DOI:10.1093/jnci/djt306
Advance Access publication November 7, 2013

ARTICLE

An International Ki67 Reproducibility Study

Mei-Yin C. Polley, Samuel C. Y. Leung, Lisa M. McShane, Dongxia Gao, Judith C. Hugh, Mauro G. Mastropasqua, Giuseppe Viale, Lila A. Zabaglo, Frédérique Penault-Llorca, John M.S. Bartlett, Allen M. Gown, W. Fraser Symmans, Tammy Piper, Erika Mehl, Rebecca A. Enos, Daniel F. Hayes, Mitch Dowsett, Torsten O. Nielsen, on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group and North American Breast Cancer Group

Manuscript received April 2, 2013; revised September 3, 2013; accepted September 16, 2013.

Correspondence to: Torsten Nielsen, MD, PhD, FRCPC, University of British Columbia Pathology and Laboratory Medicine, Anatomical Pathology, JP 1401, Vancouver Hospital & Health Sciences Centre, 855 W 12th Ave, Vancouver, BC V5Z 1M9, Canada (e-mail: torsten@mail.ubc.ca).

Background In breast cancer, immunohistochemical assessment of proliferation using the marker Ki67 has potential use in both research and clinical management. However, lack of consistency across laboratories has limited Ki67's value. A working group was assembled to devise a strategy to harmonize Ki67 analysis and increase scoring concordance. Toward that goal, we conducted a Ki67 reproducibility study.

Methods Eight laboratories received 100 breast cancer cases arranged into 1-mm core tissue microarrays—one set stained by the participating laboratory and one set stained by the central laboratory, both using antibody MIB-1. Each laboratory scored Ki67 as percentage of positively stained invasive tumor cells using its own method. Six laboratories repeated scoring of 50 locally stained cases on 3 different days. Sources of variation were analyzed using random effects models with log2-transformed measurements. Reproducibility was quantified by intraclass correlation coefficient (ICC), and the approximate two-sided 95% confidence intervals (CIs) for the true intraclass correlation coefficients in these experiments were provided.

Results Intralaboratory reproducibility was high (ICC = 0.94; 95% CI = 0.93 to 0.97). Interlaboratory reproducibility was only moderate (central staining: ICC = 0.71, 95% CI = 0.47 to 0.78; local staining: ICC = 0.59, 95% CI = 0.37 to 0.68). Geometric mean of Ki67 values for each laboratory across the 100 cases ranged 7.1% to 23.9% with central staining and 6.1% to 30.1% with local staining. Factors contributing to interlaboratory discordance included tumor region selection, counting method, and subjective assessment of staining positivity. Formal counting methods gave more consistent results than visual estimation.

Conclusions Substantial variability in Ki67 scoring was observed among some of the world's most experienced laboratories. Ki67 values and cutoffs for clinical decision-making cannot be transferred between laboratories without standardizing scoring methodology because analytical validity is limited.

J Natl Cancer Inst;2013;105:1897–1906

Uncontrolled proliferation is a key feature of malignancy. The nuclear proliferation marker Ki67 is of interest for various potential uses in the clinical management of breast cancer (eg, prognosis, prediction, and monitoring of response) (1–9). The most commonly used assay to assess Ki67 is immunohistochemical (IHC) staining with the MIB-1 antibody. However, interlaboratory methodology is inconsistent, and, despite the apparent prognostic utility of Ki67, routine use of this tumor biomarker has not been widely recommended by consensus guidelines panels such as that convened by the American Society of Clinical Oncology, mainly because of concerns regarding analytical validity (10).

With the goal of harmonizing Ki67 analytical methodology, Dowsett et al., on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group and North American Breast Cancer Group, provided an overview of the current state of the art of Ki67 evaluation and proposed a set of guidelines for analysis and reporting of Ki67 (1). Although those guidelines aimed to reduce preanalytical and analytical variations, the Working Group recognized that actual scoring procedures varied substantially, contributing to a lack of consensus regarding optimal cutoffs that should be applied in various research and clinical decision-making settings. This lack of consistency has prevented direct comparisons of Ki67 across laboratories and clinical trials.

In an effort to harmonize Ki67 analysis, the Working Group studied intra- and interlaboratory reproducibility of IHC assays for Ki67 in breast cancer among a group of highly experienced pathology laboratories. A secondary aim was to identify key sources of variation, particularly those introduced by different scoring methodologies.

Methods

One hundred breast cancer cases were arranged into 1-mm single core tissue microarrays (TMAs), with 50 cases represented on each of two TMA blocks. Specimens were representative clinical cases of invasive breast carcinomas diagnosed in British Columbia during 2009 and 2010, 50 from an academic teaching hospital and 50 from a community hospital. Cases were centrally reviewed; 79% were estrogen receptor positive, and Nottingham grade (11) distribution was 32% grade 1, 44% grade 2, and 24% grade 3. The study was approved by the BC Cancer Agency Clinical Research Ethics Board (protocol H10-03420). Cases were anonymized (treatment, outcome, or follow-up are not part of this study), and the requirement for informed consent was waived.

Eight laboratories from North America and Europe participated. Each laboratory director has a track record of publishing one or more peer-reviewed papers regarding the clinical utility of Ki67. Three experiments were conducted: one examining intralaboratory variability (Experiment 1) and two examining interlaboratory variability (Experiment 2, parts A and B).

Six laboratories participated in the intralaboratory reproducibility experiment (Experiment 1). Each laboratory used its own local protocol to stain one section from a 50-case TMA block, and then the laboratory scored Ki67 on this slide, using its own standard scoring method, on 3 separate days.

Eight laboratories participated in the interlaboratory reproducibility experiments (Experiment 2). Each laboratory received two sets of TMA sections, each set containing the same 100 cores: one centrally stained for Ki67 (Experiment 2A), the second stained by each laboratory following its own local protocol within 2 weeks of cutting (Experiment 2B).

Thus, Experiment 2A assessed interlaboratory reproducibility of Ki67 on centrally stained slides, eliminating variability of staining method (although this experiment was still subject to biological heterogeneity between serial sections cut from the same TMA). Experiment 2B assessed interlaboratory reproducibility of Ki67 when both local staining and local scoring methods were used (locally stained sections from Experiment 1 formed a subset of those included in Experiment 2B).

Details of each lab's staining methodology are provided in Table 1. Central staining used the MIB-1 clone (mouse monoclonal antibody; Dako, Carpinteria, California). All labs also used MIB-1 from Dako for local staining, but dilution, incubation, and antigen retrieval and detection systems varied, as did hematoxylin counterstain supplier and time.

