npj | Breast Cancer

www.nature.com/npjbcancer

ARTICLE OPEN

Analytical validation of a standardized scoring protocol for Ki67: phase 3 of an international multicenter collaboration

Samuel CY Leung[1], Torsten O Nielsen[1], Lila Zabaglo[2], Indu Arun[3], Sunil S Badve[4], Anita L Bane[5], John MS Bartlett[6], Signe Borgquist[7], Martin C Chang[8], Andrew Dodson[2], Rebecca A Enos[9], Susan Fineberg[10], Cornelia M Focke[11], Dongxia Gao[1], Allen M Gown[12], Dorthe Grabau[7], Carolina Gutierrez[13], Judith C Hugh[14], Zuzana Kos[15], Anne-Vibeke Lænkholm[16], Ming-Gang Lin[17], Mauro G Mastropasqua[18], Takuya Moriya[19], Sharon Nofech-Mozes[20], C Kent Osborne[13], Frédérique M Penault-Llorca[21], Tammy Piper[22], Takashi Sakatani[23], Roberto Salgado[24], Jane Starczynski[25], Giuseppe Viale[26], Daniel F Hayes[27], Lisa M McShane[28], Mitch Dowsett[2] on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group and North American Breast Cancer Group (BIG-NABCG)

Pathological analysis of the nuclear proliferation biomarker Ki67 has multiple potential roles in breast and other cancers. However, clinical utility of the immunohistochemical (IHC) assay for Ki67 immunohistochemistry has been hampered by unacceptable between-laboratory analytical variability. The International Ki67 Working Group has conducted a series of studies aiming to decrease this variability and improve the evaluation of Ki67. This study tries to assess whether acceptable performance can be achieved on prestained core-cut biopsies using a standardized scoring method. Sections from 30 primary ER+ breast cancer core biopsies were centrally stained for Ki67 and circulated among 22 laboratories in 11 countries. Each laboratory scored Ki67 using three methods: (1) global (4 fields of 100 cells each); (2) weighted global (same as global but weighted by estimated percentages of total area); and (3) hot-spot (single field of 500 cells). The intraclass correlation coefficient (ICC), a measure of interlaboratory agreement, for the unweighted global method (0.87; 95% credible interval (CI): 0.81–0.93) met the prespecified success criterion for scoring reproducibility, whereas that for the weighted global (0.87; 95% CI: 0.7999–0.93) and hot-spot methods (0.84; 95% CI: 0.77–0.92) marginally failed to do so. The unweighted global assessment of Ki67 IHC analysis on core biopsies met the prespecified criterion of success for scoring reproducibility. A few cases still showed large scoring discrepancies. Establishment of external quality assessment schemes is likely to improve the agreement between laboratories further. Additional evaluations are needed to assess staining variability and clinical validity in appropriate cohorts of samples.

npj Breast Cancer (2016) **2,** 16014; doi:10.1038/npjbcancer.2016.14; published online 18 May 2016

INTRODUCTION

Assessment of the nuclear proliferation biomarker Ki67 has multiple potential roles in breast and other cancers,[1,2] either in standard clinical practice as a prognostic[3–11] and predictive[5,7,10,12] marker or in clinical trials as an eligibility criterion or as a primary end point in early-phase neoadjuvant studies.[10] Perhaps the most critical use for standard clinical care would be to determine prognosis in the context of other factors, such as nodal status, tumor size, and estrogen receptor, progesterone receptor, and HER2 status. Although gene expression multiparameter molecular assays have gained widespread use in the United States and other countries, these assays may not be an option in many clinical settings owing to availability or economic considerations. Therefore, the Ki67 immunohistochemistry assay might offer a cost-effective alternative.[13–15] The 2015 St Gallen consensus panel stated that the majority of new breast cancer cases and breast

[1]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada; [2]Academic Department of Biochemistry, Royal Marsden Hospital and Institute of Cancer Research, London, United Kingdom; [3]Department of Pathology, Tata Medical Center, Kolkata, West Bengal, India; [4]Department of Pathology and Laboratory Medicine, Indiana University Simon Cancer Center, Indianapolis, Indiana, USA; [5]Department of Pathology and Molecular Medicine, Juravinski Hospital and Cancer Centre, McMaster University, Hamilton, Ontario, Canada; [6]Transformative Pathology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada; [7]Department of Clinical Sciences, Division of Oncology and Pathology, Lund University, Lund, Sweden; [8]Department of Pathology, Mount Sinai Hospital, Toronto, Ontario, Canada; [9]The EMMES Corporation, Rockville, Maryland, USA; [10]Department of Pathology, Montefiore Medical Center and the Albert Einstein College of Medicine, Bronx, New York, USA; [11]Department of Pathology, Dietrich-Bonhoeffer Medical Center, Neubrandenburg, Mecklenburg-Vorpommern, Germany; [12]PhenoPath Laboratories, Seattle, Washington, USA; [13]Lester and Sue Smith Breast Center and Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA; [14]Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, Canada; [15]Department of Pathology and Laboratory Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada; [16]Department of Pathology, Slagelse Hospital, Slagelse, Region Sjælland, Denmark; [17]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA; [18]Division of Pathology and Laboratory Medicine, European Institute of Oncology, Milan, Italy; [19]Department of Pathology, Kawasaki Medical School, Kurashiki, Okayama Prefecture, Japan; [20]Department of Laboratory Medicine, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada; [21]Department of Pathology, Centre Jean Perrin and Université d'Auvergne, Clermont-Ferrand, France; [22]Biomarkers & Companion Diagnostics Group, Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, United Kingdom; [23]Department of Pathology, Nippon Medical School, Bunkyo-ku, Tokyo, Japan; [24]Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Brussels, Belgium; [25]Department of Cellular Pathology, Birmingham Heart of England, National Health Service, Birmingham, United Kingdom; [26]Division of Pathology and Laboratory Medicine, European Institute of Oncology and University of Milan, Milan, Italy; [27]Breast Oncology Program, Department of Internal Medicine, University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan, USA and [28]Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland, USA.
Correspondence: M Dowsett (mitch.dowsett@icr.ac.uk)
Received 23 December 2015; revised 22 February 2016; accepted 1 April 2016

npj nature partner journals

BCRF BREAST CANCER RESEARCH FOUNDATION

cancer deaths now occur in less developed regions of the world,[13] accentuating the need for low cost, widely accessible biomarkers. However, despite extensive effort spent on evaluating Ki67 as a prognostic and/or predictive marker in the past three decades, this biomarker is still not completely integrated into clinical decision making,[16] due mainly to the lack of standardization in staining techniques and scoring methods.[3,10,16]

The International Ki67 Working Group has undertaken a systematic multiphase program to determine whether Ki67 scoring can be analytically validated and standardized across laboratories.[10,17,18] In phase 1, variability in visual interpretation, as assessed by the intraclass correlation coefficient (ICC) estimate of interobserver reproducibility, was the most important source of variability (ICC = 0.71, 95% credible interval (CI): 0.47–0.78).[17] In phase 2, substantial levels of agreement were achieved when the various laboratories followed clearly defined, standardized training exercise and scoring methods.[18] Indeed, the interobserver variability observed in phase 2 (ICC = 0.94, 95% CI: 0.90–0.97) is similar to the intraobserver reproducibility (ICC = 0.94; 95% CI: 0.93–0.97) observed in phase 1. However, this level of agreement was achieved when scoring the same tumors on tissue microarrays, whereas in clinical practice biomarker decisions are made on core-cut biopsy or on surgical excision whole-section specimens. Such specimens require pathologists to select specific regions for assessment within a larger area, and so increased variability in scoring would be expected.

