

# Ki-67 Expression in Breast Cancer Tissue Microarrays

## Assessing Tumor Heterogeneity, Concordance With Full Section, and Scoring Methods

Thaer Khoury, MD,<sup>1</sup> Gary Zirpoli, PhD,<sup>2</sup> Stephanie M. Cohen, PhD,<sup>3</sup> Joseph Geradts, MD,<sup>4</sup> Angela Omilian, PhD,<sup>1</sup> Warren Davis, PhD,<sup>2</sup> Wiam Bshara, MD,<sup>1</sup> Ryan Miller, MD, PhD,<sup>3</sup> Michelle M. Mathews, MD,<sup>3</sup> Melissa Troester, MD,<sup>3</sup> Julie R. Palmer, PhD,<sup>5</sup> and Christine B. Ambrosone, PhD<sup>2</sup>

From the <sup>1</sup>Department of Pathology and <sup>2</sup>Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY; <sup>3</sup>Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill; <sup>4</sup>Department of Pathology, Brigham and Women's Hospital, Boston, MA; and <sup>5</sup>Slone Epidemiology Center, Boston University, Boston, MA.

*Am J Clin Pathol* August 2017;148:108-118

DOI: 10.1093/AJCP/AQX053

### ABSTRACT

**Objectives:** *Ki-67 has been proposed to be used as a surrogate marker to differentiate luminal breast carcinomas (BCs). The purpose of this study was to determine the utility of and best approaches for using tissue microarrays (TMAs) and Ki-67 staining to distinguish luminal subtypes in large epidemiology studies of luminal/human epidermal growth factor receptor 2 (HER2)–negative BC.*

**Methods:** *Full-section and TMA (three 0.6-mm cores and two 1.0-mm cores) slides of 109 cases were stained with Ki-67 antibody. We assessed two ways of collapsing TMA cores: a weighted approach and mitotically active approach.*

**Results:** *For cases with at least a single 0.6-mm TMA core (n = 107), 16% were misclassified using a mitotically active approach and 11% using a weighted approach. For cases with at least a single 1.0-mm TMA core (n = 101), 5% were misclassified using either approach. For the 0.6-mm core group, there were 33.3% discordant cases. The number of discordant cases increased from 18% in the group of two cores to 40% in the group of three cores (P = .039).*

**Conclusions:** *Ki-67 tumor heterogeneity was common in luminal/HER2– BC. Using a weighted approach was better than using a mitotically active approach for core to case collapsing. At least a single 1.0-mm core or three 0.6-mm cores are required when designing a study using TMA.*

Upon completion of this activity you will be able to:

- design a study that uses tissue microarray instead of full section stained with Ki-67.
- define the confounding factors in interpreting Ki-67 expression in breast cancer tissue constructed in a tissue microarray.
- apply the knowledge gained to identify and address issues related to Ki-67 interpretation in clinical practice.

The ASCP is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians. The ASCP designates this journal-based CME activity for a maximum of 1 *AMA PRA Category 1 Credit*<sup>™</sup> per article. Physicians should claim only the credit commensurate with the extent of their participation in the activity. This activity qualifies as an American Board of Pathology Maintenance of Certification Part II Self-Assessment Module.

The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Exam is located at [www.ascp.org/ajcpcme](http://www.ascp.org/ajcpcme).

Breast cancer (BC) is a heterogeneous disease that can be divided into four classes: luminal A–like, luminal B–like, human epidermal growth factor receptor 2 (HER2) enriched, and basal-like.<sup>1,2</sup> Immunohistochemistry surrogate markers have been introduced to aid in separating these molecular subtypes. Ki-67, a nuclear marker expressed in all phases of the cell cycle other than the G0 phase,<sup>3</sup> has been proposed to separate luminal A from luminal B. Urruticoechea et al<sup>4</sup> have found that in 17 of the 18 studies that included more than 200 patients, there was a statistically significant association between Ki-67 and tumor outcome. However, the cutoffs to distinguish “Ki-67 high” from “Ki-67 low” varied from 1% to 28.6%, limiting its clinical utility.<sup>5</sup>

Multiple factors are attributed to this wide range of cutoffs, including preanalytical issues such as type of biospecimen used (core biopsy specimen vs tissue microarrays [TMA] vs full section [FS]), fixation time, cold ischemia time, how the specimen was stored long term, conditions and duration of long-term sectioned slide storage, overnight delay before fixation, freezing the specimen for frozen-section analysis before fixation, use of ethanol or Bouin solution rather than neutral buffered formalin fixation, use of EDTA or acid decalcification protocols, and the choice of Ki-67 antibody and staining protocol.<sup>6-10</sup> There are analytical issues, including intraobserver and interobserver reproducibility, number of cells counted, tumor heterogeneity, and minimal threshold of staining to be considered positive.<sup>5</sup> Moreover, counting nontumor cells such as tumor infiltrating lymphocytes (TILs) may contribute to this variability.<sup>11,12</sup> Another possible confounding factor is counting noninvasive tumor such as ductal carcinoma in situ (DCIS).

To conduct large studies such as large phase III clinical trials or epidemiologic studies, arranging patient tissues into TMAs is the best way to handle high-throughput analyses. However, it is unclear if TMA cores reliably represent the FS. Moreover, the literature is sparse in terms of how to design a study in which TMAs can be used instead of FS. Thus, we conducted this methodologic study, with a particular goal to provide information for the basis of immunohistochemistry (IHC) and scoring in the context of the African American Breast Cancer Epidemiology and Risk (AMBER) consortium, a large multicenter study of risk factors for BC subtypes in African American women.

The goals of this study were (1) to evaluate the rate of discordance between FS and TMA based on the TMA core size and the number of cores, (2) to investigate the best way of collapsing cores to reflect the FS by evaluating two different collapsing methods, (3) to examine tumor heterogeneity using different cutoffs, and (4) to evaluate the effect of TILs and DCIS on tumor classification.

