MR SAMUEL C.Y. LEUNG (Orcid ID : 0000-0003-0138-7254)

DR MARTIN C. CHANG (Orcid ID : 0000-0003-0184-1314)

DR ANNA EHINGER (Orcid ID : 0000-0001-9225-7396)

DR BERT VAN DER VEGT (Orcid ID : 0000-0002-2613-1506)

Article type : Original Article

Title: Analytical validation of a standardized scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicenter collaboration.

Short running title: Standardized visual scoring of Ki67 in breast cancer

Authors: Samuel C.Y. Leung[1], Torsten O. Nielsen[1], Lila A. Zabaglo[2], Indu Arun[3], Sunil S. Badve[4], Anita L. Bane[5], John M.S. Bartlett[6,22], Signe Borgquist[7], Martin C. Chang[8], Andrew Dodson[9], Anna Ehinger[10], Susan Fineberg[11], Cornelia M. Focke[12], Dongxia Gao[1], Allen M. Gown[13], Carolina Gutierrez[14], Judith C. Hugh[15], Zuzana Kos[16], Anne-Vibeke Lænkholm[17], Mauro G. Mastropasqua[18], Takuya Moriya[19], Sharon Nofech-Mozes[20], C. Kent Osborne[14], Frédérique M. Penault-Llorca[21], Tammy Piper[22], Takashi Sakatani[23], Roberto Salgado[24,25], Jane Starczynski[26], Tomoharu Sugie[27], Bert van der Vegt[28], Giuseppe Viale[18, 29], Daniel F. Hayes[30], Lisa M. McShane[31] and 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)

Affiliations: [1]University of British Columbia, Vancouver, British Columbia, Canada; [2]The Institute of Cancer Research, London, United Kingdom; [3]Tata Medical Center, Kolkata, West Bengal, India; [4]Indiana University Simon Cancer Center, Indianapolis, Indiana, United States; [5]Juravinski Hospital and Cancer Centre, McMaster University, Hamilton, Ontario, Canada; [6]Ontario Institute for Cancer Research, Toronto, Ontario, Canada; [7]Division of Oncology and Pathology, Department of Clinical Science, Lund University, Lund, Sweden; [8]Department of Pathology & Laboratory Medicine, University of Vermont Medical Center, Burlington, Vermont, United States; [9]Ralph Lauren Centre for Breast Cancer Research, The Royal Marsden Hospital, London, United Kingdom; [10]Department of Clinical Genetics and Pathology, Skane University Hospital, Lund University, Lund, Sweden; [11]Montefiore Medical Center and the Albert Einstein College of Medicine, Bronx, New York, United States; [12]Dietrich-Bonhoeffer Medical Center, Neubrandenburg, Mecklenburg-Vorpommern, Germany; [13]PhenoPath Laboratories, Seattle, Washington, United States; [14]Lester and Sue Smith Breast Center and Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States; [15]University of Alberta, Edmonton, Alberta, Canada; [16]University of Ottawa and The Ottawa Hospital, Ottawa, Ontario, Canada; [17]Department of Surgical Pathology, Zealand University Hospital, Slagelse, Region Sjælland, Denmark; [18]European Institute of Oncology, Milan, Italy; [19]Kawasaki Medical School, Kurashiki, Okayama Prefecture, Japan; [20]University of Toronto Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada; [21]Centre Jean Perrin and Université d'Auvergne, Clermont-Ferrand, France; [22]Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, United Kingdom; [23]Nippon Medical School, Bunkyo-ku, Tokyo, Japan; [24]Department of Pathology, GZA-ZNA, Antwerp, Belgium; [25]Division of Research, Peter MacCallum Cancer Centre, Melbourne, Australia; [26]Birmingham Heart of England, National Health Service, Birmingham, United Kingdom; [27]Kansai Medical University, Hirakata, Osaka, Japan; [28]University Medical Center Groningen, Groningen, Netherlands; [29]University of Milan, Milan, Italy; [30]University of Michigan Rogel Cancer Center, Ann Arbor, Michigan, United States and [31]National Cancer Institute, Bethesda, Maryland, United States.

Corresponding author: Samuel Leung; Full postal address: Room 509, 2660 Oak Street, Jack Bell Research Center, Vancouver, BC, V6H 3Z6; Telephone: 604-875-4111 ext. 68893; Email: sam.leung@vch.ca

Acknowledgement of support: this work was supported by a generous grant from the Breast Cancer Research Foundation.

Conflict of interest statement:

Dr. Badve has participated in Scientific Advisory Boards/ Speaker for Genomic Health Inc., Dako/Agilent, Roche Diagnostics, Targos GmBH; Athenax, Konica-Minolta and received compensation. Dr. Badve has received research funding or in kind support from Dako/Agilent. Dr. Badve has intellectual property right/ownership interests with IU. He is also associated with 2 startup companies (SYSGenomics and YeSSGenomics).

Dr. Bartlett has consulted for BioNTech GmBH, Biotheranostics Inc, RNA Diagnostics, and received compensation. Dr Bartlett has participated in Scientific Advisory Boards for Biotheranostics and RNA Diagnostics and received compensation. Dr Bartlett has received research funding or in kind support from Nanostring, BIotheranostics Inc, BioNTech GmBH. Dr Bartlett has intellectual property right/ownership interests with OICR/FACIT.

Dr. Borgquist has participated in educational talks / covered scientific conferences by Roche and Novartis.

Dr. Dowsett is on the Oncology Advisory Board for Radius and has provided ad hoc advice to Orion and Gtx. He has received lecture fees from Myriad and Roche and institutional research grants from Radius, Astrazeneca and Puma. He receives income from the Institute of Cancer Research Rewards for inventors Scheme (abiraterone).

Dr. Ehinger has participated in educational talks organized by Roche but without economical compensation.

Dr. Fineberg participated in a scientific advisory board for Genomic Health and have received monetary compensation (not for salary).

Dr. Hayes reports research support from Menarini Silicon Biosystems (MSB), Merrimack, Eli Lilly, Puma Biotechnology, Pfizer, AstraZeneca. He is the named inventor of patent US 8,790,878 B2. D.H.F. which is licensed to MSB and from whom he receives royalties. He holds stock options from OncImmune LLC and InBiomotion, and he serves as a paid advisor for Cepheid, Freenome, CellWorks, Agendia, and CVS Caremark.

Dr. Lænkholm has received research funding from Nanostring Technology (not for personal salary), participated in advisory board for Roche A/S and Novartis (for purely scientific reasons; honoraria declined) and received travel expenses for congress attendance from Astra Zeneca and Roche A/S (past 2 years).

Dr. Nielsen has consulted for Nanostring and received compensation. Dr. Nielsen has intellectual property rights / ownership interests from Bioclassifier LLC.

Dr. Osborne has consulted for Astra Zeneca, Genentech and NanoString and received compensation.

Dr. Penault-Llorca has participated in Scientific Advisory Boards for Nanostring, Myriad, Genomic Health, Agendia, Astrazeneca, Roche, Sanofi, Novartis, Pfizer, BionTech, and received compensation. Dr. Penault-Llorca has received research funding or in kind support from Nanostring, Astrazeneca, Roche, BionTech.

Dr. Van der Vegt has consulted for Philips and received compensation.

All other authors declare no conflict of interest.

Word count: 2495

ABSTRACT

Aims: The nuclear proliferation marker Ki67 assayed by immunohistochemistry has multiple potential uses in breast cancer, but an unacceptable level of inter-laboratory variability has hampered its clinical utility. The International Ki67 in Breast Cancer Working Group has undertaken a systematic program to determine whether Ki67 measurement can be analytically validated and standardized across laboratories. This study addresses whether acceptable scoring reproducibility can be achieved on excision whole sections.

