Standardization of HER2 Immunohistochemistry in Breast Cancer by Automated Quantitative Analysis

Mark D. Gustavson, PhD; Brian Bourke-Martin, MS; Dylan Reilly, MS; Melissa Cregger, BS; Christine Williams, MS; Jane Mayotte, MS; Maciej Zerkowski, MS; Greg Tedeschi, BS; Robert Pinard, PhD; Jason Christiansen, PhD

Context.—There is critical need for standardization of HER2 immunohistochemistry testing in the clinical laboratory setting. Recently, the American Society of Clinical Oncology and the College of American Pathologists have submitted guidelines recommending that laboratories achieve 95% concordance between assays and observers for HER2 testing.

Objective.—As a potential aid to pathologists for achieving these new guidelines, we have conducted an examination using automated quantitative analysis (AQUA analysis) to provide a standardized HER2 immunohistochemistry expression score across instruments (sites), operators, and staining runs.

Design.—We analyzed HER2 expression by immunohistochemistry in a cohort (n = 669) of invasive breast cancers in tissue microarray format across different instruments (n = 3), operators (n = 3), and staining runs (n = 3). Using light source, instrument calibration techniques, and a new generation of image analysis software, we produced normalized AQUA scores for each parameter and examined their reproducibility.

Results.—The average percent coefficients of variation across instruments, operators, and staining runs were 1.8%, 2.0%, and 5.1%, respectively. For positive/negative classification between parameters, concordance rates ranged from 94.5% to 99.3% for all cases. Differentially classified cases only occurred around the determined cut point, not over the entire distribution.

Conclusions.—These data demonstrate that AQUA analysis can provide a standardized HER2 immunohistochemistry test that can meet current guidelines by the American Society of Clinical Oncology/College of American Pathologists. The use of AQUA analysis could allow for standardized and objective HER2 testing in clinical laboratories.

(Arch Pathol Lab Med. 2009;133:1413–1419)
method that provides objective and continuous protein expression scores in tissue using automated fluorescence microscopy and advanced image analysis algorithms.10 It uses molecular identification of compartments to quantify biomarker expression as a function of pixel intensity in specific tissues or subcellular compartments. It has been demonstrated that AQUA scores are directly proportional to molecules per unit area or protein concentration,11 and that this methodology is broadly applicable to biomarker assessment and clinical characterization.10–19 A recent publication has shown that assessment of HER2 by AQUA analysis has a significant linear relationship with FISH.20 Furthermore, it has been demonstrated that HER2 expression as determined by AQUA analysis predicts response to trastuzumab.20

Because AQUA technology is objective and strictly quantitative, we demonstrate here how the system has been enhanced such that it can be standardized across instruments (ie, laboratories) and operators (ie, observers). We have developed calibration methodologies that enable captured signal to be normalized across multiple instruments (see “Materials and Methods”). We also have developed software algorithmic methodologies that remove operator decisions from the imaging and scoring process, making it predominantly unsupervised. Here, these methodologies are applied to HER2 testing to demonstrate the performance of a standardized testing system developed using unsupervised AQUA technology.

MATERIALS AND METHODS
Cohort Description and Tissue Microarray Construction
A large breast cancer cohort in tissue microarray (TMA) format was employed in these studies in order to test standardization techniques. This cohort from the Yale Tissue Microarray Facility (YTMA49; Yale University School of Medicine, New Haven, Connecticut) has been described in detail previously.14 Briefly, the breast cohort (n = 669) of invasive ductal carcinoma was serially collected from the Yale University Department of Pathology from 1961 to 1983. Also on the array was a selection of normal tissue and cell line controls. The mean follow-up time was 12.8 years, with a mean age of diagnosis of 58.1 years. This cohort contained approximately half node-positive and half node-negative specimens. Detailed treatment information was not available for this cohort.

