Modern Pathology (2016) 29, 1155–1164

Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group

Carsten Denkert[1,2], Stephan Wienert[1,3], Audrey Poterie[4], Sibylle Loibl[5,6], Jan Budczies[1,2], Sunil Badve[7], Zsuzsanna Bago-Horvath[8], Anita Bane[9], Shahinaz Bedri[10], Jane Brock[11], Ewa Chmielik[12], Matthias Christgen[13], Cecile Colpaert[14], Sandra Demaria[15], Gert Van den Eynden[16], Giuseppe Floris[17], Stephen B Fox[18], Dongxia Gao[19], Barbara Ingold Heppner[1], S Rim Kim[20], Zuzana Kos[21], Hans H Kreipe[13], Sunil R Lakhani[22], Frederique Penault-Llorca[23], Giancarlo Pruneri[24], Nina Radosevic-Robin[23], David L Rimm[25], Stuart J Schnitt[26], Bruno V Sinn[1,27], Peter Sinn[28], Nicolas Sirtaine[29], Sandra A O'Toole[30], Giuseppe Viale[31], Koen Van de Vijver[32], Roland de Wind[33], Gunter von Minckwitz[5], Frederick Klauschen[1], Michael Untch[34], Peter A Fasching[35], Toralf Reimer[36], Karen Willard-Gallo[37], Stefan Michiels[38], Sherene Loi[39] and Roberto Salgado[16,40]

[1]Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany; [2]German Cancer Consortium (DKTK), Partner site Berlin, Berlin, Germany; [3]VMscope GmbH, Berlin, Germany; [4]Service de Biostatistique et d'Epidémiologie; Institute Gustave Roussy, Villejuif, France; [5]German Breast Group (GBG), Neu-Isenburg, Germany; [6]Clinic of Gynecology and Obstetrics, Sana Klinikum Offenbach, Offenbach, Germany; [7]Indiana University School of Medicine, Indianapolis, IN, USA; [8]Clinical Institute of Pathology, Medizinische Universität Wien, Wien, Austria; [9]Department of Pathology & Molecular Medicine, McMaster University, Hamilton, ON, Canada; [10]Weill Cornell Medical College, Doha, Qatar; [11]Harvard Medical School, Boston, MA, USA; [12]Tumor Pathology Department, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Gliwice, Poland; [13]Institut für Pathologie, Medizinische Hochschule Hannover, Hannover, Germany; [14]Pathology, GZA Ziekenhuizen, Sint-Augustinus, Wilrijk, Belgium; [15]Radiation Oncology and Pathology, Weill Cornell Medical College, New York, NY, USA; [16]Department of Pathology and Cytology GZA Hospitals, Wilrijk, Belgium; [17]Department of Pathology University Hospitals Leuven, Leuven, Belgium; [18]Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia; [19]Anatomical Pathology, Vancouver Hospital, Vancouver, BC, Canada; [20]Division of Pathology, NSABP, Pittsburgh, PA, USA; [21]Department of Pathology and Laboratory Medicine, University of Ottawa, Ottawa, ON, Canada; [22]UQ School of Medicine and Pathology Queensland, The University of Queensland Centre for Clinical Research, Brisbane, QLD, Australia; [23]ERTICa Research Team, Department of Pathology, Jean Perrin Comprehensive Cancer Center, University of Auvergne EA4677, Clermont-Ferrand, France; [24]Division of Pathology and Laboratory Medicine, University of Milan, Milan, Italy; [25]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA; [26]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA; [27]Department for Translational Molecular Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA; [28]Sektion Gynäkopathologie, Pathologisches Institut, University of Heidelberg, Heidelberg, Germany; [29]Anatomie Pathologique, Institut Jules Bordet, Brussels, Belgium; [30]Molecular Diagnostic Oncology, Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Camperdown, NSW, Australia; [31]Department of Pathology, European Institute of Oncology, University of Milan, Milan, Italy; [32]Department of Pathology, Netherlands Cancer Institute, Amsterdam, Netherlands; [33]Jules Bordet Institute, Brussels, Belgium; [34]Department of Gynecology, Helios-Klinikum Berlin-Buch, Berlin, Germany; [35]Department of Gynecology and Obstetrics, University Hospital Erlangen, Comprehensive Cancer Center Erlangen EMN, Friedrich-Alexander University Erlangen,

Correspondence: Professor C Denkert, MD, Institute of Pathology, Charité Universitätsmedizin Berlin, Charitéplatz 1, Berlin 10117, Germany.
E-mail: carsten.denkert@charite.de
Received 24 March 2016; revised 26 April 2016; accepted 1 May 2016; published online 1 July 2016

Erlangen, Germany; [36]Department of Gynecology, Klinikum Südstadt Rostock, Rostock, Germany; [37]Molecular Immunology Unit, Institut Jules Bordet, Brussels, Belgium; [38]Service de Biostatistique et d'Epidémiologie, Gustave Roussy, CESP, Inserm U1018, Univ. Paris Sud, Univ. Paris-Saclay, Villejuif, France; [39]Division of Research and Clinical Medicine, Peter MacCallum Cancer Center, Melbourne, VIC, Australia and [40]Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Brussels, Belgium