Given the encouraging result achieved on breast cancer tissue microarrays, we proceeded to phase 3 to assess whether acceptable performance can be achieved on core-cut biopsies using a similar, standardized method including two distinct approaches in selecting which area to score.

RESULTS

Interlaboratory ICC concordance of Ki67 according to method of scoring

The different-section ICC estimate for the unweighted global score was 0.87 (95%CI: 0.81–0.93), and therefore met the prespecified success criterion (lower bound of credible interval exceeding 0.8; Table 1). The different-section ICCs for the weighted global score and hot-spot score were 0.87 (95%CI: 0.7999–0.93) and 0.84 (95% CI: 0.77–0.92), respectively, and therefore both methods had ICC credible intervals that extended below the success criterion. The corresponding same-section ICC estimates for the unweighted global, weighted global and hot-spot scores were 0.88 (95% CI: 0.81–0.93), 0.87 (95% CI: 0.80–0.93) and 0.84 (95% CI: 0.77–0.92), respectively. Figure 1 displays the side-by-side boxplots of Ki67 scores across laboratories by group. Summary statistics for the Ki67 scores across the 22 laboratories are given in Supplementary Tables 2.

Variance component analyses show that, regardless of scoring method, biological variation among different patients was the largest component of the total variation, indicating that the Ki67 score is reflecting inherent properties of the tumor and that the effect on the score introduced by the immunohistochemistry assay technical variation (sectioning, staining, and scoring) is relatively small (Figure 2, Supplementary Table 5).

Interlaboratory variation of Ki67 scoring

Figure 3 displays the variation in scores across laboratories for each case, in spaghetti plot format. Each line represents scores from one laboratory. Figure 4 presents the scores in a heat map format with the columns (laboratories) sorted (within each group) by the median scores across cases and the rows (cases) sorted by the median scores across laboratories.

Overall it can be seen that most laboratories show good parallelism in the increasing Ki67 scores across the plots. In other words, laboratories measuring higher or lower than others tended to do so relatively consistently. In group 3 one lab (N) can be seen to score considerably higher than the others in both the unweighted and weighted scores particularly in the samples with the higher scores. This laboratory also showed a number of higher scores on the hot-spot method. Another laboratory in group 3 (T) also showed substantially and consistently higher hot-spot scores than the others, whereas in group 1 one laboratory (A) can be seen to score consistently lower than the others.

Table 1. Summary of ICC values for different scoring methods

	Different-section ICC	*Same-section ICC*
Unweighted global	0.87 (95% CI: 0.81–0.93)	0.88 (95% CI: 0.81–0.93)
Weighted global	0.87 (95% CI: 0.7999–0.93)	0.87 (95% CI: 0.80–0.93)
Hot-spot	0.84 (95% CI: 0.77–0.92)	0.84 (95% CI: 0.77–0.92)

Abbreviations: CI, credible interval; ICC, intraclass correlation coefficient.

Categorical concordance of Ki67 scoring

With regard to agreement on a categorical level (rather than on a continuous, 0–100% scale), considering the categories $< 10\%$, 10–20%, and $> 20\%$, the relationship between percent agreement and continuous score is shown in Supplementary Figure 3. It shows excellent to perfect agreement on cases with scores that are either much lower or higher than the intermediate range of 10–20%.

Visually, there was moderately strong agreement across laboratories in the pathologist-selected location of the hot-spots in each of the core-cut biopsies (Figure 5 shows some examples; virtual slide images of all core-cut biopsy slides used in this study and the corresponding selected fields and scores can be viewed at http://www.gpec.ubc.ca/papers/ki67p3).

After selection of the fields to score, the median times required for nuclei counting were 3 and 4 min for the global and hot-spot methods, respectively.

DISCUSSION

The overarching goal of the International Ki67 Working Group multiphase program is to build enough evidence to either support or refute the notion that Ki67 assessed by immunohistochemistry is sufficiently analytically and clinically validated to be implemented in routine clinical practice for management of breast cancer.[10] Our previous studies demonstrated substantial interlaboratory variability in Ki67 scoring among some of the world's leaders in the field, even when reading centrally stained slides (phase 1).[17] However, this variability was reduced by introducing a standardized, practical visual scoring method (phase 2)[18]—a method that does not require any special equipment beyond a desktop computer and light microscope.

In this third study, we progressed to a more 'real world' circumstance of reading Ki67 staining of core biopsies, while still controlling for variability due to preanalytical and analytical aspects of the assay.[10] We have demonstrated that it is possible, given a set of clearly defined training exercise and scoring instructions, for pathologists to achieve high interobserver agreement in scoring Ki67 on core-cut biopsies using a conventional light microscope and manual field selection, with no additional aid such as counting grid or software. The average time taken to score, once fields for scoring had been selected, was between 3 and 4 min regardless of the method and was judged to be acceptable in general practice by the participants in the current study.

We found that the global unweighted method achieved the observed highest ICC and CI (0.87, 95% CI: 0.81–0.93) compared with weighted global (0.87, 95% CI: 0.7999–0.93) and hot-spot

