cancers

MDPI

Article

ONEST (Observers Needed to Evaluate Subjective Tests) Analysis of Stromal Tumour-Infiltrating Lymphocytes (sTILs) in Breast Cancer and Its Limitations

Bálint Cserni [1], Darren Kilmartin [2], Mark O'Loughlin [2], Xavier Andreu [3], Zsuzsanna Bagó-Horváth [4], Simonetta Bianchi [5], Ewa Chmielik [6], Paulo Figueiredo [7], Giuseppe Floris [8], Maria Pia Foschini [9], Anikó Kovács [10], Päivi Heikkilä [11], Janina Kulka [12], Anne-Vibeke Laenkholm [13], Inta Liepniece-Karele [14], Caterina Marchiò [15,16], Elena Provenzano [17,18], Peter Regitnig [19], Angelika Reiner [20], Aleš Ryška [21], Anna Sapino [15,16], Elisabeth Specht Stovgaard [22], Cecily Quinn [23,24], Vasiliki Zolota [25], Mark Webber [2], Sharon A. Glynn [2], Rita Bori [26], Erika Csörgő [26], Orsolya Oláh-Németh [27], Tamás Pancsa [27], Anita Sejben [27], István Sejben [26], András Vörös [27], Tamás Zombori [27], Tibor Nyári [28], Grace Callagy [2] and Gábor Cserni [26,27,*]

1 TNG Technology Consulting GmbH, Király u. 26., 1061 Budapest, Hungary
2 Discipline of Pathology, Lambe Institute for Translational Research, School of Medicine, University of Galway, H91 TK33 Galway, Ireland
3 Pathology Department, Atryshealth Co., Ltd., 08039 Barcelona, Spain
4 Department of Pathology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria
5 Division of Pathological Anatomy, Department of Health Sciences, University of Florence, 50134 Florence, Italy
6 Tumor Pathology Department, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, 44-102 Gliwice, Poland
7 Laboratório de Anatomia Patológica, IPO Coimbra, 3000-075 Coimbra, Portugal
8 Laboratory of Translational Cell & Tissue Research and KU Leuven, Department of Imaging and Pathology, Department of Pathology, University Hospitals Leuven, University of Leuven, Oude Market 13, 3000 Leuven, Belgium
9 Unit of Anatomic Pathology, Department of Biomedical and Neuromotor Sciences, University of Bologna, Bellaria Hospital, 40139 Bologna, Italy
10 Department of Clinical Pathology, Sahlgrenska University Hospital, 41345 Gothenburg, Sweden
11 Department of Pathology, Helsinki University Central Hospital, 00029 Helsinki, Finland
12 Department of Pathology, Forensic and Insurance Medicine, Semmelweis University Budapest, Üllői út 93, 1091 Budapest, Hungary
13 Department of Surgical Pathology, Zealand University Hospital, 4000 Roskilde, Denmark
14 Department of Pathology, Riga Stradins University, Riga East Clinical University Hospital, LV-1038 Riga, Latvia
15 Unit of Pathology, Candiolo Cancer Institute FPO-IRCCS, 10060 Candiolo, Italy
16 Department of Medical Sciences, University of Turin, 10126 Turin, Italy
17 Department of Histopathology, Cambridge University Hospitals National Health Service (NHS) Foundation Trust, Cambridge CB2 0QQ, UK
18 National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge CB2 0QQ, UK
19 Diagnostic and Research Institute of Pathology, Medical University of Graz, 8010 Graz, Austria
20 Department of Pathology, Klinikum Donaustadt, 1090 Vienna, Austria
21 The Fingerland Department of Pathology, Charles University Medical Faculty and University Hospital, 50003 Hradec Kralove, Czech Republic
22 Pathology Department, Herlev University Hospital, DK-2730 Herlev, Denmark
23 Department of Histopathology, Irish National Breast Screening Programme, BreastCheck, St. Vincent's University Hospital and School of Medicine, University College Dublin, D04 T6F4 Dublin, Ireland
24 School of Medicine, University College Dublin, D04 V1W8 Dublin, Ireland
25 Department of Pathology, School of Medicine, University of Patras, 26504 Rion, Greece
26 Department of Pathology, Bács-Kiskun County Teaching Hospital, 6000 Kecskemét, Hungary
27 Department of Pathology, University of Szeged, 6720 Szeged, Hungary
28 Department of Medical Physics and Informatics, University of Szeged, 6720 Szeged, Hungary
* Correspondence: csernig@kmk.hu

check for updates

Citation: Cserni, B.; Kilmartin, D.; O'Loughlin, M.; Andreu, X.; Bagó-Horváth, Z.; Bianchi, S.; Chmielik, E.; Figueiredo, P.; Floris, G.; Foschini, M.P.; et al. ONEST (Observers Needed to Evaluate Subjective Tests) Analysis of Stromal Tumour-Infiltrating Lymphocytes (sTILs) in Breast Cancer and Its Limitations. *Cancers* **2023**, *15*, 1199. https://doi.org/10.3390/cancers15041199

Academic Editors: Enrico Cassano and Filippo Pesapane

Received: 20 January 2023
Revised: 4 February 2023
Accepted: 9 February 2023
Published: 14 February 2023

Simple Summary: Tumour-infiltrating lymphocytes (TILs) reflect the host's response against tumours. TILs have a strong prognostic effect in the so-called triple-negative (oestrogen receptor, progesterone receptor, and human epidermal growth factor receptor-2 negative) subset of breast

cancers and predict a better response when primary systemic (neoadjuvant) treatment is administered. Although they are easy to assess, their quantitative assessment is subject to some inter-observer variation. ONEST (Observers Needed to Evaluate Subjective Tests) is a new way of analysing inter-observer variability and helps in estimating the number of observers required for a more reliable estimation of this phenomenon. This aspect of reproducibility for TILs has not been explored previously. Our analysis suggests that between six and nine pathologists can give a good approximation of inter-observer agreement in TIL assessments.