Multiple independent studies have shown that tumor-infiltrating lymphocytes (TIL) are prognostic in breast cancer with potential relevance for response to immune-checkpoint inhibitor therapy. Although many groups are currently evaluating TIL, there is no standardized system for diagnostic applications. This study reports the results of two ring studies investigating TIL conducted by the International Working Group on Immuno-oncology Biomarkers. The study aim was to determine the intraclass correlation coefficient (ICC) for evaluation of TIL by different pathologists. A total of 120 slides were evaluated by a large group of pathologists with a web-based system in ring study 1 and a more advanced software-system in ring study 2 that included an integrated feedback with standardized reference images. The predefined aim for successful ring studies 1 and 2 was an ICC above 0.7 (lower limit of 95% confidence interval (CI)). In ring study 1 the prespecified endpoint was not reached (ICC: 0.70; 95% CI: 0.62–0.78). On the basis of an analysis of sources of variation, we developed a more advanced digital image evaluation system for ring study 2, which improved the ICC to 0.89 (95% CI: 0.85–0.92). The Fleiss' kappa value for < 60 *vs* ≥ 60% TIL improved from 0.45 (ring study 1) to 0.63 in RS2 and the mean concordance improved from 88 to 92%. This large international standardization project shows that reproducible evaluation of TIL is feasible in breast cancer. This opens the way for standardized reporting of tumor immunological parameters in clinical studies and diagnostic practice. The software-guided image evaluation approach used in ring study 2 may be of value as a tool for evaluation of TIL in clinical trials and diagnostic practice. The experience gained from this approach might be applicable to the standardization of other diagnostic parameters in histopathology.
Modern Pathology (2016) **29,** 1155–1164; doi:10.1038/modpathol.2016.109; published online 1 July 2016

The development and progression of malignant tumors is characterized by an interaction with other cells in the tumor microenvironment including infiltrating immune cells, which have been observed in many different tumor types.[1,2] In HER2-positive and triple-negative breast cancer, immune infiltrates are detectable in up to 75% of tumors, with up to 20% of tumors having a particularly dense infiltrate.[3–5] In some studies, these tumors have been designated lymphocyte-predominant breast cancer,[5] indicating that they contain more lymphocytes than tumor cells.

Accumulating evidence from several studies indicates that tumor-infiltrating lymphocytes (TIL) are predictive for response to neoadjuvant therapy,[4,6–9] and prognostic after adjuvant chemotherapy.[10–13] This suggests that the biology and treatment response of breast cancers varies with different lymphocyte levels, and that it may be important to evaluate TIL in clinical trial cohorts as well as in daily histopathological practice. In particular, for the upcoming clinical trials of immune-checkpoint inhibitors,[14] it is critical that we generate reliable data on tumor immune infiltrates. Although quantitative expression analysis of immune genes has been shown to be predictive for response, the correlation of immune gene expression with TIL is typically high,[7] suggesting that evaluation of TIL may be a valid, less expensive, and readily available alternative. At the present time, it is unclear, which method is more suitable for evaluation of immune parameters in breast cancer, and it is very likely that a combined approach will be necessary.

Clinical histopathology has traditionally been focused on qualitative diagnosis of defined tumor entities. In the current age of personalized medicine, the quantitative assessment of morphological biomarkers is becoming increasingly important, and these parameters are typically evaluated on a continuous scale. TIL are a typical model of a quantitative histological biomarker; therefore it is imperative to establish standardized methods for reproducible assessment of this marker.

The international Immuno-oncology Biomarker Working Group was established in 2012 with the goal of creating an internationally standardized approach for tumor-infiltrating lymphocyte evaluation in breast cancer. The members were biostatisticians, gynecologists, immunologists, oncologists, and pathologists with an interest and research experience in the evaluation of immunological infiltrates in malignant tumors. The groups included pathologists from different regions of the world with interest in biomarker research in prospective clinical trials, including the group pathologists of the major clinical trial groups in breast cancer.

The working group has developed and published a guideline document for TIL evaluation on hematoxylin–eosin stained slides as the first step.[15] During two meetings held at the San Antonio Breast Cancer Symposium in 2013 and 2014 the best implementation strategies were discussed. The consensus from the group was that a multicenter international ring study was necessary to generate data on interobserver variability of tumor-infiltrating lymphocyte

assessment. Therefore, two ring studies were conducted.

The primary aim of ring studies 1 and 2 was the evaluation of the intraclass correlation coefficient (ICC) for decentralized assessment of TIL by a large group of international pathologists. As secondary aims, the two ring study methods were compared by the evaluation of kappa values and the concordance analysis.

Materials and Methods

Tumor Samples

A total of 120 pretherapeutic core biopsies from the neoadjuvant GeparSixto trial were evaluated in the two ring studies (60 different samples in each ring study), the clinical details of the GeparSixto study have been published (NCT01426880).[16] It was decided to perform the ring studies with prospectively collected samples from a clinical trial to minimize any errors related to sample selection. On the basis of the current knowledge, the most interesting target populations for future immunotherapeutic approaches in breast cancer are the HER2-positive and the triple-negative subtypes. Therefore, we decided to focus on a clinical trial including these subtypes. GeparSixto was chosen because we already had existing and published central pathology data on TIL in this trial for both subtypes.[7] This existing data was not regarded as the gold standard, but was used as background information to ensure that the tumors selected for the ring study were adequately representative for different levels of TIL. All participating patients had agreed to sample collections and the use of these samples in translational biomarker research, which has also been approved by the Ethics Commission of the Charité Universitätsmedizin Berlin.

Two separate cohorts of 60 patients were selected for the two sequential ring studies. Inclusion criteria were: (1) existing hematoxylin–eosin slides from GeparSixto; (2) an adequate tissue quality for pathological assessment with at least 20% of tumor area; (3) no predominant ductal carcinoma *in situ*, and (4) existing centrally determined tumor-infiltrating lymphocyte data, so that one-third of each cohort would have lymphocyte-predominant breast cancer tumors (≥60% TIL), one third a minimal lymphocytic infiltrate (<15%), and one-third an intermediate infiltrate. For ring study 1, we started with patient 1 of the GeparSixto study and selected 60 consecutive tumor slides that matched the inclusion criteria above. For ring study 2, the selection of consecutive slides was extended and a completely independent cohort was selected with the identical inclusion criteria. From all sections slide scanning was performed and digital slides were generated for remote evaluation (VMscope, Berlin, Germany).

