Various Aspects of Ki-67 Immunohistochemistry Determined Proliferation in Breast Cancer

Ph.D. Thesis

András Vörös, M.D.

Supervisor:

Prof. Gábor Cserni, M.D., D.Sc.

Department of Pathology

University of Szeged

Szeged, Hungary

Szeged

2014

LIST OF FULL PAPERS THAT SERVED AS THE BASIS OF THE PH.D. THESIS

I. **Vörös A**, Csörgő E, Nyári T, Cserni G.
An intra- and interobserver reproducibility analysis of the Ki-67 proliferation marker assessment on core biopsies of breast cancer patients and its potential clinical implications.
Pathobiology 2013; 80:111-8.
IF: 1.948

II. **Vörös A**, Csörgő E, Kővári B, Lázár P, Kelemen Gy, Cserni G.
The use of digital images improves reproducibility of the Ki-67 labeling index as a proliferation marker in breast cancer.
Pathology & Oncology Research 2013 Nov 8. [Epub ahead of print]
IF: 1.555

III. Cserni G, **Vörös A**, Liepniece-Karele I et al.
Distribution pattern of the Ki67 labeling index in breast cancer and its implications for choosing cut-off values.
The Breast 2014; 23:259-63.
IF: 1.967

IV. **Vörös A**, Csörgő E, Kővári B, Lázár P, Kelemen Gy, Nyári T, Cserni G.
Different methods of pretreatment Ki-67 labeling index evaluation in core biopsies of breast cancer patients treated with neoadjuvant chemotherapy and their relation to response to therapy.
Pathology & Oncology Research 2014 May 25. [Epub ahead of print]
IF: 1.555

1. INTRODUCTION

Proliferation is an important feature of malignant tumors, including breast cancer. Various methods have been established for its quantification, but Ki-67 immunohistochemistry is the standard proliferation assay due to its simplicity and wide availability. This protein is expressed in all phases of the cell cycle, except G0. This feature makes it the best marker to be detected by immunohistochemistry in different malignant tumors including breast carcinoma. Proliferating tumor cells show positive nuclear reaction with anti-Ki-67 antibodies. The pathologic report generally refers to the percentage of positive tumor cells (Ki-67 labeling index, LI).

Different studies have demonstrated that a high Ki-67 LI indicates an increased risk of recurrence, metastasis and faster progression of the disease. Because of this prognostic effect, Ki-67 LIs have been used to tailor patients' adjuvant treatment. There are different previously proposed cut-off values of Ki-67 LI which have been recommended for the indication of chemotherapy for certain patients suffering of breast cancer. Ki-67 expression has not only prognostic value, but may also be predictive of the response to neoadjuvant therapy as suspected by some authors. In these studies, better response (complete pathological response / pCR or complete clinical response / cCR) has been reported in breast cancers with higher proliferative activity treated with neoadjuvant chemotherapy. The predictive value of this marker has not yet been verified. Recent recommendations (including St. Gallen International Breast Cancer Conference, 2013) contain Ki-67 evaluation as the part of the routine diagnostic procedures although its predictive potential has not yet been proven unequivocally in case of neoadjuvant chemotherapy.

Owing to the clinical relevance of Ki-67 LI mentioned above, the reproducibility of its evaluation is of obvious importance. The relatively subjective determination of Ki-67 LI can lead to a high degree of interobserver variability. There are pathologists who estimate the percentage of nuclear staining, whereas others count the number of positive cells in different fields of tumor area. The protein has a differential expression during the cell cycle, peaking during the M phase, and being much lower during G1 and S phases. Some of the labeled cells undergo apoptosis, whereas others remain in one of the gaps (G1 or G2), and still others progress to mitosis. This differential expression is one limiting factor in assessing the

proportion of immunostained cells, as the intensity the evaluating pathologist or software will include as positive is variable.

Beside variations of expression intensity mentioned above, there are also other factors that may influence the evaluation, such as tissue fixation or tissue processing. Tumor areas may show significant heterogeneity of proliferation and Ki-67 labeling. While it is common to see higher proliferation rates at the periphery of breast cancers, even the periphery may have an uneven distribution of positive cells, or areas anywhere in the tumor may proliferate more than others, leading to the appearance of so called "hot spots". There is no consensus whether Ki-67 staining evaluation should only consider the hot spots if present, just include them in a general count or avoid them altogether. Human or technical performance may also generate interobserver variability.

At this time there are no standardized methods for the elimination of the different previously mentioned factors influencing the Ki-67 LI, although efforts have been made towards standardization.

2. AIMS

2.1. To analyze the reproducibility of the Ki-67 expression levels by types of antibody and investigators in core-biopsy samples of breast cancer patients.

2.2. To investigate how the use of a standardized, partially digitalized counting method could affect reproducibility of determining the Ki-67 LI.

2.3. To evaluate the distribution of Ki-67 LIs in breast carcinomas diagnosed at different institutions by different pathologists using the method reflecting their daily practice.

2.4. To compare different therapeutic response categories with different Ki-67 evaluation methods in the cases previously analyzed for reproducibility. Within the frames of this latter analysis, to explore whether there was a Ki-67 evaluation method (among the previously used ones) which could be particularly recommended (due to its superiority) or rejected (due to its inferiority) in daily practice, and to look for a potential cut-off value which could separate tumors with high or low proliferation activity before neoadjuvant therapy.