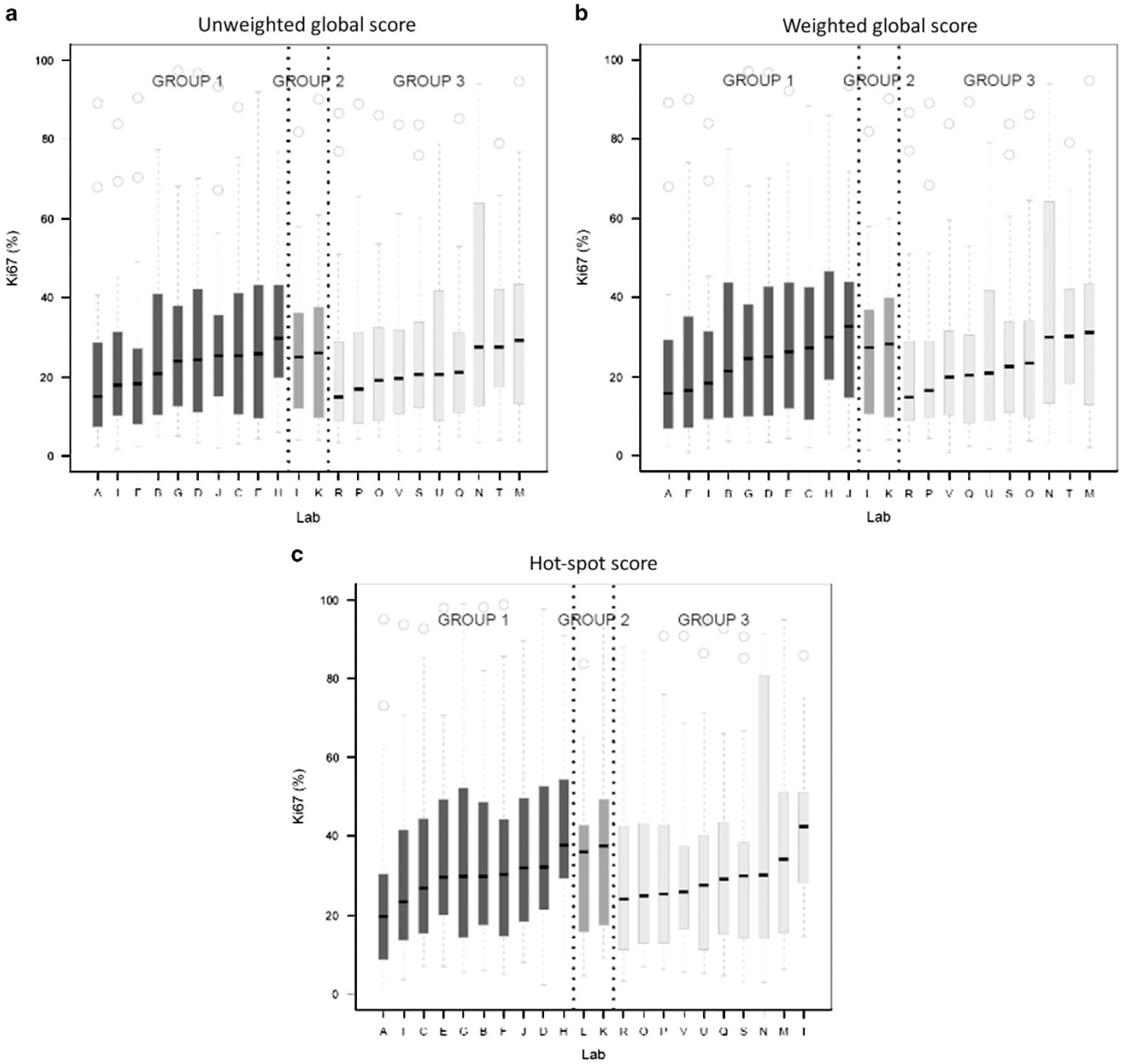

Figure 1. Ki67 scores (**a**, unweighted global; **b**, weighted global; **c**, hot-spot) of all 22 laboratories (by group): black for Group 1, medium gray for Group 2, and light gray for Group 3. Laboratories are ordered (within each group) by the median scores. The bottom/top of the box in each box plot represent the first (Q1)/third (Q3) quartiles, the bold line inside the box represents the median and the two bars outside the box represent the lowest/highest datum still within 1.5 × the interquartile range (Q3–Q1). Outliers are represented with empty circles.

methods (0.84, 95% CI: 0.77–0.92). Thus the global method was the only one to meet the prespecified criterion of success but the other methods missed this criterion by a small amount and cannot be ruled out as viable alternatives. The results do not provide sufficient evidence that the global method is significantly more reproducible, as measured by ICC, than the others. There appeared to be moderately strong agreement in the location of the selected hot-spots across laboratories (Figure 5a). However, as shown in Figure 5b, even a very slight difference in hot-spot location could result in a large difference in the Ki67 scores (8.6% vs. 26%). Our findings are in agreement with other reports, in which a marginally higher concordance among global compared with hot-spot scores was observed (ICC = 0.904 vs. ICC = 0.894, respectively).[19] We propose that differences in individual fields average out in the global method, and thus the overall score is more robust to variability introduced by the exact localization of the fields selected for scoring.

Despite the conclusion that the scoring aspect of analytical validity has been achieved based on overall assessment by ICC, there are still a few cases with large discrepancies (Figures 4a–c). To understand potential sources of these variabilities, a subsequent exploratory examination of the field selections and scores on individual fields were performed (Supplementary Document: 'Exploratory examination of scoring fields'). Five sources of variability were identified: (1) scoring of ductal carcinoma *in situ* tissue; (2) scoring of stromal cells; (3) positive nuclei being localized within a different part of the selected field; (4) need for recalibration; (5) different hot-spots within a single slide exhibiting

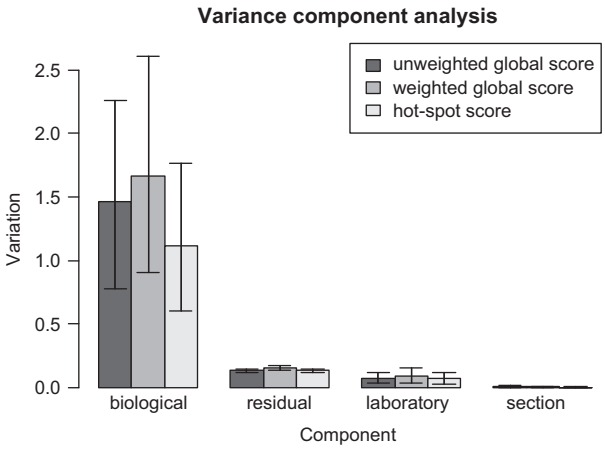

Figure 2. Variance component analysis. Variation due to different components are presented in a bar plot to show the relative magnitude differences between them. Numeric values of the variance components estimates and the corresponding credible intervals are shown in Supplementary Table 5.

different Ki67 scores. Some of these factors may be correctable/preventable (1, 2, 4) while others may be difficult to avoid (3, 5).

Among the rest of the cases, much of the scoring variability that remains between laboratories relates to relatively consistent low or high bias for scorers. Establishment of external quality assessment schemes and regular participation in such programs may improve the agreement between laboratories further, especially in the area of standardizing staining protocol (which has not yet been addressed in any of our studies).

Similar to what was observed in phase 2,[18] clinically important discrepancies persisted among laboratories for some cases in the intermediate Ki67 range between 10 and 20% (Supplementary Figure 4). These discrepancies are of concern, since this is the range in which cutoffs to distinguish high from low Ki67 levels are usually selected and used to make clinical decisions.[6,13,20] There were 19 cases for which at least one of the 22 laboratories reported an unweighted global score in the range of $10\% \leqslant Ki67 \leqslant 20\%$. Strikingly, there were no cases where all laboratories provided scores that were confined to this range. If the intermediate Ki67 range extends to $10\% \leqslant Ki67 \leqslant 30\%$, then there were 26 cases for which at least one of the 22 laboratories reported an unweighted global score in this range. There was only one case for which all laboratories provided scores in this extended cutoff range.

On the other hand, as evident from the heat maps (Figures 4a–c), exceptionally high to unanimous agreement was observed for cases with median Ki67 scores that were either much higher or lower than the intermediate range ($10\% \leqslant Ki67 \leqslant 20\%$): 100% agreement was observed with the global method (unweighted and weighted) on 11/30 (37%) cases and, with hot-spot method, 13/30 (43%) cases. Supplementary Figure 3 shows the relationship between scores and the rate of agreement on categories. It demonstrates that there is often some disagreement, but for scores that are far away from the intermediate range, for example, above 35%, everyone agreed. However, at lower levels there was increasing disagreement.

