ONEST (Observers Needed to Evaluate Subjective Tests) analysis of the reproducibility of prognostic factors in breast cancer

Ph. D. Thesis

BÁLINT CSERNI, M.Sc.

Szeged, Hungary

2023

ONEST (Observers Needed to Evaluate Subjective Tests) analysis of the reproducibility of prognostic factors in breast cancer

Ph. D. Thesis

BÁLINT CSERNI, M.Sc.

Supervisor:

Gábor Cserni, M.D., D.Sc.

Doctoral School of Multidisciplinary Medical Sciences

University of Szeged

Szeged, Hungary

2023

LIST OF FULL PAPERS THAT SERVED AS THE BASIS OF THE PH.D. THESIS

I. **Cserni B**, Bori R, Csörgő E, Oláh-Németh O, Pancsa T, Sejben A, Sejben I, Vörös A, Zombori T, Nyári T, Cserni G. 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(6):1101-1109. doi: 10.1007/s00428-021-03172-9

IF(2021): 4.535 (Scimago journal ranking: Q1)

II. **Cserni B**, Bori R, Csörgő E, Oláh-Németh O, Pancsa T, Sejben A, Sejben I, Vörös A, Zombori T, Nyári T, Cserni G. 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* 2022;229:153718. doi: 10.1016/j.prp.2021.153718.

IF(2021): 3.309 (Scimago journal ranking: Q2)

OTHER PUBLICATIONS

III. Kiscsatári L, Sárközy M, Kővári B, Varga Z, Gömöri K, Morvay N, Leprán I, Hegyesi H, Fábián G, **Cserni B**, Cserni G, Csont T, Kahán Zs. High-dose radiation induced heart damage in rat model. *In vivo* 2016;30:623-632. IF(2016): 0,953 (Scimago journal ranking: Q3)

IV. Zombori T, Lehóczky L, **Cserni B**, Nyári T, Cserni G. Emlőrákok TNM 8 szerinti anatómiai és prognosztikai stádiumainak retrospektív vizsgálata elhunyt, valaha emlőrákos betegek adatai alapján. [Evaluation of anatomic and prognostic stages of breast cancer according to the 8th edition of the TNM staging system - Retrospective analysis based on data from deceased patients once diagnosed with breast cancer]. *Orv Hetil* 2017; 158:1373-1381. IF(2017): 0,322 (Scimago journal ranking: Q4)

V. Cserni G, Chmielik E, **Cserni B**, Tot T. The new TNM based staging of breast cancer. *Virchows Arch* 2018;472(5):697-703. DOI: 10.1007/s00428-018-2301-9 IF(2018): 2,868 (Scimago journal ranking: Q1)

VI. Kovács ZZA, Szűcs G, Freiwan M, Kovács MG, Márványkövi FM, Dinh H, Siska A, Farkas K, Kovács F, Kriston A, Horváth P, Kővári B, **Cserni BG**, Cserni G, Földesi I, Csont T, Sárközy M. Comparison of the antiremodeling effects of losartan and mirabegron in a rat model of uremic cardiomyopathy. *Sci Rep* 2021;11(1):17495. IF(2021): 4,996 (Scimago journal ranking: Q1)

VII. Christgen M, Kandt LD, Antonopoulos W, Bartels S, van Bockstal MR, Bredt M, Brito MJ, Christgen H, Colpaert C, **Cserni B**, Cserni G, Daemmrich ME, Danebrock R, Dedeurwaerdere F, van Deurzen CHM, Erber R, Fathke C, Feist H, Fiche M, Gonzalez CA, ter Hoeve N, Kooreman L, Krech T, Kristiansen G, Kulka J, Laenger F, Lafos M, Lehmann U, Martin-Martinez MD, Mueller S, Pelz E, Raap M, Ravarino A, Reineke-Plaass T, Schaumann N, Schelfhout AM, de Schepper M, Schlue J, van de Vijver K, Waelput W, Wellmann A, Graeser M, Gluz O, Kuemmel S, Nitz U, Harbeck N, Desmedt C, Giuseppe Floris, Derksen PWB, van Diest PJ, Vincent-Salomon A, Kreipe H. Inter-observer agreement for the histological diagnosis of invasive lobular breast cancer. *J Pathol Clin Res* 2022;8:191-205.. IF(2021): 4,373 (Scimago journal ranking: Q1)

TABLE OF CONTENTS

List of abbreviations 1

1. Introduction 2

   1.1. General introduction 2

   1.2. Estrogen and progesterone receptors, Ki67 4

   1.3. Histological grade 6

   1.4. ONEST (Observers Needed to Evaluate Subjective Tests) 8

2. Aims 10

3. Materials and methods 10

4. Results 13

5. Discussion 27

   5.1. ONEST for estrogen and progesterone receptors, Ki67 27

   5.2. ONEST for histological grade 29

   5.3. Further considerations 31

6. Conclusions 32

7. Acknowledgements 33

8. References 34

Appendix

LIST OF ABBREVIATIONS

ER:	Estrogen Receptor
HE:	Hematoxylin and Eosin
HER2:	Human Epidermal growth factor Receptor-2
IHC:	Immunohistochemistry
ONEST:	Observers Needed to Evaluate Subjective Tests
OPA:	Overall Percent Agreement
OPA(n):	Overall Percent Agreement with n observers
OPAC:	Overall Percent Agreement Curve
PD-L1:	Programmed Death – Ligand 1
PR:	Progesterone Receptor
TIL:	Tumor Infiltrating Lymphocyte
TNBC:	Triple-Negative Breast Cancer
WHO:	World Health Organization
+:	positive
-:	negative

1. INTRODUCTION

1.1. GENERAL INTRODUCTION

Excluding skin malignancies, breast cancer is the most common malignant tumor among women in Europe and worldwide and accounts for a significant number of cancer related deaths [1, 2]. In developed countries, breast cancer mortality is much lower (about one fifth) than its incidence [1, 2]. This discrepancy in the incidence and mortality reflects the effectiveness of breast cancer screening related early recognition and efficient treatment modalities.

Breast cancer, despite a name suggestive of a single disease, is not a single entity. It can be classified along several parameters. A traditional categorization is based on the histological look of the disease, and this histological typing is still the basis of the World Health Organization (WHO) classification of the disease [3, 4]. Most features used for classification involve the prognosis or the treatment of the disease, and histological type is not an exception to this, although its prognostic value is limited; e.g. some special types of breast cancer, like tubular carcinoma or mucinous carcinoma have an excellent prognosis.

Prognosis traditionally reflected the outcome of a disease during its natural course, i.e., without treatment. As breast cancers are treated with different modalities, in the current context, prognosis reflects the outcome of the disease treated with identical or similar modalities, and its prediction is approached by means of prognostic factors. It is common to distinguish between prognostic and predictive factors, of which the latter have a role in predicting response to a given therapy. Prognostic and predictive factors cannot be sharply separated from each other, as there is considerable overlap between them. For example, estrogen receptor (ER) negativity generally reflects a worse prognosis than ER positivity [5], although there are some special types of ER-negative (ER-) carcinomas that have good prognosis [6]; however, ER is mainly considered a predictive factor, as ER negativity makes a cancer unlikely to respond to endocrine treatment targeting the ER pathway.

Some factors may be found to be of prognostic value when used alone (in univariable analyses), but may lose this prognostic value when used in conjunction with other prognosticators (in multivariable analyses). Recognized prognostic (and predictive) factors, that keep their importance even in multivariate models make the obligatory / recommended part of histopathology reports [7-10], and are listed in Table 1.

Table 1. Prognostic and/or predictive factors of invasive breast cancer to be reported on the basis of international and/or national guidelines

Prognostic / predictive factor	Comment
Tumor size	Continuous variable commonly made categorical and being the basis of (p)T categories of the TNM system [11, 12]
Lymph node status	One of the most important prognosticator of breast cancer [13], part of the (p)N categories of the TNM system
Distant metastasis	The basis of the M1 category of the TNM system; distribution and number have further influence on outcome with oligometastatic disease having better prognosis and treatment options [14]
Grade (differentiation)	Reflects the biology of breast cancer, is based on tubule formation, nuclear pleomorphism and mitotic count [15]
Histological type	Has limited prognostic value, but certain types are associated with excellent prognosis
(Lympho)vascular invasion	Reflects the risk of recurrence in node-negative carcinomas
Surgical margins	Positive margins defined as ink on the tumor reflect higher risk of recurrence [9, 10]
ER	Weak as a prognostic factor, but predictive of the response to endocrine therapy
PR	Weak association with prognosis, but along with ER, is predictive of the response to endocrine therapy; ER+PR+ tumors are most likely to respond
HER2	Originally reflecting poor prognosis; predictive of response to HER2 targeting therapies; with the use of these targeted treatments, HER2+ tumors have significantly improved survival
Ki67	Proliferation marker reflecting the proportion of cells in the cell cycle; it is of proven prognostic value, but the distinction between high and low proliferation on the basis of Ki67 labeling is still subject to debate
TILs	Continuous variable, classified using several cut-off levels; prognostic only in TNBC and HER2+ breast cancer; can predict response to neoadjuvant chemotherapy [16, 17]

ER: estrogen receptor, PR: progesterone receptor, HER2: human epidermal growth factor receptor 2; TILs: tumor infiltrating lymphocytes, TNBC: triple-negative breast cancer; +: positive; TNM: Tumor, node, metastasis

Most of the prognostic factors listed in Table 1 are determined with the examination of histological slides stained by conventional histological stains (hematoxylin and eosin; HE) or by immunohistochemistry (IHC). The interpretation of these parameters contains subjective elements, and is therefore subject to interobserver variability.

This doctoral thesis deals with some aspects of the reproducibility of the prognostic factors detailed below.

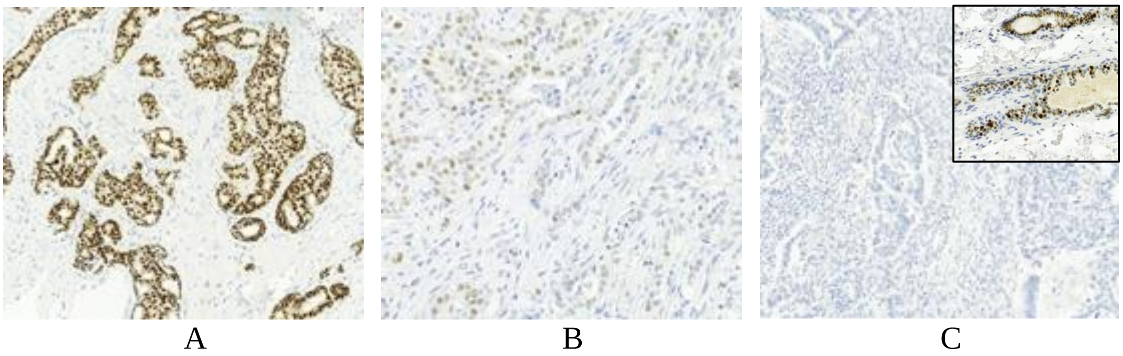
1.2. ESTROGEN AND PROGESTERONE RECEPTORS, KI67

Of the classifications of breast cancer, one of the most important is the segregation of carcinomas into ER+ and ER- groups, of which only the first is likely to benefit from endocrine treatments. Currently, ER status is universally determined by immunohistochemistry (IHC) (Figure 1), and the judgement of what constitutes an ER+ and ER- status is somewhat arbitrary, and may depend on a number of pre-analytical and analytical issues, which are attempted to be minimalized by regularly updated guidelines such as the American Society of Clinical Oncology (ASCO) recommendations [18]. ER positivity had often been defined by an inclusive cut-off value of 10% of tumor cells staining [19, 20], then 1% [21]. At present, it is acknowledged that ER+ cancers with 1-10% ER expression may respond to endocrine treatment, but their response might be below expectations, and therefore these tumors have been allocated to the category of low-ER expressing carcinomas [18, 21]. Indeed, the level of ER expression reflects the degree of endocrine responsiveness as exemplified by the response to adjuvant tamoxifen therapy in function of the Allred-scores (derived sum of the intensity subscores 0-3 and semiquantitative percentage of positive cells subscores 0-5) [22]); the greater the score, the better the response [23].

