Breast Cancer Research and Treatment (2018) 171:1–9
https://doi.org/10.1007/s10549-018-4825-8

EDITORIAL

Reproducibility and predictive value of scoring stromal tumour infiltrating lymphocytes in triple-negative breast cancer: a multi-institutional study

Mark O'Loughlin[1] · Xavier Andreu[2] · Simonetta Bianchi[3] · Ewa Chemielik[4] · Alicia Cordoba[5] · Gábor Cserni[6,7] · Paulo Figueiredo[8] · Giuseppe Floris[9] · Maria P. Foschini[10] · Päivi Heikkilä[11] · Janina Kulka[12] · Inta Liepniece-Karele[13] · Peter Regitnig[14] · Angelika Reiner[15] · Ales Ryska[16] · Anna Sapino[19] · Aliaa Shalaby[1] · Elisabeth Specht Stovgaard[20] · Cecily Quinn[17,18] · Elaine M. Walsh[1] · Vicky Zolota[21] · Sharon A. Glynn[1] · Grace Callagy[1,22]

Received: 4 May 2018 / Accepted: 8 May 2018 / Published online: 17 May 2018

Abstract

Background Several studies have demonstrated a prognostic role for stromal tumour infiltrating lymphocytes (sTILs) in triple-negative breast cancer (TNBC). The reproducibility of scoring sTILs is variable with potentially excellent concordance being achievable using a software tool. We examined agreement between breast pathologists across Europe scoring sTILs on H&E-stained sections without software, an approach that is easily applied in clinical practice. The association between sTILs and response to anthracycline-taxane NACT was also examined.

Methodology Pathologists from the European Working Group for Breast Screening Pathology scored sTILs in 84 slides from 75 TNBCs using the immune-oncology biomarker working group guidance in two circulations. There were 16 participants in the first and 19 in the second circulation.

Results Moderate agreement was achieved for absolute sTILs scores (intraclass correlation coefficient (ICC) = 0.683, 95% CI 0.601–0.767, p-value < 0.001). Agreement was less when a 25% threshold was used (ICC 0.509, 95% CI 0.416–0.614, p-value < 0.001) and for lymphocyte predominant breast cancer (LPBC) (ICC 0.504, 95% CI 0.412–0.610, p-value < 0.001). Intra-observer agreement was strong for absolute sTIL values (Spearman $\rho = 0.727$); fair for sTILs ≥ 25% ($\kappa = 0.53$) and for LPBC ($\kappa = 0.49$), but poor for sTILs as 10% increments ($\kappa = 0.24$). Increasing sTILs was significantly associated with an increased likelihood of a pathological complete response (pCR) on multivariable analysis.

Conclusion Increasing sTILs in TNBCs improves the likelihood of a pCR. However, inter-observer agreement is such that H&E-based assessment is not sufficiently reproducible for clinical application. Other methodologies should be explored, but may be at the cost of ease of application.

Keywords Triple-negative breast cancer · Stromal tumour infiltrating lymphocytes · sTILs · Neoadjuvant chemotherapy · Inter-observer agreement · Pathological complete response

Background

The role of the immune system in the pathogenesis and clinical course of cancer is well established [1, 2] and has received renewed attention with the success of immunotherapies for several solid organ cancers such as melanoma and lung cancer. The assessment of tumour infiltrating lymphocytes (TILs) within a tumour has been used as a surrogate measure of the immune response and several studies from the 1980s onwards have reported on the prognostic role of TILs in a variety of different organ systems [3–5]. Breast cancer has historically been regarded as a non-immunogenic tumour although a dense lymphoid infiltrate has long been observed in the rare medullary subtype [6], which is associated with a favourable outcome despite its otherwise high-grade morphological features.

The stromal TIL (sTIL) component in breast cancer has been examined in a number of recent clinical studies and a prognostic role has been most consistently observed in

✉ Grace Callagy
grace.callagy@nuigalway.ie

Extended author information available on the last page of the article

triple-negative breast cancer (TNBC) and HER2-positive cancers compared to other subtypes in both the adjuvant and neo-adjuvant setting [7–16]. In two adjuvant series of TNBCs, each 10% incremental increase in sTILs was associated with a 14–19% reduction in risk for recurrence or death [13, 15]. sTIL evaluation was included as a secondary endpoint of the Gepar-sixto trial and, similarly, incremental increases of sTILs were positively associated with a pathological complete response (pCR) in TNBC patients. In that study, tumours with a dense sTIL component, termed lymphocyte predominant breast cancer (LPBC), were associated with the highest pCR rate of 74% in patients who received carboplatin [7]. LPBCs, whilst not representing a specific subtype, have sTILs occupying over 50 or 60% of the stroma and are uncommon amongst breast cancers [7, 9–13, 17]. Gene expression-based analysis of TNBC also shows that the TIL component in TNBCs is highly correlated with an immune-rich expression profile that is favourably prognostic for relapse-free survival [18].

In order for the potential of a biomarker to be realised in clinical practice, it must meet standards for analytic validity in terms of the reliability, accuracy and reproducibility of the assay. In breast cancer, sTILs are most commonly scored on haematoxylin and eosin (H&E)-stained tumour sections. Reports of the reproducibility for this methodology vary from moderate to excellent [7, 10, 13–15, 17, 19]. In 2015, an international immuno-oncology biomarker working group produced guidance aimed at standardising sTILs reporting in breast cancer [19]; and subsequently reported very high inter-observer agreement in a large ring study when this guidance was combined with an interactive software tool [20].

