Breast Cancer Research and Treatment (2019) 176:557–568
https://doi.org/10.1007/s10549-018-05112-9

PRECLINICAL STUDY

Clinical and analytical validation of Ki-67 in 9069 patients from IBCSG VIII + IX, BIG1-98 and GeparTrio trial: systematic modulation of interobserver variance in a comprehensive in silico ring trial

Carsten Denkert[1,2] · Jan Budczies[1,3] · Meredith M. Regan[4] · Sibylle Loibl[5] · Patrizia Dell'Orto[6] · Gunter von Minckwitz[5] · Mauro G. Mastropasqua[6] · Christine Solbach[7] · Beat Thürlimann[8,9] · Keyur Mehta[5] · Jens-Uwe Blohmer[10] · Marco Colleoni[11,12] · Volkmar Müller[13] · Frederick Klauschen[1] · Beyhan Ataseven[14,15] · Knut Engels[16] · Roswitha Kammler[17] · Berit M. Pfitzner[1] · Manfred Dietel[1] · Peter A. Fasching[18] · Giuseppe Viale[19,20]

Received: 14 December 2018 / Accepted: 18 December 2018 / Published online: 7 May 2019

Abstract

Purpose Ki-67 has been clinically validated for risk assessment in breast cancer, but the analytical validation and cutpoint-definition remain a challenge. Intraclass correlation coefficients (ICCs) are a statistical parameter for Ki-67 interobserver performance. However, the maximum degree of variance among pathologists allowed for meaningful biomarker results has not been defined.

Methods Different amounts of variance were added to central pathology Ki-67 data ($n = 9069$) from three cohorts (IBCSG VIII + IX, BIG1-98, GeparTrio) by simulation of 4500 evaluations for each cohort, which were grouped by ICCs, ranging from excellent (ICC = 0.9) to poor concordance (ICC = 0.1). Endpoints were disease-free survival (DFS) and pathological complete response (pCR, GeparTrio).

Results Ki-67 was a significant continuous prognostic marker for DFS over a wide range of cutpoints between 8% and 30% in all three cohorts. In our modelling approach, Ki-67 was a stable prognostic marker despite increased interpathologist variance. Even for a poor ICC of 0.5, one or more significant Ki-67 cutoffs were detected in 86.8% (GeparTrio), 92.4% (IBCSG VIII + IX) and 100% of analyses (BIG1-98). Similarly, in GeparTrio, even with an extremely low ICC of 0.2, 99.6% of analyses were significant for pCR.

Conclusions Our study shows that Ki-67 is a continuous marker which is extremely robust to pathologist variation. Even if only 50% of variance is attributable to true Ki-67-based proliferation (ICC = 0.5), this information is sufficient to obtain statistically significant differences in clinical cohorts. This stable performance explains the observation that many Ki-67 studies achieve significant results despite relevant interobserver variance and points to a high clinical validity of this biomarker. For clinical decisions based on analysis of individual patient data, ongoing efforts to further reduce interobserver variability, including ring trials and standardized guidelines as well as image analysis approaches, should be continued.

Keywords Ki-67 · Breast cancer · Prediction · Prognosis · Neoadjuvant · Adjuvant

Introduction

Despite more than 30 years of research [1], the utility of the proliferation marker Ki-67 for clinical decisions is still under discussion [2–5]. Several metaanalyses have summarized the available data on Ki-67 in different clinical cohorts [6–11]. The main conclusion is that an increased Ki-67 is significantly linked to an improved response to neoadjuvant chemotherapy [12] and—at the same time—is a negative prognostic marker. The group of tumors with increased Ki-67 therefore generally has an impaired prognosis, even though

Carsten Denkert and Jan Budczies contributed equally to the publication.

✉ Carsten Denkert
Carsten.denkert@uni-marburg.de

Extended author information available on the last page of the article

some tumors exist in this group that show an increased response to chemotherapy. In different studies, Ki-67 cutpoints between 1% and 28% have been used [10, 13, 14].

The main limitation for a broad clinical implementation of Ki-67 is the limited analytical validity and, in particular, the interobserver variability between different pathologists.

The International Ki-67 Working Group [15] has conducted several Ki-67 ring trials [16–18], and additional validation projects are ongoing. In these projects, the intraclass correlation coefficient (ICC) has emerged as the single most useful parameter for evaluation of Ki-67 interobserver variance. The ICC–with values between 0 and 1–reflects the variation in the data that is based on the true biological variability of the marker. A high ICC suggests that the main variance derives from the true biological differences, while a low ICC indicates a considerable contribution of pathologist variation to the measurement of the marker.

In the first Ki-67 ring trial, ICCs between 0.94 (for intralaboratory reproducibility) and 0.59 (for interlaboratory reproducibility with local staining and evaluation) have been reported [16]. In the ring trial phase 2, the ICC for central staining and local evaluation was 0.94 after a mandatory calibration exercise [17].

For the successful conduction of ring trials, an observed ICC significantly over 0.70 (for the second phase) [17] or >0.8 (for the third phase) was predefined [18]. These ICC cutpoints are based on expert opinion and not derived from analysis of data, and the amount of pathologist variance that is allowed for meaningful biomarker results is still not defined. The problem is that ring trials typically evaluate a small number of tumors so that clinical endpoints cannot be analyzed. On the other hand, large studies with clinical outcome data are typically performed in a central laboratory, and it is not feasible to perform the evaluation of large datasets in parallel by several pathologists with the aim to determine an ICC.

In this study, we investigated the degree of analytical variation among pathologists that is needed for a clinically meaningful biomarker result for Ki-67. We asked the questions: How would the results of a Ki-67 biomarker study change with increasing magnitude of variation among pathologists? What degree of variation between pathologist leads to a loss of significance of Ki-67 for clinical endpoints such as chemotherapy response and prognosis?

The traditional approach to address this question would be the evaluation of Ki-67 in a clinical study cohort with existing outcome data by large number of pathologists and a subsequent statistical evaluation for different pathologist groups with different ICCs. However, this approach is not feasible in the real world, as the workload for the participating pathologists in an adequately powered clinical study cohort would be too high. Furthermore, if the only purpose of this evaluation is to add different degrees of background noise to the Ki-67 data, it is not necessary to involve real human pathologists.

Therefore, we used a Monte Carlo simulation approach with mathematical modeling for an "in silico Ki-67 ring trial" based on existing Ki-67 datasets from three clinical trials with a total of 9069 participants. Carrying out the in silico ring trial, we perturbed the existing data to simulate a total of 4500 virtual pathologists, each of them has evaluated each single tumor from the three independent clinical cohorts. This results in a total of 40.810.500 in silico Ki-67 measurements. The pathologists were divided into nine groups of 500 pathologists with ICCs for each group between 0.1 and 0.9, separately for each clinical trial cohort. We evaluated the influence of the different ICCs on the significance of the clinical endpoints pathological complete response (pCR) in the neoadjuvant GeparTrio cohort as well as disease-free survival (DFS) in all three clinical studies. The analysis was performed with Ki-67 as a continuous marker and with a systematical analysis of different cutpoints using the cutoff finder approach [19].

