

An international study to increase concordance in Ki67 scoring

Mei-Yin C Polley¹, Samuel CY Leung², Dongxia Gao², Mauro G Mastropasqua³, Lila A Zabaglo⁴, John MS Bartlett⁵, Lisa M McShane¹, Rebecca A Enos⁶, Sunil S Badve⁷, Anita L Bane⁸, Signe Borgquist⁹, Susan Fineberg¹⁰, Ming-Gang Lin¹¹, Allen M Gown¹², Dorte Grabau⁹, Carolina Gutierrez¹³, Judith C Hugh¹⁴, Takuya Moriya¹⁵, Yasuyo Ohi¹⁶, C Kent Osborne¹³, Frédérique M Penault-Llorca¹⁷, Tammy Piper¹⁸, Peggy L Porter¹¹, Takashi Sakatani¹⁹, Roberto Salgado²⁰, Jane Starczynski²¹, Anne-Vibeke Lænkholm²², Giuseppe Viale²³, Mitch Dowsett²⁴, Daniel F Hayes²⁵, Torsten O Nielsen²
on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group and North American Breast Cancer Group (BIG-NABCG)

¹Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland, USA; ²Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada; ³Division of Pathology and Laboratory Medicine, European Institute of Oncology, Milan, Italy; ⁴Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, UK; ⁵Transformative Pathology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada; ⁶The Emmes Corporation, Rockville, Maryland, USA; ⁷Indiana University Simon Cancer Center, Indianapolis, Indiana, USA; ⁸Department of Pathology and Molecular Medicine, Juravinski Hospital and Cancer Centre, McMaster University, Hamilton, Ontario, Canada; ⁹Department of Clinical Sciences, Division of Oncology and Pathology, Lund University, Lund, Sweden; ¹⁰Montefiore Medical Center and the Albert Einstein College of Medicine, Bronx, New York, USA; ¹¹Fred Hutchinson Cancer Research Center, Seattle, Washington, USA; ¹²PhenoPath Laboratories, Seattle, Washington, USA; ¹³Lester and Sue Smith Breast Center and Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA; ¹⁴Department of Laboratory Medicine and Pathology, University of Alberta, Alberta, Canada; ¹⁵Department of Pathology, Kawasaki Medical School, Kurashiki City, Japan; ¹⁶Department of Pathology, Hakuai Sagar Hospital, Matsubara-cho, Kagoshima, Japan; ¹⁷Department of Pathology, Centre Jean Perrin and Université d'Auvergne, Clermont-Ferrand, France; ¹⁸Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, UK; ¹⁹Department of Pathology, Jichi Medical University, Shimotsuke, Tochigi, Japan; ²⁰Institut Jules Bordet, Brussels, Belgium; ²¹Birmingham Heart of England, National Health Service, Birmingham, UK; ²²Slagelse Hospital, Slagelse, Region Sjælland, Denmark; ²³Division of Pathology and Laboratory Medicine, European Institute of Oncology, and University of Milan, Milan, Italy; ²⁴Academic Department of Biochemistry, Royal Marsden Hospital, London, UK and ²⁵Breast Oncology Program, University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan, USA

Although an important biomarker in breast cancer, Ki67 lacks scoring standardization, which has limited its clinical use. Our previous study found variability when laboratories used their own scoring methods on centrally stained tissue microarray slides. In this current study, 16 laboratories from eight countries calibrated to a specific Ki67 scoring method and then scored 50 centrally MIB-1 stained tissue microarray cases. Simple instructions prescribed scoring pattern and staining thresholds for determination of the percentage of stained tumor cells. To calibrate, laboratories scored 18 'training' and 'test' web-based images. Software tracked object selection and scoring. Success for the calibration was prespecified as Root Mean Square Error of scores compared with reference < 0.6 and Maximum Absolute Deviation from reference < 1.0 (log2-transformed data). Prespecified success criteria for tissue microarray scoring required intraclass correlation significantly > 0.70 but aiming for observed intraclass correlation \geq 0.90. Laboratory performance showed non-significant but promising trends of

Correspondence: Dr TO Nielsen, MD/PhD FRCPC, Pathology and Laboratory Medicine, University of British Columbia, Anatomical Pathology, JP 1401, Vancouver Hospital and Health Sciences Centre, 855 W 12th Avenue, Vancouver, British Columbia V5Z 1M9, Canada. E-mail: torsten@mail.ubc.ca

Received 14 October 2014; revised 17 December 2014; accepted 18 December 2014; published online 20 February 2015

improvement through the calibration exercise (mean Root Mean Square Error decreased from 0.6 to 0.4, Maximum Absolute Deviation from 1.6 to 0.9; paired *t*-test: $P = 0.07$ for Root Mean Square Error, 0.06 for Maximum Absolute Deviation). For tissue microarray scoring, the intraclass correlation estimate was 0.94 (95% credible interval: 0.90–0.97), markedly and significantly >0.70 , the prespecified minimum target for success. Some discrepancies persisted, including around clinically relevant cutoffs. After calibrating to a common scoring method via a web-based tool, laboratories can achieve high inter-laboratory reproducibility in Ki67 scoring on centrally stained tissue microarray slides. Although these data are potentially encouraging, suggesting that it may be possible to standardize scoring of Ki67 among pathology laboratories, clinically important discrepancies persist. Before this biomarker could be recommended for clinical use, future research will need to extend this approach to biopsies and whole sections, account for staining variability, and link to outcomes.