Progesterone receptors (PR) also influence endocrine responsiveness. Earlier thought to reflect only the integrity of the ER-pathway [21], recently they have been proposed to be actively involved in this pathway [24]. The evaluation of PR and its interpretation is similar to that of ER, and the Allred scoring is also applicable.

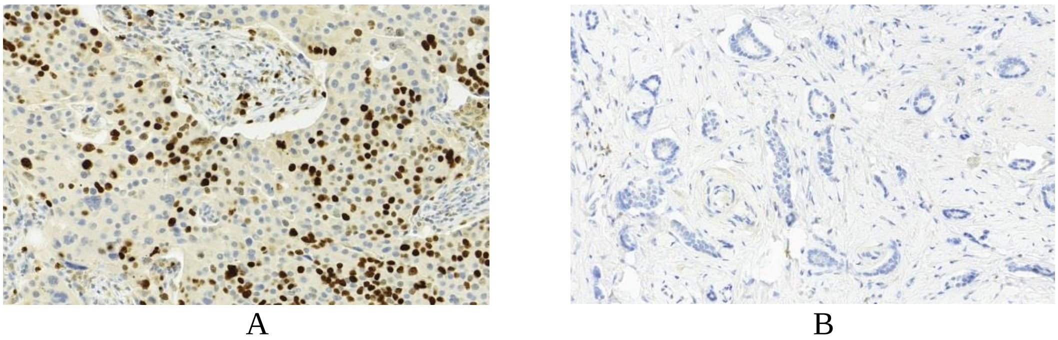
Ki67 is a protein which is expressed in variable amounts through the cell cycle, except in the G0 phase, and is a proliferation marker of prognostic significance [25] (Figure 2).

Figure 1. ER status determined by IHC



A: Case with 7/9 100% ER+ rating (study I, case 53), B: Case with 75-95% ER+ rating (study I, case 75), and C: Case with 9/9 0% ER+, i.e., ER- rating (study I, case 58). (A, C: x10, B: x20; inset normal structures serving as internal control, x10; all Novocastra Lab., Leica, 6F11).

Figure 2: Ki67 staining of breast cancers: examples of high and low proliferation



A: Case with ratings between 20-55% (Study I, case 61); B: Case with ratings between <1% to 2% (Study I, case 78)

Several cut-offs have been suggested to divide ER+ tumors into the low proliferation good prognosis (luminal A-like) category and the more aggressive, more proliferative (luminal B-like) one [26-29]. Despite the accepted prognostic role, owing to concerns about standardization, Ki67 is not part of general recommendations, although it is part of the IHC4 prognostic classifier [30]. As an estimate of proliferative tumor cells, it is also part of Hungarian guidelines for assessing breast carcinomas [9].

ER, PR and Ki67 assessment by microscopy requires the quantification of nuclei that stain with the relevant antibodies. The common method of doing this is by eyeballing, i.e., having a look at the slide and estimating the amount of tumor cells staining. This may be tuned by estimating the area occupied by 100-200 cells, made more precise by counting 500-

2000 cells [31], facilitating the count with an application [32, 33], or by using digital image analysis [34-36] or artificial intelligence [37]. Because of the costs and time required for the latter methods deemed more precise and reproducible, eyeballing is probably the most generally used method worldwide, and is not obviously worse than some forms of digital image analyses [38].

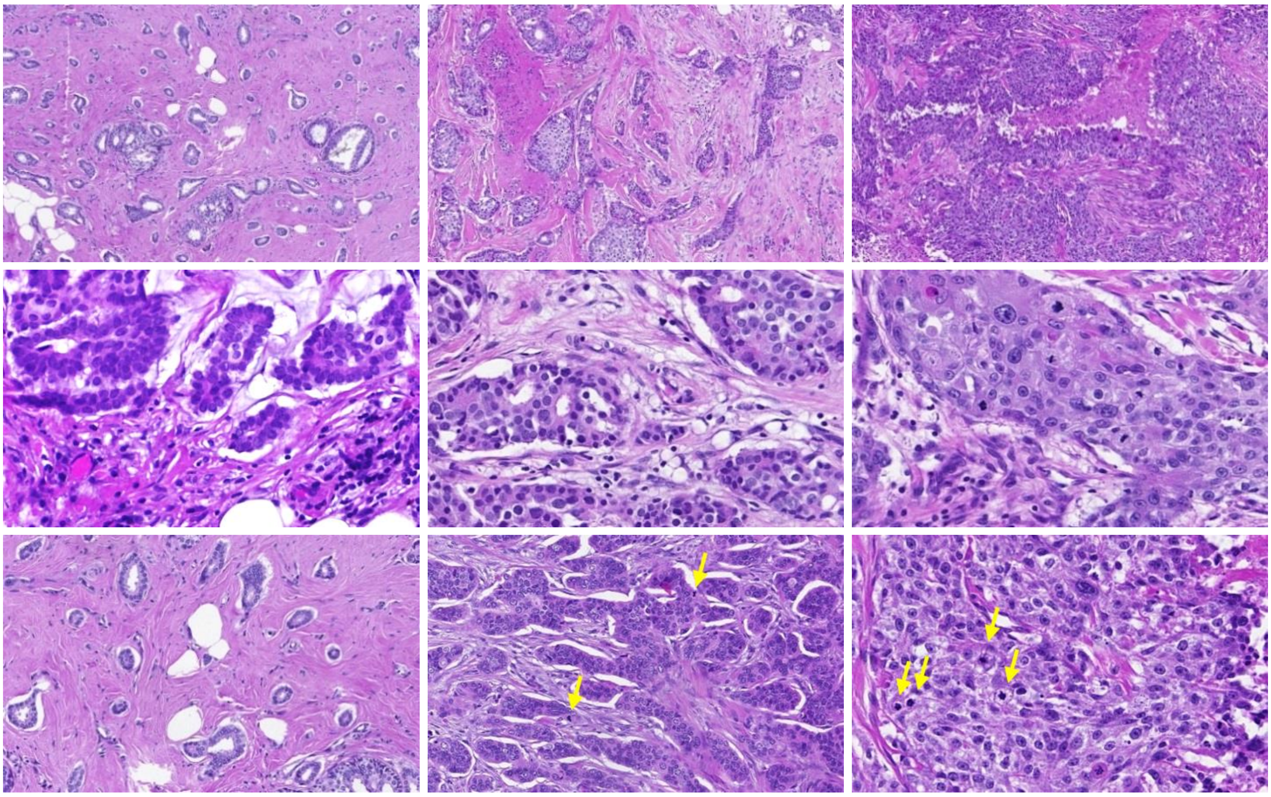
Reproducibility issues have been analyzed by multiple groups. In general, the interobserver agreement for ER and PR assessment for clinical management issues has been excellent for ER-negative cases and fair or good for strongly positive cases, with the worst consistency in allocating tumors to the moderate and low level of receptor positivity [39]. The interobserver consistency has most commonly been assessed by kappa statistics or intraclass correlation coefficients. In study I, we sought at investigating these predictive/prognostic tests by ONEST (Observers Needed to Evaluate Subjective Tests) [40].

1.3. HISTOLOGICAL GRADE

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 [41]. This grading scheme stemming from the original publications of Patey and Scarff from 1928 [42] and Bloom and Richardson from 1957 [43], was refined and standardized according to the Nottingham protocol [15], and is still part of the mandatory items of breast cancer reporting [3, 8, 44]. As a factor with proven prognostic impact [45], it is also part of a number of multivariable analysis derived multiparameter prognostic tools like the Nottingham Prognostic Index [46], Adjuvant!Online! [47] and Predict [48] or the prognostic staging of breast carcinomas defined by the 8th edition of the American Joint Committee on Cancer [12].

Histological grade is determined by the sum of 3 subscores reflecting *glandular differentiation* (“tubule formation”) (1: in >75% of the tumor, 2: in 10-75% of the tumor, 3: in <10% of the tumor), *nuclear pleomorphism* (1: small /1.5x normal ductal cell size/, regular, uniform nuclei with homogeneous chromatin; 2: moderately increased size /1.5-2x normal/ nuclei with moderate variation in size and shape, visible nucleoli; 3: large />2x normal/ nuclei with marked variability in size and shape, vesicular nuclei, multiple nucleoli) and *mitotic activity* (1-3 depending on the mitotic count per 10 high power fields adjusted to the field diameter/area, or more recently 1 mm^2) [3] (Figure 3).

Figure 3. Examples of different histological grade components from study II



Rows demonstrate gland (tubule) formation (T) at x10 magnification (1st row), nuclear pleomorphism (P) at x40 magnification (2nd row) and mitotic activity (M) with images at x20, x20 and x40 magnifications, respectively (3rd row) – arrows point to mitotic figures; columns represent scores 1, 2 and 3 (1st, 2nd and 3rd column, respectively).By row from left to right: case 77 (9/9 T1), case 53 (8/9 T2), case 52 (9/9 T3); case 63 (6/9 P1), case 53 (8/9 P2), case 52 (9/9 P3); case 77 (9/9 M1), case 53 (8/9 M2) and case 52 (9/9 M3)

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.[45]. Most often, overall agreement and kappa statistics have been used to reflect the consistency of defining the histological grade of breast cancers. It has been known for a long time that 2 observers generally agree on all cases better than 3 or more observers.

In study II, 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.

1.4. ONEST (OBSERVERS NEEDED TO EVALUATE SUBJECTIVE TESTS)

ONEST is a recently developed method to characterize how a subjective test requiring quantitative estimations of microscopic images can be reproduced by multiple observers. It has been created to analyze the performance of the atezolizumab related PD-L1(programmed death ligand 1) evaluation algorithm in breast cancer. More precisely, it was introduced to characterize how the estimation of the tumor area occupied by PD-L1 IHC stained immune cells being at least 1% (positive) or less (negative) could be reproduced by multiple observers. This PD-L1 assessment assay has been claimed to have 95% overall percent (proportion) agreement (OPA; i.e., the proportion of cases with full agreement on classification) on the basis of 2 observers [40], but empiricism suggested that the diagnostic test was less reproducible. ONEST is based on plotting the OPAs (0-1; corresponding to 0-100% agreement) against the increasing number of pathologists (observers) of 100 permutations randomly selected from all possible permutations of pathologists. Each plotted OPA for a given permutation results in an OPA curve (OPAC), and the 100 OPACs represent the full ONEST plot.

Based on this description, the author created a tool to help ONEST calculations by designing *Algorithm I*, which was used to calculate ONEST for various inputs.

In *Step 1*, *N* distinct permutations of observer numbers are generated. In our case $|S|$ equals 9, as there are nine pathologists participating in the assessment. This allows for $9! = 362\,880$ distinct permutations, out of which we select $N = 100$. These permutations define the one hundred OPACs of the ONEST plot.

In *Step 2* each OPAC is calculated based on its corresponding permutation. Each permutation represents a certain order of observers participating in the test, and each OPA at the I_{th} index of the OPAC shows the number of cases where the first I observers are in agreement relative to the total number of cases ($|C|$). The $I = 1$ case can be skipped as a single observer is always in agreement with his- or herself (OPA can be considered 1 in this case).

As an example, let us consider permutation 123456789. For $I = 2$ we are comparing the measurements of *Observer 1* and *Observer 2* over all the cases, and count the number of times they are in agreement. For $I = 3$ we also include *Observer 3*, and count the cases in which they all agree. If the permutation was 453216798, then for $I = 3$ we would consider *Observers 4*, *5* and *3*.

Algorithm I: Calculating ONEST

Input: Measurement matrix *M*, where the columns represent the observers, the rows represent the cases. The *M[c, r]* cell at the intersection of column *c* and row *r* store the value measured by observer *c* for case *r*.