## Materials and Methods

### Case Selection

The BC patient database at Roswell Park Cancer Institute was searched for cases that were estrogen receptor (ER) positive and/or progesterone receptor (PR) positive and HER2 BC negative between 2006 and 2013. ER/PR staining was considered positive when the Allred score

was more than 2. The Allred score is a semiquantitative scoring system incorporating the staining intensity and the percentage of positive cells.<sup>13</sup> HER2 scoring was conducted using the American Society of Clinical Oncology/College of American Pathologists 2007 guidelines, as these cases were selected between 2007 and 2013 when the recent guidelines had not been published yet. HER2 was considered negative by IHC when the score was 0 or 1+ or by fluorescence in situ hybridization when the *HER2/Cep17* ratio was less than 1.8.<sup>14</sup> The cases were selected on the basis of having enough tissue to provide five cores for the TMA blocks: three 0.6-mm cores and two 1.0-mm cores. Therefore, we reviewed only tumors larger than 20 mm. The number and size of the TMA cores were chosen to reflect the AMBER consortium study, which includes two TMA core sizes: 0.6 mm and 1.0 mm. The area size is 0.85 mm<sup>2</sup> for three 0.6-mm cores and 0.78 mm<sup>2</sup> for a single 1.0-mm core. The slides of the candidate cases (n = 109) were annotated. TILs were scored by one pathologist (T.K.) following the international working group recommendations from 0% to 100%.<sup>15</sup>

### TMA Construction and Slide Sectioning

First, four FS from the blocks were cut and stored in a -80°C chamber before punching the blocks for TMA construction. Then, TMAs blocks were constructed using the Beecher tissue puncher and array system (Beecher Instruments, Silver Spring, MD). The core sites were randomly chosen in the tumor regions to avoid bias. FS slides were cut at the time when Ki-67 staining was validated and the TMA technician was ready to perform the TMA blocks construction.

### Ki-67 Staining

Ki-67 (M7240, clone MIB1) antibody was purchased from DAKO (Carpinteria, CA). IHC was performed at the University of North Carolina, Translational Pathology Laboratory using the Bond fully automated slide staining system (Leica Microsystems, Norwell, MA). All staining was optimized under pathologist supervision (T.K.), and final conditions were independently reviewed by an additional pathologist (J.G.).

TMA and FS slides were deparaffinized in Bond Dewax solution (AR9222; Leica Microsystems) and hydrated in Bond Wash solution (AR9590; Leica Microsystems). Antigen retrieval was performed at 100°C for 30 minutes in Bond epitope retrieval solution 1 at pH 6.0 (AR9961; Leica Microsystems). Primary antibody against Ki-67 (1:200) was applied for 15 minutes. Antibody detection of Ki-67 was performed using

the Bond Polymer Refine Detection System (DS9800; Leica Microsystems). A control TMA containing negative and positive (range of intensities) breast tissue samples was constructed and used as a staining control. Negative controls (no primary antibody) were also included.

Stained slides were digitally imaged at  $\times 20$  using the Aperio ScanScope XT (Aperio Technologies, Vista, CA). Digital images were stored in the Aperio Spectrum eSlide Database.

### Slide Annotation and Scoring

Regions of interest on images of whole tissue sections were manually annotated using ImageScope (Aperio Technologies). Image analysis (Aperio XT; Aperio Technologies) was performed. Three annotation layers (FS, DCIS, and lymphocytic aggregate) were created by the breast pathologist (T.K.). FS is defined as annotation around the tumor, including all components, invasive tumor cells, DCIS, TILs, and stromal cells. The final Ki-67 score in the FS is defined as the mean score among all of these constituents, not a specific area (eg, hot spot). DCIS is defined as tumor confined to the basement membrane. Lymphocytic aggregate is defined as grouped lymphocytes separated from the tumor cells. TILs that were intermingled with the tumor were not annotated. All cells in annotated regions were analyzed separately using the Aperio Nuclear V9 algorithm (Aperio Technologies) with minor adjustments for cell shape and stain optical densities. Tissue cores in images from TMA slides were segmented using TMA lab (Aperio Technologies) and were analyzed with the same algorithm that was used for whole tissue sections.

First, the quality of the cores was assessed by one of the pathologists (T.K.). Cores were excluded from the analysis due to folded or substantial absence of the tissue ( $>75\%$  of the core). In addition, to exclude these components as the reason for discordance between FS and TMA cores, cores with pure DCIS ( $n = 5$  cores) or heavy TILs ( $n = 4$  cores) were excluded from the analysis. Heavy TILs are defined as a TIL percentage of more than 20%. The total number of evaluable cores was 89 for 0.6-mm core 1, 87 for 0.6-mm core 2, 95 for 0.6-mm core 3, 91 for 1.0-mm core 1, and 87 for 1.0-mm core 2.

A tumor cell was considered positive by automated image analysis when there was staining at any intensity above the positivity threshold recognized by the Aperio algorithm (marked yellow, orange, or red, with the latter being the most intense staining). We first used a range of cutoffs from 5% to 20% to evaluate accuracy and rate of discordance with each cutoff sequentially. Then, we

used more commonly employed cutoffs (10% and 14%) for more detailed analyses. A 14% cutoff was suggested to separate the tumors into luminal A and luminal B.<sup>16</sup> However, we acknowledge the fact that this cutoff has not been validated. To maximize our ability to address potential issues that could arise in the AMBER consortium data or other large studies that use TMA techniques, we did the following comparisons in duplicate (using a 14% cutoff and a 10% cutoff) using  $2 \times 2$  tables: (1) FS vs two annotated layers (DCIS and aggregated lymphocytes), (2) FS vs every single core (three 0.6-mm cores and two 1.0-mm cores), and (3) evaluate Ki-67 expression heterogeneity.

When there were at least two cores to collapse, we tested two core-to-case collapsing methods to define Ki-67 status for each case, the weighted approach and the mitotically active approach, as previously described.<sup>17</sup> The first assigned case-level status using a tumor cellularity-weighted approach. The weighted average of percent positivity was calculated by summing the product of percent positivity and core weight across all cores per case. Core weight was defined as the number of nuclei in a given core divided by the total number of tumor nuclei across all cores for that case. Two thresholds ( $\geq 14\%$  and  $\geq 10\%$ ) were subsequently and separately applied to define a dichotomous status. The second core-to-case collapsing method classified the case as mitotically active if any core had Ki-67 expression greater than or equal to the cutoff (14% and 10% in two separate analyses). For example, if there are two cores from a single case with Ki-67, and the score in core 1 was 100 positive cells/1,000 total cells (10%) and in core 2 was 20 positive cells/100 total cells (20%), there will be two scores using the weighted approach and the mitotically active approach. In the weighted approach, the final score equals the average of these two scores, taking into consideration the weight of each core. Therefore, the final score equals  $100 + 20$  positive cells/1,000 + 100 total cells = 10.9%. In the mitotically active approach, the final score is the highest among the cores regardless of the number of cells represented in each. In this example, the final score is 20%.