Modern Pathology (2015) **28**, 778–786; doi:10.1038/modpathol.2015.38; published online 20 February 2015

The nuclear proliferation marker Ki67 has potential utility in the clinical management of breast cancer.^{1–5} Oncologists frequently consider Ki67 index in making decisions about whether or not to administer chemotherapy. The American Society of Clinical Oncology, however, has refrained from recommending the use of Ki67 in clinical practice in large part due to lack of standardization in Ki67 assessment.⁶ A recent, comprehensive review found that, although Ki67 has clinical validity when scored by central laboratories, analytical standardization is needed.⁷ Despite such reservations, a recent consensus of international breast cancer experts has recommended using Ki67 via immunohistochemistry to guide clinical decisions regarding the use of adjuvant chemotherapy.⁸

In 2009, the International Ki67 in Breast Cancer Working Group of the Breast International Group and North American Breast Cancer Group formed to address the problem of inter-laboratory inconsistency in Ki67 assessment. It proposed guidelines for preanalytical and analytical conditions for Ki67 assessment, as well as for interpretation, scoring, and related data handling.¹ In 2011–2012, the Working Group carried out an international reproducibility study focusing on how much variability among Ki67 scores from different laboratories was contributed by scoring and staining procedures ('Phase 1'). Findings from that study, which included eight highly experienced laboratories applying their own best practices to tissue microarray slides, revealed substantial variation in Ki67 scores among laboratories, on both centrally and locally stained sections (intraclass correlation of 0.71 and 0.59, respectively).⁹

To determine whether standardization of Ki67 scoring methodology could result in high concordance in scores among laboratories, the Working Group in 2012–2013 undertook a second study ('Phase 2'), in which laboratories were asked to calibrate to a specified Ki67 scoring method before scoring centrally stained tissue microarray slides. The findings from this second study are reported here.

Materials and methods

This study was approved by the British Columbia Cancer Agency Clinical Research Ethics Board

(protocol H10-03420). All tissue microarray cases represented clinical cases of invasive breast carcinomas diagnosed in British Columbia during 2009–2010, described previously.⁹

Central staining used the MIB-1 clone, mouse monoclonal antibody (Dako, Carpinteria, California). Ventana CC1 was used for the antigen retrieval (1:50 MIB-1 dilution, 32 min MIB-1 incubation), and UltraMap for detection. Counterstaining was performed using hematoxylin Gill1 (Ventana), 8 min staining time.

Web-Based Calibration Exercise

Laboratories calibrated to a prescribed scoring method using a web-based tool. Nine 'training' and nine 'test' tissue microarray cases representing a wide range of Ki67 scores served as the calibration cases, chosen from one of the 50-core (1 mm) centrally MIB-1-stained breast cancer tissue microarrays used in Phase 1. Images of the cores were uploaded to a website designed by the Genetic Pathology Evaluation Centre at the University of British Columbia.¹⁰

The website interface allowed laboratories to score the web-image cores via computer mouse click on individual cells. The website software tracked which cells were scored, how they were scored (positive or negative), and overall Ki67 score for each core. Images of hematoxylin and eosin-stained sections were also uploaded for reference.

The specific Ki67 scoring procedure prescribed for this study involved counting 250 cells at the top of the tissue microarray core and 250 cells at the bottom, with any level of brown nuclear staining in invasive cancer cells considered 'positive' (Supplementary Information S1). This formal counting method was chosen as it was a simple method that showed high intra-laboratory reproducibility in Phase 1.⁹ Our Phase 1, as well as studies by others, found counting to give more consistent results than visual estimation.^{9,11,12} During the calibration exercise, refinements were made to the scoring instructions. For example, a 'sampler' of images was provided to demonstrate lightest to darkest staining that should be considered 'positive' or 'negative'. The scorers were also reminded not to score ductal

carcinoma *in situ*, with example images provided (Supplementary Information S1).

Two laboratories (Laboratories E and H) yielding the highest internal consistency in Phase 1⁹ served as reference laboratories for the calibration exercise. The average of their scores for each core served as the reference scores. Seventeen non-reference laboratories scored the training and test cases following the instructions described above. A laboratory 'passed' the training or testing if it met preset performance criteria (see Statistical Analyses and Supplementary Information S2). If a non-reference laboratory did not 'pass' the training, it was asked to learn from its discrepancies with reference scoring by comparing its scored images with scored reference images on the website, where the scorer could zoom in or out as needed (Figure 1). Laboratories were allowed multiple attempts on the training set and had to 'pass' the training to proceed to the test cases. In the test set, only one attempt was allowed, and reference images were not shown.

Scoring-on-Glass Experiment

Sixteen (including the two calibration reference laboratories) of the 19 laboratories participated in the 'scoring on glass' study after calibrating to the prescribed method. Eight countries were represented.