Methods and results: Adjacent sections from 30 primary ER+ breast cancers were centrally stained for Ki67 and sections were circulated among 23 pathologists in 12 countries. All pathologists scored Ki67 by two methods: (a) global: 4 fields of 100 tumor cells each were selected to reflect observed heterogeneity in nuclear staining; (b) hot-spot: the field with highest apparent Ki67 index was selected and up to 500 cells scored. The intraclass correlation coefficient (ICC) for the global method (0.87; 95%CI: 0.799-0.93) marginally met the prespecified success criterion (lower 95%CI ≥ 0.8) while the ICC for the hot-spot method (0.83; 95%CI: 0.74-0.90) did not. Visually, inter-observer concordance in location of selected hot-spots varies between cases. The median times for scoring were 9 and 6 minutes for global and hot-spot methods, respectively.

Conclusions: The global scoring method demonstrates adequate reproducibility to warrant next steps toward evaluation for technical and clinical validity in appropriate cohorts of cases. The time taken for scoring by either method is practical using counting software we are making publicly available. Establishment of external quality assessment schemes is likely to improve the reproducibility between laboratories further.

Keywords: Ki67, immunohistochemistry, pathology, scoring protocol, analytical validity, inter-observer variability, inter-observer reproducibility

INTRODUCTION

The nuclear antigen recognized by the Ki67 antibody is expressed in proliferating cells but absent in resting cells[1]. Since its discovery in 1983 by Gerdes *et al.*,[1] Ki67 assessed by immunostaining has been studied extensively as a prognostic[2-11] and predictive[4,6,9,12,13] marker, predominantly in hormone-receptor positive breast cancer but also in other tumors as well[14-18]. For example, pre-surgical Ki67 has been shown to be a marker for recurrence free survival[19] and in the neoadjuvant setting, a marker for endocrine resistant tumor that may require more aggressive treatment[20]. Excellent *intra*-observer reproducibility under controlled pre-analytic and staining conditions[21] has contributed to the body of evidence showing the potential of Ki67 immunohistochemistry assay to be implemented in hospital laboratories as a cost effective part of clinical management[22-24]. However, poor inter-observer reproducibility and variability due to technical aspects of the assay has limited its adoption in clinical practice[4,9,25-28].

The International Ki67 Working Group (IKWG) has undertaken a systematic multiphase program to determine whether Ki67 *scoring* can be standardized and analytically validated across laboratories[9,21,29,30]. In phase 1, as assessed by the intraclass correlation coefficient (ICC) estimate of inter-observer reproducibility, differences in pathologists' visual interpretation were the main source of variability (ICC = 0.71, 95% credible interval (CI): 0.47–0.78)[21]. In phase 2, greater concordance was achieved, at least on tissue microarrays, when pathologists trained to calibrate and standardize scoring according to a clearly defined methodology (ICC = 0.94, 95% CI: 0.90–0.97)[29]. However, in clinical practice, decisions are made on core-cut biopsy or on excision specimens which require general assessment of the entire sample and selection of areas for formal counting. Therefore, in phase 3A, we assessed whether acceptable performance could be achieved on core-cut biopsies using a standardized method with two distinct methods of scoring field selection: global (four representative fields, counting 100 nuclei each) and hot-spot (one field with highest Ki67, counting 500 nuclei). The global method achieved acceptable inter-observer reproducibility (ICC = 0.87; 95% CI: 0.81–0.93) according to our prespecified criteria, whereas the hot-spot method did not (ICC = 0.84; CI: 0.77–0.92)[30].

This current study represents the final phase (3B) of the visual scoring analytical validity program, wherein we assess whether acceptable performance can be achieved on centrally stained excision whole sections using the scoring method established on core-cut biopsies.

Future studies would be required to evaluate variability due to staining and pre-analytical aspects of the assay.

MATERIALS AND METHODS

This study was approved by the British Columbia Cancer Agency Clinical Research Ethics Board (H10-03420). All specimens used in this study were donated by patients who signed institutionally-appropriate consent forms, were excess to diagnostic requirements and ethically available for quality control studies.

Case selection and sample preparation

Excision blocks from 30 estrogen receptor (ER) positive breast cancer cases were selected: 15 from the phase 3A study[30] and 15 from Kawasaki Medical School Hospital, Kurashiki, Japan (Supplemental Figure 1). Case selection was irrespective to patients' age at diagnosis, tumor grade, size or nodal status. The clinicopathological characteristics of these 30 cases are shown in Supplemental Table 1. All blocks were 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, Arizona, USA) according to criteria established by the IKWG[9]. Sections from the same block were stained in a single immunohistochemistry run except for four cases where the staining was done in two different runs. This approach effectively controls for any technical variation in staining.

Sample distribution

Twenty-four volunteer pathologists representing 24 institutions from 12 countries, most of whom participated in the phase 3A study, were invited to participate.

Six adjacent sections from each of the 30 excision blocks were centrally stained: the first with H&E, the second with p63 (myoepithelial marker, to assist the identification of invasive foci) and the third to sixth with Ki67 (designated as slide sets 1–4). To facilitate application to the general histopathology laboratory environment, physical glass slides (as opposed to virtual slide images) were distributed to the volunteer pathologists. Because the accumulated delays

required, if all pathologists reviewed the same physical glass slides, would have made the study impractical, participating pathologists were divided into four groups and were given one of the four sets of Ki67 slides to score. The H&E and p63 reference slides were made available online as digital images. Twenty-three pathologists successfully completed the study.

Scoring protocol

All pathologists were specifically trained to score Ki67 with emphasis on having a very low threshold for appreciating "brown stain" and the principles of standardized regions for nuclei counting, through the publicly available proficiency training module (http://www.gpec.ubc.ca/calibrator) that was initially used in the phase 2 study[29]. The detailed scoring protocol is found in supplemental document: "ki67p3b_scoring_protocol.pdf". A modified version of the scoring software used in this study is available freely from the Google Play and Apple iTunes store (search term: "Ki67").

Scoring methods

The scoring methods used are the same ones that were employed in the phase 3A study[30]: 1) 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; 2) an unweighted global assessment; and 3) assessment of Ki67 only in a "hot-spot" area.

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 fashion, but the final Ki67 score was derived differently. Adapted from a scoring protocol that has been used routinely in the Dowsett laboratory[31], 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 staining frequencies. Based on these estimates, an algorithm (Supplemental Figure 2) dictates the required number of fields to select and score for each Ki67 staining frequency (irrespective of staining intensity; totaling up to four fields). This algorithm was designed such that the four (or less) selected scoring fields would capture the full range of staining frequencies while at the same time, be reflective of the proportion in staining

frequencies heterogeneity. Up to 100 invasive tumor nuclei within each field are counted using a "typewriter" pattern (Supplemental Figure 3), similar to how a tissue microarray core was scored in the phase 2 study[29].

The hot-spot method requires the pathologist to visually select one high-power field with the highest apparent staining rate and, within that area only, count up to 500 invasive tumor nuclei in a "typewriter" pattern.

Statistical analyses

Pre-specified criterion for success

Prior to data collection it was hypothesized that at least one of the scoring methods would have an associated ICC statistically greater than 0.80 (ICC of 0.8 being considered as good concordance[32]). For planning purposes, power calculations 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 at least 21 participating pathologists scoring 30 cases, there would be 80% power to exclude ICCs lower than the pre-specified ICC of 0.8 from a 95% credible interval for a given scoring method.