Abstract: Tumour-infiltrating lymphocytes (TILs) reflect antitumour immunity. Their evaluation of histopathology specimens is influenced by several factors and is subject to issues of reproducibility. ONEST (Observers Needed to Evaluate Subjective Tests) helps in determining the number of observers that would be sufficient for the reliable estimation of inter-observer agreement of TIL categorisation. This has not been explored previously in relation to TILs. ONEST analyses, using an open-source software developed by the first author, were performed on TIL quantification in breast cancers taken from two previous studies. These were one reproducibility study involving 49 breast cancers, 23 in the first circulation and 14 pathologists in the second circulation, and one study involving 100 cases and 9 pathologists. In addition to the estimates of the number of observers required, other factors influencing the results of ONEST were examined. The analyses reveal that between six and nine observers (range 2–11) are most commonly needed to give a robust estimate of reproducibility. In addition, the number and experience of observers, the distribution of values around or away from the extremes, and outliers in the classification also influence the results. Due to the simplicity and the potentially relevant information it may give, we propose ONEST to be a part of new reproducibility analyses.

Keywords: ONEST; observers needed to evaluate subjective tests; TILs; sTILs; tumour-infiltrating lymphocytes; triple-negative; breast cancer; reproducibility; international immuno-oncology biomarker working group; European Working Group for Breast Screening Pathology

1. Introduction

Tumour-infiltrating lymphocytes (TILs) are a reflection of antitumour immunity. Different compartments and populations are recognised; for breast carcinomas, stromal lymphocytes have been accepted as the most practically assessable compartment of TILs, and their quantity correlates with that of intra-epithelial TILs [1]. On the basis of meta-analyses, stromal TILs (sTILs) have been proven to be predictive of the response to neoadjuvant chemotherapy [2] and to be associated with better prognosis after adjuvant treatment of triple-negative breast carcinomas (TNBCs) [3]. TILs have also been linked to the rare phenomenon of spontaneous regression in TNBC [4]. The accumulated data on the value of TILs have matured enough to recommend this biomarker for implementation in daily routine [5].

However, there are a number of other events (e.g., necrosis or previous biopsy) that lead to the accumulation of inflammatory cells, and these have been taken into consideration when defining the rules for quantifying the amount of sTILs relevant for antitumour immunity. This has led to the formulation of guidelines recommending that sTILs should be evaluated as the average proportion of the stromal area occupied by TILs, including both lymphocytes and plasma cells. In the assessment, the total stromal area excludes areas of regressive hyalinisation, necrosis, and previous needle biopsy sites. Mononuclear cells around in situ carcinoma and normal structures should also be excluded, and all estimations should be restricted to the tumour area [6]. A later addendum suggested that the invasive front (1 mm at the edge of the tumour) should also be included [7]. The human brain tries to simplify things; therefore, the rules for quantifying sTILs predispose this biomarker to being poorly reproducible. Nevertheless, good reproducibility was docu-

mented by the International Immuno-oncology Biomarker Working Group (IIOBMWG) after the introduction of a direct online feedback software helping in the calibration of sTIL percentages in pre-selected fields of view (FOVs) [8].

Members of the European Working Group for Breast Screening Pathology (EWGBSP) have also assessed the reproducibility of scoring sTILs on digitised needle core biopsy specimens using the same performance-improving online tool that was used for training by Denkert et al. [8,9] and found moderate reproducibility for biopsy specimens (intraclass correlation coefficient, ICC 0.634, 95% CI 0.539–0.735) but good reproducibility for selected triplets of FOVs (ICC 0.798, 95% CI 0.727–0.864) [10]. In the present work, we use the same data to perform an ONEST (Observers Needed to Evaluate Subjective Tests) analysis of sTILs.

ONEST is a recently developed method that complements inter-observer agreement studies by helping to estimate the number of observers required for a reliable estimation of reproducibility [11]. ONEST uses 100 randomly selected permutations of all participating pathologists (observers or raters) and plots the overall percent agreement (OPA) values for an increasing number of observers, looking for the worst (lowest) curve to reach a plateau, beyond which an increasing number of observers does not have a substantial effect on agreement [11–13]. Additionally, ONEST has been recognised to be valuable as a visual complement to demonstrate the degree of reproducibility of subjectively evaluated parameters such as oestrogen receptor (ER) quantification, Ki-67 labelling, or histological grade, as well as the difference between observers and how these compare to the overall percent agreement (OPA) of all observers [12,13]. The aim of this study is to evaluate sTIL quantification using ONEST and to estimate the number of observers needed for a reliable evaluation of its reproducibility. The relatively large number of observers in our previous study [9] allows for a better evaluation of ONEST itself as a method.