Ki67 was scored as the percentage of invasive tumor cells positively stained.

Statistical Analysis

Ki67 data were visualized using boxplots and dot plots. Pairwise intralaboratory and interlaboratory concordance were plotted using Bland–Altman plots, which graph the difference in Ki67 between any two paired observations against their mean (12). If there is high agreement in Ki67, the differences are expected to be centered about zero, with a small standard deviation. Crossed random effects models were fitted to formally quantify the amount of variability in Ki67 measurements contributed by each source of variation in the three experiments (13). Details are provided in the Supplementary Methods (available online). Because these random effects models rely on the data being normally distributed with constant variance, Ki67 data were log2-transformed to approximate a normal distribution and stabilize the variance (1). Specifically, a value of 0.1% was first added, and then a log base 2 transformation was applied to all observations. For example, for a Ki67 score of 30%, the recorded transformed value would be log2(30.1).

The intraclass correlation coefficient (ICC) was used as a summary measure of reproducibility. The ICC has a range of 0 to 1, with 1 denoting highest agreement. For Experiment 1, we estimated the ICC for repeated Ki67 measurements made on the same patient by the same laboratory (Supplementary Methods, available online). A credible interval of the ICC was obtained using Markov Chain Monte Carlo routines for fitting generalized linear mixed models, using the MCMCglmm package in R (14). For Experiments 2A and 2B, we estimated the ICC for Ki67 measurements made on the same patient by different laboratories (Supplementary Methods, available online). The approximate two-sided 95% confidence intervals (CIs) for the true ICCs in these experiments were computed using a closed-form formula (15). This ICC has a particularly intuitive interpretation as the proportion of the total variance in the measurements across patients and labs attributable to the true biological variability between patients.

All data analyses were performed using the R language version 2.11.1 (16). The crossed random effects models described in Model 1 and Model 2 (Supplementary Methods, available online) were fitted using the lme4 package (17).

Results

Intralaboratory Reproducibility (Experiment 1)

Figure 1 presents Bland–Altman plots for each possible pairing of Ki67 scores (on the original scale as percentage of positively stained tumor cells) measured on different days by the same laboratory. Four laboratories (Laboratories C, D, E, and H) exhibited the highest degree of internal consistency. Two of these laboratories (Laboratories E and H) used formal point (individual cell) counting methods, whereas Laboratories C and D used visual estimation. Laboratory C was clearly estimating in increments of 5%, at estimates of 5% or greater, and this rounding may have contributed to a heightened impression of internal consistency. For further analyses, we considered Laboratories D, E, and H to have best demonstrated internal consistency. Supplementary Table 1 (available online) provides summary statistics for log2-transformed Ki67 scores by laboratory and day.

The estimates of the variance terms (ν_1, ν_2, ν_3, θ) in Model 1 (Supplementary Methods, available online) were 1.91, 0.29, 0.39, and 0.14, respectively. The largest source of the total variation appears to have been patient biology (as expected, reflecting biological differences in Ki67 levels among different tumors), followed by the interaction between patient and laboratory effects, laboratory, and residual errors. The ICC estimate was 0.94 ([1.91 + 0.29 + 0.39] / [1.91 + 0.29 + 0.39 + 0.14] = 0.94; 95% CI = 0.93 to 0.97). Therefore, the intralaboratory reproducibility was very high, indicating that the laboratories could deliver internally consistent results.

Table 1. Local scoring and staining methodologies used by each laboratory*

	Scoring	Ki67 staining (all used MIB-1 antibody, from Dako)				Counterstain	
Lab	Scoring method	Antigen retrieval	MIB-1 dilution	MIB-1 incubation	Detection system	Hematoxylin type and supplier	Hematoxylin staining time
Lab A	Visual estimate 1% increments (<10%), 5% increments (10%–30%), 10% increments (>30%)	High pH in PT Link	1:200	20 min	Dako Flex	Harris Hem Intermedico	45 sec
Lab B	Visual estimate (1% increments)	Pressure cooker in citrate buffer pH 6 for 6 min	1:50	30 min	Dako Envision	Harris acidic, Thermo Fisher	30–60 sec
Lab C	Visual estimate 1% increments (<5%), 5% increments (10%–40%), 10% increments (>40%)	Ventana CC1 short (30 min)	1:100	32 min 37°C	UltraView	Hematoxylin Ventana /Roche reference:760–2021	4 min
Lab D	Visual estimate (1% increments)	Ventana CC1	1:50	32 min	UltraMap	Hematoxylin Gill 1, Ventana	8 min
Lab E	Counting 500 cells using a hemocytometer; any level of stain is positive	EDTA pH 8	1:50	30 min	Dako Envision	Dako product No. S3301	15 min
Lab F	Counting ~200 cells using a keypad Web application	Leica BONDmax TRIS	1:100	15 min	Leica BONDmax DAB	Leica kit	8 min
Lab G	Visual estimate 10% increments (exceptions made for 1% and 95%: 28 cases scored in 1% increments and one case scored as 95%)	pH 9 Tris-EDTA 25 min (TFS pretreatment module)	1:250	20 min	UltraVision	Type: Mayer (75% Mayer, 25% water) Supplier: BBC Biocare product No.: 3580	5 min on Dako Autostainer
Lab H	Counting entire invasive tumor in each core using a 10 × 10 grid in a ×10 eyepiece graticule	Microwave Dako Epitope Retrieval Solution 10 min	1:50	20 min	Dako Real	Haemalum Mayer TCS Biosciences Ltd. catalog No. HS315	1 min

* EDTA = Ethylenediaminetetraacetic acid; PT = PT Link is a pre-treatment system from Dako; TFS = Thermo Fisher Scientific; TRIS = Tris(hydroxymethyl)aminomethane.

Interlaboratory Reproducibility

Experiment 2A (Central Staining, Local Scoring Method). Table 2 provides, for each laboratory, summary statistics of log2-transformed Ki67 scores observed when all laboratories used their own local scoring methods on centrally stained sections. The geometric mean of Ki67 (by taking the antilogs of the means on log2 scale) across the 100 cases ranged from 7.1% (Laboratory G) to 23.9% (Laboratories D and F) (the arithmetic mean ranged from 15.6% to 31.1%). Such a range indicates substantial differences in Ki67 measurement across laboratories on centrally stained slides from the same cases. Figure 2A presents side-by-side boxplots. Figure 3A plots the individual Ki67 measurements made for each of the 100 patients by each laboratory. Some laboratories tended to score Ki67 at several discrete values, whereas others tended to score on a more continuous scale (Figure 3A). In general, Figure 2 and Figure 3 suggest that high variability in Ki67 scoring was detected among the laboratories.