Ki67 score

The Ki67 score was defined as in the phase 3A study[30]. Positive staining was defined as any brown stain in the nucleus above background, with reference available as needed to provide standard sample images; negative staining was scored when an 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. 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. As in our previous studies, to satisfy model assumptions of normality and constant variance, for statistical analyses the Ki67 score is converted to a logarithmic scale by adding 0.1% and applying a log base 2 transformation.

ICC estimates (ranging from 0 to 1, with 1 representing perfect reproducibility) were computed as previously reported in the phase 3A study[30]. Briefly, variance component analyses were performed to quantify the contributions from the following sources of variability: scoring pathologist (observer), patient tumor (biological variation – each excision block represents a unique patient) and section of the excision block. Similar to the phase 3A study, same-section and different-section ICCs were computed. Same-section refers to pathologists scoring the same excision whole section physical slides, while different-section refers to pathologists scoring different physical slides that represent serial sections cut from the same original excision blocks. Credible intervals 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 3.3.2[33]. 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.

RESULTS

ICC of Ki67 according to scoring method.

The different-section ICC estimate for the weighted global scores was 0.87 (95%CI: 0.799–0.93), at the margin of the pre-specified success criterion (lower bound of credible interval exceeding 0.8) (Table 1). The different-section ICCs for the unweighted global scores and hot-spot scores were 0.86 (95%CI: 0.793–0.92) and 0.83 (95 %CI: 0.74–0.90), respectively, and therefore both these methods had ICC credible intervals that extended below the success criterion at the lower 95% limit. The corresponding same-section ICC estimates for the weighted global, unweighted global and hot-spot scores were virtually identical 0.87 (95% CI: 0.799–0.92), 0.86 (95% CI: 0.79–0.92) and 0.83 (95% CI: 0.74–0.90) respectively, supporting that differences between serial sections were minimal. Figure 1 displays the side-by-side boxplots of Ki67 scores across pathologists (hereafter referred to as “observers”) by group. Summary statistics for the Ki67 scores across the 23 observers are given in Supplemental Tables 2 to 4.

The median number of nuclei counted per slide (across all observers and cases) is 400 and 500 for the global and hot-spot methods respectively. The corresponding minimum number of nuclei counted is 300 and 138. Eighteen percent of the hot-spot scores were based on <500 nuclei counts. Among these 126 hot-spot scores, the median number of nuclei counted is 375.

In a context where preanalytical and staining factors are held constant, variance component analyses show that, regardless of scoring method, biological variation among different patients was the largest component of the total variation on these centrally stained slides, indicating that the Ki67 score is reflecting inherent properties of the tumor (Figure 2, Supplemental Table 5).

Inter-observer variation of Ki67 scoring.

Figure 3 displays the variation in scores across observers for cases in slide set 1 as spaghetti plots. The corresponding plots for slide set 2-4 are displayed in Supplemental Figure 4. Figure 4 presents the scores in a heat map format with the columns (observers) ordered (within each slide set) by the median scores across cases and the rows (cases) sorted by the median scores across observers.

Overall it can be seen that most observers show good parallelism in the increasing Ki67 scores across the plots. In other words, observers measuring higher or lower than others tended to do so relatively consistently.

Categorical concordance of Ki67 scoring.

Regarding concordance on a categorical level (<10%, 10-20% and >20%), the relationship between concordance and continuous score is shown in Supplemental Figure 5. It shows excellent to perfect concordance on cases with scores that are either much lower or higher than the intermediate range (10-20%).

Based on visual inspection of captured images, locations of the hot-spot selections tended to cluster in the same region across observers within each of the excision whole section slides (Figure 5 shows some examples; virtual slide images of all slides used in this study and the

corresponding selected fields and scores can be viewed at http://www.gpec.ubc.ca/papers/ki67p3b).

The median scoring time (field selection and nuclear counting) was 9 (interquartile range: 7-11) and 6 (interquartile range: 4-8) minutes for global and hot-spot methods, respectively.

DISCUSSION

The IKWG has demonstrated that it is possible, when controlling stringently for variability due to preanalytical and analytical aspects of the Ki67 immunohistochemistry assay[9], and given a set of clearly defined training exercise and scoring instructions, for pathologists to achieve high inter-observer concordance in Ki67 scoring on core-cut biopsies and now on excision whole sections using a conventional light microscope and manual field selection, with no additional aid such as counting grid.

Due to the limited sample size, we were unable to assess whether any specific method (weighted global, unweighted global or hot-spot) is significantly more reproducible than others. However, the observed ICCs for global score (weighted: 0.87; unweighted: 0.86) are relatively higher compared to hot-spot score (0.83) suggesting that a sufficiently powered study might be able to show more convincingly whether global scores are more reproducible. This result is consistent with findings on core biopsies[30].

Can this level of concordance be clinically adequate? The POETIC[11] study assessed Ki67 (cut point at 10%) as a prognostic marker. Applying this cut point to the data in our current study, 17 (out of 30) cases have at most one discordance in weighted global score (Figure 4a). There are cases with major discrepancies: TB036, on the same physical slide (set 2), received a weighted global score of 4% and of 21% from observer A and L respectively. However, it is apparent (Figure 4) that cases far away from the intermediate range (10-20%) tend to have good agreement. Considering that cases in our current study are a random sampling of the general ER+ breast cancer population, one could expect that about half of these cases would fall away from the intermediate range and hence Ki67 may provide clinically adequate information, provided that the staining and pre-analytical factors do not add too much variability.

Are the proposed scoring methods practical? The median scoring time is 6-9 minutes depending on the method used. However, an adaptive scoring protocol can be used to reduce scoring time if the purpose is to assess whether Ki67 is above or below a specific cut point. For example, considering the global scoring method, where the maximum nuclei count is pre-specified (i.e. 400), to determine whether a case has unweighted global score ≥ 10%, the pathologist can stop counting if the first field he/she scored is ≥ 40%. For cases with very low Ki67 score, one would likely still need to count all 400 nuclei.

The proposed scoring protocols do not make any recommendation concerning the required minimum tumor nuclei count. This is a limitation of this study and in practice, it will be up to the discretion of the scoring pathologist to assess if too few tumor nuclei are available for an adequate Ki67 assessment. This will depend on the percentage of positive cells scored in the cells available and the clinical context for the measurement.

External quality assessment program (e.g. NordiQC[34]), involving comparing laboratory scores with reference scores in periodic assessment challenges, will likely improve inter-observer reproducibility further. Recent studies suggest that an even higher level of concordance can be achieved with automated image analysis[35-38]. The IKWG is actively conducting studies in this area to assess how artificial intelligence may help standardize Ki67 assessment[35,38]. Also, concordance between Ki67 scores on core biopsies and excision specimens is currently being investigated.

In conclusion, this study demonstrates an adequately high level of inter-observer concordance can be achieved by visual assessment of Ki67 using practical scoring methods, although some cases with large discrepancies remain. A two-tier assessment approach may be worthy of further study as a means to reduce scoring burden and further address challenging cases: if the Ki67 value from the initial scoring falls on a grey zone (e.g., cut point +/- 5%), scoring by a second pathologist or alternative test could be pursued. Preanalytical and analytical aspects of the immunohistochemistry assay, areas that still need standardization before the clinical utility of this marker can be proven, will likely add more variability. A clinical validation study employing analytically reproducible methodology would also need to be completed in

appropriate cohorts of cases to determine whether Ki67 can be recommended for patient care decisions.