The laboratories scored 50 prestained (MIB-1, Dako, Carpinteria, California) 1-mm cores on a tissue microarray glass slide following scoring instructions similar to those used in the calibration exercise (Supplementary Information S3). The 50 cases

resided on the Phase 1 tissue microarray block not used for selecting the calibration cases.

Three Ki67-stained sections of this tissue microarray (from the same ribbon of serial sections) were mounted on glass slides and distributed among the 16 laboratories. Laboratories were assigned to Groups based on which specific tissue microarray section they scored: Group 1 (7 laboratories), Group 2 (4 laboratories), Group 3 (5 laboratories). All laboratories that had participated in our prior Phase 1 study⁹ were placed into Group 1.

Laboratories used a web-based counter application while scoring the glass slide under a microscope. This counter captured each laboratory's Ki67 score per core (including numbers of cells scored positive or negative), time spent, and any comments. Images of hematoxylin and eosin-stained sections from the cores were also made available on the website for reference.

Statistical Analyses

Calculation of Ki67 score. Ki67 score was defined as the percentage of invasive cancer cells positively stained in the examined region. For the statistical analyses, these scores were transformed to a logarithmic scale by adding 0.1% and applying a log base 2 transformation to satisfy model assumptions of normality and constant variance.¹ For example, a Ki67 score of 30% would be transformed to $\log_2(30.1) = 4.912$.

Web-based calibration exercise. Prespecified criteria for evaluating whether a laboratory passed the calibration 'training' and 'testing' reflected deviation

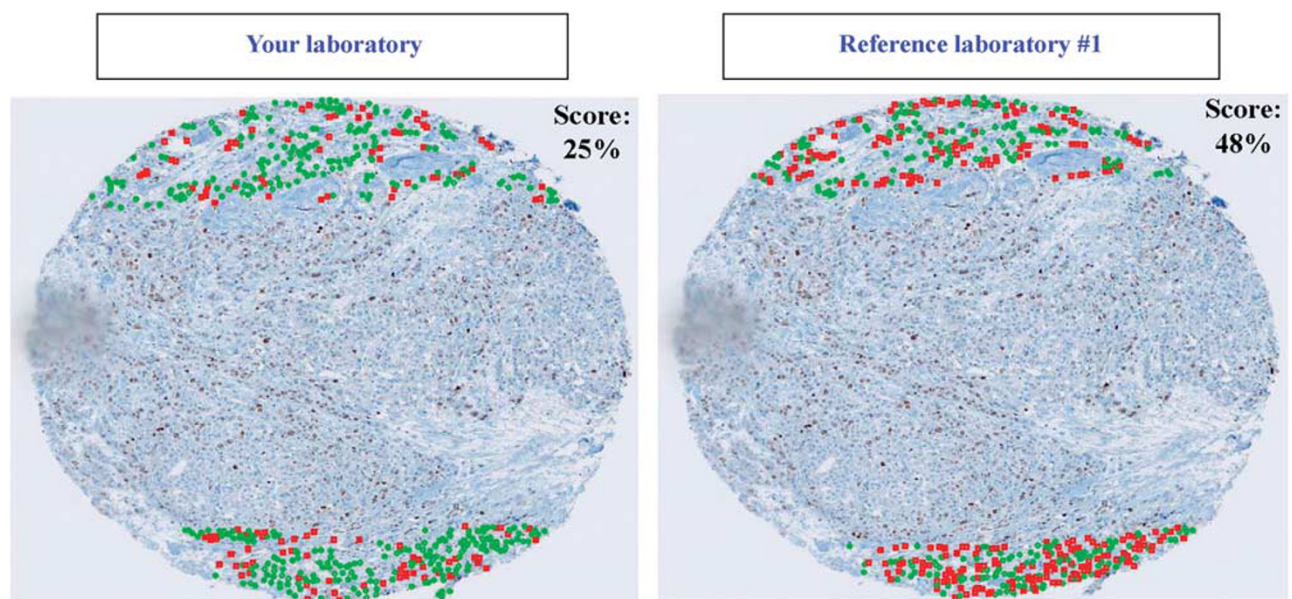


Figure 1 Example of feedback provided to scorer in the calibration exercise. Red = Scorer assessed nucleus as Ki67 positive. Green = Scorer assessed nucleus as Ki67 negative.

of scores from reference scores (Supplementary Information S2):

- Root Mean Square Error < 0.6 (log2-transformed data): Root Mean Square Error in a laboratory's scores compared with reference scores.
- Maximum Absolute Deviation < 1.0 (log 2-transformed data): Maximum Absolute Deviation of the laboratory's scores from the reference scores.

The paired *t*-test was used to compare the performance measures from the training (first attempt) and testing.

Scoring-on-glass experiment. 'Success' for the scoring-on-glass was prespecified as an observed intraclass correlation statistically significantly > 0.70, the approximate intraclass correlation observed in our Phase 1 study,⁹ in which experts used their own scoring methods (see Supplementary Information S4 for discussion of statistical power). An intraclass correlation ranges from 0 to 1, with 1 representing perfect reproducibility.

