Interobserver Agreement Among Pathologists for Semiquantitative Hormone Receptor Scoring in Breast Carcinoma

David A. Cohen,1 David J. Dabbs, MD,1 Kristine L. Cooper, MS,2 Milon Amin, MD,1 Terrell E. Jones,1 Mirka W. Jones, MD,1 Mamatha Chivukula, MD,1 Giuliana A. Trucco, MD,1 and Rohit Bhargava, MD1

Key Words: Semiquantitative H-score; Interobserver agreement; Hormone receptors; Breast carcinoma

DOI: 10.1309/AJCP6DKRND5CKVDD

Abstract

The American Society of Clinical Oncology (ASCO) in conjunction with the College of American Pathologists (CAP) recently produced guidelines for estrogen receptor (ER) and progesterone receptor (PR) testing in breast cancer.1 The 2010 guidelines recommend that ER and PR should be recorded in a semiquantitative manner. This decision reflects the consensus view that the degree of ER staining is a valuable prognostic indicator of breast cancer aggressiveness and treatment strategy choice. With its widespread use since the 1990s, immunohistochemistry has largely replaced ligand-binding assays because of its superiority in providing prognostic value, ability to be performed on small tissue sections, and lower cost.2-4

The American Society of Clinical Oncology (ASCO) recommends reporting of hormone receptor test results in a semiquantitative manner. This study used 74 resected estrogen receptor (ER)-positive invasive breast cancers to determine reproducibility of semiquantitative scoring of hormone receptors using the H-score method. Four pathologists independently scored each slide. Agreement among observers was analyzed via Fleiss κ statistics on ER and progesterone receptor (PR) categorical scores. Intraclass correlation coefficient (ICC) was used to estimate the interobserver agreement for ER and PR H-scores on a continuous scale (0-300). There was 100% agreement for categorical ER results (κ = 1) and 97% agreement (κ = 0.823, P < .001) for categorical PR results. For quantitative H-scores, ICC agreement was 0.85 (95% confidence interval [CI] = 0.79-0.90) for ER and 0.87 (95% CI = 0.82-0.92) for PR. Because the H-score provides a continuous measure of tumor hormone receptor content, we suggest universal adoption of this method.
“H-score.”7 The H-score for breast cancer ER status provides an overall score (0-300) based on the sum of ordinal weighted percentiles of cells stained weak, moderate, and strong. Some institutions, including our own, prefer the H-score because of its wide dynamic range and use of weighted percentiles.

With the preference for semiquantitative reporting, there is a renewed interobserver agreement in the reporting structure of these methods. Currently, there is a dearth of studies for evaluating postanalytic interobserver agreement for semiquantitative hormone receptor scoring methods. This study reports the fidelity of interobserver variability for semiquantitative hormone receptor scoring using the H-score in a large academic women’s center with breast specialty sign-out.

Materials and Methods

Tissue Specimens and Immunohistochemistry

A total of 74 resected invasive breast cancer specimens, previously ER+ on core biopsy, were obtained from the pathology files of Magee-Womens Hospital of the University of Pittsburgh Medical Center (Pittsburgh, PA). All specimens were fixed in 10% neutral phosphate-buffered formalin according to ASCO/CAP guidelines. Hormone receptor testing was performed on 4-μm-thick whole-slide tissue sections using ER clone SP1 and PR clone 1E2 and using iVIEW detection on the Benchmark XT system (Ventana, Tucson, AZ).

H-Score and Interobserver Agreement

Hormone receptor immunohistochemical semiquantitation was performed using the H-score.7-9 The H-score is given as the sum of the percentage of staining multiplied by an ordinal value corresponding to the intensity level (0 = none, 1 = weak, 2 = moderate, 3 = strong). With 4 intensity levels, the resulting score ranged from 0 (no staining in the tumor) to 300 (diffuse intense staining of the tumor). Examples of H-scores can be seen in our recently published study.10 In accordance with ASCO/CAP guidelines, an H-score of 1 or more was considered a positive cutoff for ER and PR.1 Four breast pathologists independently, in a blinded manner, scored each slide and recorded the H-scores. Three of the observers (G.A.T., M.W.J., and R.B.) scored all 74 cases. One observer (M.C.) scored 71 cases. The 4 pathologists have an average experience of 14 years (range 6-23 years) in breast pathology. To avoid any inadvertent bias in estimating H-score, each pathologist electronically mailed his or her scores to a fifth pathologist (D.J.D.) who kept the data until the results were forwarded to the statistician (K.L.C.).