Participating Pathologists

In the power calculation it was calculated that for a two-sided test at alpha = 0.05 with 20 readers and a power of 80%, 60 samples are needed with an ICC = 0.8 under the alternative hypothesis. On the basis of this power calculation, it was decided to evaluate 60 slides and to invite 36 pathologists for ring study 1. The acceptance rate of the invitation was higher than expected, and 32 pathologists (89%) from 27 institutions in 9 countries, agreed to participate in ring study 1. The invitations to join ring study 2 were restricted to the 32 participants of the first ring study. Overall, 28 of them (88%) agreed to participate in the second ring study, as well. Participation was completely voluntary and the remaining four pathologists were not required to give any reason for their decision not to participate again. The results of each pathologist in either ring trial were kept strictly confidential with no individual feedback provided to participants and only anonymous data being published. In particular, pathologists were not aware of their individual performance in ring study 1 when participating in ring study 2. To exclude any systematical error due to the slightly different pathologist groups in both ring studies, an additional exploratory analysis was performed for ring study 1 that included only those 28 pathologists that had participated in ring study 2, as well.

Web-Based Evaluation of Digital Slides in First Ring Study

In the first ring study, digital slides were uploaded to a central website where the invited participants could log in and enter their assessment of each slide as a semiquantitative percentage of stromal TIL. This evaluation was based on the previously published guideline paper and the tutorial that was included as Supplemental Material. The key steps of this completely predefined and standardized assessment were: (1) definition of the tumor area and exclusion of tissue outside of the tumor borders; (2) exclusion of large areas of necrosis and fibrosis in the center of the tumor; (3) focus on stromal TIL that are located in the connective tissue next to the tumor cell nests; and (4) exclusion of granulocytes in necrotic areas; (5) reporting of the percentage of stromal TIL as a single average value based on the standardized images provided in the guideline tutorial. For the ring study, the pathologists were simply instructed that their assessment should follow these published guidelines[14] and that they must read the tutorial before beginning their evaluation. The stromal lymphocyte values were directly entered to the website by the pathologists and used for the statistical evaluation. The results of the individual performance in the ring trials were not reported back to the pathologists, to avoid any bias for the subsequent ring study 2.

Standardized Software-Based Evaluation with Integrated Feedback in Second Ring Study. In the first ring study a considerable agreement was reached, however the prespecified endpoint was not met. With the aim to better understand the reasons for disagreement, an error analysis was performed (see Results section). The results of this error analysis were used to prepare the second ring study with a more standardized approach.

For the second ring study, completely new slides were selected. The selection criteria were identical to ring study 1. The digital slides were evaluated using a new designed image presentation software program that guided the pathologists through the different evaluation steps. In general, the evaluation was identical to the predefined evaluation for ring study 1 (see above). There were only two major differences: (1) To address potential random variation due to intratumoral heterogeneity and random errors, three areas of predefined size were evaluated for each tumor in the second ring study. When data entry is initiated, the magnification is adjusted and the screen automatically displays an area of 1 mm^2, to insure that the area evaluated is standardized. A minimum of three different regions had to be evaluated; the pathologists had the option to add more regions. All elevated regions were marked on the slide, to avoid evaluation of the same region twice. It was not possible to evaluate less than three regions or to select a large or smaller region size, but it was possible to evaluate additional regions. (2) For each lymphocyte value that was entered, a direct continuous visual reference feedback was presented in the form of reference images displaying an example of the selected tumor-infiltrating lymphocyte density. The reference images were based on those tumors from ring study 1 that had the highest agreement and therefore represent the consensus of a large international group of pathologists. Lymphocyte values were entered using a rotary control directly connected to this set of reference images, allowing a direct visual comparison of the individual case with the reference image and adjustment of the rotary control until the best match was achieved. The pathologists had the option to modify their assessment based on comparison of the actual slide and the reference image until they felt that their reading best matched with the reference image.

The stromal lymphocyte values for each region were recorded and the mean was used for the statistical evaluation. The software used in ring study 2 has been integrated in the VMscope slide explorer (VMscope) for use in future studies and routine clinical practice.

Statistical Analysis

The predefined primary endpoint for interobserver variance in the two ring studies was the ICC,[17] which is the proportion of total variance (in measurements across patients and laboratories) that is attributable to the biological variability among patients' tumors. The ICC has a range from 0 to 1 with a score of 1 having the maximum agreement. The interpretation of an ICC of, for example, 0.7 would be that 70% of the variance in the dataset results from the actual biology (eg, the differences in lymphocyte levels across different tumors), whereas 30% of the variance would be introduced as an artifact due to the differences between pathologists. In our predefined ring study protocol, a successful analysis was defined as an observed ICC that is statistically significantly >0.7. Therefore, the 95% confidence interval (CI) must exclude an ICC of 0.70. We have decided to use 0.7 in our study protocol because the Ki67 ring studies are using this cutpoint, as well, and this would allow a comparison of the results.[18]

The ICC for single measures was calculated using the mixed model and absolute agreement. There were five missing values (0.3% of the 1920 evaluations) in ring study 1, which were replaced by the mean for the ICC and kappa analysis. An exploratory analysis without this replacement did not alter the results of ring study 1 (not shown). In ring study 2 there were no missing values.

Because the pathologist groups in ring study 1 and 2 were slightly different, it was decided to perform an additional exploratory evaluation for ring study 1, to exclude that the different results were simply due to the exclusion of a few pathologists. This additional evaluation was therefore restricted to those 28 pathologists that had participated in both ring studies, to use a common data basis.

As a secondary endpoint, Fleiss' kappa value was evaluated comparing groups of tumors with different lymphocyte levels: <60 *vs* $\geq 60\%$; <50 *vs* $\geq 50\%$ as well as 0–20% *vs* 21–49% *vs* $\geq 50\%$. The concordance analysis per pathologist was evaluated by comparing each individual reader with the gold standard. In this setting, the gold standard was defined as the median value of the rating for each tumor slide across all pathologists.

