CEBP Focus: Biomarkers and Biospecimens

Assessment of Automated Image Analysis of Breast Cancer Tissue Microarrays for Epidemiologic Studies

Kelly L. Bolton[1,6], Montserrat Garcia-Closas[1], Ruth M. Pfeiffer[1], Máire A. Duggan[4], William J. Howat[5], Stephen M. Hewitt[2], Xiaohong R. Yang[1], Robert Cornelison[3], Sarah L. Anzick[3], Paul Meltzer[3], Sean Davis[3], Petra Lenz[2], Jonine D. Figueroa[1], Paul D.P. Pharoah[6], and Mark E. Sherman[1]

Abstract

Background: A major challenge in studies of etiologic heterogeneity in breast cancer has been the limited throughput, accuracy, and reproducibility of measuring tissue markers. Computerized image analysis systems may help address these concerns, but published reports of their use are limited. We assessed agreement between automated and pathologist scores of a diverse set of immunohistochemical assays done on breast cancer tissue microarrays (TMA).

Methods: TMAs of 440 breast cancers previously stained for estrogen receptor (ER)-α, progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), ER-β, and aromatase were independently scored by two pathologists and three automated systems (TMALab II, TMAx, and Ariol). Agreement between automated and pathologist scores of negative/positive was measured using the area under the receiver operating characteristics curve (AUC) and weighted κ statistics for categorical scores. We also investigated the correlation between immunohistochemical scores and mRNA expression levels.

Results: Agreement between pathologist and automated negative/positive and categorical scores was excellent for ER-α and PR (AUC range = 0.98-0.99; κ range = 0.86-0.91). Lower levels of agreement were seen for ER-β categorical scores (AUC = 0.99-1.0; κ = 0.80-0.86) and both negative/positive and categorical scores for aromatase (AUC = 0.85-0.96; κ = 0.41-0.67) and HER2 (AUC = 0.94-0.97; κ = 0.53-0.72). For ER-α and PR, there was a strong correlation between mRNA levels and automated (ρ = 0.67-0.74) and pathologist immunohistochemical scores (ρ = 0.67-0.77). HER2 mRNA levels were more strongly correlated with pathologist (ρ = 0.63) than automated immunohistochemical scores (ρ = 0.41-0.49).

Conclusions: Automated analysis of immunohistochemical markers is a promising approach for scoring large numbers of breast cancer tissues in epidemiologic investigations. This would facilitate studies of etiologic heterogeneity, which ultimately may allow improved risk prediction and better prevention approaches. *Cancer Epidemiol Biomarkers Prev; 19(4); 992–9. ©2010 AACR.*

Introduction

There is increasing evidence that risk factor associations for breast cancer vary by tumor subgroups defined by morphology and immunohistochemical expression of tumor markers (1-4), but our knowledge of these relationships is incomplete. Refining our understanding of etiologic heterogeneity may permit improved risk assessment and allow better prevention and screening approaches. Performing risk factor analyses by tumor subgroups requires the study of a large number of cases, creating a need for reproducible, high-throughput methods for scoring immunohistochemical stains. The development of tissue microarray (TMA) technology has partly addressed this need by providing a platform for doing standardized, rapid immunohistochemical staining of many tumors. However, optimizing methods for scoring immunohistochemical stains is challenging. Interpretation of immunohistochemical stains by pathologists remains the current standard but is limited by suboptimal interobserver agreement (5), reliance on a semiquantitative scoring metric, and a taxing workload.

Automated image analysis systems offer a potential solution by providing objective, rapid, reproducible,

Authors' Affiliations: [1]Division of Cancer Epidemiology and Genetics, [2]Laboratory of Pathology, and [3]Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland; [4]University of Calgary, Calgary, Alberta, Canada; and [5]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre; [6]Cancer Research UK, Department of Oncology, University of Cambridge, Strangeways Research Laboratory, Cambridge, United Kingdom

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

This work would not be possible without the dedicated efforts of the physicians, nurses, interviewers, and study participants.

Corresponding Author: Kelly L. Bolton, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, United Kingdom. Phone: 44-301-637-2147; Fax: 44-301-402-0916. E-mail: boltonk@mail.nih.gov

doi: 10.1158/1055-9965.EPI-09-1023

quantitative measurements of immunohistochemical stains. Several commercially available systems are in use, but published reports of their performance are limited (6, 7). Given that the performance of these systems may vary by tissue type and marker, comprehensive validation is required before applying these methods to large-scale epidemiologic investigations. In this report, we assess the performance of three automated systems for immunohistochemical scoring: TMAx (Beecher Instruments), Ariol (Applied Imaging), and TMALab II (Aperio). These were applied to a diverse set of immunohistochemical stains with relevance to breast cancer: estrogen receptors α and β (ER-α and ER-β), progesterone receptor (PR), aromatase, and human epidermal growth factor receptor 2 (HER2). These antigens were chosen to capture a diversity of staining patterns (nuclear, membranous, and cytoplasmic) and challenges in interpretation.

In the absence of a true "gold standard" for quantifying immunohistochemical staining, we used two complementary approaches to assess the performance of these automated scoring systems. First, we measured the agreement between automated and pathologists' scores, comparing it with the level of interpathologist and intrapathologist agreement. Second, we investigated differences in the strength of the association between immunohistochemical scores derived with each method and mRNA expression levels determined in frozen tumor tissues available from a subset of patients.