Statistical Analysis

ER and PR were evaluated as continuous scores (ER, PR scale: 0-300) and as categorical scores (negative: H-score <1, positive: H-score ≥1). The intraclass correlation coefficient (ICC) was used on the continuous scores to estimate the agreement among the observers.11-13 Assuming a 2-way random effects model, the ICC was used to consider the random sample of n slides all scored by a random sample of k observers. The ICC is an estimate of the proportion of variation that results from the object being measured. If the agreement among raters is perfect, then the total variation would be explained by the variation of the measurement and the ICC would be 1. Under the null hypothesis of no agreement, the ICC estimate is 0. ICC(A,1) measures the absolute agreement of the raters’ scores. In contrast, ICC(C,1) measures the consistency of the scores. For example, if the k raters all score a particular sample relatively higher than the other samples, then their scores are consistent even if the actual values themselves are different. Two-sided tests and 95% confidence intervals are reported, and the statistical analysis was performed using the software R version 2.14.2 (2012-02-29; R Foundation for Statistical Computing, Vienna, Austria). The Fleiss κ statistic was used to estimate the agreement of multiple raters for categorical data. A Fleiss κ value was interpreted as described earlier.14-16 A Fleiss κ score equal to 0.41 to 0.60 was regarded as a moderate agreement, 0.61 to 0.80 as good agreement, and 0.81 to 1.00 as very good agreement.

Results

Three of the observers scored 74 cases. One observer scored 71 cases. Although the majority of ER+ cases demonstrated diffuse strong expression, variable H-score range was noted rather than a bimodal distribution. The mean (for 4 observers) percentage of cases with ER H-score of 100 or less was 4.8% (range, 1.4%-8.1%), ER H-score of 101 to 200 was 21.5% (range, 13.5%-27%), and ER H-score of more than 200 was 73.7% (range, 64.9%-89.4%). The mean (for 4 observers) percentage of cases with PR H-score of 100 or less was 38.3% (range, 33.8%-45%), PR H-score of 101 to 200 was 40.6% (range, 36.6%-45.9%), and PR H-score of more than 200 was 21.1% (range, 18.3%-25.7%).

For ER H-scores analyzed as categorical scores (H-score of 1 or more as positive), the agreement among all raters was perfect. The Fleiss κ statistic was 1. For PR categorical scores, the agreement was 97%, with a κ score of 0.823 (P < .001). Of the PR results for which there was disagreement (2 cases), the cases showed weak expression in less than 10% of the tumor cells (H-scores ranging from 1-7) according to 2 observers. The other 2 observers either failed to see these cells or judged these to represent less than 1% of the entire tumor.
An invasive ductal carcinoma showing an area of weakly positive cells for progesterone receptor. This case was interpreted as positive by 2 pathologists and negative by 2 other pathologists. The brown staining in the lower left corner in A is hemosiderin (A, ×40; B, ×200).

Table 1
Intraclass Correlation Coefficients for Agreement and Consistency Among 4 Pathologists

<table>
<thead>
<tr>
<th></th>
<th>ICC (Agreement) (95% CI)</th>
<th>ICC (Consistent) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrogen receptor</td>
<td>0.85 (0.79–0.90)</td>
<td>0.86 (0.80–0.90)</td>
</tr>
<tr>
<td>Progesterone receptor</td>
<td>0.87 (0.82–0.92)</td>
<td>0.89 (0.84–0.92)</td>
</tr>
</tbody>
</table>

CI, confidence interval; ICC, intraclass correlation coefficient.

Table 2
Correlation Coefficients for Each Observer Pair for ER

<table>
<thead>
<tr>
<th>ER</th>
<th>Observer 1</th>
<th>Observer 2</th>
<th>Observer 3</th>
<th>Observer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 1</td>
<td>1.000</td>
<td>0.860</td>
<td>0.839</td>
<td>0.843</td>
</tr>
<tr>
<td>Observer 2</td>
<td>0.860</td>
<td>1.000</td>
<td>0.879</td>
<td>0.912</td>
</tr>
<tr>
<td>Observer 3</td>
<td>0.839</td>
<td>0.879</td>
<td>1.000</td>
<td>0.908</td>
</tr>
<tr>
<td>Observer 4</td>
<td>0.843</td>
<td>0.912</td>
<td>0.908</td>
<td>1.000</td>
</tr>
</tbody>
</table>

ER, estrogen receptor.