Results

Evaluation of the First Ring Study

In ring study 1 the 34 pathologists evaluated 60 slides for stromal TIL using a web-based virtual slide presentation system (Figure 1a).[12] The ICC was 0.70 (95% CI: 0.62–0.78, Figure 2). In an exploratory analysis restricted to the 28 pathologists who also participated in RS2 the result was similar (ICC: 0.71; 95% CI: 0.63–0.79, Figure 2).

Although very close, the lower limit of the 95% CI of the ICC was below the value of 0.7. Therefore, the prespecified endpoint was not met. We consequently analyzed the main sources of discordance. Figure 3a shows the mean and s.d. across all tumors in ring study 1, which ranged from 12 to 44% for the different pathologists, suggesting a systematic scaling variation between pathologists. This scaling error

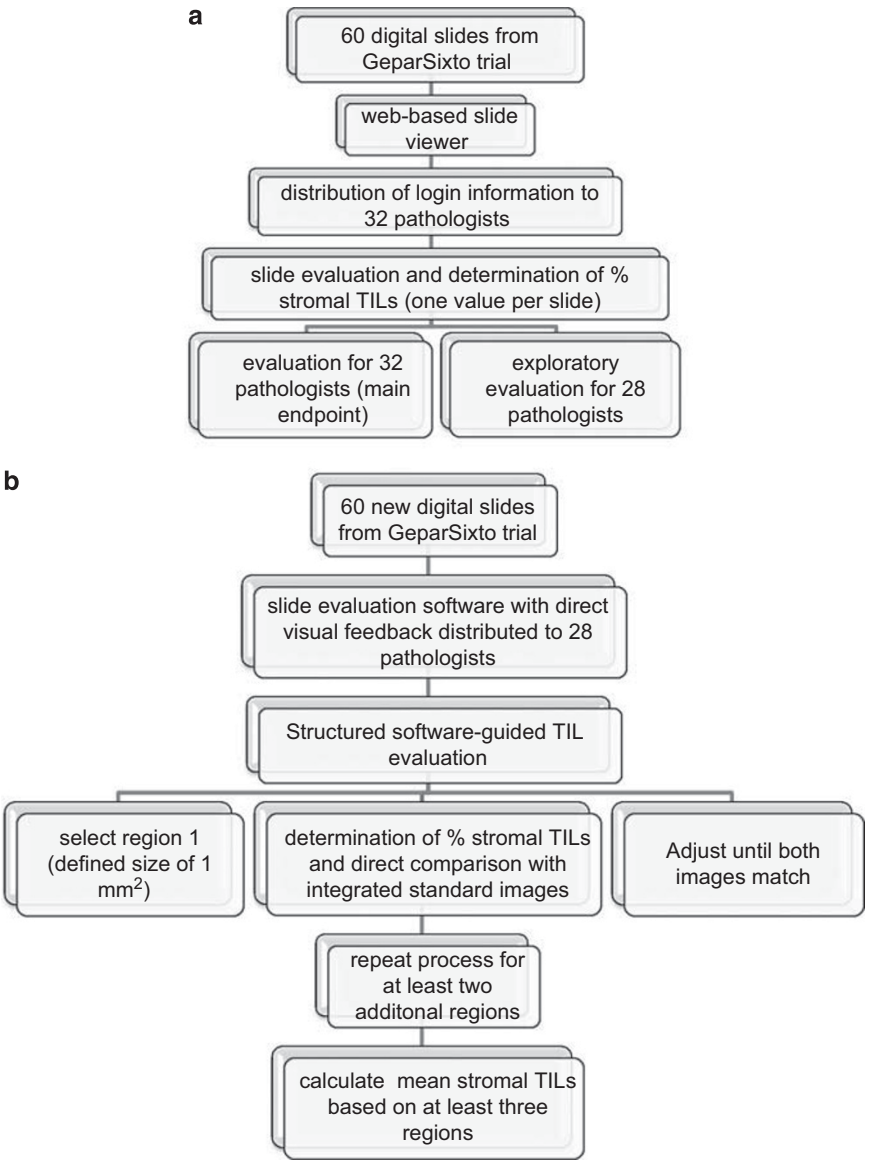

Figure 1 Structure of the two separate ring studies. Ring study 1 is a simple evaluation of digital slides using a web-based slide-viewer. For each slide, a single value for % stromal TIL is entered by the pathologist (**a**). Ring study 2 uses a software-system evaluation that guides pathologists though the evaluation. Three areas for defined size have to be evaluated and TIL levels are directly compared with integrated reference slides, allowing adjustment of the values until the best match with the control is achieved (**b**). TIL, tumor-infiltrating lymphocytes.

also becomes evident in Figure 3c, which shows all values that have been generated during ring study 1. There was a systematical shift of mean values between the pathologists indicating that the individual cutpoints for a given percentage of TIL are different among pathologists. For example, the pathologist in the top row has very high individual cutpoints, so that most of the slides were assessed as having low levels of TIL. In contrast, the pathologist in the bottom row has a very low cutpoint, so that most of the tumors have high lymphocyte values.

In addition, individual outliers were observed as a second source of variation. For example, in the lower left section of Figure 3c there is a single pathologist who has entered a value of 70% (blue), whereas all the other pathologists had values in the lower range (orange/yellow) for the same tumor. Similarly, for several tumors with high TIL (blue) in the right section of Figure 3c, some pathologists have entered very low percentages (single yellow/orange dots). These random variations could be due to intratumoral heterogeneity or simply represent individual random errors in assessment or data entry.