Patients and methods

Clinical study cohorts

This study is based on Ki-67 evaluations in a total of 9069 tumor samples from three clinical cohorts (IBCSG Trials VIII + IX, IBCSG-led BIG 1–98, GeparTrio) [20–22]. The BIG trial 1–98 (NCT00004205) was a randomized adjuvant phase III trial comparing 5 years of letrozole or tamoxifen, or sequential treatment with 2 years of one agent followed by 3 years of the other, in postmenopausal women with hormone-receptor positive early breast cancer. From the 8010 patients randomized between 1998 and 2003, a total of 6090 (76%) paraffin blocks were available for central Ki-67 [23, 24].

The IBCSG Trials VIII + IX included a total of 2732 patients enrolled between 1988 and 1999 and compared adjuvant endocrine therapy alone with sequential chemo-endocrine therapy in premenopausal (VIII, $n=1063$) and postmenopausal (IX, $n=1669$) women with node-negative breast cancer. Randomization was stratified by estrogen receptor (ER) status, and in 1998, enrollment was limited to ER + disease. A total of 1907 samples (70%) were evaluated centrally for Ki-67 [11]. In the BIG 1–98 and IBCSG VIII + IX trials, hormone receptor positive status was defined as ER $\geq 1\%$ and/or PR $\geq 1\%$.

The GeparTrio trial (NCT00544765) was a neoadjuvant trial including primary breast cancer. Details of the eligibility criteria and chemotherapy have been published before [25] Ki-67 was determined centrally in 1166 tumor samples (56%) [14]. In GeparTrio, hormone receptor positive status

was defined as at least 10% positive tumor cells. The current dataset was derived from the updated GBG MetaDB V20160226.

For all trials, institutional review boards/ethics committees reviewed and approved the trial protocols, and informed consent was obtained according to the criteria established within the individual countries.

Definition of endpoints

In GeparTrio, pCR was defined as no invasive or noninvasive residual cancer in breast and lymph nodes (ypT0, ypN0). DFS was defined as the time from randomization to the earliest time of invasive recurrence at local, regional or distant sites, a new invasive cancer in the contralateral breast, any second (non-breast) malignancy, or death from any cause. In the absence of an event, DFS was censored at the date of last follow-up.

The analysis was performed based on a predefined statistical analysis plan that was agreed upon before analysis by the collaborating clinical groups. An evaluation of different therapies was not part of this analysis plan, because this had already been reported previously for each trial.

Cutoff finder approach

As a first step, we used the cutoff finder approach for a systematical evaluation of all possible Ki-67 cutoffs in the three clinical trial cohorts. The influence of the cutoff point for Ki-67 positivity on ORs for pCR and on HRs for DFS was investigated using the cutoff finder method [26]. Plots of ORs and HRs including 95% confidence intervals were generated for each possible cutoff point.

Statistical analysis–mathematical modeling approach

In a systematic modeling approach, we simulated Ki-67 analysis of each single tumor by a total of 4500 pathologists by perturbing the existing central pathology Ki-67 datasets (Fig. 1). The analysis was conducted independently for each clinical cohort. Monte-Carlo simulations ("in silico ring trials") were carried out to analyze the effect of interpathologist variance on Ki-67 evaluation in GeparTrio, BIG 1–98 and IBCSG VIII-IX. To this end, Ki-67 data were perturbed to simulate nine groups of pathologists with 500 observers in each of the groups. Perturbation strengths were increased to result in decreasing intraclass correlations of ICC = 0.9, 0.8, …, 0.1 in the nine groups. ICCs were calculated using the function icc from the R package irr with the arguments model = twoway, type = agreement and unit = single [27]. This corresponds to the version ICC(2, 1) of intraclass correlation coefficients in the notation of Shrout and Fleiss [28]. Performing the perturbations, it was taken into account that Ki-67 is easier to evaluate when very low or very high (close to 0% or 100%) and more difficult to evaluate when there are both many positive and many negative cells. In detail, given a value of x% for the percentage of Ki-67 positive cells, an additive perturbation was drawn from a normal distribution with standard deviation proportional to x for $x \leq 50\%$ and proportional to $100\% - x$ for $x > 50\%$. Values smaller than 0% after perturbation were set to 0%, and values larger than 100% after perturbation were set to 100%.

These evaluations were grouped into 9 groups of 500 pathologists with defined ICCs, ranging from very good concordance (ICC = 0.9) to extremely poor concordance (ICC = 0.1). As endpoints, we used HR for DFS for all three study cohorts and in addition OR for pCR for the neoadjuvant cohort. As a first step, we evaluated Ki-67 as a continuous marker in the different pathologist groups with different ICCs.

Results

Clinical study cohorts and baseline parameters

As shown in Table 1, the clinical cohorts were different regarding clinicopathological variables, which is in line with the different clinical settings. In GeparTrio, 32.8% of patients had T3–T4 tumors, compared to only 1.1% in IBCSG VIII + IX and 2.8% in BIG 1–98. IBCSG VIII + IX included only patients with node negative disease, while in BIG 1-98 (42%) and GeparTrio (54%) also patients with node-positive tumors were included. In BIG 1-98, 98.5% of patients had centrally confirmed hormone receptor positive tumors (ER ≥ 1% and/or PgR ≥ 1%), while IBCSG VIII + IX had 80.4% and GeparTrio 68.5% hormone receptor positive tumors.

Reflecting the more aggressive tumor phenotype in GeparTrio, the mean Ki-67 was 32% in this trial, compared to 25% in IBCSG VIII + IX and 14% in BIG-1-98. Interestingly, the mean Ki-67 values were more similar in the three cohorts if stratified for HR and HER2 status (details see Table 1). The distribution of Ki-67 values is shown in Fig. 2a–c.

Systematical evaluation of different Ki-67 cutpoints in all three clinical cohorts

The cutoff finder tool has been designed for a convenient and comprehensive evaluation of quantitative biomarkers in cancer. In a previous study, we have used this tool to evaluate different cutoffs for Ki-67 in the GeparTrio cohort. We have shown that many different cutoffs to define Ki-67 positive vs. negative are significant for pCR as well as DFS and that