### Statistical Analysis

Agreement between scores was measured with Cohen's  $\kappa$ . Nonparametric associations were measured with Spearman's rank correlation coefficient. Rank-sum tests were used to compare median tumor cellularity between cases with all cores concordant vs cases with at least one discordant core, and  $\chi^2$  tests were used to compare rates of biomarker discordance among cases with

two and three TMA cores. All analyses were performed using SAS 9.4 statistical software (SAS Institute, Cary, NC). A *P* value of less than .05 was considered statistically significant, and all statistical tests were two-sided.

## Results

### TILs and Their Correlation With Ki-67 Score

There were only five cases with TILs more than 20%. Most cases (*n* = 100, 92%) had TILs of 10% or less. There was no correlation between the degree of TILs and Ki-67 expression (data not shown).

### Automated Analysis of FS and Annotated Layers

The median (range) of Ki-67 score was 8.2% (0.56%-28.3%) for FS, 6.0% (0.21%-32.9%) for DCIS, and 4% (0%-9.9%) for aggregated lymphocytes. We compared FS with the annotated layers. The findings were similar using either cutoff (14% or 10%). DCIS had a higher Ki-67 score than the FS in three (5.5%) of 55 cases (Supplemental Figure 1; all supplemental materials can be found at *American Journal of Clinical Pathology* online) and a lower score in six (10.9%) (Supplemental Figure 2). Aggregated lymphocytes had a higher Ki-67 score in zero of 74 cases and a lower score in 14 (18.9%). If DCIS were scored instead of invasive carcinoma, nine (8.2%) cases would have had different results from the FS. When DCIS and aggregated lymphocytes layers were subtracted from the FS, the outcome was not affected using either cutoff.

### Comparisons for FS vs TMA Cores

#### *FS vs Single TMA Cores*

*Using a 14% cutoff.* For 0.6-mm cores, the number of cores with discordant scores was 14 (16%) for core 1, 11 (13%) for core 2, and 18 (19%) for core 3. For 1.0-mm cores, the number of cores with discordant scores was seven (8%) for each core. While 0.6-mm cores had better sensitivity, 1.0-mm cores had better specificity. However, core 3 had a very low sensitivity of 24% that is due to the small number of cases that had Ki-67 of 14% or more. The overall accuracy for 1.0-mm cores was better than that for 0.6-mm cores (92% vs 81% to 87%) **Table 1** and **Figure 1**.

*Using a 10% cutoff.* For 0.6-mm cores, the number of cores with discordant scores was 23 (26%) for core 1, 21 (24%) for core 2, and 27 (28%) for core 3. For 1.0-mm cores, the number of cores with discordant scores was 18 (19%) for core 1 and 21 (24%) for core 2 (**Table 1** and **Figure 1**).

Overall, after excluding core 3, the sensitivity was better for 0.6-mm cores and specificity was better for 1.0-mm cores. The overall accuracy was slightly better for 1.0-mm cores vs 0.6-mm cores (80% and 76% vs 76% vs 74%).

When using a range of cutoffs from 5% to 20%, there was an overall trend of better accuracy and less discordance toward higher cutoffs regardless of the core size. However, the variability was more pronounced in 0.6-mm cores (Supplemental Figure 3).

### Core-to-Case Collapsing Analyses

Three 0.6-mm cores from each case were assembled in three TMA blocks. However, the number of cases that had three evaluable cores was 68 (63%), two evaluable cores was 28 (26%), and a single evaluable core was 11 (10%). Two 1.0-mm cores from each case were assembled in two TMA blocks. However, the number of cases that had a single evaluable core was 24 (22%), and two evaluable cores was 77 (71%). The number of cases that did not have any evaluable core was one (1%) for 0.6-mm cores and seven (6%) for 1.0-mm cores.

In this analysis, we included all cases that had at least a single 0.6-mm evaluable core. We used two approaches to collapse cores, the mitotically active approach and the weighted approach **Table 2** and **Figure 2**.

- 0.6-mm cores using a 14% cutoff:* The number of cases with discordant scores was 16 (15%) using the mitotically active approach and 11 (11%) using the weighted approach with accuracies of 85% and 90%, respectively.
- 0.6-mm cores using a 10% cutoff:* The number of cases with discordant scores was 28 (26%) using the mitotically active approach and 20 (18%) using the weighted approach with accuracies of 74% and 81%, respectively.
- 1.0-mm cores using a 14% cutoff:* A similar approach was conducted for 1.0-mm cores. The number of cases with discordant scores was five (5%) using either approach, with an accuracy of 95% for each.
- 1.0-mm cores using a 10% cutoff:* The number of cases with discordant scores was 22 (22%) using the mitotically active approach and 19 (18%) using the weighted approach with accuracies of 78% and 81%, respectively.

When using a range of cutoffs from 5% to 20%, there was a trend of better accuracy and less discordance toward higher cutoffs in 1.0-mm cores when the weighted approach was used. This trend was less obvious in the