Ki67 IHC might be used for one of many possible applications, including for determination of breast cancer intrinsic subtype,[21] use in IHC-based multiparameter assays to approximate results from gene expression assays such as the 21-gene recurrence score,[14,15] and use in IHC-based prognostic models.[22,23] Regardless, Ki67 is usually interpreted in the context of other clinicopathological parameters, such as tumor size, lymph node status and grade, or biomarkers, such as ER, PR and HER2 status. In this regard, Denkert *et al.* noted that treatment decisions for individual patients should not be made based on small differences of Ki67 around a given cutpoint.[16] Further studies in the impact of Ki67 scoring variability on multiparameter clinical application would be beneficial. Regardless, the increasing scoring concordance we have observed through our three phases of consensus training suggests some progress toward Ki67 immunohistochemistry applicability in the standard of care setting, assuming proper training and adherence to proficiency testing.

While our study shows that Ki67 visual scoring systems can be standardized to reach high levels of interobserver agreement (as measured by ICC) in centrally stained core-cut biopsy samples, it has several limitations. In clinical practice, additional preanalytical and analytical aspects, such as staining protocol differences,[10] will add substantial variability, as will moving from core-cut biopsies onto whole sections. In addition, the clinical validity (and therefore clinical utility) of this specific scoring system has yet to be confirmed. The data from the current phase 3 study are sufficiently positive to support proceeding to evaluation of these other aspects by consortium members in a series of planned studies.

In conclusion, we believe we are one step closer to standardizing the Ki67 immunohistochemistry assay for use in breast cancer. However, at this stage, we cannot yet recommend this assay platform to be used to drive patient-care decisions in clinical practice.

MATERIALS AND METHODS

This study was approved by the British Columbia Cancer Agency Clinical Research Ethics Board (protocol H10-03420). All samples used in this study were donated by patients who signed a generic consent. All core-cut biopsy material used in this study was excess to diagnostic requirements and ethically available for quality control studies.

Case selection

One hundred and ten cases of estrogen receptor (ER) positive breast cancer were selected from the Academic Department of Biochemistry (ADB) tumor bank at the Royal Marsden Hospital, UK. Sixty-nine of these were further selected for initial sectioning, based on visual estimation of the available material, Haematoxylin and Eosin (H&E) and Ki67 staining using Academic Biochemistry protocols.[24] Quality of each section (for example, crush artifacts and cellularity) was assessed and the percentage of Ki67 positivity was estimated. A set of 40 core-cut biopsy blocks was sectioned and stained in the Royal Marsden Hospital Histopathology Department using monoclonal antibody MIB1 at dilution 1:50 (DAKO UK, Cambridgeshire, UK) using an automated staining system (Ventana Medical Systems, Tucson, AZ, USA) according to the criteria established by consensus of the International Ki67 Working Group.[10] The final set of 30 core-cut biopsy sections was selected on the basis of sufficient cell numbers and quality of staining (Supplementary Figure 1). The distribution of clinicopathological parameters among these 30 cases is shown in Supplementary Table 1.

Sample preparation and distribution

Twenty-four volunteer laboratories, most of whom participated in phase 1 or 2 of the International Ki67 Working Group initiatives, representing 23 institutions from 11 countries, were invited to participate in phase 3.

Five adjacent sections from each of the 30 core-cut biopsy source blocks were centrally stained. The first section was stained with H&E, the second with a myoepithelial marker (p63) and the third to fifth sections with Ki67. Because the time required to have all laboratories review the same slide would have been prohibitive, the latter three Ki67-stained sections were prepared and are designated Groups 1, 2, and 3. Each group of slides included 30 sections, one from each of the 30 patients. The participating laboratories were initially divided into three groups (eight laboratories in each group) and members within the same group were given the same group of slides to score. Because the slides were damaged en route to the third volunteer laboratory in Group 2, members within this group who had not yet scored were subsequently reassigned to Group 1 or 3. Two

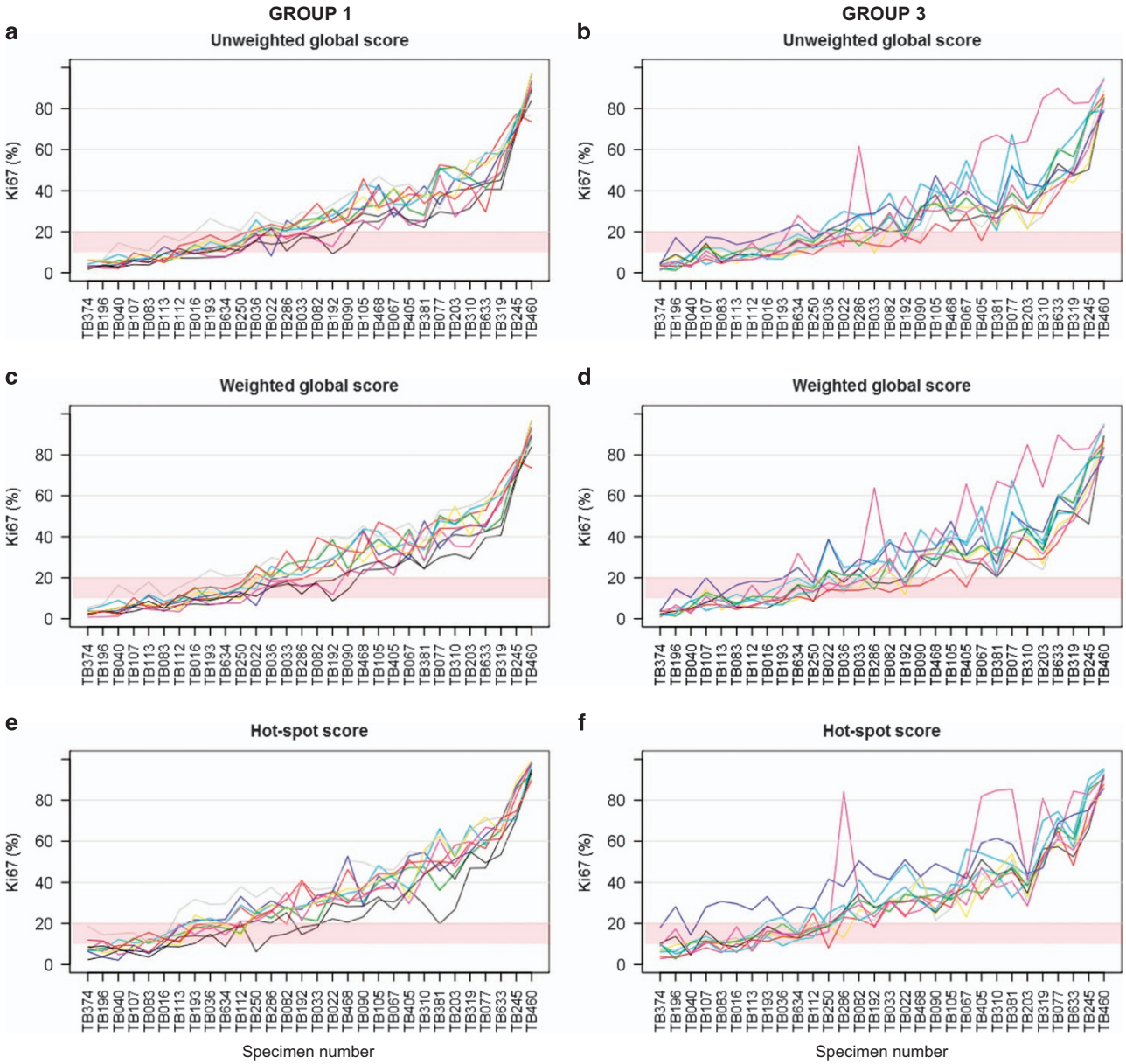

Figure 3. Variability in Ki67 scores (**a**, **c** and **e** correspond to Group 1; **b**, **d** and **f** correspond to Group 3). Each line represents Ki67 scores from one laboratory. Shaded region indicates Ki67 scores between 10 and 20%. Scores from Group 2 are not shown since there are only two laboratories in this group.

volunteer laboratories did not complete the study in time for the analysis. Twenty-two laboratories successfully completed the study: 10 laboratories in Group 1, two in Group 2 and 10 in Group 3.