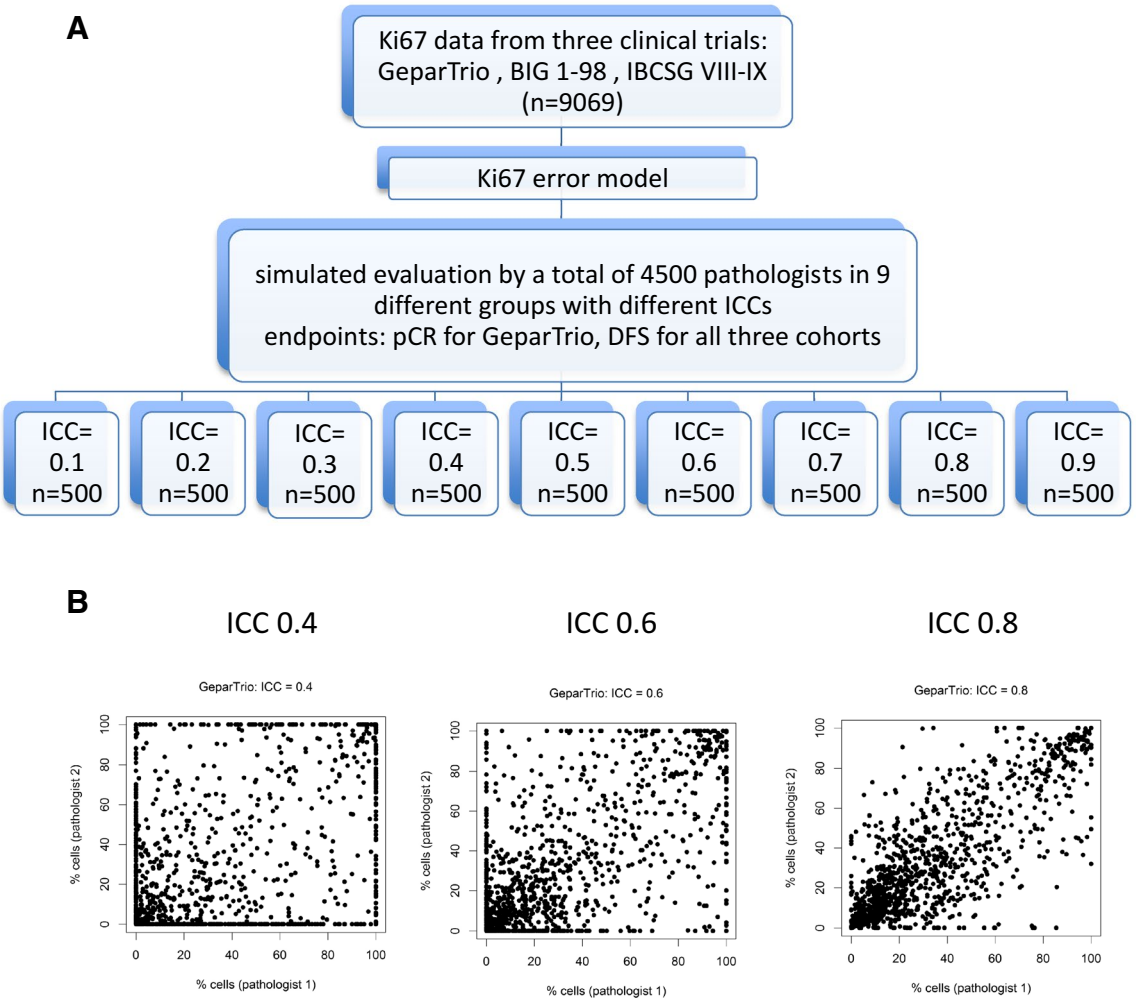

Fig. 1 **a** Overview. The in silico ring trial evaluates the actual clinical relevance of different levels of pathologist concordance. The three Ki-67 datasets were perturbed to simulate the evaluation by a total of 4500 pathologists that were divided into 9 pathologist groups with different ICCs, three examples for ICC 0.4, 0.6 and 0.8 are shown below. **b** Examples for comparison of two pathologists with defined ICCs of 0.4, 0.6 and 0,.8; illustrating the increased concordance with higher ICCs. The examples are from the GeparTrio dataset

it is not possible to define the "best" cutoff. The additional datasets from BIG 1-98 and IBCSG VIII + IX provide the opportunity to validate this observation for DFS.

As shown in Fig. 2d, Ki-67 was significant over a wide range of cutpoints for DFS in all three studies. The data on GeparTrio, with significant *p* values for DFS for all Ki-67 cutoffs values between 5% and 52%, have been published before and are shown for comparison. For IBCSG VIII + IX, the results are similar to GeparTrio, with significant *p* values for Ki-67 cutoffs between 7.5% and 32.5% (Fig. 2).

For the large hormone-receptor positive BIG 1-98 cohort investigating endocrine therapies, all Ki-67 cutpoints between 1.5% and 79% were significant for prediction of DFS. In BIG 1-98 the hazard ratio for reduced DFS was continuously increased with increased Ki-67 values, similar to previous evaluations [23].

In silico ring trial for evaluation of Ki-67 as a marker for pCR and prognosis

In the in silico ring trial approach presented here, we used mathematical modeling to simulate the evaluation of Ki-67 as a biomarker in the complete dataset of 9069 tumors by 4500 individual pathologists. This comprehensive in silico modeling approach resulted in a complex dataset of 40,810,500 individual Ki-67 values. To assess the impact of different degrees of variability between pathologists, the 4500 pathologists were divided into 9 groups of each 500 pathologists, with different ICCs for each group (Fig. 1a). These ICCs ranged from 0.9, which is well accepted as a high concordance, to 0.1, which indicates a very poor concordance. Typical examples of pathologists with ICCs of 0.4, 0.6 and 0.8 are shown in Fig. 1b, visualizing the

Table 1 Clinicopathological parameters of the three cohorts

Parameter	GeparTrio cohort *N* (%)	IBCSG VIII + IX cohort *N* (%)	BIG 1-98 cohort *N* (%)
Number of samples	1166	1926	6107
Age (years)			
< 40	187 (16.0%)	152 (7.9%)	6 (0.1%)
40–49	378 (32.4%)	499 (25.9%)	256 (4.2%)
50–59	336 (28.8%)	603 (31.35)	2368 (38.8%)
≥ 60	265 (22.7%)	672 (34.9%)	3477 (56.9%)
Median age	50	55	61
Tumor stage[†]			
T1	15 (1.3%)	1129 (58.6%)	3779 (61.9%)
T2	764 (65.5%)	729 (38.3%)	2117 (34.7%)
T3	224 (19.2%)	22 (1.1%)	171 (2.8%)
T4	159 (13.6%)	0 (0%)	0 (0%)
Unknown	4 (0.3%)	39 (1.9%)	40 (0.7%)
Nodal status[†]			
N0	530 (45.5%)	1926 (100%)	3481 (57%)
N +	624 (53.5%)	0 (0%)	2566 (42%)
Unknown	12 (1.0%)	0 (0%)	60 (1.0%)
Tumor grade			
G1	46 (3.9%)	268 (13.9%)	1235 (20.2%)
G2	654 (56.1%)	910 (47.2%)	3438 (56.3%)
G3	407 (34.9%)	738 (38.3%)	1423 (23.3%)
Unknown	59 (5.1%)	36 (1.9%)	11 (0.2%)
Hormone receptor status			
Negative	359 (30.8%)	376 (19.5%)	77 (1.3%)
Positive	782 (67.1%)	1548 (80.4%)	6015 (98.5%)
Unknown	25 (2.1%)	2 (0.1%)	11 (0.2%)
HER2 status			
Negative	819 (70.2%)	1601 (69.2%)	5694 (93.2%)
Positive	271 (23.2%)	308 (16.0%)	405 (6.6%)
Unknown	76 (6.5%)	17 (0.9%)	8 (0.1%)
Molecular tumor type			
HR+/HER2−	569 (48.8%)	1332 (69.2%)	5629 (92.2%)
HR+/HER2+	166 (14.2%)	201 (10.4%)	384 (6.3%)
HR−/HER2+	101 (8.7%)	107 (5.6%)	20 (0.3%)
TNBC	236 (20.2%)	267 (13.9%)	57 (0.9%)
Unknown	94 (8.1%)	19 (1.0%)	17 (0.3%)
Mean Ki67 (%)	32	25	14
Mean Ki67 (%) in molecular subtype			
HR+/HER2−	23	20	13
HR+/HER2+	27	25	22
HR−/HER2+	37	34	28
TNBC	53	47	35
Chemotherapy response			
pCR	184 (15.8%)	na	na
No pCR	982 (84.2%)	na	na
Not applicable	0 (0%)	1926 (100%)	6107 (100%)
Mean DFS (years)	7.79 ± 0.12	8.41 ± 0.07	8.52 ± 0.04
Type of therapy			
Neoadjuvant chemotherapy	1166 (100%)	na	na
Adjuvant chemotherapy			
Yes	0%	1094 (56.8%)	1396 (22.9%)
No	0%	832 (43.2%)	4711 (77.1%)