3. PATIENTS AND METHODS

3.1 REPRODUCIBILITY AND CORRELATION ANALYSIS OF THE KI-67 EXPRESSION LEVELS BY TYPES OF ANTIBODY AND INVESTIGATORS

Core-biopsy samples of patients with operable T2≥3 cm or T3-4 and/or N1-2 and M0 breast cancer candidate for neoadjuvant docetaxel-epirubicin with/without capecitabine chemotherapy were retrospectively analyzed. Samples had been taken between January 2003 and December 2011, at the Department of Radiology, University of Szeged or Bács-Kiskun County Teaching Hospital. The tumor samples were fixed in buffered formalin and embedded in paraffin. Samples were routinely stained with hematoxylin and eosin (HE) and routinely immunostained for estrogen receptor (ER), progesterone receptor (PR), HER-2 and topoisomerase II-alpha. For the purpose of this study they were immunostained for Ki-67 with the following 3 antibodies: SP6 (monoclonal rabbit antibody, Hisztopatologia Kft., Pécs, Hungary), B56 (monoclonal mouse antibody, Hisztopatologia Kft., Pécs, Hungary) and MIB-1 (monoclonal mouse antibody, Dako, Glostrup, Denmark). Wet antigen retrieval consisted of pretreatment of all samples in microwave oven in a citrate buffer with pH6 for 30, 30 and 50 minutes in case of MIB-1, B56 and SP6, respectively. All antibodies were diluted at 1:100. Expression of Ki-67 was determined using Dako EnVision FLEX/HRP, DAB+ Chromogen (Dako, Glostrup, Denmark).

All samples were assessed independently by 3 pathologists at high-power magnification (X400). The proportion of Ki-67 positive cells was established with 5% accuracy; therefore only values ending with 5 or 0 were recorded. Each observer was asked to use his/her daily evaluation approach to quantify the proportion of Ki-67 positive cells and to perform the evaluation of all cases stained with one antibody at first, followed by all cases stained with the second and third antibody in order to avoid bias arising from remembering the LI of a given sample. The assessments were done twice with an interval of at least two months between the two evaluations. Each investigator's results were analyzed for all pairs of antibodies, and the results of the different antibodies for all pairs of investigators. To assess how the parameters correlate with each other, Spearman's rank correlation was used in these pairwise analyses. Similarly, supposing the ideal linear relationship between Ki-67 LI values, Pearson's coefficients were also calculated for the same pairs. To investigate the influence of the observers and the antibody on the Ki-67 LI value further, two-way ANOVA was also

performed. The computations were done with the statistical software package SPSS 15.0 for Windows (SPSS Inc., Chicago, Illinois). Ki-67 scores were also divided into four quarters (0-25%, 26-50%, 51-75% and 76-100% Ki-67 LI) to allow categorical data analyses.

Four equally sized categories were arbitrarily chosen to limit their number and to allow a better analysis of the consistency of rating into a given category. As the 15% limit has been proposed as a cut-off for the low proliferation category and 30% for the high proliferative category, grouping according to these two marginal values (into three categories of unequal size) was also evaluated. Interobserver, interreagent and intraobserver agreements were assessed with kappa statistics according to Fleiss. Interobserver reproducibility was also evaluated by determining pairwise weighted kappa values. These weighted kappas do not only take into account the classification into another category (non-agreement) but also give weigh to the "distance" between the ordinal categories that have been used for classifying the Ki-67 values (e.g. having two ratings into neighboring categories is better than having them into categories separated by another category.) The kappa values were interpreted as reflecting slight (0-0.2), fair (0.21-0.4), moderate (0.41-0.6), substantial (0.61-0.8) and almost perfect (>0.8) agreement between observations according to Landis and Koch.

3.2 INVESTIGATION OF A STANDARDIZED, PARTIALLY DIGITALIZED COUNTING METHOD

In this part of the study, the same immunostained core biopsy samples were used as the ones described in section 3.1. Microphotographs were taken of each immunostained slide of the core biopsies at an identical magnification (x200). The hot-spot area was photographed in all cases where such a hot spot could be identified. More than one photograph was taken of each sample (range 2 to 4) and the best was selected for the study. The pictures were entered in a Microsoft PowerPoint file. Four different investigators first determined the Ki-67 LI by estimating the proportion of stained cells with 5% precision in the same areas (i.e. the same digital image displayed on a screen). No counting was involved in this assessment. Time needed for the evaluation was recorded in series of cases for all investigators. In a second round, a uniform grid composed of equidistant parallel horizontal lines was laid on all digital images, previously used for estimation. The observers were asked to count the tumor cells crossed by the lines or touching the lines. The lines of the grid can be followed and the

touching or crossed cells can be recorded (counted) continuously without the doubt of double counting or omitting single cells. Both immunohistochemically negative and positive nuclei were counted. Non-cancerous cells (stromal elements, lymphocytes etc.) were ignored as much as possible. The ratio of positive cells was derived from these values. In further analyses, rounded values (to the next integer) were used. Evaluation time was also recorded for this method. In all cases, the participating pathologists were asked to consider positive any cell with a brown (stained) rather than blue (unstained) hue. Comparisons were performed between the estimated and counted values of each investigator. Different investigators' values were also compared with each other. Kappa statistics were used to evaluate the interobserver reproducibility regarding estimation and counting. The following cut-off values were used (taking the values mentioned by the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer in 2009): 0-15%, 16-30% and >30%. Beside this categorization, Ki-67 values were divided into four quarters (0-25%, 26-50%, 51-75% and 76-100% Ki-67 LI). Spearman and Pearson correlation analyses were also used in order to compare the intra- and interobserver estimated and calculated values. Coefficients were categorized as follows: values between 0.9-1 show excellent, 0.75-0.9 good, 0.5-0.75 moderate, 0.25-0.5 week correlation, whereas values between 0.0-0.25 reflect lack of correlation. Comparisons were made for each antibody alone (30 values per observer) and for the three different antibodies combined (90 values per observer) for each pair of investigators. The analyses were performed both for the estimated and the calculated values, using software package SPSS 15.0 for Windows (SPSS Inc., Chicago, Illinois).

3.3 EVALUATION OF THE DISTRIBUTION OF KI-67 LIs IN BREAST CARCINOMAS DIAGNOSED AT DIFFERENT INSTITUTIONS

Contributors were asked to give the Ki-67 values of approximately 100 consecutive breast cancers where the staining was performed. Additional information collected in parallel included the ER, PR and HER2 status, the age of the patient, the histological type and grade of the tumor, the mitotic score component of the combined histological grade and the specimen type on which the Ki-67 values were determined. In a questionnaire, data about details of staining and evaluation were also collected. In a second round, another questionnaire specifically assessed the relation of the assessors to the recommendations of the

International Ki-67 in Breast Cancer Working Group. Data were collected between December 2012 and March 2013.