**Table 1**

**Comparisons Between Full Section and Single TMA Cores Using 14% and 10% Cutoffs**

TMA cores	Full Section						
	<14%, No. (%)	≥14%, No. (%)	Total, No. (%)	Sensitivity, %	Specificity, %	Accuracy, %	κ (95% CI)
0.6 mm							
Core 1							
<14%	59 (66)	4 (5)	63 (71)	80	86	84	0.59 (0.40-0.78)
≥14%	10 (11)	16 (18)	26 (29)				
Total	69 (77)	20 (23)	89 (100)				
Core 2							
<14%	62 (71)	6 (7)	68 (78)	70	93	87	0.64 (0.44-0.83)
≥14%	5 (6)	14 (16)	19 (22)				
Total	67 (77)	20 (23)	87 (100)				
Core 3							
<14%	72 (76)	16 (17)	88 (93)	24	97	81	0.28 (0.05-0.50)
≥14%	2 (2)	5 (5)	7 (7)				
Total	74 (78)	21 (22)	95 (100)				
1.0 mm							
Core 1							
<14%	71 (78)	6 (7)	77 (85)	68	99	92	0.74 (0.56-0.92)
≥14%	1 (1)	13 (14)	14 (15)				
Total	72 (79)	19 (21)	91 (100)				
Core 2							
<14%	68 (78)	6 (7)	74 (85)	67	99	92	0.73 (0.54-0.92)
≥14%	1 (1)	12 (14)	13 (15)				
Total	69 (79)	18 (21)	87 (100)				
0.6 mm							
Core 1							
<10%	39 (44)	8 (9)	47 (53)	77	72	74	0.48 (0.30-0.66)
≥10%	15 (17)	27 (30)	42 (47)				
Total	54 (61)	35 (39)	89 (100)				
Core 2							
<10%	41 (47)	10 (11)	51 (59)	71	79	76	0.50 (0.32-0.69)
≥10%	11 (13)	25 (29)	36 (41)				
Total	52 (60)	35 (40)	87 (100)				
Core 3							
<10%	54 (57)	23 (24)	77 (81)	38	93	72	0.34 (0.16-0.52)
≥10%	4 (4)	14 (15)	18 (19)				
Total	58 (61)	37 (39)	95 (100)				
1.0 mm							
Core 1							
<10%	50 (55)	13 (14)	63 (69)	64	91	80	0.57 (0.40-0.74)
≥10%	5 (5)	23 (25)	28 (31)				
Total	55 (60)	36 (40)	91 (100)				
Core 2							
<10%	45 (52)	12 (14)	57 (66)	64	83	76	0.48 (0.29-0.67)
≥10%	9 (10)	21 (24)	30 (34)				
Total	54 (62)	33 (38)	87 (100)				

CI, confidence interval; TMA, tissue microarray.

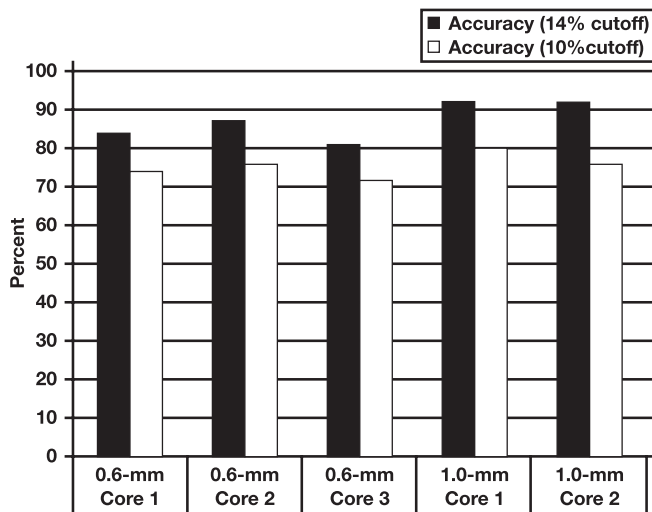
other categories (0.6 mm using any of the two approaches or 1.0 mm using the mitotically active approach) (Supplemental Figure 4).

### Ki-67 Expression Tumor Heterogeneity

We explored tumor heterogeneity in this analysis. A case was considered heterogeneous when at least one core in the same core size group (0.6 mm or 1.0 mm) had a score different from the other core(s). This analysis was performed using two different cutoffs (10% and 14%). For

example, if a 0.6-mm core had a score of 11% and the second core had a score of 20%, the case was considered heterogeneous using a 14% cutoff and homogeneous using a 10% cutoff **Table 3**.

1. *Using a 14% cutoff:* In the 0.6-mm core group that had at least two cores per case, there were 32 (33%) of 96 discordant cases. The number of discordant cases increased from five (18%) of 28 in the group of two cores to 27 (40%) of 68 in the group of three cores ( $P = .039$ ). This



**Figure 1** Bar plot reflecting the findings in Table 1, showing accuracy of scores in each core compared to frozen section, using 14% or 10% cutoffs. Note the following: the accuracy of 1.0-mm cores is better than 0.6-mm cores using a 14% cutoff, the accuracy is slightly better using a 10% cutoff, and the overall accuracy is better across all cores using a 14% cutoff than a 10% cutoff.

calculation could not be performed for the 1.0-mm core group, as a maximum of two cores were constructed in the TMA blocks. The median cellularity was not statistically significantly different between cases with concordant cores and those with any discordant cores using either core size ( $P = .085$  for the 0.6-mm core group and  $.128$  for the 1.0-mm core group).

- Using a 10% cutoff: In the 0.6-mm core group that had at least two cores per case, there were 32 (33%) of 96 discordant cases. The number of discordant cases increased from six (21%) of 28 in the group of two cores to 36 (53%) of 68 in the group of three cores ( $P = .005$ ). The median cellularity was not statistically significantly different between cases with concordant cores and those with any discordant cores using either core size ( $P = .682$  for the 0.6-mm core group and  $.24$  for the 1.0-mm core group).

## Discussion

Overall, using 1.0-mm cores was better than using 0.6-mm cores to reflect Ki-67 scores in FS (Table 1 and Figure 1). The incidence of tumor heterogeneity was higher in 0.6-mm cores than in 1.0-mm cores (Table 3). The weighted approach was better than the mitotically active approach when collapsing core to case in three of

the four comparisons: 0.6 mm (14% cutoff), 0.6 mm (10% cutoff), and 1.0 mm (10% cutoff). For 1.0 mm (14% cutoff), the accuracy was the same in both approaches (Table 2 and Figure 2).

As mentioned above, many preanalytical and analytical factors could affect Ki-67 scoring. To control these factors, we chose cases from a single institution where the time in formalin, cold ischemia time, type of fixation, and the conditions of tissue block storage are well controlled. We have developed a system in our institution to ensure proper tissue handling, and we published a series of studies on the effect of cold ischemia time (also known as delay to formalin fixation) and time in formalin on breast biomarkers.<sup>18-20</sup> We did not use any frozen tissue or tissue that was previously used for frozen-section diagnosis. All tissues were uniformly fixed in 10% neutral buffered formalin. To avoid slide prolonged section storage, we cut the FS slides immediately before constructing the TMA blocks. The staining protocol was verified, and staining started 2 weeks after sectioning. Meanwhile, the FS slides were stored in a  $-80^{\circ}\text{C}$  chamber. We chose the recommended antibody for Ki-67 (MIB-1) by the international Ki-67 in BC working group.<sup>5</sup> To avoid interobserver and intraobserver variability, we performed automated image analysis.