Output: The collection of OPACs (i.e., the ONEST plot).

Let $C = \{ 1,\ldots,rows(M) \}$ be the set of case indices, $S = \{ 1,\ldots,columns(M) \}$ the set of observer indices.

1. Generate $N \le factorial(|S|)$ distinct permutations of set *S*.

2. For each permutation *P* calculate the corresponding OPAC by assigning the OPA value to each element of *P* at index $I \ge 2$ as follows:

```
1. function OPA(M, C, S, P, I):
2.     c := 0
3.     for each i in C:
4.         for each j in S where j < I:
5.             if M[P[j], i] ≠ M[P[j + 1], i]:
6.                 goto 3 and continue with the next i
7.         c := c + 1
8.     return c / |C|
```

The resulting OPACs can be plotted by having the number of observers on the X, and the OPA values on the Y axis. Examples follow in the Results section. (For further details and the visual impression, please, look up explanations under Figure 4 on page 17.)

For reproducible classifications, the resulting ONEST plots converge at a certain OPA level, reaching a plateau once past a certain number of observers. ONEST, therefore, suggests the number of observers needed to give adequate estimations of reproducibility, with the plateauing value estimating the overall percent agreement that can be expected. The plots can also be used to visualize the greatest difference in agreement between two observers (wide versus narrow curve ranges; bandwidth).

By applying ONEST to the PD-L1 algorithm tested, 41% agreement plateau was reached with 9 observers [40]. Well reproducible tests have high values of OPA(n) with low numbers of raters to reach the plateau and small difference between the best and worst agreement of two raters, i.e., small bandwidth.

It should be noted that it is also possible to have no cases on which all the participants agree, suggesting that the classification is not reproducible. Similarly, it is also possible that no plateau is reached and the OPACs keep decreasing down to the last observer, implying that more observers would be needed before reproducibility could be determined.

Initially, it was our perception that the plateau only needs to be approached and not necessarily to be reached, and whenever the minimum curve reflecting the lowest OPAs reaches a point from which the decline is minimal (the minimum curve is visually less steep), the number of observers needed (ONEST value) can be acceptably estimated. This approach had been used in our first interpretations, but when the software was fully developed, we opted to define the ONEST value when the plateau is reached for a more uniform interpretation of the results. It must be noted however, that the full ONEST plot always conveys more information than the numeric ONEST value, and one should consider the differences in OPAs between observer counts before finally deciding on the number of observers to use in real applications.

Study I was performed to evaluate the assessment of 3 IHC based biomarkers with nuclear staining by means of ONEST.

In Study II, grade was investigated with ONEST, and previous reproducibility studies and results were assessed in the light of the results.

2. AIMS

To develop a universal computer program for ONEST calculation, and use it to estimate the number of observers needed for a reliable evaluation of reproducibility of some prognostic and predictive factors in breast cancer, notably the assessment of estrogen receptors, progesterone receptors and Ki67 with immunohistochemistry [49], and the determination of histological grade and its components on HE stained sections [50].

3. MATERIALS AND METHODS

From the archives of the Bács-Kiskun County Teaching Hospital, 100 breast cancer cases with routine determination of ER, PR and Ki67 were selected. The cases included 50 core biopsy samples which were taken with a policy to obtain at least 3 cores by 14G needle biopsy gun (CNB) and 50 samples from unrelated resected tumor specimens (EXC). These cases were relatively consecutive, but some ER-PR- cases were discarded to allow better variation of the ER and PR values.

The IHC was performed with monoclonal antibodies 6F11 (Novocastra, Leica, Newcastle, UK) for ER, PgR312 (Novocastra, Leica, Newcastle, UK) for PR and MIB1 (Dako-Agilent, Glostrup, Denmark) for Ki67. Participants were asked to report the percentage of tumor cells staining for all three IHC reactions, along with the average staining intensity and Allred scores for ER and PR.

The ER and PR data were categorized as negative (<1% staining), weekly positive (1-10%) and positive (>10%). Mean intensity scores were given as nil (0), weak (1), medium (2) or strong (3). The Allred scores were categorized into broader groups (0, 2 vs 3-4 vs 5-6 vs 7-8), following the European Working Group for Breast Screening Pathology earlier practice [39].

The Ki67 values were assessed following the Hungarian breast pathology recommendations, which allow for eye-balling based estimation of the Ki67 labelling fraction with rounding to the closest 5%. Individual practice includes estimation similar to ER and PR, but also more quantitative estimations like delineation of groups of about 100 cells and counting labelled cells in a few such sized groups. Five categorizations were evaluated: (1) with the same percentages as for ER and PR – although this has no practical value, it makes the results directly comparable with the steroid hormone receptor values; (2) with cut-offs suggested by the 2009 St Gallen consensus (i.e., ≤15%, 16-30% and >30% for low, intermediate and high proliferation)[26]; (3) with a cut-off suggested by the 2011 St Gallen consensus (i.e., ≤13% and >13% for low and high proliferation)[27]; (4) with a cut-off suggested by the 2013 St Gallen consensus (i.e., ≤20% and >20% for low and high proliferation)[28], and finally (5) with cut-offs suggested by the 2015 St Gallen consensus (i.e., at least 10% less than the median labelling of ER+ breast cancers for low labelling, at least 10% more than this median value for high proliferation, and the range in between for intermediate labelling)[29]. For this, the median Ki67 labelling (15%) of ER+ cases diagnosed in 2020 (n=170) was used.

Rating reliability was analyzed by the intraclass correlation coefficient (two-way random effects, absolute agreement, single rater/measurement; ICC(2,1) [51]).

In parallel, all observers were asked to also grade the 100 cases according to current practice, as recommended by the most recent WHO Classification of breast tumors [3] and report the scores for tubule/gland formation, nuclear pleomorphism and mitotic counts, along with the histological grade of the tumors.

For the analysis of reproducibility for grade descriptive statistics, the ICC(2,1) and Fleiss kappa values [52] 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 [51]. 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 [53].

ONEST, as initially described by Reisenbichler et al [40], was calculated for a randomly selected 100 permutations of the 362,880 (=9!) possible permutations of ranked pathologists. The computer program used for the calculation was developed by the author using the C++ programming language and the wxWidgets graphical user interface (GUI) library. Some calculations during the studies were performed with an earlier, but algorithmically identical version of the software.

The Kruskal–Wallis test was applied to characterize and compare minimum values (i.e., minimum OPACs, the lowest plots – the "worst performances"); p-values <0.05 were considered statistically significant. The calculations were performed with the Real Statistics Resource Pack Excel add-in [54].

In the light of our findings, previous reproducibility studies of the histological grading using the Nottingham modification of the original Scarff Bloom Richardson scheme [15] 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 [45] 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.

4. RESULTS

Nine pathologists, including 2 residents trained in breast pathology have evaluated the 100 cases. They all had experience in the field of breast pathology, ranging from >1 to >25 years.

As the consistency of classifying the cases is dependent on the percentage of cells staining, with 0% and 100% being the easiest to categorize unanimously, Supplementary figure 1 demonstrates the boxplots for the main descriptive statistical features of the 50 CNB and 50 EXC specimens for the 3 nuclear markers assessed. As the cases were continuous but with exclusion of some ER- cases, the median scores for the markers are only characteristic for the cases assessed; but to some extent they also reflect breast cancer cases encountered in routine practice. The median percentage (interquartile range) of ER+, PR+ and Ki67+ cells as assessed by the 9 pathologists in biopsies vs excision specimens were: 95 (30) vs 95 (15) (ER), 60 (89) vs 73 (95) (PR) and 20 (85) vs 10 (20) (Ki67), respectively. These values highlight that most nuclei stained for ER, less nuclei labelled with PR and the least with Ki67.

The OPAs per diagnostic categories are displayed in Supplementary Tables 1 and 2. The 100% agreement per diagnostic categories for ER and PR were high in both CNB and EXC specimens (38 to 47/50 cases), but were somewhat worse for a similar distribution of Ki67 (31/50) on CNB and less than 50% (22/50) for Ki67 on EXC (Supplementary Table 1). With different St Gallen recommendations on interpreting Ki67 labelling values, consensus on categorization was best on CNB with the 2011 two-tiered-classification: 30/50 cases were classified with 100% agreement (Supplementary Table 2).

The ICC values for the evaluated parameters are shown in Table 2. According to these, most classifications relating to the ER and PR status of the tumors have excellent or good to excellent level of reliability. In contrast, all Ki67 related classifications have moderate or moderate to good reliability. The difference in ICC values of the 3-category-based (1% and 10% cut-off) classification of ER or PR vs Ki67 is striking, whereas the difference in ICC values of different Ki67 categorizations is less prominent. No major or consistent differences are seen between the ICC values of CNB and EXC specimens.

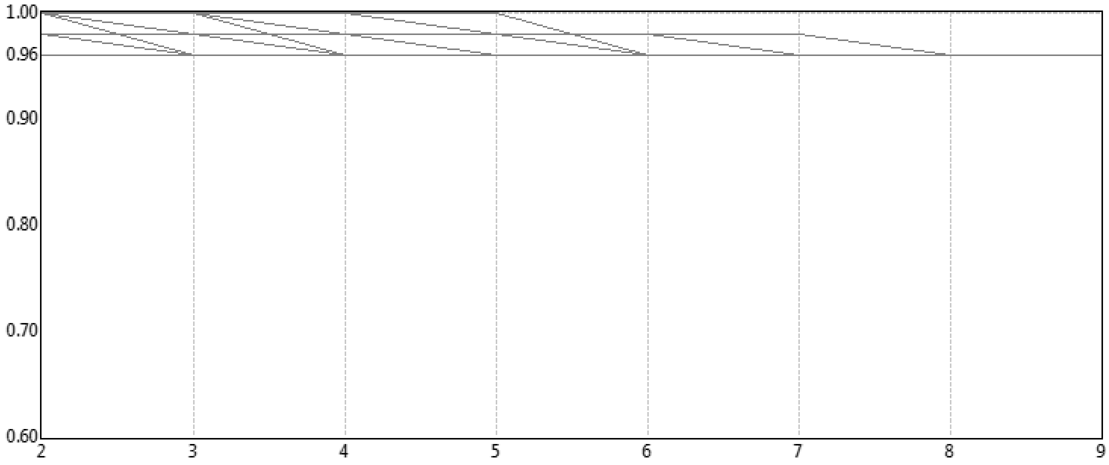
Table 2. ICC (95% confidence interval, CI) values for the investigated categories

	CNB	EXC
ER intensity	0.813 (0.740-0.876)	0.873 (0.815-0.919)
PR intensity	0.782 (0.705-0.851)	0.830 (0.765-0.886)
ER (QS)	0.924 (0.890-0.951)	0.979 (0.968-0.987)
PR (QS)	0.920 (0.886-0.948)	0.927 (0.896-0.953)
ER (%)	0.909 (0.870-0.941)	0.969 (0.954-0.981)
PR (%)	0.942 (0.917-0.941)	0.935 (0.907-0.958)
Ki67 (%)	0.874 (0.812-0.921)	0.812 (0.742-0.874)
ER (% - 3 categories)	0.918 (0.883-0.947)	0.996 (0.994-0.997)
PR (% - 3 categories)	0.946 (0.922-0.965)	0.918 (0.883-0.947)
Ki67 (% - 3 categories)	0.673 (0.576-0.768)	0.625 (0.520-0.731)
Ki67 (St Gallen-2009)	0.760 (0.677-0.836)	0.707 (0.614-0.796)
Ki67 (St Gallen-2011)	0.654 (0.555-0.753)	0.629 (0.525-0.735)
Ki67 (St Gallen-2013)	0.629 (0.526-0.733)	0.649 (0.546-0.751)
Ki67 (St Gallen-2015)	0.698 (0.600-0.790)	0.700 (0.603-0.791)