Table 1 (continued)

Parameter	GeparTrio cohort *N* (%)	IBCSG VIII + IX cohort *N* (%)	BIG 1-98 cohort *N* (%)
Adjuvant endocrine therapy			
Yes	–	1667 (86.5%)	6107 (100%)
No	–	259 (13.4%)*	0 (0%)
No data in CRF	1166 (100%)**		

*The IBCSG Trial VIII included also a third treatment group receiving chemotherapy alone

**For GeparTrio, endocrine therapy was not captured but was recommended according to the current guidelines

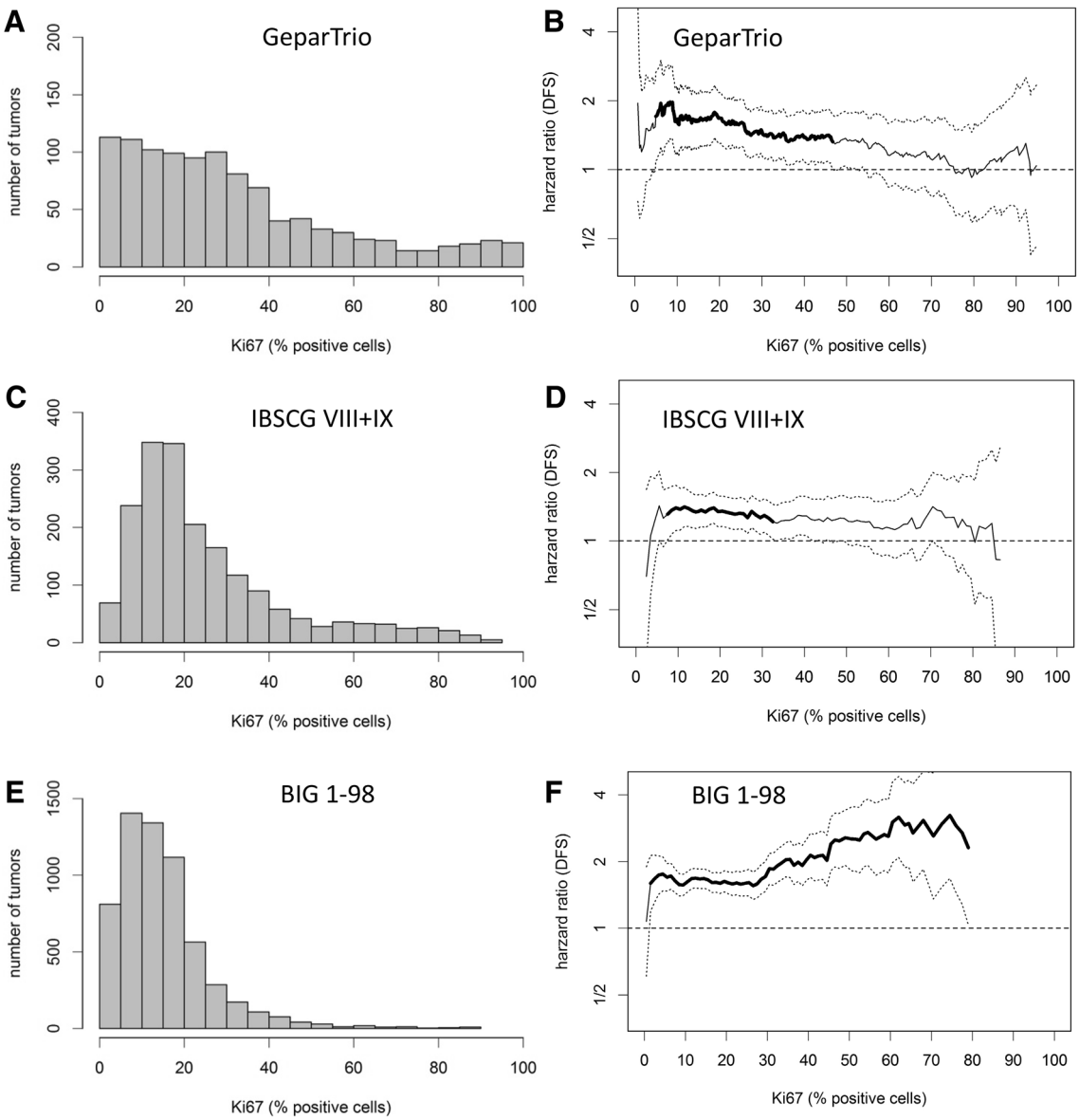

Fig. 2 Distribution of Ki-67 values (**a, c, e**) and systematical evaluation of cutpoints (**b, d, f**) in the three clinical trial cohorts **a, b** GeparTrio, **c, d** IBCSG VIII + IX, **e, f** BIG 1–98. **b, d, f** Cutoff finder approach: Systematical evaluation of cutoff points for the percentage of Ki-67 positive cells to define a tumor as Ki-67 positive. DFS HRs (Ki-67 positive vs. negative) with 95% confidence intervals are plotted each possible cutoff point. The range of cutoffs resulting in statistically significant ($p < 0.05$) differences of prognostic association of Ki-67 positivity with of DFS is marked by thick lines. In all three cohorts, at least all cutpoints between 8% and 30% were significant