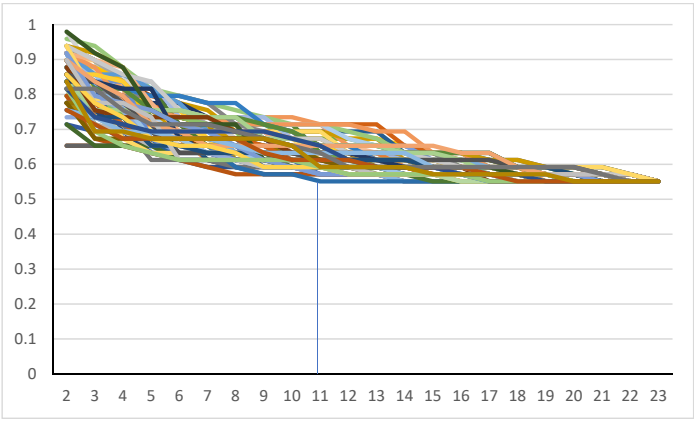
2. Materials and Methods

We used anonymised results from the EWGBSP analysis of reproducibility [9]. In that study, 23 pathologists assessed 49 core needle biopsies from TNBCs in circulation 1 (C1), and 14 pathologists, as a subset, assessed both C1 (this subset of C1 denoted as C1s) in addition to 3 pre-selected digital FOVs of the same 49 cases with different labels to prevent comparisons (C2). The corresponding author of this previous study (Grace Callagy) has released the sTIL percentage values reported by the 23 and 14 participants for each case in a tabulated format, with rows representing cases and columns representing one or the other observer, and these values were used for the ONEST analyses of C1 and C2, respectively. There were 2 missing values in all circulations (C1, C1s, and C2) which were replaced by mean sTIL percentages rounded to the closest integer. For the ONEST analysis, as per the introduction of the method and its subsequent uses [11–13], 100 randomly selected permutations were selected for the values of the ONEST plots. Four selected cut-offs were used to define categories: <60% vs. ≥60%, e.g., [14], and <50% vs. ≥50%, e.g., [15,16], to match two different definitions of lymphocyte-predominant breast cancers, which are the likeliest responders to neoadjuvant treatment [6]; <30% vs. ≥30% to match a cut-off proposed for a strong prognostic role in the adjuvant setting [3]; and 0–20%, 21–49%, and ≥50% to match a three-tiered classification used in the IIOBWG ring studies [8].

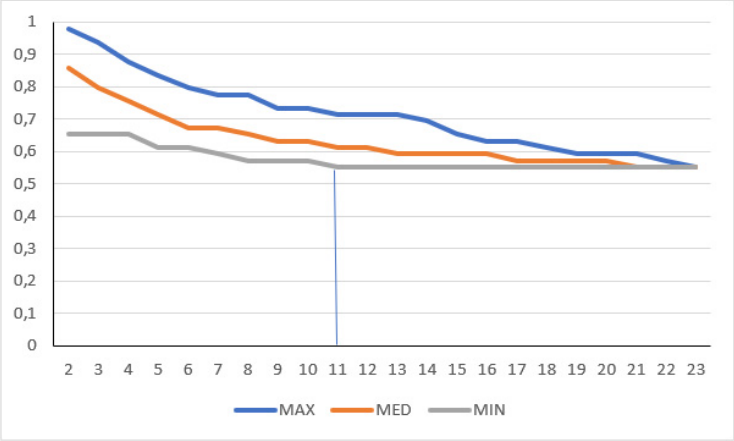
In a previous study, 9 pathologists assessed the ER, the progesterone receptor (PR) status, Ki67 labelling [12], and histological grade [13] of breast cancers in 50 core needle biopsies and 50 resection specimens represented on a full-face glass slide for each case. While assessing these parameters, the participants were also asked to document sTILs based on the IIOBMWG recommendations [6,7], which are also part of the Hungarian recommendation [17,18]. These results have never been analysed previously and were also used for a separate ONEST analysis as circulation 3 (C3).

A full ONEST plot includes all OPA values per increasing number of observers for the 100 randomly obtained permutations of observers, i.e., it represents 100 OPA curves (OPACs), each representing the OPA values of a given permutation (Figure 1A). We also

introduced a simplified ONEST plot, which includes only the maximum OPA values (maximum curve—best scenario), the minimum OPA values (minimum curve—worst scenario), and a median value curve. The maximum and minimum curves do not necessarily represent an OPAC from the 100 randomly selected permutations, but they obviously coincide with an OPAC from all possible permutations. Figure 1A and 1B compare the full and simplified ONEST plots of the same entity studied. The ONEST value is the integer from axis *x* (the number of pathologists), which reflects the minimum curve OPA value beyond which there is no more relevant decrease in OPA values with further increase in observers. Bandwidth is defined as the difference between the highest and lowest OPA values with 2 pathologists assessing sTILs, i.e., this is the difference in OPA of the maximum and minimum curves with 2 observers. Finally, OPA(n) is the OPA value for all observers, the percentage of cases upon which all assessing observers agree. Good reproducibility implies a high OPA(n), a low ONEST value, and narrow bandwidth, whereas the opposite is true for poor reproducibility. The worst scenario is when OPA(n) = 0, i.e., there are no cases on which all observers agree. This latter scenario is unacceptable for biomarker studies or subjective tests on relevant issues in general and should be remedied by improving reproducibility or dropping the test and substituting it with a better one. An open-source software designed by the first author for randomly selecting 100 permutations from all possible ones and making a basic ONEST analysis is available at github.com (accessed on 12 November 2022) [19].

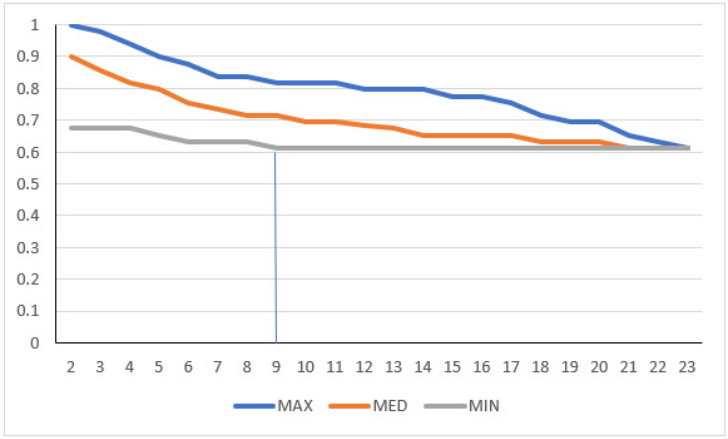


(**a**)

