Pathology - Research and Practice 229 (2022) 153718

Contents lists available at ScienceDirect

Pathology - Research and Practice

journal homepage: www.elsevier.com/locate/prp

ONEST (Observers Needed to Evaluate Subjective Tests) suggests four or more observers for a reliable assessment of the consistency of histological grading of invasive breast carcinoma: A reproducibility study with a retrospective view on previous studies

Bálint Cserni [a], Rita Bori [b], Erika Csörgő [b], Orsolya Oláh-Németh [c], Tamás Pancsa [c], Anita Sejben [c], István Sejben [b], András Vörös [c], Tamás Zombori [c], Tibor Nyári [d], Gábor Cserni [b,c,*,**]

[a] TNG Technology Consulting GmbH, Király u. 26., 1061 Budapest, Hungary
[b] Bács-Kiskun County Teaching Hospital, Department of Pathology, Nyíri út 38., 6000 Kecskemét, Hungary
[c] University of Szeged, Department of Pathology, Állomás u. 1., 6720 Szeged, Hungary
[d] University of Szeged, Department of Medical Physics and Informatics, Korányi fasor 9., 6720 Szeged, Hungary

ARTICLE INFO

Keywords:
Breast cancer
Histological grade
Reproducibility
ONEST (Observers Needed to Evaluate Subjective Tests)

ABSTRACT

Histological grade is one of the most important prognosticators of breast cancer which is available for nearly all cases. It also makes part of several multivariable analysis derived combined prognostic profiles despite concerns about its reproducibility. The aims included a reproducibility study of grading in the light of a recently described statistical approach, ONEST (Observers Needed to Evaluate Subjective Tests) and review earlier reproducibility studies in the light of the ONEST analysis. Nine pathologists reviewed 50 core needle biopsies and 50 slides from different excision specimens and recorded the scores for gland (tubule) formation, nuclear pleomorphism and mitotic activity as well as histological grade. Overall percent agreement, Fleiss kappa and the intraclass correlation coefficient (ICC) were used for the analysis of reproducibility. ONEST data and curves were generated from 100 random permutations of the participants. ONEST suggested a minimum of 4 observers for the reliable evaluation of reproducibility for both the scored components and grade in either type of specimen. Our results suggested moderate or moderate to good reproducibility of grading (kappa values of 0.51 for excisions, and 0.54 for biopsies and ICCs of 0.70 and 0.69, respectively) with gland formation being the most and nuclear pleomorphism the worst consistently evaluated feature. In studies with sufficient participants (at least 4) and non-pairwise comparisons in the analysis, the reproducibility of histological grading is fair to moderate, whereas studies with fewer participants or pairwise kappa analysis suggest moderate to almost prefect agreement of the results. ONEST is a valuable complementation of reproducibility analyses.

1. Introduction

The grade of differentiation is a prognostic parameter reflecting the biology of the tumor, and histologic grade has been part of breast cancer classification since the first edition of the World Health Organization (WHO) histological typing of breast tumors [1]. This grading scheme stemming from the original publications of Patey and Scarff from 1928 [2] and Bloom and Richardson from 1957 [3], was refined and standardized according to the Nottingham protocol [4], and is still part of the mandatory items of breast cancer reporting [5–7]. As a factor with proven prognostic impact [8], it is also part of a number of multivariable analysis derived multiparameter prognostic tools like the Nottingham Prognostic Index [9], Adjuvant!Online! [10] and Predict [11] or the prognostic staging of breast carcinomas defined by the 8th edition of the Cancer Staging Manual by the American Joint Committee on Cancer [12].

Despite the recognized prognostic impact of histological grade, issues about the less than perfect reproducibility of grading have been the subject of several publications summarized by van Dooijeweert et al.

* Correspondence to: University of Szeged, Department of Pathology, Állomás u. 1., 6720 Szeged, Hungary.
** Correspondence to: Bács-Kiskun County Teaching Hospital, Department of Pathology, Nyíri út 38., 6000 Kecskemét, Hungary.
E-mail address: cserni@freemail.hu (G. Cserni).

https://doi.org/10.1016/j.prp.2021.153718
Received 23 October 2021; Received in revised form 21 November 2021; Accepted 25 November 2021
Available online 6 December 2021

Nomenclature

CNB	Core Needle Biopsy
ER	Estrogen Receptor
EXC	EXCision (specimen)
ICC	Intraclass Correlation Coefficient
ONEST	Observers Needed to Evaluate Subjective Tests
OPA	Overall Percent Agreement
OPAC	Overall Percent Agreement Curve
PR	Progesterone Receptor
WHO	World Health Organization

[8]. Most often, overall agreement and kappa statistics have been used to reflect the consistency of defining the histological grade of breast cancers. ONEST (Observers Needed to Evaluate Subjective Tests) is a recently described complementary method to characterize reproducibility [13,14]. It has been known for a long time that 2 observers generally agree on all cases better than 3 or more observers, and ONEST tries to approximate the number of observers required for a sufficiently confident assessment of reproducibility.

In brief, ONEST involves the plot (ONEST plot) of overall percent agreement (OPA, the proportion of cases on which the given number of raters agree) as a function of the increasing number of observers from a random selection of 100 permutations of observers (resulting in 100 OPA curves (OPACs) for one ONEST plot; see also Fig. 2 in the Results section for illustration). A given (n) number of observers can be listed in ranked order in n! (factorial product of n) series. In a previous work, we have shown that the ONEST plot from 100 random permutations reliably reflect all OPACs from 9! = 362,880 possible permutations when 9 observers were involved in evaluating the estrogen receptor (ER) and progesterone receptor (PR) status along with the Ki67 labeling index of breast carcinomas. Therefore, ONEST is a good approximation which is easy to handle with only 100 permutations to analyze.

