalin-fixed, paraffin-processed cell line has been designed.26 This could serve as a more standard base to evaluate the interlaboratory differences. According to the National Surgical Adjunctive Breast and Bowel Project B31 clinical trial, interlaboratory reproducibility will improve when the same staining kit is used in qualified laboratories.27 Although HercepTest is neither the best nor the only suitable immunohistochemical kit for the HER-2/neu test, it provides standard reagents, procedures, controls, and a scoring method. Otherwise, good quality assurance procedures should be set up when in-house assays are used.

The kit assay protocol and the scoring method of the HercepTest should be followed strictly. Local modifications of techniques can lead to false-positive and false-negative assay results. The scoring method provided with the kit uses a semiquantitative system based on the intensity and percentage of positive cells. Interobserver variation in the assessment of staining can lead to misclassification of HER-2/neu status. Each individual should standardize scoring against known positive and negative cases, and interobserver agreement will improve. In our interobserver reproducibility study, we found it was easy and consistent to distinguish negative (0, 1+) from positive (2+, 3+) HER-2/neu overexpression using standard immunohistochemical methods with appropriate quality controls such as the HercepTest. By combining quality and quantity assessment, we achieved more convincing results in strongly positive cases. As shown in previous reports, 70% to 80% of patients with invasive breast carcinomas are HER-2/neu–negative and the negative predictive value for immunohistochemical analysis is very high; thus, using immunohistochemical analysis as a screening test is effective.28-30 Further confirmation of the positive cases using FISH is reasonable.16,17,31,32

From the Departments of Pathology and Laboratory Medicine, Veterans General Hospital-Taipei, Taipei, Taiwan, and Pathology, National Yang-Ming University School of Medicine, Taipei, Taiwan, Republic of China.

Address reprint requests to Dr Hsu: Dept of Pathology and Laboratory Medicine, Veterans General Hospital-Taipei, 201 Shih-Pai Rd, Sec 2, Taipei, 11217 Taiwan, Republic of China.
References