Adaptation of the System for Increased Standardization

The results from the first ring study were discussed at the working group meeting at the San Antonio

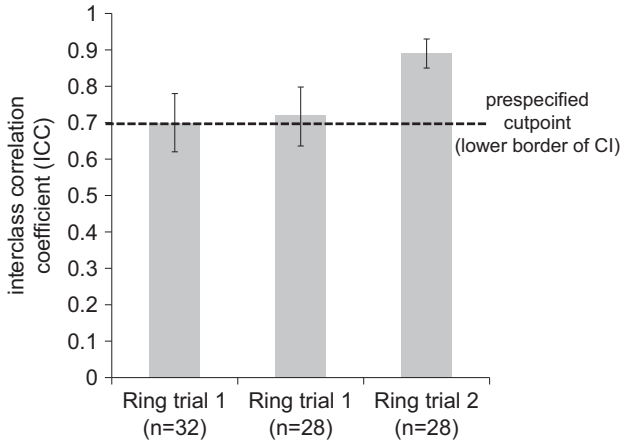

Figure 2 Results of ring studies 1 and 2 (ICC). The primary analysis for ring study 1 ($n=32$) and 2 ($n=28$) is shown. In addition, we have performed and exploratory analysis for ring study 1 including only those 28 pathologists that had also participated in ring study 2 (middle column). ICC, intraclass correlation coefficient.

Breast Cancer Symposium in December 2014. The consensus was that, although the results were promising, the approach used was not yet ready for clinical practice, because the prespecified endpoint was not met.

In an effort to reduce both systematic and random variation, a more user-friendly system with guided slide evaluation was developed (Figure 1b). Systematic scaling variation between pathologists was addressed by providing continuous feedback by integrated reference images. Tumor-infiltrating lymphocyte values were entered using a rotary control directly connected to this set of reference images, allowing a direct visual comparison of the individual case with the reference image, and adjustment until the best match was achieved (Figure 4). To address variation due to intratumoral heterogeneity and random errors, at least three areas of predefined size were evaluated for each tumor.

Evaluation of the Second Ring Study

An independent set of 60 additional slides was evaluated in ring study 2 resulting in an ICC of 0.89 (95% CI: 0.85–0.92, Figure 2). The prespecified endpoint was met in ring study 2, with an improved range of mean values (between 24 and 35%, Figure 3b). Both types of variation, ie, the systematic as well as the random variation, were reduced in ring study 2, as shown in Figure 3b and d by the more homogenous distribution.

Comparison of Kappa Values and Interobserver Concordance

Fleiss' kappa values in ring study 1 for the three cutpoints (<60 *vs* $\geq 60\%$; <50 *vs* $\geq 50\%$; 0–20% *vs* 21–49% *vs* $\geq 50\%$) were 0.45, 0.51, and 0.46, respectively, corresponding to a moderate agreement (Table 1). In ring study 2, the Fleiss' kappa values were 0.63, 0.72, and 0.65, respectively, which correspond to a substantial agreement. Concordance for each pathologist was evaluated by comparing each reader with the median. The mean and s.d. for these individual values is shown in Table 1. The mean concordance rates for the three lymphocyte cutpoints were 0.88, 0.89, and 0.78, respectively, for ring study 1, and they improved to 0.92, 0.93, and 0.85, respectively, in ring study 2.

Discussion

There is a growing pressure on pathologists to provide reliable data on quantitative tissue parameters. This represents a challenge, because the human eye, although excellent in pattern recognition, has limitations regarding quantitative assessments. The traditional approach to standardization has been to create written evaluation guidelines for pathologists. The exact process of the evaluation is, however, typically not standardized. In most situations, the pathologist simply looks through the microscope and writes down his or her assessment in a report.

Our experience with ring studies 1 and 2 was interesting for comparison of different workflows in diagnostic pathology. Ring study 1 was performed using the traditional approach, which could be phrased simply as 'read the guideline and evaluate the slide'. The exact way how to apply the guidelines to the individual slides was not defined and was left open to the individual pathologist. Considering these very basic instructions, it is remarkable that agreement within ring study 1 was still relatively good.

The approach in ring study 2 was a considerable improvement and increased the ICC to 0.89. This means that 89% of the lymphocyte variance is based on true biological variance and differences between individual tumors, and only 11% of the variance is generated by diagnostic differences between pathologists. The observed concordance rates of 92% in ring study 2 are very similar to the concordance rates described for HER2 testing at centers of excellence, eg, central pathology labs in clinical studies.[19]

The improved ICC in ring study 2 suggests that in addition to written guidelines, computer-assisted diagnosis and reporting systems with integrated reference feedback might be valuable tools for pathologists to standardize the process of evaluation of TIL. Important elements of these assistance systems are the systematic evaluation of different regions to reduce the influence from tumor heterogeneity, and probably also the random errors caused by incorrect data entry. In addition, an integrated visual feedback from reference values helps to adjust the cutpoints for assessment of quantitative parameters. The standardized control values used in ring study 2 were based on the assessment by an

