Virtual Double Staining: A Digital Approach to Immunohistochemical Quantification of Estrogen Receptor Protein in Breast Carcinoma Specimens

Nina Lykkegaard Andersen, MD, Anja Brügmann, MD, PhD, Giedrius Lelkaitis, MD, Søren Nielsen, HT, Michael Friis Lippert, MSc, and Mogens Vyberg, MD

Abstract: Visual assessment of immunohistochemically detected estrogen receptor protein is prone to interobserver and intraobserver variation due to its subjective evaluation. The aim of this study was to validate a new image analysis system based on virtual double staining (VDS) by comparing visual and automated scorings of ER in tissue microarrays of breast carcinomas. Tissue microarrays were constructed of 112 consecutive resection specimens of breast carcinomas. Immunohistochemistry assays for ER and pancytokeratin were applied on separate serial sections. ER scoring was visually performed by 5 observers using the histoscore (H-score) method. The Visiopharm ER image analysis protocol (APP) software application using VDS technique was applied separating stromal cells from carcinoma and other epithelial cells based on the pancytokeratin reaction. Using color deconvolution, polynomial filters, and nuclear segmentation the APP determined the percentage of positive cells and their intensity, and calculated the resulting H-score. On the basis of 1% cutoff VDS was perfectly correlated with visual assessment (k = 1). Using H-score, a very high agreement between VDS and visual ER assessment was seen (R² = 0.950). Image analysis has the attributes to eliminate the shortcomings of visual ER evaluation by generating automated, reproducible, and objective results of ER assessment.

Key Words: breast cancer, estrogen receptor, image analysis, virtual IHC double staining, histoscore

In patients with early breast cancer and positive estrogen receptor (ER) status, 5 years treatment with endocrine therapy has been shown to lower the risk of recurrence and mortality significantly.1,2

According to American Society of Clinical Oncology and the College of American Pathologists, weak expression of ER as determined by immunohistochemistry (IHC) in ≥1% of the tumor nuclei in a breast carcinoma sample is sufficient to predict a significant benefit from endocrine therapy,3,4 and is recommended as cutoff value. For ER-positive carcinomas treatment efficacy correlates with ER-expression levels as shown by longer overall survival and disease-free survival (DFS).5-8

This indicates, that exact quantitative assessment of ER status in breast carcinoma is a clinical relevant factor when planning antihormonal therapy and assessing therapeutic response and survival.

Several methods of semiquantitative assessment of IHC-based ER expression have been introduced. The current guideline by the American Society of Clinical Oncology and the College of American Pathologists recommend using a scoring method including both proportion of stained tumor nuclei and the intensity of staining. This allows for composite scoring as described for Allred score and H-score.9

Visual scoring methods are by definition based on subjective assessments of ER and therefore prone to interobserver and interobserver variation.10 Observer variation is especially a concern in cases with low ER expression due to the fact that subtle differences in staining intensity is difficult for the human eye to detect, when using a continuous scale in composite scoring systems.11,12

Image analysis (IA) has the attributes to eliminate the shortcomings of visual ER evaluation by generating automated, reproducible, and objective results of ER assessment.13

Several validation studies of different IA applications for ER detection in breast carcinomas, have been published within the last 15 years. In general, good concordance has been seen between visual ER assessment and assessment by the different IA software systems.14-22

However, while the trained pathologist can differentiate between epithelial and stromal components in ER-stained sections, IA generally lacks this ability. To avoid the inclusion of the stromal component the region of interest (ROI) must be outlined manually, which may be cumbersome and imprecise. Alternatively, the carcinoma cells must be identified using pattern recognition systems, which pose many challenges and are generally not available.23 A third principle is IHC double staining using ER in combination with, for example, pancytokeratin (PCK).
When double staining is performed physically on 1 tissue section, overlapping chromogens make a precise assessment difficult both by visual scoring and IA.\textsuperscript{24,25}

A software system developed by Visiopharm (Hoersholm, Denmark) implements a VDS technique to distinguish between ER-positive epithelial and stromal cell nuclei by aligning digitized serial sections stained for ER and PCK. Only the ER-positive epithelial invasive tumor cells contribute to the reported ER percentage. The PCK-stained slide identifies positive tumor cells within the invasive tumor component and is superimposed on the ER-positive slide the ratio of ER-positive nuclei can be quantified. Still, normal epithelial cells and carcinoma in situ must be excluded by defining ROIs.

