Pathol. Oncol. Res. (2018) 24:115–127
DOI 10.1007/s12253-017-0220-8

ORIGINAL ARTICLE

Reproducibility and Prognostic Potential of Ki-67 Proliferation Index when Comparing Digital-Image Analysis with Standard Semi-Quantitative Evaluation in Breast Cancer

Balázs Ács[1] · Lilla Madaras[1] · Kristóf Attila Kovács[1] · Tamás Micsik[2] · Anna-Mária Tőkés[3] · Balázs Győrffy[4,5] · Janina Kulka[1] · Attila Marcell Szász[1,4]

Received: 28 February 2017 / Accepted: 30 March 2017 / Published online: 11 April 2017

Abstract In this study, the reproducibility of Ki-67 proliferation index (KIPI) was investigated by comparing the semi-quantitative (SQ) results of three assessors with those of digital image-analysis (DIA) methods. The prognostic significance of the two approaches was also correlated with clinical outcome. Tissue microarrays of duplicate 2 mm cores were constructed from representative areas of formalin-fixed and paraffin-embedded tumor blocks of 347 breast cancer patients. SQ evaluation of Ki-67 (MIB1 clone) immunostained slides was performed independently by three pathologists. DIA was completed using a fully automated histological pattern and cell recognition module for KIPI detection (DIA-1) and an adjustable module (DIA-2) with the possibility of manual corrections. To compare SQ and DIA evaluations intra-class correlation (ICC) and concordance correlation coefficients (CCC) were determined. The three SQ evaluations demonstrated a remarkable ICC (0.853). Significant difference and poor concordance occurred between SQ-1 and SQ-2 as well as between SQ-1 and SQ-3 ($p \leq 0.001$, CCC ≤ 0.827 for both comparisons). Thus, the reference KIPI value (SQ-RV) was generated from the mean values of SQ-2 and SQ-3. SQ-RV and DIA-2 results showed substantial concordance (CCC = 0.963, at $p = 0.754$), while SQ-RV and DIA-1 values differed ($p \leq 0.001$) at only moderate concordance (CCC = 0.906). In multivariate analysis, lymph node status and SQ-2 assessment were significantly associated with clinical outcome ($p \leq 0.012$ for both comparisons). Our results confirm that KIPI is a significant prognostic marker in breast cancer, which can be can be reliably reproduced by using an adjustable DIA-2 image analysis module.

This study was presented at the San Antonio Breast Cancer Symposium 2015 and won the American Association for Cancer Research (AACR) Scholar-in-Training Award.

✉ Attila Marcell Szász
cac@korb2.sote.hu

Balázs Ács
acs.balazs.se@gmail.com

Lilla Madaras
madaras.lilla@med.semmelweis-univ.hu

Kristóf Attila Kovács
kovacs.attila@med.semmelweis-univ.hu

Tamás Micsik
micsikt@gmail.com

Anna-Mária Tőkés
tokesa1972@yahoo.co.uk

Balázs Győrffy
zsalab2@yahoo.com

Janina Kulka
kulka.janina@med.semmelweis-univ.hu

[1] 2nd Department of Pathology, Semmelweis University, 93 Üllői út, Budapest 1091, Hungary

[2] 1st Department of Pathology and Experimental Cancer Research, Semmelweis University, 26 Üllői út, Budapest 1085, Hungary

[3] MTA-SE Tumor Progression Research Group, 2nd Department of Pathology, Semmelweis University, 93 Üllői út, Budapest 1091, Hungary

[4] MTA-TTK Lendület Cancer Biomarker Research Group, 2 Magyar tudósok körútja, Budapest 1117, Hungary

[5] 2nd Department of Pediatrics, Semmelweis University, Tűzoltó u. 7-9, Budapest, Hungary

Keywords Ki-67 proliferation index · Breast cancer · Digital image analysis · Intra-class correlation · Concordance correlation · Prognosis

Introduction

Parameters that currently define treatment recommendations for early breast cancer are based on the patients' clinical data and on tumor biology. Among them, particularly important features are patient's age, tumor type, tumor size, grade, involvement of regional lymph nodes and distant organs (stage), as well as estrogen-, progesterone receptor and HER2 status [1]. Besides these, elevated proliferation can also define progression, treatment response and prognosis of breast cancer [2]. In addition to mitotic index, Ki-67 proliferation index (KIPI) assessed on formalin fixed, paraffin-embedded tissue samples is widely used in pathological practice [3]. Furthermore, according to the St. Gallen Consensus, both of 2013 and 2015, Ki-67 assessment is recommended to distinguish Luminal A-like and Luminal B-like cancers [4, 5].

However, clinical application of Ki-67 is still on debate due to lack of standardized immunohistochemical detection and scoring-system, which was denoted in the guidelines of the American Society of Clinical Oncology and International Ki67 in Breast Cancer Working Group [6–8]. Parameters that predominantly influence the immunohistochemical results of Ki-67 include pre-analytical, analytical, interpretation and scoring, and data analysis steps [7]. Poor results were linked with the consistency of sampling, fixation, antigen retrieval and staining methods [9, 10]. Although, recommendations published in 2011 provide a suitable landmark to improve pre-analytical and analytical validity, related protocols still show high variety [7]. Difficulties in evaluating immunoreactions can also be responsible for discrepancies of Ki-67 scoring reproducibility. KIPI values are usually defined as the percentage of positive tumor cell nuclei, counted in 3–10 high-power fields by testing 500–1000 tumor cells [11]. Another method is to estimate the mean KIPI in the entire lesion. Both methods are monotonous, time-consuming and exhausting with a chance of leading to controversial results and inaccurate reproducibility [11].

Rapid development of digital microscopy by now allows fast digitalization of histological slides at high-resolution, which can firmly support education, research and diagnostics in pathology [12, 13]. The emergence of digital image analysis platforms improved the capacity, precision and reproducibility of in situ biomarker evaluation [14]. However, these features alone may not enough for diagnostic accuracy, which must be based on histological pattern recognition as the most relevant requirement of precise sample selection and assessment of immunoreactions [15].

Digital image analysis (DIA) platforms are able to assess KIPI, however it has not been clarified yet, if their results can meet the requirements of the daily diagnostic practice [16]. In this study, reproducibility of KIPI was investigated among three pathologists, based on their traditional visual estimation. These semi-quantitative evaluations served as reference KIPI values for testing their agreement with two DIA assessment modules. Finally, the outcome prediction potential of each semi-quantitative and DIA assessments were determined and compared to that of conventional clinicopathological factors.

Material and Methods

Patients

This study was performed using 347 consecutive female breast cancer cases, diagnosed between January of 1999 and December of 2002 at the Buda MÁV Hospital, Budapest, Hungary. All patients' breast cancers had been surgically removed. Pathological features that could not be assessed on tissue microarrays (TMA) were retrieved from the pathology reports or the original H&E stained slides were reviewed. Treatment and follow-up data were retrieved from patients' medical records. Regarding the definition of surrogate molecular subtypes of breast cancer, we referred to the St. Gallen recommendations from 2013 that include five categories (Luminal A, Luminal B/HER2-, Luminal B/HER2+, HER2+ and triple negative [5].

Tissue Preparation

Tissue microarrays were built from 10% buffered formalin-fixed, paraffin-embedded tissue blocks of the 347 cases. Tumor areas were selected by pathologists on the basis of HE stained slides. Duplicate cores (each 2 mm in diameter) were punched (TMA Master, 3DHISTECH Ltd., Budapest, Hungary) from each case, resulting 10 TMA blocks.