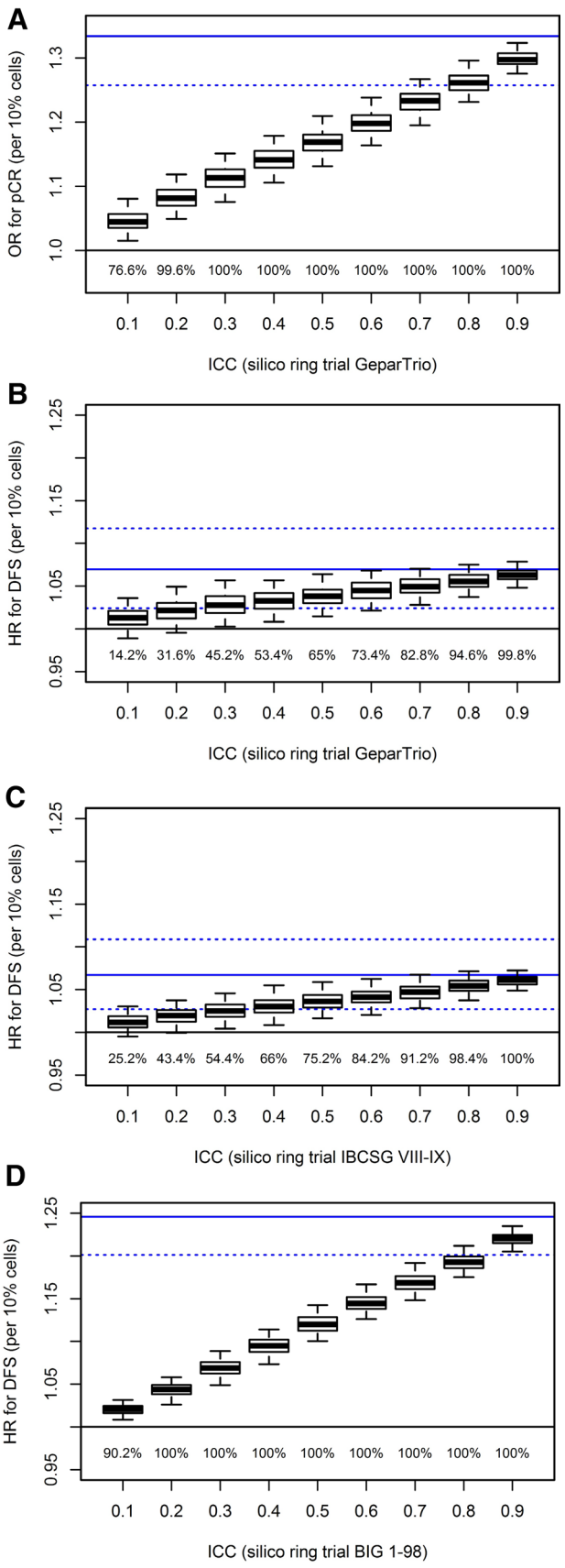

Fig. 3 Mathematical modeling of differences in pathologist variation for Ki-67 ("in silico ring trial"). The existing data from three clinical trial cohorts were used to model different degrees of pathologist variation in 9 different simulated groups, each comprising 500 pathologists. The results are shown for Ki-67 as a continuous variable for the association of Ki-67 with response to neoadjuvant chemotherapy (pCR) in GeparTrio (**a**), as well as for DFS in all three cohorts: GeparTrio (**b**), IBCSG VIII + IX (**c**) and BIG 1-98 (**d**). Each box visualizes the results, as OR or HR for the association of Ki-67 with the clinical endpoint, of a group of 500 virtual pathologists (5%, 25%, 50%, 75% and 95% quantiles). The numbers below are the percentage of pathologists in each ICC group with statistically significant results for association of Ki-67 with pCR or DFS. The blue lines indicate the actual OR or HR results of the clinical trial cohort in the original central analysis. The ORs and HRs are remarkably stable over a wide range of different ICCs. *ICC* intraclass correlation coefficient, *OR* odds ratio, *CI* confidence interval

different degrees of agreement. The analysis was performed separately for each clinical trial cohort.

As the next step, we performed a statistical analysis to assess whether Ki-67 was a statistically significant predictor for pCR (GeparTrio) as well as prognostic marker for DFS (all three cohorts) for each of the 4500 pathologists. This approach allows us to analyze Ki-67 as a biomarker for pCR and/or DFS for each single pathologist and to correlate the ORs or HRs with the interobserver variance measured as ICC. As a first step, the analysis was performed using Ki-67 as a continuous marker. Figure 3 shows the odds ratio for the association of Ki-67 with pCR (A, GeparTrio) as well as the DFS hazard ratios (B–D, all three cohorts) for nine different groups of pathologists with nine different ICCs. Each box plot shows the results of 500 pathologists within the group defined by ICC values (box plots show 5%, 25%, 50%, 75% and 95% quantiles of the 500 estimated odds ratios or hazard ratios). The number below indicates the percentage of pathologists from each ICC group, whose Ki-67 scoring results show a statistically significant association with pCR or DFS. For example, for GeparTrio and the pCR endpoint, 100% of the pathologists with an ICC of 0.5 showed a statistically significant association of Ki-67 with pCR. We note, however, that—while the *p* values are still significant—the odds ratio decreased with lower ICC values, indicating that information on the size of the effect is diminishing.

Evaluation of different cutpoints in all three study cohorts

In addition to the analyses of Ki-67 as a continuous marker, we also evaluated the impact of pathologist variation on the association of Ki-67 positive versus negative with pCR and DFS using different Ki-67 cutoffs. The results are shown for pCR (GeparTrio) in Fig. 4 and for DFS (all three cohorts) in Fig. 5. Three of the 9 pathologist groups with high (ICC = 0.8) moderate (ICC = 0.6) and poor (ICC = 0.4) agreement are shown. This analysis shows that