Materials and Methods

Study population

The Polish Breast Cancer Study included women between 20 and 74 y of age who resided in Warsaw or Lodz, Poland from 2000 to 2003 (4). Breast cancer cases were identified through a rapid identification system organized at five participating hospitals and through cancer registries. At total of 2,386 cases agreed to participate and provided informed consent under a protocol approved at the National Cancer Institute and local Institutional Review Boards in Poland. This report includes a subset of 440 invasive carcinomas with available formalin-fixed, paraffin-embedded tumor blocks that had been previously prepared as a TMA.

Immunohistochemistry

Routinely prepared formalin-fixed, paraffin-embedded blocks of 440 cases with invasive breast cancer were used to construct TMA blocks with 2-fold representation as 0.6-mm-diameter cores (Beecher Instruments). Methods for doing immunohistochemical stains for ER-α, ER-β, PR, HER2, and aromatase have been detailed elsewhere (8).

Immunohistochemical scoring

TMA slides were digitized via whole-slide scanning with 20× objective using two systems: the Aperio T2 scanner (Aperio Technologies) and the Ariol SL-50 scanner (Genetix). The digital images of immunohistochemically stained slides generated with the Aperio system were independently scored by two pathologists (M.E.S. and M.A.D.). The percentage (0%, 1%, 5%, 10%, 20%… 100%) of tumor cells with positive staining and the average staining intensity (0 = negative, 1 = weak, 2 = intermediate, and 3 = strong) were recorded for each marker. Stains for aromatase were either negative or diffusely positive; therefore, we only assessed intensity for this marker. Combined scores based on the percentage of cells stained times intensity (ranging from 0 to 300) were generated for each pathologist and then averaged. To measure the repeatability of visual scores, one pathologist (M.A.D.) rescored a random sample of 10% of the spots masked to her previous scores. Stains for ER-α, PR, and ER-β were considered positive if the average combined score was ≥10.5. Positive stains for aromatase and HER2 were defined as having an average intensity score of ≥1.5 (i.e., at least one of the two tissue cores with 2+ score). Semiquantitative categories (negative, low, moderate, and strong staining) for the average pathologist scores were also created. The combined score cut points for these categories for ER-α, PR, and ER-β are as follows: none (<0.5), low (0.5-10.5), moderate (10.5-100.4), and strong (≥100.5). The intensity score cut points for HER2 and aromatase are as follows: none (<0.5), low (0.5-1.5), moderate (1.5-2.5), and strong (2.5-3). Based on the pathologists' assessments of core quality, we excluded missing tissue cores or cores that were uninterpretable secondary to artifacts (25%). For cases with tumors with two satisfactory cores, the results were averaged (59%); for cases with tumors with one poor-quality spot, results were based on the interpretable core (28%). The remaining cases with two tumor cores that could not be interpreted were excluded (13%), leaving a total of 339 cases available for analyses, on average, for each stain.

Aperio-derived images were scored using two systems: TMAx and TMALab II. The Ariol system was used for both scanning and scoring. We excluded cores with blurred images (2%). Before analysis, we adjusted automated scoring algorithms for size and shape parameters in an effort to limit analysis to carcinoma cells within each spot and adjusted intensity thresholds to distinguish positive immunohistochemical reactions from background counterstaining. For the TMAx system, algorithm tuning and analysis were done by the vendor. TMALab II algorithms were initially set by the vendor and then refined independently by two pathologists (P.L. and S.M.H.). The Ariol system algorithm was tuned by an image analysis expert (W.J.H.) with the support of a pathologist (M.E.S.). Algorithms for nuclear stains were used to score ER-α, ER-β, and PR, and cell membrane algorithms were used to score HER2. At the time of the analysis, refined automated scoring algorithms for cytoplasmic markers were not available and aromatase staining was quantified as the

average positive staining intensity of the entire spot. For ER-α, ER-β, and PR, the systems calculated the percent of cells stained (1-100%) and average positive stain intensity as a continuous measure. For HER2 and aromatase, only the continuous intensity measure was used. As in the pathologists' scores, the product of percentage and intensity was used as the main staining measure for ER-α, ER-β, and PR.

mRNA expression

Among the breast cancer cases included in this analysis, samples of 84 tumors had been snap frozen, stored in liquid nitrogen (–196°C), and subsequently profiled for mRNA expression. Briefly, ~30 mg of frozen tissue were processed to isolate RNA with Trizol reagent (Invitrogen), and the resulting RNA was purified with Qiagen RNeasy Mini columns. Aliquots of 250 ng of input RNA were amplified and labeled using the Illumina TotalPrep RNA Amplification kit (Applied Biosystems/Ambion) according to the manufacturer's protocol. The biotin-labeled cRNAs were quantitated using RiboGreen RNA Quantitation reagent (Molecular Probes), and 750 ng was hybridized to Illumina HumanRef-8 v2 Expression Beadchip microarrays (Illumina).