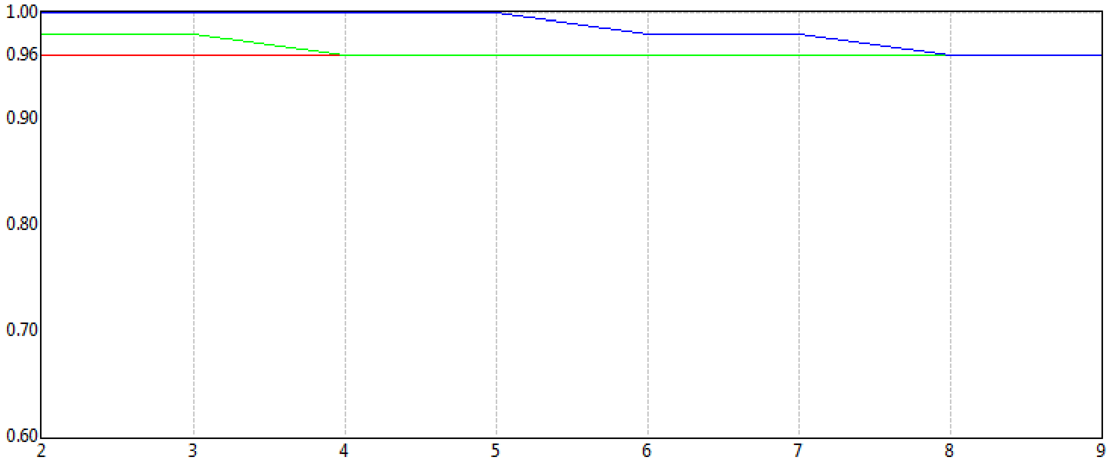
ER: estrogen receptor; PR: progesterone receptor; QS: quick score or Allred score; intensity refers to average intensity scorings; (%) refers to the recorded percentage values with all different values representing a different category; 3 categories refer to <1%, 1-10% and >10% categorization; St Gallen – year refers to the categories of low/(intermediate)/high Ki67 labelling as defined by the St Gallen Consensus Conference of the given year (see Methods). The greyscale reflects the categorization of the level of reliability into excellent (ICC>0.9), good to excellent, good (ICC>0.75-0.9), moderate to good and moderate (ICC>0.5-0.75) from white to deeper shades of grey; the 95% CIs are taken into account for the categorization [51].

As demonstrative examples, ONEST plots of the ER, PR and Ki67 classifications of CNB samples reflected in Supplementary Table 1 (i.e., with categories <1%, 1-10% and >10%) are shown in Figure 4. The A1, B1 and C1 parts of the figure demonstrate OPACs of ER (A1), PR (B1) and Ki67 (C1) classifications of 100 randomly selected permutations of 9 pathologists, whereas only the minimum, median and maximum values of these OPAs are shown in the A2, B2 and C2 parts. Rather than demonstrating all possible ONEST plots, the minimum, maximum and median OPA values are shown in Supplementary Table 3, and the differences between the maximum and minimum OPAs, the OPA for all 9 pathologists, the number of pathologists to reach the plateau are shown in Table 3.

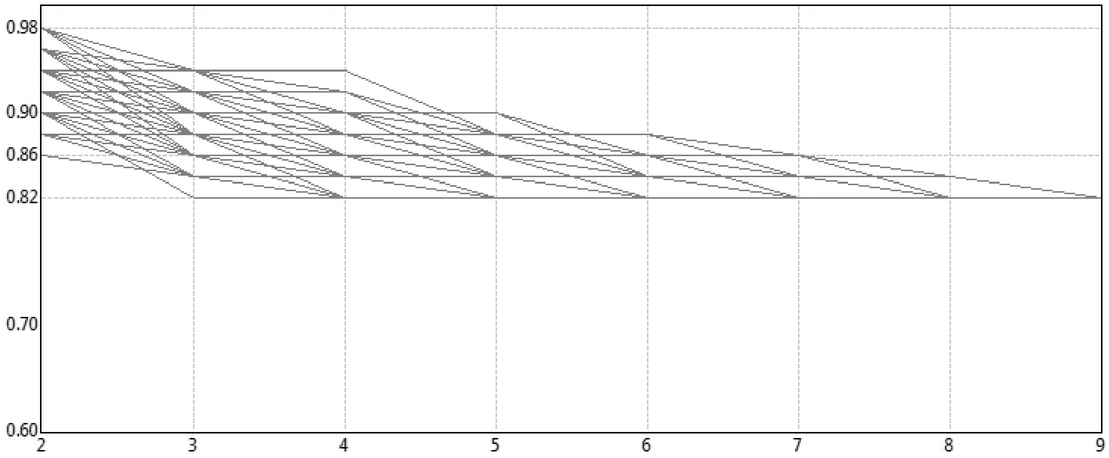
Figure 4. ONEST plots of ER (A), PR (B) and Ki67 (C) classifications into <1%, 1-10% and >10% categories on CNB with all 100 random permutations of pathologists (A-B-C 1) and just the best (blue), worst (red) and median (green) OPA values from these 100 permutations (i.e., "simplified" ONEST plots; A-B-C 2)



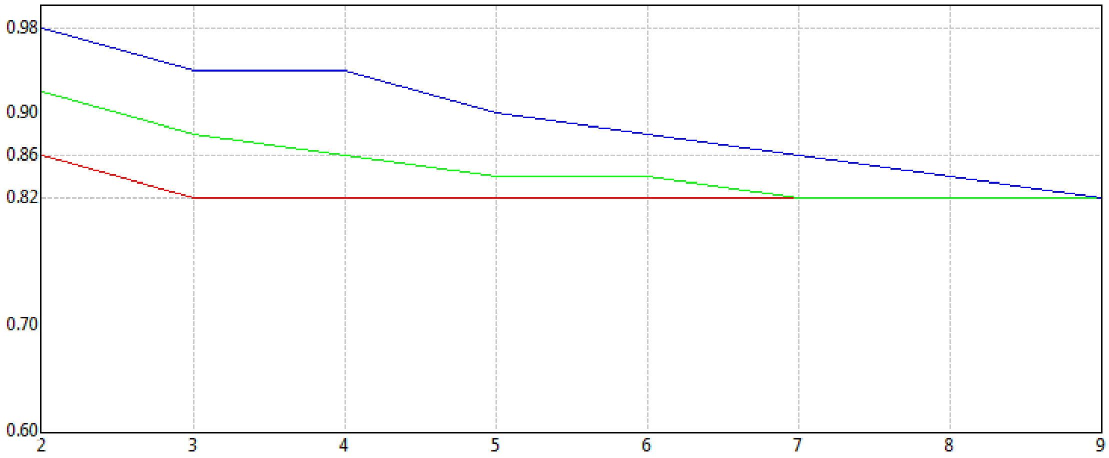
A1: ONEST plot for ER



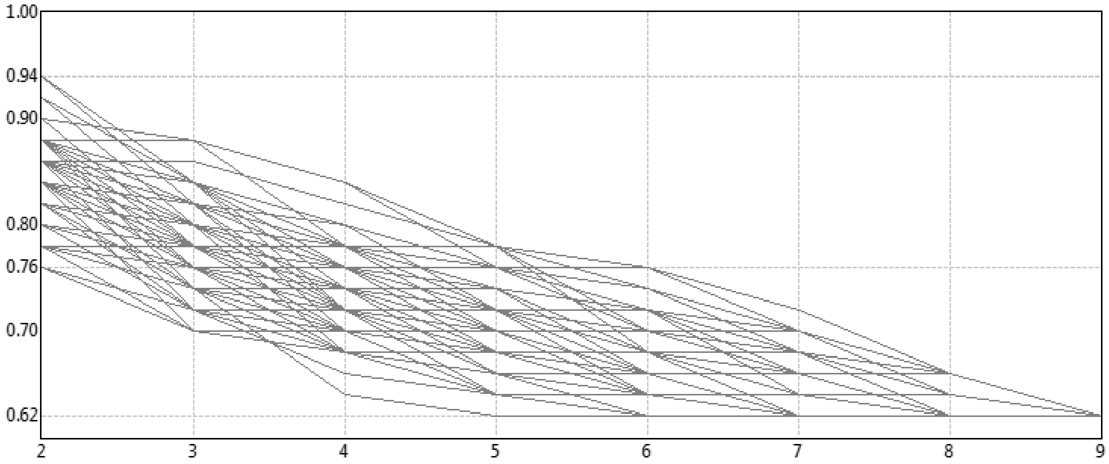
A2: Simplified ONEST plot for ER



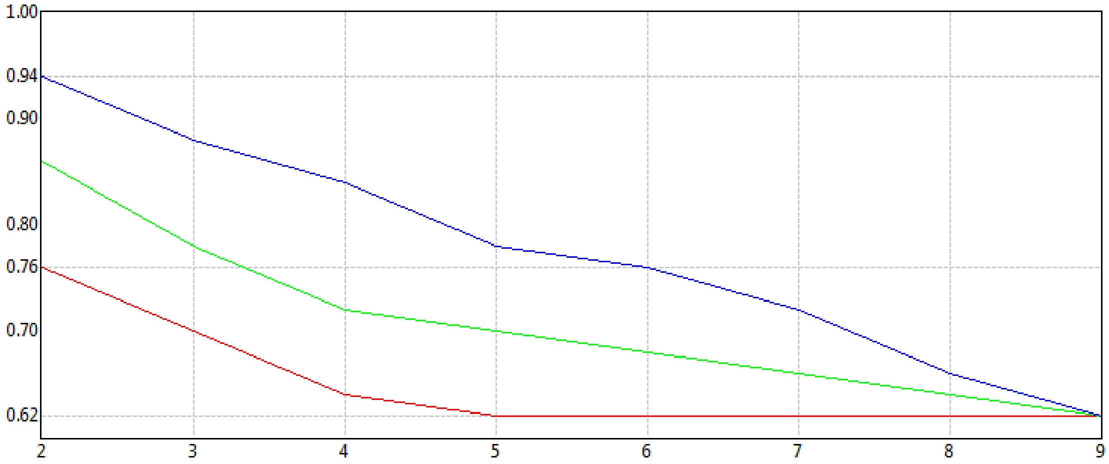
B1: ONEST plot for PR



B2: Simplified ONEST plot for PR



C1: ONEST plot for Ki67



C2: Simplified ONEST plot for Ki67

Notes:

On the ONEST plots the X axis reflects the number of observers and the Y axis the OPA values.

C2 demonstrates best that with increasing number of pathologists, the OPA decreases till approaching a plateau with 4 (or reaching it with 5) pathologists. The classification can be characterized with the distance between the minimum and maximum OPA with 2 pathologists (0.94-0.76=0.18; bandwidth), the number of pathologists required to reach the plateau (5; ONEST value), the value of the plateau (0.62) and the OPA value for all observers (0.62; OPA(n)). The OPA(n) coincides with the value of the plateau whenever there is a plateau. Categorizations with good reproducibility have a narrow gap between the maximum and minimum values (narrow bandwidth), reach the plateau with few pathologists (low ONEST value) and have a high OPA for all pathologists (high OPA(n); e.g.: A1, A2).

While A1, B1, C1 demonstrate 100 OPACs each, A2, B2, C2 show the minimum and maximum OPA values out of these that do not necessarily overlap with an OPAC from the 100 permutations, but obviously overlap with an OPAC from all permutations.

The worst scenario (i.e., the minimum OPA values) was selected to characterize the categorizations.