Acs et al<sup>11</sup> found that the mitotic activity of stromal cells including TILs in luminal/HER2 correlated with high Oncotype DX recurrence score. They proposed that high Ki-67 index in these cells may have contributed to false high Oncotype DX scores. However, in a study conducted by us where we found a similar correlation, we proposed that the reason for that could be due to tumor genetic instability, making the tumors more prone to better response to chemotherapy.<sup>12</sup> Nonetheless, since stromal cells may express Ki-67 and potentially could be a confounding factor in the rate of concordance between the TMAs and the FSs, we evaluated Ki-67 in TILs. For that, we estimated TIL density following the international TILs Working Group<sup>15</sup> and annotated the clusters of lymphocytes in the FSs separately. We found that most of the cases had low TILs (>90% had TILs in  $\leq 10\%$  of the stromal spaces) as we previously reported.<sup>21</sup> Ki-67 index was relatively low. Moreover, when we subtracted this component from the FS scores, the concordance with the TMA scores did not change using either cutoff (10% or 14%). DCIS has not been studied in terms of comparing Ki-67 index between DCIS and the concurrent invasive carcinoma in the same tissue section or evaluated in terms of its role in the discordance between FSs and TMAs. For that, we also annotated DCIS separately. We found that the Ki-67 index varied between DCIS and the

**Table 2**

**Core-to-Case Collapse Analyses Using 14% and 10% Cutoffs<sup>a</sup>**

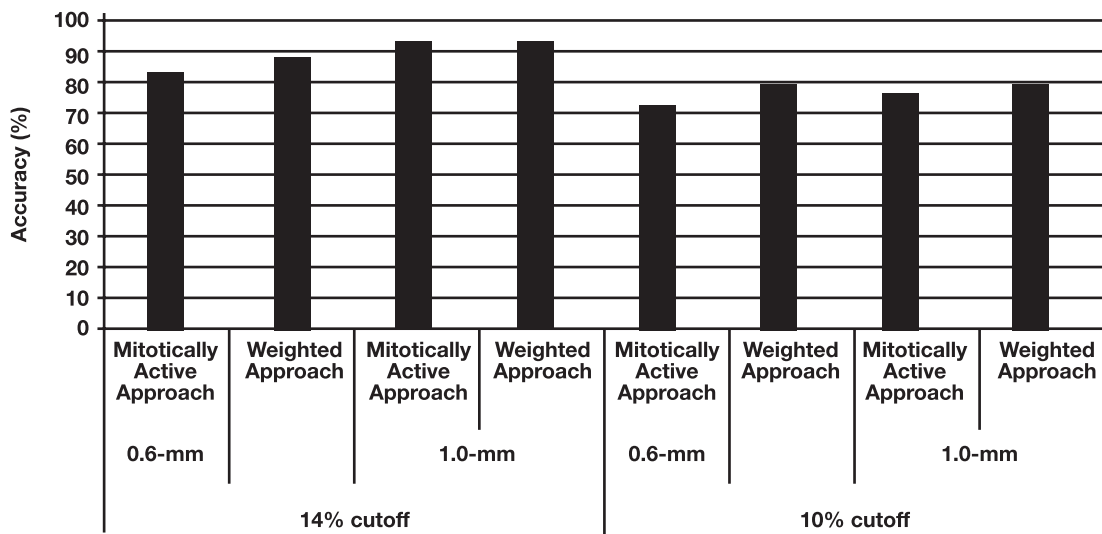
TMA Cores	Full Section						
	<14%, No. (%)	≥14%, No. (%)	Total, No. (%)	Sensitivity, %	Specificity, %	Accuracy, %	κ (95% CI)
0.6 mm							
<14% <sup>b</sup>	70 (65)	0 (0)	70 (65)	100	81	85	0.63 (0.48-0.79)
≥14% <sup>b</sup>	16 (15)	21 (20)	37 (35)				
Total <sup>b</sup>	86 (80)	21 (20)	107 (100)				
<14% <sup>c</sup>	81 (76)	6 (6)	87 (81)	71	94	90	0.67 (0.49-0.85)
≥14% <sup>c</sup>	5 (5)	15 (14)	20 (19)				
Total <sup>c</sup>	86 (80)	21 (20)	107 (100)				
1.0 mm							
<14% <sup>b</sup>	79 (78)	3 (3)	82 (81)	85	98	95	0.84 (0.71-0.98)
≥14% <sup>b</sup>	2 (2)	17 (17)	19 (19)				
Total <sup>b</sup>	81 (80)	20 (20)	101 (100)				
<14% <sup>c</sup>	81 (80)	5 (5)	86 (85)	75	100	95	0.83 (0.68-0.97)
≥14% <sup>c</sup>	0 (0)	15 (15)	15 (15)				
Total <sup>c</sup>	81 (80)	20 (20)	101 (100)				
0.6 mm							
<10% <sup>b</sup>	45 (42)	5 (5)	50 (47)	87	66	74	0.49 (0.33-0.64)
≥10% <sup>b</sup>	23 (21)	34 (32)	57 (53)				
Total <sup>b</sup>	68 (64)	39 (36)	107 (100)				
<10% <sup>c</sup>	60 (56)	12 (11)	72 (67)	69	88	81	0.59 (0.43-0.75)
≥10% <sup>c</sup>	8 (7)	27 (25)	35 (33)				
Total <sup>c</sup>	68 (64)	39 (36)	107 (100)				
1.0 mm							
<10% <sup>b</sup>	51 (50)	10 (10)	61 (60)	74	81	78	0.54 (0.37-0.71)
≥10% <sup>b</sup>	12 (12)	28 (28)	40 (40)				
Total <sup>b</sup>	63 (62)	38 (38)	101 (100)				
<10% <sup>c</sup>	57 (56)	13 (13)	70 (69)	66	90	81	0.58 (0.42-0.75)
≥10% <sup>c</sup>	6 (6)	25 (25)	31 (31)				
Total <sup>c</sup>	63 (62)	38 (38)	101 (100)				

CI, confidence interval; TMA, tissue microarray.

<sup>a</sup>At least one matching core is required to be included in the comparison.

<sup>b</sup>Comparison is made using mitotically active approach.

<sup>c</sup>Comparison is made using weighted approach.