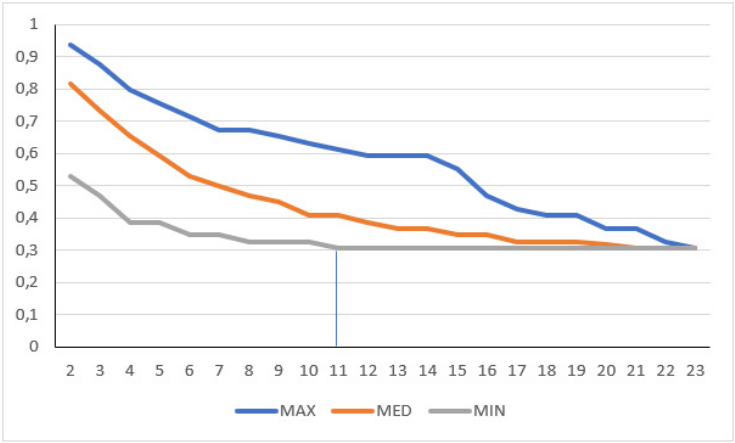


(**b**)

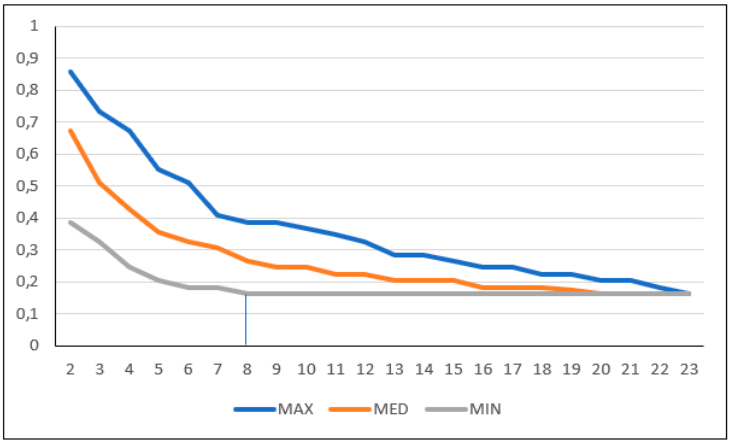
Figure 1. *Cont.*



(**c**)



(**d**)



(**e**)

Figure 1. ONEST plots of different cut-off values for 23 pathologists. (**a**) Full and (**b**) simplified ONEST plots for the 49 cases assessed by 23 pathologists for a cut-off of <50% vs. ≥50% sTILs. (**c**–**e**) Simplified ONEST plots for further cut-off values studied: (**c**) <60% vs. ≥60%; (**d**) <30% vs. ≥30%; and (**e**) <20%, 21–50%, >50%. Readings from the plots are included in Table 1. OPA (n = 23) values are the OPA values at the right side of the plots and reflect the proportion of cases with full agreement. ONEST values correspond to the number of observers on the *x*-axis, where the minimum curve levels off, and no substantial decrease is noted with further increase in the number of observers (this is highlighted by vertical segments between the *x*-axis value and the minimum curve). The bandwidth of the ONEST plot is visualised on the left side of the plot as the difference between the maximum and the minimum curves with 2 observers; this is the largest difference in agreement between two observers.

Table 1. ONEST analyses of different circulations and cut-off values of sTILs.

<50% vs. ≥50%	**C1**	**C1 without Divergent Raters 7 and 20**	**C1s**	**C2**	**C2 without Divergent Raters 4 and 13**	**C3**
n	23	21	14	14	12	9
OPA(n)	0.551	0.612	0.571	0.776	0.816	0.89
Bandwidth	0.327	0.245	0.265	0.184	0.143	0.07
ONEST	11	7	8	6	3	6
<60% vs. ≥60%						
n	23	21	14	14	12	9
OPA(n)	0.612	0.796	0.612	0.796	0.837	0.91
Bandwidth	0.327	0.286	0.612	0.163	0.163	0.07
ONEST	9	7	4	6	2	2
<30% vs. ≥30%						
n	23	21	14	14	12	9
OPA(n)	0.306	0.347	0.327	0.551	0.592	0.81
Bandwidth	0.408	0.306	0.306	0.306	0.204	0.09
ONEST	11	8	9	8	7	6
≤20%, 21–49%, ≥50%						
n	23	21	14	14	12	9
OPA(n)	0.163	0.204	0.408	0.408	0.449	0.74
Bandwidth	0.469	0.388	0.143	0.265	0.245	0.12
ONEST	8	7	7	5	6	6

C1: circulation 1 with 23 pathologists and 49 digital slides of core needle biopsy samples; C1s: subset of C1 with the 14 pathologists taking part in C2; C2: circulation 2 with 14 pathologists and 3 preselected fields of view of the 49 cases viewed in C1; C3: circulation 3 is independent from C1 and C2 and involves 9 pathologists assessing 100 cases, half from core needle biopsies and half from excision specimens. For further details, see the Materials and Methods section.

For the analysis with a cut-off value of <50% vs. ≥50% sTILs, ONEST analyses were repeated 3 times (3 random selections of 100 permutations in which the chances of identical permutations are practically nil), and the minimum curves obtained were compared by means of the Kruskal–Wallis test. In the original series, two pathologists (numbers 7 and 20) substantially diverged in their opinions from the rest of the group in C1, whereas two pathologists (numbers 4 and 13) diverged from others in C2. To test the influence of these divergently classifying pathologists, 3 and 3 ONEST plots for the same cut-off values (<50% vs. ≥50%) were also generated after the removal of results by these observers, and the ONEST values were determined from all plots. The ONEST values obtained with or without the deviant classifiers were compared by means of the two sample Wilcoxon rank sum test. Statistical analyses were performed in Excel with the Real Statistics Add-Ins [20] and STATA Software version 17.0 (StataCorp LP, College Station, TX, USA).

