Development of an Evidence-Based Approach to External Quality Assurance for Breast Cancer Hormone Receptor Immunohistochemistry

Comparison of Reference Values

Nikita Makretsov, MD, PhD; C. Blake Gilks, MD; Reza Alaghehbandan, MD, MSc; John Garratt, RT; Louise Quenneville, MD; Joel Mercer, MSc; Dragana Palavdzic, MD; Emina E. Torlakovic, MD, PhD; on behalf of the Canadian Immunohistochemistry Quality Control Program, Canadian Association of Pathologists

- Context.—External quality assurance and proficiency testing programs for breast cancer predictive biomarkers are based largely on traditional ad hoc design; at present there is no universal consensus on definition of a standard reference value for samples used in external quality assurance programs.

Objective.—To explore reference values for estrogen receptor and progesterone receptor immunohistochemistry in order to develop an evidence-based analytic platform for external quality assurance.

Design.—There were 31 participating laboratories, 4 of which were previously designated as “expert” laboratories. Each participant tested a tissue microarray slide with 44 breast carcinomas for estrogen receptor and progesterone receptor and submitted it to the Canadian Immunohistochemistry Quality Control Program for analysis. Nuclear staining in 1% or more of the tumor cells was a positive score. Five methods for determining reference values were compared.

Results.—All reference values showed 100% agreement for estrogen receptor and progesterone receptor scores, when indeterminate results were excluded. Individual laboratory performance (agreement rates, test sensitivity, test specificity, positive predictive value, negative predictive value, and $k$ value) was very similar for all reference values. Identification of suboptimal performance by all methods was identical for 30 of 31 laboratories. Estrogen receptor assessment of 1 laboratory was discordant: agreement was less than 90% for 3 of 5 reference values and greater than 90% with the use of 2 other reference values.

Conclusions.—Various reference values provide equivalent laboratory rating. In addition to descriptive feedback, our approach allows calculation of technical test sensitivity and specificity, positive and negative predictive values, agreement rates, and $k$ values to guide corrective actions.

(Arch Pathol Lab Med. 2011;135:874–881)
clinical sensitivity and specificity as it fulfills the following conditions: (1) analytic sensitivity of the laboratory test (which refers to the lowest amount of analyte in a sample that can be detected) is evaluated through use of a range of positive and negative reference samples (including those with low level of expression of the biomarker of interest); (2) the proportion of positive versus negative cases reflects usual distribution of clinical cases in Canada; and (3) there is greater than 95% concordance for both positive and negative results between reference laboratories. In this study, based on use of TMAs, we use the terms test sensitivity and test specificity to refer to clinical sensitivity and specificity, as described above.4

There is an ongoing rapid evolution of immunohistochemistry as a technique, with introduction of automated stainers, antigen retrieval and improved visualization methods. With the variability in staining systems and reagents in use, standardization of methods is not entirely possible and there is emphasis, instead, on standardization of laboratory performance and results achieved, with results monitored by internal quality assurance by using appropriate and preferably standardized positive and negative controls. Monitoring by PT and peer comparison plays an essential role as to date, positive and negative controls for clinical IHC tests are insufficiently standardized.1,3–8 If clinical sensitivity and specificity are to be directly addressed by PT programs, the number of samples needs to be large enough to enable valid statistical calculations, as well as reasonably representative of true distribution of results in patient populations. To complicate matters further, the interpretation of staining is driven by the quantitative or semiquantitative cut points defined in prior studies, and for ER and PR these cut points have been changing as new scientific evidence has become available.8–10

There is no universally accepted definition of reference values for use in external quality assurance (EQA) programs for IHC tests, including testing for class II breast cancer markers.1,3–5,7–11 These reference values are critically important as they are used to score laboratory performance; in multifissure samples, such as TMAs, they can be used to calculate agreement rates, test sensitivity, specificity, positive predictive value, negative predictive value, and κ values.2,3 Two approaches have evolved: use of consensus results from multiple laboratories, or use of a single reference laboratory to determine the correct result. Both approaches have significant potential weaknesses. For example, consensus results will only work as reference values if most laboratories are obtaining the correct result. With respect to reference laboratories, there is no uniformly accepted definition of a reference immunohistochemistry laboratory in Canada.