Scoring protocol

All laboratories were required to complete the phase 2 web-based calibration exercise[18] prior to the phase 3 scoring. This calibrator is publicly accessible at http://www.gpec.ubc.ca/calibrator. The detailed scoring protocol is found in Supplementary Document: 'Instructions for Ki67 Reproducibility Study Phase 3: Core Biopsies'. A modified version of the scoring software (modified for offline use) used in this study can be downloaded at: http://www.gpec.ubc.ca/papers/ki67p3.

Scoring methods

Three scoring methods were assessed in this study: (1) an unweighted global assessment of Ki67 staining; (2) a global assessment that is weighted according to the estimated percentage of the total cancer area covered by each of high, medium, low, or negligible Ki67 staining levels and (3) assessment of Ki67 only in 'hot-spots.'

Global methods attempt to derive an average score across all the tissue available for assessment. In the weighted and unweighted global methods, Ki67 index counting was performed in the same manner, but the final Ki67 score was derived differently. Adapted from a scoring protocol that has been used routinely in the Dowsett ADB laboratory,[24] these two global methods require the pathologist to first assess staining heterogeneity by estimating the percentages of the invasive tumor component of the slide exhibiting relatively high, medium, low or negligible Ki67 scores. On the basis of these estimates, a standard algorithm (Supplementary Figure 2) determined the required number of fields to score for each Ki67 score level (total up to four fields). The pathologist was then asked to count up to 100 invasive tumor nuclei within each field, using a 'typewriter' pattern, similar to how a tissue microarray core was scored in the phase 2 study.[18]

Variations on hot-spot assessments are often used by pathologists for mitotic counting, where the pathologist identifies what appears

a

Specimen number	(median score)	GROUP 1										GROUP 2		GROUP 3								
		A	I	F	B	G	D	J	C	E	H	L	K	R	P	O	V	S	U	Q	N	T	M
TB374 (4)	2	2	4	6	6	4	2	3	4	6	4	4	4	4	5	1	2	2	5	4	4	4	22:0:0
TB196 (5)	4	4	2	5	6	3	3	4	6	6	4	8	3	6	8	5	1	2	9	6	17	4	21:1:0
TB040 (5)	3	3	2	5	5	4	2	6	9	15	12	7	4	8	5	3	6	9	5	4	10	9	20:2:0
TB083 (7)	4	5	5	5	10	6	8	7	7	11	11	10	4	7	8	6	8	7	5	5	17	12	18:4:0
TB113 (8)	7	10	6	6	7	13	5	5	9	18	10	7	7	8	5	6	10	6	9	8	14	9	18:4:0
TB107 (8)	4	6	6	7	8	6	10	6	6	12	13	8	7	7	14	11	13	4	14	8	18	12	14:8:0
TB112 (9)	7	12	8	10	7	9	13	8	9	16	12	9	6	8	8	6	12	10	9	14	15	8	14:8:0
TB016 (10)	7	9	10	10	12	11	15	14	10	20	12	13	9	10	9	8	11	7	8	6	18	13	12:10:0
TB193 (11)	8	10	8	11	16	12	18	10	15	26	12	15	9	8	8	11	11	7	11	15	20	15	6:14:2
TB634 (13)	8	12	8	10	13	13	15	12	12	23	9	13	11	12	9	15	17	12	16	28	24	19	4:15:3
TB250 (14)	12	11	10	15	13	13	19	15	15	20	16	10	9	14	14	13	15	12	12	21	17	16	2:18:2
TB022 (19)	14	21	22	18	18	8	24	22	18	25	18	20	15	16	16	18	20	21	22	13	24	30	1:12:9
TB036 (19)	15	16	16	18	23	19	21	20	26	30	24	24	14	12	16	16	21	14	18	19	21	25	0:13:9
TB033 (20)	19	17	20	19	22	21	26	26	22	30	26	28	14	20	10	20	19	20	22	18	29	28	1:10:11
TB286 (21)	15	11	16	18	24	26	22	20	21	24	20	29	15	16	24	20	13	28	19	62	28	28	0:10:12
TB192 (21)	9	19	13	31	25	26	25	31	28	33	28	18	18	17	17	15	18	18	21	37	27	22	1:9:12
TB082 (26)	17	17	16	22	26	23	34	26	26	29	28	34	13	24	22	30	29	27	20	22	34	39	0:5:17
TB090 (28)	14	23	24	24	30	31	27	25	34	38	36	38	14	16	30	31	32	24	31	27	26	44	0:3:19
TB381 (29)	25	22	26	38	38	42	34	28	38	37	27	24	30	31	22	24	29	20	27	67	30	33	0:0:22
TB468 (32)	25	28	21	41	31	43	32	33	41	47	24	30	20	32	32	32	29	30	25	44	34	36	0:1:21
TB405 (33)	26	25	23	38	36	35	42	31	33	43	38	29	16	22	32	34	30	33	28	64	33	38	0:1:21
TB105 (34)	25	29	26	30	37	31	46	32	43	41	30	38	24	32	34	30	34	43	38	36	48	40	0:0:22
TB067 (34)	30	32	32	34	41	27	34	41	33	42	32	38	29	27	32	30	36	49	26	38	40	55	0:0:22
TB203 (36)	29	40	27	51	37	46	36	52	45	45	31	43	29	22	21	32	31	36	30	64	44	36	0:0:22
TB077 (39)	30	37	48	52	38	33	40	50	51	31	38	43	33	38	33	43	39	52	32	62	52	68	0:0:22
TB310 (41)	32	41	34	48	55	42	42	46	46	53	36	37	29	28	36	38	41	42	39	85	42	46	0:0:22
TB633 (49)	41	44	44	54	52	45	30	42	58	58	51	57	39	43	47	43	60	46	53	90	50	58	0:0:22
TB319 (51)	41	45	49	66	60	59	56	49	58	61	42	50	51	48	44	48	56	52	48	82	48	67	0:0:22
TB245 (70)	68	70	70	78	68	70	67	76	73	75	58	61	77	66	54	61	76	78	50	83	66	77	0:0:22
TB460 (88)	89	84	90	74	98	97	94	88	92	77	82	90	87	89	86	84	84	79	85	94	79	95	0:0:22