3.4 DIFFERENT METHODS OF PRETREATMENT KI-67 LABELING INDEX EVALUATION IN CORE BIOPSIES OF BREAST CANCER PATIENTS TREATED WITH NEOADJUVANT CHEMOTHERAPY AND THEIR RELATION TO RESPONSE TO THERAPY

Core-biopsy samples described in section 3.1 and 3.2 were chosen by considering regression of the primary tumor after neoadjuvant chemotherapy. All patients received combination chemotherapy containing docetaxel and an anthracycline (epirubicin for 29 and doxorubicin for one patient) with or without capecitabine.

Three categories were defined: no regression (or progression), partial regression (histological signs of regression) and pathological complete regression (no residual invasive tumor). Each category included different types of breast carcinoma regarding histological subtype, grade (G), hormone receptor (HR) and HER-2 status. Histological response was assessed according to the *European guidelines for quality assurance in breast cancer screening and diagnosis*. Each group consisted of 10 samples.

The previously mentioned three methods described in section 3.1 and 3.2 (counting on glass slide - method 1; counting on digital images – method 2; estimation on digital images – method 3) were completed with a fourth, in which the proportion of Ki-67 labeled cells was estimated by eyeballing on the glass slides – method 4. The same four investigators estimated the proportion of positive tumor cells on the slides by quick inspection of the slides, without counting. The average Ki-67 LIs of the respective 10 cases of all observers, antibodies and assessment methods were used for comparisons of clinical outcomes (tumor regression categories). The Kruskal-Wallis test and analyses of variance (one-way and two-way ANOVA) were performed to investigate the differences in Ki-67 LIs between different clinical outcomes. Receiver operating characteristic (ROC) curve analyses were applied to broadly approach a distinction between Ki-67 values of cases with complete response to therapy and those without it. The possible connection between ER receptor status and type of tumor regression (omitting the cases with partial regression) following neoadjuvant

chemotherapy was also analyzed applying Fisher's exact test. All analyses were performed with the statistical software package SPSS 15.0 for Windows (SPSS Inc., Chicago, Illinois).

4. RESULTS

4.1 REPRODUCIBILITY AND CORRELATION ANALYSIS OF THE KI-67 EXPRESSION LEVELS BY TYPES OF ANTIBODY AND INVESTIGATORS

Thirty core-biopsy samples of breast cancer patients were analyzed. The mean ± SE age of the population was 46 ± 2 (range: 26–70) years. Samples included 28 invasive ductal carcinomas, and 2 invasive lobular carcinomas. Altogether 270 evaluations were made by 3 independent pathologists with expertise in breast pathology (CG, VA, CE), and these were repeated a second time to better assess intraobserver variability. Mean±SE Ki-67 LI values of the first and the second evaluations by the 3 pathologist were 45±2%, 52±2% and 56±2% for the SP6, B56 and MIB-1 antibodies, respectively.

Spearman's rank correlation was used for comparison of the different observers and antibodies. Each analysis was performed on the first and the second evaluation separately and then combined. Interobserver comparisons for the dual (combined) evaluation suggested that the correlation between ratings was good to excellent (coefficients ranging between 0.74 and 0.91, $p<0.0001$). Interantibody comparisons yielded coefficients ranging between 0.8 and 0.92 ($p<0.0001$) suggesting excellent interantibody correlation. Pearson's coefficients ranged from 0.73 to 0.91 ($p<0.0001$.) for the pairs of observers and from 0.79 to 0.93 ($p<0.0001$) for the pairs of antibodies, suggesting similarly good to excellent correlation. When all 18 ratings (3 observers, 3 antibodies, twice) of the 30 cases were considered, the majority determined allocation of the cases was as follows: 1st quarter (0-25% Ki-67 LI) – 5 cases; 2nd quarter – 12 cases; 3rd quarter – 6 cases; 4th quarter – 7 cases. Only 2 tumors had all rating falling into the same quarter: one belonged to the lowest, one to the highest proliferation group. For the other categorization, the majority classifications suggested that 3 cases had a Ki-67 LI not greater than 15%, 4 cases fell into the 16-30% LI category and 23 cases had a high proliferation with a LI>30%. All the 10 cases with 100% agreement in the classification belonged to the highly proliferative tumors.

On the basis of the kappa values, the reproducibility of the Ki-67 LI values by quarter category distribution was generally moderate to substantial in the lowest and the highest quarters, but was only slight to fair in the middle quarter categories.

This resulted in overall kappa values reflecting only fair reproducibility both for given observers and for given antibodies. The same statement can be made after the evaluation of the dual assessment of all cases with all antibodies (18 evaluations per case), (kappa values: 0.3–0.44). Pairwise weighted kappa values showed somewhat better interobserver consistency (range 0.43-0.75) reflecting moderate to substantial reproducibility. When looking at the distributions determined by the 15% and 30% cut-off values, the reproducibility of the highly proliferative classification remained moderate/substantial, but the low and intermediate proliferation groups were less reproducible (kappa values: 0.37–0.52). Pairwise weighted kappa values ranged between 0.3 and 0.78 for these categories and also suggested fair to substantial reproducibility. The better reproducibility reflected by the weighted kappa values indicate that inconsistent categorizations were often one category away from each other, i.e. low proliferation rates were unlikely to be ranked as high and vice versa.

Similar statements can be made concerning the intraobserver agreements determined for all 90 immunostained samples: tumors with a high and a low Ki-67 LI can be more reproducibly identified than tumors falling into the intermediate (26-75% or 16-30%) range both on the basis of quarter based four-tiered distribution or the two cut-offs defined three-tiered distribution (kappa values: 0.26-0.42 and 0.36-0.61). However, differences between observers can also be identified. Observer 2 demonstrated somewhat worse performance than the other two. This observer had only 47% identical categorizations on the basis of the four-tiered categorization (in contrast to 57% and 58% for the two others) and 63% identical categorizations on the basis of the three-tiered classification (in contrast to 80% and 83% for the two others).