This study compares different methods of determining PT reference values in a tissue microarray–based breast cancer biomarker EQA proficiency testing program, and examines the consequences that the selection of reference value has on participating laboratories scores, including calculated agreement rates, test sensitivity and specificity, positive and negative predictive values, and κ values for each participating laboratory. Our hypothesis is that various definitions for determination of reference values, based on either consensus or designated reference laboratory values, would be comparable in identifying laboratories with suboptimal ER and PR IHC testing results for class II IHC tests, in a setting where laboratories are required to comply with published guidelines and most participating laboratories have already demonstrated very high agreements with reference laboratories.

**MATERIALS AND METHODS**

**Case Series and TMAs for Run 5 CIQC**

Forty-four invasive breast carcinoma cases were randomly selected. All cases were reviewed to ensure that tissue fixation and processing were handled appropriately and that internal control staining (of normal breast epithelium) for ER and PR was present. The tissue microarrays were designed and constructed by using manual tissue arrayer MTA1 (Beecher Instruments Inc, Silver Spring, Massachusetts). Each case was represented by a single 1-mm tissue core.

Thirty-one laboratories participated in assessment of ER and PR in this study. The serially cut TMA slides were mailed to all participating laboratories, with each laboratory receiving 1 slide per test. The stained slides (each laboratory used its own routine protocol) were then returned to CIQC for scanning, Web posting, and analysis. All glass slides were analyzed by a panel of 5 pathologists (CIQC assessors) simultaneously, by using a multiheaded microscope. All panel members were blinded to the identity of participating laboratories.

**Interpretation of the Immunostaining Results**

Estrogen receptor and PR were scored either as negative (cut point <1% of the invasive tumor cell nuclei weakly positive, equivalent to Allred scores 0 and 2) or positive (1% and greater positive nuclei of any intensity, equivalent to Allred scores 3–8). Any discrepant opinions among the CIQC panel members while reviewing cases were resolved immediately by a simple majority vote. All results were recorded instantly in Excel spreadsheets (Microsoft, Redmond, Washington). Once the analysis was completed for ER and PR, all scores were visualized on color plots ("garattograms"12) and cases with discrepant results, relative to the reference values, were then reviewed by the same panel to ensure that there was no data entry error and these discrepancies indeed reflected different staining results. Therefore, the discrepancies were either due to true differences in sensitivity of the test or, extremely rarely, due to intrinsic tissue heterogeneity, with cells positively stained on 1 slide not present for analysis on another slide. With respect to this latter possibility, only 1 such case was detected and it was excluded from calculations.

**Identification of “Expert” Laboratories**

Four accredited laboratories were selected by the CIQC as competent in performing ER and PR IHC testing and designated “expert” reference laboratories. This was based on the following characteristics: publication records demonstrating that their staining correlated with patient outcomes and/or biochemical determination of hormone receptor status,12 large volume breast cancer biomarker immunohistochemical testing with long-standing internal quality assurance programs, and regular and successful participation in EQA programs. In addition, US Food and Drug Administration (FDA)–approved IHC Dako (Dako, Glostrup, Denmark) kit for ER and PR detection was used as a “reference method,” as none of the 4 designated reference laboratories use this method. Slides from the reference laboratories were scored by the CIQC assessors at the same time as all participants’ slides. The assessors were blinded for the source of the slides.

**Reference Value Definitions**

We assigned 5 reference values for ER and PR, for each sample, as follows.

**Reference Value 1.**—Consensus result obtained by 4 of 4 expert laboratories, or 3 of 3 expert laboratories, if 1 laboratory was not able to assess a given sample. If fewer than 3 laboratories were able to evaluate a sample, no consensus score was possible.
and the case would not be included for statistical analysis and would be flagged as “indeterminate.”

Reference Value 2.—Consensus result obtained by a majority (eg, 3 of 4, 4 of 4, 2 of 3, or 3 of 3) of expert laboratories. If there was no agreement between a majority of the expert laboratories, then no consensus score was possible and the case would not be included for statistical analysis and would be flagged as “indeterminate.”