Laboratory

b

Specimen number	(median score)	GROUP 1										GROUP 2		GROUP 3								
		A	F	I	B	G	D	E	C	H	J	L	K	R	P	V	Q	U	S	O	N	T	M
TB374 (3)	2	1	2	4	3	4	4	2	5	2	5	4	4	4	1	2	2	2	4	4	4	2	22:0:0
TB196 (4)	4	1	4	4	4	3	6	4	7	4	6	8	4	5	7	4	2	1	4	6	14	4	21:1:0
TB040 (5)	2	1	3	5	5	4	9	5	16	2	1	8	4	8	3	5	9	6	5	4	10	9	20:2:0
TB083 (7)	4	4	5	5	7	6	8	7	11	7	8	10	4	7	6	6	7	8	8	5	16	12	19:3:0
TB107 (8)	4	6	6	7	7	6	5	6	12	10	10	8	7	7	11	8	4	13	14	8	20	11	14:7:1
TB113 (8)	6	5	9	6	7	11	12	5	18	5	10	7	7	10	6	11	6	10	4	9	12	9	16:6:0
TB112 (9)	7	3	12	7	4	8	10	9	16	11	8	7	6	10	6	5	10	11	5	16	18	9	15:7:0
TB016 (10)	7	10	9	10	12	10	9	15	19	15	11	13	9	9	8	5	7	11	10	7	18	13	12:10:0
TB193 (10)	7	7	9	10	13	11	16	9	24	15	12	12	9	7	10	9	7	10	7	15	20	16	10:11:1
TB634 (12)	8	8	11	10	11	13	12	12	21	15	9	13	11	12	15	17	12	17	10	32	25	20	4:15:3
TB250 (14)	12	8	11	16	10	13	15	15	21	18	17	10	9	14	13	8	13	15	15	21	18	16	5:15:2
TB036 (19)	15	16	16	18	23	19	26	20	30	21	27	24	14	12	16	18	14	21	16	19	22	25	0:13:9
TB022 (20)	11	14	18	22	20	6	15	26	24	26	13	20	14	15	18	23	21	24	24	13	39	38	1:10:11
TB033 (21)	17	19	18	19	21	20	22	27	28	33	26	28	14	19	18	24	21	22	15	24	29	26	0:8:14
TB192 (22)	9	12	19	29	26	26	29	38	40	36	29	19	16	16	16	20	18	17	12	42	33	24	1:9:12
TB286 (23)	16	16	12	20	25	26	21	28	24	23	24	28	15	17	21	18	28	14	24	64	27	29	0:8:14
TB082 (26)	17	17	18	22	24	24	27	29	30	40	30	35	13	23	30	17	27	29	23	22	37	39	0:5:17
TB381 (30)	25	28	24	44	38	48	34	28	38	35	30	28	35	30	21	20	20	31	27	67	27	33	0:0:22
TB090 (31)	14	24	23	21	36	34	34	24	39	33	37	38	16	16	31	30	24	31	31	28	33	44	0:3:19
TB468 (31)	24	22	26	43	28	43	44	33	45	32	27	30	20	29	32	25	30	29	34	44	34	36	0:1:21
TB405 (31)	25	21	25	31	35	35	34	33	43	43	37	31	16	20	28	28	36	31	29	66	31	37	0:1:21
TB067 (36)	30	42	32	32	31	29	37	44	47	36	35	40	29	25	30	27	49	36	35	42	40	55	0:0:22
TB105 (36)	24	28	28	27	36	31	42	39	40	47	35	38	24	33	30	38	43	34	36	36	48	40	0:0:22
TB203 (39)	29	35	40	51	40	46	53	52	55	45	38	43	29	24	32	33	36	34	27	64	42	38	0:0:22
TB077 (43)	30	44	37	49	41	34	48	51	53	44	39	44	33	35	40	30	52	42	32	64	52	68	0:0:22
TB310 (43)	32	35	41	48	55	43	46	46	53	44	38	41	29	28	38	44	42	44	40	85	45	46	0:0:22
TB633 (51)	40	46	43	53	57	45	56	43	59	46	51	57	37	43	43	53	52	60	46	90	60	58	0:0:22
TB319 (53)	41	57	45	67	62	59	61	49	66	58	44	50	51	51	48	52	52	56	50	82	53	67	0:0:22
TB245 (71)	68	74	70	78	68	70	74	76	75	72	58	60	77	68	60	46	78	76	65	83	67	77	0:0:22
TB460 (89)	89	90	84	74	97	97	92	88	86	94	82	90	87	89	84	89	79	84	86	94	79	95	0:0:22

Laboratory

Figure 4. Heat map of Ki67 scores (**a**: unweighted global; **b**: weighted global; **c**: hot-spot). Rows represent cases and columns represent laboratories. Green color indicate that the score is $<10\%$, yellow 10–20%, and red $>20\%$. Cases are ordered by the median scores (across laboratories), which are shown in parentheses beside the specimen number. Laboratories are ordered (within each group) by the median scores (across cases). The three colon-separated numbers to the right of the table represent the number of laboratories giving scores falling into different ranges: $<10\%$ (left-most), 10–20% (middle) and $>20\%$ (right-most). For example, '15:6:1' indicates that 15 laboratories gave a score of $<10\%$, six laboratories between 10 and 20% and one laboratory $>20\%$.

to be the most active area of cell division. The hot-spot method required the pathologist to select one high-power field with high staining rate and count up to 500 invasive tumor nuclei in a 'typewriter' pattern.

Statistical analyses

Prespecified criterion for success. Prior to data collection, it was hypothesized that at least one of the scoring methods would have an associated ICC of at least 0.80. For planning purposes, power calculations