**Figure 2** Bar plot showing the difference between using the weighted approach and the mitotically active approach. Note that the accuracy is better for the weighted approach vs the mitotically active approach regardless of the cutoff (14% or 10%) used or the size of cores (0.6 mm vs 1.0 mm), except for the 1.0-mm core when a 14% cutoff is used, where both approaches have similar accuracy rates (95%, third and fourth bars from the left).

**Table 3**

**Impact of Tumor Sampling on Ki-67 Discordance Between Tissue Microarray Cores Including All Cases With at Least Two Cores Using 14 and 10% Cutoffs**

Characteristic	0.6-mm Cores				1.0-mm Cores			
	No. (%)	Conc (n = 64)	Disc (n = 32)	P Value	No. (%)	Conc (n = 69)	Disc (n = 8)	P Value
14% cutoff								
Cellularity, median (IQR)	107 (100)	3,350 (2,752-4,254)	3,789 (3,156-4,358)	.085	101 (100)	7,540 (5,945-9,479)	9,675 (7,809-10,454)	.128
Core number, No. (%)								
1	11 (10)	NA	NA		24 (24)	NA	NA	
2	28 (26)	23 (82)	5 (18)	.039	77 (76)	69 (90)	8 (10)	NA
3	68 (64)	41 (60)	27 (40)					
10% cutoff								
Cellularity, median (IQR)	107 (100)	3,475 (2,828-4,263)	3,632 (2,760-4,346)	.682	101 (100)	7,540 (6,022-9,744)	8,644 (6,156-10,101)	.24
Core number, No. (%)								
1	11 (10)	NA	NA		24 (24)	NA	NA	
2	28 (26)	22 (79)	6 (21)	.005	77 (76)	61 (79)	16 (21)	NA
3	68 (64)	32 (47)	36 (52)					

Conc, concordant Ki-67 status between all cores for a given case; Disc, discordant Ki-67 status between any cores for a given case; IQR, interquartile range; NA, not applicable.

concurrent invasive carcinoma in the FSs in 16.4% of the cases (Supplemental Figures 1 and 2). However, when DCIS scores were subtracted from the FS scores, it had no effect on the concordance between the TMAs and FSs. We conclude that these two variables are not confounding factors affecting the concordance rate between FSs and TMAs when classifying luminal tumors.

It is controversial to use TMAs instead of FSs, particularly in clinical trials and epidemiologic studies, to evaluate Ki-67 scores.<sup>5</sup> The issues related to using TMAs are how representative the TMA cores are of the FSs and what is the best way to design the study in terms of the core size and number of cores, as well as the best way of core-to-case collapsing when more than a single core is included in the TMA. For that, we chose three 0.6-mm cores and two 1.0-mm cores and compared their Ki-67 scores with the FS. When more than a single core was included in the analysis, we performed two methods of core-to-case collapsing. The first method, the weighted approach, accounts for the cellularity of the core, whereas the second method, the mitotically active approach, accounts for the degree of positivity in every core separately regardless of the number of cells in each core.

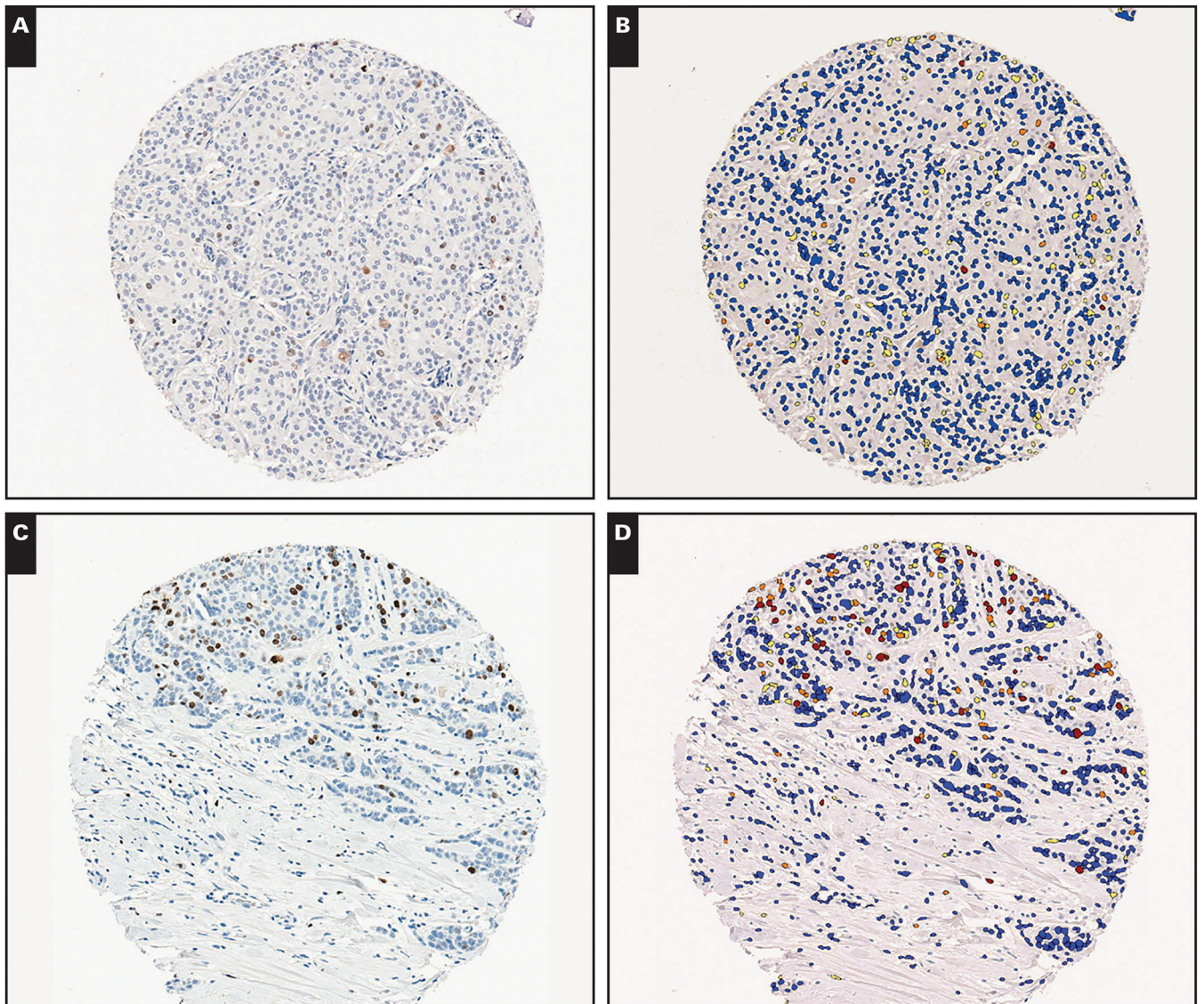
We found significant tumor heterogeneity in the invasive carcinoma whereby up to 40% (using a 14% cutoff) and 52% (using a 10% cutoff) of the cases had at least one discordant core in the 0.6-mm core group (Table 3). Two factors could affect the evaluation and interpretation of Ki-67 heterogeneity: the degree of TILs and DCIS. However, as described above, these two factors did not have any effect on the concordance between the TMA scores and FSs. Therefore, we conclude that Ki-67