Among laboratories scoring the exact same tissue microarray section, the intraclass correlation represents the proportion of the total variability attributable to biological variability among the 50 patient tumor spots represented on the tissue microarray section. This intraclass correlation will be referred to as the same-section intraclass correlation.

Comparing between Groups of laboratories scoring different sections of the same tissue microarray, a slightly different intraclass correlation is computed reflecting the proportion of variability attributable to biological differences between patient tumors. In this situation, the same patient tumor is represented on three different tissue microarray sections, which may introduce additional variability, for example, due to intratumoral heterogeneity or assay variation. This second type of intraclass correlation will be referred to as the different-section intraclass correlation and is expected to be less than the same-section intraclass correlation.

The same-section and different-section intraclass correlations were calculated from a variance components model that quantified the contributions from different sources of variability (see Supplementary Information S4 for details). The sources of variability (variance components) considered in the model include patient tumor (ν_{Patient}), spot from the same patient tumor on different tissue microarray sections (ν_{Spot}), scoring laboratory (ν_{Lab}), and residual error (ν_{Error}). Residual error includes within-scorer variation and any other variation not explained by the other sources of variability.

Intraclass correlation estimates were computed as same-section intraclass correlation = $(\nu_{\text{Patient}} + \nu_{\text{Spot}}) / (\nu_{\text{Patient}} + \nu_{\text{Lab}} + \nu_{\text{Spot}} + \nu_{\text{Error}})$ and different-section intraclass correlation = $\nu_{\text{Patient}} / (\nu_{\text{Patient}} + \nu_{\text{Lab}} + \nu_{\text{Spot}} + \nu_{\text{Error}})$, where each of ν_{Patient} , ν_{Lab} , ν_{Spot} , and ν_{Error} are replaced by their estimated values obtained by fitting

the variance components model. Credible intervals for the variance components and these intraclass correlations were obtained using the Markov Chain Monte Carlo routines for fitting generalized linear mixed models (MCMCglmm package in R; see Supplementary Information S4).

Secondary analyses examined concordance among laboratories for dichotomized Ki67 scores ('low' versus 'high') for several cutoffs commonly used in clinical practice or reported in the literature. To obtain a simple statistic for each cutoff, a kappa statistic was computed over all 16 laboratories.¹³

Exploratory analyses compared performance in Phase 1⁹ to performance in Phase 2.

All data analyses were performed using the R language version 2.15.2. Sources of experimental variation in log2-transformed Ki67 scores were analyzed using random effects models as implemented in the R packages lme4 and MCMCglmm. Data were visualized using boxplots and spaghetti plots.

Results

Web-Based Calibration Exercise

Twelve of the 17 non-reference laboratories participating in the calibration exercise met the prespecified passing criteria on their one allowed attempt at scoring the calibration test set (Supplementary Table S1). Failures to pass were all due to divergence from reference scoring on the lowest Ki67 case. Table 1 presents performance measure summary statistics for training (first attempt) and testing. Overall performance of the laboratories through the calibration exercise showed encouraging, but non-significant, trends of improvement from first training attempt to testing: mean Root Mean Square Error across laboratories decreased from 0.6 to 0.4 ($P=0.07$) and mean Maximum Absolute Deviation decreased from 1.6 to 0.9 ($P=0.06$) (paired *t*-test).

Scoring-on-Glass Experiment

Summary statistics for log2-transformed Ki67 scores across the 16 laboratories participating in the scoring-on-glass experiment are given in Table 2. Figure 2 displays the side-by-side boxplots of untransformed Ki67 scores across laboratories by Group. A similar plot in which laboratories are grouped by whether or not they passed the calibration test is provided in Supplementary Figure S1.

The same-section intraclass correlation estimate was 0.94 (95% credible interval: 0.90–0.97), and the different-section intraclass correlation estimate was 0.92 (95% credible interval: 0.88–0.96). Biological variation among different patients was the largest component of the total variation ($\nu_{\text{Patient}}=1.40$, 95% credible interval: 0.91–2.12), followed by residual variation ($\nu_{\text{Error}}=0.05$, 95% credible interval: 0.04–0.05), then laboratory-to-laboratory variation

Table 1 Summary performance statistics across laboratories, calibration exercise

| | Root Mean Square Error (first training attempt) | Root Mean Square Error (test) | Maximum Absolute Deviation (first training attempt) | Maximum Absolute Deviation (test) |
|-----------------|--|----------------------------------|--|--------------------------------------|
| Minimum | 0.205 | 0.14 | 0.349 | 0.288 |
| 25th percentile | 0.309 | 0.286 | 0.656 | 0.552 |
| Median | 0.517 | 0.407 | 1 | 0.9 |
| Mean | 0.634 | 0.393 | 1.569 | 0.87 |
| 75th percentile | 0.959 | 0.531 | 2.63 | 1.189 |
| Maximum | 1.469 | 0.592 | 3.82 | 1.407 |
| S.d. | 0.408 | 0.148 | 1.217 | 0.369 |

Note: Data from 17 laboratories participating in the calibration showed a statistically non-significant downward trend in average Root Mean Square Error (from 0.6 to 0.4, $P=0.07$) and Maximum Absolute Deviation (from 1.6 to 0.9, $P=0.06$) from first training attempt to test phase of the calibration. The prespecified passing value was set at <0.6 for Root Mean Square Error and <1.0 for Maximum Absolute Deviation.

