

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).

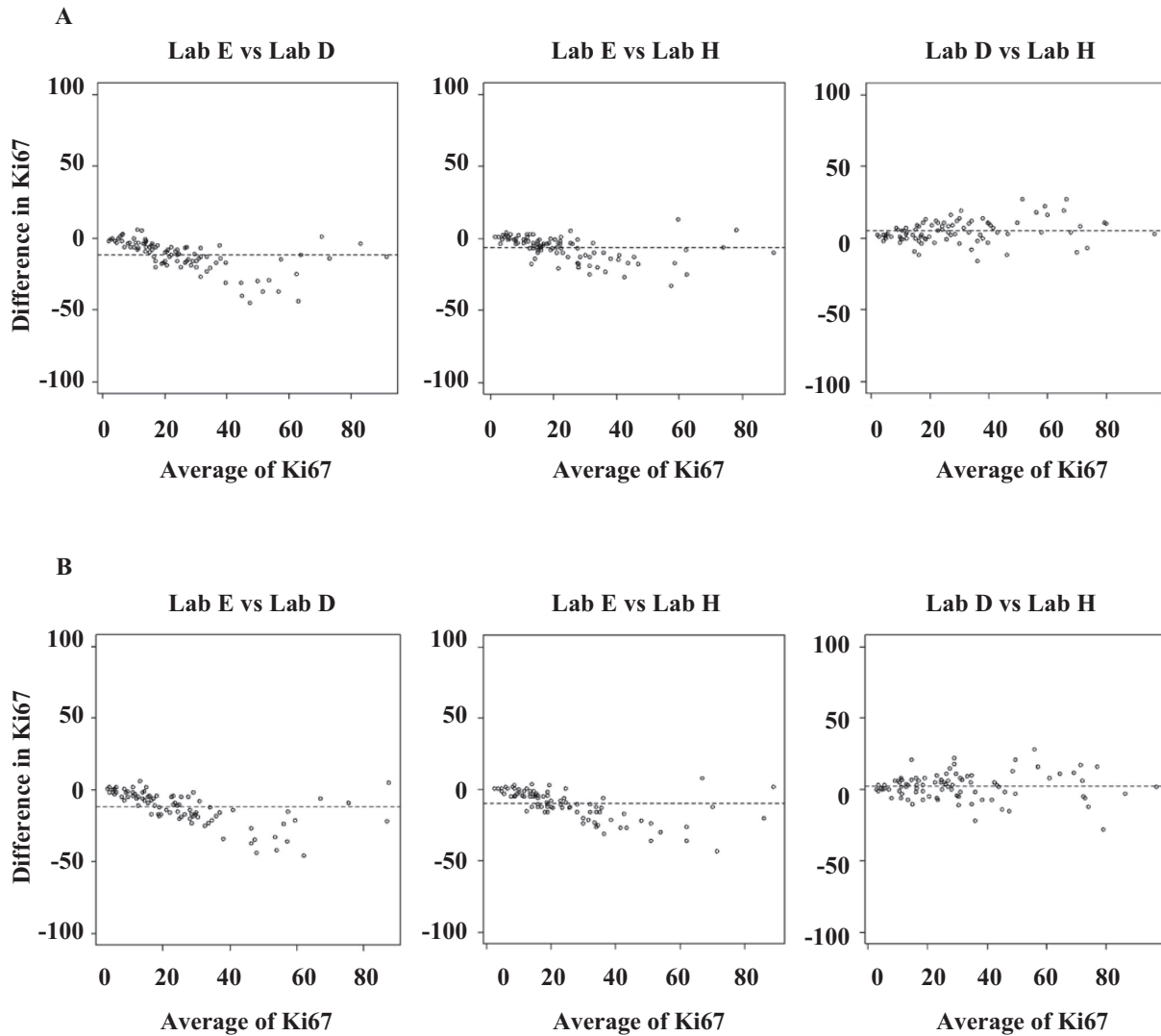


Figure 4. Pairwise Bland–Altman plots showing high interlaboratory reproducibility among three laboratories (Laboratories D, E, and H). **A)** Experiment 2A: central staining, local scoring method. **B)** Experiment 2B: local staining, local scoring method. The y-axis (difference in Ki67) represents 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

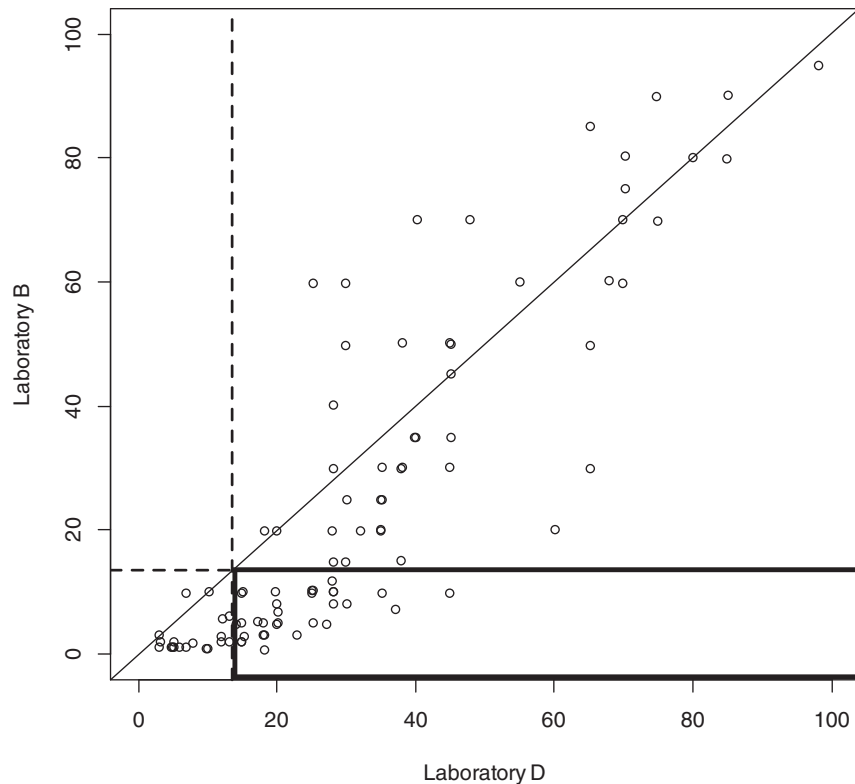


Figure 5. Degree of concordance (**bolded box**) between Laboratory B and Laboratory D at the St. Gallen–recommended Ki67 percentage of positively stained tumor cells cutoff of 13.5%. At a hypothetical 13.5% cutoff, there are 32.3% cases that Laboratory D would call high Ki67 but Laboratory B would call 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 postmenopausal 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 pancreaticobiliary 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/accommodations/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 Kandalaf, Inès Raouf, 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.