Using statistical Model 2 (Supplementary Methods, available online), we obtained variance estimates to describe biological variation between patients and variation between laboratories. The estimates of these variances (ν_1', ν_2', θ') in Model 2 were 1.78, 0.34, and 0.39, respectively. The largest variation appears to have come from individual patients, followed by residual errors and laboratory. The ICC estimate was 0.71 (1.78 / [1.78 + 0.34 + 0.39] = 0.71;

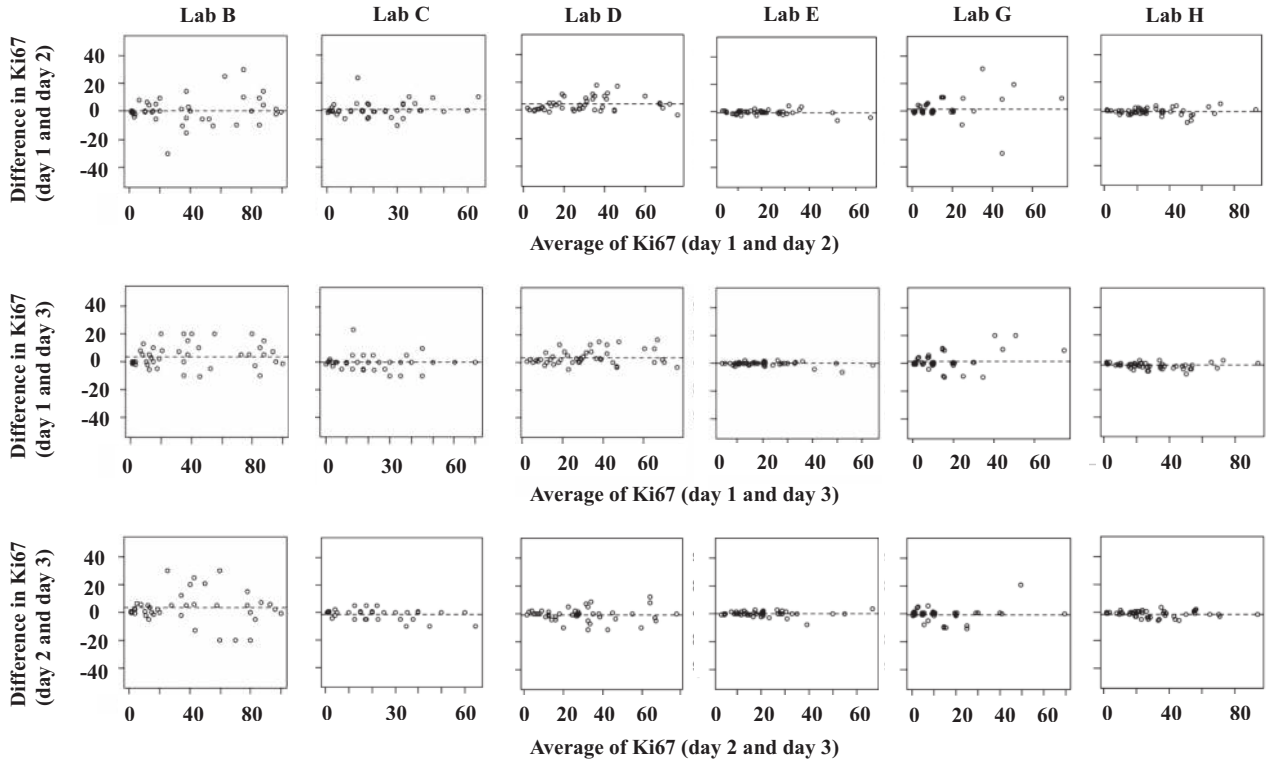

Figure 1. Bland–Altman plots for all pairs of days by laboratory (Experiment 1: same laboratory, different days). The y-axis (difference in Ki67) represents the Ki67 value (percentage of positively stained tumor cells) of the former day minus that of the latter day. The **middle dashed line** represents the average of the differences across all observations. Hence, a middle dashed line greater than 0 would indicate that the average Ki67 value of the former day is greater than that of the latter day, and vice versa. When 2 days have the same Ki67 value, random jittering is used to displace the points vertically to aid visualization. All plots were based on 50 data points, except for those for Laboratory B (n = 46), Laboratory G (n = 49), and Laboratory H (n = 49).

Table 2. Summary statistics of log2-transformed Ki67 measurements by laboratory, Experiment 2A (central staining, local scoring method)*

Laboratory	Min	Q1	Median	Mean	Q3	Max	SD	No. of missing observations
Laboratory A	0.14	3.34	3.92	3.59	4.33	6.65	1.65	0
Laboratory B	0.14	2.35	3.34	3.61	5.18	6.57	1.95	4
Laboratory C	0.14	3.34	4.65	4.44	5.33	6.65	1.39	0
Laboratory D	1.63	3.92	4.81	4.56	5.33	6.62	1.17	1
Laboratory E	0.14	3.34	3.92	3.89	4.59	6.41	1.11	0
Laboratory F	1.19	4.16	4.73	4.58	5.25	6.43	1.03	0
Laboratory G	0.14	0.14	3.34	2.82	4.33	6.57	1.97	1
Laboratory H	0.14	3.57	4.33	4.26	5.18	6.57	1.25	0

* max = maximum; min = minimum; Q1 = first quartile; Q3 = third quartile; SD = standard deviation.