Table 2 Summary statistics of log2-transformed Ki67 scores (percentage of invasive cancer cells scored positive), scoring on glass tissue microarray slides (16 laboratories, 50 cases)

| Laboratory | Minimum | Q1 | Median | Mean | Q3 | Maximum | S.d. | Missing ^a |
|----------------------------|---------|------|--------|------|------|---------|------|----------------------|
| <i>Group 1^b</i> | | | | | | | | |
| G | 0.59 | 3.52 | 4.34 | 4.23 | 5.26 | 6.25 | 1.34 | 1 |
| A | 1.20 | 3.98 | 4.69 | 4.51 | 5.27 | 6.24 | 1.17 | 0 |
| H | 0.93 | 4.29 | 4.96 | 4.70 | 5.52 | 6.24 | 1.16 | 0 |
| E | 0.93 | 3.57 | 4.32 | 4.17 | 4.95 | 5.95 | 1.08 | 0 |
| C | 1.32 | 4.38 | 5.07 | 4.86 | 5.66 | 6.35 | 1.15 | 4 |
| B | 1.43 | 4.02 | 4.81 | 4.57 | 5.32 | 6.10 | 1.04 | 0 |
| D | 0.38 | 3.97 | 4.69 | 4.41 | 5.16 | 6.31 | 1.26 | 0 |
| <i>Group 2^b</i> | | | | | | | | |
| L | 0.14 | 3.94 | 4.54 | 4.42 | 5.28 | 6.33 | 1.28 | 2 |
| M | 0.14 | 4.11 | 4.74 | 4.46 | 5.27 | 6.31 | 1.35 | 0 |
| N | 1.07 | 4.26 | 4.86 | 4.71 | 5.48 | 6.47 | 1.20 | 1 |
| O | -0.15 | 3.80 | 4.57 | 4.29 | 5.21 | 6.30 | 1.38 | 0 |
| <i>Group 3^b</i> | | | | | | | | |
| P | 0.38 | 3.73 | 4.57 | 4.25 | 5.03 | 5.81 | 1.19 | 0 |
| Q | 1.32 | 4.17 | 4.95 | 4.78 | 5.56 | 6.51 | 1.11 | 0 |
| R | 1.07 | 4.34 | 4.97 | 4.67 | 5.42 | 6.42 | 1.20 | 0 |
| S | 0.93 | 3.89 | 4.64 | 4.40 | 5.20 | 6.18 | 1.17 | 0 |
| T | 0.59 | 3.99 | 4.68 | 4.44 | 5.27 | 6.13 | 1.25 | 0 |

^aNumber of cores not scored among the 50 cores. ^bGroups of laboratories were defined by which section of the tissue microarray they received for scoring.

($\nu_{\text{Lab}}=0.04$, 95% credible interval: 0.02–0.09), and finally tissue spot within patient tumor ($\nu_{\text{Spot}}=0.02$, 95% credible interval: 0.01–0.03). The overall mean level of Ki67 (μ) was 4.49 (log2-transformed data).

All intraclass correlations estimated in this study—both same-section and different-section—met the prespecified criteria for success as both 95% credible intervals fell completely above intraclass correlation = 0.70.

Figure 3 displays the variation in scores across laboratories for the 50 cases. Highlighted (gray lines) are the 26 cases for which at least one of the 16 laboratories reported a score in the clinically relevant range of $10\% \leq \text{Ki67} \leq 20\%$. In only one of those 26 cases did all laboratories report scores confined to that range.

In a secondary analysis, six well-known Ki67 cutoffs were applied to the scoring-on-glass data. Inter-laboratory concordance was found to be substantial for all the cutoffs and was best for the $\leq 5\%$ cutoff, where observed agreement was nearly perfect (Table 3).

Exploratory comparison of the different-section intraclass correlation = 0.92 (95% credible interval: 0.88–0.96) estimated in this study to the intraclass correlation = 0.71 (95% credible interval: 0.47–0.78) reported for the centrally stained sections from our prior Phase 1 study⁹ suggests that there was improvement in reproducibility. Although differences in the two studies' designs prevented direct, formal statistical testing for improvement in intraclass correlation between the studies, the intraclass