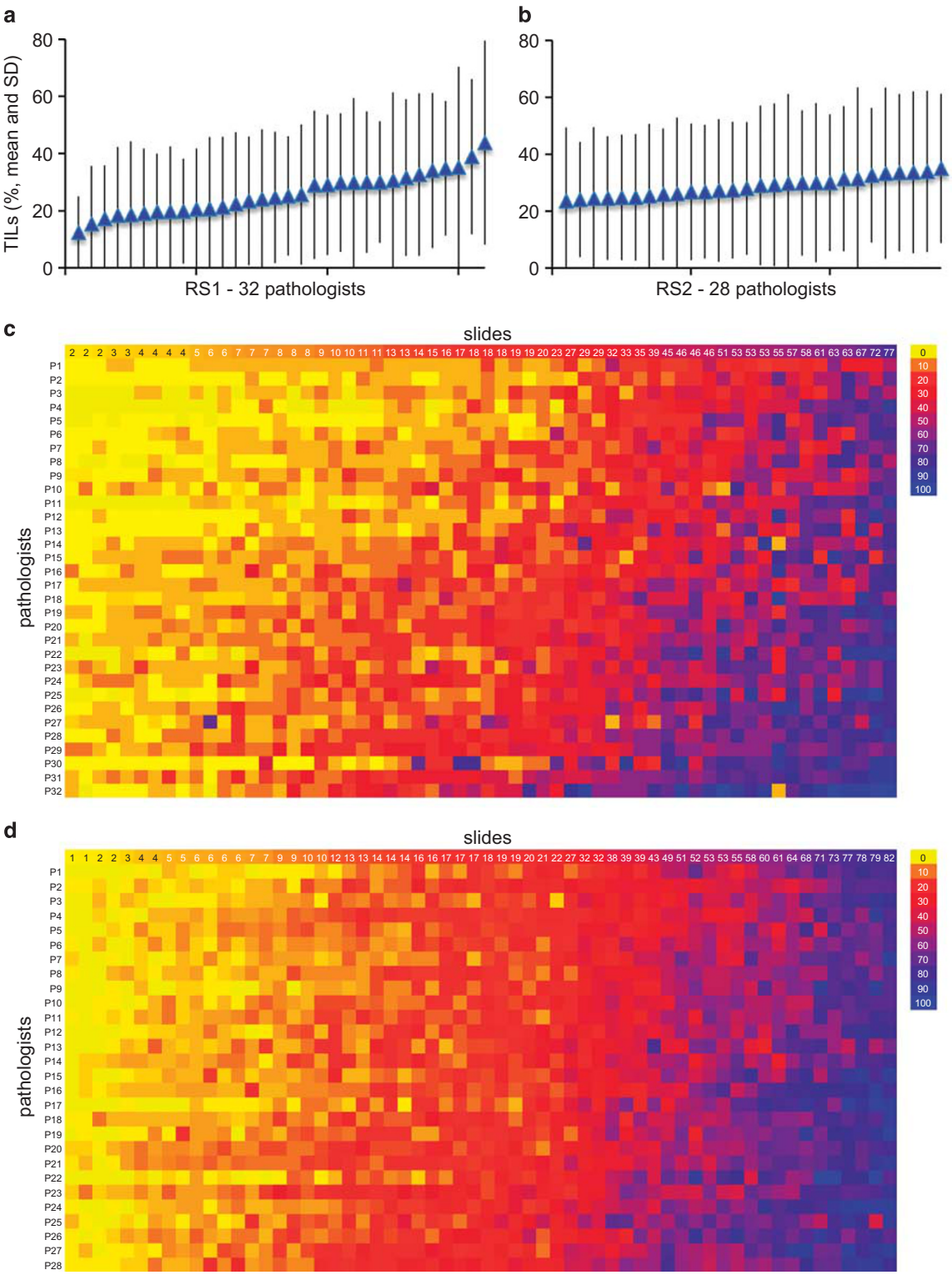

Figure 3 Comparison of results for both ring studies. The mean stromal TIL and s.d. for each pathologist are shown ring study 1 (**a**) and 2 (**b**). In ring study 2, there is considerably less variation in the means between pathologists. In **c** and **d**, all values from ring study 1 (**c**) and 2 (**d**) are shown. The values are sorted from left to right based on ascending mean stromal lymphocyte values for the different slides (top row), and from top to bottom based on ascending mean stromal lymphocyte values for each pathologist. Different color-coding indicates single outlier values. The comparison shows that there are more outlier values in ring study 1 and that the distribution of results is more homogenous in ring study 2. TIL, tumor-infiltrating lymphocytes.