The two way ANOVA has confirmed the fact that Observer 2 rated the cases significantly lower than the two others ($p=0.004$ first observations and $p=0.008$ second observations), and also suggested that one antibody (SP6) also resulted in significantly lower mean Ki-67 LI values (45.2 % overall versus 52.3% for B56 and 55.7% for MIB-1; $p=0.017$ first round observations and $p=0.01$ second round observations).

The individual evaluation patterns were also evaluated. It turned out that all observers aimed to quantify the area with the highest number of stained cells; very weekly, faintly staining nuclei were generally discarded. All observers estimated the proportion of positive cells in relation to approximately 100 tumor cells of the chosen area, but there were also differences in the way of estimating this proportion. Observer 1 counted 100 cells and recounted the number of positive nuclei in the same 100 cells. Care was taken to keep in mind

the tumor structures included and excluded in the 100-cell-containing area to allow the count of the positive nuclei in the same 100 cells. Observer 2 counted 10 tumor cells to get an impression of the area these cells occupied and included an area of tumor cells ten times greater, than this basic count, and finally effectively counted the number of positive nuclei in the area estimated to contain 100 tumor cells. Observer 3 made an approaching guess in a larger area of the hot spot, than counted the proportion of stained nuclei in 20 or 30 cells, and multiplied by 5 or 3 to get a percentage estimate. For cases with a very high LI a reverse count was made, by evaluating the unstained nuclei. Therefore the used methods were comparable in some respects, but differed in their details of cell counting. All were generally based on the different approximations of the proportion of nuclear staining in relation to 100 cancer cells in hot spot areas, and neither tried to better estimate this ratio by counting 1000 cells as suggested by some authors.

4.2 INVESTIGATION OF A STANDARDIZED, PARTIALLY DIGITALIZED COUNTING METHOD

The patients and core biopsy samples were identical with the set studied in part 1 (under headings 3.1 and 4.1) and the basic characteristics were described there.

Altogether 720 evaluations were made by 4 independent pathologists with special interest in breast pathology (GC, AV, EC, BK).

The calculated Ki-67 LI was based on the assessment of an average of 75-91 cells. There were no major differences in the cells counted from the same set of digital images by different investigators. The range of cells counted on the grid marked images was 9 to 194, as a single image with the same grid was used for each tumor and core biopsy.

Good to excellent correlation was observed both with the Pearson's and the Spearman's methods when comparing the estimated and the calculated Ki-67 values of each investigator when analyzing all antibodies (90 cases per observer), (Pearson's coefficients: 0.899-0.920; Spearman's coefficients: 0.886-0.925).

The 90 assessments (all antibodies included) of each pair of investigators were also compared both for the estimated and the calculated Ki-67 LIs. The interobserver correlation coefficients demonstrate an excellent correlation (in case of estimation Pearson's coefficients:

0.909-0.964; Spearman's coefficients: 0.916-0.956; in case of calculation: 0.954-0.979 and 0.951-0.977).

Similar correlation analyses were repeated in case of each antibody, one by one (i.e. only 30 cases with the same antibody per observer), in order to see whether different antibodies were associated with different correlations. To compare the effect of the antibodies on the intraobserver correlation of estimated and calculated Ki-67 LIs, as a basic approach, the 4 correlation coefficients by observers were averaged. In case of SP6 the mean±SE values of the Pearson's and Spearman's correlation coefficients for the estimated Ki-67 LIs were 0.855±0.044 and 0.857±0.035, respectively. These values were 0.922±0.004 and 0.926±0.008 in case of B56, and 0.904±0.012 and 0.879±0.026 in case of MIB-1, respectively.

These results suggest that SP6 might have at least a trend for a slightly weaker intra-observer correlation, than the others. With interobserver analyses by antibody type, comparing the calculated values (6 pairs of investigators) in case of SP6, the mean±SE values of correlation coefficients derived from the Pearson and Spearmen tests were 0.871±0.044 and 0.881±0.039, respectively. In case of B56 these values were 0.967±0.005 and 0.947±0.005, respectively, while for MIB-1 they were 0.957±0.006 and 0.960±0.006, respectively. For the estimated values, the following results were found using Pearson and Spearman tests, respectively (mean±SE): 0.942±0.008 and 0.943±0.009 for SP6, 0.947±0.008 and 0.962±0.006 for B56, finally 0.926±0.012 and 0.896±0.013 for MIB-1.

These results also suggest that SP6 might be associated with less concordance between observers, but only in case of the calculated Ki-67 LIs while in case of MIB-1, the estimation of Ki-67 LIs showed lower correlations.

When all Ki-67 values (30 cases stained with 3 antibodies, i.e. 90 values per observer) estimated by eyeballing were considered, reproducibility of the proliferative activity was substantial both for the classification into 4 equal quarters (kappa: 0.68) and the classification into three categories (kappa: 0.65). The kappas were 0.67 and 0.73, respectively in case of the calculated Ki-67 values, all corresponding to substantial agreement.

Examining the antibodies one by one, using four categories and estimated Ki-67 values, the kappas were 0.65, 0.69 and 0.64 for the MIB-1, B56 and SP6 antibodies, respectively. For the three-tiered estimated Ki-67 categories, kappa values were 0.59, 0.69 and 0.67, respectively. Analyzing the calculated Ki-67 LIs, kappas were 0.66, 0.71 and 0.65, respectively for the four categories and 0.90, 0.60 and 0.69, respectively for the three categories. The agreement of the Ki-67 LIs gained by different antibodies was therefore

almost always substantial, with two instances suggesting moderate reproducibility but falling just short of the substantial agreement category, and one instance with an almost perfect agreement.

The mean time to evaluate the Ki-67 LI on a single digital image was calculated on the basis of the time used for the investigation of 30 biopsy samples stained by a given antibody. By eyeballing this time ranged between 18 and 50 seconds per investigator, and this range was between 90 and 180 seconds when the cells were counted, and the Ki-67 LI was derived from the calculated proportion of stained and all tumor cells.