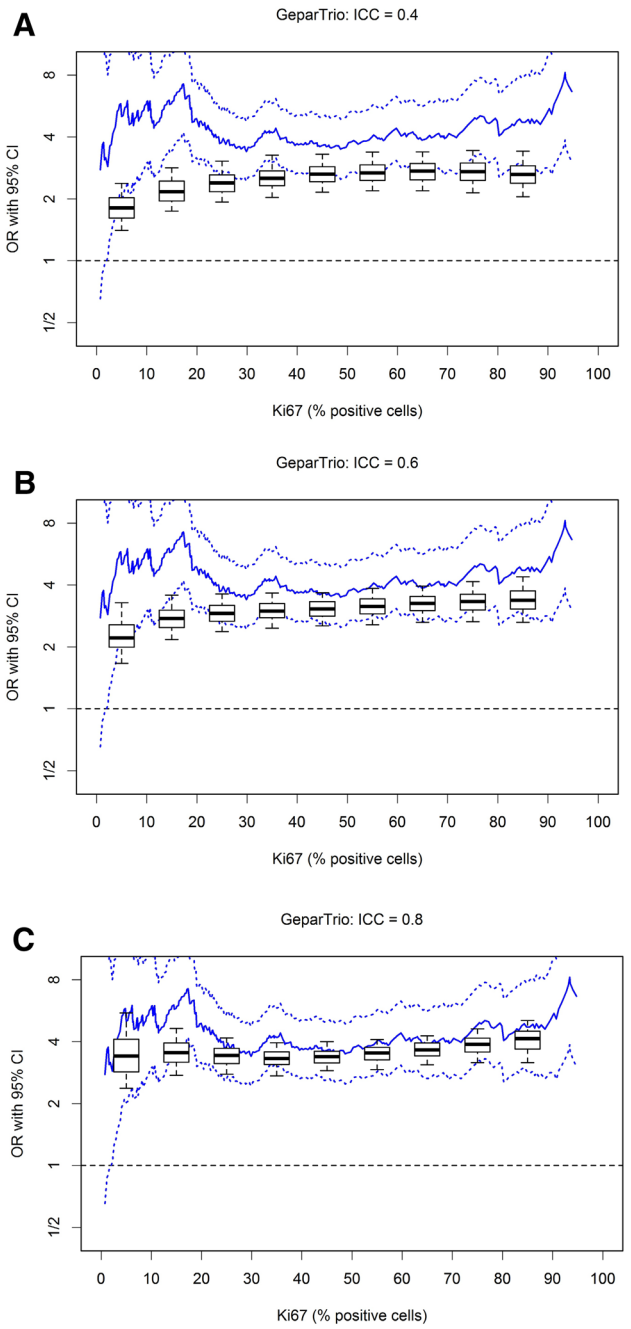

Fig. 4 Mathematical modeling of differences in pathologist variation–relevance of different Ki-67 cutpoints for pCR in GeparTrio. Three different virtual pathologist groups of 500 pathologists with either poor performance (ICC = 0.4; **a**), or moderate performance (ICC = 0.6, **b**) or good performance (ICC = 0.8, **c**) are shown. Each box visualizes the OR for prediction of pCR for the results of a group of 500 virtual pathologists (5%, 25%, 50%, 75% and 95% quantiles) and a given Ki-67 cutpoint. Different cutpoints between Ki-67 of 5% and 95% are systematically evaluated. The blue lines indicate the actual OR for prediction of pCR (with 95% CI) based on the central Ki-67 evaluation in the GeparTrio trial. The ORs are remarkably stable over a wide range of different ICCs. *ICC* intraclass correlation coefficient, *OR* odds ratio, *CI* confidence interval

the prognostic significance of Ki-67 is observed in the large majority of evaluations, even with poor pathologist concordance of 0.4, though again, with reduced HRs, that indicate an attenuation of the "true" estimation of prognostic association of Ki-67 positivity with the endpoints.

Discussion

Our results suggest that the prognostic role of Ki-67 on a trial basis is extremely robust to variation between pathologists. This stable performance of Ki-67 provides an explanation for the observation that many Ki-67 studies achieve statistically significant prognostic association despite the interobserver variance and heterogeneity issues. It might also suggest a relevant clinical utility for Ki-67 despite considerable variation introduced in the evaluation. Our analysis shows that for a clinical trial cohort, the introduction of Ki-67 testing will provide significant prognostic information that would not be available without Ki-67 testing and that for the cohort of patients the prognostic assessment is significant despite the variation among pathologists. This indicates the tumor proliferation is a strong driver of prognosis, and therefore—from our point of view—the proliferation rate should be known for each patient as a part of the biological assessment of the tumor.

As a cautionary statement, it should be noted that this stable performance of Ki-67 is only observed if we evaluate this marker on a trial level looking at a large cohort of patients. On a patient level, there might still be situations where there is a considerable disagreement of different pathologists in Ki-67 evaluation. This is very similar to the situation observed for clinical validation of a new therapy: The new therapy could be significantly superior to the standard-of-care in a large clinical trial, nevertheless there will always be individual patients that do not benefit from this therapy, that will develop adverse effects and that might even show a progress of the disease during the new therapeutic approach. For an oncological therapy, we are used to the fact that not all patients do respond, but for a biomarker, the general expectance is that there should be a 100% agreement in all situations. Therefore, discussions on the limited validity of Ki-67 typically focus on a small number of patient with discrepant results, instead of looking at the generally superior outcome of the complete cohort. It should be noted that on the one hand, we should take all necessary steps to ensure that each single patient will receive the best possible Ki-67 measurement, but on the other hand the existence of single patients with an discordant Ki-67 measurement does not allow the conclusion that the marker is not a useful addition to the diagnostic assessment. This conclusions should