The aim of our multi-institutional study was to evaluate the reproducibility between experienced breast pathologists across Europe for scoring sTILs in TNBCs in routine practice. Our assessment of sTILs was confined to light microscopy using the guidance of the immuno-oncology biomarker working group on the basis that this methodology is simple to perform and could be easily applied in routine practice; the software tool was not used because this would add a level of complexity that would make it more difficult to roll out in clinical practice. The case series was limited to TNBCs for two reasons: there is consistent evidence supporting a prognostic and predictive association for sTILs in this subtype, and because any differences between subtypes would not then be a cause of variation. As a secondary endpoint, the association between sTILs and the likelihood of attaining a pCR in TNBC was examined.

Materials and methods

The series comprised 75 consecutive TNBCs diagnosed in 72 patients in a symptomatic breast service of a single tertiary referral centre between 2004 and 2015 (Table 1). All but one patient received NACT. Three of the 72 patients had multiple synchronous TNBCs. Nine patients had more than one core biopsy taken from the same tumour and these additional biopsies were included to evaluate intratumoural heterogeneity. A representative H&E-stained section of the 84 needle core biopsies (NCBs) from 75 tumours was selected and slides were scanned using an Olympus VS120 slide scanner. The digitised slides were anonymised and were uploaded to the PathXL online repository. pCR breast was defined as ypT0/is and pCR breast/axilla as ypT0/isN0 [21].

Table 1 Characteristics of the series (75 TNBCs in 72 patients)

Parameter	Total n	n	%
Patient age (years)			
Median (range)	72	48 (24–73)	
Tumour type	75		
Ductal (NST)		69	92
Metaplastic		2	3
Medullary-like		4	5
Tumour grade	75		
1		0	0
2		20	27
3		55	73
NACT	71		
Anthracycline and taxane		41	58
Anthracycline, taxane, and carboplatin		28	39
Unknown regimen		2	3
No NACT	1		
ypT[a]	71		
Is		7	10
0		30	42
1		19	27
2		7	10
3		3	4
4		5	7
ypN[a]	71		
X		1	1
0		50	70
1		7	10
2		9	13
3		4	6
pCR (Breast)	71		
pCR		37	52
Non-pCR		34	48
pCR (breast and axilla)	71		
pCR		33	46
Non-pCR		38	54

[a]Reference [21]; *n* number, *NACT* neoadjuvant chemotherapy, *pCR* pathological complete response

The study consisted of two circulations. In the first circulation, an email with instructions for the study was sent to 35 consultant pathologists who were members of the European Working Group for Breast Screening Pathology (EWG-BSP). The email included the review and the online tutorial from the immuno-oncology biomarker working group [19], links to the digitised slides, and a MS Excel template. Participants were asked to read the tutorial and the review before scoring sTILs in each slide and to record the absolute percentage of sTILs for each slide in the Excel template provided, which was then returned by the individual pathologists to the organising pathologist. Participants were also asked to record the length of time taken to complete the exercise.

After the first circulation, an independent pathologist, who was not part of the inter-observer study, reviewed those digital slides for which there was a noticeable variance in scores and noted the features pertinent to these cases e.g., necrosis, difficult boundary, tumour heterogeneity. The 84 digitised slides were relabelled and reordered at random on the PathXL online repository. 4 months after the completion of the first circulation, an email that contained links to the re-ordered slides, the TILWG tutorial and an Excel template was circulated to members of the EWG-BSP. The email for this second circulation highlighted the specific guidance in the working group tutorial that pertained to those features in the slides for which there was most disagreement in sTIL scores in the first circulation. Participants were asked to review the working group tutorial again and then record the absolute percentage of sTILs for each case in the Excel template and to return it to the organiser.

Slide selection, scanning and anonymisation were performed by a senior technician and an independent pathologist, neither of whom participated in the inter-observer study. All participating pathologists were blinded to the scores of other pathologists.

Statistical analysis

The relationships between the pathologists' scores in the different circulations and between each other were assessed as continuous variables (raw scores); as increments of ten percent; and as dichotomous categorical variables using a threshold of ≥ 25 and of $\geq 50\%$, the latter defined as LPBC. The intraclass correlation coefficient (ICC) was used to assess how closely the measurements of sTILs by different pathologists resembled each other for each slide [22, 23]. The two-way mixed single measures figure was used as it reflects the values for a single typical rater. Spearman's correlation coefficient (ρ) was used to measure the strength of the relationship between scores in circulation one and circulation two for each individual pathologist for the raw sTIL scores that were given. Cohen's kappa statistic (κ) was used to measure the strength of association between circulation one and circulation two scores for each individual pathologist for the sTILs as a categorical variable. Univariate and multivariate Logistic regression analysis was used to calculate odds ratio (OR) and 95% confidence intervals (CI) to adjust for prognostic variables. The p-values reported were two tailed and a p-value of less than 0.05 was considered statistically significant. Pearson χ^2 testing was also used to assess the association between sTILs categories and pCR. The sTIL results were collated in Microsoft Excel and were subsequently analysed in SPSS 24 and Stata/IC (v14.0).