4.3 EVALUATION OF THE DISTRIBUTION OF KI-67 LIs IN BREAST CARCINOMAS DIAGNOSED AT DIFFERENT INSTITUTIONS

Altogether 19 departments related to the European Working Group for Breast Screening Pathology referred data on 1782 tumors, of which 73 from one centre had to be excluded because of categorical Ki-67 values (<15, 15-30, >30 per cent LI), leaving 1709 tumors for further analysis. Ki-67 staining was automated in 15 of the laboratories providing the results, and MIB-1 antibodies were used in the majority (n=14). The antibody dilutions varied between 1:20 and 1:500. Full tumor sections were used to establish the Ki-67 LI in 72% of the cases (n=1233). Data on 1473 tumors were the results of routine staining in consecutive breast carcinomas, whereas data on 309 tumors (from 4 departments) were consecutive staining of nonconsecutive cases (in these departments not all breast cancers were stained, but all stained cases of a given period were included). Eleven of the 19 centers / pathologists reporting data (1009 tumors) counted the Ki-67 LI, whereas 8 of them used an eyeballing based estimation (773 tumors). Some method of rounding of the obtained Ki-67 LI values was used in the case of 873 tumors, and no rounding of the values was made for 909 tumors. Five pathologists assessed the proportion of staining in 100 cells, 3 in 200 cells, 1 in 300 cells, 1 in 500 cells and 1 in 1000 cells. Hot spots were included by 18/19 pathologists and were the only areas assessed when present in the practice of half of them. The recommendations of the International Ki-67 in Breast Cancer Working Group were known in 16 of the 19 laboratories, but were adhered to in only a minority. Overall, the median Ki-67 LI of the 1709 breast carcinomas analyzed further was 17% with a mean±SD of 23.4±21% (range 0% to 100%). When arranged in quarters or thirds, the mean±SD Ki-67 LIs (%) for quarters and thirds were 4.2±2.1, 12.4±2.4, 23.5±4.2, 53.6±18.5 and 5.6±3.1, 17.4±4.3, 47.3±19.4, respectively. The

lower half had a mean±SD LI of 8.3±4.7%, whereas the upper half was characterized by a mean±SD of 38.6±20.1%. The Ki-67 LI values were higher in cases in which the mitotic score component of the combined histological grade were higher. Each score was characterized by the following respective median and mean±SD Ki-67 LIs: score 1: 10 and 13±12; score 2: 23 and 27±18; score 3: 45 and 48±24. The Ki-67 LI values showed clustering at numbers ending with 5 or 0, 1084 values (63%) clustered at zeros and fives and 653 values clustered at zeros (38%).

As such clustering is to be expected with estimated values, the data were divided according to the method of evaluation (counting versus estimation) and the subset of values obtained from consecutive tumors with counting of the stained cells and no rounding of the values (n=600) was separately analyzed. Similar clustering of Ki-67 LI values was present in 199 (33%) and 119 (20%) cases, respectively.

As Ki-67 LI is often used to tailor treatment in ER-positive and HER2-negative tumors, this subset has been analyzed separately. The median Ki-67 LI was 14% (mean±SD: 17±15) in the 1248 patients having this type of carcinoma. Clustering of the values was seen in both the subsets assessed by counting (n=745) and estimation (n=503).

4.4 DIFFERENT METHODS OF PRETREATMENT KI-67 LABELING INDEX EVALUATION IN CORE BIOPSIES OF BREAST CANCER PATIENTS TREATED WITH NEOADJUVANT CHEMOTHERAPY AND THEIR RELATION TO RESPONSE TO THERAPY

Altogether 1350 evaluations were analyzed in our study (as one investigator did not take part in the first count, i.e. Method 1). The overall mean value of the Ki-67 LI (with all assessment methods, antibodies and observers considered) was 54.22 (95% CI: 53.13-55.31).

According to the statistical analyses, values of Ki-67 were significantly different in the different regression groups: values of the group without regression were significantly lower than the values of the group showing complete regression ($p<0.0001$). The mean values of Ki-67 LI taking into account all methods of assessment were 66.61 (95% CI: 64.71-68.50), 51.32 (95% CI: 49.43-53.21) and 44.72 (95% CI: 42.83-46.62) for the groups with complete, partial and no regression, respectively.

The two way analysis of variance using the methods of evaluation and the response categories as factors showed significant differences in mean Ki-67 values according to both

factors. The mean Ki-67 value was the lowest in Method 1 and highest in Method 3. Mean Ki-67 LI values were 46.63 (95% CI: 44.21-49.05), 54.62 (95% CI: 52.52-56.72), 61.59 (95% CI: 59.49-63.69) and 54.04 (95% CI: 51.94-56.17) in case of Methods 1 through 4, respectively.

We found similar results after taking the clinical response into consideration. Mean value of Ki-67 LIs by Method 1 were 58.72 (95% CI: 54.53-62.92), 46.39 (95% CI: 42.19-50.59) and 34.78 (95% CI: 30.58-38.98) in case of complete, partial and no regression, respectively, whereas these values were 77.18 (95% CI: 73.54-80.18), 56.59 (95% CI: 52.95-60.13) and 51.00 (95% CI: 47.37-54.64), respectively with Method 3.

By applying an ROC curve analysis to the mean Ki-67 data, the distinction between cases showing complete regression versus cases showing no regression (and omitting the cases with partial regression) on the basis of Ki-67 values (gained by any method) gives an area under the curve (AUC) of 0.969, whereas comparing complete regression values with the cases with partial regression and those showing no regression lumped together, the AUC turns to 0.93. Here the suggested best cut-off value to predict complete regression would have been around 56% LI (sensitivity: 0.89; specificity: 0.81).

According to the Fisher's exact test cases showing complete pathological regression were more likely negative for ER receptor immunohistochemistry (p=0.024).

5. DISCUSSION

Proliferation has been proposed as an important prognosticator of breast carcinomas, and the Ki-67 LI has been implied as a factor enabling a distinction between tumors with a high and a low proliferation. On this basis, Ki-67 has been used in various settings including the addition of adjuvant chemotherapy to hormonal therapy in the treatment of hormone sensitive breast carcinomas [St. Gallen 2009], the distinction between luminal A and luminal B subsets of estrogen receptor positive and HER-2 negative carcinomas or the distinction between histological grade 2 carcinomas with an outcome similar to grade 1 or grade 3 cancers.