Immunohistochemistry

Paraffin sections of 3 μm thickness were cut from the TMA blocks for immunohistochemistry. The mouse monoclonal antibody clone Mib1 (1:100, Dako, Glostrup, Denmark) was used to detect Ki-67 protein in an automated immunostainer (Ventana Benchmark XT, Roche, Basel, Switzerland) according to the manufacturer's protocol (at 42 °C for 32 min) after antigen retrieval using the pH 9.0 CC1 buffer at 42 °C for 30 min. For antibody visualization, UltraView DAB Detection kit (Ventana, Tucson, USA) was applied.

Semi-Quantitative Evaluation of Ki67 Reactions

Semi-quantitative (SQ) evaluation was performed on digital slides using the TMA Module software on the PannoramicViewer (v1.11.49.0) platform (all 3DHISTECH, Budapest, Hungary) by 3 pathologists (SQ-1, SQ-2, SQ-3) independently as follows: KIPI was defined as the percentage of positive tumor cell nuclei, estimated on average in each core. Any nuclear positivity was considered, including nuclear, nucleolar or nuclear membrane localization irrespective of the pattern (granular or diffuse) in a range of 100–500 cells, depending on the cellularity of the TMA cores. Duplicate cores were evaluated separately and their mean KIPI was finally analyzed. The three pathologists involved in our study have considerable but different level of experience in Ki-67 scoring of breast cancer. SQ1 is the youngest with a pathology specialist status for a year only. SQ-2 and SQ-3 are consultant pathologists with substantial experience in diagnostic practice and special focus on breast pathology. Dichotomization of KIPI values either at14% or 20% thresholds was also performed [4, 5].

Digital Image Analysis

TMA slides were digitized with Pannoramic Flash II slide scanner using ×20 objective (NA = 0.83), collecting sharp signals from 7 focal planes in "Extended-focus" mode through the 3 μm section thickness at 80 jpeg image quality factor. DIA was performed using the PatternQuant (PQ) software of the QuantCenter package module enabling automated tissue pattern recognition by separating epithelial elements from stroma. All digital hardware and software tools were from 3DHISTECH (Budapest, Hungary). Designation of training tissue patterns to be recognized and the calibration were done in co-operation by a pathologist and an IT expert to achieve the best recognition pattern (achieved at a PQ training magnification of 1.5×; a gamma level of 1; dilution of 3; a contour of 0). So as the detection and quantification of tumor cell nuclei using NuclearQuant (NQ) at the following settings: Blur: 15; Radius minimum: 1.5; Radius maximum: 8; Area min: 15; Intensity minimum: 30; Contrast minimum 30 (Fig. 1). The brown DAB and the hematoxylin counterstain were separated with digital color deconvolution [17]. Based on these settings of PQ and NQ, automated Ki-67 evaluation was performed on each core (DIA-1 analysis). In the other DIA test, automated annotations were assessed by pathologists on each core, and when it was necessary, DIA settings were adjusted independently (from the KIPI results of DIA-1, SQ-1, SQ-2, SQ-3) to exclude artifacts, underestimation or overestimation of positive/negative cells and false detections (DIA-2 analysis).

Statistical Analysis

For statistical analysis SPSS 22 software (IBM, Armonk, USA) and MedCalc 13.3.3.0 (MedCalc Software, Ostend, Belgium) software were used. Interobserver variability was estimated with intra-class correlation coefficient (ICC) and concordance correlation coefficient (CCC). Altman's guideline was followed for the interpretation of ICC [18]. CCC was interpreted according to McBride [19]. Degree of agreement among different observers (SQ-1, SQ-2, SQ-3, DIA-1, and DIA-2) was evaluated by using Cohen's kappa and Bland-Altman Plot. To assess statistical differences between observers the Wilcoxon signed-rank and McNemar tests were

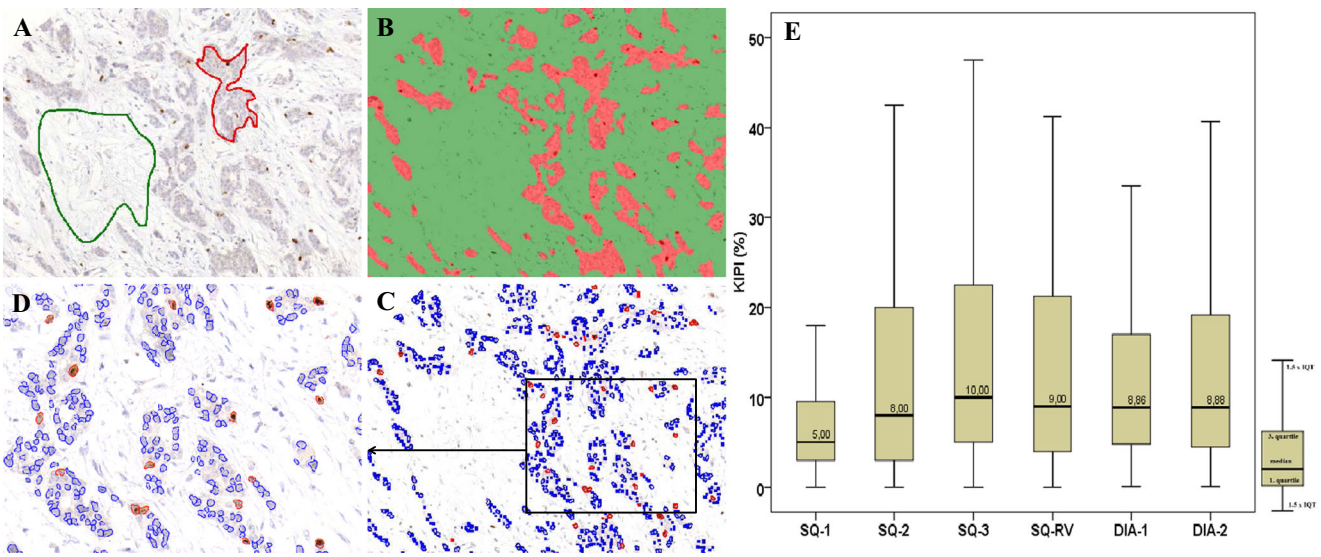

Fig. 1 Workflow of 3DHistech DIA assessment and boxplot of KIPI assessments. Examples of desired tissue patterns were given, demarcated with the red and green lines (red = epithel pattern, green = stroma pattern) (**a**), that we wanted to be recognized and distinguished by the software named PatternQuant (**b**). Then the software named NuclearQuant counts the recognized negative (blue) and positive (red) cells only in the annotations designated by PatternQuant (red areas on picture B) (**c**, **d**).SQ-1 differed mostly from the other KIPI evaluations (**e**)

applied, since our data were not normally-distributed, even after log-transformation (Shapiro-Wilk and Kolmogorov-Smirnov tests). Kaplan-Meier analysis supported with log-rank test was executed using disease-free survival (DFS). To compare prognosis prediction potential, multivariate Cox-regression analysis was applied. In all statistical analysis the level of significance was set at $p < 0.05$.

Results

Aggregate clinicopathological features of the cases are displayed in Table 1. Mean patient age was 58.93 years (SD ± 13.05; range: 27–90). Most of the cases were pT1 and pT2, the majority with low mitotic index and histological grade of 1 or 2 and of Luminal A– like subtype. In 87 cases we didn't have data about the axillary status, but in the known cases, the majority of cases' axillary stage was pN0–1. More than half of the patients (200/347, 57.7%) underwent postoperative breast irradiation, while 42.4% (147/347) of the patients received adjuvant chemotherapy in this cohort. All ER positive breast cancer patients were treated with anti-estrogens. Median follow-up time was 99.85 months.