ACKNOWLEDGEMENTS

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

CONTRIBUTIONS

Samuel C.Y. Leung: study design, data collection, statistical analysis, manuscript drafting & review

Torsten O. Nielsen: study design, manuscript drafting & review

Lila A. Zabaglo: study design, collection and preparation of samples, data collection, manuscript drafting & review

Indu Arun: study design, data collection, manuscript drafting & review

Sunil S. Badve: study design, data collection, manuscript drafting & review

Anita L. Bane: study design, data collection, manuscript drafting & review

John M.S. Bartlett: study design, manuscript drafting & review

Signe Borgquist: study design, manuscript drafting & review

Martin C. Chang: study design, data collection, manuscript drafting & review

Andrew Dodson: study design, collection and preparation of samples, manuscript drafting & review

Anna Ehinger: study design, data collection, manuscript drafting & review

Susan Fineberg: study design, data collection, manuscript drafting & review

Cornelia M. Focke: study design, data collection, manuscript drafting & review

Dongxia Gao: study design, data collection, manuscript drafting & review

Allen M. Gown: study design, data collection, manuscript drafting & review

Carolina Gutierrez: study design, data collection, manuscript drafting & review

Judith C. Hugh: study design, data collection, manuscript drafting & review

Zuzana Kos: study design, data collection, manuscript drafting & review

Anne-Vibeke Lænkholm: study design, data collection, manuscript drafting & review

Mauro G. Mastropasqua: study design, data collection, manuscript drafting & review

Takuya Moriya: study design, data collection, manuscript drafting & review

Sharon Nofech-Mozes: study design, data collection, manuscript drafting & review

C. Kent Osborne: study design, manuscript drafting & review

Frédérique M. Penault-Llorca: study design, data collection, manuscript drafting & review

Tammy Piper: study design, data collection, manuscript drafting & review

Takashi Sakatani: study design, data collection, manuscript drafting & review

Roberto Salgado: study design, data collection, manuscript drafting & review

Jane Starczynski: study design, data collection, manuscript drafting & review

Tomoharu Sugie: study design, manuscript drafting & review

Bert van der Vegt: study design, data collection, manuscript drafting & review

Giuseppe Viale: study design, manuscript drafting & review

Daniel F. Hayes: study design, manuscript drafting & review

Lisa M. McShane: study design, statistical analysis, manuscript drafting & review

Mitch Dowsett: study design, manuscript drafting & review

FUNDING

This work was supported by a generous grant from the Breast Cancer Research Foundation.

REFERENCES

(1) Gerdes J, Schwab U, Lemke H, Stein H. Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. Int J Cancer 1983 Jan 15;31(1):13-20.

(2) Luporsi E, Andre F, Spyratos 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 Apr;132(3):895-915.

(3) de Azambuja E, Cardoso F, de Castro G,Jr et al. Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. Br J Cancer 2007 May 21;96(10):1504-1513.

(4) Denkert C, Budczies J, von Minckwitz G, Wienert S, Loibl S, Klauschen F. Strategies for developing Ki67 as a useful biomarker in breast cancer. Breast 2015 Aug 14.

(5) 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 Jun;139(2):539-552.

(6) Viale G, Regan MM, Maiorano E 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 2007 Sep 1;25(25):3846-3852.

(7) Viale G, Regan MM, Mastropasqua MG 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 2008 Feb 6;100(3):207-212.

(8) Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. Lancet Oncol 2010 Feb;11(2):174-183.

(9) 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 Nov 16;103(22):1656-1664.

(10) 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 2015 Sep 4.

(11) Robertson JFR, Dowsett M, Bliss JM et al. Peri-operative aromatase inhibitor treatment in determining or predicting longterm outcome in early breast cancer - The POETIC Trial. San Antonio Breast Cancer Symposium 2017 presented December 6, 2017;abstract GS1-03.

(12) Criscitiello C, Disalvatore D, De Laurentiis M 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 2014 Feb;23(1):69-75.

(13) Cohen AL, Factor RE, Mooney K et al. POWERPIINC (PreOperative Window of Endocrine TheRapy Provides Information to Increase Compliance) trial: Changes in tumor proliferation index and quality of life with 7 days of preoperative tamoxifen. Breast 2017 Feb;31:219-223.

(14) Lei Y, Li Z, Qi L et al. The Prognostic Role of Ki-67/MIB-1 in Upper Urinary-Tract Urothelial Carcinomas: A Systematic Review and Meta-analysis. J Endourol 2015 Jul 23.

(15) Desouki MM, Chamberlain BK, Li Z. The role of immunohistochemistry in the evaluation of gynecologic pathology part 2: a comparative study between two academic institutes. Ann Diagn Pathol 2015 Jun 11.

(16) He Y, Wang N, Zhou X et al. Prognostic value of ki67 in BCG-treated non-muscle invasive bladder cancer: a meta-analysis and systematic review. BMJ Open 2018 Apr 17;8(4):e019635-2017-019635.

(17) Richardsen E, Andersen S, Al-Saad S et al. Evaluation of the proliferation marker Ki-67 in a large prostatectomy cohort. PLoS One 2017 Nov 15;12(11):e0186852.

(18) Xie Y, Chen L, Ma X et al. Prognostic and clinicopathological role of high Ki-67 expression in patients with renal cell carcinoma: a systematic review and meta-analysis. Sci Rep 2017 Mar 13;7:44281.

(19) Dowsett M, Smith IE, Ebbs SR et al. Prognostic value of Ki67 expression after short-term presurgical endocrine therapy for primary breast cancer. J Natl Cancer Inst 2007 Jan 17;99(2):167-170.

(20) Ellis MJ, Suman VJ, Hoog J et al. Ki67 Proliferation Index as a Tool for Chemotherapy Decisions During and After Neoadjuvant Aromatase Inhibitor Treatment of Breast Cancer: Results From the American College of Surgeons Oncology Group Z1031 Trial (Alliance). J Clin Oncol 2017 Apr 1;35(10):1061-1069.

(21) Polley MY, Leung SC, McShane LM et al. An international Ki67 reproducibility study. J Natl Cancer Inst 2013 Dec 18;105(24):1897-1906.

(22) Iwamoto T, Katagiri T, Niikura N et al. Immunohistochemical Ki67 after short-term hormone therapy identifies low-risk breast cancers as reliably as genomic markers. Oncotarget 2017 Apr 18;8(16):26122-26128.

(23) Thakur SS, Li H, Chan AMY et al. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS One 2018 Jan 5;13(1):e0188983.

(24) Reinert T, Goncalves R, Ellis MJ. Current Status of Neoadjuvant Endocrine Therapy in Early Stage Breast Cancer. Curr Treat Options Oncol 2018 Apr 16;19(5):23-018-0538-9.

(25) Laenkholm AV, Grabau D, Moller Talman ML et al. An inter-observer Ki67 reproducibility study applying two different assessment methods: on behalf of the Danish Scientific Committee of Pathology, Danish breast cancer cooperative group (DBCG). Acta Oncol 2018 Jan;57(1):83-89.

(26) Focke CM, Burger H, van Diest PJ et al. Interlaboratory variability of Ki67 staining in breast cancer. Eur J Cancer 2017 Oct;84:219-227.

(27) Mengel M, von Wasielewski R, Wiese B, Rudiger T, Muller-Hermelink HK, Kreipe H. Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial. J Pathol 2002 Nov;198(3):292-299.

(28) Ekholm M, Grabau D, Bendahl PO et al. Highly reproducible results of breast cancer biomarkers when analysed in accordance with national guidelines - a Swedish survey with central re-assessment. Acta Oncol 2015 Jul;54(7):1040-1048.

(29) Polley MY, Leung SC, Gao D et al. An international study to increase concordance in Ki67 scoring. Mod Pathol 2015 Jun;28(6):778-786.