Results

Sixteen pathologists participated in the first circulation; nineteen participated in the second circulation, comprising all sixteen pathologists who partook in the first circulation and an additional three pathologists. The average time taken by participants to score sTILs in each slide was 4 min (median 3 min; range 1–10 min).

Inter-observer agreement

The distribution of sTIL scores given by each of the 16 pathologists who participated in both circulations is shown in Fig. 1 and the distribution of sTIL scores recorded for all 84 slides is shown in Fig. 2. In circulation 1, the two-way mixed single measures ICC indicated fair agreement for absolute sTIL scores (ICC 0.595, 95% CI 0.517–0.679, p-value < 0.001) and for the sTILs $\geq 25\%$ (ICC 0.437, 95% CI 0.356–0.531, p-value < 0.001) and for the LPBC category (ICC 0.415, 95% CI 0.336–0.508, p-value < 0.001). Assessing agreement between just the original 16 participants in circulation 2, the single measures ICC increased for the absolute sTIL scores (ICC 0.683, 95% CI 0.601–0.767, p-value < 0.001) reflecting good agreement but agreement remained fair for both the sTIL $\geq 25\%$ category (ICC 0.509, 95% CI 0.416–0.614, p-value < 0.001) and for the LPBC group (ICC 0.504, 95% CI 0.412–0.610, p-value < 0.001).

When data from the 19 pathologists participating in circulation 2 was evaluated, the single measures ICC for the absolute sTIL scores was slightly less than that for the original 16 participants (ICC 0.660, 95% CI 0.577–0.747, p-value < 0.001); agreement was fair for sTILs $\geq 25\%$ (ICC 0.501, 95% CI 0.411–0.606, p-value < 0.001) and also for LPBC (ICC 0.481, 95% CI 0.391–0.588, p-value < 0.001).

Intra-observer agreement

The intra-observer agreement for the original 16 pathologists who partook in both circulations ranged from weak to very strong correlation for absolute sTIL values (Spearman $\rho = 0.314$ to 0.970; p-values range from < 0.001 to 0.015)

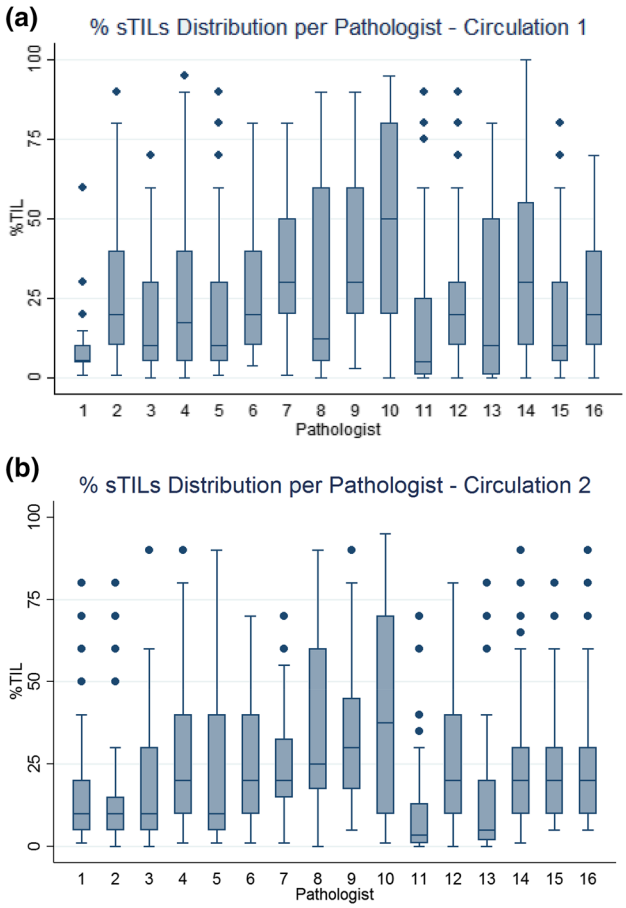

Fig. 1 Distribution of sTILs scores given by each of the 16 participating pathologists for the 84 slides in circulation 1 (**a**) and in circulation 2 (**b**). There was greater variation in the range of scores given by the 16 pathologists in circulation 1 than in circulation 2 and the range of scores given by pathologists changed between the two circulations. Pathologist 1 gave a narrow range of scores relative to other participants in circulation 1 and gave a wider range in circulation 2 that was more in line with that of other participants; the converse was observed for pathologist 11. The range of scores given by pathologist 10 was wide relative to others in both circulations. The distribution of scores given by pathologists 14, 15, and 16 converged to become very similar in circulation 2

with a strong average correlation (Spearman $\rho = 0.727$). The lowest intra-observer agreement for one pathologist (Spearman $\rho = 0.314$, p-value = 0.015) reflected a move from poor agreement between this pathologist's scores and those of the other participants in the first circulation (average inter-item correlation = 0.356) to strong agreement in the second circulation average inter-item correlation = 0.740. Overall intra-observer agreement was fair using the 25% threshold ($\kappa = 0.53$; range 0.158–0.947) and for the LPBC category ($\kappa = 0.49$; range 0.021–0.868) but agreement was poor for sTILs as 10% increments ($\kappa = 0.24$; range 0.069–0.545).