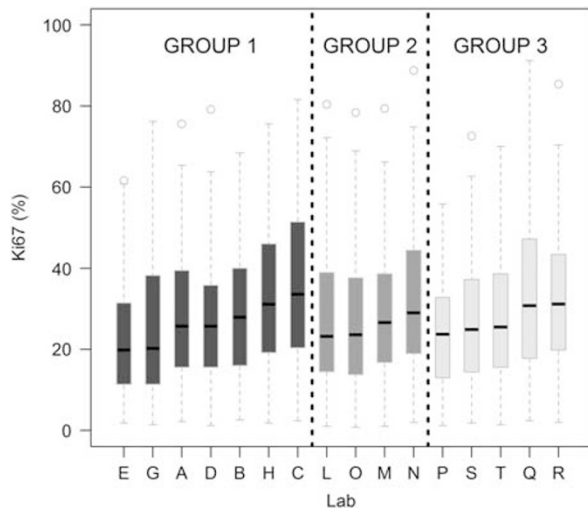


Figure 2 Ki67 scores (percentage of invasive cancer cells scored positive) of all 16 laboratories scoring on glass, by Group (all laboratories within a Group scored the same tissue microarray section). Laboratories have been ordered according to increasing median Ki67 values within each Group. Each Group represents a given tissue microarray section (boxplot shading: black for Group 1, medium gray for Group 2, and light gray for Group 3). The bottom and top of the box in each box plot 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 $1.5 \times (Q3 - Q1)$ of Q1, and the highest datum still within $1.5 \times (Q3 - Q1)$ of Q3 (ie, $1.5 \times$ the inter-quartile range). Any data not within the two bars are outliers and represented with empty circles.

correlation observed in Phase 2 was statistically significantly >0.70 , with the lower bound on the credible interval substantially higher than 0.70. Recalculating the intraclass correlation for Phase 1 using the same statistical approach used in this study, restricted to the 7 laboratories and 50 cases common to both studies, the impression was similar: an intraclass correlation of 0.71 (95% credible interval: 0.54–0.86).

Laboratories spent a median of 5.6 min per core to score the requisite 500 cells (range of laboratory medians 3.0–12.3 min). Scoring all 50 cases required a median of 5.0 h (range 2.9–10.8 h).

Discussion

Substantial interest surrounds use of immunohistochemistry for Ki67 to estimate cellular proliferation rates to guide therapy for patients with breast cancer.^{1–5,7,8} However, for a tumor biomarker assay to be accepted for routine clinical application, it must have both analytical validity as well as high levels of evidence that it has clinical utility for a given use context.¹⁴ Several reports have demonstrated that immunohistochemistry assays for Ki67 have clinical validity for a number of use contexts in early-stage breast cancer, such as prognosis, or as a surrogate end point for response in clinical trials,

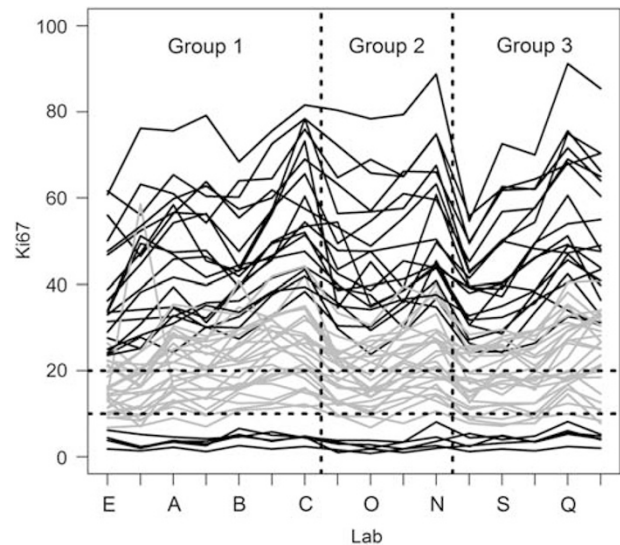


Figure 3 Variability in Ki67 scores (percentage of invasive cancer cells scored positive) across the 16 laboratories for the 50 cases. Each line represents Ki67 scores for the same case. Lighter-colored lines represent data for the 26 cases for which at least one of the 16 laboratories reported a score in the range $10\% \leq \text{Ki67} \leq 20\%$ (one case per line); the darker lines are the remaining cases. Within each Group, laboratories were looking at the exact same tissue microarray section; between Groups, the laboratories were looking at different sections derived from the same tissue microarray block.

Table 3 Kappa values for commonly used Ki67 cutoffs when applied to our study data (16 laboratories scoring tissue microarray glass slides)

| Cutoff | Kappa ^a |
|---|--------------------|
| $\leq 2.7\%$ (low) vs $> 2.7\%$ (high) (relevant to preoperative endocrine prognostic index (PEPI) score = 0) ²¹ | 0.78 |
| $\leq 5\%$ (low) vs $> 5\%$ (high) | 0.91 ^b |
| $\leq 10\%$ (low) vs $> 10\%$ (high) | 0.75 |
| $\leq 13.25\%$ (low) vs $> 13.25\%$ (high) (Cheang <i>et al</i>) ²² | 0.74 |
| $< 14\%$ (low) vs $\geq 14\%$ (high) (St Gallen Consensus ⁸) | 0.73 |
| $\leq 20\%$ (low) vs $> 20\%$ (high) | 0.70 |

^aKappa indicates degree of agreement, incorporating a correction for the extent of agreement expected by chance alone. These analyses combined data over all 16 laboratories (three Groups). ^bObserved agreement was nearly perfect.