In this study, as a new approach, we have used ONEST to characterize histologic grading, and have looked at previous studies in the light of the number of observers involved in them and their reliability to assess reproducibility.

2. Materials and methods

A series of breast cancer cases including 50 nearly consecutive core needle biopsies (CNB) and 50 single slides from different excision (EXC) specimens used previously for an ER, PR and Ki67 assessment study were used for determining the histological grade of the tumors by 9 observers involved in breast cancer reporting for at least more than 1 year up to more than 25 years. Details of case selection are not of specific importance, and are described in the previous study [13].

All observers were asked to grade the 100 cases according to current practice, as recommended by the most recent WHO Classification of breast tumors [7] and report the scores for tubule/gland formation, nuclear pleomorphism and mitotic counts, along with the histological grade of the tumors.

Beside descriptive statistics, the intraclass correlation coefficients (two-way random effects, absolute agreement, single rater/measurement; ICC(2,1) [15]), Fleiss kappa values [16] and ONEST analysis [13,14] were used. For the ICC values (based on the lower 95% confidence interval, CI), the following categorical interpretation was used: < 0.5, poor; 0.5–0.749, moderate; 0.750–0.9, good and > 0.9, excellent agreement [15]. For kappas, the interpretation by Landis and Koch was considered with values < 0 reflecting poor, 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial and 0.81–1 almost perfect (i.e. excellent) agreement [17]. For ONEST, instead of using the 100 individual OPACs, the minimum and maximum values were drawn along with an average curve derived from the OPA values belonging to the 100 random permutations [14]. For the comparison of ICCs, the Kruskal-Wallis test was used, with $p < 0.05$ used as significance level. Calculations were performed in Excel with the Real Statistics Add-Ins [18].

In the light of our findings, previous reproducibility studies of the histological grading using the Nottingham modification of the original Scarff Bloom Richardson scheme [4] were looked at, and their results analyzed on the basis of their statistical approaches, and the number of observers involved in generating the figures. A recent review by van Dooijeweert, van Diest and Ellis [8] was used to identify the relevant reproducibility studies, with additional ones from the references of these studies or personal involvement.

No ethical permission was deemed necessary for this retrospective non-interventional study, which did not involve any patient data; all slides used were anonymous.

3. Results

The individual ratings for the component scores of histological grade and the grade itself are represented in Fig. 1. Less than third (n = 29) of the cases were unanimously graded with a rather equal distribution of cases within each grade. As the majority grades were G1 (22 =9 +13), G2 (50 =26 +24) and G3 (28 =15 +13) on CNB and EXC cases, the proportion of uniformly graded cases also reflects the worse consistency of determining the middle category of G2. The majority grades are also reflected on Fig. 1.

The kappa and ICC values are shown in Tables 1 and 2, respectively. These values reflect that the reproducibility of histological grading was moderate or moderate to good, with individual components being less reproducible; tubule / gland formation being the most consistently assessed feature. Interestingly, the consistency of scoring tubule formation and nuclear pleomorphism assessment was somewhat better on excision specimens. Pleomorphism was the least reproducibly scored component of histological grade. In general, the middle categories were less reproducible than the extremes (Table 1).

Regarding ONEST, the plots are reproduced in Fig. 2, and the main values are shown in Table 3. The graphs and table are in keeping with previous analyses based on kappa and ICC values, demonstrating that tubule formation is the most consistently reproducible part of histological grading, and nuclear pleomorphism is the least consistent one. About one quarter of the cases on both CNB and EXC specimens are differently graded by 2 pathologists in the worst scenario, whereas 78% (EXC) to 80% (CNB) are identically graded in the best one; on average, two pathologists are agreeing on the grade in 70% of the cases. Importantly, the ONEST plots suggest that a minimum of 4 pathologists would be required for the reliable assessment of grade reproducibility; this is where the minimum OPACs start to level off and reach a plateau (Table 3, Fig. 2).

The comparison of the average OPA values for tubule formation, nuclear pleomorphism, mitotic activity and grade demonstrated no significant difference between CNB and EXC specimens (Kruskal Wallis test, p ranging between 0.14 and 0.85). For the minimum OPA values, these were significantly different for CNB and EXC specimens in the cases of nuclear pleomorphism ($p = 0.006$) and histological grade ($p = 0.042$), being worse for CNB specimens in the first, and better for CNB specimens in the second.

Tables 4 and 5 demonstrate the results of previous studies on histological grading on the basis of kappa values (Table 4) [19–35] and OPA of all observers (Table 5) [19,21,22,24–27,29–32,36,37]. Both of these