Table 3 Main results of the ONEST analyses of different parameters

	Maximum OPA differences (Bandwidth)	Pathologists needed for plateau (ONEST value) *	OPA with 9 pathologists OPA(9)
ER categories (<1%, 1-10%, >10%) CNB	0.04	2 (2)	0.96
ER categories (<1%, 1-10%, >10%) EXC	0.02	2 (2)	0.98
ER intensity CNB	0.32	5 (6)	0.48
ER intensity EXC	0.36	4 (5)	0.38
ER Allred scores (0,2; 3-4; 5-6; 7-8) CNB	0.12	4 (4)	0.72
ER Allred scores (0,2; 3-4; 5-6; 7-8) EXC	0.10	2 (2)	0.90
PR categories (<1%, 1-10%, >10%) CNB	0.12	3 (3)	0.82
PR categories (<1%, 1-10%, >10%) EXC	0.18	3 (3)	0.76
PR intensity CNB	0.36	4 (5)	0.38
PR intensity EXC	0.42	4 (5)	0.36
PR Allred scores (0,2; 3-4; 5-6; 7-8) CNB	0.22	5 (6)	0.48
PR Allred scores (0,2; 3-4; 5-6; 7-8) EXC	0.20	3 (4)	0.58
Ki67 categories (<1%, 1-10%, >10%) CNB	0.18	4 (5)	0.62
Ki67 categories (<1%, 1-10%, >10%) EXC	0.26	4 (5)	0.44
Ki67 St Gallen 2009 CNB	0.30	4 (4)	0.32
Ki67 St Gallen 2009 EXC	0.28	4 (5)	0.38
Ki67 St Gallen 2011 CNB	0.18	5 (6)	0.6
Ki67 St Gallen 2011 EXC	0.24	4 (5)	0.5
Ki67 St Gallen 2013 CNB	0.22	5 (6)	0.52
Ki67 St Gallen 2013 EXC	0.26	5 (5)	0.54
Ki67 St Gallen 2015 CNB	0.3	4 (5)	0.32
Ki67 St Gallen 2015 EXC	0.34	5 (5)	0.26

* The values given are those gained with visual impression on the basis of the less steep decline of the minimum curves; values in parenthesis are those where the plateau is definitely reached.

As concerns the classifications according to the 1% and 10% cut-offs or the different St Gallen criteria, the intensity scores for ER and PR, and the Allred scores lumped into 4 categories, there were no significant differences (Kruskal–Wallis, p>0.05) between CNB and EXP sample OPAs for the PR intensity scores and the Ki67 categories according to the St Gallen 2013 criteria; all the other classifications significantly differed in OPAs for CNB and EXC specimens. Agreement was better on CNB specimens for ER intensity, PR status, Ki67 categories with 1% and 10% cut-offs, St Gallen 2011 and 2015 cut-offs, and was better on

EXC specimens for ER status, ER and PR Allred scores, Ki67 classification according to St Gallen 2009.

Using the <1%, 1-10% and >10% cut-offs for categorization, there were significant differences in the minimum OPA values from the ONEST plots between any pairs of ER, PR, and Ki67s both on CNB and EXC specimens.

The 4-category (0, 2 vs 3-4 vs 5-6 vs 7-8) Allred score grouping minimum OPA values were also significantly different for ER and PR on both CNB and EXC specimens, whereas these values for the scores for average intensity of staining showed significant differences only for CNB specimens and not for EXC specimens (Kruskal–Wallis, p=0.44).

As concerns the classification of Ki67 labeling indices into low vs high (vs intermediate if defined) proliferation according to different definitions proposed by consecutive St Gallen consensus conferences, the highest OPA was noted with the 2013 proposal, i.e., a classification based on ≤20% vs >20%, and this was significantly better than any other St Gallen recommendation based segregation. However, ICC values still suggested moderate to good (CNB) or good (EXC) level of reliability (Table 2).

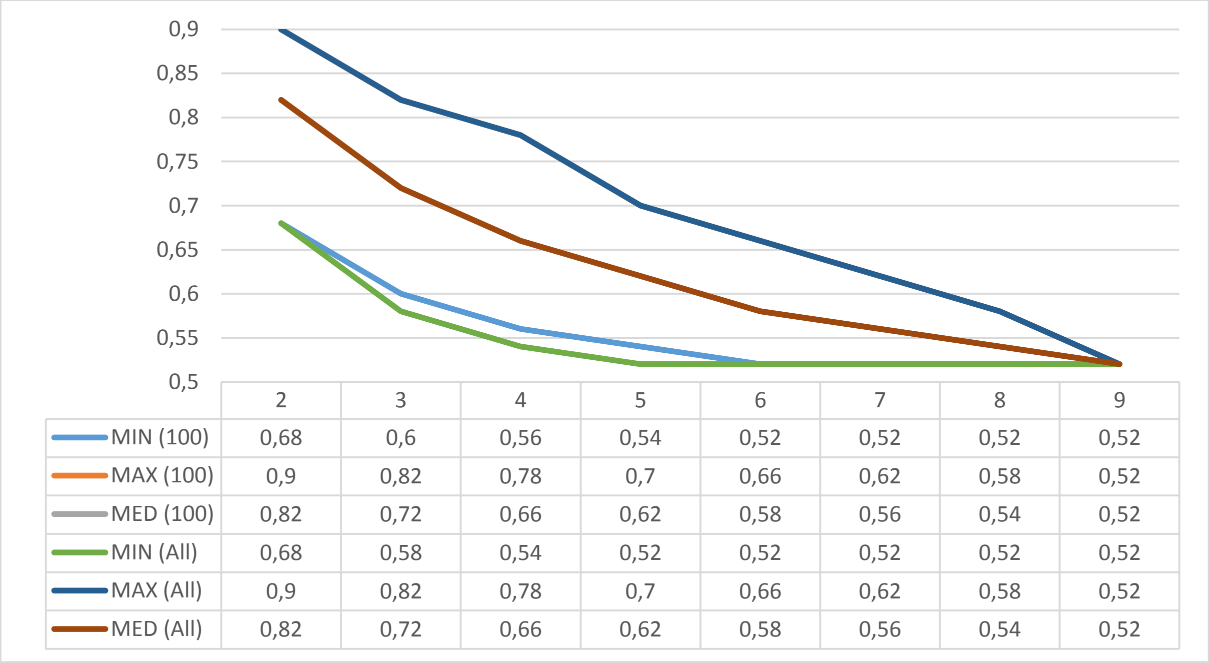
As 9! (362,880) is still a manageable number, the minimum OPA values per number of observers from the 100 random permutations were compared with the minimum OPA values per number of observers form all permutations (i.e., the lowest OPAC). No significant differences were noted, most comparisons (Kruskal–Wallis) yielded p=1, and p values ranged from 0.64 to 1. The classification with the lowest p value is illustrated on Figure 5.

The individual ratings for the component scores of histological grade and the grade itself are represented in Figure 6. 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 Figure 6.

The kappa and ICC values are shown in Tables 4 and 5, 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 4).

Figure 5. Comparison of OPAs derived from 100 and all permutations of pathologist for Ki67 categorization according to the St Gallen 2013 recommendation

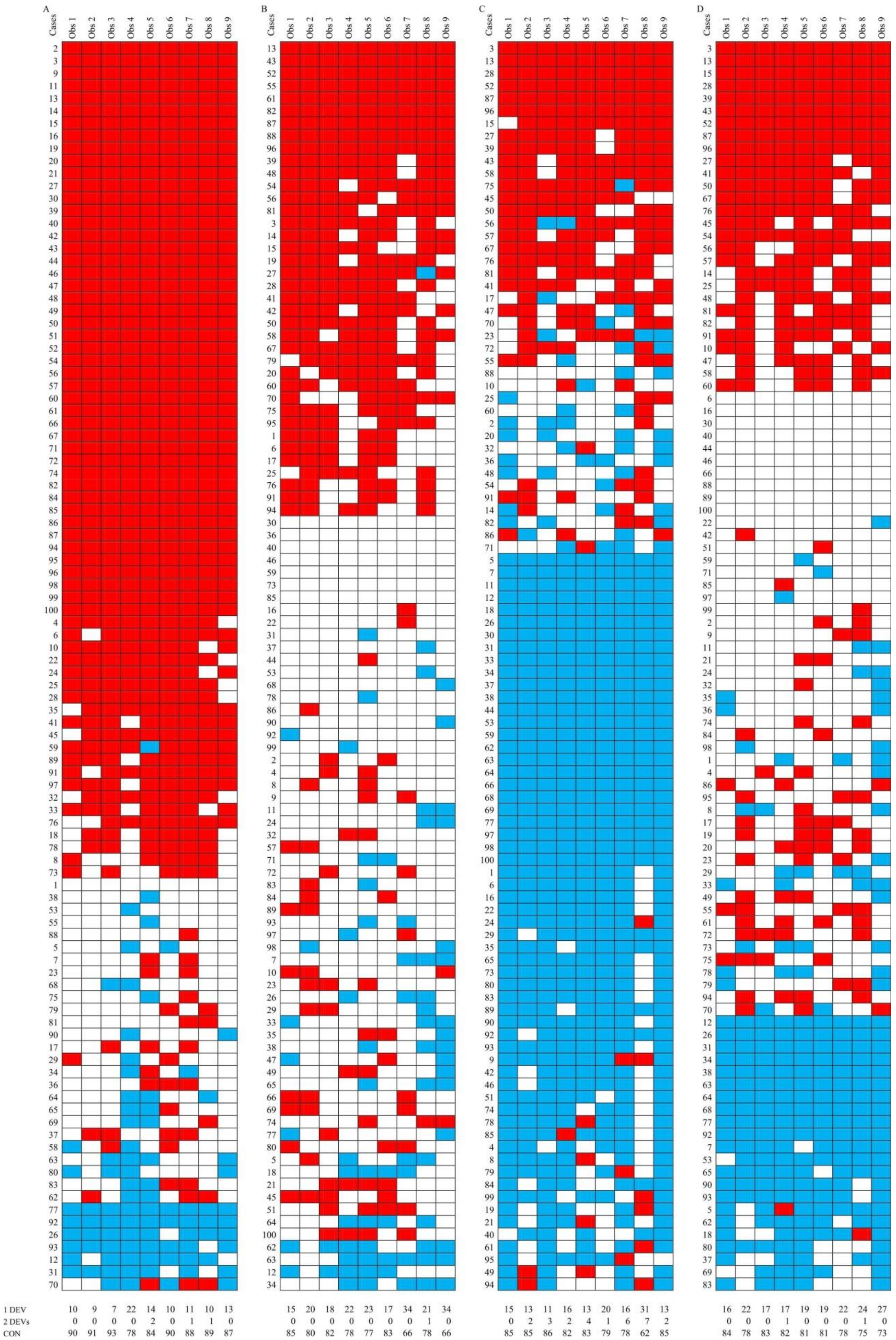


	2	3	4	5	6	7	8	9
MIN (100)	0,68	0,6	0,56	0,54	0,52	0,52	0,52	0,52
MAX (100)	0,9	0,82	0,78	0,7	0,66	0,62	0,58	0,52
MED (100)	0,82	0,72	0,66	0,62	0,58	0,56	0,54	0,52
MIN (All)	0,68	0,58	0,54	0,52	0,52	0,52	0,52	0,52
MAX (All)	0,9	0,82	0,78	0,7	0,66	0,62	0,58	0,52
MED (All)	0,82	0,72	0,66	0,62	0,58	0,56	0,54	0,52

MIN: minimum; MAX: maximum; MED: median; (100): for the 100 permutations, (All) for all 9! permutations. The MIN(All) and MAX(All) represent the worst and best OPAC, whereas the MIN(100) and MAX(100) curves lay on the worst and best OPA values and do not necessarily represent an OPAC from the 100 curves plotted. The MED values are derived from the 100 or 9! OPA values belonging to the respective number of pathologists on the x axis.