and perhaps as a predictive factor for benefit from adjuvant chemotherapy.¹ However, the analytical validity of the Ki67 immunohistochemistry assay in diverse clinical laboratories remains poor.⁷ Thus this tumor biomarker test cannot currently be recommended for routine decision-making to guide treatment decisions in standard contemporary oncology practice.¹⁵

Our previous study (Phase 1) revealed substantial variation in Ki67 scoring when laboratories used their own scoring methods on centrally stained slides from the same tissue microarray block,

indicating that absolute values and cutoffs for Ki67 cannot be transferred between laboratories without careful standardization of scoring methodology.⁹ In this second study, we eliminated variability due to preanalytical staining issues by providing centrally stained tissue microarray slides, thus permitting laboratories to calibrate to a specified scoring method. We found that a standardized, practicable visual scoring method could deliver high inter-laboratory reproducibility in Ki67 scoring on centrally stained tissue microarray slides. The intraclass correlation when all laboratories were scoring the exact same tissue microarray section was 0.94 (95% credible interval: 0.90–0.97), and the intraclass correlation when laboratories were scoring different sections from the same tissue microarray block was 0.92 (95% credible interval: 0.88–0.96). This latter intraclass correlation (0.92) suggests an improvement over intraclass correlation = 0.71 (95% credible interval: 0.54–0.86) calculated in Phase 1 (restricting to the 50 cases and 7 laboratories common to both studies).⁹ As the study was not specifically designed to make this comparison, and a formal statistical test was not performed, it should be interpreted with some caution. Nonetheless, it appears to represent a decrease from 29 to 8% in the proportion of variation in Ki67 scores. The design of this study, using centrally stained slides, suggests that this decrease can be attributed to standardized scoring methodology. If this can be translated into a true clinically applicable context, Ki67 differences between individual patients may well reflect biological differences between tumors, which the assay is intended to detect.

The very light level of staining considered sufficient to be called 'positive' in our scoring instructions was also seen as a potential limitation by scorers. Although they found the provided visual examples helpful in both the calibration and the glass-slide scoring, at the very light end of the spectrum it seemed to require very subjective assessment to deem a cell positive or negative. This 'lightening' of the 'positive' threshold may account for the overall mean of laboratory Ki67 means increasing from 25.7% in Phase 1 to 29.1% in Phase 2 among the Group 1 laboratories, who participated in both studies (Supplementary Table S2).

Finally, although all laboratories were given an opportunity to review their calibration images vs reference images shortly before they scored on glass, different amounts of time passed between completion of the calibration and the subsequent scoring-on-glass among the laboratories (range < 1 month to 7 months). Nonetheless, overall good concordance was still observed; therefore, we believe that some learning effect obtained from the calibration was retained.

This calibration exercise demonstrated that laboratories could be 'trained' to follow a specific scoring pattern. We also learned that visual depictions of the levels of staining intensity considered 'positive' or

'negative' were considered helpful by the scoring pathologists. Nonetheless, clinically important discrepancies persisted among laboratories for some cases in the critical Ki67 range between 10 and 20%, in which the cutoff to distinguish high from low is often selected.^{1,3} There were 26 cases for which at least one of the 16 laboratories reported a score in the range $10\% \leq \text{Ki67} \leq 20\%$. In only one of these 26 cases did all laboratories report scores confined to that range.

Our analyses of concordance after application of cutoffs suggest that the observed degree of inter-laboratory agreement is influenced by the cutoff chosen and argue for caution in dichotomizing Ki67 scores into 'high Ki67' and 'low Ki67'. In fact, as discussed in our earlier guidelines,¹ cutoffs for Ki67 should be context-dependent. Moreover, the establishment of such cutoffs will require additional data to permit examination of associations between Ki67, clinical outcome, and other clinical and pathological factors. As with other continuous variables in pathology, different laboratories use different cutoffs for the same context; but uniform cutoffs can be problematic as well if measurements differ substantially across laboratories. The impact on clinical decision-making varies according to context and, in some circumstances, may merit the reporting of confidence intervals or the definition of equivocal values that fall within a range of uncertainty.

Our study does not have immediate clinical application, as we chose to use a highly standardized specimen format (centrally stained tissue microarray cores) not typically used in clinical practice. However, we conclude that we were able to achieve satisfactory levels of scoring agreement in the highly stylized context in which this study was performed. Having done so, our next logical next step will be to see if similar agreement is possible on core biopsies and if this proves satisfactory proceed to a similar strategy with whole sections, in which the reader will have broader discretion regarding what portion of the tissue should be read. Finally, we plan to return to a local staining methodology strategy, which we expect will re-introduce additional variability into the analyses. Our goal is to standardize each of these points of potential variability among the participants of this set of exercises. If we can maintain the satisfactory concordance we have observed in the current study, we will provide the methodological strategy and protocol to the general pathological community, in a manner similar to what has been accomplished by the American Society of Clinical Oncology and College of American Pathologists joint Committees to establish guidelines for HER2 and hormone receptor analyses.^{16–19}