SQ Evaluations

We examined the 3 SQ KIPI evaluations (SQ-1, SQ-2, SQ-3), and the following median values were observed: 5 (SQ-1), 8 (SQ-2), 10 (SQ-3) (Fig. 1). Significant difference was found between all the 3 SQ KIPI assessments expressed in percentage (p values for all comparisons ≤0.001). However, they showed a very good consistency (ICC = 0.853) concerning the relative difference between cases. The best interobserver variability was found between SQ-2 and SQ-3 (CCC = 0.935), while SQ-1 showed poor concordance with SQ-2 and SQ-3

Table 1 Clinicopathological data of the patients

Patients				
		(n, %)	347	100%
Age	(mean ± SD, range)		58.93 ± 13.05	27–90
Tumor size (mm)	(mean ± SD)		24.37 ± 15.77	
Mitotic index (n/10HPF)	(mean ± SD)		9.52 ± 11.08	
Grade	1	(n, %)	123	35.4%
	2	(n, %)	138	39.8%
	3	(n, %)	86	24.8%
Subtype	LUMA	(n, %)	233	67.1%
	LUMB	(n, %)	46	13.3%
	HER2	(n, %)	20	5.8%
	TNBC	(n, %)	47	13.5%
	no data	(n, %)	1	0.3%
Lymph node status (TNM 7)	0	(n, %)	121	34.9%
	1	(n, %)	75	21.6%
	2	(n, %)	38	10.9%
	3	(n, %)	26	7.5%
	no data	(n, %)	87	25.1%
Vascular Invasion	none	(n, %)	103	29.7%
	present	(n, %)	235	67.7%
	no data	(n, %)	9	2.6%
Necrosis	none	(n, %)	249	71.8%
	present	(n, %)	92	26.5%
	no data	(n, %)	6	1.7%
Chemotherapy	no	(n, %)	194	55.9%
	yes	(n, %)	147	42.4%
	no data	(n, %)	6	1.7%
Irradiation	no	(n, %)	141	40.6%
	yes	(n, %)	200	57.7%
	no data	(n, %)	6	1.7%
Follow-up time	(n, median, IQT*)		306, 99.85	57.17

*interquartile range

(CCC = 0.817, CCC = 0.827, respectively). We also investigated the agreement of the three SQ evaluations using Bland-Altman plots (Fig. 2). Significant bias was observed in all comparisons. The lowest bias and the narrowest range of agreement were found between SQ-2 and SQ-3 without a proportional error, however, the variability of differences still

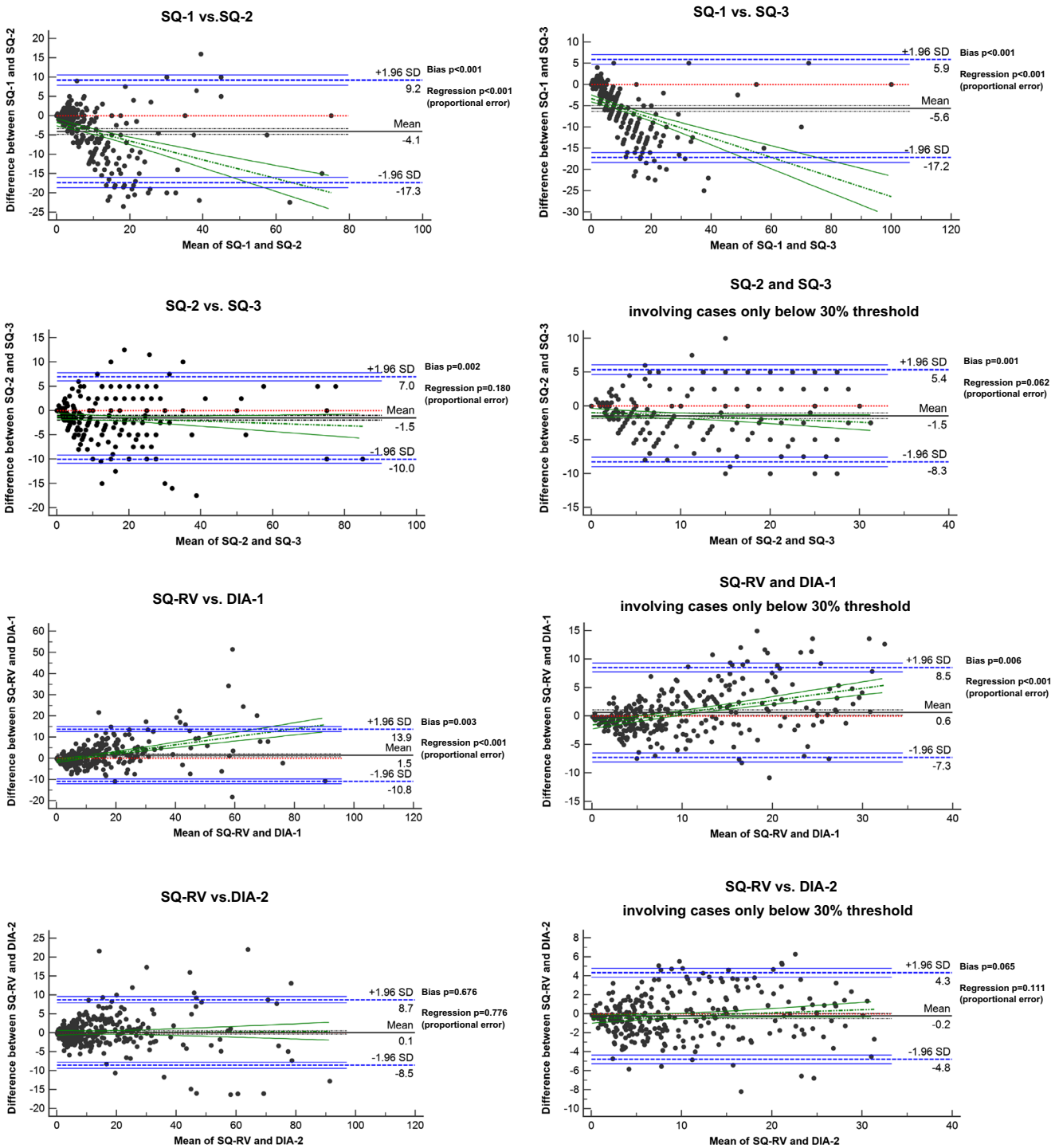

Fig. 2 Bland-Altman plots comparing KIPI evaluations and Bland-Altman plots comparing KIPI evaluations involving cases only below 30% cut-off point. Red dashed line corresponds the expected mean zero difference between Ki67 LI scores of the antibodies. Black line represents the observed mean difference between Ki67 LI scores of the antibodies, namely the observed bias (black dashed lines are the CI of the observed mean difference). Blue dashed lines illustrate the range of agreement (lower and upper limit of agreement) based on 95% of differences (blue lines are the CI of the limits of agreement). Green dashed line is the fitted regression line to detect potential proportional error (green lines are the CI of the regression line)

showed an increasing trend, proportional to the magnitude of KIPI. Bland-Altman plots were also created for cases of <30% KIPI values since these were overrepresented in our cohort, with still covering all clinically relevant thresholds. The same trends were observed between SQ evaluations with lower bias and narrower range of agreement.