Features associated with poor agreement in sTIL scores

An independent pathologist selected the slides for which there was greatest inter-observer disagreement in sTIL scores on the basis of a standard deviation for absolute scores in the top 25%. The features that could explain this variation were intra-tumoural heterogeneity of sTILs ($n = 11$), necrosis ($n = 5$), fragmentation of the biopsy ($n = 4$), difficulties delineating the tumour boarder ($n = 4$), low ($n = 3$) and high ($n = 2$) tumour cellularity; some of these features co-existed in the same case. When sTIL scores for different biopsies from the same tumour were examined ($n = 9$), there was overall moderate agreement (Spearman $\rho = 0.511$) that was weak in three cases (lowest Spearman $\rho = 0.276$, p-value = 0.268).

Association between sTILs and response to NACT

For the 72 patients, the median sTIL score was 20% (range 1–80%) in circulation 1 and 15% (range 1–80%) in circulation 2. The distribution of sTIL categories across the 72 patients is shown in Table 2. The median sTIL score for each case from circulation 2 was used to examine the association between sTILs and response to NACT. Increasing sTILs was paralleled by an increased likelihood of both pCR breast and pCR breast/axilla by univariate and multivariable analysis (Table 3). Increasing 10% increments of sTILs improved the likelihood of both a pCR breast and pCR breast/axilla by over 40% on univariate analysis (p-value = 0.020 and p-value = 0.022, respectively). LPBC was associated with the greatest likelihood of a pCR breast and pCR breast/axilla (OR 9.1, 95% CI 1.07–77.2, p-value = 0.043; OR 11.8, 95% CI 1.39–100.6, p-value = 0.024 respectively), with the caveat that there were only nine LPBCs and a very wide 95% CI was observed. By multivariable analysis, increasing 10% increments of sTILs was an independent predictor of both pCR endpoints when adjusted for age at diagnosis, tumour grade and tumour type. Again, the magnitude of the association between sTILs and pCR on multivariable analysis was greatest for LPBC and significant for pCR breast/axilla. LPBC was associated with a higher rate of pCR breast and breast/axilla (both 89%; $n = 8$) than non-LPBC (47%; $n = 29$; Pearson χ^2 5.59. $p = 0.018$ and 40% ($n = 25$); Pearson χ^2 7.45 p-value = 0.006, respectively).

Discussion

sTILs have emerged as a potential prognostic and predictive marker in TNBC in the adjuvant and neoadjuvant setting. The consistency of scoring sTILs varies with excellent reproducibility reported when a software tool is used along