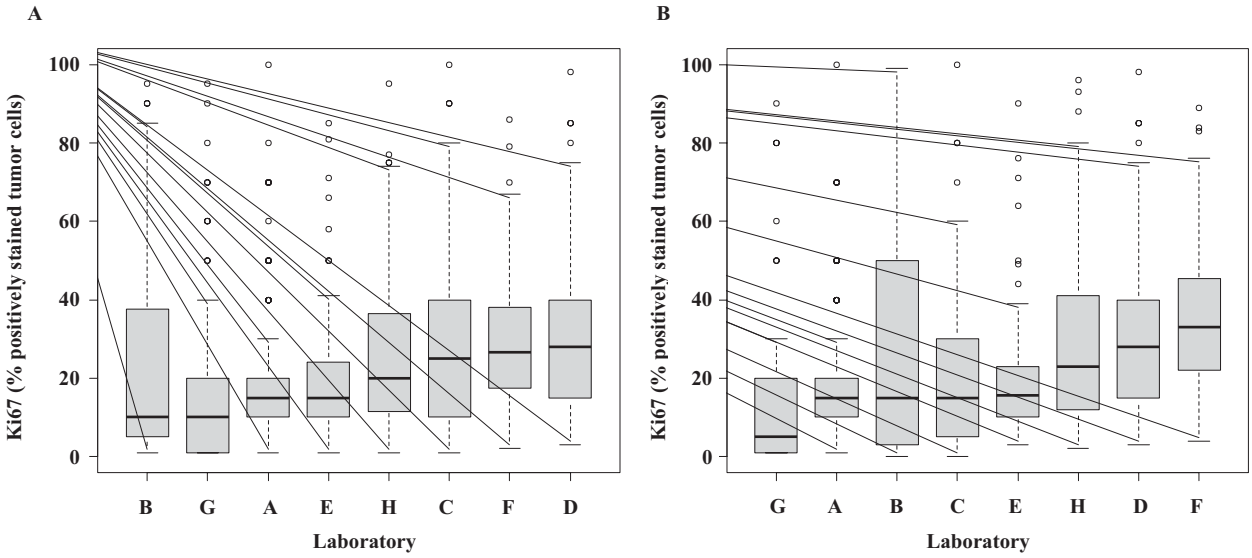

Figure 2. Side-by-side boxplots of Ki67 distribution from Experiments 2A and 2B. **A**) Centrally stained, local scoring method. **B**) Locally stained, local scoring method. Laboratories are reordered according to increasing median Ki67 value. In Experiment 2A, the median of Ki67 ranged from 10% (Laboratories B and G) to 28% (Laboratory D). In Experiment 2B, the median of Ki67 ranged from 5% (Laboratory G) to 33% (Laboratory F). Note that the **bottom and top of the box** in each boxplot represent the first (Q1) and third (Q3) quartiles, and the **bold line** inside the box represents the median of the distribution. The **two bars** outside the box represent the lowest datum still within 1.5 × (Q3 – Q1) of Q1, and the highest datum still within 1.5 × (Q3 – Q1) of Q3 (ie, 1.5 × the interquartile range). Any data not within the two bars are outliers and are represented with **empty circles**.

95% CI = 0.47 to 0.78). Therefore, Ki67 achieved only moderate reproducibility across the laboratories when they used their own scoring methodology on sections stained in a central laboratory.

Experiment 2B (Local Staining, Local Scoring Method). Table 3 provides, for each laboratory, the summary statistics of log2-transformed Ki67 scores that were observed when all laboratories used their own local scoring methods on sections they stained according to their own local staining methods. The geometric mean of Ki67 across the 100 cases ranged from 6.1% (Laboratory G) to 30.1% (Laboratory F) (the arithmetic mean ranged from 12.6% to 35.5%), suggesting a large interlaboratory variability in Ki67 measurement.

Figure 2B presents side-by-side boxplots. Figure 3B presents a dot plot. In both figures, laboratories are ordered by increasing median Ki67 values. Again, there appears to be a large interlaboratory variability. Although laboratory order is not identical in Experiments 2A and 2B when laboratories are arranged by increasing median Ki67 values, the general distribution of Ki67 data for individual laboratories seems to be largely preserved between the two experiments (Figure 3).

To quantify the sources of variation in the data, we again fit a two-way crossed random effects model specified in Model 2 (Supplementary Methods, available online). The estimates of the variance terms (ν_1', ν_2', θ') in Model 2 were 1.71, 0.46, and 0.71, respectively. Similar to Experiment 2A, the largest variation was attributed to biological variability between patients, followed by residual errors and laboratory-to-laboratory variability. Using the variance estimates, the ICC was computed as 0.59 (1.71 / [1.71 + 0.46 + 0.71] = 0.59; 95% CI = 0.37 to 0.68). Therefore, allowing each laboratory to stain the slides based on its own local staining protocol (in addition to scoring them based on its own local scoring method) further decreased the interlaboratory reproducibility compared with Experiment 2A.

The three laboratories that demonstrated the highest level of intralaboratory reproducibility in Experiment 1 (Laboratories D, E, and H) also tended to generate Ki67 values most similar to one another (Figure 4). In both experiments, Laboratory D (which used visual estimation) scored slightly higher in terms of median Ki67 index than Laboratory H, followed by Laboratory E (the latter two used formal counting methods).

In an exploratory analysis, we found that interlaboratory variation was lower among laboratories using formal counting approaches (n = 3 labs) compared with those using visual estimation (n = 5 labs). Spaghetti plots graphically demonstrate this lab-to-lab variability per case (Supplementary Figures 1, A and B, and 2, A and B, available online). ICC was also higher among labs using formal counting: 0.82 (95% CI = 0.19 to 0.87) by counting vs 0.72 (95% CI = 0.32 to 0.80) by visual estimation in Experiment 2A, and 0.75 (95% CI = 0.10 to 0.84) by counting vs 0.64 (95% CI = 0.31 to 0.72) by visual estimation in Experiment 2B. ICC confidence intervals are wide in these subgroup analyses because of the limited number of laboratories.

Discussion

In this study, we observed large variation among a group of highly experienced analysts in determination of levels of Ki67, a biomarker that has been incorporated into clinical care by pathology laboratories worldwide. Although intralaboratory reproducibility of Ki67 evaluation by experienced laboratories was generally good, suggesting that analytical validity may be achievable, interlaboratory reproducibility was only moderate and was even worse when both staining and scoring were done locally. These results support