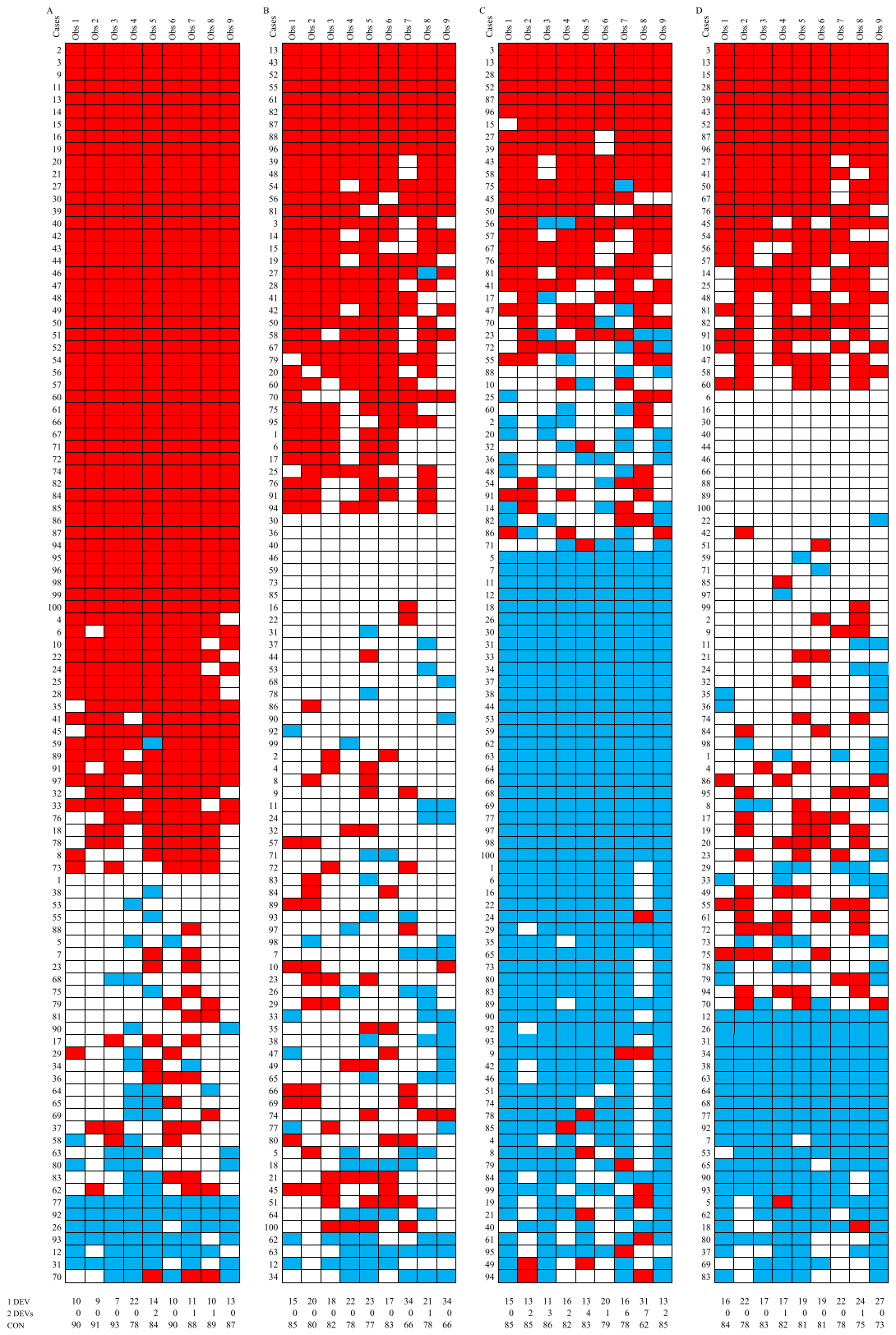

Fig. 1. Representation of individual scores for tubule/gland formation (A), nuclear pleomorphism (B), mitotic activity (C) and histological grade (D). Obs: observer, 1 DEV: 1 deviation from majority rating; 2 DEVs: 2 deviations from majority rating; CON: concordance with majority. Note: red represents 3, white 2 and blue 1. Cases are represented from top to bottom as majority scores/grade 3, 2 and 1 with decreasing number of majority ratings and case serial numbers; cases 1–50 are CNB (core needle biopsy specimens) and 51–100 are EXC (excision specimens). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article. In greyscale, red becomes black, white stays white and blue turns to grey.)