3. Results

The results of the C1 were selected to be represented by the ONEST plots in Figure 1. Readings of this and other ONEST analyses from C2 and C3 are represented in Table 1. With different approaches, pathologists, and numbers of pathologists, the ONEST values varied between 2 and 11 (Table 1). There were two pathologists, in both circulations C1 and C2, who substantially deviated from the overall average ratings; separate ONEST analyses were also performed without these participants. Not surprisingly, not only did the OPA(n) values increase, but the bandwidth became smaller, and the ONEST values decreased. With the exception of the C1 (n = 23 pathologists) for the <50% vs. ≥50% and the <30% vs. ≥30% categorisations, the ONEST values were not greater than 9; the number of pathologists involved in the C3 yielded the best OPA(n) values, i.e., the best reproducibility (Table 1).

One of the sTIL categorisations was used to test the ONEST plots. Three random selections of 100 permutations were compared with the Kruskal–Wallis test for the chosen (<50% vs. ≥50%) sTIL categorisation for C1, C1 without the two substantially divergent raters, C1s, C2, and C2 without the two substantially divergent raters. Although sometimes there was a small shift in the ONEST and other values, these permutations were not statistically different with regard to the minimum curves; their p-values were 0.937, 0.271, 0.877, 0.855, and 1, respectively (Figure 2).

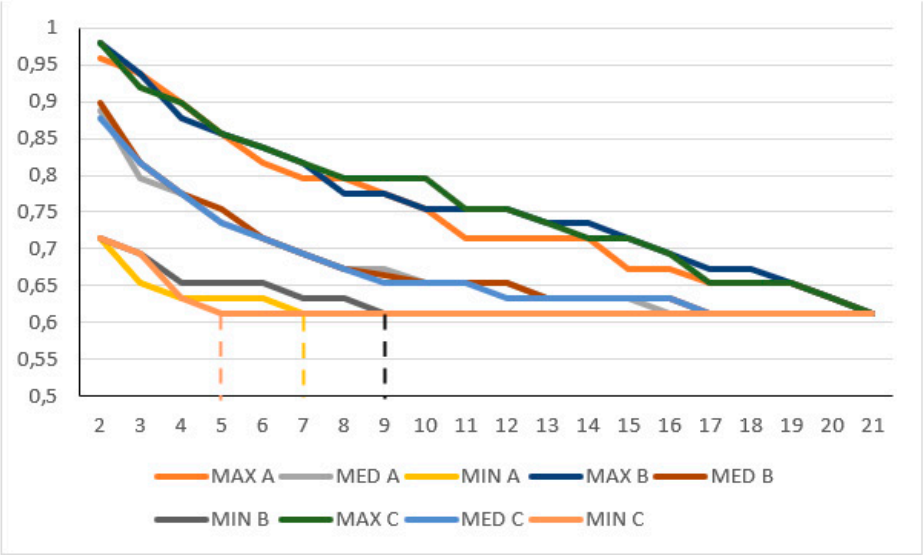


Figure 2. Partly overlapping simplified ONEST plots of 3 randomly selected 100 permutations (A, B, and C) for the <50% sTIL or more classification in C1 circulation without the two divergent classifiers; this example showed the lowest *p*-value in the Kruskal–Wallis test. Note: the *y*-axis only represents values between 0.5 and 1; despite not being statistically significantly different, the 3 randomly selected ONEST plots of 100 permutations yield 3 different ONEST values: 7 (A), 9 (B), and 5 (C) (to ease reading of the values, these are highlighted by vertical dashed segments between the *x*-axis value and the minimum curves), whereas the bandwidth is very similar (0.245 A, 0.265 B and C), and by definition, the OPA(21) value is identical (0.612). MAX: maximum curve; MED: median curve; MIN: minimum curve.

Furthermore, the three random permutations from C1 and C1 without the outlying classifiers; C2 and C2 without the outlying classifiers; and finally, C1 and C2, were also compared with the Wilcoxon rank sum test for the ONEST values that could be derived from them, and this demonstrated significant differences ($p = 0.046$, $p = 0.034$, and $p = 0.043$, respectively) for each of these comparisons.

4. Discussion

ONEST is a recently described additional analysis that can complement reproducibility studies [11–13]. Although it was introduced to estimate the minimum number of observers required to provide a reliable estimate of the reproducibility of a given classification [11], it also gives a visual impression of how much agreement is reached when categorising items into predefined classes and the difference one can expect between two observers. However, as a complementary tool, ONEST is not independent of the studied "population" and the observers.

It is generally accepted that two-tiered classifications are more reproducible than those with more than two categories, e.g., [21]. This also applies to ONEST, as reported for PD-L1 [9] and Ki67 [10], and this is also supported by our analysis of the three-tiered classification in the present study, which demonstrated the worst OPA(n) values in nearly all circulations (Table 1).

Although our attempt to analyse the data without the two observers who substantially deviated from the majority opinion resulted in "improved" results in both C1 and C2 (i.e., greater OPA(n), narrower bandwidth, and lower ONEST values), the analyses without these outliers may not reflect real-life assessments. It is well accepted that populations are generally described with their average values of measurable things, but they also have members that are above and below the average. Therefore, if one wishes to estimate the real-life performance of a classification, all raters, and not only the best raters, should be included in the analysis.

Reproducibility is also dependent on the distribution of the parameter being evaluated in the cases. While assessing three nuclear immunostains for ER, PR, and Ki67 in a different study, we found that using the same cut-off values for all three biomarkers resulted in different reproducibility and ONEST estimations [12]. This was explained by the difference in the number of cases close to or away from the extreme values (0% and 100%). Most values for ER staining were in the 90–100% or 0% range, whereas PR values showed more divergence, Ki67 scores were distributed over a wider range, and ONEST values increased in a respective manner. This phenomenon is likely to be the most important contributor to the surprisingly good results observed for the C3 circulation in the present study (Table 1). Indeed, in C3, there were only 45/900 ratings for sTILs $\geq$ 50% involving 8/100 cases.