Immunohistochemical/Immunofluorescence Tissue Staining
Chromagen-based immunohistochemical staining and scoring for cases in YTMA49 was performed as described previously.13 YTMA49 was stained using a modified indirect immunofluorescence protocol.10 In brief, precut, paraffin-coated TMA slides were deparaffinized and antigen retrieved by heat-induced epitope retrieval in 10 mM Tris (pH 9.0). Using an autostainer (LabVision, Fremont, California), slides were preincubated with Background Sniper (BioCare Medical, Concord, California). Slides were then incubated with primary antibodies against HER2 (1:8000; rabbit polyclonal Hercept Test antibody, catalog no. A0485, DAKO, Carpinteria, California) and cytokeratin (1:200; rabbit polyclonal pan-cytokeratin, catalog no. M3515, DAKO) diluted in DaVinci Green (BioCare Medical) for 1 hour at room temperature. Slides were washed 3 x 5 minutes with 1× Tris-buffered saline containing 0.05% Tween-20. Corresponding secondary antibodies were diluted in Da Vinci Green and incubated for 30 minutes at room temperature. These included either antibodies directly conjugated to a fluorophore for anti-keratin (1:100; Alexa 555–conjugated goat anti-rabbit, Molecular Probes, Eugene, Oregon) and/or conjugated to a horseradish peroxidase via anti-mouse or anti-rabbit Envision (DAKO). Slides were then incubated with a fluorescent chromagen amplification system (Cy-5-tyramide, NEN LifeScience Products, Boston, Massachusetts) which, like diaminobenzidine, is activated by horseradish peroxidase and results in the deposition of numerous covalently associated Cy-5 dyes immediately adjacent to

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Representative HER2-positive (A through C) and HER2-negative (D through F) tumor sample images used for automated quantitative analysis (AQUA). Shown are micrographs of cytokeratin/Cy3 (A and D), HER2/Cy5 (B and E), and their subsequent merge (C and F). Orange/yellow pixels in merged images represent pixels quantified for HER2 AQUA score generation (original magnification ×60).
Standardization of HER2 Immunohistochemistry—Gustavson et al

Figure 2. Comparison of automated quantitative analysis (AQUA) scoring to traditional chromagen-based immunohistochemistry (IHC) scoring. A, Box plot of log2-transformed HER2 AQUA scores (y-axis) categorized by traditional IHC scoring (x-axis) for 425 cases with both AQUA and IHC scoring information. One-way analysis of variance for comparison of means across all categories was significant (P < .001). Posthoc analysis using a Tamhane T2 statistic for multiple comparisons shows significant differences between each category (all P values < .05). B, The 2 × 2 contingency tables comparing positive (POS) versus negative (NEG) population segregation–based HER2 IHC scoring cut points (NEG: less than +3; POS: +3) and AQUA POS/NEG cut point as generated by X-tile (Figure 4). Also shown are overall concordance, positive agreement, and negative agreement rates with 95% confidence intervals (95CI).

Microscopy System and Image Acquisition

The PM2000 system, commercialized by HistoRx, is based on a system described previously. In brief, it comprises the Olympus BX51 epi-fluorescence microscope (Olympus America Inc, Center Valley, Pennsylvania), which is equipped with a motorized nosepiece to control selection of objectives (ie, 4×, 10×, 20×, 40×, and 60×); a motorized filter turret to control selection of different filter cubes (ie, 4',6'-diamidino-2-phenylindole, Cy2, Cy3, Cy5, and Cy7, or equivalent wavelengths); a motorized stage to control stage movements (Prior Scientific Inc, Rockland, Massachusetts); an X-Cite 120 mercury/metal halide light source (EXFO Life Sciences & Industrial Division, Mississauga, Ontario, Canada); and a QUANTFIRE monochromatic digital camera (Optronics Inc, Goleta, California).

Automated image capture was performed by the HistoRx PM-2000 using the AQUAsition software package. High-resolution, 8-bit (resulting in 256 discrete intensity values per pixel of an acquired image) digital images of the cytokeratin staining visualized with Cy3, 4',6'-diamidino-2-phenylindole, and target (HER2) staining with Cy5 were captured and saved for every histospot on the array. Pixels were written to image files as a function of power (Power = [Pixel Intensity/256]/exposure time) to help compensate for experimental variations in staining intensity and exposure times.