Statistical methods

We assessed agreement between automated immunohistochemical scores and pathologists' scores of negative versus positive and categorical strength of staining. To quantify the agreement between automated scores and pathologists' negative/positive scores, we evaluated the area under the curve (AUC) of the receiver operating characteristic (ROC) graphs for each instrument, considering the pathologists' result of negative or positive as the reference. The ROC curve plots the true-positive versus false-positive fraction for each possible cutoff point that could have been used to define negative versus positive tumors. The AUC of the ROC graph represents the probability that an automated score will be higher for a randomly chosen true-positive sample (defined by the pathologists) than for a randomly chosen true-negative sample. An AUC of 1.0 would represent perfect discrimination of the pathologists' negative/positive categorization by an automated instrument and an AUC of 0.5 would correspond to no discriminatory accuracy. The AUCs for the three automated systems were compared using a nonparametric method (9).

To assess the agreement between the continuous scores of the automated instruments and the categorical scores of the pathologists for strength of staining, we converted the automated scores into the four categories used by the pathologists (negative, low, moderate, and strong staining). This was done by aligning the distributions of the pathologist semiquantitative scores with the automated scores. For example, if 30% of samples for a marker were categorized as negative by the pathologists, then the automated scores corresponding to the lowest 30th percentile of the automated results were categorized as negative. We measured the agreement using a weighted κ statistic, which represents the agreement exceeding that expected by chance. For comparison, we calculated intraobserver agreement for one pathologist (M.A.D.) and interobserver agreement between the two pathologists (M.E.S. and M.A.D.). We interpreted κ values of 0.8 to 1.0 as almost perfect agreement, 0.6 to 0.8 as substantial agreement, 0.4 to 0.6 as moderate agreement, 0.2 to 0.4 as fair agreement, and 0.0 to 0.2 as slight agreement (10). To assess the effect of removing spots of poor quality from our main analyses, we compared the agreement between automated systems in spots of adequate quality with the agreement in spots of low quality.

Standard Illumina preprocessing was applied to the mRNA expression data. Specifically, the variance stabilization transformation was used followed by quantile normalization. We estimated correlations between immunohistochemical scores and mRNA levels by marker and immunohistochemical scoring method using Spearman's rank correlation test and used the Fisher r-to-z transformation to generate confidence intervals (11).

Table 1. AUC for automated systems discriminating between pathologist negative/positive scores

Stain	AUC (95% CI)	*P**
ER-α		
Aperio	0.98 (0.97-0.99)	
TMAx	0.99 (0.97-1.00)	
Ariol	0.99 (0.98-1.00)	0.08
PR		
Aperio	0.99 (0.98-0.99)	
TMAx	0.99 (0.98-1.00)	
Ariol	0.99 (0.99-1.00)	0.16
HER2		
Aperio	0.93 (0.89-0.97)	
TMAx	0.97 (0.94-1.00)	
Ariol	0.96 (0.92-0.99)	0.006
ER-β		
Aperio	0.99 (0.98-1.00)	
TMAx	1.00 (0.99-1.00)	
Ariol	0.99 (0.98-1.00)	0.32
Aromatase		
Aperio	0.96 (0.93-0.99)	
TMAx	0.96 (0.93-0.98)	
Ariol	0.85 (0.80-0.90)	0.0001

NOTE: For ER-α, PR, and ER-β, the combined (intensity × percent) score was used. For HER2 and aromatase, the intensity score was used.

**P* value is for the test of the null hypothesis that the AUCs for all three automated systems are equal.

Results

ROC analysis

The AUC for all the three automated systems was roughly 0.99 for discrimination of pathologist positive/negative categorization of ER-α, PR, and ER-β staining (Table 1). There was some difference in the capacity of the automated systems to discriminate between pathologist negative/positive categories for HER2 and aromatase. In the case of HER2 staining, TMALab II showed less agreement with pathologists' scores (AUC = 0.93) than the other systems (AUC = 0.96-0.97). Ariol showed lower agreement with pathologist scores for the aromatase stain (AUC = 0.85) compared with the other systems (AUC = 0.96). Results for the percent-only scores compared with the combined percent-intensity scores for the nuclear markers (ER-α, ER-β, and PR) yielded similar results for ROC analyses, apart from a slightly lower agreement for the TMALab II system with pathologists for ER-β (AUC = 0.90).

Agreement between pathologist's semiquantitative scores and automated systems

Agreement between pathologists' semiquantitative and automated scores for ER-α and PR were almost perfect for all systems (κ = 0.86-0.91; Fig. 1). The intrapathologist reproducibility (κ = 0.86) and interpathologist reproducibility (κ = 0.91-0.93) were also high. There was substantial agreement between automated and pathologist interpretations of HER2 staining levels for TMAx and Ariol (κ = 0.69-0.72) but less (κ = 0.53) for TMALab II. Interpathologist (κ = 0.86) and intrapathologist agreement (κ = 0.95) for HER2 staining levels was excellent. When we restricted the analysis to 296 tumors scored as negative (0+) or strongly positive (3+), agreement improved: TMAx and Ariol (κ = 0.83-0.90), Aperio (κ = 0.69), interpathologist (κ = 0.98). The agreement between automated scores and pathologists' scores for ER-β was excellent with minimal differences among instruments (κ = 0.80-0.86). Intraobserver agreement (κ = 0.84) was substantially higher than interobserver concordance for this marker (κ = 0.69). Of the five markers, aromatase