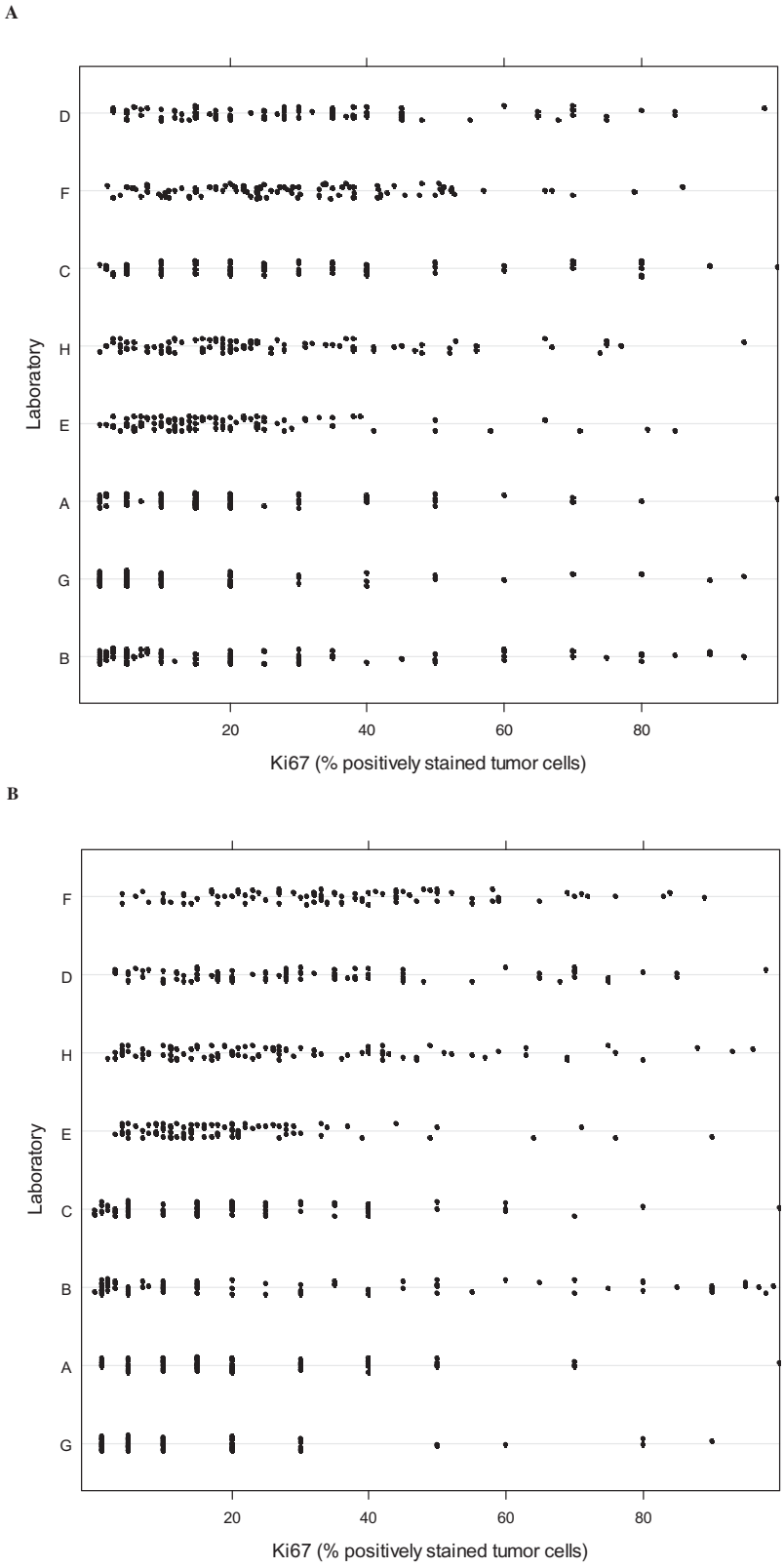

Figure 3. Dot plot of Ki67 measurements by laboratory from Experiments 2A and 2B. **A**) Centrally stained, local scoring method. **B**) Locally stained, local scoring method. Laboratories are reordered according to increasing median Ki67 value (percentange of positively stained tumor cells). When multiple observations have the same Ki67 value, random jittering is used to displace the points vertically to aid visualization.

Table 3. Summary statistics of log2-transformed Ki67 measurements by laboratory, Experiment 2B (local staining, local scoring method)*

Laboratory	Min	Q1	Median	Mean	Q3	Max	SD	No. of missing observations
Laboratory A	0.14	3.34	3.92	3.74	4.33	6.65	1.42	1
Laboratory B	−3.32	1.63	3.92	3.69	5.65	6.63	2.29	9
Laboratory C	−3.32	2.35	3.92	3.54	4.85	6.65	2.08	2
Laboratory D	1.63	3.92	4.81	4.56	5.33	6.62	1.17	1
Laboratory E	1.63	3.34	3.96	3.94	4.51	6.49	1.01	2
Laboratory F	2.04	4.47	5.05	4.91	5.51	6.48	0.92	1
Laboratory G	0.14	0.14	2.35	2.6	4.33	6.49	1.85	2
Laboratory H	1.07	3.6	4.53	4.4	5.36	6.59	1.25	1

* max = maximum; min = minimum; Q1 = first quartile; Q3 = third quartile; SD = standard deviation.

the conservative decisions of evidence-based guidelines bodies regarding routine use of breast cancer biomarkers to make clinical decisions (10,18).

Ki67 levels might be used to determine prognosis or residual risk after primary therapy, to predict activity of systemic therapies, or to monitor patients for sustained response or resistance to delivered therapies (1). However, according to terminology suggested by the Evaluation of Genomic Applications in Practice and Prevention initiative (19), an assay cannot have clinical utility for any of these uses unless its analytical validity has been demonstrated. Our results suggest that even among some of the world's experts in IHC staining and evaluation of Ki67, the analytical validity for this assay is unacceptably poor. Unless an individual pathology laboratory has demonstrated that its staining and scoring methodology, including cutoff determination, meet the highest level of evidence for clinical utility (20), clinicians should use Ki67 results with great caution.

Although interlaboratory differences in staining methodology contributed to Ki67 variability, we also observed a striking heterogeneity in scoring interpretation of centrally stained slides, even using a TMA platform that reduces concern about selection of tissue areas for reading. We did not assess Ki67 in core biopsies or whole sections, diagnostic formats that add complexities regarding which area to score and how to handle "hot spots" of proliferation. If high levels of interobserver concordance cannot be achieved with TMAs, it is even less likely they would be achievable using standard clinical diagnostic formats. Indeed, in a recently reported study in which breast cancer whole sections were distributed to 15 pathologists, Varga et al. observed problematically high interobserver variability among cases in the Ki67 midrange (Ki67 index of 8% to 15%), precisely the region in which most cutoffs are located for making clinical decisions (21). In their study, no single factor (counting method, threshold for positivity, area chosen to score, or staining methodology) explained these differences. There is strong evidence, however, that future approaches using agreed-upon consensus guidelines may improve observer variability and assist standardization (22), an issue that we will address in the next phase of our studies.

Even within the TMA format in our study, several other sources of variability contributed to poor interlaboratory agreement, including whether the laboratory used formal counting of nuclei vs visual estimation, unavoidably subjective assessments of which nuclei represent invasive cancer cells, and what threshold to use for "positive" staining. Although the contributions from these factors are not rigorously separable in our data, the data distributions do suggest that laboratories using formal counting methods gave more consistent results than those using visual estimation.