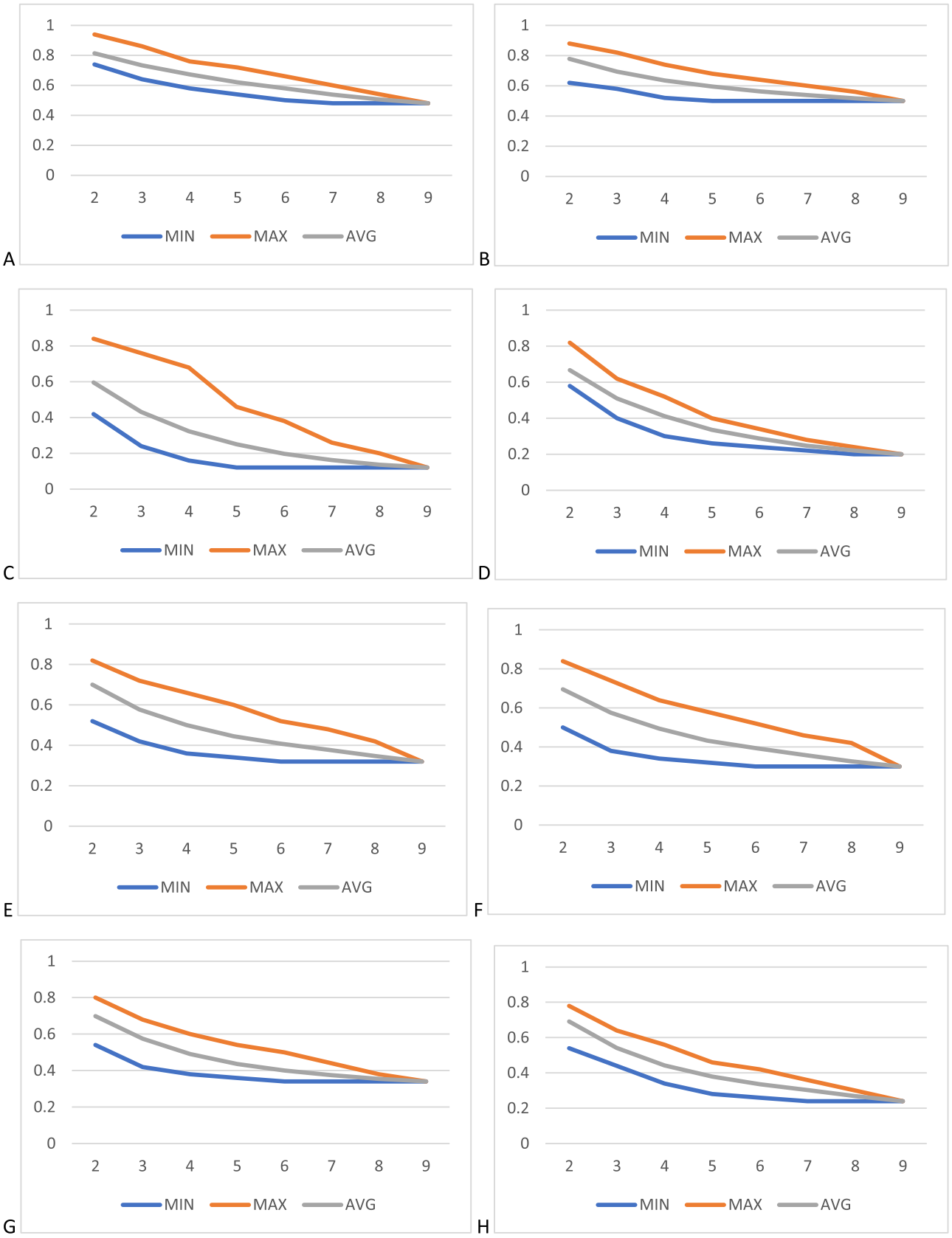

Fig. 2. ONEST plots with minimum (MIN), average (AVG) and maximum (MAX) OPA values for tubule/gland formation (A and B), nuclear pleomorphism (C and D), mitotic activity (E and F) and histological grade (G and H) for CNB (A, C, E and G) and EXC (B, D, F and H) specimens.

Table 1
Kappa values for component scores of histological grade and grade itself.

Fleiss kappa					Kappa scale	Interpretation [17]
Parameter	Score 1	Score 2	Score 3	Overall	-1 to -0.01	poor
Tubule scores (CNB)	0.64	0.45	0.62	0.56	0-0.10	slight
Tubule scores (EXC)	0.76	0.50	0.47	0.61	0.11-0.20	slight
Pleomorphism score (CNB)	0.44	0.25	0.19	0.32	0.21-0.30	fair
Pleomorphism score (EXC)	0.54	0.35	0.25	0.42	0.31-0.40	fair
Mitosis score (CNB)	0.58	0.18	0.58	0.48	0.41-0.50	moderate
Mitosis score (EXC)	0.57	0.16	0.58	0.47	0.51-0.60	moderate
Grade (CNB)	0.59	0.43	0.62	0.54	0.61-0.70	substantial
Grade (EXC)	0.50	0.37	0.70	0.51	0.71-0.80	substantial
					0.81-0.90	almost perfect
					0.91-1	almost perfect

Table 2
ICC values for component scores of histological grade and grade itself.

Parameter	ICC	95%CI	ICC Scale	Interpretation [15]
Tubule scores (All	0.735	(0.673-0.795)	<0.5	poor
Tubule scores (CNB)	0.732	(0.644-0.814)	95%CI<0.5	moderate to poor
Tubule scores (EXC)	0.733	(0.642-0.817)	0.5-0.749	moderate
Pleomorphism score (All)	0.507	(0.426-0.594)	95%CI>0.749	good to moderate
Pleomorphism score (CNB)	0.459	(0.346-0.588)	0.75-0.9	good
Pleomorphism score (EXC)	0.561	(0.452-0.676)	>0.9	excellent
Mitosis score (All)	0.673	(0.600-0.744)		
Mitosis score (CNB)	0.685	(0.587-0.779)		
Mitosis score (EXC)	0.667	(0.565-0.765)		
Grade (All)	0.692	(0.623-0.758)		
Grade (CNB)	0.687	(0.588-0.781)		
Grade (EXC)	0.700	(0.605-0.791)		

Table 3
Main data from the ONEST analysis.

	Minimum observer needed: ONEST curve reading (MIN)		Maximum difference in OPA for 2 observers		Average OPA for 2 observers		Overall agreement of all observers	
	CNB	EXC	CNB	EXC	CNB	EXC	CNB	EXC
TUB	4	4	20%	26%	81%	78%	48%	50%
PLEOM	4	4	42%	24%	60%	67%	12%	20%
MIT	4	4	30%	34%	70%	70%	32%	30%
Grade	3	4	26%	24%	70%	69%	34%	24%

CNB: core needle biopsy; EXC: excision; MIN: minimum curve/values; MIT: mitoses; OPA: overall proportion agreement, PLEOM: pleomorphism; TUB: tubule formation.

tables suggest that reproducibility figures gained with less than 4 observers or by pairwise comparisons are better.

4. Discussion

Histological grade is one of the most important traditional prognosticators of breast cancer. Semi-quantitatively reflecting how much a tumor deviates from normal lumen forming breast parenchyma, how much the nuclei enlarge and become different in shape from the normal epithelial cells, and how much it is proliferating on the basis of its mitotic activity, grade gives a morphological assessment of the potential biological behavior of the given carcinoma. Despite concerns about the less than perfect reproducibility of grading, this factor has retained its importance over the years and has been included in several

Table 4
Kappa values gained in different studies of histological grade reproducibility.