Reference Value 3.—Fifty-one percent consensus (simple majority) of the results (positive or negative) from all participating laboratories, including the expert laboratories. Only the cases in which at least 80% of participating laboratories were able to provide score results were considered in the analysis; they were otherwise considered “indeterminate.”

Reference Value 4.—Eighty percent consensus among all participating laboratories. If there was less than 80% agreement for a given case, then no reference value could be determined and these cases were not included in the analysis but were flagged as “indeterminate.” Only the cases in which at least 80% of participating laboratories were able to provide score results were considered in the analysis; they were otherwise considered indeterminate (same as for the reference value 3).

Reference Value 5.—All results from the single expert laboratory designated as a reference laboratory, based on their use of FDA-approved kits, as described above. Only uninterpretable or missing cases were excluded from the statistical analysis.

Handling of Missing and Discrepant Data

Missing data were typically a result of either a core not sticking to the slide during processing, or the core on that particular TMA slide not containing tumor, or not containing sufficient tumor to allow assessment. Arbitrarily, CIQC requires at least 50 tumor cells for assessment.

“Indeterminate” reference values were occasionally observed. They are defined above under the reference methods description. They were not included into the statistical analysis for each specific reference value.

Data Analysis.—The spreadsheets with all data for ER and PR were subjected to statistical analysis by SPSS 17.0 (Chicago, Illinois) package. The following values were then calculated for each reference method used: agreement rate, test sensitivity, test specificity, positive predictive value, negative predictive value, and interval \( \kappa \) for each participating laboratory. \( \chi \) Statistic 95% confidence intervals were calculated using the following formula: estimate \( \pm 1.96 \times \) standard error. We measured agreement between all the laboratories and the reference values for each biomarker by using interval \( \kappa \) value methodology (within 1 standard deviation), which allowed us to provide an unbiased measurement of agreement between the laboratories with slightly different sets of interpretable tissue cores on each slide (owing to lost or uninterpretable tissue cores). A \( \kappa \) value equal to 1.0 and concordance of 100% were considered as perfect. A \( \kappa \) value greater than 0.8 but less than 1.0 was considered as near perfect, while \( \kappa \) greater than 0.5 but less than or equal to 0.8, as substantial agreement. A \( \kappa \) value below 0.5 was considered fair to poor agreement. For the purposes of this study, agreement rates of less than 100% but at least 90% were considered near optimal, while agreement rates inferior to 90% were considered suboptimal.

RESULTS

Comparison of 5 Reference Values

The summary of reference values 1 through 5 for ER and PR, for each of the 44 cases, is shown as color plots (garattograms) (Figure 1, A and B, respectively, right-hand sections). These demonstrate complete (100%) agreement for ER and complete (100%) agreement for PR between reference values determined by 5 different methods, when cases with indeterminate results are excluded in calculation of agreement rates (\( P < .001 \), Spearman \( \rho = 1 \)). Within this case series, 73% to 75% of cases were ER positive (including 7%–10% of low-positive ER cases, ie, 1%–10% weakly positive cases), while 52% to 56% of cases were PR positive (including 12%–19% of low-positive PR cases, respectively), depending on the reference value used.

ER Results

The ER test results for all cases and all participating laboratories are presented graphically in Figure 1, A (left-hand section). Examples of the ER immunohistochemical staining images obtained from the several laboratories are shown in Figure 2. The clustered bar charts in Figure 3 depict ER test agreement, test sensitivity, test specificity, positive and negative predictive values, and \( \kappa \) values for each laboratory, as calculated with 5 different reference values. Most of the participating laboratories demonstrated either complete or near-complete agreement with the reference values, irrespective of the reference value used for calculation (Figure 2, A). Twenty-one of 31 laboratories showed perfect (ie, 100%) agreement with all reference values. Seven laboratories showed near-optimal results (>90% but <100% agreement), while 2 laboratories showed less than 90% agreement or suboptimal results. A single laboratory (laboratory 17) had less than 90% concordance with 3 of the reference values and greater than 90% concordance with 2 of the reference values. All ER test failures were attributable to lower test sensitivity (ie, false-negative results) (Figure 3, B), with no problems detected in the test specificity of ER staining (ie, false-positive results) (Figure 3, C). Positive and negative predictive values of the ER tests mirrored sensitivity and specificity results (Figure 3, D and E). The results for each laboratory were highly consistent (being either optimal or not), with no or minimal variation of the results depending on the reference method used for comparison, with the exception of laboratory 17, as noted previously. Interval \( \kappa \) values, depicting a range of 1 standard deviation, were highly concordant for all 5 reference values used. For only 1 laboratory did the interval \( \kappa \) values clearly fall outside the perfect or near-perfect range (<0.80) (Figure 3, F).