Clinical decision-making regarding treatment options in breast cancer often relies on the application of a Ki67 cutoff to classify patients into "Ki67 high" or "Ki67 low" risk groups. Widely varying cutoff values, however, further impede the clinical utility of Ki67 and make it difficult to compare Ki67 data across different studies. Reviews of multiple studies in early breast cancer show that cutoffs ranging from 0% to 28.6% have been used (23,24). The 2011 St. Gallen International Consensus Meeting Conference Panel recommended a cutoff of 13.5% to distinguish between "luminal A" and "luminal B/HER2-negative" subtypes in patients with node-negative invasive breast cancer (2). Our data suggest that even if a common Ki67 cutoff is agreed upon, lack of interlaboratory reproducibility in Ki67 measurements represents a major obstacle to confident use of Ki67 for clinical decisions. For example, if the cutoff of 13.5% were applied to the two laboratories that had substantially discordant Ki67 measurements from our Experiment 2A (central staining, local scoring method), 31 of 96 patients (32.3%) would be classified as "Ki67 high" by Laboratory D but as "Ki67 low" by Laboratory B (Figure 5). Further, when the 13.5% cutoff is applied to all Experiment 2A laboratories, the laboratory-specific percentage of patients who would be classified as luminal A varies widely: 56.0% (Laboratory A), 47.9% (Laboratory B), 74.0% (Laboratory C), 77.8% (Laboratory D), 57.0% (Laboratory E), 81.0% (Laboratory F), 30.3% (Laboratory G), 69.0% (Laboratory H). Application of this or other cutoffs for selecting patients for chemotherapy is inappropriate without rigorous analytical standardization.

Computerized digital image analysis has been suggested as a potential solution to problems of analytical subjectivity and interobserver variability in Ki67 assessment (25–28). We have intentionally limited our assessments to visual methods requiring no special equipment because these could be readily and inexpensively adopted by laboratories around the world. Image analysis methods could be a subject of future studies if visual assessments cannot achieve sufficient analytical validity. Recent studies in the neuroendocrine tumor literature, where Ki67 scoring is part of World Health Organization–recommended grading systems, report that digital image analysis (of this comparatively homogeneous tumor type) performs as well as or better than visual counting, with both superior to visual estimation (29–31).

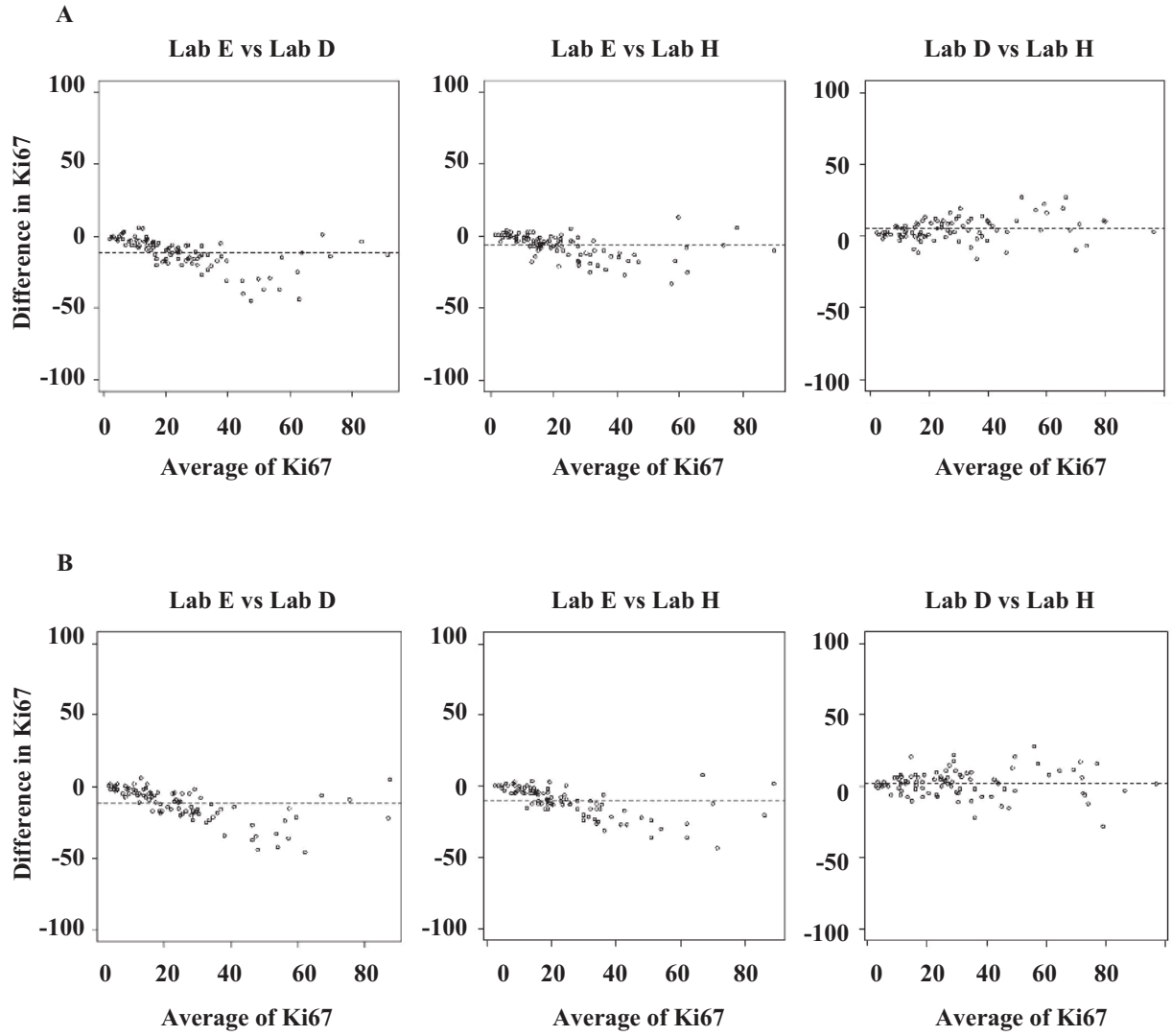

Figure 4. Pairwise Bland–Altman plots showing high interlaboratory reproducibility among three laboratories (Laboratories D, E, and H). **A**) Experiment 2A: central staining, local scoring method. **B**) Experiment 2B: local staining, local scoring method. The y-axis (difference in Ki67) represents the Ki67 value (percentage of positively stained tumor cells) of the former laboratory minus that of the latter laboratory. The **middle dashed line** represents the average of the differences across all observations. Hence, a middle dashed line greater than 0 would indicate that the average Ki67 value of the former laboratory is greater than that of the latter laboratory, and vice versa.