(30) Leung SCY, Nielsen TO, Zabaglo L et al. Analytical validation of a standardized scoring protocol for Ki67: phase 3 of an international multicenter collaboration. NPJ Breast Cancer 2016 May 18;2:16014.

(31) Zabaglo L, Salter J, Anderson H et al. Comparative validation of the SP6 antibody to Ki67 in breast cancer. J Clin Pathol 2010 Sep;63(9):800-804.

(32) Kirkegaard T, Edwards J, Tovey S et al. Observer variation in immunohistochemical analysis of protein expression, time for a change? Histopathology 2006 Jun;48(7):787-794.

(33) R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.

(34) Vyberg M, Møller J, Røge R. Nordic immunohistochemical Quality Control - Ki67 assessment. 2018; Available at: http://www.nordiqc.org/epitope.php?id=1, 2018.

(35) Acs B, Pelekanou V, Bai Y et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. Lab Invest 2018 Sep 4.

(36) Stalhammar G, Robertson S, Wedlund L et al. Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. Histopathology 2018 May;72(6):974-989.

(37) Koopman T, Buikema HJ, Hollema H, de Bock GH, van der Vegt B. Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. Breast Cancer Res Treat 2018 May;169(1):33-42.

(38) Rimm DL, Leung SCY, McShane LM et al. An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer. Mod Pathol 2018 Aug 24.

TABLE

Table 1. Summary of ICC values for different scoring methods.

	Different-section ICC	Same-section ICC
Weighted global	0.87 (95%CI: 0.799–0.93)	0.87 (95% CI: 0.799–0.92)
Unweighted global	0.86 (95%CI: 0.79–0.92)	0.86 (95% CI: 0.79–0.92)
Hot-spot	0.83 (95 %CI: 0.74–0.90)	0.83 (95% CI: 0.74–0.90)

FIGURE LEGENDS

Figure 1. Ki67 scores of all 23 observers (by slide set). Observers are ordered (within each group) by the median scores. The bottom/top of the box in each box plot represent the first

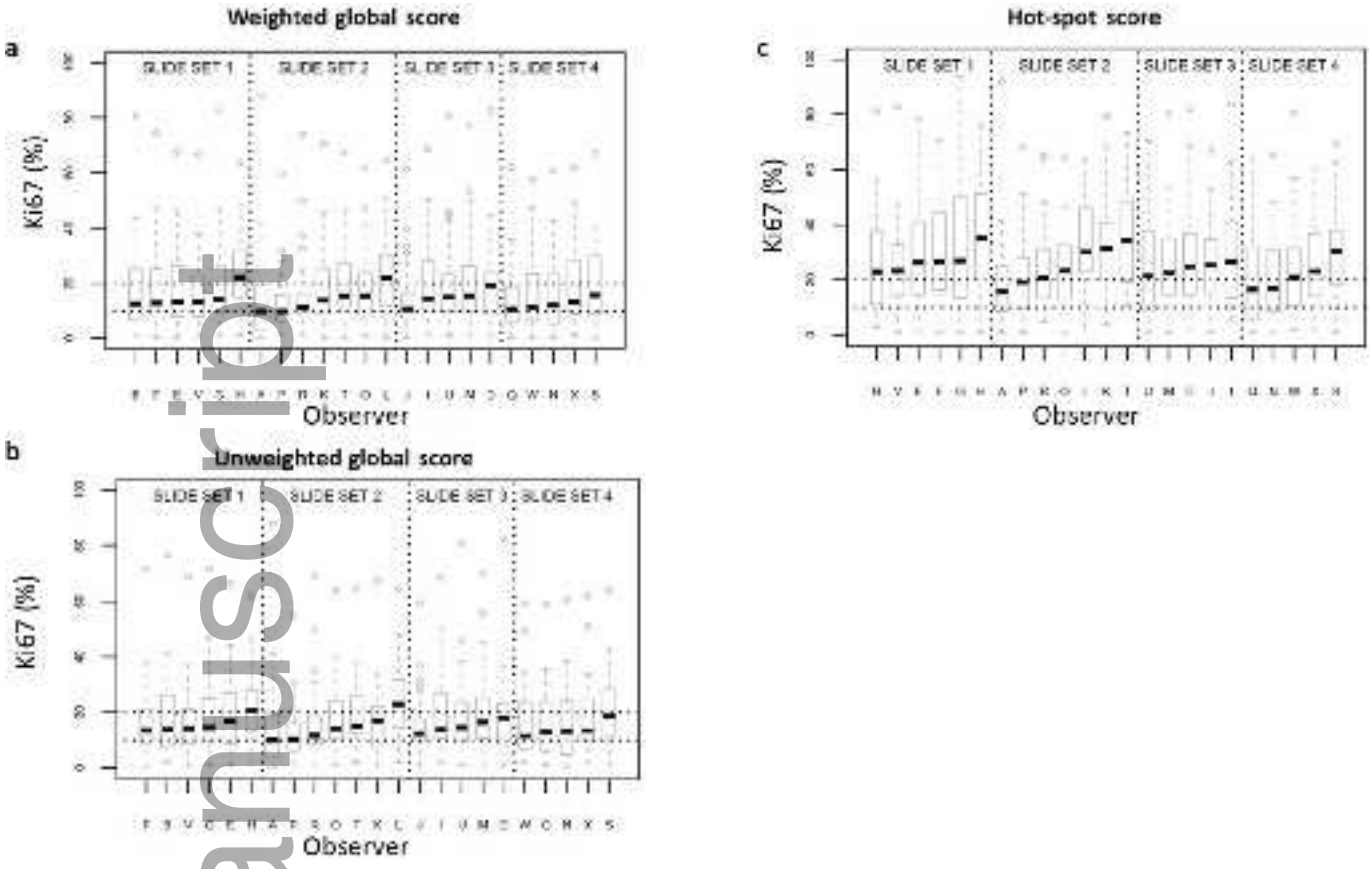
(Q1)/third (Q3) quartiles, the bold line inside the box represents the median and the two bars outside the box represent the lowest/highest datum still within 1.5 × the inter-quartile range (Q3-Q1). Outliers are represented with empty circles.