In conclusion, Ki67 is a widely accessible biomarker of proliferation with strong evidence for clinical validity for prognosis and prediction in breast and other cancers.¹ Its clinical utility has been limited by poor inter-observer reproducibility, with differences in scoring methodology a major contributor to

variability.^{9,20} We demonstrate that improved inter-observer reproducibility in Ki67 scoring can be achieved on centrally stained tissue microarray slides after training observers on a standardized, practicable visual scoring method that can be performed relatively quickly. Our results suggest that strategies can be developed to at least partially overcome inter-observer variation, which has limited the utility of this marker. The potentially encouraging results demonstrated in this study remain tempered by the recognition that achieving agreement among laboratories in clinical practice remains a significant challenge, as these promising findings need to be translated to core biopsies and/or whole sections, with local staining, and clinical validity confirmed. Therefore, we maintain that, unless and until preanalytical and analytical features for immunohistochemistry of Ki67 can be standardized, this assay platform should not be used to drive patient-care decisions in clinical practice. If we are successful in our next set of studies, then immunohistochemistry for Ki67 could be incorporated into standard of care. If not, then evaluation of proliferation for prognostic and predictive considerations in early breast cancer will need to be performed with other assay platforms that may have more analytical validity but which may be more difficult to apply in practices around the world.

Acknowledgments

This work was supported by the Breast Cancer Research Foundation. Additional funding for the UK laboratories 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. Judith Hugh is the Lilian McCullough Chair in Breast Cancer Surgery Research and the CBCF Prairies/NWT Chapter. We are grateful to the Breast International Group and North American Breast Cancer Group (BIG-NABCG) collaboration, including the leadership of Dr Nancy Davidson, Dr Martine Piccart, and Dr Larry Norton, and for the support of the Breast Cancer Research Foundation.

Disclosure/conflict of interest

Mitch Dowsett: Advisory boards for Nanostring and Genoptix. Lecture fees from Ventana. Daniel F Hayes: Research Advisory Panel: Pfizer; Consultancy: Eli Lilly; Sponsored Clinical Research: Janssen Diagnostics; Stock Options: Oncimmune, Inbiomotion; Patents: 3 approved (none related to the submitted work). Torsten Nielsen: Patents: Bioclassifier: Intellectual property in a gene expression test that is not part of the submitted work. Consultancy: Nanostring Technologies: Intellectual

property in a gene expression test that is not part of the submitted work. The other authors declare no conflict of interest.

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;103:1656–1664.
- 2 Goldhirsch A, Wood WC, Coates AS *et al*. 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;22:1736–1747.
- 3 Inwald EC, Klinkhammer-Schalke M, Hofstadter F *et al*. Ki-67 is a prognostic parameter in breast cancer patients: results of a large population-based cohort of a cancer registry. *Breast Cancer Res Treat* 2013;139: 539–552.
- 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: 5569–5575.
- 5 Yerushalmi R, Woods R, Ravdin PM *et al*. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol* 2010;11:174–183.
- 6 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;25:5287–5312.
- 7 Luporsi E, Andre F, Spyrtos F *et al*. Ki-67: level of evidence and methodological considerations for its role in the clinical management of breast cancer: analytical and critical review. *Breast Cancer Res Treat* 2012;132: 895–915.
- 8 Goldhirsch A, Winer EP, Coates AS *et al*. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann Oncol* 2013;24:2206–2223.
- 9 Polley MY, Leung SC, McShane LM *et al*. An international Ki67 reproducibility study. *J Natl Cancer Inst* 2013;105:1897–1906.
- 10 Genetic Pathology Evaluation Centre Ki67-QC calibrator [http://www.gpec.ubc.ca/calibrator].
- 11 Mikami Y, Ueno T, Yoshimura K *et al*. Interobserver concordance of Ki67 labeling index in breast cancer: Japan Breast Cancer Research Group Ki67 Ring Study. *Cancer Sci* 2013;104:1539–1543.
- 12 Tang LH, Gonen M, Hedvat C *et al*. 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;36:1761–1770.
- 13 Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–382.
- 14 Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst* 2009;101:1446–1452.
- 15 Hayes DF. From genome to bedside: are we lost in translation? *Breast* 2013;22 Suppl 2:S22–S26.

- 16 Wolff AC, Hammond ME, Hicks DG *et al*. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J Clin Oncol* 2013;31:3997–4013.
- 17 Wolff AC, Hammond ME, Hicks DG *et al*. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *Arch Pathol Lab Med* 2014;138:241–256.
- 18 Hammond ME, Hayes DF, Dowsett M *et al*. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Arch Pathol Lab Med* 2010;134:907–922.
- 19 Hammond ME, Hayes DF, Dowsett M *et al*. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* 2010;28:2784–2795.
- 20 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:e37379.
- 21 Ellis MJ, Tao Y, Luo J *et al*. Outcome prediction for estrogen receptor-positive breast cancer based on postneoadjuvant endocrine therapy tumor characteristics. *J Natl Cancer Inst* 2008;100:1380–1388.
- 22 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.

Supplementary Information accompanies the paper on Modern Pathology website (<http://www.nature.com/modpathol>)