The homogeneity versus heterogeneity of the entity being observed also influences reproducibility, and this is substantiated by earlier studies. ER staining is generally more homogeneous than Ki67, as reflected in the lower inter-observer agreement for the latter [12]. On the other hand, sTILs often have a heterogeneous distribution, making it more difficult to assess the overall average distribution. This phenomenon, i.e., heterogenous distribution, was identified as the main contributor to the weaker reproducibility for some cases in our previous study [9] and was also reported by others [22]. Scoring preselected FOVs (C2) eliminates the variability associated with the observers selecting the areas to score in the case of heterogeneously distributed sTILs and results in substantially better reproducibility (ICC for absolute sTILs with preselected FOVs vs. the case when observers had selected their FOVs to be assessed: 0.798 vs. 0.634) [9]. This improved reproducibility was also reflected by key values of ONEST plot analyses: higher OPA(n) values, lower bandwidth, and lower ONEST values in C2 vs. C1 for all categorical classifications.

The number of observers may also influence reproducibility and ONEST plots. For example, C2 versus C1, without the discordant raters (with 12 observers of the former all included in the 21 of the latter), resulted in different OPA(n) values (82% vs. 61% agreement for, e.g., sTILs $\geq$ 50% or fewer). The number of observers also greatly impacts the number of possible permutations, being 2.585×10^{22} for C1 (n = 23), 87,178,291,200 for C2 (n = 14), and "only" 362,880 for C3 (n = 9). In a previous study, also with nine observers [12], we verified that the minimum curve of the 100 randomly selected permutations does not significantly differ from the minimum OPAC of all permutations. In the present analysis, three random ONEST plots were examined for all circulations with one of the cut-offs (<50% vs. the $\geq$50%), and no significant difference was found between their minimum curves. This is also reflected in Figure 2, in which the minimum (and the maximum and median) curves of the three plots substantially overlap with each other. Despite this, there were minor alterations in the bandwidths and ONEST values from the three analyses of the same datasets. This leads us to conclude that even ONEST readings are just estimations and might have a range, but depending on how close the ONEST value is to 2, we can estimate how a reproducibility study with a low number of participants may reflect real-life performance for the test in question. An early study of TILs with 99 cases suggested an 85% (95% CI, 76% to 91%) agreement with no more than a 10% difference in absolute sTIL ratings between two observers [23]. Kojima and colleagues reported an 81% agreement between two observers when classifying sTILs into three categories in 129 cases [24]. A report on 100 cases and >90% mean pairwise agreement on sTILs, by any of six pathologists, with a seventh pathologist serving as the main reviewer for a study, also suggests excellent reproducibility [25]. However, Figure 1A clearly shows that two observers randomly

selected from a pool of observers or pairwise comparisons may have minimal discrepancies or no discrepancy at all, but the bandwidth may be much wider than this. Four pathologists also achieved a good agreement scoring sTILs in 121 cases [26] and substantial agreement in 75 cases [27], but Table 1 suggests that this number is still prone to underestimating real-life conditions. Certainly, two observers [23,24,28,29] do not accurately reflect inter-observer agreement [11], and most readings from the ONEST plots (Table 1) with a different number or quality of readers suggest that between 6 and 11 readers are required for a reasonable estimation of inter-observer agreement.

As a limitation, ONEST analyses can only be performed for categorical classifications. Agreement for scoring some markers (e.g., sTILs) as a continuous variable is generally better than the agreement observed using categories defined by given cut-off values [30]. On the other hand, therapeutic decisions are generally made using cut-off values for a biomarker.

Finally, after considering the factors influencing the reproducibility of a subjective test, such as scoring sTILs in breast cancer, it is the case that other variables (e.g., number and experience of observers, distribution of the cases around or away from the extremes, and heterogeneity between fields to assess) also influence ONEST analyses and the ONEST values. Therefore, we can state that two to four observers are certainly not sufficient to reflect the actual inter-observer agreement for evaluating sTILs in breast cancer, but between 6 and 11 observers would be sufficient. The studies by the IIOBMWG largely fulfil this requirement, and their reported values of good reproducibility should be considered reliable [8]. Notwithstanding, the finding that our group, also with a sufficient number of pathologists, was only able to match their high ICC values when scoring sTILs on preselected FOVs, but not when full digital slides were scored, clearly means that factors other than the number of observers contribute to reproducibility [11]. This is also substantiated by another study involving 41 cases of digitised core needle biopsies scored by 40 pathologists, where the ICC values ranged between −0.376 and 0.947, with a mean of 0.659 [31]. In addition to applying methods such as ONEST, the development of tools that can quantify other contributors to lower reproducibility will be useful in the design of reproducibility studies. Due to its simplicity and the data it gives, we also propose that an ONEST analysis could be a part of reproducibility studies to explore the reliability of the results presented or published previously, as not all reports satisfy the suggested minimum number of observers to reach the best possible conclusions. However, the limitations described in the present article must be kept in mind.

5. Conclusions

The reproducibility of sTIL assessments in breast cancer has been examined in several studies. Our results using ONEST indicate that between six and nine observers are expected to give a good estimate of inter-observer variability, and studies involving fewer than these numbers may overestimate agreement between observers. As sTIL evaluation becomes part of daily practice [5], efforts to characterise factors interfering with the reproducibility of scoring are welcome.