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

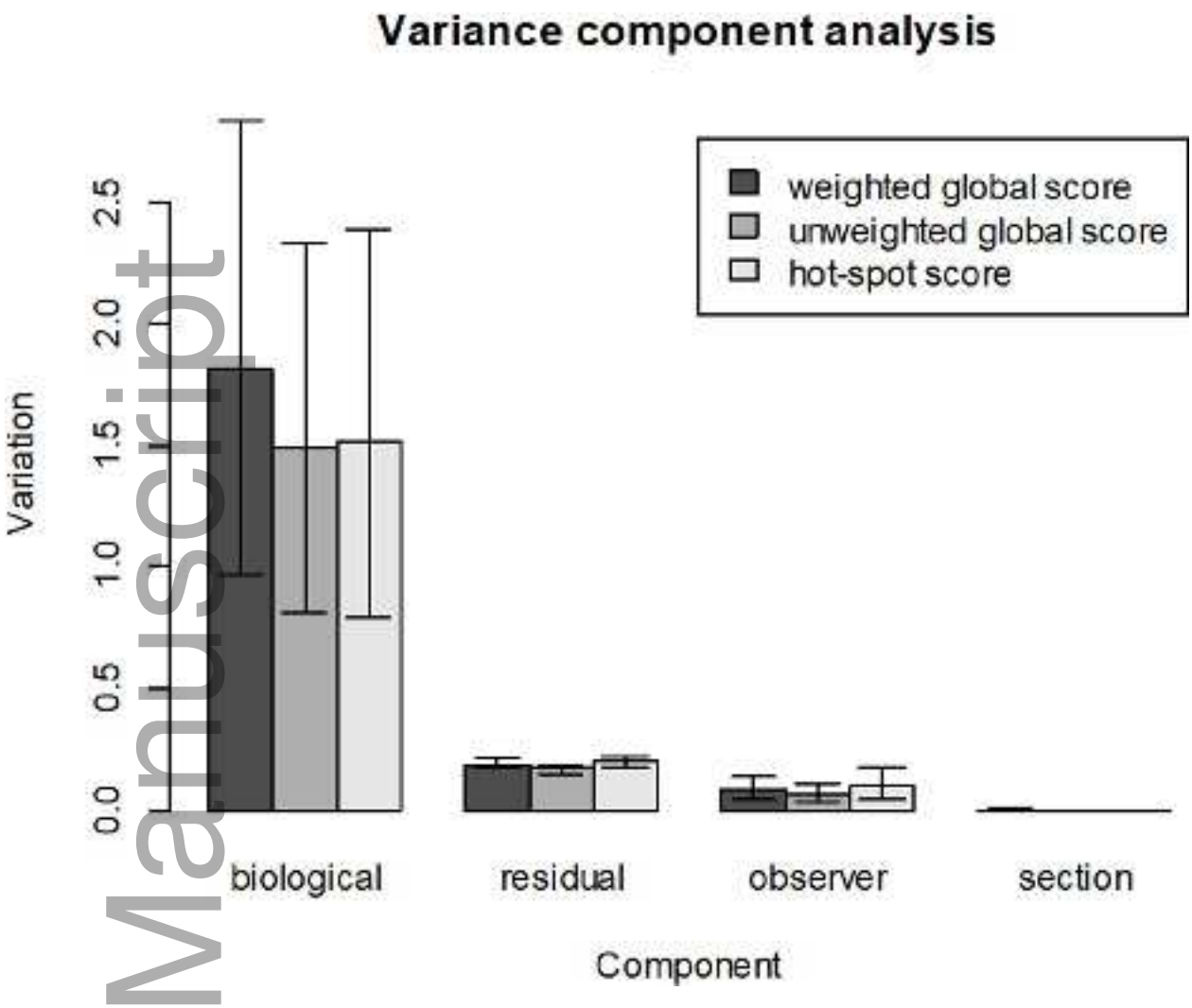
Figure 3. Variability in Ki67 scores (slide set 1 only). Each line represents Ki67 scores from one observer. Shaded region indicates Ki67 scores between 10-20%. Scores on slide set 2-4 are shown in Supplemental Figure 4.

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

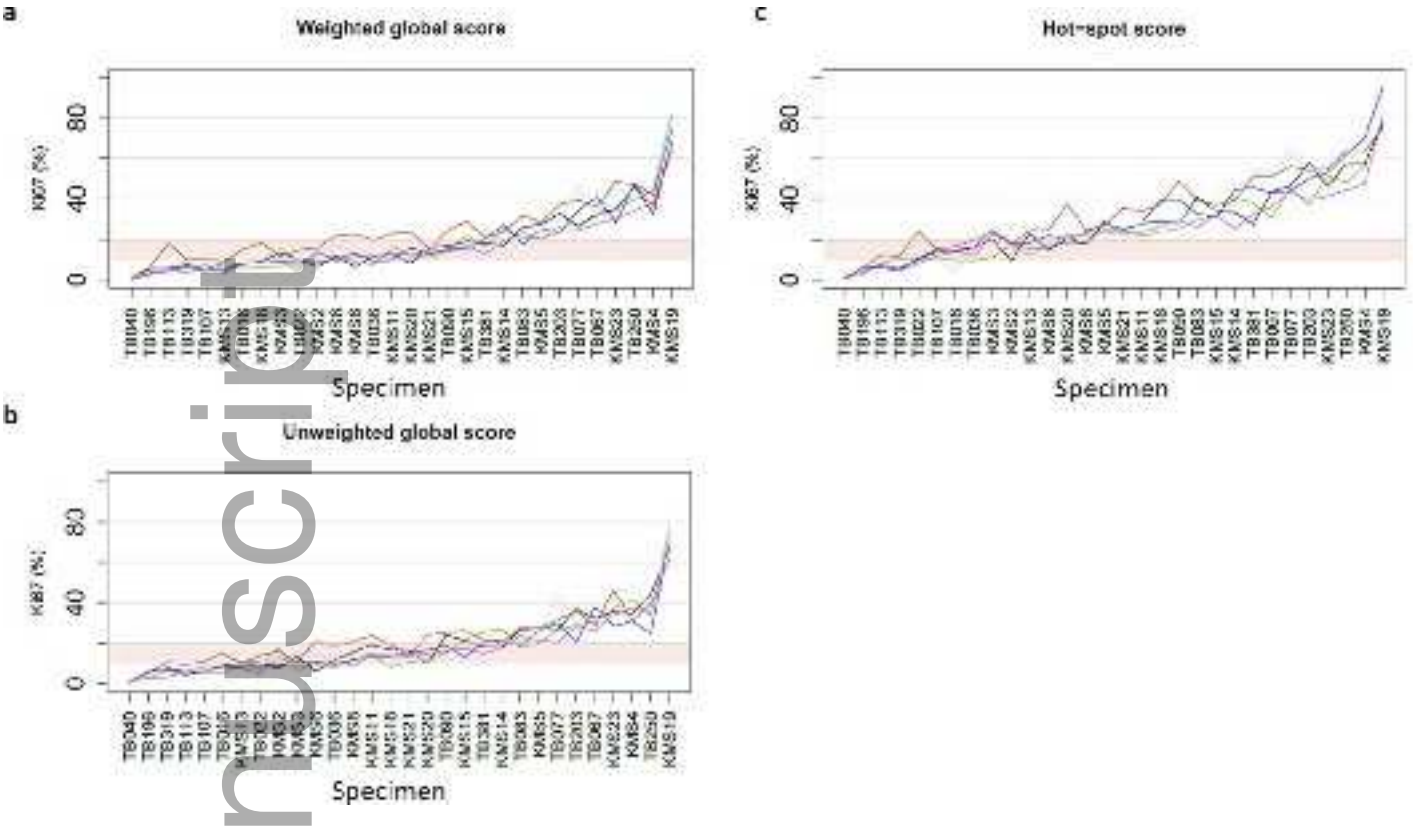
Figure 5. Hot-spot field selection by different observers on the same excision whole section slide. Figure 5a shows selections (indicated by red circles) on some example excision whole section slides. Figure 5b is an example of a single excision whole section slide (median score: 18%) with zoomed-in fields. Each observer was asked to circle the area considered to be the hot spot (b-i). Most observers honed in on the same general area of the slide, although individual selected scoring fields do not always overlap. Figure 5b-iii and 5b-iv represent segments of the same area chosen by two different observers to read Ki67. Figure 5b-v represents the "outlier" field selected by only one observer as the hot-spot.