Interpretation [17] Scale	Reference	Observers	Cases	Kappa	poor	slight		fair		moderate		substantial		almost perfect		Scale
Jacquemier [19]		2 (21)	24	a							0.53					<0
Sicca [20]		2 (3)	40	b								0.68-		-0.83		0-0.10
Meyer [21]		2 (7)	49	c						0.50-	-0.59					0.11-0.20
Robbins [22]*		2 (5)	50	d									0.73			0.21-0.30
Robbins [22]**		2 (5)	50	d							0.58					0.31-0.40
Anderson [23]		2	52	b							0.54					0.41-0.50
Frierson [24]		2 (6)	75	b						0.43-			-0.74			0.51-0.60
Rabe [25]		2 (6)	100	b							0.58-			-0.86		0.61-0.70
Ginter [26]		2 (6)	143	b					0.35-			-0.68				0.71-0.80
Postma [27]		2	310	e									0.80			0.81-0.90
Reed [28]		2	613	b								0.69				0.91-1
Bueno-de-Mesquita [29]		2	694	f							0.56					
Cserni [30]		3	75	g						0.41						
Rabe [25]		6	100	g								0.68				
Ginter [26]		6	143	g						0.50						
Boiesen [31]		7	93	g							0.54					
present (CNB)		9	50	g							0.54					
present (EXC)		9	50	g							0.51					
Longacre [32]		13	35	h					0.40-			-0.70				
Sloane [33]		23	57	i							0.53					
Ellis [34]***		>200	76	j				0.24-	-0.36							
Ellis [34]****				j						0.45-	-0.53					
Rakha [35]		>600	104	j					0.34-		-0.56					

CNB: core needle biopsy; EXC: excision.
a: mean (expert vs non-expert); b: pairwise; c: average pairwise in 5 consecutive tests of 10–23 cases; d: 3 pathologists' consensus vs 2 pathologists' consensus; e: central vs local; f: mean (local vs central); g: Fleiss; category specific kappa (0.40 for G2, 0.7 for G1 and G3); i: weighted; j: Fleiss kappa (overall) range for consecutive circulations.
*B5-fixed; ** buffered formal saline fixed; *** before application of revised guidelines; **** after application of revised guidelines.

multivariable analysis derived combined prognosticators [9–12], proving that the degree of subjectivity in its determination does not interfere with its independency in multivariable models.

Our study reproduced several previous observations on the reproducibility of histological grading. In keeping with the long-term experience of the United Kingdom external quality assurance scheme in breast pathology, tubule formation is the best reproducible component of the 3 elements, and nuclear pleomorphism is the worst [35]. The middle categories are generally less reproducible than the extremes (the low and the high score categories), and the middle category of mitotic activity was the worst reproducible element [35]. Overall, we found that the reproducibility of grading was moderate (kappa values >0.50, but <0.6; Table 1) or good to moderate (ICC values 0.687–0.700, Table 2). OPA values would suggest a somewhat poorer reproducibility with full agreement of all 9 observers seen in only 29% (with fewer cases in EXC specimens than in CNBs), but 47% of the cases had 9/9 or 8/9 majority grade allocation. Deviations from majority opinion were generally of one grade with only 2/450 ratings showing the opposite: both of these were G3 allocations for two different cases by two different pathologists for 6/9 and 7/9 majority grade 1 lesions, respectively (Fig. 1). The fact that discrepant grade allocations are always or almost always only at one grade difference from majority rating is also a common finding in previous reproducibility studies [19–37]. In 1994, Dalton et al. have assumed that virtually all pathologists should be able to adequately grade breast cancers. However, breast cancer grading requires experience and routine: the least experienced participant of this study had the highest deviation rate from majority ratings, whereas the most experienced one had the least deviation. Training and adequate guidelines are also necessary, since the revised guidelines led to a relevant improvement in the consistency of grading in the United Kingdom external quality assurance scheme (Table 4) [34]. Following a Dutch nationwide study documenting relevant inter- and intradepartmental variations in the distribution of histological grades [38], both anonymized specific feedback to the laboratories and pathologists [39] and e-learning [40] have helped to decrease this variation. Not least, optimal tissue preservation is also required for adequate grading, better fixation has been

Table 5
Overall proportion agreement (OPA) values gained in different studies of histological grade reproducibility.

Reference	Observers	Cases	Comment	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1
Robbins [22]*	5	50	a									0.83	
Robbins [22]**	5	50	a								0.74		
Postma [27]	2	310										0.88	
Bueno-de-Mesquita [29]	2	694									0.72		
Cserni [30]	3	75						0.44					
Theissig [36]	3	166									0.72		
Frierson [24]	6	75					0.40						
Rabe [25]	6	100	b						0.54				
Ginter [26]	6	143				0.30							
Meyer [21]	5-7	72	c							0.67	0.74		
Boiesen [31]	7	93					0.31						
present (CNB)	9	50					0.34						
present (EXC)	9	50				0.24							
Longacre [32]	13	35					0.40						
Jacquemier [19]	21	24								0.69			
Dalton [37]	25	10				0.30							

Scale

Scale
0-0.10
0.11-0.20
0.21-0.30
0.31-0.40
0.41-0.50
0.51-0.60
0.61-0.70
0.71-0.80
0.81-0.90
0.91-1

CNB: core needle biopsy; EXC: excision.
*B5-fixed; ** buffered formal saline fixed.
a: statistically different OPA between the reproducibility with the two fixatives; b: 2 rounds evaluated, 54% agreement in round 1 and 58% in round 2; c: best and worst OPA from 5 consecutive tests of 10–23 cases by 5–7 pathologists.

documented to result in better grading consistency [22]. When looking at national databases or greater cohorts of patients, it also appears that the frequency of a given grade may be different. For example, the previously cited Dutch nationwide report on grade suggests that the proportion of grade 1 tumors makes up 28% of 33,043 patients [38], whereas data from the Survival Epidemiology and End Results database suggest only 21% of 746,507 breast cancers being of this grade [41], and the United Kingdom registry data of 5694 breast cancers from which the multivariable predictive tool PREDICT had been derived from included only less than 18% grade 1 tumors [42]. Although such differences may stem from differences in populations, as a higher rate of screen detected cancers leads to a greater proportion of well differentiated carcinomas [43], differences in training, teaching may also contribute to differences in grade distribution by countries. As all participants of the present study were from the same country, no such factor had to be considered.