Author Contributions: Conceptualisation, B.C. and G.C. (Gábor Cserni); methodology, B.C., G.C. (Gábor Cserni), and T.N.; scoring cases (investigation), X.A., R.B., Z.B.-H., S.B., E.C. (Erika Csörgő), E.C. (Ewa Chmielik), G.C. (Grace Callagy), G.C. (Gábor Cserni), P.F., G.F., M.P.F., A.K., P.H., J.K., A.-V.L., I.L.-K., C.M., O.O.-N., T.P., E.P., P.R., A.R. (Aleš Ryška), A.R. (Angelika Reiner), A.S. (Anna Sapino), A.S. (Anita Sejben), E.S.S., I.S., C.Q., A.V., V.Z. and T.Z.; formal analysis and data curation, B.C., G.C. (Gábor Cserni), T.N.; D.K., M.O. and S.A.G.; digitisation, case distribution, M.W.; writing—original draft preparation, B.C. and G.C. (Gábor Cserni); writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki; it involved no patient data, and no ethical approval was deemed necessary.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the corresponding author and will be released upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. El Bairi, K.; Haynes, H.R.; Blackley, E.; Fineberg, S.; Shear, J.; Turner, S.; de Freitas, J.R.; Sur, D.; Amendola, L.C.; Gharib, M.; et al. The tale of TILs in breast cancer: A report from The International Immuno-Oncology Biomarker Working Group. *N.P.J. Breast Cancer.* **2021**, *7*, 150. [CrossRef] [PubMed]
2. Denkert, C.; von Minckwitz, G.; Darb-Esfahani, S.; Lederer, B.; Heppner, B.I.; Weber, K.E.; Budczies, J.; Huober, J.; Klauschen, F.; Furlanetto, J.; et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **2018**, *19*, 40–50. [CrossRef]
3. Loi, S.; Drubay, D.; Adams, S.; Pruneri, G.; Francis, P.A.; Lacroix-Triki, M.; Joensuu, H.; Dieci, M.V.; Badve, S.; Demaria, S.; et al. Tumor-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers. *J. Clin. Oncol.* **2019**, *37*, 559–569. [CrossRef] [PubMed]
4. Cserni, G.; Serfőző, O.; Ambrózay, É.; Markó, L.; Krenács, L. Spontaneous pathological complete regression of a high grade triple negative breast cancer with axillary metastasis–report of a case. *Pol. J. Pathol.* **2019**, *70*, 139–143. [CrossRef] [PubMed]
5. Laenkholm, A.V.; Callagy, G.; Balancin, M.; Bartlett, J.M.S.; Sotiriou, C.; Marchio, C.; Kok, M.; Dos Anjos, C.H.; Salgado, R. Incorporation of TILs in daily breast cancer care: How much evidence can we bear? *Virchows Arch.* **2022**, *480*, 147–162. [CrossRef] [PubMed]
6. Salgado, R.; Denkert, C.; Demaria, S.; Sirtaine, N.; Klauschen, F.; Pruneri, G.; Wienert, S.; Van den Eynden, G.; Baehner, F.L.; Penault-Llorca, F.; et al. The Evaluation of Tumor-Infiltrating Lymphocytes (TILs) in Breast Cancer: Recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **2015**, *26*, 259–271. [CrossRef]
7. Dieci, M.V.; Radosevic-Robin, N.; Fineberg, S.; van den Eynden, G.; Ternes, N.; Penault-Llorca, F.; Pruneri, G.; D'Alfonso, T.M.; Demaria, S.; Castaneda, C.; et al. Update on Tumor-Infiltrating Lymphocytes (TILs) in Breast Cancer, Including Recommendations to Assess TILs in Residual Disease After Neoadjuvant Therapy and in Carcinoma In Situ: A Report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *Semin. Cancer Biol.* **2018**, *52*, 16–25.
8. Denkert, C.; Wienert, S.; Poterie, A.; Loibl, S.; Budczies, J.; Badve, S.; Bago-Horvath, Z.; Bane, A.; Bedri, S.; Brock, J.; et al. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: Results of the ring studies of the international immuno-oncology biomarker working group. *Mod. Pathol.* **2016**, *29*, 1155–1164. [CrossRef]
9. Kilmartin, D.; O'Loughlin, M.; Andreu, X.; Bagó-Horváth, Z.; Bianchi, S.; Chmielik, E.; Cserni, G.; Figueiredo, P.; Floris, G.; Foschini, M.P.; et al. Intra-Tumour Heterogeneity Is One of the Main Sources of Inter-Observer Variation in Scoring Stromal Tumour Infiltrating Lymphocytes in Triple Negative Breast Cancer. *Cancers* **2021**, *13*, 4410. [CrossRef]
10. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
11. Reisenbichler, E.S.; Han, G.; Bellizzi, A.; Bossuyt, V.; Brock, J.; Cole, K.; Fadare, O.; Hameed, O.; Hanley, K.; Harrison, B.T.; et al. Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer. *Mod. Pathol.* **2020**, *33*, 1746–1752. [CrossRef] [PubMed]
12. Cserni, B.; Bori, R.; Csörgő, E.; Oláh-Németh, O.; Pancsa, T.; Sejben, A.; Sejben, I.; Vörös, A.; Zombori, T.; Nyári, T.; et al. 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**, *479*, 1101–1109. [CrossRef]
13. Cserni, B.; Bori, R.; Csörgő, E.; Oláh-Németh, O.; Pancsa, T.; Sejben, A.; Sejben, I.; Vörös, A.; Zombori, T.; Nyári, T.; et al. 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. *Pathol. Res. Pract.* **2021**, *229*, 153718. [CrossRef]
14. Stanton, S.E.; Disis, M.L. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J. Immunother. Cancer* **2016**, *4*, 59. [CrossRef] [PubMed]
15. Loi, S.; Michiels, S.; Salgado, R.; Sirtaine, N.; Jose, V.; Fumagalli, D.; Kellokumpu-Lehtinen, P.L.; Bono, P.; Kataja, V.; Desmedt, C.; et al. 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.* **2014**, *25*, 1544–1550. [CrossRef]
16. Sasaki, R.; Horimoto, Y.; Yanai, Y.; Kurisaki-Arakawa, A.; Arakawa, A.; Nakai, K.; Saito, M.; Saito, T. Molecular Characteristics of Lymphocyte-predominant Triple-negative Breast Cancer. *Anticancer Res.* **2021**, *41*, 2133–2140. [CrossRef] [PubMed]
17. Cserni, G.; Francz, M.; Járay, B.; Kálmán, E.; Kovács, I.; Krenács, T.; Udvarhelyi, N.; Tóth, E.; Vass, L.; Vörös, A.; et al. Pathological diagnosis, work-up and reporting of breast cancer. Recommendations from the 4th Breast Cancer Consensus Conference. *Magy. Onkol.* **2020**, *64*, 301–328. (In Hungarian) [PubMed]
18. Cserni, G.; Francz, M.; Járay, B.; Kálmán, E.; Kovács, I.; Krenács, T.; Udvarhelyi, N.; Tóth, E.; Vass, L.; Vörös, A.; et al. Pathological Diagnosis, Work-Up and Reporting of Breast Cancer 1st Central-Eastern European Professional Consensus Statement on Breast. *Cancer. Pathol. Oncol. Res.* **2022**, *28*, 1610373. [CrossRef]