c

Specimen number (median score)	GROUP 1										GROUP 2		GROUP 3								
	A	I	C	E	G	B	F	J	D	H	L	K	R	O	P	V	U	Q	S	N	M	T	
TB196 (7)	4	9	8	7	6	7	11	11	4	15	5	16	3	10	7	17	5	14	3	4	6	28	15:6:1
TB040 (8)	7	7	8	12	7	9	5	8	2	15	13	9	6	10	10	6	8	5	11	5	11	14	14:8:0
TB374 (8)	2	8	7	7	9	7	7	12	6	19	11	12	4	7	8	9	10	10	10	3	6	18	15:7:0
TB083 (10)	6	4	10	11	12	6	12	16	5	11	13	9	10	8	11	7	6	10	11	6	10	31	10:11:1
TB107 (11)	7	5	13	10	8	9	6	12	9	15	17	15	10	12	11	11	11	16	12	8	14	28	7:14:1
TB016 (12)	9	8	13	14	10	12	11	12	11	13	15	14	6	12	13	18	6	9	11	10	11	30	5:16:1
TB113 (12)	9	13	15	18	9	11	19	11	18	26	16	11	12	7	6	7	8	13	13	8	15	27	7:13:2
TB634 (15)	9	16	18	20	19	20	14	20	22	30	14	17	13	13	11	15	12	13	14	15	15	28	1:17:4
TB193 (18)	10	14	18	21	24	18	15	19	22	32	20	23	11	16	15	16	14	19	15	19	21	33	0:14:8
TB036 (19)	15	14	19	22	21	17	19	20	21	29	21	23	14	13	14	15	9	15	20	15	23	24	1:13:8
TB112 (20)	21	19	15	29	14	28	21	18	33	38	15	18	25	20	20	18	13	16	17	14	28	27	0:11:11
TB250 (21)	6	21	26	24	26	22	21	23	26	33	21	26	8	19	21	20	25	19	19	16	18	42	2:6:14
TB286 (26)	14	20	23	29	29	26	25	26	31	38	22	31	23	13	28	25	29	26	25	84	23	38	0:2:20
TB192 (27)	18	15	23	28	28	40	21	41	27	39	38	32	19	25	22	23	25	28	28	18	30	44	0:4:18
TB082 (29)	15	25	28	27	27	20	35	32	27	31	32	36	22	27	23	29	21	35	32	31	41	51	0:2:20
TB033 (31)	19	18	21	30	30	29	34	32	34	37	34	39	31	32	30	31	33	31	26	30	40	42	0:2:20
TB090 (33)	23	28	31	36	37	34	33	38	29	50	41	40	31	24	22	32	26	25	34	27	37	49	0:0:22
TB468 (34)	21	28	32	35	37	46	31	30	53	51	38	43	33	43	43	26	30	31	33	33	37	43	0:0:22
TB022 (34)	22	30	36	32	34	31	32	36	35	37	43	40	23	30	35	24	38	31	34	30	49	51	0:0:22
TB105 (37)	31	33	40	48	41	37	38	44	43	47	38	39	28	37	27	36	39	34	32	35	32	46	0:0:22
TB067 (41)	30	27	43	42	45	37	40	44	37	46	41	44	45	23	36	29	38	38	36	45	56	42	0:0:22
TB203 (42)	27	42	44	52	52	58	47	49	51	54	39	52	39	33	30	28	40	35	38	41	43	44	0:0:22
TB310 (45)	29	44	47	49	56	44	44	50	55	54	44	50	43	45	37	37	44	44	41	85	51	61	0:0:22
TB405 (47)	36	36	47	37	32	51	30	50	53	55	47	45	32	41	44	47	46	51	35	82	54	59	0:0:22
TB381 (50)	20	50	36	66	63	49	61	50	46	65	51	49	45	54	51	41	33	46	47	85	49	59	0:1:21
TB319 (55)	47	55	54	68	65	60	58	59	55	66	47	54	53	52	49	52	57	56	50	81	70	47	0:0:22
TB077 (61)	47	50	60	58	72	60	67	57	63	70	43	55	65	58	63	61	71	57	67	62	74	69	0:0:22
TB633 (62)	62	54	65	70	65	61	66	71	70	72	58	61	48	58	61	56	57	53	61	84	64	73	0:0:22
TB245 (76)	73	71	85	71	89	82	86	75	87	73	65	77	76	70	76	69	86	66	85	83	90	75	0:0:22
TB460 (93)	95	94	93	98	99	98	99	90	98	91	84	93	88	87	91	91	94	93	91	91	95	86	0:0:22

Laboratory

Figure 4. (Continued)

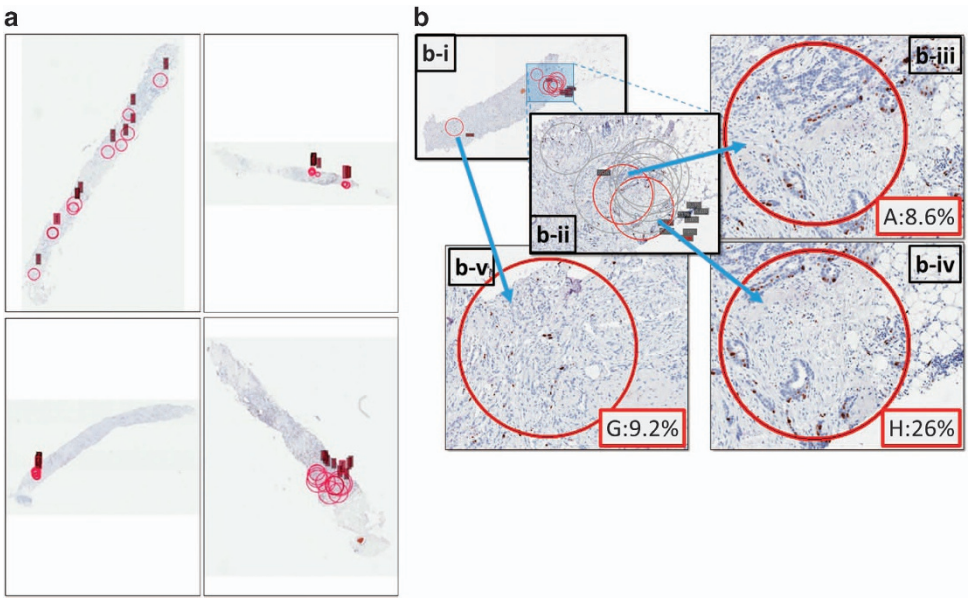

Figure 5. Hot-spot field selection by different laboratories on the same core-cut biopsy slide. (**a**) Selections (indicated by red circles) on some example core biopsies. (**b**) Example of a single-core biopsy (median score: 12%) with zoomed-in fields. Each laboratory was asked to circle the area considered by that laboratory to be the hot-spot (**b-i**). Most pathologists honed in on the same area of the core, although individual-selected circular scoring fields do not always overlap. (**b-iii**, **b-iv**) Segments of the same area chosen by two different laboratories to read Ki67. (**b-v**) The 'outlier' field selected by only one laboratory as the hot-spot.

performed under a variety of scenarios considered to represent good reproducibility and similar to the results observed in the phase 2 study showed that with 21 laboratories there would be 80% power to exclude ICCs lower than the prespecified ICC of 0.8 from a 95% credible interval for a given scoring method. This success criterion of '0.8' was chosen using criteria similar to those for Kappa value interpretation: 0.81–1 indicating 'almost perfect' agreement.[25]

Ki67 scoring. The Ki67 score was defined as the percentage of invasive tumor cells positively stained in the examined field(s). Positive staining was defined as any brown stain in the nucleus above background, illustrated by sample images; negative staining was scored when the invasive cancer cell showed only a blue counterstained nucleus. The unweighted global and hot-spot scores were simply the total number of positively stained tumor nuclei counted divided by the total number of tumor nuclei counted (in

the one hot-spot field, or across all fields for the global method). The weighted global score was derived with tumor nuclei counts in each assessed field weighted by the estimated percentage of the total cancer area covered by each of high, medium, low, or negligible Ki67 staining levels. For example, consider a slide estimated to have 10% of its area covered by relatively high Ki67 index regions, while 90% of the area is covered by relatively low-level Ki67 regions. By the algorithm (Supplementary Figure 2), the required number of fields to select and score is one high field and three low fields. Suppose the number of positive/total tumor nuclei counted in the high field is 85/100 and the three low fields are 30/100, 20/80, and 18/90. The weighted Ki67 score would be $0.1 \times (85/100) + 0.9 \times ((30+20+18)/(100+80+90)) = 31\%$, whereas the unweighted score would be $(85+30+20+18)/(100+100+80+90) = 41\%$. For the statistical analysis, the Ki67 score was 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.[10]

ICC estimates (ranging from 0 to 1, with 1 representing perfect reproducibility) were computed as previously reported in the phase 2 study.[18] Briefly, variance component analyses were performed to quantify the contributions from the following sources of variability: scoring laboratory, patient tumor (biological variation—each core-cut biopsy block represents a unique patient) and section of the core-cut biopsy block. Similar to the phase 2 study, same-section and different-section ICC were computed. Same-section refers to scoring laboratories scoring the same set of core-cut biopsy slides, whereas different-section refers to scoring laboratories scoring different sections of the same core-cut biopsy blocks. CI for the variance components and the ICCs were obtained using the Markov Chain Monte Carlo routines for fitting generalized linear mixed models.