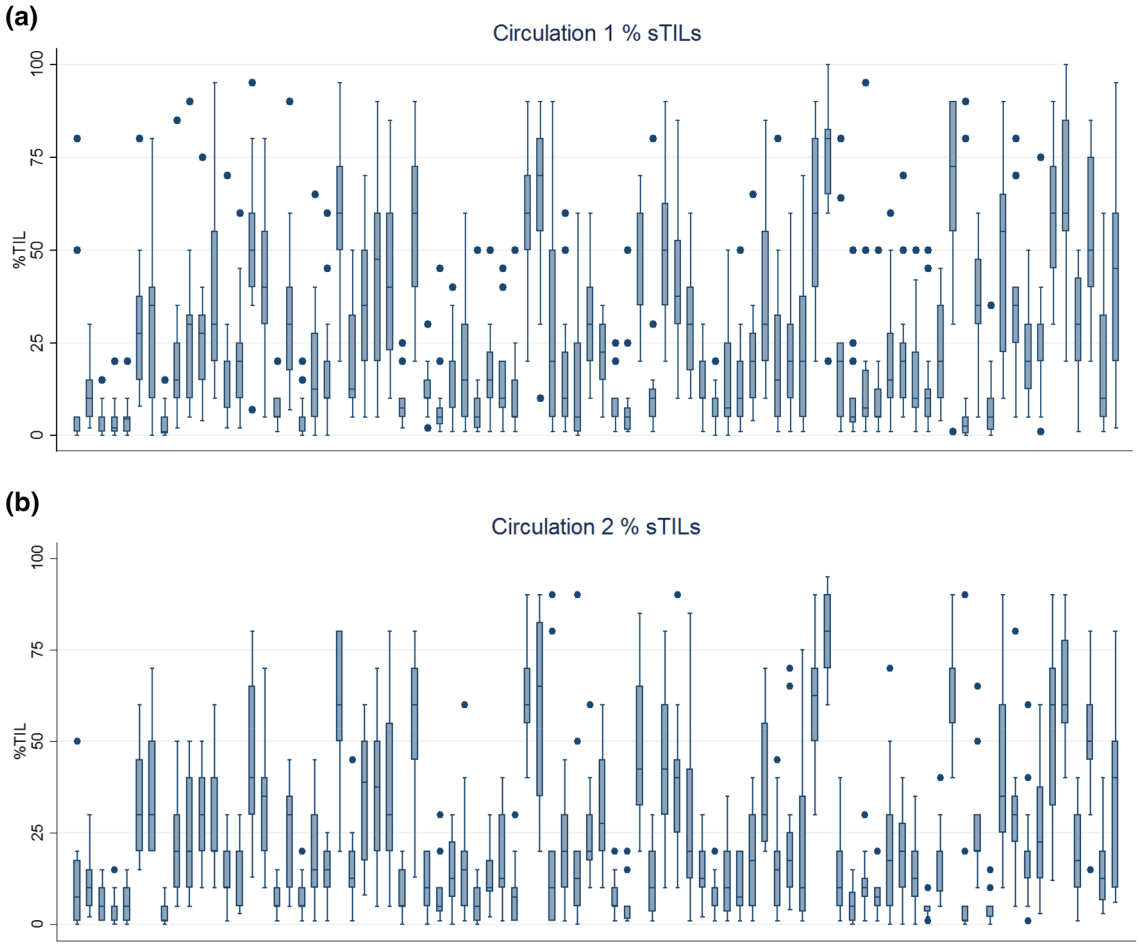

Fig. 2 Distribution of sTIL scores for all slides in circulation 1 (**a**) and in circulation 2 (**b**). The distribution of sTIL scores was less heterogeneous in circulation 2 than in circulation 1. In circulation 2, there were fewer outlier scores and there was a narrow range of scores for those cases with a low sTIL population (sTILs < 20%) indicating a better level of agreement for these cases

Table 2 Distribution of sTIL categories in 72 cases of TNBC

sTILs		sTILs	
Binary categories	*n* (%)	10% increments	*n* (%)
< 10%	34 (47)	1–10	34 (47)
≥ 10%	38 (53)	11–20	16 (22)
< 25%	50 (69)	21–30	7 (10)
≥ 25	22 (31)	31–40	3 (4)
Non-LPBC	63 (88)	41–50	4 (6)
LPBC	9 (12)	51–60	6 (8)
		61–70	1 (1)
		71–80	1 (1)

with guidance from an expert group [19]. In our study, the reproducibility of sTILs assessment in TNBC was examined using this guidance but without the software tool in order to examine reproducibility of a methodology that could be easily applied in routine practice. Our data affirm the predictive importance of sTILs in the neoadjuvant setting whereby increasing levels of sTILs are associated with increased odds of a pCR following treatment with anthracycline-based NACT. However, there was only moderate agreement at best between experienced pathologists for scoring sTILs.

The distribution of sTIL scores in our series was similar to that reported by others. The median sTIL value of 15% (range 1–80%) in circulation 2 is in line with other reports of a median of 15–23% in TNBC (9,11,12,14,15,17) and higher than that observed by some (13). LPBC was observed in 12% of cases, which was within the range of 4.4–28% noted by others in TNBCs [7, 9–13, 17]. Increasing sTILs was paralleled by an increased likelihood of a pCR on both univariate and multivariable analysis. Each 10% increase in sTILs improved the likelihood of a pCR by over 40%, which is higher than 15–23% described by others [7, 9, 11–15]. Although the number of LPBCs was small, our data suggest that the predictive relevance of sTILs may be greatest for these tumours. This is consistent with data pertaining to the

Table 3 Logistic regression of the association between sTILs and achieving a pCR

	$OR_{(pCR)}$	95% CI	P value
Univariate logistic regression analysis			
pCR Breast			
sTILs Absolute %	1.04	1.01–1.08	0.010
sTILs 10% increments	1.45	1.06–1.98	0.020
sTILs > 10%	2.64	1.01–6.89	0.048
sTILs ≥ 25%	3.56	1.18–10.64	0.023
LPBC	9.10	1.07–77.2	0.043
pCR breast and axilla			
sTILs Absolute %	1.04	1.01–1.07	0.013
sTILs 10% increments	1.41	1.05–1.88	0.022
sTILs > 10%	2.16	0.83–5.61	0.114
sTILs ≥ 25%	2.76	0.97–7.83	0.056
LPBC	11.84	1.39–100.6	0.024
Multivariable logistic regression analysis[a]			
pCR breast			
sTILs absolute %	1.05	1.01–1.09	0.016
sTILs 10% increments	1.49	1.03–2.15	0.035
sTILs > 10%	2.08	0.71–6.12	0.185
sTILs ≥ 25%	3.93	1.01–13.95	0.034
LPBC	8.29	0.81–84.55	0.074
pCR breast and axilla			
sTILs absolute %	1.04	1.01–1.07	0.013
sTILs 10% increments	1.41	1.05–1.88	0.022
sTILs > 10%	1.64	0.54–4.99	0.387
sTILs ≥ 25%	3.07	0.88–10.67	0.077
LPBC	17.6	1.19–259.35	0.037

[a]Adjusted for age at diagnosis, tumour grade, tumour type

Logistic regression with pCR = 1; non-pCR = 0

TNBC subset of the Geparsixto study, in which LPBC was associated with the greatest increase in likelihood of pCR (OR 2.17. CI 1.27–3.73, *p*-value = 0.05) [7].