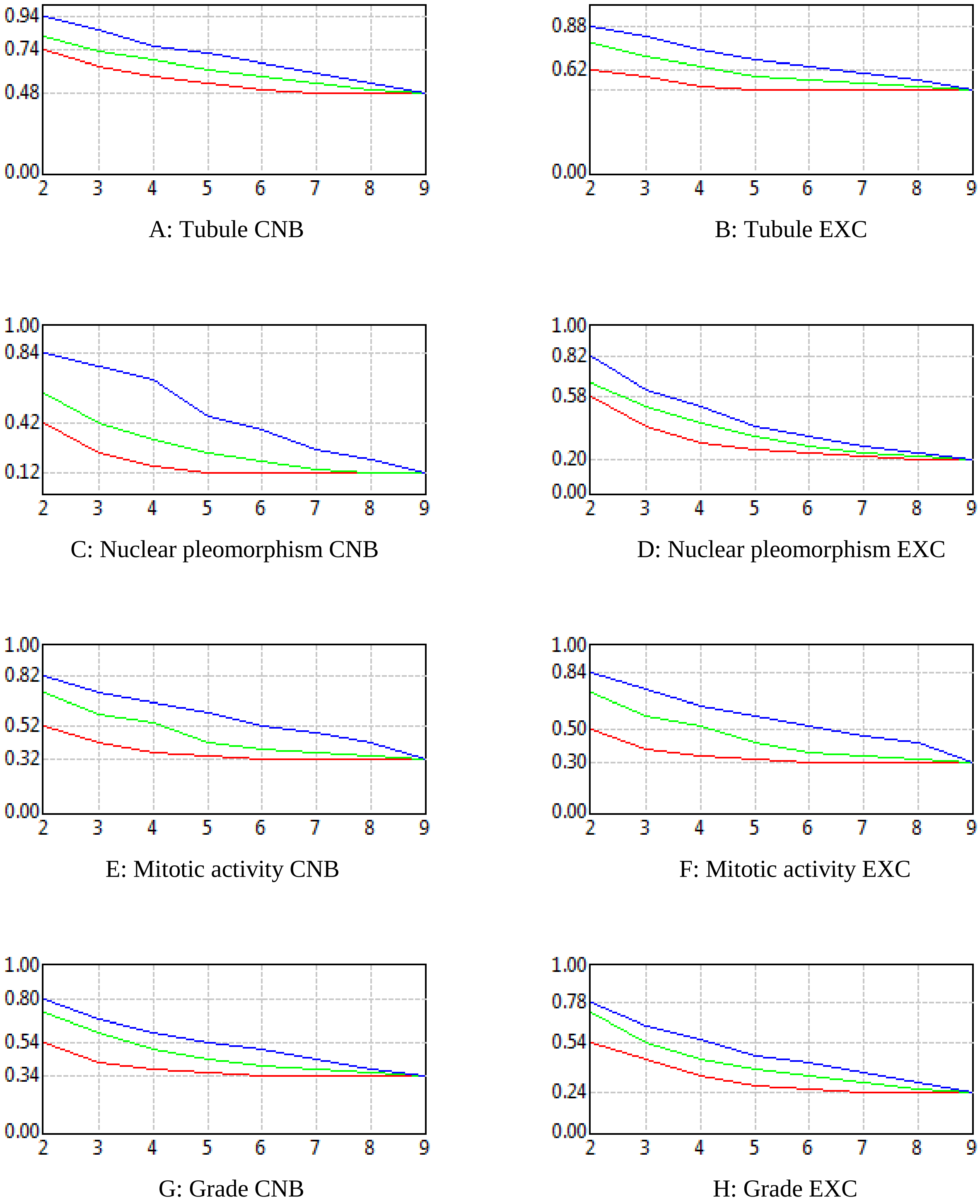
The MAX and MED values (curves) overlap completely. The MIN(100) vs MIN(All) curves deviate slightly, but the differences are not significant (Kruskal–Wallis, p=0.64)

Figure 6 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. 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).

Figure 7. ONEST plots with minimum (red), median (green) and maximum (blue) 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



A: Tubule CNB

B: Tubule EXC

C: Nuclear pleomorphism CNB

D: Nuclear pleomorphism EXC

E: Mitotic activity CNB

F: Mitotic activity EXC

G: Grade CNB

H: Grade EXC

Note: Like in Figure 4, the X axis of the ONEST plots represents the number of observers and the Y axis the OPA values.

Table 4. Kappa values for component scores of histological grade and grade itself

Fleiss kappa				
Parameter	Score 1	Score 2	Score 3	Overall
Tubule scores (CNB)	0.64	0.45	0.62	0.56
Tubule scores (EXC)	0.76	0.50	0.47	0.61
Pleomorphism score (CNB)	0.44	0.25	0.19	0.32
Pleomorphism score (EXC)	0.54	0.35	0.25	0.42
Mitosis score (CNB)	0.58	0.18	0.58	0.48
Mitosis score (EXC)	0.57	0.16	0.58	0.47
Grade (CNB)	0.59	0.43	0.62	0.54
Grade (EXC)	0.50	0.37	0.70	0.51

Kappa scale	Interpretation [53]
-1 to -0.01	poor
0-0.10	slight
0.11-0.20	slight
0.21-0.30	fair
0.31-0.40	fair
0.41-0.50	moderate
0.51-0.60	moderate
0.61-0.70	substantial
0.71-0.80	substantial
0.81-0.90	almost perfect
0.91-1	almost perfect

Table 5. ICC values for component scores of histological grade and grade itself

Parameter	ICC	95%CI
Tubule scores (All	0.735	(0.673-0.795)
Tubule scores (CNB)	0.732	(0.644-0.814)
Tubule scores (EXC)	0.733	(0.642-0.817)
Pleomorphism score (All)	0.507	(0.426-0.594)
Pleomorphism score (CNB)	0.459	(0.346-0.588)
Pleomorphism score (EXC)	0.561	(0.452-0.676)
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)

ICC Scale	Interpretation [51]
<0.5	poor
95%CI<0.5	moderate to poor
0.5-0.749	moderate
95%CI>0.749	good to moderate
0.75-0.9	good
>0.9	excellent

Table 6. Main data from the ONEST analysis

	Minimum observer needed: ONEST curve reading (MIN)*		Maximum difference in OPA for 2 observers (bandwidth)		Median OPA for 2 observers		Overall agreement of all observers, i.e., OPA(9)	
	CNB	EXC	CNB	EXC	CNB	EXC	CNB	EXC
TUB	4 (7)	4 (5)	20%	26%	82%	78%	48%	50%
PLEOM	4 (5)	4 (8)	42%	24%	60%	66%	12%	20%
MIT	4 (6)	4 (6)	30%	34%	72%	72%	32%	30%
Grade	3 (6)	4 (7)	26%	24%	72%	72%	34%	24%

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

* The values given are those gained with visual impression on the basis of the less steep decline of the minimum curves; values in parenthesis are those where the plateau is definitely reached.

Regarding ONEST, the plots are reproduced in Figure 7, and the main values are shown in Table 6. 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; the median value reflects that two pathologists are agreeing on the grade in 72% of the cases. Importantly, the ONEST plots suggest that at least 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 they reach a plateau at 6 (CNB specimens) or 7 (EXC specimens) observers (Table 6, Figure 7).

For the minimum OPA values, there were significant differences between CNB and EXC specimens in the cases of nuclear pleomorphism (Kruskal–Wallis, 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. The minimum OPACs for other parameters (i.e., scores for tubule formation and mitotic rate) were not statistically different in CNB and EXC specimens.

Tables 7 and 8 demonstrate the results of previous studies on histological grading on the basis of kappa values (Table 7) [55-71] and OPA of all observers (Table 8) [55, 57, 60-63, 65-68, 72, 73]. Both of these tables suggest that reproducibility figures gained with less than 4 observers (i.e., the ONEST value) or by pairwise comparisons (virtually) reflect better agreement.

Table 7. Kappa values gained in different studies of histological grade reproducibility

Interpretation [53]				poor	slight		fair		moderate		substantial		near-perfect		
Scale				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	
Reference	Observers	Cases	Kappa												Scale
Jacquemier [55]	2 (21)	24	a							0.53					<0
Sikka [56]	2 (3)	40	b								0.68-	▨	-0.83		0-0.10
Meyer [57]	2 (7)	49	c						0.50-	-0.59					0.11-0.20
Robbins [58]*	2 (5)	50	d									0.73			0.21-0.30
Robbins [58]**	2 (5)	50	d							0.58					0.31-0.40
Anderson [59]	2	52	b							0.54					0.41-0.50
Frierson [60]	2 (6)	75	b						0.43-	▨	▨	-0.74			0.51-0.60
Rabe [61]	2 (6)	100	b							0.58-	▨	▨	-0.86		0.61-0.70
Ginter [62]	2 (6)	143	b					0.35-	▨	▨	-0.68				0.71-0.80
Postma [63]	2	310	e									0.80			0.81-0.90
Reed [64]	2	613	b								0.69				0.91-1
Bueno-de-Mesquita [65]	2	694	f							0.56					
Cserni [66]	3	75	g						0.41						
Rabe [61]	6	100	g								0.68				
Ginter [62]	6	143	g						0.50						
Boiesen [67]	7	93	g							0.54					
present (CNB)	9	50	g							0.54					
present (EXC)	9	50	g							0.51					
Longacre [68]	13	35	h					0.40-			-0.70				
Sloane [69]	23	57	i							0.53					
Ellis [70]***	>200	76	j				0.24-	-0.36							
Ellis [70]****			j						0.45-	-0.53					
Rakha [71]	>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; h: 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.

Table 8. Overall proportion agreement (OPA) values gained in different studies of histological grade reproducibility

Scale

Reference	Observers	Cases	Comment	🟪	🟦	🟩	🟩	🟨	🟨	🟧	🟧	🟥	🟥
Robbins [58]*	5	50	a									0.83	
Robbins [58]**	5	50	a								0.74		
Postma [63]	2	310										0.88	
Bueno-de-Mesquita [65]	2	694									0.72		
Cserni [66]	3	75						0.44					
Theissig [72]	3	166									0.72		
Frierson [60]	6	75					0.40						
Rabe [61]	6	100	b						0.54				
Ginter [62]	6	143				0.30							
Meyer [57]	5-7	72	c							0.67	0.74		
Boiesen [67]	7	93					0.31						
present (CNB)	9	50					0.34						
present (EXC)	9	50				0.24							
Longacre [68]	13	35					0.40						
Jacquemier [55]	21	24								0.69			
Dalton [73]	25	10				0.30							

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.

5. DISCUSSION

5.1. ONEST FOR ESTROGEN AND PROGESTERONE RECEPTORS, KI67

It is recognized that many factors influence the assessment of ER, PR and Ki67 by IHC. This study concentrated on interpretational issues only, although two different types of material were evaluated in parallel: in contrast to whole section excision material, core biopsies have better fixation parameters and a smaller overall area to evaluate, potentially diminishing the discrepancies between observers.

With 100 cases mostly reflecting daily routine, ER and PR statuses (negative vs low positive vs positive) were the most reproducible with excellent or excellent to good classification of reliability. ONEST suggested that the categorization of ER showed the highest rates of OPA, and even 2 observers were sufficient to reflect reproducibility of assessment of the ER status, whereas PR was characterized by slightly lower OPA values and by 3 observers required for reflecting reproducibility (Tables 2 and 3, Figure 4). The results suggest that these tests are valuable as assessed in daily practice. Although no recommendation exists to use Ki67 with <1%, 1-10% and >10% categories, to allow better comparison with the determination of ER and PR, the virtual exercise of classifying cases according to these cut-offs was also done: the ICC suggested moderate or moderate to good reproducibility, the OPAs per increasing number of pathologists were lower, and the number of observers required for better assessment of reproducibility was at least 4 (Tables 2 and 3, Figure 4). As all tests reflected the estimation of the percentage of stained tumor cell nuclei (without the influence of staining intensity) and their classification according to the same cut-off limits, the difference between the individual tests was only the proportion of stained nuclei and the size of the specimen (greater for EXC than CNB). It has been found in several studies that intermediate categories are less reproducible than categories at the extremes [39, 74, 75], and indeed, as indicated in the results (see also Supplementary Figure 1), Ki67 staining proportions were often away from the extremes, which seems typical for this marker [76].