19. Cserni, B. ONEST Calculator. Available online: https://github.com/csernib/onest (accessed on 12 November 2022).
20. Zaiontz, C. Real Statistics Resource Pack | Real Statistics Using Excel. Available online: https://real-statistics.com (accessed on 22 September 2022).
21. Tramm, T.; Di Caterino, T.; Jylling, A.-M.B.; Lelkaitis, G.; Lænkholm, A.-V.; Ragó, P.; Tabor, T.P.; Talman, M.-L.M.; Vouza, E.; Scientific Committee of Pathology, Danish Breast Cancer Group (DBCG). Standardized assessment of tumor-infiltrating lymphocytes in breast cancer: An evaluation of inter-observer agreement between pathologists. *Acta. Oncol.* **2018**, *57*, 90–94. [CrossRef]
22. Kos, Z.; Roblin, E.; Kim, R.S.; Michiels, S.; Gallas, B.D.; Chen, W.; van de Vijver, K.K.; Goel, S.; Adams, S.; Demaria, S.; et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *N.P.J. Breast Cancer* **2020**, *6*, 17. [CrossRef]
23. Adams, S.; Gray, R.J.; Demaria, S.; Goldstein, L.; Perez, E.A.; Shulman, L.N.; Martino, S.; Wang, M.; Jones, V.E.; Saphner, T.J.; 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. [CrossRef] [PubMed]
24. Kojima, Y.A.; Wang, X.; Sun, H.; Compton, F.; Covinsky, M.; Zhang, S. Reproducible evaluation of tumor-infiltrating lymphocytes (TILs) using the recommendations of International TILs Working Group 2014. *Ann. Diagn. Pathol.* **2018**, *35*, 77–79. [CrossRef] [PubMed]
25. Kim, R.S.; Song, N.; Gavin, P.G.; Salgado, R.; Bandos, H.; Kos, Z.; Floris, G.; Eynden, G.G.G.M.V.D.; Badve, S.; Demaria, S.; et al. Stromal Tumor-infiltrating Lymphocytes in NRG Oncology/NSABP B-31 Adjuvant Trial for Early-Stage HER2-Positive Breast Cancer. *J. Natl. Cancer Inst.* **2019**, *111*, 867–871. [CrossRef] [PubMed]
26. Buisseret, L.; Desmedt, C.; Garaud, S.; Fornili, M.; Wang, X.; Van den Eyden, G.; de Wind, A.; Duquenne, S.; Boisson, A.; Naveaux, C.; et al. Reliability of tumor-infiltrating lymphocyte and tertiary lymphoid structure assessment in human breast cancer. *Mod. Pathol.* **2017**, *30*, 1204–1212. [CrossRef]
27. Swisher, S.K.; Wu, Y.; Castaneda, C.A.; Lyons, G.R.; Yang, F.; Tapia, C.; Wang, X.; Casavilca, S.A.; Bassett, R.; Castillo, M.; et al. Interobserver Agreement Between Pathologists Assessing Tumor-Infiltrating Lymphocytes (TILs) in Breast Cancer Using Methodology Proposed by the International TILs Working Group. *Ann. Surg. Oncol.* **2016**, *23*, 2242–2248. [CrossRef] [PubMed]
28. Cabuk, F.K.; Aktepe, F.; Kapucuoglu, F.N.; Coban, I.; Sarsenov, D.; Ozmen, V. Interobserver reproducibility of tumor-infiltrating lymphocyte evaluations in breast cancer. *Indian J. Pathol. Microbiol.* **2018**, *61*, 181–186. [CrossRef]
29. Khoury, T.; Peng, X.; Yan, L.; Wang, D.; Nagrale, V. Tumor-Infiltrating Lymphocytes in Breast Cancer: Evaluating Interobserver Variability, Heterogeneity, and Fidelity of Scoring Core Biopsies. *Am. J. Clin. Pathol.* **2018**, *150*, 441–450. [CrossRef]
30. O'Loughlin, M.; Andreu, X.; Bianchi, S.; Chemielik, E.; Cordoba, A.; Cserni, G.; Figueiredo, P.; Floris, G.; Foschini, M.P.; Heikkilä, P.; et al. Reproducibility and predictive value of scoring stromal tumour infiltrating lymphocytes in triple-negative breast cancer: A multi-institutional study. *Breast Cancer Res. Treat.* **2018**, *171*, 1–9. [CrossRef]
31. Van Bockstal, M.R.; François, A.; Altinay, S.; Arnould, L.; Balkenhol, M.; Broeckx, G.; Burguès, O.; Colpaert, C.; Dedeurwaerdere, F.; Dessauvagie, B.; et al. Interobserver variability in the assessment of stromal tumor-infiltrating lymphocytes (sTILs) in triple-negative invasive breast carcinoma influences the association with pathological complete response: The IVITA study. *Mod. Pathol.* **2021**, *34*, 2130–2140. [CrossRef]