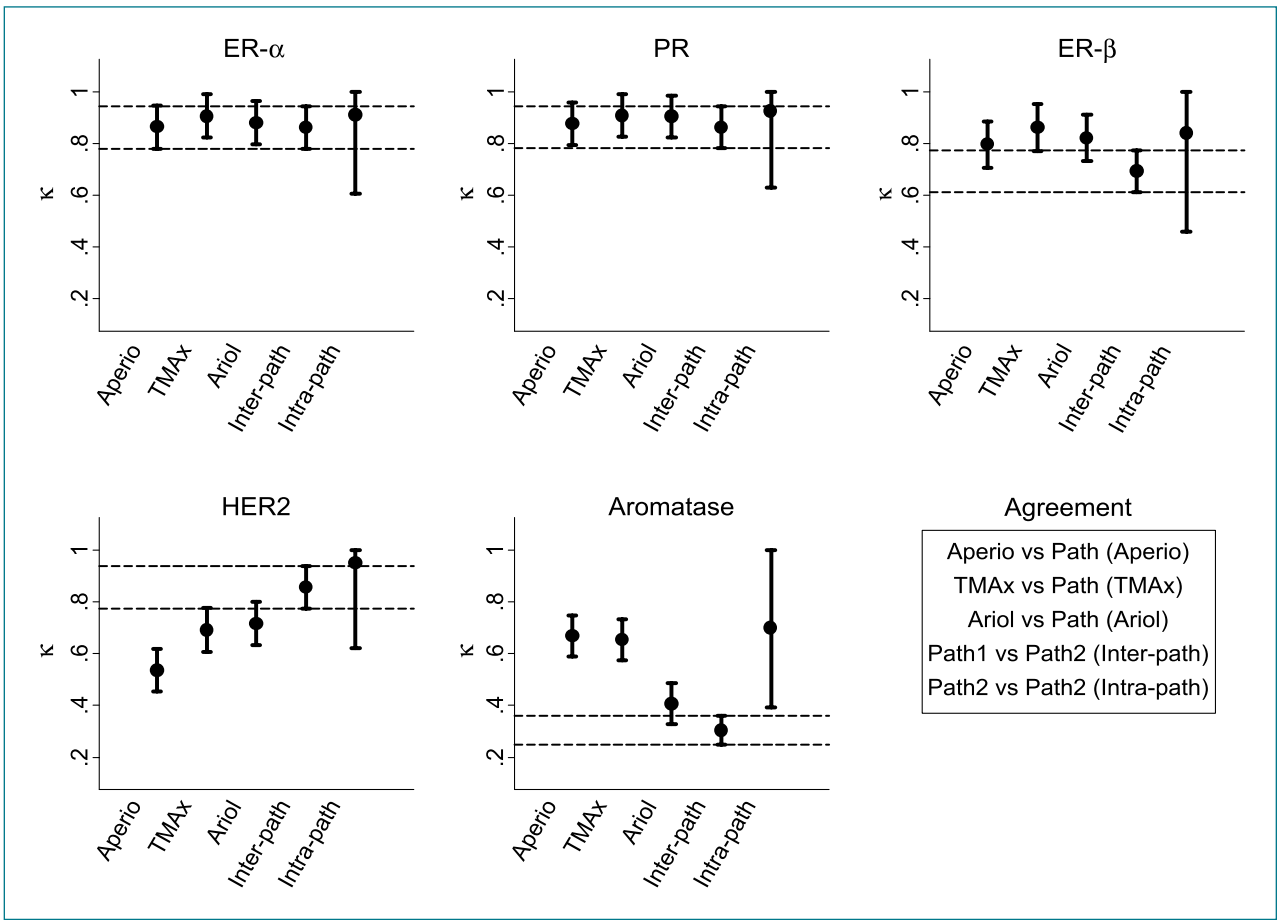

Figure 1. Agreement between pathologist and automated staining categories. For ER-α, PR, and ER-β, the combined (intensity × percent) score was used; for HER2 and aromatase, the intensity score was used. Bars indicate the 95% confidence interval for the κ statistic. Dashed lines indicate the 95% confidence limits of the κ statistic for the interpathologist agreement.