Upon dichotomizing KIPI values at 14% and 20% thresholds, SQ-1 still differed considerably from SQ-2 ($p \leq 0.001$, $p \leq 0.001$, respectively) and SQ-3 ($p \leq 0.001$, $p \leq 0.001$, respectively) with a moderate agreement (SQ-2 κ/14%/= 0.462, SQ-2 κ/20%/= 0.490, SQ-3 κ/14%/= 0.452, SQ-3 κ/20%/= 0.473). However, no significant difference ($p = 0.708$, $p = 0.082$, respectively), and a substantial agreement (κ/14%/= 0.741, κ/20%/= 0.727) was found between SQ-2 and SQ-3, at these thresholds.

DIA Evaluations

The median values for DIA evaluations were the following: 8.86 (DIA-1) 8.88 (DIA-2) (Fig. 1). For the comparison with DIA assessments, a reference SQ KIPI value was generated (SQ-RV) as the mean of SQ-2 and SQ-3, since SQ-1 differed considerably from those. SQ-RV and automated DIA-1 differed ($p \leq 0.001$) and showed moderate concordance (CCC = 0.906). SQ-RV and adjustable DIA-2 showed no significant difference ($p = 0.754$), and represented a substantial concordance (CCC = 0.963). Significant difference ($p \leq 0.001$) but substantial concordance (CCC: 0.943) was found when DIA-1 was compared to DIA-2. Using Bland-Altman plots showed a significant bias and proportional error between SQ-RV and DIA-1 values, which was not seen between SQ-RV and DIA-2 values and the range of agreement was also superior in the latter case (Fig. 2). Moreover, in the comparison of SQ-RV and DIA-2, the variability of differences did not show an increasing trend, proportional to the magnitude of KIPI. The same results were found at 30% threshold between SQ-RV and DIA evaluations, but the range of agreement became narrower in all comparisons (Fig. 2).

At 14% and 20% thresholds, though DIA-1 differed from SQ-RV significantly ($p = 0.010$, $p \leq 0.001$, respectively), DIA-2 and SQ-RV values did not ($p = 0.337$, $p = 0.701$, respectively). Both DIA methods showed substantial (DIA-1 κ/14%/= 0.743, κ/20%/= 0.775) or outstanding agreement (DIA-2 κ/14%/= 0.849, κ/20%/= 0.868) with SQ-RV. Though significant difference occurred between DIA-1 and DIA-2 (p/14%/= 0.019, p/20%/≤ 0.001), agreements were high (κ/14%/ = 0.894, κ/20%/ = 0.852). Interobserver variability within DIA (DIA-1, DIA-2) and SQ (SQ-1, SQ-2, and SQ-3) evaluations referred to a very good consistency (ICC = 0.886).

Comparing SQ and DIA Evaluations in Prognosis Prediction

For prognosis, all KIPI evaluations (DIA-1 $p = 0.031$, DIA-2 $p = 0.018$, SQ-1 $p = 0.022$, SQ-2 $p = 0.008$) but SQ-3 ($p = 0.062$) were able to perform statistically significant splitting of our cohort into 2 patients' group with distinct DFS at 14% threshold (Fig. 3). At 20% cut-off point, KIPI evaluations of DIA-2 ($p = 0.004$), SQ-2 ($p \leq 0.001$) and SQ-3 ($p = 0.013$) were able to sort patients into good and unfavorable prognostic groups, while SQ-1 ($p = 0.085$) and DIA-1 ($p = 0.055$) did not (Fig. 3). KIPI assessments were also tested as potential independent predictors of DFS adjusted by IHC subtypes, lymph node status, histological grade, mitotic index, vascular invasion and necrosis. At 14% cut-off no KIPI evaluation but only lymph node status ($p = 0.001$) showed independent association with DFS. However, at 20% threshold, both lymph node status and SQ-2 were significantly linked to DFS ($p = 0.012$). Tables 2 and 3 show results of Cox regression analysis for DIA and SQ evaluations and for clinicopathological factors. Effect of different treatment protocols on clinical outcome was also investigated (Table 2). Patients who underwent surgical intervention only had the longest DFS, while patients who received surgery + irradiation + chemotherapy combination had the most unfavorable prognosis (Fig. 4). KIPI assessments could not split our cohort into distinct DFS in groups in patient groups treated either with surgery + irradiation, or with surgery + irradiation + chemotherapy combination (Table 4). All KIPI evaluations but SQ-1 were able to significantly distinguish good and unfavorable prognosis at 20% threshold in patients who underwent surgery only (SQ-1 $p = 0.085$, SQ-2 $p < 0.001$, SQ-3 $p = 0.020$, DIA-1 $p = 0.034$, DIA-2 $p = 0.010$, Table 4). In the patients treated with surgery + chemotherapy, statistically significant prognostic results were seen only with SQ-2 evaluation ($p = 0.049$, Table 4). Multivariate analyses of KIPI assessments within treatment subgroups were not performed due to the low number of cases compared to relatively numerous clinicopathological factors.

Discussion

Uncontrolled proliferation is a hallmark of cancer acquired during multistep development of human neoplasias [2]. To define proliferation, mitotic rate and Ki-67 proliferation index have been widely used in histopathology practice [11, 20]. In contrast to mitotic figures, Ki-67 is expressed not only in M phase but all phases of cell cycle except G0, capturing a larger proportion of proliferating cells [20]. However, clinical utility of KIPI is still uncertain. The ongoing debate related to the use of Ki-67 IHC in oncology decision making has been emphasized by the International Ki67 in Breast Cancer Working