ONEST is a recently introduced method to complement other measures of reproducibility assessment [13,14]. Our analysis suggested that a minimum of 4 observers are needed to adequately reflect reproducibility both for components of grade and the grade itself. In keeping with this figure, our tables reflecting the literature highlight that OPA figures from studies with less than 4 raters are somewhat better than those gained with more observers (Table 5). Studies reporting kappa statistics reflect the same trend (Table 4). Many studies on the reproducibility of grading have used Cohen's kappa, which is devised for 2 observers, [16] therefore pairwise comparisons were made, and the range or average was reported (Table 4), but these basically reflect data derived from 2 observers, which may mirror a better performance than what the ONEST analysis implies. Although Cohen's kappa has a weighted version, where more weight is given to greater deviations, the unweighted and weighted kappas should be very similar for grade, because this has no or just minimum deviation of two steps from majority opinion. Weighted kappa may be more important for the components of grade, especially for mitoses and pleomorphism which had the most deviations. Seldom has a weighted kappa been used to reflect greater deviations [33]. Fleiss has also devised a kappa coefficient for multiple observers [16], and this has been used in several studies with more than two raters (Table 4), and seems more appropriate in this setting.

5. Conclusions

Considering the ONEST analysis, it is suggested that 4 or more observers are needed to reliably assess the reproducibility of grading, and consistently with this finding, previous studies with fewer observers or pairwise comparisons show a somewhat better consistency for grading either on the basis of OPA values or on the basis of Fleiss kappa values. Our results are fitting the results of previous studies with more than 3 observers, and suggest that grading has moderate or moderate to good reproducibility, and this still allows histological grade to be part of multivariable analysis derived combined prognostic tools of breast cancer. Variability in grading needs to be accepted [35], but can be diminished with training, feedback and dedicated assessment.

CRediT authorship contribution statement

BC, GC: Conceptualization. **BC, TN**: Data curation. **GC, BC, TN**: Formal analysis. **BC, RB, EC, OO-N, TP, AS, IS, AV, TZ, GC**: Investigation, Resources. **BC, GC**: Methodology. **GC**: Project administration. **BC**: Software. **BC, GC, TN**: Validation. **BC, GC**: Visualization, Roles/Writing – original draft. **BC, RB, EC, OO-N, TP, AS, IS, AV, TZ, TN, GC**: Writing – review & editing.

Conflict of interest statement

The authors have no conflict of interest to disclose.

Funding

This study received no funding.

Ethical review

This study involved no patient data, no ethical approval was deemed necessary.

Acknowledgements

The publication of this work was supported by the University of Szeged Open Access Fund - Grant No. 5580.