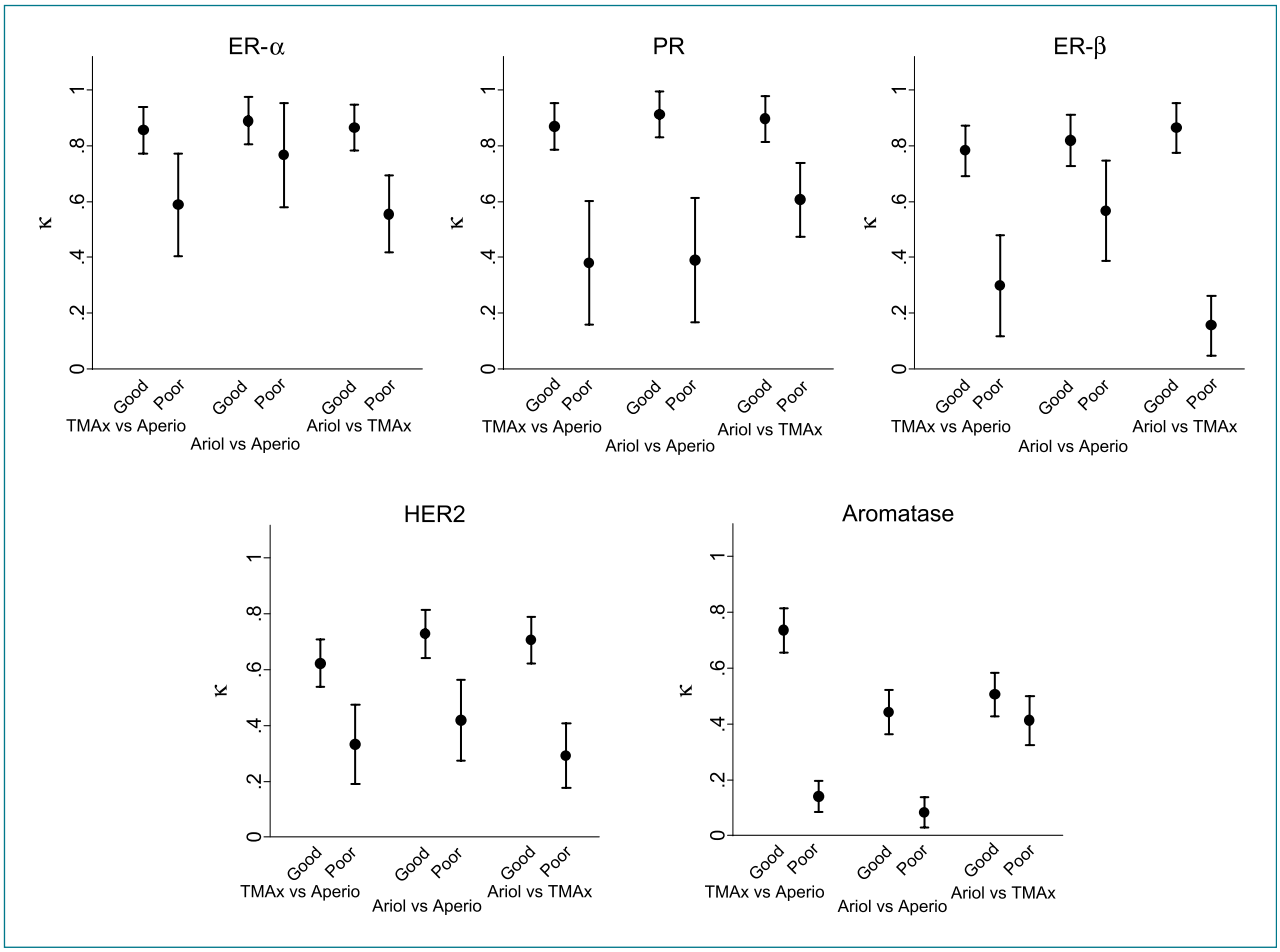

Figure 2. Agreement between automated systems by spot quality. For ER-α, PR, and ER-β, the combined (intensity × percent) score was used; for HER2 and aromatase, the intensity score was used. Spots were considered to have poor quality due to technical issues with the core (core partially missing, few carcinoma cells, thick or folded) or staining issues (partially stained core, difficult to interpret staining patterns). The percentage of poor-quality spots for each stain is as follows: ER-α and PR (18%), ER-β (33%), HER2 (29%), and aromatase (40%).

automated scores showed the poorest agreement with pathologist scores, with Ariol having lower agreement (κ = 0.41) than TMALab II and TMAx (κ = 0.65-0.67). The interpathologist (κ = 0.30) and intrapathologist agreement (κ = 0.70) was also lowest for this marker.

Agreement between pathologist semiquantitative scores for percent of positive staining cells did not differ substantially from the agreement levels seen for the combined score for the nuclear markers ER-α, PR, and ER-β. The two exceptions to this were a decrease in the agreement between TMALab II and pathologist ER-β percent scores (κ = 0.51) and an increase in the interpathologist agreement (κ = 0.82) when using percent-only ER-β scores.

We saw minimal differences in the agreement between pathologist and TMALab II staining categories generated by the three different users across all five markers (results not shown). The agreement between automated systems exhibited similar patterns to the agreement between pathologists and automated systems, with ER-α and PR showing the highest (κ = 0.85-0.91) and aromatase showing the lowest agreement (κ = 0.44-0.74). In general, there was a marked improvement in the agreement between automated systems in spots of adequate quality compared with spots of poor quality (Fig. 2). ER-β and aromatase had the highest proportion of spots, which were difficult to interpret by the pathologists. ER-β, although a nuclear marker, displayed a variable amount of cytoplasmic and background staining, which made evaluation difficult. Aromatase generally showed weak diffuse cytoplasmic staining with frequent concurrent staining of the stroma, which was also difficult to visually interpret (Supplementary Fig. S1). In general, across all five markers, disagreement between pathologists and automated systems was limited to one-category discordances (e.g., weak versus moderate staining). When we examined the instances of extreme discordance between pathologists and the Ariol system, a wide variety of issues seemed to be driving the disagreement, including misclassification of normal cells as tumor cells, staining artifacts, and equivocal staining (Supplementary Fig. S2).