heterogeneity is real. Tumor heterogeneity, therefore, is believed to be responsible for the higher number of discordant cases in the 0.6-mm core group vs the 1.0-mm core group, as the degree of tumor heterogeneity was higher in the 0.6-mm group. Similarly, tumor heterogeneity was responsible for the higher degree of discordance when a 10% cutoff was used vs 14%, as tumor heterogeneity was higher when a 10% cutoff was used.

When there were two or more cores, the best way of core-to-case collapsing that reflected the FS score was the weighted approach (Image 1). When the mitotically active approach was used for 0.6-mm cores and setting 14% as a cutoff, 15% of the cases had Ki-67 of 14% or more in at least one core and less than 14% in the FS with an accuracy of 85%. This number became 26% using a 10% cutoff with an accuracy of 74%. When cores are small (0.6 mm) and many (up to three) taken from a tumor that has a high level of heterogeneity, the chance of having at least a single discordant core is high. There was no difference in the number of discordant cases (n = 5) or the accuracy rate (95%) for 1.0-mm cores using either approach when a 14% cutoff was used. However, the accuracy rate decreased when a 10% cutoff was used. It is worth noting that using the weighted approach was better than the mitotically active approach (78% vs 81%) (Table 2 and Figure 2). That is due to the increased tumor heterogeneity using a 10% cutoff vs a 14% cutoff (Table 3).

Knutsvik et al<sup>22</sup> compared FS vs TMA (1.0 mm in triplicate) cores in all types of BC. They did not explicitly report the difference between the two samples in the luminal/HER2- subtype (n = 350). However, Ki-67 expression was lower in the TMA cores than in FS. When they used





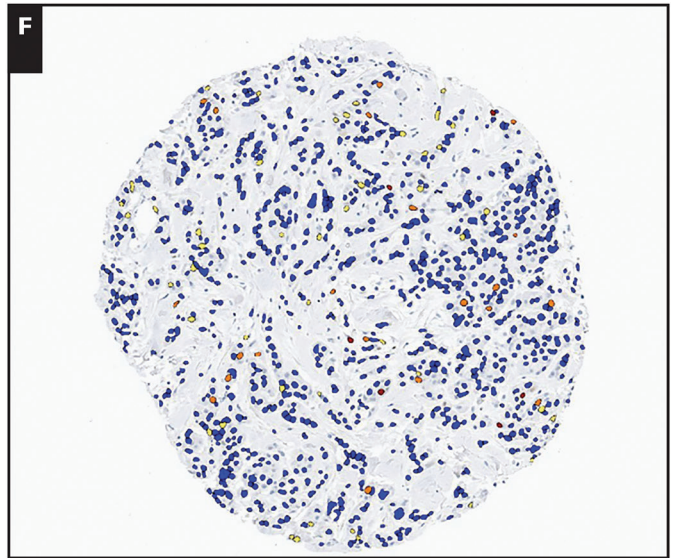
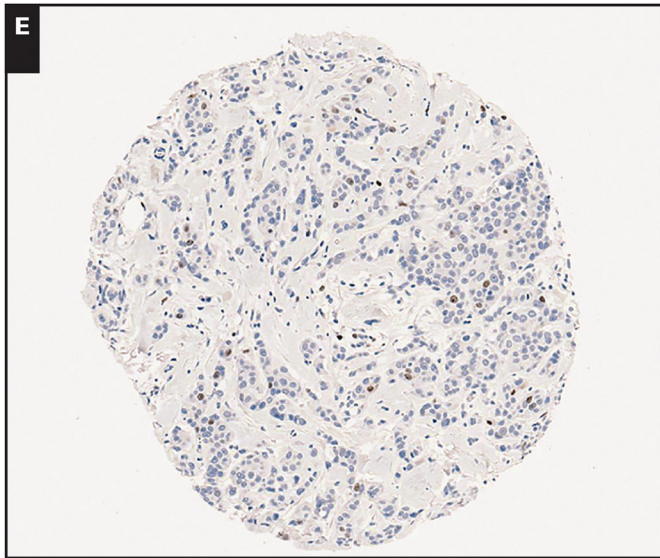
**Image 1** Ki-67 tumor heterogeneity effect on discordance with 0.6-mm cores and the advantage of using a weighted approach. Negative nuclei are shown in blue, low intensity in yellow, medium intensity in orange, and strong intensity in red. **A, B**, Ki-67 = 7.9%. **C, D**, Ki-67 = 14.4%.

a 14% cutoff, 40.9% of luminal B-like cases (Ki-67  $\geq 14\%$ ) were misclassified in the TMA as luminal A compared with the FS, and only 0.3% were misclassified as luminal B. There are a few issues with this study. They manually counted 500 cells by a single pathologist. Although there was minimal interobserver and intraobserver variability in this study as they described, it is known that when Ki-67 is manually counted, the pathologist tends to choose a hot spot area. This would bias the results to higher Ki-67 expression in the FS vs the TMA cores. In our study, we found variability between TMA cores, and FS goes both ways. That is mainly due to using image analysis of FS to count all cells in the TMA core, which prevents selection bias. There were cases where the TMA core had a higher

Ki-67 score than the FS, particularly with single 0.6-mm cores (Tables 1 and 2).