Therefore, the reproducibility of Ki-67 LI is important. Our results suggest that interobserver reproducibility of the Ki-67 LI may only be fair in everyday pathology practice, although the correlation between assessments by different observers or using different antibodies is good or excellent. This may cast some doubt about the general usefulness of this

marker in its present state. The data also suggest that the classification of tumors into low and high proliferation categories is better (moderate to substantial) than that of tumors with an intermediate proliferative activity. This proved true both for cut off values of 25 and 75 per cents (per equal quarters assessment) and for cut-offs at 15 and 30 per cents as proposed by the St Gallen experts' consensus in 2009. This may also mean that the range where the distinction is clinically important is characterized by better reproducibility. We also noted a rather consistent intraobserver reproducibility variation. The kappa values relating to Observer 2 were consistently lower than the values of the two other observers, although the high and low proliferation categories had higher kappa values even for this observer. This may reflect the differences in the technique of evaluating the Ki-67 LI. All observers made the LI estimation on the basis of approximately 100 cells in a hot spot area. Observer 1 counted a real percentage values by counting 100 cells and recounting the positive (or at times the negative) nuclei in the same cells. Observer 2 approximated the area containing 100 cells by first delineating an area with one tenth of this population, and then counted the positive tumor cell nuclei in this area. Finally, observer 3 counted the proportion of immunostained tumor cell nuclei in one fifth or one third of the area and then extrapolated this result to 100 cells. The two latter techniques spare time. Interestingly, Observer 3 had kappa values very comparable with those of Observer 1. The most "time consuming" evaluation, that of Observer 1, took 3 to 4 minutes per case, and 90 to 100 minutes for scoring the 30 cases in the series once.

MIB-1 is generally used in everyday practice to determine the Ki-67 LI on formalin fixed and paraffin embedded material. This antibody was generated using recombinant technology and reacts with the immunodominant area of the Ki-67 nuclear antigen. B56 is also directed against the same area and also represent an IgG1 mouse monoclonal antibody. Although SP6 is a rabbit monoclonal antibody directed against the C-terminus of the Ki-67 protein, it has recently been reported to correlate well with MIB-1 staining. SP6 showed significantly lower mean Ki-67 LI values than the other antibodies in both the first and the second round assessments, but this did not influence reproducibility. In keeping with the above, the three antibodies used had very similar overall kappa values reflecting fair to moderate reproducibility for the whole range of the cases.

Several studies casted doubts about the reproducibility of Ki-67 LI evaluation. Although the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer (2009 and 2011) recommended an accurate methodical proceeding (e.g. counting 1000 tumor

cells), and the International Ki67 in Breast Cancer Working Group also suggested a similar approach, the everyday use and ability to follow this recommendation is questionable. As suggested in the present series, even intra-laboratory differences in assessment may exist; pathologists may tend to use less time-consuming methods. Despite only fair overall reproducibility of the Ki-67 LI into categories of different proliferative activity, the results between the observers and the antibodies were at least moderately correlated, and the identification of tumors with the highest proliferation was better (moderate to substantial), than that of tumors with intermediate proliferation according to both approaches tested. (Although this is the result of a universal statistical occurrence, we should be more aware of this phenomenon in clinical decision making, and acknowledge that intermediate or "grey-zone" categories are generally less reproducible than low and high value categories, as exemplified by our data). It therefore seems that clinically important categorizations (high proliferation versus non-high) can be more consistently made even with the methods described, although it is believed that substantial to excellent reproducibility would make the use of Ki-67 as a biomarker more confident.

Although an international consensus recommends the examination of at least 500, but optimally 1000 cells for deriving the Ki-67 LI, this practice is rarely followed, and counting about 100 cells or estimating the overall stained proportion of tumor cells are common methods of assessing proliferation as suggested by the data provided by pathologists from different European institutions in part 3 of this thesis. Indeed, 13 out of 19 participants followed these latter practices, i.e. estimated the proportion of stained nuclei or produced the percentage of stained cells on the basis of 100 cells. The lack of time is one of the most important factors deviating from the counting of high number of cells. The presented results regarding digital image analysis support that by choosing a limited area of high proliferation from the tumor as represented by a digital image, and by helping to choose which cells to count with a grid, improves reproducibility of determining the Ki-67 LI. Indeed, the interobserver agreement on the Ki-67 LI reached on the real slides of the same cases and derived on the basis of about 100 cells from the area with the highest staining proportion was only fair on the basis of the overall kappa values <0.4, but changed to substantial (overall kappa >0.6) for the digital images. Such an improvement in reproducibility was achieved by counting somewhat less cells on average than the 100 cells in the previous investigation, but the present study did not assess how many cells needed to be evaluated to reflect the proliferative activity of a tumor on the basis of a needle core biopsy, it only concentrated on

reproducibility issues. It may well be, that several digital images would be required to reflect tumor proliferation. We expect similarly acceptable reproducibility with 2, 3 or more images. On the dark side of such an improvement, we must accept a loss in time. Making digital images of given areas of a tumor histology slide and adding a standard grid to the image may be fast in some settings, but may also take too much time to be affordable. The evaluation itself is also somewhat time-taking, requiring 2 to 3 minutes per digital image, depending on the cell density. Therefore, the finding that a rough estimate of the stained proportion of tumor cells may be as reproducible as the calculated LI is of interest. Varga et al. in a very carefully designed study showed, that better reproducibility could be achieved by estimation than by accurate counting: there was much less deviation of data from a central mean value when the observers estimated the proportion of stained cells than when they counted it. Our results are in keeping with this observation, as good to excellent correlation was found between the estimated and the counted Ki-67 LI values and the overall kappa values also suggested substantial reproducibility for the eyeballing based estimation of the Ki-67 LI. Eyeballing obviously require less time, as supported by our data.