All data analyses were performed using R version 3.2.1.[26] Sources of 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 heat maps, boxplots and spaghetti plots.

ACKNOWLEDGMENTS

This work was supported by a generous grant from the Breast Cancer Research Foundation (D.F.H.). 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 Nancy Davidson, Thomas Buchholz, Martine Piccart, and Larry Norton. This work was supported by a generous grant from the Breast Cancer Research Foundation.

CONTRIBUTIONS

S.C.Y.L.: study design, data collection, manuscript drafting and review. T.O.N.: study design, manuscript drafting and review. L.Z.: study design, collection and preparation of samples, data collection, manuscript drafting and review. I.A.: study design, data collection, manuscript drafting and review. S.S.B.: study design, manuscript drafting and review. A.L.B.: study design, data collection, manuscript drafting and review. J.M.S.B.: study design, manuscript drafting and review. S.B.: study design, data collection, manuscript drafting and review. M.C.C.: study design, data collection, manuscript drafting and review. A.D.: study design, data collection, manuscript drafting and review. R.A.E.: study design, data collection, manuscript drafting and review. S.F.: study design, data collection, manuscript drafting and review. C.M.F.: study design, data collection, manuscript drafting and review. D.G.: study design, data collection, manuscript drafting and review. A.M.G.: study design, data collection, manuscript drafting and review. D.G.: study design, data collection, manuscript drafting and review. C.G.: study design, data collection, manuscript drafting and review. J.C.H.: study design, data collection, manuscript drafting and review. Z.K.: study design, data collection, manuscript drafting and review. A-V.L.: study design, data collection, manuscript drafting and review. M-G.L: study design, data collection, manuscript drafting and review. M.G.M.: study design, data collection, manuscript drafting and review. T.M.: study design, data collection, manuscript drafting and review. S.N-M.: study design, data collection, manuscript drafting and review. C.K.O.: study design, manuscript drafting and review. F.M.P-L.: study design, data collection, manuscript drafting and review. T.P.: study design, data collection, manuscript drafting and review. T.S.: study design, data collection, manuscript drafting and review. R.S.: study design, data collection, manuscript drafting and review. J.S.: study design, data collection, manuscript drafting and review. G.V.: study design, manuscript drafting and review. D.F.H.: study design, manuscript drafting and review. L.M.McS.: study design, statistical analysis, manuscript drafting and review. M.D.: study design, manuscript drafting and review.

COMPETING INTERESTS

J.M.S.B. has consulted for Insight Genetics and BioNTech and received compensation. T.O.N. has consulted for Nanostring and received compensation. C.K.O. has consulted for Astra Zeneca, Genentech and NanoString and received compensation. The remaining authors declare no conflict of interest.

REFERENCES

1. Lei, Y. *et al.* The prognostic role of Ki-67/MIB-1 in upper urinary-tract urothelial carcinomas: a systematic review and meta-analysis. *J. Endourol.* **29**, 1302–1308 (2015).
2. Desouki, M. M., Chamberlain, B. K. & Li, Z. The role of immunohistochemistry in the evaluation of gynecologic pathology part 2: a comparative study between two academic institutes. *Ann. Diagn. Pathol.* **19**, 296–300 (2015).
3. Luporsi, E. *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.* **132**, 895–915 (2012).
4. de Azambuja, E. *et al.* Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. *Br. J. Cancer* **96**, 1504–1513 (2007).
5. Denkert, C. *et al.* Ki67 levels as predictive and prognostic parameters in pre-therapeutic breast cancer core biopsies: a translational investigation in the neoadjuvant GeparTrio trial. *Ann. Oncol.* **24**, 2786–2793 (2013).
6. Inwald, E. C. *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.* **139**, 539–552 (2013).
7. Viale, G. *et al.* Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. *J. Clin. Oncol.* **25**, 3846–3852 (2007).
8. Viale, G. *et al.* Predictive value of tumor Ki-67 expression in two randomized trials of adjuvant chemoendocrine therapy for node-negative breast cancer. *J. Natl Cancer Inst.* **100**, 207–212 (2008).
9. Yerushalmi, R., Woods, R., Ravdin, P. M., Hayes, M. M. & Gelmon, K. A. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol.* **11**, 174–183 (2010).
10. Dowsett, M. *et al.* Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J. Natl Cancer Inst.* **103**, 1656–1664 (2011).
11. Petrelli, F., Viale, G., Cabiddu, M. & Barni, S. Prognostic value of different cut-off levels of Ki-67 in breast cancer: a systematic review and meta-analysis of 64,196 patients. *Breast Cancer Res. Treat.* **153**, 477–491 (2015).
12. Criscitiello, C. *et al.* High Ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in luminal B HER2 negative and node-positive breast cancer. *Breast* **23**, 69–75 (2014).
13. Coates, A. S. *et al.* Tailoring therapies-improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer. *Ann. Oncol.* **26**, 1533–1546 (2015).
14. Allison, K. H., Kandalaft, P. L., Sitlani, C. M., Dintzis, S. M. & Gown, A. M. Routine pathologic parameters can predict Oncotype DX recurrence scores in subsets of ER positive patients: who does not always need testing? *Breast Cancer Res. Treat.* **131**, 413–424 (2012).
15. Turner, B. M. *et al.* Use of modified Magee equations and histologic criteria to predict the Oncotype DX recurrence score. *Mod. Pathol.* **28**, 921–931 (2015).
16. Denkert, C. *et al.* Strategies for developing Ki67 as a useful biomarker in breast cancer. *Breast* **24 Suppl 2**, S67–S72 (2015).
17. Polley, M. Y. *et al.* An international Ki67 reproducibility study. *J. Natl Cancer Inst.* **105**, 1897–1906 (2013).
18. Polley, M. Y. *et al.* An international study to increase concordance in Ki67 scoring. *Mod. Pathol.* **28**, 778–786 (2015).
19. Shui, R., Yu, B., Bi, R., Yang, F. & Yang, W. An interobserver reproducibility analysis of Ki67 visual assessment in breast cancer. *PLoS ONE* **10**, e0125131 (2015).
20. Stuart-Harris, R., Caldas, C., Pinder, S. E. & Pharoah, P. Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. *Breast* **17**, 323–334 (2008).
21. Maisonneuve, P. *et al.* Proposed new clinicopathological surrogate definitions of luminal A and luminal B (HER2-negative) intrinsic breast cancer subtypes. *Breast Cancer Res.* **16**, R65 (2014).

22. Cuzick, J. *et al.* Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *J. Clin. Oncol.* **29**, 4273–4278 (2011).
23. Miglietta, L. *et al.* A prognostic model based on combining estrogen receptor expression and Ki-67 value after neoadjuvant chemotherapy predicts clinical outcome in locally advanced breast cancer: extension and analysis of a previously reported cohort of patients. *Eur. J. Surg. Oncol.* **39**, 1046–1052 (2013).
24. Zabaglo, L. *et al.* Comparative validation of the SP6 antibody to Ki67 in breast cancer. *J. Clin. Pathol.* **63**, 800–804 (2010).
25. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
26. R Core Team. *R: A Language And Environment For Statistical Computing* (R Foundation for Statistical Computing, Austria, 2015).

Supplementary Information accompanies the paper on the *npj Breast Cancer* website (http://www.nature.com/npjbcancer)