AQUA Score Generation

Images were validated for percent area tumor (tumors showing < 5% area per field were redacted), out of focus, and debris. Of the 669 tumor samples on YTM49, 86 samples were redacted (12.8%), leaving 583 samples for subsequent scoring and analysis. Compartment-specific AQUA scores for HER2 for each histospot were generated based on the pixel-based locale assignment for compartmentalization of expression algorithm as described previously. To remove operator-to-operator bias for threshold setting, we have developed an unsupervised pixel-based clustering algorithm for optimal image segmentation for use in the pixel-based locale assignment for compartmentalization of expression algorithm.

Instrument and AQUA Score Normalization

For machine and AQUA score normalization, 3 calibration factors were developed: calibration cube factor (CC factor), light source factor (LS factor), and Cy5 optical path factor (OP factor). Calculation of these factors is based on pixel intensity measurements given by images acquired under described conditions. All
Light passed through a specific microscope objective/filter combination relative to the measured incoming light intensity. For these measurements, a standard sample is required that can be transferred between different machines and maintain reproducibility in its construction. A commercially available blue fluorescent standard slide was selected for this purpose (Omega Optical Inc, Brattleboro, Vermont). To produce this measurement, a ratio is taken between the incident light intensity and the measured intensity of the standard sample. The Cy5 OP factor is the quotient of the average total light intensity of 16 images taken for each cube/sample combination. The CC and OP factors are intrinsic to the specific hardware system being studied and need only be calculated once or at an interval where one would suspect some type of modification in the optics has occurred.

The normalized AQUA score is shown in equation (1):

\[
\text{Normalized AQUA score} = \text{Raw AQUA score} \times \text{CC factor} \times \text{LS factor} \times \text{OP factor}
\]

where the CC and OP factors are defined upon system setup/construction, and the LS factor is measured simultaneously. Thus, the system standardization results described here do not add significant extra time to the data acquisition from the current methods.

**Statistical Analysis**

For analysis, AQUA scores were log, transformed. Statistical analysis and output were performed using SPSS 15.0 (SPSS Inc, Chicago, Illinois) unless otherwise noted. Optimal cut points for continuous HER2 AQUA data for 5-year disease-specific death were generated as a function of survival using the software package X-tile as described previously.22 X-tile performs Monte Carlo simulations (ie, cross-validation23) to produce corrected P values to assess statistic significance of data generated by multiple cut points. The software also generates training/validation subsets for additional P value estimation. Agreement percentages with 95% confidence intervals for 2×2 contingency tables were determined using the Web-based tool JvStat for 2-way contingency table analysis (http://statpages.org/ctab2×2.html; last date accessed: October 20, 2008).

**RESULTS**

**HER2 Immunofluorescence Staining**

AQUA analysis employs multiplexed fluorescent stains to compartmentalize and measure expression of specific biomarkers. For HER2, only pixels within the cytokeratin-derived (epithelium-specific) compartment are considered for analysis, thus differentiating tumor from stromal HER2 signal as well as membrane/cytoplasm from nuclear HER2 signal. Sample images are provided in Figure 1, where an example of a HER2-positive tumor (Figure 1, A through C) and a HER2-negative tumor (Figure 1, D through F) are given. As described, only HER2 pixels that coincide with cytokeratin pixels are used to generate an AQUA score (orange/yellow pixels in merged images; Figure 1, C and F).

**AQUA Score Correlation With Traditional IHC Scoring Methods**

HER2 AQUA scores show a moderate but highly significant correlation with categorical IHC scoring methods (0, +1, +2, and +3), with a Spearman Rho value of 0.46 (P < .001). Multinomial regression analysis (for comparison of categorical versus continuous data) showed a highly significant correlation (P < .001), with a pseudo-R\(^2\) value of 0.56. Figure 2, A, categorizes HER2 AQUA scores as a function of IHC scores using a box plot. Although there is a highly significant difference in population mean (analysis of variance [ANOVA] P < .001), because AQUA provides a continuous expression score, significant overlap of AQUA scores is observed across the range of traditional.
Figure 5. The 2 × 2 contingency tables comparing positive (POS) versus negative (NEG) population segregation based on X-tile cut points generated for the reference (ie, instrument 1) for each indicated instrument set (A and B), operator set (C and D), and run set (E and F). Also shown are overall concordance, positive agreement, and negative agreement rates with 95% confidence intervals (95CI).