Despite the strong favourable association between sTILs and response to NACT, the inter-observer agreement in our study suggests that sTIL evaluation using this methodology is not sufficiently reproducible for application in TNBCs in routine practice. Inter-observer agreement for the absolute percentage of sTILs was only moderately reproducible, reflected by an ICC of 0.683 with a lower limit of the 95% CI of 0.601. Higher levels of agreement for absolute sTIL values are reported by others but in studies involving only two or three pathologists with recorded ICC values of 0.92 [7] and 0.97 [17]; 85% agreement [13]; and strong correlation [14]. However, in studies involving more participants, reproducibility was comparable to that achieved in this work with an ICC value of 0.62 between four pathologists [24]; and an ICC of 0.71 in a ring study with 32 pathologists [20]. Both our study and the latter ring study [20] included a large number of participants and scored sTILs on full face sections of pre-treatment NCBs. Case mix and selection differed between the studies in that the ring study included cases from the Geparsixto trial of both TNBCs and HER2-positive tumours that were selected to ensure equal representation of tumours with different levels of sTILs. Our cases comprised consecutive TNBC biopsies from routine practice with no pre-selection criteria other than sufficient tumour for diagnosis. Thus, the slightly lower level of reproducibility for absolute sTIL values reported here may be more reflective of what is achievable in routine practice for TNBCs than that reported in the ring study. Despite the recommendation to score sTILs as a continuous variable [20], and good reproducibility for scoring absolute sTIL values, the clinical utility of the absolute percentage of sTILs is uncertain and has only a negligible effect on response to NACT. In contrast, 10% incremental increases in sTILs are consistently associated with response to NACT but the utility of this measure is likely to be confounded by the poor intra-observer agreement that we observed ($\kappa = 0.24$).

Denkert et al. reported excellent reproducibility when an interactive software tool was used to aid sTILs assessment [20]. This was achieved for absolute (ICC 0.89) and for categorical measures in a cohort pre-selected to include equal numbers of cases with low, intermediate and high sTIL levels. These data are very promising and the software has now become accessible on line (http://www.tilsinbreastcancer.org). The software gives the scorer integrated feedback by showing pre-calibrated reference images against which an image from a case is assessed. However, this approach increases the complexity of the evaluation process and the time taken; the user is required to create and upload three high-power static images from each case to generate a sTIL score. In our study, even without this step, the median time taken to score each slide was 4 min, which is not insignificant in the context of a busy routine practice; others report a time of between three and nine minutes to score a case using the working group guidance [25]. This is significantly longer than the time taken to score other biomarkers in breast cancer e.g., oestrogen or progesterone receptor and HER2, where estimates of positivity are given around pre-defined thresholds.

Some of the features that we observed in cases with the greatest variation in sTIL scores were highlighted by the expert group i.e., necrosis, difficulty delineating the boarder [19]; others were not e.g., fragmentation of the core and tumour cellularity. Many of these features are not uncommon in TNBC biopsies and, consequently, may hamper attempts to improve reproducibility of scoring in this subtype. Intra-tumoural heterogeneity was seen most often and we observed only moderate agreement (ranging from weak to strong) between sTIL scores in paired biopsies from the same tumour. It will be difficult to mitigate the effect of

heterogeneity on the analytic and clinical validity of sTILs in TNBCs because of the potential for sampling bias arising from the reliance on NCBs in the neoadjuvant setting. The working group guidance recommends giving an average sTIL score for a case; however, there have been no formal studies examining either the effect of heterogeneity on reproducibility or the relative clinical importance of a sTIL score derived from the average sTIL population, the hot-spots or the area with the lowest sTILs.

Our study has limitations. The number of cases was small and, as in many studies, the small number of LPBCs hampered the interpretation of their significance. We restricted our analysis to TNBC, in which the potential clinical relevance of sTILs has been shown most consistently, and our data may not be pertinent to other subtypes of breast cancer. For example, in a previous study on 454 breast cancers treated with NACT, we showed that the presence or absence of sTILs, with a cut off of 1%, significantly impacted on response to treatment in the Luminal B oestrogen receptor-positive/HER2-negative subtype [26]. Nonetheless, our series is an accurate reflection of TNBCs that are diagnosed in routine practice with no pre-selection applied. Participating pathologists did not receive formal training in scoring sTILs in the interval between the two circulations. Formal training of pathologists is emphasised to improve the consistency of reporting many of the newer predictive immunohistochemical biomarkers in other cancer types [27] and it could be argued that training could have improved consistency for scoring sTILs in this work. In our second circulation, we observed slightly better agreement between the sixteen pathologists who had already undertaken the exercise in the first circulation (ICC 0.683, 95% CI 0.601–0.767, $p < 0.001$) than we observed when scores from three new participants were included (ICC 0.660, 95% CI 0.577–0.747, $p < 0.001$). Notwithstanding, the participants were experienced breast pathologists from across Europe that would be representative of international best practice. Finally, the aim of our study was to assess concordance in measuring the whole sTIL population; we did not examine subpopulations of lymphoid cells which may provide a more functional assessment of the immune infiltrate [28–31].