Table 3
Correlation Coefficients for Each Observer Pair for PR

<table>
<thead>
<tr>
<th>PR</th>
<th>Observer 1</th>
<th>Observer 2</th>
<th>Observer 3</th>
<th>Observer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 1</td>
<td>1.000</td>
<td>0.871</td>
<td>0.896</td>
<td>0.867</td>
</tr>
<tr>
<td>Observer 2</td>
<td>0.871</td>
<td>1.000</td>
<td>0.893</td>
<td>0.905</td>
</tr>
<tr>
<td>Observer 3</td>
<td>0.896</td>
<td>0.893</td>
<td>1.000</td>
<td>0.921</td>
</tr>
<tr>
<td>Observer 4</td>
<td>0.867</td>
<td>0.905</td>
<td>0.921</td>
<td>1.000</td>
</tr>
</tbody>
</table>

PR, progesterone receptor.

shows a case with scattered weakly positive cells, which was scored as negative by 2 observers and positive by the other 2 observers, and had H-scores of 1 and 3.

Table 1 shows the ICC for agreement (2-way random effects model measuring absolute agreement of values) and consistency (2-way random effects model measuring consistency of scores) for quantitative H-scores on ER and PR for which the proportion and intensity of positively stained nuclei were assessed. The overall ICC for agreement among all 4 pathologists was very good, with an ICC agreement of 0.85 and 0.87 for ER and PR, respectively. Table 2 and Table 3 show the correlation coefficients for each observer pair among the 4 pathologists scoring ER and PR, respectively. The data for observer pairs for ER and PR are illustrated in Figure 1 and in Figure 2, respectively.

Discussion

ER and PR status have an established prognostic value for responsiveness of invasive breast cancer to endocrine therapy. Patients with ER-positive tumors treated with tamoxifen have a substantially reduced risk of recurrence and an overall survival advantage. Results from the Early Breast Cancer Trialists’ Collaborative Group study showed that tamoxifen strongly benefits ER-positive patients, whereas women with ER-negative results do not benefit. PR expression, while not as strong, provides additional clinical value independent of ER levels, especially in premenopausal women. Since the early 1990s, immunohistochemistry has largely replaced...
Figure 1  Representation of H-scores for estrogen receptor between different observer (OB) pairs.

Figure 2  Representation of H-scores for progesterone receptor between different observer (OB) pairs.
ligand-binding assays. The trade-off was replacement of a quantitative method of measurement with a qualitative chromogenic assay.

The recently released ASCO/CAP guidelines recommend reporting of estrogen and progesterone hormone receptor results in a semiquantitative manner. The guidelines suggest a semiquantitative vs dichotomous method of reporting because of the superior prognostic value in the degree of immunohistochemical ER and PR staining. The degree of positivity corresponds significantly with breast cancer response to hormonal therapy in the time to recurrence and overall mortality reduction. Moreover, combination of standard histologic and semiquantitative immunohistochemical results for prognostic/predictive markers in breast cancer can help prognosticate individual patient risk and also has the potential to predict standard chemotherapy benefits. However, a standard for reporting staining in a composite score could not be agreed upon in the ASCO/CAP. Moreover, there exists a paucity of studies evaluating assay reproducibility for the semiquantitative immunohistochemistry reporting. Recent CAP checklist guidelines for laboratory accreditation require that the medical director demonstrate reproducibility of hormone receptor semiquantitation.

Previous immunohistochemical studies of interobserver reproducibility for ER and PR status have shown inconsistent results. Often, interobserver variance may be masked by different laboratory protocols used among scoring pathologists. Parker et al found moderate interlaboratory agreement, with an overall κ score of 0.54 for 0 to 3+ categorical staining and a κ score of 0.84 for dichotomous reporting of ER. Other studies have found high levels of erroneous negative hormone receptor reporting. An earlier study on interlaboratory variance in reporting ER staining of low to medium intensity found a false-negative rate of 30% to 60% in this range among 200 laboratories in Europe (the NEQAS-ICC consortium). Other authors have similarly found low agreement among poorly staining tumors, which represent a small population of patients who respond to hormone therapy. Based on this concern of false-negative reporting, National Comprehensive Cancer Network guidelines recommend retesting ER-negative low-grade, lobular, tubular, and mucinous breast tumors. In contrast, other studies have found high concordance. In the original study by Harvey et al, in which 2 pathologists used the Allred method for scoring, the weighted κ statistic for concordance was 0.87 (P < .0001). In 2 large quality assurance Canadian studies, the target κ values were greater than 0.8 among 18 laboratories for more than 85% of results. The Eastern Cooperative Oncology Group E2197 study also found a 90% concordance among immunohistochemical results compared with a central laboratory. All of these studies used microarrays or virtual microscopies disseminated to laboratories for independent scoring. None of these concordance studies specifically examined H-score composite scoring.