The false-negative results included ER cases of any level of positivity, as determined by reference values used.

PR Results

The PR test results for all cases and all participating laboratories are presented graphically in Figure 1, B (left-hand section). Examples of the PR immunohistochemical staining images from several laboratories are shown in Figure 2. A minority of the laboratories demonstrated 100% agreement (11 of 31), in contrast to ER results (Figure 4, A). Nineteen laboratories showed near-optimal results (>90% agreement), while only 1 laboratory showed less than 90% agreement (suboptimal results). Disagreement with reference values reflected occasional problems with PR test sensitivity, test specificity, or both (this is mirrored in the respective positive and negative predictive values) (Figure 4, B through E). Nevertheless, interval \( \kappa \) values for PR testing showed that, for all laboratories, results extended into the range defined for the purposes of this study as being perfect or near-perfect agreement (\( \kappa > 0.80 \)) (Figure 4, F). The false-negative PR results were discordant with any PR positive standard scores of any intensity, as determined by the reference values used.
Figure 1. Results for estrogen receptor (A) and progesterone receptor (B) for the 44 cases, with 5 different methods for determination of reference values. Each row is an individual case, while each column is a different participating laboratory. Reference values for each case, determined through 5 different methods, are shown at the right, for comparison (reference [ref] values 1 to 5). Red, positive (p); white, negative (n); yellow, indeterminate (i).
All 5 reference values correlated significantly between each other; there were no statistically significant differences between them. Nevertheless, we observed a trend toward higher agreement rates for reference value 4 (ie, 80% agreement level between all participating laboratories). This was due to the exclusion of the cases for which 80% agreement cannot be reached. For this reason, the reference value 4 appears less strict and resulted in slightly higher concordance rates, although this was statistically insignificant. This is represented in Figures 3, F (for ER) and 4, F (for PR), which show that respective interval $k$ values are overlapping.

COMMENT

External quality assurance in IHC is an essential part of routine practice in anatomic pathology. Participation in such programs aims to help laboratories detect problems not identified by internal quality control. Assessment of laboratories in PT programs requires development of scientific methodology, which should allow for minimal reliance on subjective assessment and should detect deficiencies in individual laboratories by comparing their results to reference results in an unbiased fashion.

In this study we compared different reference values for ER and PR EQA proficiency testing. There is currently no generally accepted method for determination of reference values in steroid hormone receptor IHC,\textsuperscript{1,2,9} unlike HER2/neu IHC, for which control tissues may be verified by testing their respective gene amplification status by either fluorescence or chromogenic in situ hybridization.\textsuperscript{1,10,13–15} Biochemical assays can be used for determination of hormone receptor expression in breast cancer, but they have been shown to be inferior to IHC in predicting response to targeted therapy, and are therefore not suitable for use as an independent method for determination of reference values.\textsuperscript{7,10} Thus, determination of reference values for ER and PR currently is done using the same test technique, that is, IHC. We compared a number of methods for determination of reference values, principally based on consensus results obtained by different laboratories (not necessarily using the same equipment or protocols) and they were equivalent. This is a reassuring result, which suggests that the selection of a reference value—be it consensus result, some modification of the consensus result, or a single reference laboratory—probably does not have a major impact on achieved laboratory score in PT.