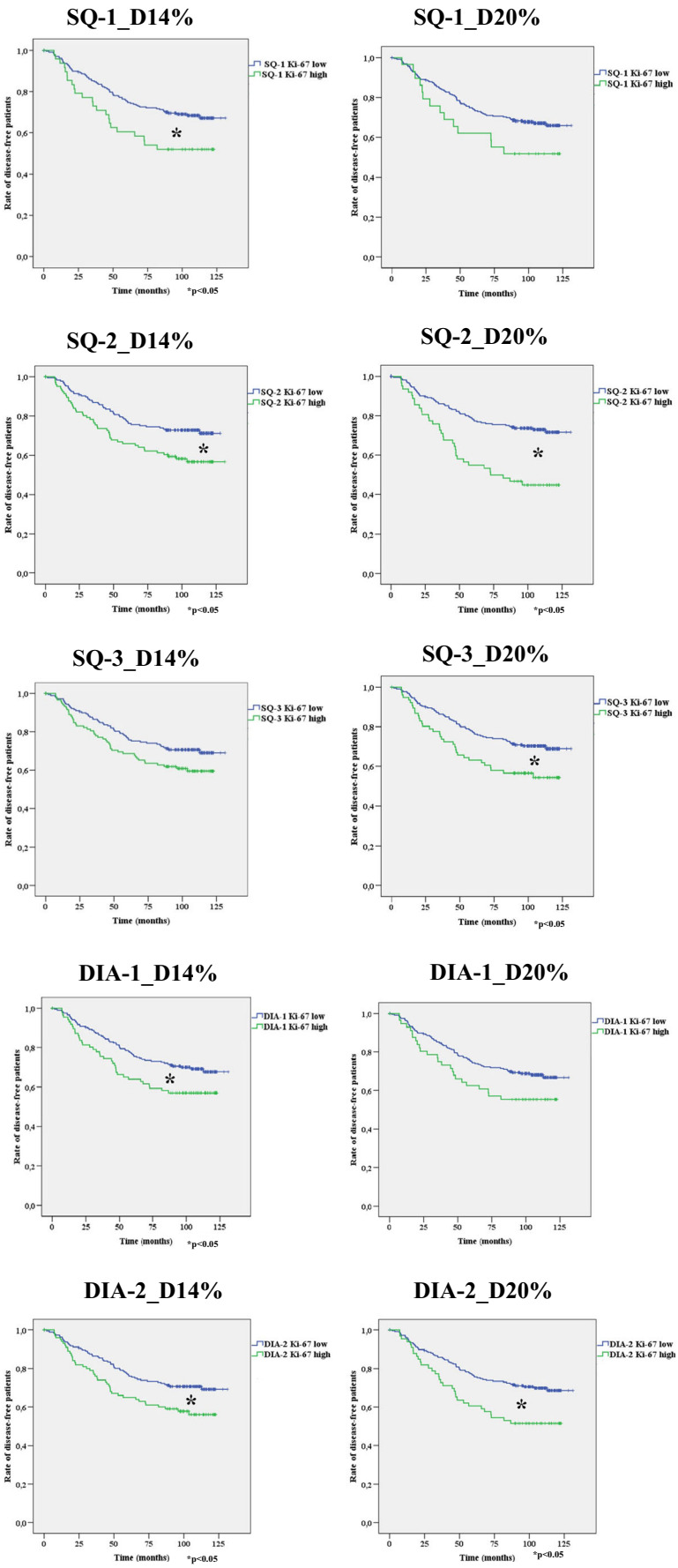

Fig. 3 Various KIPI evaluations and disease free survival

Table 2 Univariate Cox regression analysis of KIPI assessments and pathological factors

Prognostic factor	Subgroups	Univariate Cox regression analysis		
		HR	95% CI	*p*-value
Age	–	0.915	0.596–1.406	0.685
Tumor size	<2 cm	–	–	0.629
	2-5 cm	1.245	0.796–1.945	0.337
	>5 cm	1.176	0.518–2.671	0.698
IHC Subtype	Luminal-A like	–	–	**≤0.001**
	Luminal-B like	**1.963**	**1.154–3.340**	**0.013**
	HER2	**3.278**	**1.791–6.001**	**≤0.001**
	TNBC	1.728	0.991–3.013	0.054
Histological grade	1	–	–	**0.017**
	2	1.509	0.931–2.446	0.095
	3	**2.060**	**1.253–3.387**	**0.004**
Lymphnode status (TNM 7)	0	–	–	**0.002**
	1	**2.186**	**1.230–3.888**	**0.008**
	2	**2.295**	**1.167–4.516**	**0.016**
	3	**3.686**	**1.842–7.374**	**≤0.001**
Mitotic index	≤9	–	–	0.053
	10–19	1.320	0.837–2.080	0.232
	>19	**1.858**	**1.113–3.100**	**0.018**
Vascular invasion	–	**2.035**	**1.245–3.324**	**0.005**
Necrosis	–	**1.602**	**1.066–2.407**	**0.023**
Treatment	Surgical treatment only	–	–	**≤0.001**
	Surgery + Irradiation	1.596	0.811–3.143	0.176
	Surgery + Chemotherapy	**2.970**	**1.395–6.321**	**0.005**
	Surgery + Chemotherapy + Irradiation	**3.174**	**1.760–5.721**	**≤0.001**
SQ-1_D14%	–	**1.730**	**1.084–2.762**	**0.022**
SQ-1_D20%	–	1.645	0.934–2.898	0.085
SQ-2_D14%	–	**1.723**	**1.149–2.583**	**0.008**
SQ-2_D20%	–	**2.445**	**1.604–3.727**	**≤0.001**
SQ-3_D14%	–	1.454	0.981–2.153	0.062
SQ-3_D20%	–	**1.693**	**1.119–2.561**	**0.013**
DIA-1_D14%	–	**1.563**	**1.041–2.346**	**0.031**
DIA-1_D20%	–	1.557	0.990–2.449	0.055
DIA-2_D14%	–	**1.611**	**1.084–2.394**	**0.018**
DIA-2_D20%	–	**1.844**	**1.211–2.808**	**0.004**

The bold values highlight the statistically significant prognostic factors

D14% dichotomized at 14% threshold, *D20%* dichotomized at 20% threshold

Group of the Breast International Group in 2011 [7]. According to this expert group, variability of Ki-67 IHC largely emanates from preanalytical, as well as analytical factors and from the discrepancy between observers in KIPI scorings [7]. We considered that preanalytical and analytical issues didn't bias our results since all of our observers evaluated the same slides, thus discrepancies between final KIPI values were derived from variability of each observer's evaluations.

The other reason why this group of experts did not advise the application of Ki-67 IHC results in therapy decision making is interobserver variability [9]. Ring studies showed that moderate intraclass correlation (0.59–0.71) achieved between observers performing SQ evaluations, could be inmproved to 0.92 [8, 9], based on systematic training and following the guidelines [8]. Although we found a very good consistency between SQ evaluations, statistically significant difference and poor concordance also occurred between SQ-1 and SQ-2 as well as between SQ-1 and SQ-3. Besides this, the variability of differences between SQ-1 and SQ-2 as well as SQ-1 and SQ-3 represented a proportional error. The possible explanation for the discrepancy might be that SQ-1 has the least experience in daily diagnostic practice. This observation

Table 3 Multivariate Cox regression analysis of KIPI assessments and pathological factors

Prognostic Factors	Multivariate Cox regression analysis involving KIPI assessments and clinicopathological factors		
	HR	95% CI	*p*-value
Age	-	-	-
Tumor size	-	-	-
IHC Subtype	1.078	0.910–1.277	0.385
Histological grade	1.064	0.699–1.620	0.771
Lymphnode status (TNM 7)	**1.435**	**1.133–1.817**	**0.001**
Mitotic index	1.154	0.782–1.701	0.471
Vascular invasion	1.016	0.534–1.934	0.961
Necrosis	1.237	0.688–2.227	0.477
SQ-1_D14%	1.481	0.773–2.838	0.237
SQ-2_D14%	1.296	0.713–2.355	0.395
SQ-3_D14%	-	-	-
DIA-1_D14%	1.265	0.709–2.257	0.426
DIA-2_D14%	1.489	0.839–2.642	0.174
SQ-1_D20%	-	-	-
SQ-2_D20%	**2.287**	**1.199–4.364**	**0.012**
SQ-3_D20%	1.048	0.570–1.924	0.881
DIA-1_D20%	-	-	-
DIA-2_D20%	1.460	0.797–2.674	0.221

The bold values highlight the statistically significant prognostic factors

D14% dichotomized at 14% threshold, *D20%* dichotomized at 20% threshold

might emphasize the relevance of consecutive experience and training in breast pathology. In the 2013 Ki-67 ring study of the Japan Breast Cancer Research Group intraclass correlation ranged from 0.57 to 0.66 when pathologists scored whole slides applying counting and visual estimate methods [21]. When they evaluated printed photographs of Ki-67 stained slides to exclude variations by assessment of varied microscopic field, 0.82–0.94 of intraclass correlation was observed [21]. This study has claimed that the standardization of assessment area might be the essential point to evaluate Ki-67 with high reproducibility [21]. We performed KIPI evaluation on TMA slides to avoid variation in scorings by different microscopic fields. Concerning the relative difference between cases, a very good intraclass correlation was observed between our pathologists, suggesting the area of interest to be assessed is essential regarding KIPI. This conclusion was also implied in a recent study where the highest agreement between pathologists was observed when regions of interest were defined on whole slides to be assessed for KIPI [22].