The results from our study emphasize the major differences that exist even among experienced laboratories in results from Ki67 analyses in breast cancer and focus attention on the need for harmonization of scoring procedures if this biomarker's potential is to be realized in breast cancer. We are currently studying whether a prespecified target of success (ICC = 0.9) can be achieved when scorers "train" using a calibration tool designed to mitigate causes of systematic Ki67 scoring differences. Because of widespread use of Ki67 in many research and clinical settings, however, we feel that it is imperative at the present time to report the high interlaboratory variability in Ki67 we observed in this study.

Our study is limited in the sense that it pertains specifically to analytical validity and does not touch on clinical validity. As stated above, we used TMA slides rather than specimen formats typically used in clinical practice (core biopsies or whole sections). Because we used different serial sections of the (same) TMA, some variability in scoring could conceivably be attributable to the section received (although we believe this to be minimal). Although lower than the variability introduced by scoring differences, variability introduced by methodological differences in the IHC staining processes could not be assigned to individual steps within this multistage procedure. Although we succeeded in underscoring an existing problem, we are not yet able to offer a solution. As we note above, we are actively carrying out follow-on studies examining the effect of training labs on a common scoring method, with a view to developing a standardized approach. If these studies are successful, we would extend our approach to core biopsies and whole sections and link findings to patient outcomes to confirm clinical utility.

In summary, although there are multiple potential applications for Ki67 in research and clinical management that are supported by an extensive literature (3,4,23), the clinical utility of Ki67 in breast cancer remains elusive because of analytical concerns. Variability among laboratories in their approaches to scoring is

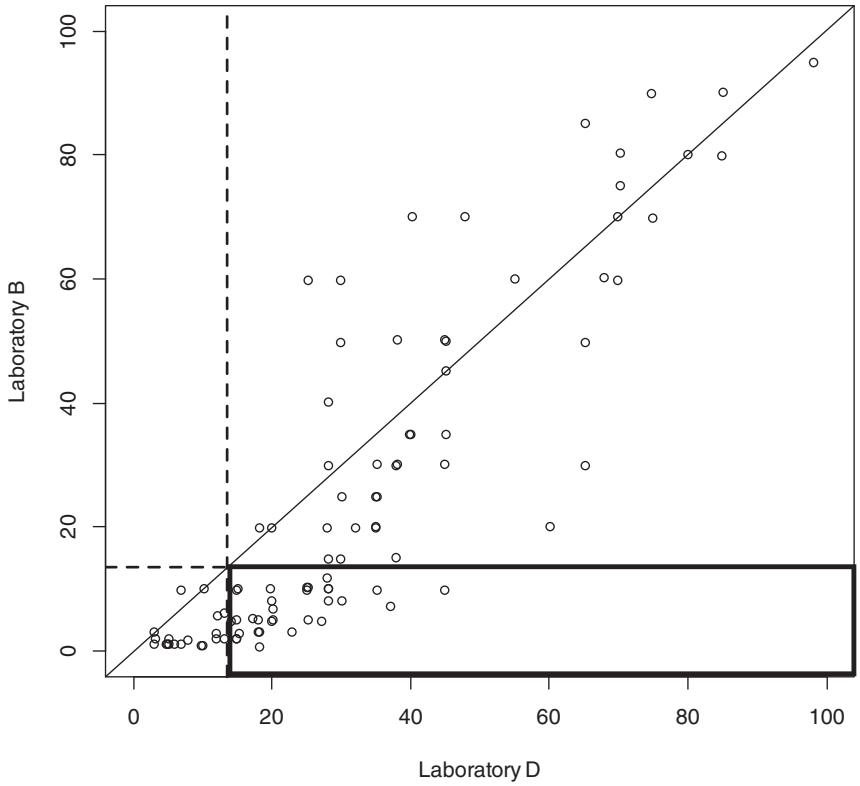

Figure 5. Degree of concordance (**bolded box**) between Laboratory B and Laboratory D at the St. Gallen–recommended Ki67 percentage of positively stained tumor cells cutoff of 13.5%. At a hypothetical 13.5% cutoff, there are 32.3% cases that Laboratory D would call high Ki67 but Laboratory B would call low Ki67. The plot is based on 96 cores (three cores were not scored by one of the labs, and one core was not scored by both labs). When more than one core obtained the same paired Ki67 measurements from the two labs, random jittering is used to displace the points vertically to aid visualization (the small amount of noise added to break ties was generated from a uniform distribution between *-d/5* and *d/5* where *d* is the smallest difference among the original values within each lab).

a major contributor to discordance in results. We recommend that caution be exercised at present when comparing Ki67 results across different laboratories or studies, and we echo the sentiments of the 2007 American Society of Clinical Oncology Tumor Marker Guidelines Committee against using Ki67 in routine clinical practice (10). The following recommendation from our 2011 policy paper therefore still holds for assessment of Ki67 index in breast cancer, when performed by visual assessment of glass slides: "Cut points for prognosis, prediction, and monitoring should only be applied if the results from local practice have been validated against those in studies that have defined the cutoff for the intended use of the Ki67 result" (1).

References

1. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst*. 2011;103(22):1656–1664.
2. Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thurlimann B, Senn HJ. Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol*. 2011;22(8):1736–1747.
3. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol*. 2010;11(2):174–183.
4. Viale G, Giobbie-Hurder A, Regan MM, et al. Prognostic and predictive value of centrally reviewed Ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: results from Breast International Group Trial 1–98 comparing adjuvant tamoxifen with letrozole. *J Clin Oncol*. 2008;26(34):5569–5575.
5. Dowsett M, Ebbs SR, Dixon JM, et al. Biomarker changes during neoadjuvant anastrozole, tamoxifen, or the combination: influence of hormonal status and HER-2 in breast cancer—a study from the IMPACT trialists. *J Clin Oncol*. 2005;23(11):2477–2492.
6. Ellis MJ, Suman VJ, Hoog J, et al. Randomized phase II neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: clinical and biomarker outcomes and predictive value of the baseline PAM50-based intrinsic subtype—ACOSOG Z1031. *J Clin Oncol*. 2011;29(17):2342–2349.
7. Freedman OC, Amir E, Hanna W, et al. A randomized trial exploring the biomarker effects of neoadjuvant sequential treatment with exemestane and anastrozole in post-menopausal women with hormone receptor-positive breast cancer. *Breast Cancer Res Treat*. 2010;119(1):155–161.
8. Murray J, Young OE, Renshaw L, et al. A randomised study of the effects of letrozole and anastrozole on oestrogen receptor positive breast cancers in postmenopausal women. *Breast Cancer Res Treat*. 2009;114(3):495–501.
9. Smith IE, Walsh G, Skene A, et al. A phase II placebo-controlled trial of neoadjuvant anastrozole alone or with gefitinib in early breast cancer. *J Clin Oncol*. 2007;25(25):3816–3822.
10. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol*. 2007;25(33):5287–5312.
11. Galea MH, Blamey RW, Elston CE, et al. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat*. 1992;22(3):207–219.
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
13. Rabe-Hesketh S, Skrondal A. *Multilevel and longitudinal modeling using Stata*. 2nd ed. College Station, TX: Stata Press; 2008.