The purpose of using TMA design for EQA samples for PT for class II IHC tests is to measure not only analytic performance of the laboratory (eg, signal to noise ratio, intensity of staining, morphology) but also to measure test performance quantitatively by testing multiple samples with a range of expression levels reflecting those encountered in practice. As well, the nature of the samples used, that is, formalin-fixed, paraffin-embedded tumor samples, reflect the samples tested in practice. This use of multiple samples representative of cases encountered in practice allows descriptive statistical assessment of the performance of individual laboratories. It should be emphasized that performance as estimated by this method may underestimate performance in practice, as the size of the test tissue cores is small, and cases with equivocal results cannot be repeated, as would happen in practice; nonetheless, it is a starting point from which to move toward more evidence-based practice.

Measurement of agreement between the reference values and each individual laboratory results can be done by using multiple methods. In this study we assessed agreement rates, test sensitivity, test specificity, positive predictive value, and negative predictive value. Not surprisingly, the results for individual laboratories were essentially equivalent with all assessment tools. However,
Figure 3. Estrogen receptor (ER) results (clustered bar charts). Bars are calculated results for an individual laboratory, with each laboratory represented by a cluster of 5 bars derived by use of the 5 different reference values: blue, reference value 1 (consensus between 4 expert laboratories); green, reference value 2 (consensus between 3 most sensitive expert laboratories); grey, reference value 3 (consensus by a simple majority of 51% of the laboratories); brown, reference value 4 (consensus between 80% of the laboratories); and yellow, reference value 5 (single expert laboratory). Vertical axis: percentage value for each of (A) agreement rate (with the red bar indicating 90% agreement level); (B) test sensitivity; (C) test specificity; (D) positive predictive value; (E) negative predictive value; and (F) interval \( \kappa \) values (with the red bar at \( \kappa = 0.8 \), indicating near-perfect agreement). Red dots at \( \kappa = 1.0 \) indicate the laboratories with complete agreement with all 5 reference values used.
Figure 4. Progesterone receptor (PR) results (clustered bar charts). Refer to Figure 3, A through F, for a detailed description (graphs are organized by using the same principle as for ER results).
they do provide slightly different information. For example, discrepant ER results, compared to the reference value, were all false negatives (ie, decreased sensitivity, with perfect specificity), while PR testing was associated with both false-negative and false-positive results. This information is in addition to that derived if only agreement rates are assessed. We also used an interval $\kappa$ methodology, which considers variation in a number of pair-wise comparisons between the laboratories and the standard values. We preferred to use interval $\kappa$ results (ie, low-high thresholds for each individual $\kappa$ within 1 standard deviation) as the most objective measure of agreement in cases of random data loss. As expected, these results showed that agreement fell within a broad range, based on our sample size. This raises the question of how large a sample size is needed to detect suboptimal testing results. The CIQC TMAs include about 40 cases because, from a statistical point of view, a sample size of at least 30 valid data points would be required for a minimally adequate statistical analysis; moreover, as suggested by the American Society of Clinical Oncology (ASCO) and College of American Pathologists (CAP) expert panel, when using the 90% benchmark agreement rate, 40 samples would provide conditions under which a laboratory operating at 85% rate of true agreement with reference values would have only a 15% chance of meeting or exceeding the 90% agreement benchmark on the set of test cases.\(^{14}\)

Any error in diagnostic IHC can be perceived as unacceptable by the public, yet error remains unavoidable, in practical terms. The question remains, “What is an acceptable frequency of error according to the current standard of care?” The ASCO/CAP guidelines for ER and PR testing\(^{16}\) recommend that all laboratories performing ER or PR testing participate in external proficiency testing and show 90% correct responses (ie, agreement with reference values). The figure of 90% appears to be completely arbitrary; we are not aware of any evidence supporting this choice of cutoff. In our opinion, the agreement rate cutoff should be determined by current best practice, with a goal of improvement (ie, decreasing the acceptable error rate) moving forward, as technical innovations are made. As such, development of testing programs and analytic tools, and collection of data, are foundational aspects for determining cutoffs for deeming the practice of an individual laboratory to be suboptimal. The first step in this process should be the assessment of methods for measurement of laboratory performance, based upon validated reference methods. The current study demonstrates the equivalence of a number of different methods for determination of reference values. Although not a primary objective of this study, we also demonstrate that achievement of greater than 90% agreement with reference values is possible for diagnostic IHC laboratories performing ER and PR testing.

References