In conclusion, our data affirm the predictive significance of sTILs with respect to pCR in TNBC. Quantification of sTILs by light microscopy is simple and would be suitable for widespread clinical application; however, our data show considerable inter- and intra-observer variability between experienced breast pathologists in the assessment of sTILs using the immuno-oncology biomarker working group guidance alone. Tumour heterogeneity contributed to reproducibility issues. Other methodologies may improve the consistency for scoring sTILs and should continue to be explored. The software tool of the immuno-oncology biomarker working group has the potential to improve standardisation but at the expense of decreased ease of use. Further studies will need to validate this tool for scoring sTILs in cases from routine practice, without the pre-selection of cases applied in the ring study (21), and to determine if can overcome the effect on heterogeneity on reproducibility. Formal studies evaluating the clinical importance of sTIL heterogeneity with data to support guidance on how heterogeneous cases should be scored is required. Methodological studies aimed at improving the consistency of reporting sTILs may need to be designed with specific tumour subtypes and clinical endpoints in mind.

Funding This work was funded by Breast Cancer Now (2013MayPR019 and 2015NovPhD643), Science Foundation Ireland (17/CDA/4638) and Irish Cancer Society (SG); Breast Cancer Research (EW), and an NUI Galway School of Medicine Scholarship (EW).

Data availability Data that support this study are available upon reasonable request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval This work complies with regulations governing ethical standards. Informed consent was obtained from patients who participated in this study and the project was approved by the Clinical Research Ethics Committee, Galway University Hospital (Ref. CA1012) on 23rd January 2014.

References

1. Pagès F, Galon J, Dieu-Nosjean M-C, Tartour E, Sautès-Fridman C, Fridman W-H (2010) Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene 29(8):1093–1102
2. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C et al (2006) Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 313(5795):1960–1964
3. Jass JR (1986) Lymphocytic infiltration and survival in rectal cancer. J Clin Pathol 39(6):585–589
4. Marrogi AJ, Munshi A, Merogi AJ, Ohadike Y, El-Habashi A, Marrogi OL et al (1997) Study of tumor infiltrating lymphocytes and transforming growth factor-β as prognostic factors in breast carcinoma. Int J Cancer 74(5):492–501
5. Sharma P, Shen Y, Wen S, Yamada S, Jungbluth AA, Gnjatic S et al (2007) CD8 tumor-infiltrating lymphocytes are predictive of survival in muscle-invasive urothelial carcinoma. Proc Natl Acad Sci USA 104(10):3967–3972

6. Moore OS, Foote FW (1949) The relatively favorable prognosis of medullary carcinoma of the breast. Cancer 2(4):635–642
7. Denkert C, Von Minckwitz G, Brase JC, Sinn BV, Gade S, Kronenwett R et al (2015) 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 33(9):983–991
8. Denkert C, Loibl S, Noske A, Roller M, Müller BM, Komor M et al (2010) Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol 28(1):105–113
9. Dieci MV, Criscitiello C, Goubar A, Viale G, Conte P, Guarneri V et al (2014) Prognostic value of tumor-infiltrating lymphocytes on residual disease after primary chemotherapy for triple-negative breast cancer: a retrospective multicenter study. Ann Oncol 25(3):611–618
10. Issa-Nummer Y, Darb-Esfahani S, Loibl S, Kunz G, Nekljudova V, Schrader I et al (2013) 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 8(12)
11. Pruneri G, Gray KP, Vingiani A, Viale G, Curigliano G, Criscitiello C et al (2016) Tumor-infiltrating lymphocytes (TILs) are a powerful prognostic marker in patients with triple-negative breast cancer enrolled in the IBCSG phase III randomized clinical trial 22-00. Breast Cancer Res Treat 158(2):323–331
12. Dieci MV, Mathieu MC, Guarneri V, Conte P, Delaloge S, Andre F et al (2015) Prognostic and predictive value of tumor-infiltrating lymphocytes in two phase III randomized adjuvant breast cancer trials. Ann Oncol 26(8):1698–1704
13. Adams S, Gray RJ, Demaria S, Goldstein L, Perez EA, Shulman LN et al (2014) 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 32(27):2959–2966
14. Loi S, Michiels S, Salgado R, Sirtaine N, Jose V, Fumagalli D et al (2014) Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial. Ann Oncol 25(8):1544–1550
15. Loi S, Sirtaine N, Piette F, Salgado R, Viale G, Van Eenoo F et al (2013) 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 31(7):860–867
16. Luen SJ, Savas P, Fox SB, Salgado R, Loi S (2017) Tumour-infiltrating lymphocytes and the emerging role of immunotherapy in breast cancer. Pathology 49(2):141–155
17. Pruneri G, Vingiani A, Bagnardi V, Rotmensz N, De Rose A, Palazzo A et al (2016) Clinical validity of tumor-infiltrating lymphocytes analysis in patients with triple-negative breast cancer. Ann Oncol 27(2):249–256
18. Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y et al (2016) Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. PLoS ONE 11(6):e0157368
19. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G et al (2015) The evaluation of tumor-infiltrating lymphocytes (TILS) in breast cancer: Recommendations by an International TILS Working Group 2014. Ann Oncol 26(2):259–271
20. Denkert C, Wienert S, Poterie A, Loibl S, Budczies J, Badve S et al (2016) Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. Mod Pathol 29(10):1155–1164
21. Wittekind C, Brierley JD, Gospodarowicz MK (2017) TNM Classification of Malignant Tumours. 8th edn. Wiley
22. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6(4):284–290
23. Hallgren K (2012) Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 8(1):23–24
24. Swisher SK, Wu Y, Castaneda CA, Lyons GR, Yang F, Tapia C et al (2016) Interobserver agreement between pathologists assessing tumor-infiltrating lymphocytes (TILs) in breast cancer using methodology proposed by the International TILs Working Group. Ann Surg Oncol 23(7):2242–2248
25. Hida AI, Ohi Y (2015) Evaluation of tumor-infiltrating lymphocytes in breast cancer; proposal of a simpler method. Ann Oncol 26(11):2351
26. Balmativola D, Marchiò C, Maule M, Chiusa L, Annaratone L, Maletta F et al (2014) Pathological non-response to chemotherapy in a neoadjuvant setting of breast cancer: an inter-institutional study. Breast Cancer Res Treat 148(3):511–523
27. Cooper WA, Russell PA, Cherian M, Duhig EE, Godbolt D, Jessup PJ et al (2017) Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non-small cell lung cancer. Clin Cancer Res 23(16):4569–4577
28. Ali HR, Provenzano E, Dawson S-J, Blows FM, Liu B, Shah M et al (2014) Association between CD8+T-cell infiltration and breast cancer survival in 12,439 patients. Ann Oncol 25(8):1536–1543
29. Liu S, Foulkes WD, Leung S, Gao D, Lau S, Kos Z et al (2014) Prognostic significance of FOXP3+ tumor-infiltrating lymphocytes in breast cancer depends on estrogen receptor and human epidermal growth factor receptor-2 expression status and concurrent cytotoxic T-cell infiltration. Breast Cancer Res 16(5)
30. Liu S, Lachapelle J, Leung S, Gao D, Foulkes WD, Nielsen TO (2012) CD8+ lymphocyte infiltration is an independent favorable prognostic indicator in basal-like breast cancer. Breast Cancer Res 14(2):48
31. West NR, Milne K, Truong PT, Macpherson N, Nelson BH, Watson PH (2011) Tumor-infiltrating lymphocytes predict response to anthracycline-based chemotherapy in estrogen receptor-negative breast cancer. Breast Cancer Res 13(6):126