14. Hadfield J. MCMC methods for multi-response generalised linear mixed models: the MCMCglmm R Package (R package version 2–12). *J Stat Softw*. 2010;33(2):1–22.
15. Cappelleri JC, Ting N. A modified large-sample approach to approximate interval estimation for a particular intraclass correlation coefficient. *Stat Med*. 2003;22(11):1861–1877.
16. R Foundation for Statistical Computing. *R: A language and environment for statistical computing*. Vienna, Austria: R Development Core Team; 2011.
17. Bates D, Maechler M, Bolker B. Linear mixed-effects models using S4 classes. R package. Version 0.999375–39. 2011. Available at: http://cran.r-project.org/web/packages/lme4/index.html. Accessed October 25, 2013.
18. Carlson RW, Allred DC, Anderson BO, et al. Invasive breast cancer. *J Natl Compr Canc Netw*. 2011;9(2):136–222.
19. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med*. 2009;11(1):3–14.
20. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*. 2009;101(21):1446–1452.
21. Varga Z, Diebold J, Dommann-Scherrer C, et al. How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists. *PLoS One*. 2012;7(5):e37379.
22. Kirkegaard T, Edwards J, Tovey S, et al. Observer variation in immunohistochemical analysis of protein expression, time for a change? *Histopathology*. 2006;48(7):787–794.
23. Stuart-Harris R, Caldas C, Pinder SE, Pharoah P. Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. *Breast*. 2008;17(4):323–334.
24. Urruticoechea A, Smith IE, Dowsett M. Proliferation marker Ki-67 in early breast cancer. *J Clin Oncol*. 2005;23(28):7212–7220.
25. Fasanella S, Leonardi E, Cantaloni C, et al. Proliferative activity in human breast cancer: Ki-67 automated evaluation and the influence of different Ki-67 equivalent antibodies. *Diagn Pathol*. 2011;6(Suppl 1):S7.
26. Konsti J, Lundin M, Joensuu H, et al. Development and evaluation of a virtual microscopy application for automated assessment of Ki-67 expression in breast cancer. *BMC Clin Pathol*. 2011;25(11):3. doi:10.1186/1472-6890-11-3.
27. Laurinavicius A, Laurinaviciene A, Dasevicius D, et al. Digital image analysis in pathology: benefits and obligation. *Anal Cell Pathol (Amst)*. 2012;35(2):75–78.
28. Gudlaugsson E, Skaland I, Janssen EA, et al. Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology*. 2012;61(6):1134–1144.
29. Adsay V. Ki67 labeling index in neuroendocrine tumors of the gastrointestinal and pancreatobiliary tract: to count or not to count is not the question, but rather how to count. *Am J Surg Pathol*. 2012;36(12):1743–1746.
30. Remes SM, Tuominen VJ, Helin H, Isola J, Arola J. Grading of neuroendocrine tumors with Ki-67 requires high-quality assessment practices. *Am J Surg Pathol*. 2012;36(9):1359–1363.
31. Tang LH, Gonen M, Hedvat C, Modlin IM, Klimstra DS. Objective quantification of the ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: a comparison of digital image analysis with manual methods. *Am J Surg Pathol*. 2012;36(12):1761–1770.

Funding

This work was supported by the Breast Cancer Research Foundation. Additional funding for the UK labs was received from Breakthrough Breast Cancer and the National Institute for Health Research Biomedical Research Centre at the Royal Marsden Hospital. Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario.

Notes

The study sponsors had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

The authors wish to disclose the following: JC Hugh has worked as a consultant for NanoString Technologies. WF Symmans has received travel/accomodations/meeting expenses from San Antonio Breast Cancer Symposium/AACR and planned patents, royalties, and stock/stock options with Nuvera Biosciences for prognostic/predictive genomic signatures. DF Hayes has stock/stock options with InBiomotion and OncImmune; has grants/grants pending or contracts with Veridex/Janssen; and has planned patents relating to detection and characterization of circulating tumor cells. M. Dowsett has worked as a consultant for Genoptix and Nanostring. TO Nielsen has a grant or contract with Breast Cancer Research Foundation; has patents with Bioclassifier LLC related to IP in a gene expression test that is not part of the submitted work; and has worked as a consultant for Nanostring Technologies related to IP in a gene expression test that is not part of the submitted work.

We are grateful for the contributions of Drs Patricia Kandalaft, Inès Raoelfils, Nancy Davidson, Martine Piccart, and Larry Norton, and the Breast Cancer Research Foundation.

Affiliations of authors: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD (MCP, LMM); Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada (SCYL, DG, EM, TON); Department of Laboratory Medicine and Pathology, University of Alberta, Alberta, Canada (JCH); Division of Pathology and Laboratory Medicine, European Institute of Oncology, Milan, Italy (MGM); Division of Pathology and Laboratory Medicine, European Institute of Oncology, and University of Milan, Milan, Italy (GV); Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, United Kingdom (LAZ); Department of Pathology, Centre Jean Perrin, Clermont-Ferrand, and Université d'Auvergne, France (FP-L); Transformative Pathology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada (JMSB); PhenoPath Laboratories, Seattle, WA (AMG); Department of Pathology, MD Anderson Cancer Center, Houston, TX (WFS); Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, United Kingdom (TP); The EMMES Corporation, Rockville, MD (RAE); Breast Oncology Program, University of Michigan Comprehensive Cancer Center, Ann Arbor, MI (DFH); Academic Department of Biochemistry, Royal Marsden Hospital, London, United Kingdom (MD); on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group and North American Breast Cancer Group.