Categoric scoring. To demonstrate accuracy of AQUA with respect to HER2 quantification, we examined positive/negative concordance between traditional (clinical) IHC scoring and AQUA scoring. Figure 2, B, shows 2 × 2 concordance tables demonstrating overall percent agreement between traditional IHC scoring and AQUA analysis at 94.8%.

**Standardized HER2 Expression Scores**
Three serial sections of a cohort (n = 669) of invasive breast cancers were fluorescently stained for HER2 as described in “Materials and Methods.” The first serial section was used to assess AQUA score variability across 3 different instruments (intersite variability) and 3 different operators (single instrument). The second and third serial sections were stained on separate days to assess run-to-run variability. Figure 3 gives box plots showing AQUA score distributions for each indicated acquisition parameter (instrument [Figure 3, A], operator [Figure 3, B], and independent staining runs [Figure 3, C]). For 583 patient samples, the average percent coefficient of variation (%CV) was 1.8% (minimum, 0.04%; maximum, 10.7%).
Figure 6. Frequency distributions separated into negative agreement, positive agreement, and nonagreement cases for (A) instrument 2 (automated quantitative analysis [AQUA] scores) to instrument 1 (cut point); (B) operator 2 (AQUA scores) to operator 1 (cut point); and (C) run 2 (AQUA scores) to run 1 (cut point) to demonstrate where disagreement occurs within the population of breast cancer cases. Cases that disagree reside in and around the indicated cut points and do not span over the entire distribution.

across instruments; 2.0% (minimum, 0.06%; maximum, 15.6%) across operators; and 5.1% (minimum, 0.12%; maximum, 29.7%) across independent staining runs. These %CVs rival those of in vitro immunoassays, such as the enzyme-linked immunosorbent assay.24

**Positive/Negative Concordance**

The critical parameter for HER2 testing in the clinical laboratory is the ability to reproducibly classify patients as positive or negative. Using survival as a surrogate marker for positive/negative classification,23 we established an optimal AQUA score cut point using X-tile22 for HER2 AQUA scores produced for instrument 1, operator 1, and staining run 1. Figure 4, A, shows Kaplan-Meier survival analysis of positive/negative HER2 classification for instrument 1. As described in “Materials and Methods,” the cut point was validated with significance by Monte Carlo simulation (\(P < .001\)) and training/validation subsets (\(P = .002\)). This validated cut point was applied to AQUA scores generated on instruments 2 and 3 with significances \(P < .001\) and \(P = .004\), respectively (Figure 4, B and C). We observed similar reproducibility across independent operator and staining run acquisitions, with \(P\) values all .01 or less (data not shown).

Current ASCO-CAP guidelines suggest laboratories achieve 95% positive/negative concordance for current HER2 assay methodologies.6 A recent study shows that for HER2 IHC-based scoring, concordance between observers ranges from 54% to 85%, falling short of these guidelines.9 We examined positive/negative concordance for AQUA scoring across instruments, operators, and staining days using the cut points established above. As shown in Figure 5, overall concordance ranged from 94.5% (instrument 1 to instrument 3; Figure 5, B) to 99.3% (operator 1 to operator 2; Figure 5, C). It is important to note that these analyses include all cases, including those that would be considered equivocal. We did observe lower positive agreement rates compared with negative agreement. Positive agreement ranged from 72.2% to 98.6%, whereas negative agreement ranged from 96.5% to 99.8%. This disparity is expected, given the relatively low prevalence of HER2-positive cases within the given population (AQUA, 12.3%–15.6%; IHC, 11.1%).