It is also important to highlight that our findings were retrieved on core-biopsy samples. Core biopsies might not always contain 1000 tumor cells for counting as non-tumor tissue may be common part of such specimens. Of the different factors potentially influencing the Ki-67 LI, tissue fixation, heterogeneity of different tumor areas, heterogeneity of nuclear labeling intensity are infrequent in such small samples, and this may improve reproducibility, although, on the other side, the representative nature of a small sample may be questionable. For example, Romero at al showed significant differences between core biopsy and surgical sample proliferation values, what may confirm our supposition that core biopsy specimens allow better reproducibility on the basis of a smaller sample with lesser variability.

Beside reproducibility analysis, it seemed to be interesting and informative to evaluate the distribution of Ki-67 LIs in breast carcinomas diagnosed at different institutions. The collected data demonstrate that the distribution of Ki-67 LI clusters at lower values (the median proportion of Ki-67 stained proliferative cells was 17% for all tumors and 14% for the ER-positive and HER2-negative subset), which reflects a tumor biology related influence on the distribution of the values. However the Ki-67 LI values also cluster at numbers ending with 5 or 10, which is a human evaluation related item.

Several factors influencing the determination of Ki-67 LI have been mentioned previously. In this series based on daily pathological methods and practices, the human factor

is reflected by the predilection for some LI values being more common than others in both the low and the high proliferation ranges. As the methods of evaluation are somewhat heterogeneous and the interobserver reproducibility of evaluating the proportion of Ki-67 labeled tumor cells is less then optimal, it is not surprising that the suggested cut-offs are also variable. Any cut-off value will separate higher and lower proliferative groups of tumors, but generalization of cut-off values from a specific study carries in itself a potential misclassification of some patients.

Published data suggest that the reproducibility of Ki-67 stained cell evaluation is not worse when the proportion is estimated rather than calculated after meticulous counting. Therefore the clustering of the values illustrated on the figures of this report (which can probably be generalized) suggest that if a cut-off between highly proliferating tumors and tumors with low proliferation is to be used to allow a selection of patients at higher risk of relapse and as a factor influencing the application of systemic chemotherapy, it should consider realistic distribution of the cases, and should probably be an inclusive or non-inclusive number ending with 5 or 0 (like 10%, 15% or 20%), or more preferably ending with 0 (like 10% or 20%). Taking into consideration the distribution of Ki-67 values presented in this study in comparison to previously presented cut-offs, some existing recommendations for Ki-67 LI in therapeutic decision making (e.g. the previous St Gallen recommendations using a cut-off of <14% for determining luminal A tumours by IHC) do not seem to be reasonable. In line with our findings, the latest St Gallen recommendations have appeared and mention an inclusive 20% cut-off for discriminating between luminal A and HER2-negative luminal B breast carcinomas. Different cut-offs may be generated for different clinical purposes.

As mentioned previously, the Ki-67 LI has prognostic value in breast cancer and may indicate the need for additional chemotherapy in special cases. Cytotoxic therapy acts on proliferating cells, therefore it would be very logical to hypothesize that tumors with a higher baseline proliferation may better benefit from chemotherapy than those with a low baseline proliferation. In keeping with this theoretical approach, a better response to neoadjuvant chemotherapy was reported in tumors with high baseline Ki-67 LI. In contrast, tumors with a high Ki-67 LI showed similar response rate (77%) to chemoendocrine therapy than those with a low Ki-67 LI (81%) in a retrospective series from the Royal Marsden Hospital. It is also likely that Ki-67 applied as a dynamic marker (i.e. taken as a baseline value and also during neoadjuvant therapy) may better predict the response and the outcome of disease, but multiple biopsies are not always easy to obtain. Our results show that breast cancers showing pCR

consistently had a higher mean Ki-67 LI than those not responding or showing only partial pathological response whatever the method of evaluation or the antibody used were or whoever the observer was. Although this difference was seen among the small groups showing regression, looking into the details demonstrates that the regression of individual cases of breast carcinoma cannot be precisely predicted, and prediction on the basis of high versus low Ki-67 LI does not work for the individual cases. The only thing that can be stated is that tumors with higher Ki-67 LIs are more likely to regress completely, independently of the method of evaluation.

The methods compared in this study had different reproducibility. Although a good to excellent correlation was found between the results gained by different methods, antibodies and observers, the interobserver and intraobserver reproducibility of the everyday practice based counting on glass slides (Method 1) was only fair to moderate. Reproducibility of the Ki-67 LI has improved when the field to be assessed was limited to a digital image, chosen to represent the highest proliferation in the core biopsy sample (Method 2). Interestingly, reproducibility was not worse when the proportion of stained cells was only estimated (Method 3) rather than counted (Methods 1 and 2). This is in keeping with the results of the carefully designed study by Varga et al, cited earlier: estimation does not seem to be worse than counting, it may even be better. This suggests that the simple approach of estimating by eyeballing has also diagnostic value, and may be better received by the pathological society as a simple and fast method, taking 4 to 12 times less time than counting. The daily use of estimating rather than counting the Ki-67 LI by at least some pathologists was also highlighted in the series evaluating the distribution of Ki-67 LI values in different pathology departments.

Although all of our study cases received neoadjuvant chemotherapy on the basis of locally advanced stage according to general practice and recommendations, biological markers, including those reflecting a high proliferation rate might be more suitable to select patients with potential benefit from this treatment. Pathological complete regression is associated with better prognosis, and tumors achieving pCR have a higher mean Ki-67 LI than those which do not achieve pCR, in keeping with earlier works. However, proliferation alone is not sufficient to predict the response to neoadjuvant chemotherapy and it is also a matter of debate what cut-off value should separate tumors with high and low Ki-67 LIs. Fasching et al used a >13% cut-off, in keeping with the value formerly suggested for the immunohistochemistry based separation of luminal A and luminal B carcinomas. As