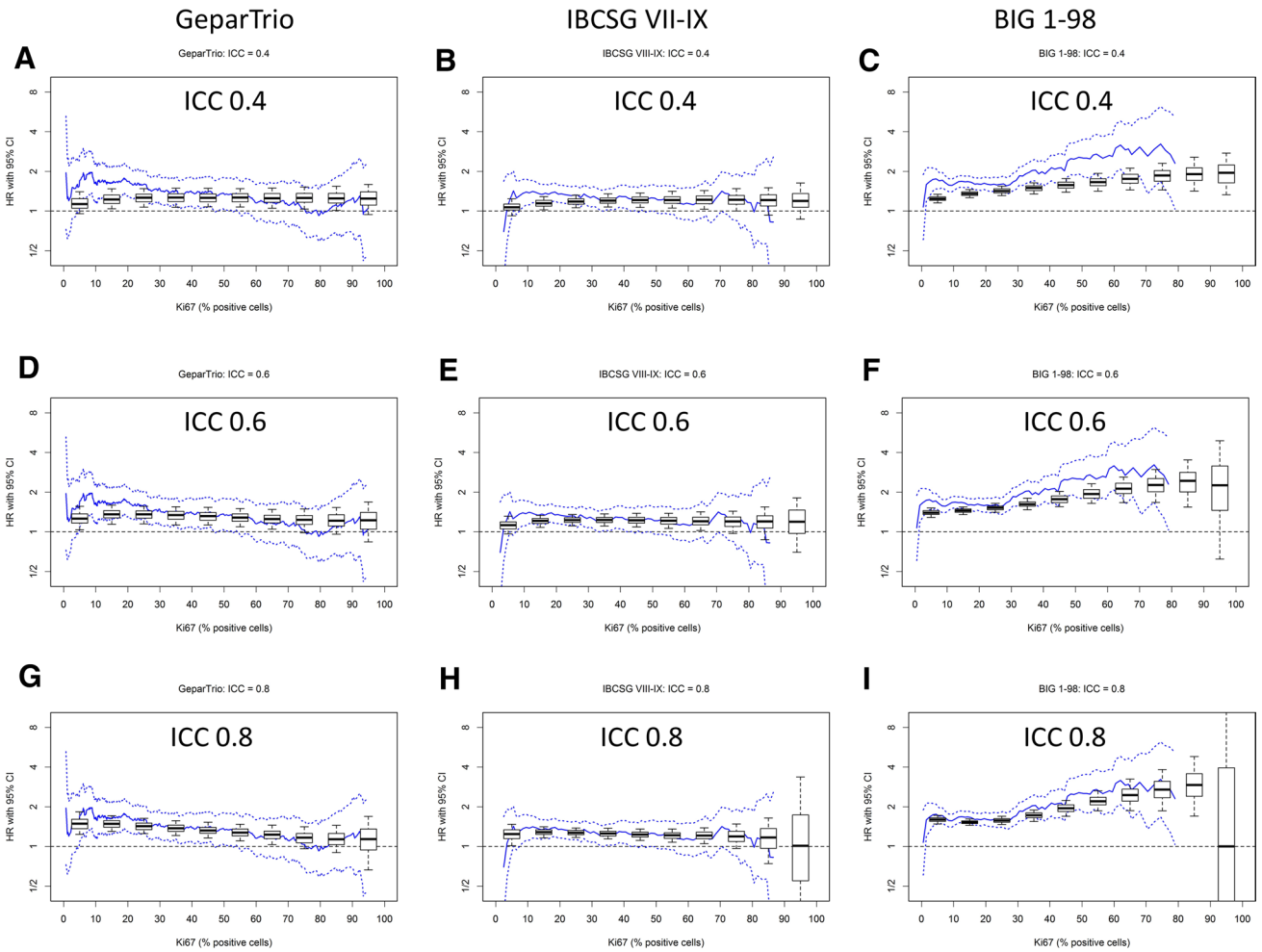

Fig. 5 Mathematical modeling of differences in pathologist variation–systematical evaluation of different Ki-67 cutpoints for DFS in three clinical studies. Three different virtual pathologist groups of each 500 pathologists with either poor performance (ICC = 0.4; **a**–**c**), or moderate performance (ICC = 0.6, **d**–**f**) or good performance (ICC = 0.8, **g**–**i**) are shown for GeparTrio (**a, d, g**), IVCSG VIII-IX (**b, e, h**) and BIG 1–98 (**c, f, i**). Each box visualizes the OR for prediction of pCR for the results of a group of 500 virtual pathologists (5%, 25%, 50%, 75% and 95% quantiles) and a given Ki-67 cutpoint. Different cutpoints between Ki-67 of 5% and 95% are systematically evaluated. Blue lines indicate the actual HR for DFS (with 95% CI) based on the central Ki-67 evaluation in the original clinical trial dataset. The HRs are remarkably stable over a wide range of different ICCs. *ICC* intraclass correlation coefficient, *HR* hazard ratio, *CI* confidence interval

only be based on the results of clinical trials, which are—as shown in out study—highly robust to pathologist variation.

It must be emphasized that the integration of Ki-67 into clinical decisions requires an extremely careful assessment for each individual patient—similarly to the administration of a new therapeutic strategy. To achieve this high standard of assessment several precautions have to be taken: The evaluation must be based on the recommendations of the International Ki-67 Working Group, the staining conditions and the preanalytical variables must be carefully controlled, and the quantification method must allow an exact determination of the number of positive cells.

Pathologists and clinicians should be aware of the limitations of the method. In particular, for Ki-67 values in the intermediate range between 10% and 25%, where the interobserver heterogeneity is most relevant according to the Ki-67 ring trials [17, 29]; it might be necessary to use additional, more advanced prognostic assays, including gene expression assays.

As established by the International Ki-67 Working group, the best indicator of interobserver performance for a continuous biomarker is the intraclass correlation coefficient (ICC). The ICC refers to the combined overall performance across the dataset and across the complete range of Ki-67 values. Therefore, even with a high ICC of 0.8, individual tumors may exist for which the Ki-67 values differ substantially among individual observers. The ICC as a statistical performance parameter does not require the minimization of

all individual outlier values. In clinical practice, the problem of individual outliers remains and must be considered for individual clinical decisions. Thus, clinicians should be aware of the fact that there may be a small number of outliers in any Ki-67 cohort, which have larger absolute differences between pathologists.

In a previous study on the GeparTrio cohort, we had already pointed out that Ki-67 is a continuous parameter and that is it not possible to define an optimal cutpoint to define positive vs. negative Ki-67. We have now validated this result in the additional datasets from BIG 1-98 and IBCSG VIII + IX, that also showed very similar hazard ratios for the association of Ki-67 positivity with DFS across a wide range of Ki-67 cutpoints. This is in line with the statement of the International Ki-67 guideline published in 2011, which stated that the experts were not able to come to a conclusion regarding valid Ki-67 cutpoints [15]. This can be easily explained by the biology of tumor cell proliferation, which is a continuous parameter when observed on the level of a tumor cell population. In addition, the proliferation rate of a tumor is not fixed at a defined value, but is dependent on the proliferative capacity of the tumor cells, but also in the local microenvironment including oxygen and energy supply, constitution of the tumor stroma and vascular density. This variability results in temporal and spatial heterogeneity of proliferation and thus Ki-67 expression. As a conclusion, international efforts to define cutpoints for Ki-67 should be stopped and clinicians should get used to the interpretation of this marker as a continuous rather than a discrete parameter.

In conclusion, it is important to emphasize that our study is evaluating the performance of Ki-67 as a prognostic marker in large clinical trial cohorts. In contrast, in the daily clinical diagnostic setting, the aim is to provide a reliable Ki-67 value for a single patient. To achieve the best result for each patient, it is necessary that the local Ki-67 staining is standardized to achieve optimal performance. Therefore, ongoing efforts to further reduce interobserver variability, including ring trials and standardized guidelines as well as image analysis approaches, should be continued [30, 31].

Acknowledgements We thank the patients, pathologists and investigators who contributed to these trials. We thank the International Breast Cancer Study Group (IBCSG) for contributing data for IBCSG Trials VIII and IX and BIG 1-98.

Funding This study was funded by a Grant from the German Cancer Aid Translational Oncology Program in the TransLUMINAL-B project. Central pathology assessment of Trials VIII and IX was funded by the IBCSG. Translational research in BIG 1-98 has been funded in part by Novartis and Susan G. Komen for the Cure Promise Grant (KG080081 to GV and MMR). The IBCSG Statistical Center received funding from US NIH Grant CA075362 (MMR).

Compliance with ethical standards

Conflict of interest Author C. Denkert has been cofounder and shareholder of Sividon Diagnostics and co-inventor of the vmScope digital image analysis system; he was an advisor/consultant for Teva, Roche, AstraZeneca, Celgene, Pfizer, Novartis, Daiichi and MSD. Author B. Thürlimann is an advisor/consultant of Pfizer, Amgen, Genomic Health Inc, Eli Lilly and reports stock ownership of Novartis and Roche. Author M. Dietel is an advisor/consultant of provitro AG. Author B. Ataseven is an advisor/consultant of Roche and Amgen and reports funding form Tesaro and Astra Zeneca. All other authors report no conflicts of interest.

Research involving human participants and/or animals All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the clinical trials.

References

1. Gerdes J, Schwab U, Lemke H et al (1983) Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. Int J Cancer 31(1):13–20
2. Andre F, Arnedos M, Goubar A et al (2015) Ki-67–no evidence for its use in node-positive breast cancer. Nat Rev Clin Oncol 12(5):296–301
3. Harris LN, Ismaila N, McShane LM et al (2016) Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline. J Clin Oncol 34(10):1134–1150
4. Goldhirsch A, Wood WC, Coates AS et al (2011) Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 22(8):1736–1747
5. Goldhirsch A, Winer EP, Coates AS et al (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
6. Stuart-Harris R, Caldas C, Pinder SE et al (2008) Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. Breast 17(4):323–334
7. de Azambuja E, Cardoso F, de Castro G Jr et al (2007) Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. Br J Cancer 96(10):1504–1513
8. Yerushalmi R, Woods R, Ravdin PM et al (2010) Ki-67 in breast cancer: prognostic and predictive potential. Lancet Oncol 11(2):174–183
9. Luporsi E, André F, Spyratos F et al (2012) Ki-67: level of evidence and methodological considerations for its role in the clinicalmanagement of breast cancer: analytical and critical review. Breast Cancer ResTreat 132(3):895–915
10. Petrelli F, Viale G, Cabiddu M et al (2015) Prognostic value of different cut-off levels of Ki-67 in breast cancer: a systematic review