References

[1] R.W. Scarff, H. Torloni. Histological Typing of Breast Tumours, first ed, World Health Organization, Geneva, 1968.
[2] D.H. Patey, R.W. Scarff, The position of histology in the prognosis of carcinoma of the breast, Lancet 211 (5460) (1928) 801–804.
[3] H.J. Bloom, W.W. Richardson, Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years, Br. J. Cancer 11 (1957) 359–377.
[4] C.W. Elston, I.O. Ellis, Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up, Histopathology 19 (1991) 403–410, https://doi.org/10.1111/j.1365-2559.1991.tb00229.x.
[5] I.O. Ellis, S. Al-Sam, N. Anderson, P. Carder, R. Deb, A. Girling, S. Hales, A. Hanby, M. Ibrahim, A.H.S. Lee, R. Liebmann, E. Mallon, S.E. Pinder, E. Provenzano, C. Quinn, E. Rakha, D. Rowlands, T. Stephenson, C.A. Wells, Pathology reporting of breast disease in surgical excision specimens incorporating the dataset for histological reporting of breast cancer; June 2016. ⟨https://www.rcpath.org/uploads/assets/7763be1c-d330–40e8–95d08f955752792a/G148_BreastDataset-hires-Jun16.pdf⟩, 2016. (Accessed 23 October 2021).
[6] I. Amendoeira, N. Apostolikas, J.P. Bellocq, S. Bianchi, W. Boecker, B. Borisch, G. Bussolati, C.E. Connolly, G. Cserni, T. Decker, P. Dervan, M. Drijkoningen, I. O. Ellis, C.W. Elston, V. Eusebi, D. Faverly, P. Heikkila, R. Holland, H. Kerner, J. Kulka, J. Jacquemier, M. Lacerda, J. Martinez-Penuela, C. De Miguel, H. Nordgren, J.L. Peterse, F. Rank, P. Regitnig, A. Reiner, A. Sapino, B. Sigal-Zafrani, A.M. Tanous, S. Thorstenson, E. Zozaya, C.A. Wells, EC Working Group on Breast Screening Pathology, Quality assurance guidelines for pathology, in: N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, L. von Karsa (Eds.), European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis, fourth ed, European Commission, Luxemburg, 2006, pp. 219–311.
[7] WHO Classification of Tumours Editorial Board (Eds.), WHO Classification of Tumours. – Breast Tumours, fifth ed., International Agency for Research on Cancer, Lyon, 2019.
[8] C. Van Dooijeweert, P.J. van Diest, I.O. Ellis, Grading of invasive breast carcinoma: the way forward, Virchows Arch. (2021), https://doi.org/10.1007/s00428-021-03141-2.
[9] J.L. Haybittle, R.W. Blamey, C.W. Elston, J. Johnson, P.J. Doyle, F.C. Campbell, R. I. Nicholson, K. Griffiths, A prognostic index in primary breast cancer, Br. J. Cancer 45 (1982) 361–366, https://doi.org/10.1038/bjc.1982.62.
[10] P.M. Ravdin, L.A. Siminoff, G.J. Davis, M.B. Mercer, J. Hewlett, N. Gerson, H. L. Parker, Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer, J. Clin. Oncol. 19 (2001) 980–991, https://doi.org/10.1200/JCO.2001.19.4.980.
[11] G.C. Wishart, E.M. Azzato, D.C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, P.D. Pharoah, PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer, Breast Cancer Res. 12 (2010) R1, https://doi.org/10.1186/bcr2464.
[12] G. Hortobagyi, J.L. Connolly, C.J. D'Orsi, S.B. Edge, E.A. Mittendorf, H.S. Rugo, L. J. Solin, D.L. Weaver, D.J. Winchester, A.E. Giuliano, Breast, in: M.B. Amin, S. B. Edge, F.L. Greene, D.L. Byrd, R.K. Brookland, M.K. Washington, J. E. Gershenwald, C.C. Compton, K.R. Hess, D.C. Sullivan (Eds.), AJCC Cancer Staging Manual, eight ed, Springer, New York, 2017, pp. 587–628.
[13] E.S. Reisenbichler, G. Han, A. Bellizzi, V. Bossuyt, J. Brock, K. Cole, O. Fadare, O. Hameed, K. Hanley, B.T. Harrison, M.G. Kuba, A. Ly, D. Miller, M. Podoll, A. C. Roden, K. Singh, M.A. Sanders, S. Wei, H. Wen, V. Pelekanou, V. Yaghoobi, F. Ahmed, L. Pusztai, D.L. Rimm, Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer, Mod. Pathol. 33 (2020) 1746–1752, https://doi.org/10.1038/s41379-020-0544-x.
[14] B. Cserni, R. Bori, E. Csörgő, O. Oláh-Németh, T. Pancsa, A. Sejben, I. Sejben, A. Vörös, T. Zombori, T. Nyári, G. Cserni, The additional value of ONEST (Observers Needed to Evaluate Subjective Tests) in assessing reproducibility of oestrogen receptor, progesterone receptor and Ki67 classification in breast cancer, Virchows Arch. (2021), https://doi.org/10.1007/s00428-021-03172-9.
[15] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2016) 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.
[16] J.L. Fleiss. Statistical Methods for Rates and Proportions, second ed, John Wiley and Sons, New York, 1980.
[17] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[18] C. Zaiontz, Real Statistics Resource Pack | Real Statistics Using Excel (⟨https://real-statistics.com⟩). (Accessed 23 October 2021).
[19] J. Jacquemier, C. Charpin, Reproducibility of histoprognostic grades of invasive breast cancer [in French], Ann. Pathol. 18 (1998) 385–390.
[20] M. Sikka, S. Agarwal, A. Bhatia, Interobserver agreement of the Nottingham histologic grading scheme for infiltrating duct carcinoma breast, Indian J. Cancer 36 (1999) 149–153.
[21] J.S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. A. Zehnbauer, K. Lister, Parwaresch, R Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index, Mod. Pathol. 18 (2005) 1067–1078, https://doi.org/10.1038/modpathol.3800388.
[22] P. Robbins, S. Pinder, N. de Klerk, Histological grading of breast carcinomas: a study of interobserver agreement, Hum. Pathol. 26 (1995) 873–879.
[23] T.J. Anderson, F.E. Alexander, J. Lamb, A. Smith, A.P. Forrest, Pathology characteristics that optimize outcome prediction of a breast screening trial, Br. J. Cancer 83 (2000) 487–492, https://doi.org/10.1054/bjoc.2000.1286.
[24] H.F. Frierson Jr, R.A. Wolber, K.W. Berean, D.W. Franquemont, M.J. Gaffey, J. C. Boyd, D.C. Wilbur, Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma, Am. J. Clin. Pathol. 103 (1995) 195–198, https://doi.org/10.1093/ajcp/103.2.195.
[25] K. Rabe, O.L. Snir, V. Bossuyt, M. Harigopal, R. Celli, E.S. Reisenbichler, Interobserver variability in breast carcinoma grading results in prognostic stage differences, Hum. Pathol. 94 (2019) 51–57, https://doi.org/10.1016/j.humpath.2019.09.006.
[26] P.S. Ginter, R. Idress, T.M. D'Alfonso, S. Fineberg, S. Jaffer, A.K. Sattar, A. Chagpar, P. Wilson, M. Harigopal, Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy, Mod. Pathol. 34 (2021) 701–709, https://doi.org/10.1038/s41379-020-00698-2.
[27] E.L. Postma, H.M. Verkooijen, P.J. van Diest, S.M. Willems, M.A. van den Bosch, R. van Hillegersberg, Discrepancy between routine and expert pathologists' assessment of non-palpable breast cancer and its impact on locoregional and systemic treatment, Eur. J. Pharmacol. 717 (2013) 31–35, https://doi.org/10.1016/j.ejphar.2012.12.033.
[28] W. Reed, E. Hannisdal, P.J. Boehler, S. Gundersen, H. Host, J. Marthin, The prognostic value of p53 and c-erb B-2 immunostaining is overrated for patients with lymph node negative breast carcinoma: a multivariate analysis of prognostic factors in 613 patients with a follow-up of 14–30 years, Cancer 88 (2000) 804–813, https://doi.org/10.1002/(sici)1097-0142(20000215)88:4<804::aid-cncr11>3.0.co;2-y.
[29] J.M. Bueno-de-Mesquita, D.S. Nuyten, J. Wesseling, H. van Tinteren, S.C. Linn, M. J. van de Vijver, The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment, Ann. Oncol. 21 (2010) 40–47, https://doi.org/10.1093/annonc/mdp273.
[30] G. Cserni, L. Kocsis, P. Serényi, Grading of invasive breast cancers using the Nottingham modification of the Bloom and Richardson scheme. Study on reproducibility [in Hungarian], Magy. Onkol. 40 (1996) 188–191.
[31] P. Boiesen, P.O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, S. Johansson, O. Ljungberg, A. Ringberg, G. Ostberg, M. Fernö, Histologic grading in breast cancer–reproducibility between seven pathologic departments. South Sweden Breast Cancer Group, Acta Oncol. 39 (2000) 41–45, https://doi.org/10.1080/028418600430950.
[32] T.A. Longacre, M. Ennis, L.A. Quenneville, A.L. Bane, I.J. Bleiweiss, B.A. Carter, E. Catelano, M.R. Hendrickson, H. Hibshoosh, L.J. Layfield, L. Memeo, H. Wu, F. P. O'malley, Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study, Mod. Pathol. 19 (2006) 195–207, https://doi.org/10.1038/modpathol.3800496.
[33] J.P. Sloane, I. Amendoeira, N. Apostolikas, J.P. Bellocq, S. Bianchi, W. Boecker, G. Bussolati, D. Coleman, C.E. Connolly, V. Eusebi, C. De Miguel, P. Dervan, R. Drijkoningen, C.W. Elston, D. Faverly, A. Gad, J. Jacquemier, M. Lacerda, J. Martinez-Penuela, C. Munt, J.L. Peterse, F. Rank, M. Sylvan, V. Tsakraklides, B. Zafrani, Consistency achieved by 23 European pathologists from 12 countries in diagnosing breast disease and reporting prognostic features of carcinomas.