suggested by our analysis of the data relating to the distribution of Ki-67 LIs gained with the participation of members of the European Working Group for Breast Screening Pathology, the distribution pattern of 1709 Ki-67 LI values had peaks at values ending with 0 or 5, and therefore an inclusive or non-inclusive cut-off ending with these values would seem more realistic, and independently, the latest St Gallen recommendations have happily adopted a 20% cut-off in keeping with this result. Our results would suggest that with most of the methods included, a cut-off of 50% would delineate high proliferation, and the ROC curve analysis also supports a similar value (i.e. 55%) but the analysis of the distribution pattern of Ki-67 LIs suggests that only a minority of breast cancers would fall into this category, as the median Ki-67 LI of the large cohort was 17%, whereas the estrogen receptor-positive cases had a median of only 14%. Therefore, the small subset of locally advanced breast cancers analyzed in the present work would fall into the higher end of the Ki-67 LI values of a general breast cancer cohort, and might not be representative enough. The fact that the determination of the Ki-67 LI in the present analyses was concentrating on the most proliferative parts of the tumors represented in the core needle biopsy samples rather than on both highly and less proliferating areas together may also partially explain these differences. These considerations would also counteract the determination of a cut-off value. Individual cases with high proliferation (e.g. Ki-67 LIs above 50%) would still have overlapping Ki-67 LI values with non-regressing tumors.

As to the best method, statistics do not allow to suggest that one of these methods is better than the other, they are just different from each other regarding to Ki-67 LIs. It seems obvious that Method 3 (the estimation made on the digital image taken from the area thought to represent the highest proliferation in the core biopsy sample) had substantial reproducibility, is fast and simple, and seems to separate responders from non-responders better than the other methods. However, individual cases would fail to follow the prediction even on the basis of this small sample. (The number of data evaluated in this study did not allow an ROC curve analysis of sufficient value to estimate the best cut-off value for this individual method.) Response to therapy is obviously a phenomenon depending on multiple factors of which the number of cells in the cycle is only one. For example, the predictive role of the hormone receptor status has also been widely investigated. A recent meta-analysis by Houssami also concluded that ER-negative tumors are more likely to achieve pCR. The data of this small series are in keeping with this, since ER-negative tumors were more common in the pCR group than in the rest of the patients.

6. CONCLUSIONS

6.1 Our results suggest that reproducibility of the Ki-67 LI is less than optimal even in core needle biopsies of breast cancer patients. However, the reproducibility of classifying tumors into a clinically more important highly proliferative category (like >15% or >25% or >30%) is better than that of the overall classification and is not very much influenced by the antibody used (at least true for the 3 antibodies tested).

6.2 At present, the consistency of Ki-67 LI determination in the routine work of some (probably numerous) laboratories (including ours) does not allow error free therapeutic decision making on a yes or no basis. Our results also indicate that similar but slightly differing individual practices of Ki-67 LI evaluation by different observers may influence reproducibility, therefore reasonable standardization and the recommendation of a workable uniform method should be encouraged.

6.3 The use of a simple digital technology, taking microphotographs of proliferating areas of breast cancers and adding a grid to the pictures to better delineate which cells to consider in the count makes possible for different investigators to examine the same area and the same cells when determining the Ki-67 LI. This can significantly improve reproducibility. We found that calculating the LI on the basis of such grid labeled digital images results in better reproducibility than the frustratingly low one found on the basis of counting stained cells on the histology slides of the same core biopsy specimens.

6.4 Estimating the proportion of Ki-67 stained cells on the same digital images is not only faster than counting the stained and unstained cells, but also results in acceptable and substantial reproducibility and the estimated and counted values correlate strongly. Therefore estimation should not be considered an inadequate method of establishing the Ki-67 LI, simply because it is not based on objective numbers.

6.5 On the basis of the distribution pattern of a larger series of Ki-67 LI values, some suggested Ki-67 LI cut-off values are not realistic, and it is proposed to select more realistic values ending with 0 or 5.

6.6 Tumors achieving pCR after neoadjuvant chemotherapy have a higher mean Ki-67 LI than those that achieve either partial pathological response or do not show regression.

This statement is true for all the 3 antibodies tested, all the 4 observers and all the methods evaluated. Therefore, the simplest methods of evaluating the Ki-67 LI by eyeballing rather than time-consuming counting could be a good alternative to assess the proliferation rate. Despite the association of a higher Ki-67 LI with pCR, a high Ki-67 LI alone is not sufficient to predict response to neoadjuvant chemotherapy, since other factors (including the ER status of the tumor) also influence the response to treatment.

6.7 Based on our studies, different evaluation methods may require different cut-off values to distinguish between tumors with high and low proliferation, and identify tumors with a higher chance of regression following neoadjuvant chemotherapy. Our data do not allow the suggestion of any cut-off value. The estimation of a cut-off value would probably require a larger series of cases.

7. ACKNOWLEDGEMENTS

This study was supported by **TÁMOP-4.2.2.A-11/1/KONV-2012-0**.

I wish to express special thanks to my supervisor **Professor Gábor Cserni** from the Department of Pathology, University of Szeged and Head of the Department of Pathology, Bács-Kiskun County Teaching Hospital in Kecskemét, for his support and scientific guidance of my work.

I would like to thank **Professor Béla Iványi** director and **Professor Tivadar Mikó** former director of the Department of Pathology, University of Szeged, who provided excellent working conditions for me at the institute.

I am also grateful to **István Pálka** who introduced me into breast pathology for the first time.

Special thanks are to the present and former staff members of the breast tumor board: **Professor Zsuzsanna Kahán**, **Professor György Lázár**, **Attila Paszt**, **Zsolt Simonka, Katalin Ormándi**, **Csilla Hoffmann**, **Máté Lázár**, **Gyöngyi Kelemen**, **Alíz Nikolényi**, **Erzsébet Valicsek**, **Ágnes Dobi**, **Zsófia Együd, Orsolya Rusz**, **Professor Gábor Cserni**, **Sándor Hamar**, **László Kaizer**, **Bence Kővári**, **Erika Csörgő, Péter Ragó, Levente Kuthi**, **Judit Tóth-Lipták**, **Péter Lázár** without whom this task would never have been fulfilled.

Special thanks are also due to **Tibor Nyári** for his help in the statistical analysis.

I greatly appreciate all the support and work of high standard provided by assistants, photo technician (**Mihály Dezső**) of the Department of Pathology, University of Szeged that helped this dissertation to be born.

I also thank my family and friends for encouraging and supporting me.