and meta-analysis of 64,196 patients. Breast Cancer Res Treat 153(3):477–491
11. Viale G, Regan MM, Mastropasqua MG et al (2008) Predictive value of tumor Ki-67 expression in two randomized trials of adjuvant chemoendocrine therapy for node-negative breast cancer. J Natl Cancer Inst 100(3):207–212
12. Jones RL, Salter J, A'Hern R et al (2009) The prognostic significance of Ki-67 before and after neoadjuvant chemotherapy in breast cancer. Breast Cancer Res Treat 116(1):53–68
13. Urruticoechea A, Smith IE, Dowsett M (2005) Proliferation marker Ki-67 in early breast cancer. J Clin Oncol 23(28):7212–7220
14. Denkert C, Loibl S, Müller BM et al (2013) Ki-67 levels as predictive and prognostic parameters in pretherapeutic breast cancer core biopsies: a translational investigation in the neoadjuvant GeparTrio trial. Ann Oncol 24(11):2786–2793
15. Dowsett M, Nielsen TO, A'Hern R et al (2011) Assessment of Ki-67 in breast cancer: recommendations from the International Ki-67 in Breast Cancer working group. J Natl Cancer Inst 103(22):1656–1664
16. Polley MY, Leung SC, McShane LM et al (2013) An international Ki-67 reproducibility study. J Natl Cancer Inst 105(24):1897–1906
17. Polley MY, Leung SC, Gao D et al (2015) An international study to increase concordance in Ki-67 scoring. Mod Pathol 28(6):778–786
18. Leung SYJ, Nielsen TO, Zabaglo L et al (2016) Analytical validation of a standardized scoring protocol for Ki-67: phase 3 of an international multicenter collaboration. NPJ Breast Cancer 2:16014
19. Budczies J, Klauschen F, Sinn BV et al (2012) Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS ONE 7(12):e51862
20. Karlsson P, Sun Z, Braun D et al (2011) Long-term results of International Breast Cancer Study Group Trial VIII: adjuvant chemotherapy plus goserelin compared with either therapy alone for premenopausal patients with node-negative breast cancer. Ann Oncol 22(10):2216–2226
21. Aebi S, Sun Z, Braun D, Price KN et al (2011) Differential efficacy of three cycles of CMF followed by tamoxifen in patients with ER-positive and ER-negative tumors: long-term follow up on IBCSG Trial IX. Ann Oncol 22:1981–1987
22. Regan MM, Neven P, Giobbie-Hurder A et al (2011) Assessment of letrozole and tamoxifen alone and in sequence for postmenopausal women with steroid hormone receptor-positive breast cancer: the BIG 1-98 randomised clinical trial at 8.1 years median follow-up. Lancet Oncol 12(12):1101–1108
23. Viale G, Giobbie-Hurder A, Regan MM et al (2008) Prognostic and predictive value of centrally reviewed Ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: results from trial BIG 1-98 comparing adjuvant endocrine therapy with tamoxifen versus letrozole. J Clin Oncol 26:5569–5575
24. Viale G, Regan MM, Dell'Orto P et al (2011) Which patients benefit most from adjuvant aromatase inhibitors? Results using a composite measure of prognostic risk in the BIG 1-98 randomized trial. Ann Oncol 22(10):2201–2207
25. von Minckwitz G, Blohmer JU, Costa SD et al (2013 Oct) Response-guided neoadjuvant chemotherapy for breast cancer. J Clin Oncol 10(29):3623–3630 31(
26. Budczies J, Klauschen F, Sinn BV et al (2012) Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One 7(12):e51862. https://doi.org/10.1371/journal.pone.0051862
27. Gamer M, Lemon J, Fellows I, Singh P (2012). irr: various coefficients of interrater reliability and agreement. R package version 0.84. http://CRAN.R-project.org/package=irr
28. Patrick E, Shrout, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420–428
29. Varga Z, Diebold J, Dommann-Scherrer C et al (2012) How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists. PLoS ONE 7(5):e37379
30. Klauschen F, Wienert S, Schmitt WD et al (2015) Standardized Ki-67 Diagnostics Using Automated Scoring–Clinical Validation in the GeparTrio Breast Cancer Study. Clin Cancer Res 21(16):3651–3657
31. Rimm DL, Leung SCY, McShane LM et al (2018) An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki-67 in breast cancer. Mod Pathol 32(1):59

Affiliations

Carsten Denkert[1,2] · Jan Budczies[1,3] · Meredith M. Regan[4] · Sibylle Loibl[5] · Patrizia Dell'Orto[6] · Gunter von Minckwitz[5] · Mauro G. Mastropasqua[6] · Christine Solbach[7] · Beat Thürlimann[8,9] · Keyur Mehta[5] · Jens-Uwe Blohmer[10] · Marco Colleoni[11,12] · Volkmar Müller[13] · Frederick Klauschen[1] · Beyhan Ataseven[14,15] · Knut Engels[16] · Roswitha Kammler[17] · Berit M. Pfitzner[1] · Manfred Dietel[1] · Peter A. Fasching[18] · Giuseppe Viale[19,20]

1 Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany

2 Institute of Pathology, Philipps-University Marburg, Marburg, Germany

3 Institute of Pathology, University of Heidelberg, Heidelberg, Germany

4 International Breast Cancer Study Group Statistical Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

5 German Breast Group, Neu-Isenburg, Germany

6 International Breast Cancer Study Group Central Pathology Office, Division of Pathology and Laboratory Medicine, European Institute of Oncology, IRCCS, Milan, Italy

7 Breast Center, University of Frankfurt, Frankfurt, Germany

8 Breast Center, Kantonsspital, St. Gallen, St. Gallen, Switzerland

9 Swiss Group for Clinical Cancer Research (SAKK), St. Gallen, Switzerland

10 Breast Center, Charité University Hospital, Berlin, Germany

[11] Division of Medical Senology, IEO, European Institute of Oncology IRCCS, Milan, Italy

[12] International Breast Cancer Study Group, Milan, Italy

[13] Department of Gynecology, Universitätsklinikum Hamburg- Eppendorf, Hamburg, Germany

[14] Department of Gynecology and Gynecologic Oncology, Kliniken Essen-Mitte, Essen, Germany

[15] Department of Obstetrics and Gynecology, University Hospital, LMU, Munich, Germany

[16] Zentrum für Pathologie, Zytologie und Molekularpathologie Neuss, Neuss, Germany

[17] International Breast Cancer Study Group Central Pathology Office, IBCSG Coordinating Center, Bern, Switzerland

[18] Women's Health Clinic, University Hospital, University of Erlangen, Erlangen, Germany

[19] International Breast Cancer Study Group Central Pathology Office, Department of Pathology and Laboratory Medicine, IEO European Institute of Oncology IRCCS, Milan, Italy

[20] University of Milan, Milan, Italy