In the work of the Japan Breast Cancer Research Group, counting method was slightly superior to visual estimation [21]. In our study, for SQ assessment the "eye balling" method was applied, because it was shown in two recent investigations that visual estimation could be just as good as the meticulous counting method of the ratio of positive tumor cell nuclei among all tumor cell nuclei [22–24]. Furthermore, visual estimation is less time-consuming and the possibility of miscalculation also persists as chance of error for the counting method.

Digital image analysis offers the opportunity to assess KIPI more objectively and with increased reproducibility, but concordance compared to conventional evaluations is currently under examination. In a recent study, an ICC of 0.885 was observed, when DIA KIPI and conventional SQ KIPI assessments were compared [25]. They performed KIPI evaluations on whole slides of 50 cases of breast cancer, selecting 3–5 hot spots to be assessed, and both methods were performed on identical high-power fields [25]. Similarly high concordance (ICC: 0.93) was found between the fully automated DIA assessment and SQ evaluation by Klauschen et al., who performed KIPI analysis on whole core biopsies from 1082 patients [26]. In our study, both automated DIA and adjustable DIA assessments represented substantial or outstanding agreement with SQ-RV evaluation. However, adjustable DIA seemed superior to automated DIA, since only the adjustable DIA assessments showed no proportional error compared to SQ-RV and the variability of their differences did not show an increasing trend, proportional to the magnitude of KIPI. Furthermore, significant difference was observed between automated DIA and SQ-RV evaluations, while adjustable DIA and SQ-RV did not differ significantly. This result was also observed in the study by Laurinavicius et al. who found improvement in DIA evaluation when quality assessment was achieved on the default automated DIA evaluation [15]. In our study, significant difference was found between automated DIA and adjustable DIA assessments. Basically, DIA method is more dependent on IHC staining quality, than conventional evaluation, since the human brain is able to compensate inadequate IHC quality [16]. Unequal tissue thickness and folds, cracks on glass slides, uneven coverglass glue layer might also lead to suboptimal quality in scanning slides and to false image analysis results [16]. In our opinion, significant discrepancies between our DIA methods were due to these features (Fig. 5), which can be avoided by a pathologist's adjustment and by standardization of preanalytical and analytical steps of IHC. In our investigation it has been also demonstrated, that the adjustable DIA is as robust as the visual estimation of KIPI performed by well-trained and experienced pathologists.

Several studies have compared KIPI assessment obtained by DIA to survival rates such as disease-free survival and overall survival [16]. To investigate the outcome prediction potential of KIPI, dichotomizing is needed at a well-defined cut-off point. However, former guidelines have recommended different thresholds for such dichotomization; recent studies suggest, that an optimal cut-off score for KIPI is not definable [4, 27]. Thus, local laboratory specific cut-off points or KIPI as a continuous marker should be applied to assess proliferation potential of the tumor [5].

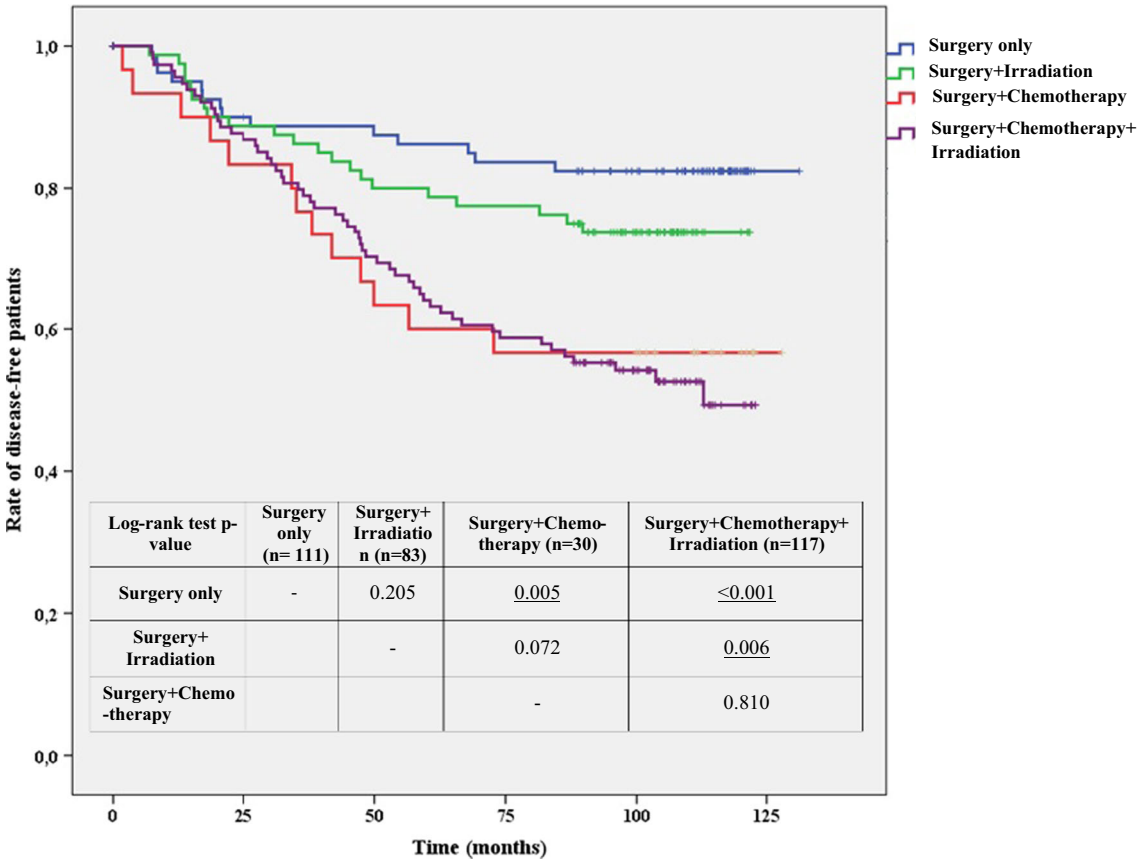

Fig. 4 Survival functions of the treatment subgroups

In a recent study, the prognosis prediction of DIA KIPI evaluation was significant in univariate analysis, although in multivariate analysis it has not remained significant compared to conventional clinicopathological factors [28]. In contrast with the results of this study, Klauschen reported that KIPI obtained by automated DIA was significantly linked to prognosis in multivariate analysis adjusted by age, grade, ER, PgR and HER2 status as well as T status [26]. To ensure comparability between SQ's and DIA's prognosis prediction potential, we have utilized the widely applied 14% and 20% cut-offs for

Table 4 Univariate Cox regression analysis of KIPI assessments and pathological factors in the different treatment groups. Only significant factors shown

Treatment groups	Prognostic factor	Subgroups	HR	95% CI	*p*-value
Surgical treatment only (n = 111)	IHC Subtype	TNBC	6.642	1.934–22.815	0.003
	Lymph node status (TNM 7)	1	2.652	1.072–6.562	0.035
	Mitotic index	>19	3.584	1.076–11.935	0.038
	SQ-1_D14%	-	4.975	1.619–15.287	0.005
	SQ-2_D20%	-	6.836	2.194–21.303	≤0.001
	SQ-3_D20%	-	3.514	1.217–10.144	0.020
	DIA-1_D20%	-	3.364	1.098–10.307	0.034
	DIA-2_D20%	-	4.181	1.402–12.467	0.010
Surgery + Irradiation (n = 83)	Tumor size	2-5 cm	2.725	1.046–7.102	0.040
Surgery + Chemotherapy (n = 30)	IHC Subtype	Luminal-B like	6.529	1.147–37.165	0.034
	SQ-2_D14%	-	3.018	1.120–9.568	0.049
Surgery + Chemotherapy + Irradiation (n = 117)	IHC Subtype	HER2	2.923	1.333–6.406	0.007