The aim of this study was to evaluate the performance of Visiopharm VDS software by comparing automated scores with visual scores of ER in tissue microarray (TMA) applied breast carcinomas specimens.

**MATERIALS AND METHODS**

One hundred twelve consecutive invasive breast carcinoma specimens were included in this study. The material was collected in the period 2007 to 2008 and archived at the Institute of Pathology, Aalborg University Hospital, Aalborg, Denmark.

The patients were from 27 to 72 years of age (mean, 57 y; median, 59 y) at the time of diagnosis. Malignancy grade I was seen in tumors of 54 women (48%). Thirty-four (30%) and 28 (25%) tumors had malignancy grade II and III, respectively. There were 79 (71%) ductal carcinomas, 15 (13%) were lobular carcinomas, and 18 (16%) had other types.

The samples spanned a broad range of ER expression as determined by the pathologist at routine sign-out (ie, 0% to 100%).

The cohort was divided at random into halves using one half as a training set and the other half as a test set. The training set was used to calibrate the software.

**TMA Construction**

Two TMAs each containing 58 tissue cores were constructed as previously described by Brügmann et al.\textsuperscript{26} Controls constituted a TMA from NordiQC run B8, 2009 including; 1 uterine cervix expressing 80% to 90% ER.

The VDS APP works with a 3-step process as follows (Fig. 2). First, alignment of the images of adjacent slides from the TMAs stained for ER and PCK. The alignment was done both on a large, whole-core scale, and on a finer detail level, to get the best possible match of the 2 tissue sections. Second, the areas containing breast cancer cells were automatically detected from the PCK-stained tissue slide and outlined as ROIs.

The ROIs were superimposed on to the aligned ER tissue slide to outline the tumor region for subsequent analysis limited to the tumor regions. Third, segmentation of all nuclei on the ER-stained image was performed in the ROIs.

The detection of nuclei was based on approximation of polynomial blob and linear filtering of the images highlighting nuclei and nuclear edges, respectively. On the filtered image nuclear separation algorithm based on shape, size, and nuclei probability was used, employing a fully automated watershed-based nuclear segmentation technique to separate positive and negative nuclei from the background.

To quantify the staining intensity color deconvolution algorithms was performed to highlight the 3,3’-Diaminobenzidine (DAB)-stained areas of the tissue and thereby distinguish positive nuclei from negative. Positive nuclei were subdivided into 3 categories based on mean expression as determined by the pathologist at routine sign-out (ie, 0% to 100%).
staining intensities of deconvolved pixels within each nucleus.

The color-deconvolved values are represented on a scale from 0 to 255, with 0 being completely DAB saturated and 255 being devoid of any hint of DAB stain. Thresholds were set and adjusted over multiple algorithm adjustment sessions by 2 pathologists (M.V. and G.L.) based on the training set (Table 2). The intensity of the nuclei and their percentage distribution was used to calculate the H-score as mentioned above.

In addition, the percentage of positive ER cells was calculated by dividing $100 \times \text{the number of positive nuclei}$ with the total number of nuclei.

**Statistical Analysis**

Interobserver agreement and the comparison between VDS and visual detected H-score was assessed through linear regression analysis and Cohens $\kappa$. $\kappa$ statistics evaluates the level of agreement adjusted for agreement expected to occur by chance alone. A $\kappa$ agreement > 0.8 corresponds to an almost perfect agreement as proposed by Landis and Koch.30

Bland-Altman analysis was used to visualize the overall agreement between H-score from VDS and visual assessment, respectively.

**RESULTS**

The training set was only used to calibrate the software and determine the thresholds.

Therefore, the results account for 58 cores. There were 3 control tissue cores. Ten cores from the test TMA were excluded due to insufficient tumor material and 2 cores were without material leaving 45 cores adequate for assessment.

Visual evaluation was carried out by 5 observers resulting in H-scores ranging from 0 to 295 (Fig. 3).