Correlation between immunohistochemical scores and mRNA expression

Immunohistochemical and mRNA levels for ER-α and PR were highly correlated (Table 2) based on both automated systems (ρ = 0.67-0.74) and visual reads by pathologists (ρ = 0.67-0.77). HER2 immunohistochemical and mRNA levels were also correlated, but the association was stronger for pathologists' scores (ρ = 0.63) than automated methods (ρ = 0.41-0.49). Immunohistochemical staining and mRNA levels for ER-β and aromatase were not correlated by any method. We investigated whether ER-β mRNA expression was correlated with immunohistochemical staining intensity or the percent of cells staining positively but did not see an association for either staining parameter (results not shown).

Table 2. Correlation between mRNA expression levels and immunohistochemical staining levels as measured by pathologists and automated systems

Stain	ρ (95% CI)	*P*
ER-α		
Pathologist	0.67 (0.45-0.89)	4×10^{-12}
Aperio	0.73 (0.51-0.95)	4×10^{-15}
TMAx	0.74 (0.52-0.96)	1×10^{-15}
Ariol	0.67 (0.45-0.89)	2×10^{-12}
PR		
Pathologist	0.77 (0.55-0.99)	6×10^{-17}
Aperio	0.69 (0.47-0.91)	6×10^{-13}
TMAx	0.73 (0.51-0.95)	1×10^{-14}
Ariol	0.73 (0.51-0.95)	5×10^{-15}
HER2*		
Pathologist	0.63 (0.41-0.85)	4×10^{-10}
Aperio	0.43 (0.21-0.65)	1×10^{-4}
TMAx	0.49 (0.27-0.71)	4×10^{-6}
Ariol	0.41 (0.19-0.63)	2×10^{-4}
ER-β		
Pathologist	–0.04 (–0.22 to 0.22)	0.76
Aperio	0.09 (–0.13 to 0.31)	0.43
TMAx	0.003 (–0.22 to 0.22)	0.97
Ariol	–0.12 (–0.34 to 0.1)	0.3
Aromatase*		
Pathologist	0.06 (–0.17 to 0.29)	0.61
Aperio	–0.03 (–0.26 to 0.2)	0.79
TMAx	0.04 (–0.19 to 0.27)	0.76
Ariol	0.09 (–0.14 to 0.32)	0.44

NOTE: For ER-α, PR, and ER-β, the combined (intensity × percent) score was used. For HER2 and aromatase, the intensity score was used.

*Results are for the expression probe with the highest correlation coefficients.

Discussion

This study shows that using automated systems to assess immunohistochemical stains done on TMAs in epidemiologic studies of breast cancer is a promising approach. The three commercial systems that we evaluated performed similarly well; however, performance was less encouraging for some markers (HER2, aromatase, and ER-β), irrespective of mode of assessment.

ROC analysis done to assess agreement between automated instruments and pathologists at the level of negative versus positive for the nuclear markers ER-α and PR showed nearly perfect agreement for all systems. In addition, agreement between automated systems and pathologists for strength of ER-α and PR staining was excellent. Previous studies have also shown close agreement between automated assessments and pathologists' scores for ER-α and PR immunohistochemical stains in breast cancer (7, 12-16). Rexhepaj et al. (15) report an AUC of 0.85 for ER-α and 0.74 for PR comparing pathologist and automated scores using an in-house image analysis system. Turbin et al. (13) report κ statistics of 0.88 to 0.90 when comparing pathologist and Ariol scores of positive/negative staining for ER-α, and Diaz et al. (12) report a κ of 0.84 using the QCA image analysis system.

After dichotomizing automated HER2 scores, we showed strong agreement with pathologists' scores of negative and positive. Agreement was good but not as strong for comparisons of the strength of HER2 staining. Concordance between TMALab II and pathologists was minimally less for HER2 staining levels than for other instruments. Recalibration of the tuning parameters by multiple users failed to improve agreement, suggesting that the result was operator independent. The lower level of agreement for HER2 staining categories could reflect difficulties in visual quantification of intermediate staining levels, a problem that is well documented (17-19). Indeed, when we restricted the analysis to unequivocal results of 0 or 3+, the agreement between automated and pathologists' scores improved across all instruments. Previously, Joshi et al. (20) reported somewhat higher levels of agreement between automated and pathologists' scores of 0, 1+, 2+, and 3+ (κ = 0.80-0.91) using an in-house image analysis system. One study using the Ariol system (6) reported κ = 0.84 between automated and pathologist scores of 0/1+, 2+, and 3+. We achieved almost identical results when we reanalyzed our Ariol data using this grouping (κ = 0.82).

Our findings about the use of automated scoring systems for ER-β and aromatase are less clear. Although, in general, there was good agreement between the automated systems and pathologists scores, we saw lower levels of interpathologist agreement, particularly for aromatase. These data are consistent with our experience that measurement of nuclear-specific markers is quite reproducible, whereas measurement of markers in other cell compartments (cell membrane for HER2; cytoplasm for aromatase) or markers that show staining of multiple compartments (i.e., ER-β) is more challenging.

We do not believe that difficulty in scoring antibodies for ER-β and aromatase is specific to the antibodies used in our study but rather is exemplary of the general challenges related to assessing nonnuclear or multicompartment stains. Reports suggest that ER-β antibodies may show both nuclear and cytoplasmic staining, which, although difficult to score, may have prognostic significance (21). Similarly, the aromatase antibody we used shows diffuse cytoplasmic staining, which presents scoring challenges, but is typical of a valid pattern of staining that one may wish to assess with automated systems if possible (22). Overall, our results for interobserver agreement and automated scoring emphasize the need to validate automated methods for specific assays, show the range of agreement that is obtained for visual and automated scoring of different markers, and provide impetus for further methodologic development of both immunohistochemcial assays and automated scoring techniques.