The intensity of staining was also assessed for ER and PR, and although the ICC values were reasonably good or even good to excellent (range 0.78-0.87), the ONEST analysis suggested that OPA values were low (0.36 to 0.48), with less than half of the pathologists agreeing, and therefore at least 4 to 5 pathologists needed to assess reproducibility. As the Allred quick scores are composed of subscores for intensity and for proportion of stained cells, these consequently had ICC values reflecting excellent (with the 95% CI, good to excellent) reliability. However, the ONEST analysis of Allred scores reflected up to 22% difference between two observers, and 2 to 6 pathologists required to assess reproducibility, with the worst results for PR assessment on CNBs (Table 3).

The comparison of ER, PR and Ki67 with the 1% and 10% cut-offs suggested that the last biomarker was the least reproducible, and this could probably be explained by the relatively wide range in the proportion of the stained cells per case. On the basis of daily practices reflected in this study, different classifications of low vs high (vs intermediate when defined) proliferation categories are not excellently reproducible (Table 2), the ICC values ranged from 0.63 to 0.76. Interestingly, the best ICC value was that of a 3-tiered classification (St Gallen 2009)[26] for CNB specimens. In keeping with the lower ICC values for any Ki67 determination (than for ER or PR staining), the ONEST analysis also suggested higher maximal differences between 2 observers (up to 34%), lower OPAs with all observers (26% as minimum), and higher number of pathologists required to reflect reproducibility (mostly 5). The two-tiered systems of St Gallen recommendations from 2011 [27] and 2013 [28] showed better results (lower maximum differences between 2 observers and higher OPAs for all observers).

It is evident from improved ICC values reported by the International Ki67 in Breast Cancer Working Group, that scoring consistency of Ki67 can also be improved by standardized reporting, even without image analysis [32], and standardization is the way forward to achieve reliable Ki67 assessments. However, this study was not devised to increase reproducibility, but reproducibility was described as basic data, and the analysis was complemented by the newly developed ONEST method, to see what this can add to studies of reproducibility in case of biomarkers deemed suitable for prognostic or predictive conclusions. As hypothesized, ONEST can complement conventional statistics of agreement. It can prove or simply visualize that a biomarker is reliable, due to its easy assessment and natural distribution (like ER in our series; high plots with narrow bandwidth, Figure 4A). It can also highlight weaknesses of biomarker assessment (high interrater differences, i.e., wide bandwidth between the top and the bottom curves, and low OPA values with all observers included, Figure 4C). This is in addition to the original aim of ONEST to determine the number of observers needed for the plot to reach a kind of plateau, i.e., the number minimally required to reliably reflect reproducibility. In this context, the results of some earlier reports may be challenged on the basis of the number of observers involved, including a work from the Institute of Pathology, University of Szeged [75]. In the referred study, only three observers were included for the categorization of Ki67 staining according to the St Gallen 2009 criteria, whereas the current ONEST analysis would suggest at least 4, preferably 5 for reliable estimations.

5.2. ONEST FOR HISTOLOGICAL GRADE

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 multivariable analysis derived combined prognosticators [12, 46-48], 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 [71]. 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 [71]. Overall, we found that the reproducibility of grading was moderate (kappa values >0.50, but <0.6; Table 4) or good to moderate (ICC values 0.687-0.700, Table 5). 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 (Figure 6). 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 [55-73]. In 1994, Dalton et al have assumed that virtually all pathologists should be able to adequately grade breast cancers [73]. 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 7) [70]. Following a Dutch nationwide study documenting relevant inter- and intradepartmental variations in the distribution of histological grades [77], both anonymized specific feedback to the laboratories and pathologists [78] and e-learning [79] have helped to decrease this variation. Not least, optimal tissue preservation is also required for adequate grading, better fixation has been documented to result in better grading consistency [58]. When looking at national databases or greater cohorts of patients, it also appears that the distribution 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 33043 patients [77], whereas data from the Survival Epidemiology and End Results database suggest only 21% of 746507 breast cancers being of this grade [80], 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 [48]. 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 [81], 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 [40, 49]. Our analysis suggested that a minimum of 4 to 7 observers are needed to adequately reflect reproducibility both for the 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 to 6 raters are somewhat better than those gained with more observers (Table 8). Studies reporting kappa statistics reflect the same trend (Table 7). Many studies on the reproducibility of grading have used Cohen’s kappa, which is devised for 2 observers [52], therefore pairwise comparisons were made, and the range or average was reported (Table 7), 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 [69]. Fleiss has also devised a kappa coefficient for multiple observers [52], and this has been used in several studies with more than two raters (Table 7), and seems more appropriate in this setting.

5.3. FURTHER CONSIDERATIONS

During our work on tumor infiltrating lymphocytes (TILs) in progress and further analysis of ONEST as a method to highlight some aspects of reproducibility for subjective tests, we have identified a number of factors that may influence the results of this analysis, and it is worth to mention them at the end of this thesis.

Conclusions from ONEST plots can be influenced by the number and experience of the observers, and the elimination of observers with substantial divergence from the others can “improve” the results, but biases real-life expectations. Indeed, in real life, not all observers have the same skills, and if one wishes to have a reflection of reproducibility, divergent classifiers should not be ignored. Further to factors identified previously, like the number of categories in the classification, or the distribution of the variables around and away from the extremes, heterogeneity in distribution can also impact on the ONEST results, just like on other measures of reproducibility.

In the publications forming the basis of the thesis, we used ONEST values read from the minimum OPACs leveling off, i.e., approaching the horizontal, because approaching the plateau with a minimal slope may also yield a sufficient approximation of the ONEST value. In the thesis, this was modified with the integration of the ONEST values that coincide with the value at which the plateau of the minimum OPAC is reached, and this is how the publicly available software was also developed [82].

6. CONCLUSIONS

In summary, we have applied ONEST for characterizing the reproducibility of three biomarkers, ER, PR and Ki67, all evaluated by estimating the proportion of immunostained nuclei on CNB and EXC specimens. The differences in reproducibility were mainly explained by the distribution of the stained nuclei around or away from the extremes (0% and 100%). ONEST gave useful supplementary information and its plots helped in visualizing the results. The minimum OPA values, the greatest difference in OPA for 2 pathologists (bandwidth) and the OPA for all pathologists, i.e., OPA(n), are all reflected in ONEST plots.

The number of observers required for the reliable estimation of reproducibility was 2 for ER and 3 for PR categorization, and ranged between 4-6 for the various Ki67 categorizations.

Considering our ONEST analyses, it is suggested that a minimum of 4, preferably 6-7 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 [71], but can be diminished with training, feedback and dedicated assessment.

ONEST, like other measures of reproducibility, is also dependent on a number of factors which may influence its results. These include the number of categories in the classification (two-tiered vs three-tiered classifications), the distribution of the parameters assessed around or away from the extremes, homogeneity in distribution, number and experience of observers, the presence of outliers with substantially divergent classification from the others. Therefore, ONEST should also be regarded as an estimation and a complementary tool for reproducibility studies.

7. ACKNOWLEDGEMENTS

First and foremost, I would like to thank my father and supervisor, Prof. Dr. Gábor Cserni, whose immense support and guidance was irreplaceable for the writing of this thesis. Thank you, father!

I thank all my co-authors, without whom this study would not have been possible. I would like to especially emphasize the help provided by Prof. Dr. Tibor Nyári, whose expertise in statistics was essential for many of our calculations.

I express my special thanks to the author of the Real Statistics Resource Pack Excel add-in, Charles Zaiontz. His tool proved very useful for our statistical calculations, and he also helped us multiple times with our questions in regard to statistics or the usage of the tool.

Development of the program used for the calculations depended on the wxWidgets library and the lest test framework, both for which I am grateful to their authors.

I am grateful to all my teachers. From the University of Szeged, I would like to especially thank my professors, instructors and tutors at the Institute of Informatics, specifically emphasizing the work of Dr. Antal Nagy, Dr. Rudolf Ferenc, Prof. Dr. Tibor Gyimóthy, Dr. István Siket and Dr. Tamás Vinkó. I would also like to express my gratitude to my teachers at Kecskemét Katona József Secondary Grammar School, most notably to my math teacher, Dr. Szablics Bálintné, and to my physics/chemistry teacher, Irén Sáróné Jéga-Szabó, as it is their preparation that allowed me to take the first obstacles at my university studies with ease.

I am thankful to my family and friends, who all supported me throughout this journey.

The publication of the articles forming the basis of the thesis were supported by the University of Szeged Open Access Fund - Grants No 5440 and 5580.

8. REFERENCES

1. International Agency for Research on Cancer. Global Cancer Observatory https://gco.iarc.fr (Last accessed 12. Sept. 2022.)

2. Europa Donna. Breast Cancer Facts. https://www.europadonna.org/breast-cancer/ (Last accessed 13. Sept. 2022.)

3. WHO Classification of Tumours Editorial Board (Eds.). WHO Classification of Tumours. – Breast Tumours, 5th ed., International Agency for Research on Cancer, Lyon, 2019.

4. Cserni G. Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica* 2020;112:25-41.

5. Fitzgibbons PL, Page DL, Weaver D, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 2000;124:966-978.

6. Cserni G, Quinn CM, Foschini MP, et al. Triple-Negative Breast Cancer Histological Subtypes with a Favourable Prognosis. *Cancers* 2021;13:5694.

7. Ellis I, Allison KH, Dang C, et al. *Invasive Carcinoma of the Breast Histopathology Reporting Guide, 2nd edition*. International Collaboration on Cancer Reporting; Sydney, 2022 https://www.iccr-cancer.org/datasets/published-datasets/breast/invasive-carcinoma-of-the-breast/ (Last accessed 02 October 2022)

8. Ellis IO, Al-Sam S, Anderson N, et al. 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 (Last accessed 02 October 2022)

9. Cserni G, Francz M, Járay B, et al. Pathological diagnosis, work-up and reporting of breast cancer. Recommendations from the 4th Breast Cancer Consensus Conference [In Hungarian]. Magy Onkol 2020;64:301-328. (http://huon.hu/2020/64/4/0301/0301a.pdf, Last accessed 21 May 2021)

10. Cserni G, Francz M, Járay B, 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.

11. Brierley JD, Gospodarowicz MK, Wittekind C (Eds). Union for International Cancer Control TNM classification of malignant tumours, 8th ed, Wiley Blackwell, Hoboken, New Jersey, 2017, pp272.

12. Hortobagyi G, Connolly JL, D'Orsi CJ, et al. AJCC Cancer Staging Manual, 8th ed., Springer, New York, 2017, pp. 587-628.

13. Lyman GH, Giuliano AE, Somerfield MR, et al. American Society of Clinical Oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer. J Clin Oncol 2005;23:7703-7720.

14. Nesbit EG, Donnelly ED, Strauss JB. Treatment Strategies for Oligometastatic Breast Cancer. Curr Treat Options Oncol 2021;22:94.

15. Elston CW, Ellis IO. 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. 1991;19:403-410.

16. Denkert C, von Minckwitz G, Darb-Esfahani S, 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.

17. Loi S, Drubay D, Adams 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.

18. Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Guideline Update. Arch Pathol Lab Med 2020;144:545-563.

19. Dowsett M, Allred C, Knox J, et al. Relationship between quantitative estrogen and progesterone receptor expression and human epidermal growth factor receptor 2 (HER-2) status with recurrence in the Arimidex, Tamoxifen, alone or in combination trial. J Clin Oncol 2008;26:1059–1065.

20. Hammond ME, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology / College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Arch Pathol Lab Med 2010;134:e48-72.

21. Fei F, Siegal GP, Wei S. Characterization of estrogen receptor-low-positive breast cancer. Breast Cancer Res Treat 2021;188:225-235.

22. Allred DC, Harvey JM, Berardo M, et al. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. Mod Pathol 1998;11:155-168.