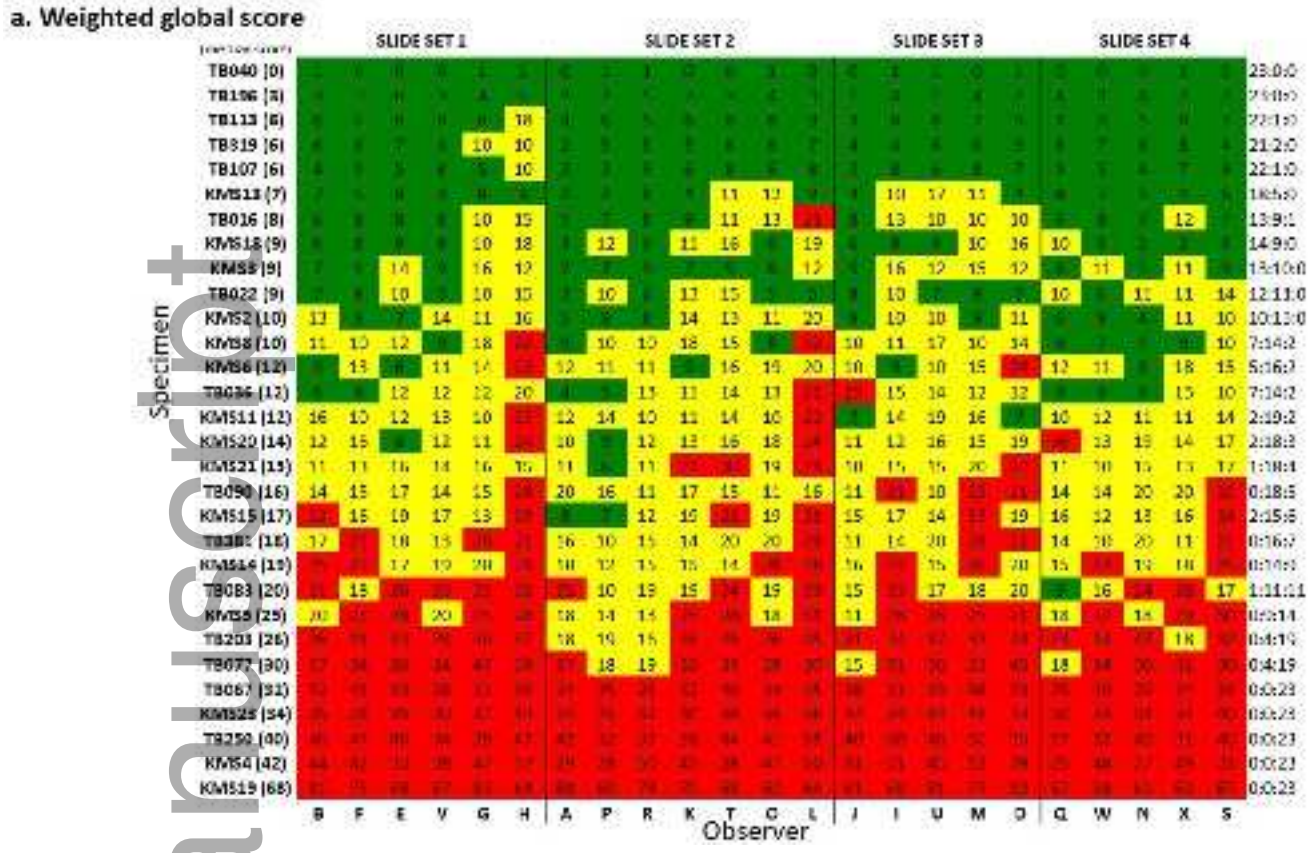
his_13880_f1.jpg



his_13880_f2.jpg

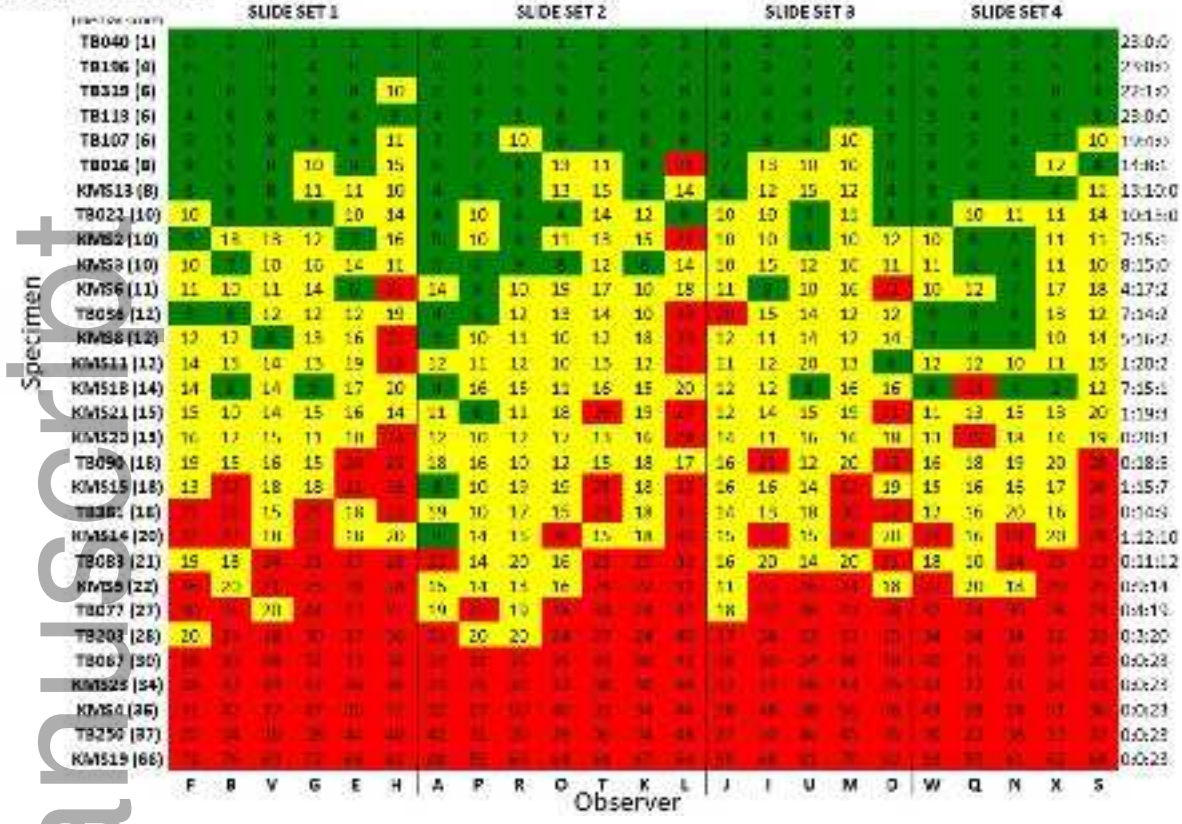


his_13880_f3.jpg



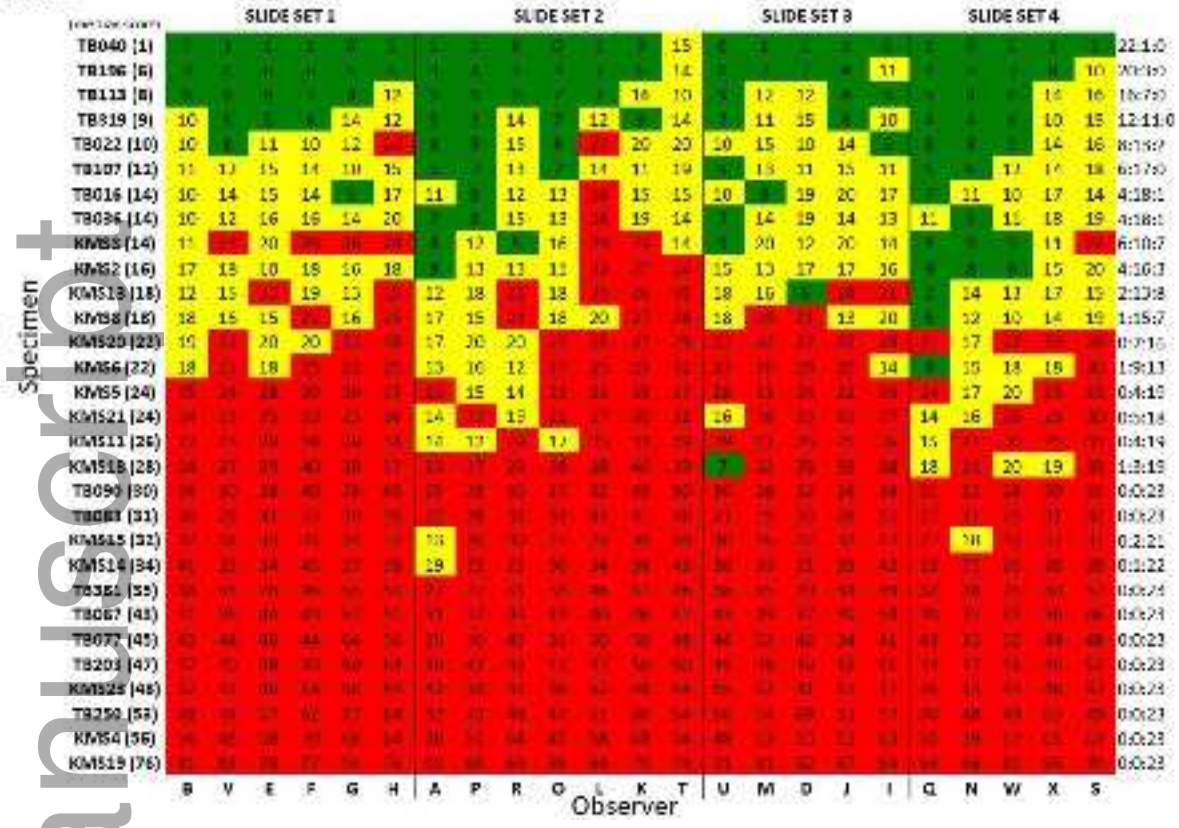
his_13880_f4a.jpg