We found substantially better agreement between scores for markers of adequate versus poor quality, indicating that triage of spots is an important quality assurance measure for automated analysis. It is generally recommended that TMAs include two to four cores from each sample to minimize the effect of tissue misrepresentation and missing results (23-25). Redundant representation of tumors in TMAs when doing automated image analysis is particularly useful because many artifacts that are not limiting for pathologists may interfere with automated reads.

In the absence of protein quantification or other means of assessing accuracy, we assessed whether automated and pathologists' immunohistochemical scores produced similar correlations with mRNA levels. Correlations for ER-α, PR, and HER2 were of particular interest because of their clinical and epidemiologic relevance. For ER-α and PR, all measures were highly correlated, suggesting equally accurate representation of gene expression. For HER2, immunohistochemical and mRNA levels were more strongly correlated with pathologist than automated scores, although results for the latter were also highly significant. In contrast, none of the immunohistochemical scores was correlated with mRNA for ER-β and aromatase. Given that expression at the mRNA and protein levels is not necessarily correlated, this result may not indicate poor immunohistochemical scoring, especially because this was seen uniformly across automated and pathologist scores.

The strengths of our study include evaluation of multiple markers and instruments, using tissues from a population-based study, the use of multiple pathologists' interpretations, and exclusion of poor images for automated analysis. There are several limitations of our study. First, we did the image analysis in a fully automated mode, although the Ariol and TMALab II systems are designed to be run on tumor-rich regions of cores marked by a trained reviewer. By including some benign tissue in the scoring, this could have negatively affected the performance of these two systems. Cores that contained benign epithelium or abundant stroma seemed to diminish agreement between the automated systems and pathologists. Indeed, spots showing a high level of discordance between pathologists and automated systems often seemed to be caused by the misclassification of normal cells as tumor cells. However, we assumed that gating on cell features of tumor cells and the preparation of TMAs using cores removed from tumor-rich areas of tissue minimized the effect of this approach on most samples. In addition, the use of different scanners and different procedures for tuning algorithms may have reduced the validity of direct comparisons between the three systems. However, we did not notice any systematic differences in the performance of the systems using Aperio-derived images (TMALab II and TMAx) compared with the Ariol system, suggesting that the scanner variation had a minimal effect. Although our data suggest that there is minimal effect from scanner variation, this subject has not been rigorously tested to date. Appropriate investigation of these issues requires cross-platform and cross-instrument comparisons to provide a systematic understanding of any potential sources of bias. Although experience with a system may enhance performance, we attempted to optimize all automated analyses and observed that performance of automated analyses was similar for different users.

Our study shows that fully automated image analysis systems can provide results that agree well with pathologist scores, particularly for robust nuclear markers such as ER-α and PR. Automated image analysis systems can greatly facilitate large-scale multicenter epidemiologic studies by providing standardized, quantitative measures of immunohistochemical staining. Although reducing misclassification of TMA scoring is crucial, it is not the only factor affecting immunohistochemical data. Other aspects influencing immunohistochemical results include delays in the time to formalin fixation (26), variation in the adequacy of formalin fixation (27), and improper storage of cut and unstained slides. Our group (28) and others (29) have previously evaluated issues related to slide storage but the difficult issue of the effect of tissue fixation on immunohistochemical data has yet to be completely addressed. Tackling each of these steps will be needed to realize the full potential of tissue-based epidemiologic research.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Steven Hashagen (Aperio Technologies) and Juha Kononen (Beecher Instruments) for doing the automated image analysis using the TMALab II and TMAx systems, respectively; Pei Chao (IMS, Silver Spring, MD) for the data management of the study; and Kirsten Colquhoun for editing the manuscript.

Grant Support

K.L. Bolton, M. Garcia-Closas, R.M. Pfeiffer, S.M. Hewitt, X.R. Yang, R. Cornelison, S.L. Anzick, P. Meltzer, S. Davis, P. Lenz, J.D. Figueroa, and M.E. Sherman were funded by the National Cancer Institute Intramural Research Program, including resources of the Applied Molecular Pathology Laboratory. M.A. Duggan was funded by the University of Calgary. P.D.P. Pharaoh and W.J. Howat were funded by Cancer Research UK.

Received 10/01/2009; revised 01/27/2010; accepted 02/02/2010; published OnlineFirst 03/23/2010.