D14% dichotomized at 14% threshold, *D20%* dichotomized at 20% threshold

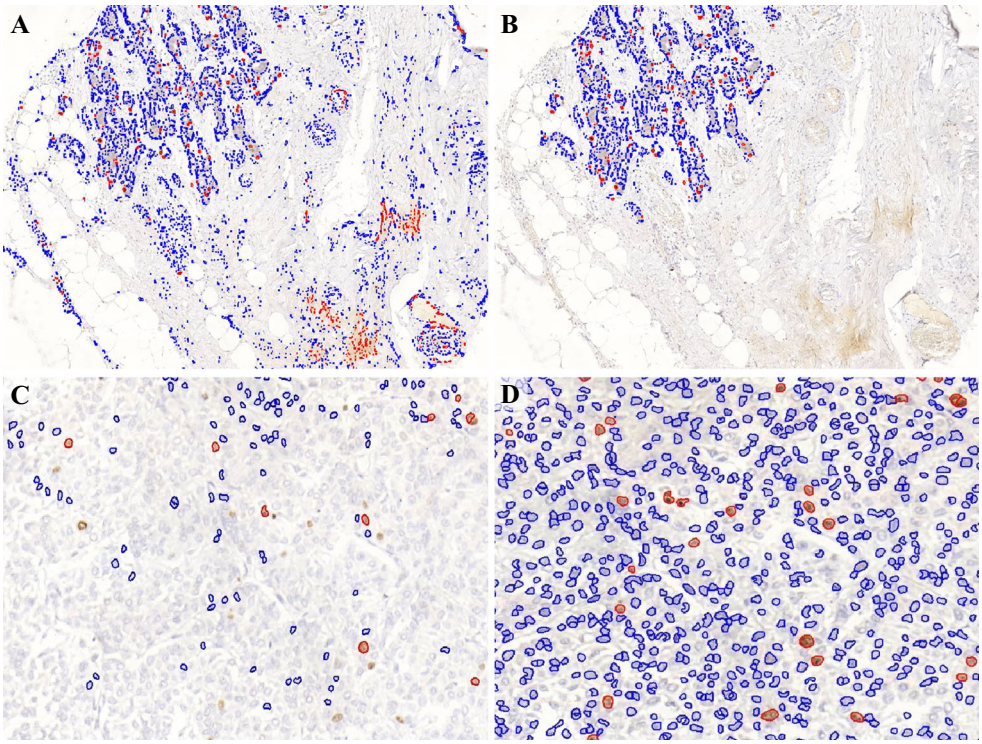

Fig. 5 False detections were observed due to irrelevant Ki-67 staining (**a**) with automated DIA (DIA-1). These issues could be controlled by adjustable DIA method (DIA-1) with the presence of the pathologist (**b**). Automated DIA (automated intensity threshold setting) was not able to recognize most of the tumor cells in some cases (**c**). The reason for underestimated cell recognition is the inadequate quality of tissue processing. However, with adjustment of DIA (DIA-2 with adjustable intensity threshold), the vast majority of tumor cells were detected (**d**)

each assessment. In our hands none of the KIPI evaluations (regardless of DIA and SQ methods) were significantly linked to DFS at 14% threshold. However at 20% threshold one of the three SQ assessments (SQ-2) was an independent prognostic factor besides lymph node status. To exclude bias related to different treatment protocols, we have also investigated the prognosis prediction potential of KIPI assessments in each treatment subgroup. All KIPI evaluations but SQ-1 were able to distinguish good and unfavorable patient cohorts at 20% cut-off in the surgical treatment only subgroup, while in the patient subgroup treated with surgery + chemotherapy SQ-2 was able to perform statistically splitting the cohort. In treatment subgroups of surgery + irradiation and surgery + irradiation + chemotherapy combination no prognosis prediction potential was observed for any KIPI evaluation.

The limitation of our retrospective study is that KIPI evaluations were performed on TMA slides, which might raise the possibility of underrepresented tumor areas related to prognosis prediction, even if we have used two cores from each case. Furthermore, we could retrieve DFS only from clinical data. Although preanalytical and analytical steps were not standardized in the contemporary terms, all of the cases were collected from a single hospital resulting in uniform preanalytical conditions. Thus fixation, tissue processing and other preanalytical issues affected the immunohistochemical results uniformly. Clinical data related to chemotherapy protocols were not available. Thus, we were not able to investigate predictive significance of KIPI for different chemotherapy regimens. In treatment stratified analyses, multivariate Cox regression was not performed due to low number of cases compared with events to relatively many clinicopathological factors.

In summary, we provide further evidence that KIPI is a significant prognostic marker in breast cancer. The pathologists' experience is essential to control and adjust DIA and to avoid false detections. We also demonstrate that the adjustable DIA can be a feasible and reproducible tool to evaluate KIPI in breast cancer which may support standardization efforts.

Acknowledgements Balazs Acs was supported by the Hungarian Talent Program, National Young Talent Award 2016 (NTP-NFTÖ-16-0496). Anna-Maria Tokes was supported by the Hungarian Society of Medical Oncology (MKOT) 2014-2016, and A. Marcell Szasz was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

We thank Gabor Kiszler Ph.D. for his innovative technical and IT support and Zsuzsanna Bodor M.D. for her excellent remarks on this study. We acknowledge Professor Zoltán Prohászka for his constructive comments on statistical analyses. We also appreciate Rita Keszthelyi's efforts made on implementation of immunohistochemical reactions.

Contribution of Authors BÁ carried out the conception and design of the study, performed digital-image analysis and statistical analyses, wrote the manuscript. LM evaluated Ki-67 IHC and revised the manuscript. KAK participated in the evaluation of Ki-67 IHC and drafted the manuscript. TM participated in digital-image analysis and helped to revise the manuscript. AMT reviewed follow-up data and drafted the manuscript. BG helped to interpret data and wrote the manuscript. JK designed and coordinated the study and revised the manuscript. AMSZ conceived of the study, and participated in its design and coordination, also evaluated Ki-67 IHC reactions, drafted the manuscript. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval The study was approved by the Institutional Review Board of Semmelweis University (IKEB, #7–1/2008). This article does not contain any studies with animals performed by any of the authors.

Informed Consent The Institutional Review Board of Semmelweis University was consulted for patient consent requirement, written approval was unnecessary due to the nature of the study, namely routine material was utilized for the evaluation of Ki-67 without any identification of the patients and their personal data. The Helsinki Declaration of 1964 was followed in any such survey.