b. Unweighted global score



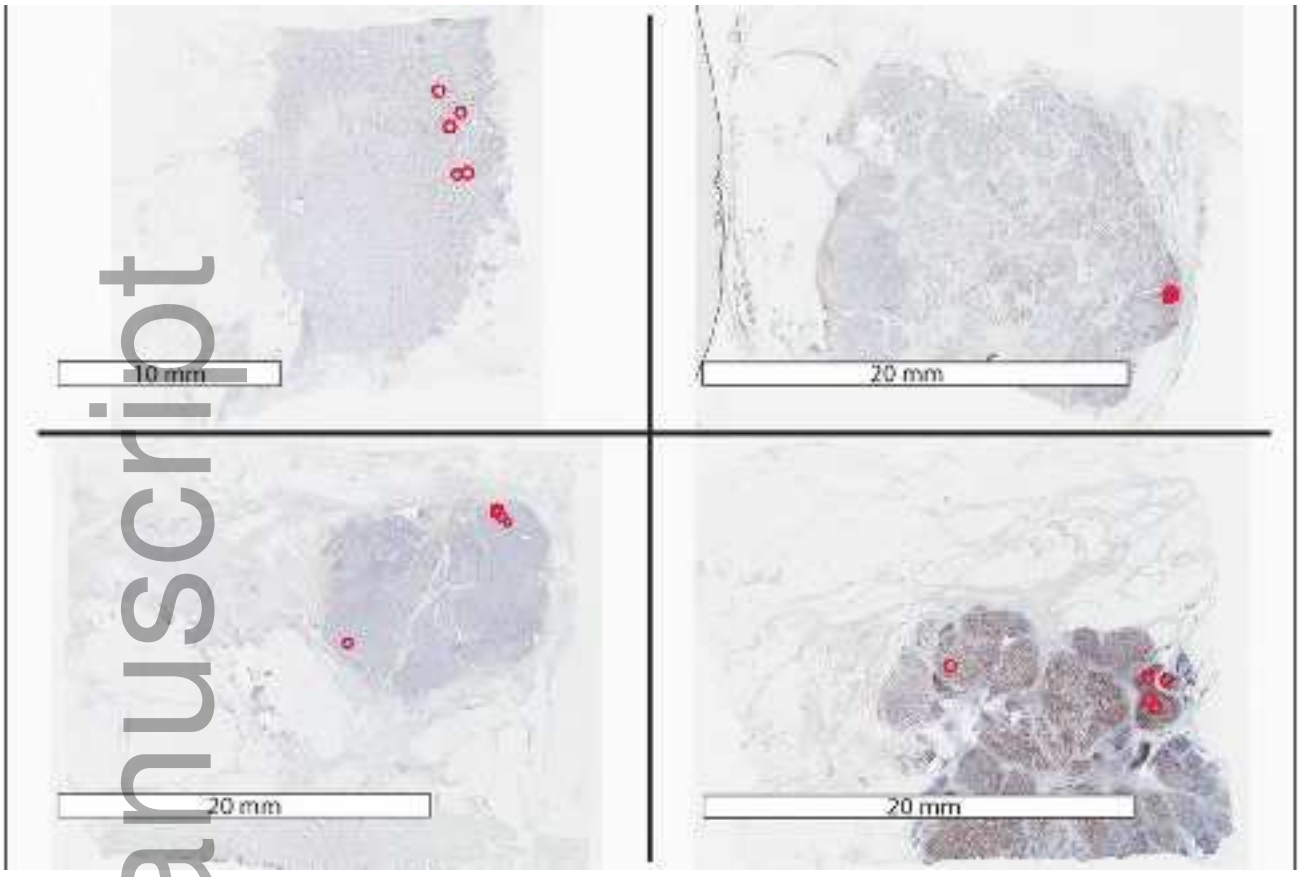
his_13880_f4b.jpg

c. Hot-spot score

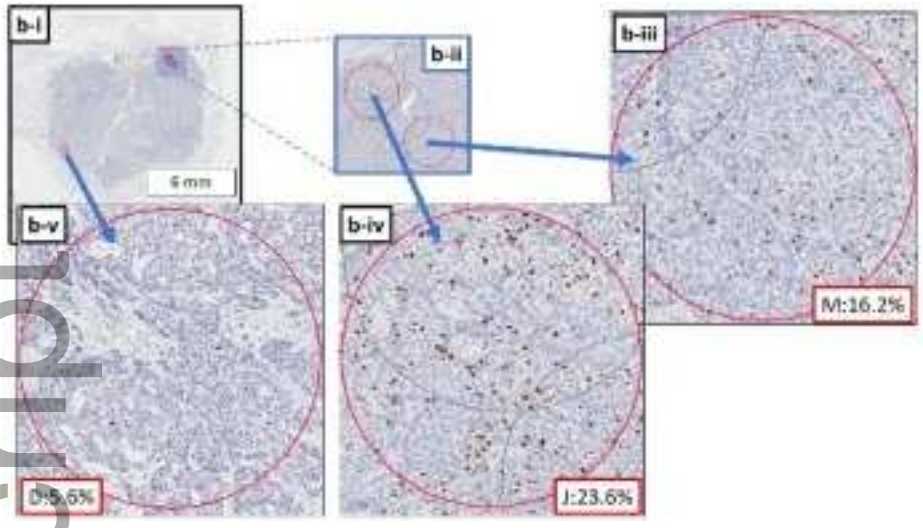


his_13880_f4c.jpg

a



his_13880_f5a.jpg



his_13880_f5b.jpg