On the basis of the recommended cutoff value in clinical settings today ($\geq 1\%$ ER-positive nuclei) agreement between the VDS APP and mean visual evaluation was excellent and both systems assessed 5 cores as negative and 40 as positive ($\kappa = 1$), corresponding to perfect agreement.

When applying a 10% cutoff value which was used as standard prior, the agreement between the VDS APP and mean visual evaluation was lower ($\kappa = 0.91$) due to 1 core visually detected as ER negative but ER positive by the VDS APP, however, still an almost perfect agreement.

Interobserver Agreement

When applying percentage of positive cells as measurement ($\geq 10\%$ ER-positive nuclei) interobserver agreement was excellent ($\kappa = 0.88$) (data not shown).

But when H-scores were applied interobserver variations in H-score values was seen.

Interobserver agreement is visualized in Figure 3 (blue plots). Each observer is compared with the other observers individually on a single-core basis from the test set. Despite an overall good agreement between the observers with a linear regression coefficient of determination, $R^2$ of 0.938 (data not shown), the figure shows several substantial outliers. As an example 1 core was

<table>
<thead>
<tr>
<th>TABLE 1. Protocol Settings for Antibodies Used</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clone and Vendor</strong></td>
</tr>
<tr>
<td>Estrogen receptor protein</td>
</tr>
<tr>
<td>Pancytokeratin</td>
</tr>
</tbody>
</table>

**FIGURE 1.** Tumor cells in the rectangles A to D represent examples of the intensity scale used for estrogen receptor staining reaction in breast biopsy sections: A: 0 (negative), B: 1+ (weak), C: 2+ (moderate), and D: 3+ (strong) nuclear staining. (Original magnification, ×40).
scored to an H-score of 155 by 1 observer and 240 by another.

Figure 3 also shows the bimodal distribution of the test set with only 4 cores receiving a H-score within the range of 50 to 100.

The overall agreement between mean observer H-score and VDS H-score resulted in $R^2 = 0.950$ (Fig. 3).

However, when visualized through the Bland-Altman plot substantial outliers are seen in the high expression range of H-score 175 to 275 (Fig. 4). Fourteen cores are above the positive SD line and 2 are under the negative SD line illustrating a tendency toward VDS over scoring compared with visual assessment.

The 8 highest outliers with a difference in H-score of 35 or higher (35 to 83) were reevaluated by 3 observers (M.V., A.B., N.L.A.).

They were characterized by a high staining intensity and thus a high H-score (VDS H-score ranging from 198 to 286 and mean visual H-score: 142 to 246). Noticeable high interobserver variation was seen especially when determining the highest ER nuclei intensity 3+ the widest range was 5% to 50%.

**DISCUSSION**

In this study we investigated a new software system based on VDS technique to measure the expression of ER in breast carcinoma specimens. We compared VDS H-score with visual assessed H-score and found an overall excellent correlation between visual and automated ER assessment.

The result of our study follows the tendency of other studies testing various IA systems but due to their significant variation in methodologies and technical properties, a direct comparison is difficult.18-20,31

In reference to the clinical setting of 1% cutoff level the VDS APP had an excellent correlation with visual assessment ($\kappa = 1$). When applying a 10% cutoff which was the previous cutoff level the agreement was a little lower due to fact that one core detected as ER negative by visual assessment was found ER positive by the VDS APP ($\kappa = 0.91$).

The group of patients displaying 1% to 9% ER positivity is a complex group due to divergent opinions regarding treatment efficacy. Raghav et al32 described that patients with breast cancer tumors expressing 1% to 5% ER receiving endocrine therapy did not have a significant effect in terms of 3-year recurrence-free survival. Patients expressing 6% to 10% had a marginal effect.

Viale et al4 found a statistically significant different 5 years DFS of the 1% to 9% group compared with the ER absent group and $\geq 10\%$ group, respectively.

In opposition, Balduzzi et al33 did not find significant difference between the $<1\%$ group and the 1% to 10% group in regards to 5-year DFS.

With these clinical outcome studies in mind, combined with potential side effects from endocrine therapy the group of patients displaying ER 1% to 9% positivity is very important to detect with high accuracy.