References

1. Senkus E, Kyriakides S, Penault-Llorca F, Poortmans P, Thompson A, Zackrisson S, Cardoso F (2013) Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 24(Suppl 6):vi7–v23. doi:10.1093/annonc/mdt284
2. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674. doi:10.1016/j.cell.2011.02.013
3. Gerdes J, Schwab U, Lemke H, Stein H (1983) Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. Int J Cancer 31(1):13–20
4. Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart M, Thurlimann B, Senn HJ (2015) Tailoring therapies-improving the management of early breast cancer: St Gallen international expert Consensus on the Primary therapy of early breast cancer 2015. Ann Oncol 26(8):1533–1546. doi:10.1093/annonc/mdv221
5. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, Senn HJ (2013) Personalizing the treatment of women with early breast cancer: highlights of the St Gallen international expert Consensus on the Primary therapy of early breast cancer 2013. Ann Oncol 24(9):2206–2223. doi:10.1093/annonc/mdt303
6. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC Jr, American Society of Clinical Oncology (2007) American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 25(33):5287–5312. doi:10.1200/JCO.2007.14.2364
7. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, Ellis M, Henry NL, Hugh JC, Lively T, McShane L, Paik S, Penault-Llorca F, Prudkin L, Regan M, Salter J, Sotiriou C, Smith IE, Viale G, Zujewski JA, Hayes DF (2011) Assessment of Ki67 in breast cancer: recommendations from the international Ki67 in breast cancer working group. J Natl Cancer Inst 103(22):1656–1664. doi:10.1093/jnci/djr393
8. Polley MY, Leung SC, Gao D, Mastropasqua MG, Zabaglo LA, Bartlett JM, McShane LM, Enos RA, Badve SS, Bane AL, Borgquist S, Fineberg S, Lin MG, Gown AM, Grabau D, Gutierrez C, Hugh JC, Moriya T, Ohi Y, Osborne CK, Penault-Llorca FM, Piper T, Porter PL, Sakatani T, Salgado R, Starczynski J, Laenkholm AV, Viale G, Dowsett M, Hayes DF, Nielsen TO (2015) An international study to increase concordance in Ki67 scoring. Mod Pathol 28(6):778–786. doi:10.1038/modpathol.2015.38
9. Polley MY, Leung SC, McShane LM, Gao D, Hugh JC, Mastropasqua MG, Viale G, Zabaglo LA, Penault-Llorca F, Bartlett JM, Gown AM, Symmans WF, Piper T, Mehl E, Enos RA, Hayes DF, Dowsett M, Nielsen TO (2013) An international Ki67 reproducibility study. J Natl Cancer Inst 105(24):1897–1906. doi:10.1093/jnci/djt306
10. Luporsi E, Andre F, Spyratos F, Martin PM, Jacquemier J, Penault-Llorca F, Tubiana-Mathieu N, Sigal-Zafrani B, Arnould L, Gompel A, Egele C, Poulet B, Clough KB, Crouet H, Fourquet A, Lefranc JP, Mathelin C, Rouyer N, Serin D, Spielmann M, Haugh M, Chenard MP, Brain E, de Cremoux P, Bellocq JP (2012) Ki-67: level of evidence and methodological considerations for its role in the clinical management of breast cancer: analytical and critical review. Breast Cancer Res Treat 132(3):895–915. doi:10.1007/s10549-011-1837-z
11. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA (2010) Ki67 in breast cancer: prognostic and predictive potential. Lancet Oncol 11(2):174–183. doi:10.1016/s1470-2045(09)70262-1
12. Soenksen D (2009) Digital pathology at the crossroads of major health care trends: corporate innovation as an engine for change. Arch Pathol Lab Med 133(4):555–559. doi:10.1043/1543-2165-133.4.555
13. Kayser K, Borkenfeld S, Kayser G (2012) How to introduce virtual microscopy (VM) in routine diagnostic pathology: constraints, ideas, and solutions. Anal Cell Pathol (Amst) 35(1):3–10. doi:10.3233/acp-2011-0044
14. Kayser K, Gortler J, Borkenfeld S, Kayser G (2011) How to measure diagnosis-associated information in virtual slides. Diagn Pathol 6(Suppl 1):S9. doi:10.1186/1746-1596-6-s1-s9
15. Laurinavicius A, Plancoulaine B, Laurinaviciene A, Herlin P, Meskauskas R, Baltrusaityte I, Besusparis J, Dasevicius D, Elie N, Iqbal Y, Bor C (2014) A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. Breast Cancer Res 16(2):R35. doi:10.1186/bcr3639
16. Riber-Hansen R, Vainer B, Steiniche T (2012) Digital image analysis: a review of reproducibility, stability and basic requirements for optimal results. APMIS 120(4):276–289. doi:10.1111/j.1600-0463.2011.02854.x
17. Ruifrok AC, Johnston DA (2001) Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 23(4):291–299
18. Altman DG (1990) Practical statistics for medical research. Chapman and Hall, London
19. McBride GB (2005) A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. NIWA Client Report: HAM2005–062
20. Kontzoglou K, Palla V, Karaolanis G, Karaiskos I, Alexiou I, Pateras I, Konstantoudakis K, Stamatakos M (2013) Correlation between Ki67 and breast cancer prognosis. Oncology 84(4):219–225. doi:10.1159/000346475
21. Mikami Y, Ueno T, Yoshimura K, Tsuda H, Kurosumi M, Masuda S, Horii R, Toi M, Sasano H (2013) Interobserver concordance of Ki67 labeling index in breast cancer: Japan breast cancer research group Ki67 ring study. Cancer Sci 104(11):1539–1543. doi:10.1111/cas.12245
22. Varga Z, Cassoly E, Li Q, Oehlschlegel C, Tapia C, Lehr HA, Klingbiel D, Thurlimann B, Ruhstaller T (2015) Standardization for Ki-67 assessment in moderately differentiated breast cancer. A retrospective analysis of the SAKK 28/12 study. PLoS One 10(4):e0123435. doi:10.1371/journal.pone.0123435
23. Cserni G, Voros A, Liepniece-Karele I, Bianchi S, Vezzosi V, Grabau D, Sapino A, Castellano I, Regitnig P, Foschini MP, Zolota V, Varga Z, Figueiredo P, Decker T, Focke C, Kulka J, Kaya H, Reiner-Concin A, Amendoeira I, Callagy G, Caffrey E, Wesseling J, Wells C (2014) Distribution pattern of the Ki67

labelling index in breast cancer and its implications for choosing cut-off values. Breast 23(3):259–263. doi:10.1016/j.breast.2014.02.003

24. Voros A, Csorgo E, Kovari B, Lazar P, Kelemen G, Rusz O, Nyari T, Cserni G (2015) 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. Pathol Oncol Res 21(1):147–155. doi:10.1007/s12253-014-9800-z
25. Maeda I, Abe K, Koizumi H, Nakajima C, Tajima S, Aoki H, Tsuchiya J, Tsuchiya S, Tsuchiya K, Shimo A, Tsugawa K, Ueno T, Tatsunami S, Takagi M (2015) Comparison between Ki67 labeling index determined using image analysis software with virtual slide system and that determined visually in breast cancer. Breast Cancer. doi:10.1007/s12282-015-0634-7
26. Klauschen F, Wienert S, Schmitt WD, Loibl S, Gerber B, Blohmer JU, Huober J, Rudiger T, Erbstosser E, Mehta K, Lederer B, Dietel M, Denkert C, von Minckwitz G (2015) Standardized Ki67 diagnostics using automated scoring-clinical validation in the GeparTrio breast cancer study. Clin Cancer Res 21(16):3651–3657. doi:10.1158/1078-0432.ccr-14-1283
27. Denkert C, Budczies J, von Minckwitz G, Wienert S, Loibl S, Klauschen F (2015) Strategies for developing Ki67 as a useful biomarker in breast cancer. Breast. doi:10.1016/j.breast.2015.07.017
28. Joshi S, Watkins J, Gazinska P, Brown JP, Gillett CE, Grigoriadis A, Pinder SE (2015) Digital imaging in the immunohistochemical evaluation of the proliferation markers Ki67, MCM2 and Geminin, in early breast cancer, and their putative prognostic value. BMC Cancer 15:546. doi:10.1186/s12885-015-1531-3