Of those cases that are differentially classified, the question becomes where along the distribution of AQUA scores do these cases occur. To address that question, we generated paneled frequency histograms to examine where differentially classified cases were occurring. As shown in Figure 6, for instrument to instrument, operator to operator, and run to run, differentially classified cases occur at the cut point and not over the entire distribution. These data suggest that the classification error concerns cut point selection, not generation and reproducibility of the HER2 AQUA score. Taken together, these data suggest that although there is an equivocal region in the vicinity of the cut point, classification of patients for interinstrument, interoperator, and interrun assessment of HER2 expression using AQUA scoring is highly reproducible, with concordance rates approaching, if not exceeding, that suggested by ASCO-CAP.6

**COMMENT**

Here, we demonstrate the use of AQUA technology to reproducibly quantify and classify HER2 expression in tissue. We took advantage of the objective and quantitative nature of AQUA technology to develop a standardization methodology that resulted in %CVs across instruments, operators, and staining runs that rival quantitative immunoassays in the clinical laboratory (ie, enzyme-linked immunosorbent assay).24 We also demonstrated that the positive/negative classification of patients based on HER2 expression could also be standardized across instruments and operators as well as independent staining runs, resulting in concordance rates (94.5%–99.3%) that approach and/or exceed guidelines now established by ASCO-CAP.6

The goal of this study was to demonstrate how AQUA can be employed to generate a standardized HER2 immunohistochemical result and how it can be used to improve the current IHC testing in the clinical setting. An important consideration is accuracy of AQUA results with
respect to outcome and positive/negative classification as it relates to the currently accepted method of assessing in situ HER2 protein expression (traditional chromagen-based IHC scoring). Using traditional IHC scoring as an imperfect standard, we demonstrate a high degree of concordance (94.8%; Figure 2) for positive/negative patient classification as ascertained with traditional IHC scoring. These data demonstrate that AQUA can ‘correctly’ classify patients, but taken together with other data presented here, this classification is highly reproducible and standardized using AQUA. In a recent study by Giltnane et al.,20 it was suggested that a possible source of discordance between IHC and FISH measurements is related to variability in IHC measurements.20 The authors of this study further demonstrate a significant linear relationship between local and central laboratories for not only HER2 strictly positive or negative results, but rather provide a reducible AQUA score means that assays in personalized evaluation methods that can be standardized across multiple laboratories and readers. Because of the subjective and qualitative nature of routine IHC testing, standardization is problematic. This has been demonstrated in studies looking at concordance rates between local and central laboratories for not only HER2 but also other biomarkers, such as estrogen receptor.26 From its inception, AQUA technology has been setting new standards for oncologic biomarker assessment by providing objective and quantitative expression scores where an observer did not have to subjectively decide total expression; rather, biomarker expression was objectively determined in both quantity (pixel intensity) and localization (pixel colocalization with other molecularly defined markers) by computer-driven image analysis software. The addition of methodologies that allow for normalization of pixel intensity (machine calibration) and also removal of operator-to-operator variability now provide an AQUA technology that is standardized and highly reproducible. It is this type of testing that is now being called for, especially in the field of HER2 testing, by professional societies and governing bodies. Thus, AQUA technology can aid the pathologist, especially in the local laboratory, by providing a HER2 test that is objective and reproducible. Although we employed overall survival, rather than response to trastuzumab, as our reference standard for positive/negative classification, it has been demonstrated that response to trastuzumab can be predicted using AQUA technology.20

It is important to note that AQUA technology and its standardization capabilities are not limited to HER2. To date, we have demonstrated the same type of reproducibility for, but not limited to, EGFR, ER, mTOR, and PTEN (data not shown) in multiple tissue types. The quantitative and continuous nature of the AQUA technology, coupled with the reproducibility described here, also means that the platform can be extended to multiplex measurements of multiple targets. Furthermore, the nature of the reproducible AQUA score means that assays in personalized medicine may not need to directly classify patients into strictly positive or negative results, but rather provide a continuous value that can be used to provide a more in-depth assessment of a biomarker’s status.

AQUA, PM2000, and AQUAsition are registered trademarks or trademarks of HistoRx. The PM2000 and/or AQUA analysis software may be covered by one or more patent applications issued or pending in the United States and worldwide.

References