References

1. Garcia-Closas M, Hall P, Nevanlinna H, et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet 2008;44:e1000054.
2. Reeves GK, Beral V, Green J, Gathani T, Bull D. Hormonal therapy for menopause and breast-cancer risk by histological type: a cohort study and meta-analysis. Lancet Oncol 2006;7:910–8.
3. Ma H, Bernstein L, Pike MC, Ursin G. Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: a meta-analysis of epidemiological studies. Breast Cancer Res 2006;84:R43.
4. Garcia-Closas M, Brinton LA, Lissowska J, et al. Established breast cancer risk factors by clinically important tumour characteristics. Br J Cancer 2006;951:123–9.
5. Kirkegaard T, Edwards J, Tovey S, et al. Observer variation in immunohistochemical analysis of protein expression, time for a change? Histopathology 2006;487:787–94.
6. Turashvili G, Leung S, Turbin D, et al. Inter-observer reproducibility of HER2 immunohistochemical assessment and concordance with fluorescent *in situ* hybridization (FISH): pathologist assessment compared to quantitative image analysis. BMC Cancer 2009;9:165.
7. Gokhale S, Rosen D, Sneige N, et al. Assessment of two automated imaging systems in evaluating estrogen receptor status in breast carcinoma. Appl Immunohistochem Mol Morphol 2007;154:451–5.
8. Yang XR, Pfeiffer RM, Garcia-Closas M, et al. Hormonal markers in breast cancer: coexpression, relationship with pathologic characteristics, and risk factor associations in a population-based study. Cancer Res 2007;67:10608–17.
9. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;443:837–45.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;331:159–74.
11. Altman DG. Practical statistics for medical researchers. New York: Chapman & Hall; 1991.
12. Diaz LK, Sahin A, Sneige N. Interobserver agreement for estrogen receptor immunohistochemical analysis in breast cancer: a comparison of manual and computer-assisted scoring methods. Ann Diagn Pathol 2004;81:23–7.
13. Turbin DA, Leung S, Cheang MC, et al. Automated quantitative analysis of estrogen receptor expression in breast carcinoma does not differ from expert pathologist scoring: a tissue microarray study of 3,484 cases. Breast Cancer Res Treat 2008;1103:417–26.
14. Sharangpani GM, Joshi AS, Porter K, et al. Semi-automated imaging system to quantitate estrogen and progesterone receptor immunoreactivity in human breast cancer. J Microsc 2007;226 (Pt 3): 244–55.
15. Rexhepaj E, Brennan DJ, Holloway P, et al. Novel image analysis approach for quantifying expression of nuclear proteins assessed by immunohistochemistry: application to measurement of oestrogen and progesterone receptor levels in breast cancer. Breast Cancer Res 2008;105:R89.
16. Mofidi R, Walsh R, Ridgway PF, et al. Objective measurement of breast cancer oestrogen receptor status through digital image analysis. Eur J Surg Oncol 2003;291:20–4.
17. Interobserver reproducibility of immunohistochemical HER-2/neu assessment in human breast cancer: an update from INQAT round III. Int J Biol Markers 2005;203:189–94.
18. McCormick SR, Lillemoe TJ, Beneke J, Schrauth J, Reinartz J. HER2 assessment by immunohistochemical analysis and fluorescence *in situ* hybridization: comparison of HercepTest and PathVysion commercial assays. Am J Clin Pathol 2002;1176:935–43.
19. Lacroix-Triki M, Mathoulin-Pelissier S, Ghnassia JP, et al. High interobserver agreement in immunohistochemical evaluation of HER-2/neu expression in breast cancer: a multicentre GEFPICS study. Eur J Cancer 2006;4217:2946–53.
20. Joshi AS, Sharangpani GM, Porter K, et al. Semi-automated imaging system to quantitate Her-2/neu membrane receptor immunoreactivity in human breast cancer. Cytometry A 2007;715:273–85.
21. Shaaban AM, Green AR, Karthik S, et al. Nuclear and cytoplasmic expression of ERβ1, ERβ2, and ERβ5 identifies distinct prognostic outcome for breast cancer patients. Clin Cancer Res 2008;14: 5228–35.
22. Sasano H, Anderson TJ, Silverberg SG, et al. The validation of new aromatase monoclonal antibodies for immunohistochemistry—a correlation with biochemical activities in 46 cases of breast cancer. J Steroid Biochem Mol Biol 2005;95:35–9.
23. Camp RL, Charette LA, Rimm DL. Validation of tissue microarray technology in breast carcinoma. Lab Invest 2000;8012:1943–9.
24. Rubin MA, Dunn R, Strawderman M, Pienta KJ. Tissue microarray sampling strategy for prostate cancer biomarker analysis. Am J Surg Pathol 2002;263:312–9.
25. Hoos A, Urist MJ, Stojadinovic A, et al. Validation of tissue microarrays for immunohistochemical profiling of cancer specimens using the example of human fibroblastic tumors. Am J Pathol 2001;1584: 1245–51.
26. Oyama T, Ishikawa Y, Hayashi M, Arihiro K, Horiguchi J. The effects of fixation, processing and evaluation criteria on immunohistochemical detection of hormone receptors in breast cancer. Breast Cancer 2007;142:182–8.
27. Goldstein NS, Ferkowicz M, Odish E, Mani A, Hastah F. Minimum formalin fixation time for consistent estrogen receptor immunohistochemical staining of invasive breast carcinoma. Am J Clin Pathol 2003;1201:86–92.
28. Fergenbaum JH, Garcia-Closas M, Hewitt SM, et al. Loss of antigenicity in stored sections of breast cancer tissue microarrays. Cancer Epidemiol Biomarkers Prev 2004;13:667–72.
29. Jacobs TW, Prioleau JE, Stillman IE, Schnitt SJ. Loss of tumor marker-immunostaining intensity on stored paraffin slides of breast cancer. J Natl Cancer Inst 1996;88:1054–9.