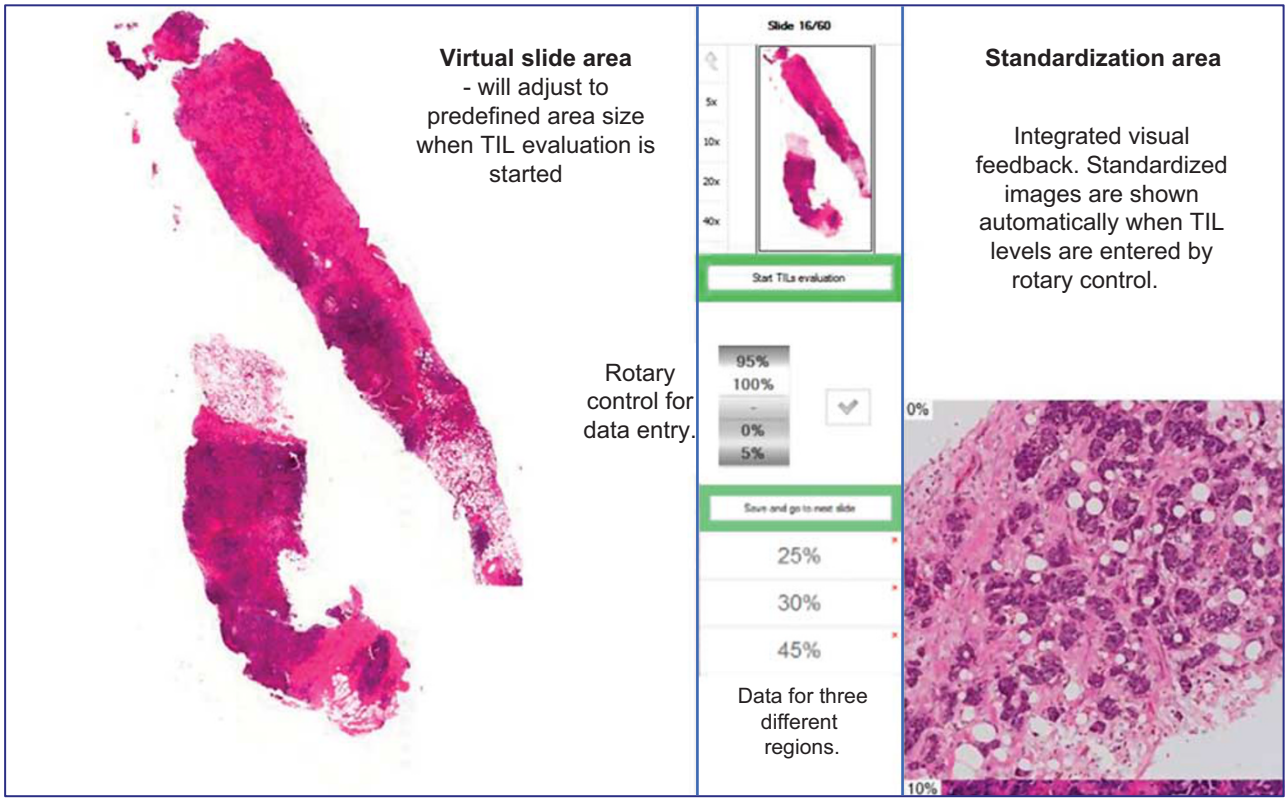

Figure 4 System used for guided TIL evaluation in the second ring trial. The digital slide is displayed on the left. The magnification of this window will adjust at the time of data entry so that only a standardized area of 1 mm^2 is evaluated. The lymphocyte levels are entered using the rotary control, which avoids any writing errors. Whenever this rotary control is moved, the standardization area on the right side will show reference images of different lymphocyte levels for direct comparison with the slide on the left. The rotary control can be adjusted until the best match between both the diagnostic slide and the reference image is achieved. The process is repeated for at least three different regions. TIL, tumor-infiltrating lymphocyte.

Table 1 Comparison of ring study 1 and 2 for primary and secondary endpoints

	Ring study 1	*Ring study 2*
ICC	0.7 (0.62–0.78)	0.89 (0.85–0.92)
Fleiss' kappa		
TILs >60 *vs* ≥60%	0.45	0.63
TILs >50 *vs* ≥50%	0.51	0.72
TILs 0–20% *vs* 21–49% *vs* ≥50%	0.46	0.65
Concordance rates[a]		
TILs >60 *vs* ≥60%	0.88 (±0.05)	0.92 (±0.03)
TILs >50 *vs* ≥50%	0.89 (±0.05)	0.93 (±0.04)
TILs 0–20% *vs* 21–49% *vs* ≥50%	0.78 (±0.07)	0.85 (±0.07)

Abbreviations: ICC, intraclass correlation coefficient; TILs, tumor-infiltrating lymphocytes.
[a]The concordance of each pathologist with the median of all pathologists was calculated for three different TIL-groups.
The values in the table represent the mean and s.d. of these concordance rates for all pathologists.

international group of 32 pathologists in ring study 1, and therefore represent a broad-based international consensus rather than the view of a single institution. To transfer this approach to future evaluations in clinical trial cohorts or in daily diagnostic practice, we have modified the software program used in ring study 2 for the use in future histopathological evaluations with or without digital slides.

The experience from the two ring studies is similar to the different phases of the international Ki67 ring studies. The first Ki67 ring study[17] had an ICC of 0.71 (95% CI: 0.47–0.78). The Ki67 working group then decided to perform a calibration test for all participants before the second ring study, and developed software to assist in data collection, resulting in an improved ICC of 0.94 (95% CI: 0.90–0.97).[18]

There are some differences between our approach for TIL and the Ki67 experience. In ring study 2, we did not include a mandatory pre-test calibration phase. Instead, we provided integrated feedback via the simultaneous presentation of reference images for calibration of each slide that was read. This approach can be translated to the diagnostic setting with the potential advantage that the pathologist is continuously recalibrated by the defined reference values whenever a tumor-infiltrating lymphocyte evaluation is performed. With a separate pre-calibration, the learning effect of the calibration might be lost after some time due to diagnostic drift. It would

be very interesting to investigate whether the integrated feedback calibration could also be useful for the evaluation of Ki67, and perhaps other quantitative markers. There are a few other differences between Ki67 and TIL: the stromal lymphocyte percentage is an estimation, although Ki67 is exactly counted. In terms of staining, our approach to lymphocyte measurement requires routine hematoxylin–eosin stained slides; therefore an adjustment to different immunohistochemistry methods in different laboratories is not necessary.

There are some limitations to our study. Theoretically, the improvement in ring study 2 could result from increased experience of pathologist after ring study 1. However, we believe that this is rather unlikely, because the slides used in study 2 were different, there was a considerable 'wash-out' time period of several months between both ring studies and the pathologists had not received any feedback on their individual performance in ring study 1. Although the two pathologist groups were not completely identical, a restriction of the ring study 1 evaluation to those 28 pathologists that had participated in both ring studies does not change the results. Furthermore, it should be noted that the two ring studies were based on core biopsies, and some adaptation to large sections might be necessary. In particular, the larger area of these sections might require the evaluation of more regions, in particular in those tumors with a heterogeneous distribution of lymphocytes. In addition, in large tumor sections, it is necessary to exclude central necrotic areas, areas with granulocytes as well as peritumoral stroma from the evaluation. This evalution should be based on the published guidelines that have been prepared by the working group.[15]

In our ring studies, we have focused on the evaluation of stromal TIL, which are defined as those lymphocytes within the tumor that are located in the stromal tissue next to the tumor-cell nests. Stromal lymphocytes have been selected for evaluation in this study because they are the predominant lymphocytes in breast cancer, and evaluation of these lymphocytes is recommended by the current international guideline.[15] Intratumoral lymphocytes that are directly infiltrating the epithelial cell nests were not evaluated, because they typically constitute just a minor fraction of the TIL in breast cancer. Further studies are necessary to evaluation the differences between stromal and intratumoral lymphocytes.

The evaluation of TIL in hematoxylin–eosin slides is a simple approach for evaluating the immune status of an individual tumor. For a more comprehensive approach, it might be necessary to evaluate the relative proportion of specific immune subpopulations as well as the spatial organization of the immune infiltrate, including tertiary lymphoid structures. In a previous study, we have shown that TIL and immune mRNAs were highly correlated.[7] The combination of TIL and molecular markers might be particularly interesting for response prediction to promising approaches such as immune-checkpoint blockade in the large group of tumors with intermediate TIL levels.