The H-score provides a wide dynamic range (0-300) for reporting ER and PR assays compared with the Allred score (0-8). This provides a theoretically higher level of prognostic information to clinicians. Furthermore, certain breast cancer specimens have a variation in staining intensity among tumor cells, which is factored into the H-score. The Allred score allows only 1 degree of scoring for intensity per specimen. For these reasons, certain institutions prefer to report the H-score over other commonly used methods such as the Allred score and the “quick” score.

In this study, 74 resected specimens were independently scored by 4 pathologists in a blinded manner with excellent interobserver agreement for categorical reporting of ER and PR results. By using the same slides, only variation among scoring by pathologists was measured. Agreement among H-score numerical values for ICC agreement was also very good, suggesting reproducibility in reporting a continuous variable score along a dynamic range of 0 to 300. For PR immunohistochemistry results, the disagreement rate was low (3%), with a corresponding κ score of 0.823. Other studies have also shown less robust correlation with PR staining, likely because of the historical increased preanalytic variance with PR antibodies. In our study, the PR disagreement was because of focal weak expression of tumor cells considered to be more than 1% by 2 observers and considered less than 1% by 2 other observers. Limitations of our study include the assessment of interobserver agreement among specialized breast pathologists from a single institution, relatively small sample size (n = 74), and low number of pathologists (n = 4).

Routine H-scoring for hormone receptors definitely has advantages but also poses some challenges in terms of pathologist training and performance review. Automated image analysis systems have been suggested as an alternative to human scoring. Nassar et al found that in 260 breast tissue specimens, digital image analysis produced substantially equivalent scoring of ER/PR compared with manual microscopy. However, such methods have not been approved by the US Food and Drug Administration, and performance studies are still needed. Moreover, the image analysis systems that are currently used for analyzing immunohistochemical ER/PR slides are incapable of reporting H-score results. Most systems provide accurate assessment of percentage of positive cells in the field of view, which is generally a very small area of the tumor unless the whole slide is scanned and interpreted. Whole-slide scanning takes time, and even if it is performed the pathologist has to make sure the invasive tumor cells are counted by the computer for reporting. Improvement in image analysis technology may overcome some of these aspects. Image analysis systems using fluorescent probes have also been described as a method of identifying immunoreactive...
tumor cells with an increased interobserver reproducibility. However, the challenge of correctly identifying invasive carcinoma using dark-field microscopy and the increased time in doing such an analysis can further impede incorporation into diagnostic pathology. Furthermore, efforts have been made to use mRNA expression levels of ER/PR through quantitative reverse transcription polymerase chain reaction (PCR) as an alternative to immunohistochemistry. The increased use of Oncotype DX (Genomic Health, Redwood City, CA) and MammaPrint (TargetPrint, Agenda, Irvine, CA) diagnostic testing may increase the popularity of these methods. However, as noted by the ASCO/CAP committee, there is a paucity of data on the concordance between messenger RNA–based assays and immunohistochemistry-based clinical validation studies. We recently reported good agreement between immunohistochemistry semiquantitative H-score results and Oncotype DX quantitative ER/PR quantitative reverse transcription PCR results, with immunohistochemistry being slightly more sensitive than PCR for both ER and PR.10 We did not find any reason to replace immunohistochemistry with PCR. Moreover, because of the speed, higher sensitivity, preservation of morphology, and widespread use in every pathology laboratory, immunohistochemistry is more desirable.

In conclusion, we report excellent interobserver agreement among breast pathologists for reporting ER and PR semiquantitative immunohistochemical results via H-score. Because of the proven advantage of quantification of hormone receptors over qualitative results, we recommend universal adoption of such methods for reporting ER and PR immunohistochemical results.

References