Plancoulaine et al<sup>23</sup> and Besusparis et al<sup>24</sup> studied the impact of sampling on the accuracy of Ki-67 in BC. They included 297 BC cases (189 cases of luminal type) and performed TMA simulation using hexagonal tiling with the size of each hexagon 0.75 mm, equivalent to an approximately 1.0-mm core. They found that to achieve a coefficient error of 10%, five to six cores were needed for homogeneous cases, 11 to 12 cores for heterogeneous cases, and in mixed tumor populations, eight TMA cores were required. It typically is not feasible to acquire at least five 1.0-mm cores from a single tumor for an epidemiologic study. In fact, in our study, we were able to design the





**Image 1** (cont) **E, F**, Ki-67 = 6.9%. Full section (FS) score = 10.2% (low expression). Using the weighted approach, the score was 9.6% (low expression); using the mitotically active approach, the score was 14.4% (high expression). The FS score was less than 14%.

study acquiring three 0.6-mm cores and two 1.0-mm cores when the tumor was larger than 20 mm to leave 50% of the tumor in the block for future studies. Moreover, many of the luminal/HER2- type BC tumors are smaller than 10 mm, making it impossible to acquire a relatively large number of cores. We found that a single 1-mm core could give an accuracy of 95% in this tumor type. Therefore, for an epidemiologic study that uses a high-throughput method, a single 1.0-mm core is practically sufficient.

We conclude that when designing a study that requires high-throughput analysis and TMA technique is elected, three 0.6-mm cores or a single 1.0-mm core is adequate, but two is preferred as the second core acts as a backup in case the first core is lost. If the size of the TMA core is 0.6 mm, it is recommended to use at least three cores. However, tumor heterogeneity might be a limiting factor. With decreasing the value of the cutoff (10% instead of 14%), the number of discordant cases increases and the accuracy rate decreases. Regardless of the cutoff, the weighted approach is a better collapsing method than the mitotically active approach, which minimizes the effect of tumor heterogeneity.

*Corresponding author: Thaer Khoury, MD, Roswell Park Cancer Institute, Elm & Carlton St, Buffalo, NY 14263; thaer.khoury@roswellpark.org.*

*The research conducted by the AMBER Consortium is funded by the following National Institutes of Health and Foundation grants: P01 CA151135, P50 CA58223, and the Breast Cancer Research Foundation.*

## References

1. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747-752.
2. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98:10869-10874.
3. Gerdes J, Lemke H, Baisch H, et al. Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody Ki-67. *J Immunol*. 1984;133:1710-1715.
4. Urruticoechea A, Smith IE, Dowsett M. Proliferation marker Ki-67 in early breast cancer. *J Clin Oncol*. 2005;23:7212-7220.
5. Dowsett M, Nielsen TO, A'Hern R, et al; International Ki-67 in Breast Cancer Working Group. Assessment of Ki67 in breast cancer: recommendations from the international Ki67 in Breast Cancer Working Group. *J Natl Cancer Inst*. 2011;103:1656-1664.
6. Pinhel IF, Macneill FA, Hills MJ, et al. Extreme loss of immunoreactive p-Akt and p-Erk1/2 during routine fixation of primary breast cancer. *Breast Cancer Res*. 2010;12:R76.
7. Bai Y, Tolles J, Cheng H, et al. Quantitative assessment shows loss of antigenic epitopes as a function of pre-analytic variables. *Lab Invest*. 2011;91:1253-1261.
8. Mengel M, von Wasielewski R, Wiese B, et al. Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial. *J Pathol*. 2002;198:292-299.
9. Benini E, Rao S, Daidone MG, et al. Immunoreactivity to mib-1 in breast cancer: methodological assessment and comparison with other proliferation indices. *Cell Prolif*. 1997;30:107-115.
10. Khoury T. Delay to formalin fixation alters morphology and immunohistochemistry for breast carcinoma. *Appl Immunohistochem Mol Morphol*. 2012;20:531-542.

11. Acs G, Kiluk J, Loftus L, et al. Comparison of Oncotype DX and Mammostrat risk estimations and correlations with histologic tumor features in low-grade, estrogen receptor-positive invasive breast carcinomas. *Mod Pathol*. 2013;26:1451-1460.
12. Khoury T, Huang X, Chen X, et al. Comprehensive histologic scoring to maximize the predictability of pathology-generated equation of breast cancer Oncotype DX recurrence score. *Appl Immunohistochem Mol Morphol*. 2016;24:703-711.
13. Harvey JM, Clark GM, Osborne CK, et al. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *J Clin Oncol*. 1999;17:1474-1481.
14. Wolff AC, Hammond ME, Schwartz JN, et al; American Society of Clinical Oncology; College of American Pathologists. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J Clin Oncol*. 2007;25:118-145.
15. Salgado R, Denkert C, Demaria S, et al; International TILs Working Group 2014. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs Working Group 2014. *Ann Oncol*. 2015;26:259-271.
16. Cheang MC, Chia SK, Voduc D, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst*. 2009;101:736-750.
17. Allott EH, Cohen SM, Geradts J, et al. Performance of three-biomarker immunohistochemistry for intrinsic breast cancer subtyping in the AMBER consortium. *Cancer Epidemiol Biomarkers Prev*. 2016;25:470-478.
18. Khoury T, Sait S, Hwang H, et al. Delay to formalin fixation effect on breast biomarkers. *Mod Pathol*. 2009;22:1457-1467.
19. Khoury T, Liu Q, Liu S. Delay to formalin fixation effect on HER2 test in breast cancer by dual-color silver-enhanced in situ hybridization (dual-ISH). *Appl Immunohistochem Mol Morphol*. 2014;22:688-695.
20. Qiu J, Kulkarni S, Chandrasekhar R, et al. Effect of delayed formalin fixation on estrogen and progesterone receptors in breast cancer: a study of three different clones. *Am J Clin Pathol*. 2010;134:813-819.
21. Khoury T, Nagrale V, Opyrchal M, et al. Prognostic significance of stromal versus intratumoral infiltrating lymphocytes in different subtypes of breast cancer treated with cytotoxic neoadjuvant chemotherapy [published online February 9, 2017]. *Appl Immunohistochem Mol Morphol*.
22. Knutsvik G, Stefansson IM, Aziz S, et al. Evaluation of Ki67 expression across distinct categories of breast cancer specimens: a population-based study of matched surgical specimens, core needle biopsies and tissue microarrays. *PLoS One*. 2014;9:e112121.
23. Plancoulaine B, Laurinaviciene A, Herlin P, et al. A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data. *Virchows Archiv* 2015;467:711-722.
24. Besusparis J, Plancoulaine B, Rasmusson A, et al. Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer. *Diagn Pathol*. 2016;11:82.