Multiple scientific publications have reported clinically significant results regarding evaluation of TIL in breast cancer and therapy response, and prognosis.[6,7,9,11,12] Thus, there is a very strong biological basis for the contribution of immunity to therapy response and tumor progression. In a recent commentary,[20] it has been emphasized that despite significant clinical results TIL evaluation may not be completely ready for introduction into routine clinical practice due to interobserver variance and the lack of standardization.

With the combined effort of a large international group of breast pathologists that we present here, we believe that this work presents a major step toward resolving these limitations. This opens the way for standardized reporting of tumor immunological parameters in diagnostic clinical practice. The participating pathologists have enthusiastically worked together for improved standardization in their field. Moreover, the experience gained from this approach might be applicable to the standardization of other diagnostic parameters in histopathology.

Acknowledgments

We would like to thank all patients, clinicians, and pathologists participating in the clinical studies and the biomaterial collection. We are grateful for the excellent technical assistance by Britta Beyer, Sylwia Handzik, Ines Koch, Petra Wachs as well as Peggy Wolkenstein. Funding was given by the European Commission FP7, Grant 278659 (RESPONSIFY) and the TRANSCAN-I, Grant BMBF-01KT1314 (TRANSCAN-I UGI 1).

Disclosure/conflict of interest

The following authors have reported disclosures: Carsten Denkert: stock or other ownership interest: Sividon Diagnostics, Cologne, Germany; consulting: Astra Zeneca; Honoraria: Roche, Teva, Celgene; Cecile Colpaert: travel, accommodation expenses: Roche; Sandra Demaria: consulting or advisory role: Sanofi US Services; Regeneron Pharmaceuticals; Gert Van den Eynden: stock or other ownership interest: Agoko NV, travel/accommodation expenses: Roche; Hans H. Kreipe: honoraria: Roche, Astra Zeneca, Novartis; consulting/advisory role: Roche, Astra Zeneca, Novartis; Giuseppe Viale: honoraria: GlaxoSmithKline; consulting or advisory role: Dako, Roche/Genentech; Speakers' Bureau: Novartis; travel, accommodations, expenses: Roche; Toralf Reimer: Consulting/advisors role for MSD Sharp Dohme, Novartis, Pfizer; travel/accommodation expenses: Novartis, Roche. The remaining authors declare no conflict of interest.

References

1 Fridman WH, Pagès F, Sautès-Fridman C *et al.* The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12:298–306.
2 Disis ML. Immune regulation of cancer. J Clin Oncol 2010;28:4531–4538.
3 Ignatiadis M, Singhal SK, Desmedt C *et al.* Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. J Clin Oncol 2012;30:1996–2004.
4 Bianchini G, Gianni L. The immune system and response to HER2-targeted treatment in breast cancer. Lancet Oncol 2014;15:e58–e68.
5 Denkert C, Loibl S, Noske A *et al.* Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol 2010;28:105–113.
6 Issa-Nummer Y, Darb-Esfahani S, Loibl S *et al.* Prospective validation of immunological infiltrate for prediction of response to neoadjuvant chemotherapy in HER2-negative breast cancer–a substudy of the neoadjuvant GeparQuinto trial. PLoS One 2013;8:e79775.
7 Denkert C, von Minckwitz G, Brase JC *et al.* Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. J Clin Oncol 2015;33:983–991.
8 West NR, Milne K, Truong PT *et al.* Tumor-infiltrating lymphocytes predict response to anthracycline-based chemotherapy in estrogen receptor-negative breast cancer. Breast Cancer Res 2011;13:R126.
9 Yamaguchi R, Tanaka M, Yano A *et al.* Tumor-infiltrating lymphocytes are important pathologic predictors for neoadjuvant chemotherapy in patients with breast cancer. Hum Pathol 2012;43:1688–1694.
10 Schmidt M, Hellwig B, Hammad S *et al.* A comprehensive analysis of human gene expression profiles identifies stromal immunoglobulin κ C as a compatible prognostic marker in human solid tumors. Clin Cancer Res 2012;18:2695–2703.
11 Loi S, Sirtaine N, Piette F *et al.* Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02-98. J Clin Oncol 2013;31:860–867.
12 Adams S, Gray RJ, Demaria S *et al.* Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. J Clin Oncol. 2014;32:2959–2966.
13 Gu-Trantien C, Loi S, Garaud S *et al.* $CD4^+$ follicular helper T cell infiltration predicts breast cancer survival. J Clin Invest 2013;123:2873–2892.
14 Disis ML, Stanton SE. Triple-negative breast cancer: immune modulation as the new treatment paradigm. Am Soc Clin Oncol Educ Book 2015;e25–e30.
15 Salgado R, Denkert C, Demaria S *et al.* The evaluation of tumor-infiltrating lymphocytes (tumor-infiltrating lymphocytes) in breast cancer: recommendations by an International tumor-infiltrating lymphocytes Working Group 2014. Ann Oncol 2015;26:259–271.
16 von Minckwitz G, Schneeweiss A, Loibl S *et al.* Neoadjuvant carboplatin in patients with triple-negative and HER2-positive early breast cancer (GeparSixto; GBG 66): a randomised phase 2 trial. Lancet Oncol 2014;15:747–756.
17 Polley MY, Leung SC, McShane LM *et al.* An international Ki67 reproducibility study. J Natl Cancer Inst 2013;105:1897–1906.
18 Polley MY, Leung SC, Gao D *et al.* An international study to increase concordance in Ki67 scoring. Mod Pathol. 2015;28:778–786.
19 Perez EA, Press MF, Dueck AC *et al.* Immunohistochemistry and fluorescence in situ hybridization assessment of HER2 in clinical trials of adjuvant therapy for breast cancer (NCCTG N9831, BCIRG 006, and BCIRG 005). Breast Cancer Res Treat 2013;138:99–108.
20 Tung NM, Winer EP. Tumor-infiltrating lymphocytes and response to platinum in triple-negative breast cancer. J Clin Oncol 2015;33:969–971.