The VDS APP performance had a tendency of scoring cores with high ER expression to a higher H-score compared with visual assessment. It was based on a number of calibration sessions with 2 observers evaluating cores from the training set, and dividing them according to the DAB intensity scale. In retrospect, a larger number of

### TABLE 2. Intensity Thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0;50]</td>
<td>3+ positive nuclei</td>
</tr>
<tr>
<td>[50;125]</td>
<td>2+ positive nuclei</td>
</tr>
<tr>
<td>[125;220]</td>
<td>1+ positive nuclei</td>
</tr>
<tr>
<td>[220;255]</td>
<td>Negative nuclei</td>
</tr>
</tbody>
</table>

**FIGURE 3. Inter observer agreement on single-core basis (blue plots). Visual mean H-score and virtual double staining (VDS) H-score plotted against each other and its linear regression (yellow plots and regression line).**

cores in the training set could have provided an even better calibration, in particularly for the category 3+.

Another reason could be visual underscoring as seen in the highest 8 outliers. There was a tendency among the observers that the higher general intensity of the ER-stained cores had, the higher DAB color intensity was needed to add nuclei to the highest intensity category 3+.

The difficulty of visually detecting subtle differences in staining intensity, particularly at the high ends of the scale, as well as the tendency to round scores, limits the accuracy of the visual assessed H-score.11

In regards to interobserver variation, there was an overall high agreement, but when focused on case-by-case assessment inconsistencies in the visual assessments were noticed underlining the variation inevitably present when visually assessing ER in breast specimens.

One observer generally provided higher H-scores than the rest of the group, which affected the overall, interobserver agreement and the mean visual assessment held against the VDS evaluation. When applying the calibrated VDS APP the ER scores approximates a continuous scale. However, the test set showed bimodal distribution due to a lack of positive low expressors in the H-score range 25 to 150. The investigated cohort of breast carcinomas did not represent the category of low ER-expressing tumors fully. Further investigations should proceed on ER negative and low ER expressors.

H-score is a composite scoring method and its purpose is to translate nominal observations into semi-quantitative results. By implementing H-score to IA software as seen in this VDS APP, ER assessment is carried out on a true continuous scale based on cell-by-cell data resulting in higher accuracy and reproducibility. Because of the wider range of the H-scores (0 to 300) compared with Allred score range (0 to 8), the H-score provides the clinicians with a theoretically more accurate information.

When detecting ER by automated IA systems, a substantial challenge for the IA system is to only assess ER in carcinoma cells in complex tissues containing both epithelial and stromal components.

There have been different approaches into segregating these cells. Some have manually defined the relevant ROIs by removing unwanted cell areas and actively circled tumor areas in every slide to be assessed, a laborious and time-consuming job.14,19,20 Several automated systems include masking algorithms based on morphologic characteristics such as, color, size of cell, or nucleus or shape etc. to sufficiently differentiate between stromal and epithelial cells.15,19,21,34,35

Because of the fact that malignant breast cells are able to assume a variety of sizes and shapes the accuracy of the above mentioned approach is compromised which may affect the result and is recognized as a weakness by some authors.19,21,35 The VDS APP separates tumor areas from stromal cells based on specific epitope recognition in each cell circumventing the above mentioned pitfall and thus may increase accuracy.

The VDS APP has the potential to improve the ER evaluation in breast carcinoma specimens in several ways. It will provide the pathologist with objective and reproducible quantitative data on ER expression and thereby eliminate observer variation. It will make standardization of ER evaluation possible and applicable.

Quantitatively measured ER expression in breast tumors has important clinical implications.

In regards to the predictive significance in treatment of breast cancer patients, the ER expression level has been shown to correlate with the efficacy of adjuvant endocrine therapy.5–8 Low levels of ER and progesterone receptor in these specimens have been proven to be predictive of the benefit of adding chemotherapy to the treatment plan in this group of patients.36

The standardized, objective, and reproducible VDS method makes it possible to reappraise earlier cases detected as ER negative and the used cutoff value to establish potential grounds for adjustment of current clinical practice. Potentially, this may improve the grounds on which the decision of breast cancer patient stratification to adjuvant therapy is based on.

In conclusion, we found excellent correlation between visual assessment and the VDS APP analysis of ER on TMA applied breast carcinoma. The VDS APP needs further validation in a set up with a larger sample size enriched with ER low expressors and follow-up to evaluate the performance.

REFERENCES