Affiliations

Mark O'Loughlin[1] · Xavier Andreu[2] · Simonetta Bianchi[3] · Ewa Chemielik[4] · Alicia Cordoba[5] · Gábor Cserni[6,7] · Paulo Figueiredo[8] · Giuseppe Floris[9] · Maria P. Foschini[10] · Päivi Heikkilä[11] · Janina Kulka[12] · Inta Liepniece-Karele[13] · Peter Regitnig[14] · Angelika Reiner[15] · Ales Ryska[16] · Anna Sapino[19] · Aliaa Shalaby[1] · Elisabeth Specht Stovgaard[20] · Cecily Quinn[17,18] · Elaine M. Walsh[1] · Vicky Zolota[21] · Sharon A. Glynn[1] · Grace Callagy[1,22]

1 Discipline of Pathology, National University of Ireland, Galway, Ireland

2 Pathology Department, UDIAT, Centre Diagnostic, Corporacio Sanitaria del Parc Taulí-Institut Universitari Parc Taulí-UAB, Sabadell, Spain

3 Department of Surgery and Translational Medicine, AOU Careggi, University of Florence, Largo G. Brambilla 3, 50134 Florence, Italy

4 Tumor Pathology Department, Maria Sklodowska-Curie Institute-Oncology Center, Gliwice, Poland

5 Department of Pathology Section A, Navarra Health Service, Hospital Complex of Navarra, Irunlarrea 4, 31008 Pamplona, Spain

6 Department of Pathology, Bács-Kiskun County Teaching Hospital, Nyiri ut 38., 6000 Kecskemét, Hungary

7 Department of Pathology, University of Szeged, Állomás u. 1, 6725 Szeged, Hungary

8 Lab Histopatologia, Av Bissaya Barreto, Apartado 2005, 3001-651 Coimbra, Portugal

9 Department of Pathology, University Hospitals Leuven, Leuven, Belgium

10 Department of Biomedical and Neuromotor Sciences, Section of Anatomic Pathology University of Bologna, Ospedale Bellaria Via Altura 3, 40139 Bologna, Italy

11 Department of Pathology, Helsinki University Central Hospital, Helsinki, Finland

12 2nd Department of Pathology, Semmelweis University Budapest, Üllői út 93, 1091 Budapest, Hungary

13 Pathology Centre, Riga East Clinical University Hospital, Riga, Latvia

14 Medizinische Universität Graz, Institut für Pathologie, Graz, Austria

15 Institute of Pathology, Danube Hospital, Langobardenstrasse 122, 1220 Vienna, Austria

16 Department of Pathology, Charles University Medical Faculty Hospital, Hradec Kralove, Czech Republic

17 Irish National Breast Screening Programme, BreastCheck, Dublin, Ireland

18 School of Medicine, University College Dublin, Dublin, Ireland

19 Dip. di Scienze Mediche, Candiolo Cancer Institute - FPO, IRCCS, Università di Torino, Turin, Italy

20 Pathology Department, Herlev University Hospital, Herlev, Denmark

21 Department of Pathology, Rion University Hospital, University of Patras, Medical School, Patras, Greece

22 Division of Anatomic Pathology, Galway University Hospital, Newcastle Road, Galway, Ireland