European Commission Working Group on Breast Screening Pathology, Virchows Arch. 434 (1999) 3–10, https://doi.org/10.1007/s004280050297.

[34] I.O. Ellis, D. Coleman, C. Wells, S. Kodikara, E.M. Paish, S. Moss, S. Al-Sam, N. Anderson, L. Bobrow, I. Buley, C.E. Connolly, N.S. Dallimore, S. Hales, A. Hanby, S. Humphreys, F. Knox, J. Lowe, J. Macartney, R. Nash, D. Parham, J. Patnick, S.E. Pinder, C.M. Quinn, A.J. Robertson, J. Shrimankar, R.A. Walker, R. Winder R, Impact of a national external quality assessment scheme for breast pathology in the UK, J. Clin. Pathol. 59 (2006) 138–145, https://doi.org/10.1136/jcp.2004.025551.

[35] E.A. Rakha, R.L. Bennett, D. Coleman, S.E. Pinder, I.O. Ellis, UK National Coordinating Committee for Breast Pathology (EQA Scheme Steering Committee), Review of the national external quality assessment (EQA) scheme for breast pathology in the UK, J. Clin. Pathol. 70 (2017) 51–57, https://doi.org/10.1136/jclinpath-2016-203800.

[36] F. Theissig, K.D. Kunze, G. Haroske, W. Meyer, Histological grading of breast cancer. Interobserver, reproducibility and prognostic significance, Pathol. Res. Pract. 186 (1990) 732–736, https://doi.org/10.1016/S0344-0338(11)80263-3.

[37] L.W. Dalton, D.L. Page, D.W. Dupont, Histologic grading of breast carcinoma. A reproducibility study, Cancer 73 (1994) 2765–2770, 10.1002/1097-0142(19940601)73:11<2765::aid-cncr2820731119>3.0.co;2-k.

[38] C. van Dooijeweert, P.J. van Diest, S.M. Willems, C.C.H.J. Kuijpers, E. van der Wall, L.I.H. Overbeek, I.A.G. Deckers, Significant inter- and intra-laboratory variation in grading of invasive breast cancer: a nationwide study of 33,043 patients in the Netherlands, Int. J. Cancer 146 (2020) 769–780, https://doi.org/10.1002/ijc.32330.

[39] C. van Dooijeweert, P.J. van Diest, I.O. Baas, E. van der Wall, I.A. Deckers, Variation in breast cancer grading: the effect of creating awareness through laboratory-specific and pathologist-specific feedback reports in 16 734 patients with breast cancer, J. Clin. Pathol. 73 (2020) 793–799, https://doi.org/10.1136/jclinpath-2019-206362.

[40] C. van Dooijeweert, I.A.G. Deckers, E.J. de Ruiter, N.D. Ter Hoeve, C.P.H. Vreuls, E. van der Wall, P.J. van Diest, The effect of an e-learning module on grading variation of (pre)malignant breast lesions, Mod. Pathol. 33 (2020) 1961–1967, https://doi.org/10.1038/s41379-020-0556-6.

[41] V. Sopik, S.A. Narod, The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer, Breast Cancer Res. Treat. 170 (2018) 647–656, https://doi.org/10.1007/s10549-018-4796-9.

[42] G.C. Wishart, E.M. Azzato, D.C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, P.D. Pharoah, PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer, Breast Cancer Res. 12 (2010) R1, https://doi.org/10.1186/bcr2464.

[43] S.J. Dawson, S.W. Duffy, F.M. Blows, K.E. Driver, E. Provenzano, J. LeQuesne, D. C. Greenberg, P. Pharoah, C. Caldas, G.C. Wishart, Molecular characteristics of screen-detected vs symptomatic breast cancers and their impact on survival, Br. J. Cancer 101 (2009) 1338–1344, https://doi.org/10.1038/sj.bjc.6605317.