23. Harvey JM, Clark GM, Osborne CK, et al. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol 1999;17:1474-1481.

24. Baird RD, Carroll JS. Understanding Oestrogen Receptor Function in Breast Cancer and its Interaction with the Progesterone Receptor. New Preclinical Findings and their Clinical Implications. Clin Oncol (R Coll Radiol) 2016;28:1-3.

25. Yerushalmi R, Woods R, Ravdin PM, et al. Ki67 in breast cancer: prognostic and predictive potential. Lancet Oncol 2010;11:174-183.

26. Goldhirsch A, Ingle JN, Gelber RD, et al. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol 2009;20:1319-1329.

27. Goldhirsch A, Wood WC, Coates AS, et al. Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 2011;22:1736-1747.

28. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 2013;24:2206-2223.

29. Coates AS, Winer EP, Goldhirsch A, et al. Tailoring therapies-improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol 2015;26:1533-1546.

30. Cuzick J, Dowsett M, Pineda S, et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2

immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. J Clin Oncol 2011;29:4273-4278.

31. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. J Natl Cancer Inst 2011;103:1656-1664.

32. Polley MY, Leung SC, Gao D, et al. An international study to increase concordance in Ki67 scoring. Mod Pathol 2015;28:778–786.

33. Arun I, Venkatesh S, Ahmed R, et al. Reliability of Ki67 visual scoring app compared to eyeball estimate and digital image analysis and its prognostic significance in hormone receptor-positive breast cancer. APMIS 2021;129:489-502.

34. Tuominen VJ, Ruotoistenmäki S, Viitanen A, et al. ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. Breast Cancer Res 2010;12:R56.

35. Ács B, Madaras L, Kovács KA, et al. Reproducibility and Prognostic Potential of Ki-67 Proliferation Index when Comparing Digital-Image Analysis with Standard Semi-Quantitative Evaluation in Breast Cancer. Pathol Oncol Res 2018;24:115-127.

36. Ács B, Pelekanou V, Bai Y, et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. Lab Invest 2019;99:107-117.

37. Cai L, Yan K, Bu H, et al. Improving Ki67 Assessment Concordance with AI-Empowered Microscope: A Multi-institutional Ring Study. Histopathology 2021;79:544-555.

38. Varga Z, Cassoly E, Li Q, et al. Standardization for Ki-67 assessment in moderately differentiated breast cancer. A retrospective analysis of the SAKK 28/12 study. PLoS One 2015;10:e0123435.

39. Wells CA, Sloane JP, Coleman D, et al. Consistency of staining and reporting of oestrogen receptor immunocytochemistry within the European Union--an inter-laboratory study. Virchows Arch 2004;445:119-128.

40. Reisenbichler ES, Han G, Bellizzi A, 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.

41. Scarff RW, Torloni H. Histological Typing of Breast Tumours, first ed., World Health Organization, Geneva, 1968.

42. Patey DH, Scarff RW. The position of histology in the prognosis of carcinoma of the breast. Lancet 1928;211(5460):801-804.

43. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 1957;11:359-377.

44. Amendoeira I, Apostolikas N, Bellocq JP, et al. Quality assurance guidelines for pathology, in: Perry N, Broeders M, de Wolf C, et al. (Eds.), European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis, fourth ed., European Commission, Luxemburg, 2006, pp. 219-311.

45. Van Dooijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. Virchows Arch 2022;480:33–43.

46. Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. Br J Cancer 1982;45:361-366.

47. Ravdin PM, Siminoff LA, Davis GJ, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001;19:980-991.

48. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res 2010;12:R1.

49. Cserni B, Bori R, Csörgő E, 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.

50. Cserni B, Bori R, Csörgő E, 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 2022;229:153718.

51. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155-163.

52. Fleiss JL. Statistical Methods for Rates and Proportions, second ed., John Wiley and Sons, New York, 1980.

53. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

54. Zaiontz C. Real Statistics Resource Pack | Real Statistics Using Excel (https://real-statistics.com) (Accessed 23 October 2021)

55. Jacquemier J, Charpin C. Reproducibility of histoprognostic grades of invasive breast cancer [in French]. Ann Pathol 1998;18:385-390.

56. Sikka M, Agarwal S, Bhatia A. Interobserver agreement of the Nottingham histologic grading scheme for infiltrating duct carcinoma breast. Indian J Cancer 1999;36:149–153.

57. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod Pathol 2005;18:1067-1078.

58. Robbins P, Pinder S, de Klerk N. Histological grading of breast carcinomas: a study of interobserver agreement. Hum Pathol 1995;26:873–879.

59. Anderson TJ, Alexander FE, Lamb J, et al. Pathology characteristics that optimize outcome prediction of a breast screening trial. Br J Cancer 2000;83:487–492.

60. Frierson HF Jr, Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol 1995;103:195-198.

61. Rabe K, Snir OL, Bossuyt V, et al. Interobserver variability in breast carcinoma grading results in prognostic stage differences. Hum Pathol 2019;94:51-57.

62. Ginter PS, Idress R, D'Alfonso TM, et al. Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy. Mod Pathol 2021;34:701-709.

63. Postma EL, Verkooijen HM, van Diest PJ, et al. Discrepancy between routine and expert pathologists’ assessment of non-palpable breast cancer and its impact on locoregional and systemic treatment. Eur J Pharmacol 2013;717:31–35.

64. Reed W, Hannisdal E, Boehler PJ, et al. 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 2000;88:804–813.

65. Bueno-de-Mesquita JM, Nuyten DS, Wesseling J, et al. 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 2010;21:40-47.

66. Cserni G, Kocsis L, Serényi P. Grading of invasive breast cancers using the Nottingham modification of the Bloom and Richardson scheme. Study on reproducibility [in Hungarian]. Magyar Onkol 1996;40:188-191.

67. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer--reproducibility between seven pathologic departments. Acta Oncol 2000;39: 41–45.

68. Longacre TA, Ennis M, Quenneville LA, et al. Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. Mod Pathol 2006;19:195-207.

69. Sloane JP, Amendoeira I, Apostolikas N, et al. Consistency achieved by 23 European pathologists from 12 countries in diagnosing breast disease and reporting prognostic features of carcinomas. Virchows Arch 1999;434:3-10.

70. Ellis IO, Coleman D, Wells C, et al. Impact of a national external quality assessment scheme for breast pathology in the UK. J Clin Pathol 2006;59:138–145.

71. Rakha EA, Bennett RL, Coleman D, et al. Review of the national external quality assessment (EQA) scheme for breast pathology in the UK. J Clin Pathol 2017;70:51-57.

72. Theissig F, Kunze KD, Haroske G, et al. Histological grading of breast cancer. Interobserver, reproducibility and prognostic significance. Pathol Res Pract 1990;186:732-736.

73. Dalton LW, Page DL, Dupont DW. Histologic grading of breast carcinoma. A reproducibility study. Cancer 1994;73:2765-2770.

74. Pu T, Shui R, Shi J et al. External quality assessment (EQA) program for the immunohistochemical detection of ER, PR and Ki-67 in breast cancer: results of an interlaboratory reproducibility ring study in China. BMC Cancer 2019;19:978.

75. Vörös A, Csörgő E, Nyári T, et al. An intra- and interobserver reproducibility analysis of the Ki-67 proliferation marker assessment on core biopsies of breast cancer patients and its potential clinical implications. Pathobiology 2013;80:111-118.

76. Cserni G, Vörös A, Liepniece-Karele I, et al. Distribution pattern of the Ki67 labelling index in breast cancer and its implications for choosing cut-off values. Breast 2014;23:259-263.

77. van Dooijeweert C, van Diest PJ, Willems SM, et al. 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 2020;146:769-80.

78. van Dooijeweert C, van Diest PJ, Baas IO, et al. 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 2020;73:793-799.

79. van Dooijeweert C, Deckers IAG, de Ruiter EJ, et al. The effect of an e-learning module on grading variation of (pre)malignant breast lesions. Mod Pathol 2020;33:1961-1967.

80. Sopik V, Narod SA. The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer. Breast Cancer Res Treat 2018;170:647-656.

81. Dawson SJ, Duffy SW, Blows FM, et al. Molecular characteristics of screen-detected vs symptomatic breast cancers and their impact on survival. Br J Cancer 2009;101:1338-1344.

82. Cserni B. ONEST Calculator. Available online: https://github.com/csernib/onest (accessed on 12. November 2022.)

APPENDIX I: MAJOR NEW FINDINGS

1. ONEST (Observers Need to Evaluate Subjective Tests) is a recently developed statistical tool to reflect how many observers might be needed to reliably assess reproducibility of a subjective test. ONEST has been used for the first time to evaluate ER, PR, Ki67 status and histological grade of breast carcinomas, suggesting that the number of pathologists involved in previous studies of reproducibility might be suboptimal in some.

2. Rather than only giving a number beyond which the number of observers (ONEST value) has no or minimal impact on the reliability of assessing reproducibility, ONEST also gives visual impressions of reproducibility, by showing the greatest difference between the ratings of two observers (bandwidth), the OPA that all observers can reach and the slope of the minimum OPA curve.

3. The terminology of the ONEST readings (ONEST value, bandwidth, OPA(n), OPAC, minimum OPAC) has been developed gradually by the author to better allow reference to these values.

4. ONEST has been originally described as a determination of the ONEST value based on 100 permutations of pathologists estimating one specific parameter. In our analysis of nuclear biomarker immunostains with 9 observers, we have demonstrated that results gained from the 100 permutations do not differ from results gained from all (9!) permutations, i.e., using 100 permutations is a suitable and fast way to make ONEST analyses and estimations. (The developed software allows for both 100 and all permutations to be taken into account)

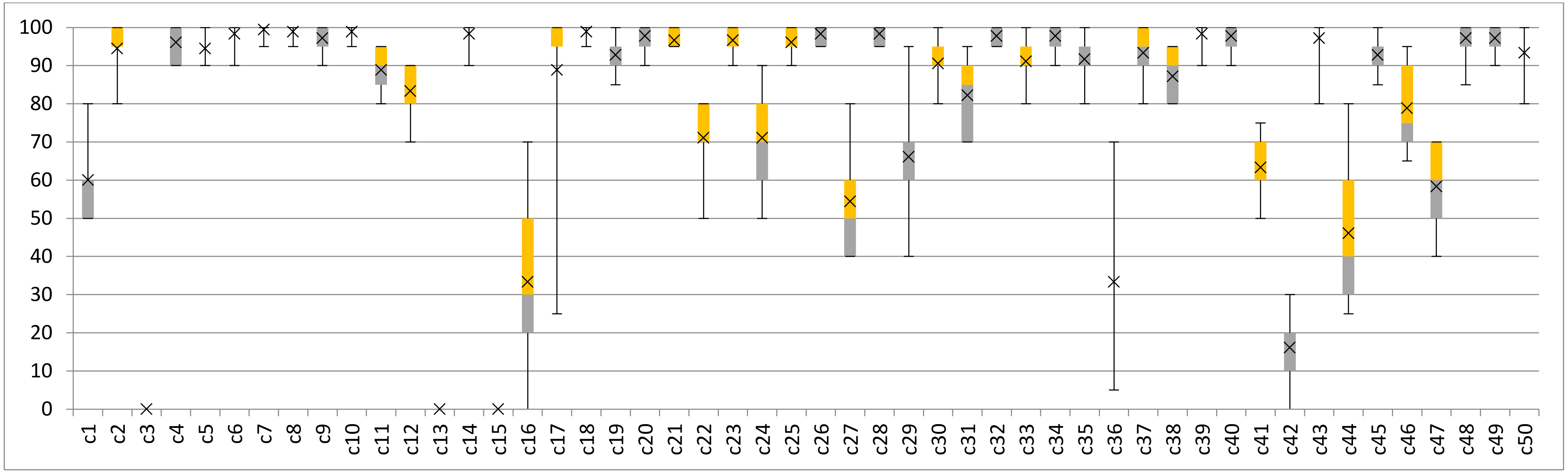
5. To allow a wider use of the method, a publicly available ONEST calculator was also developed.

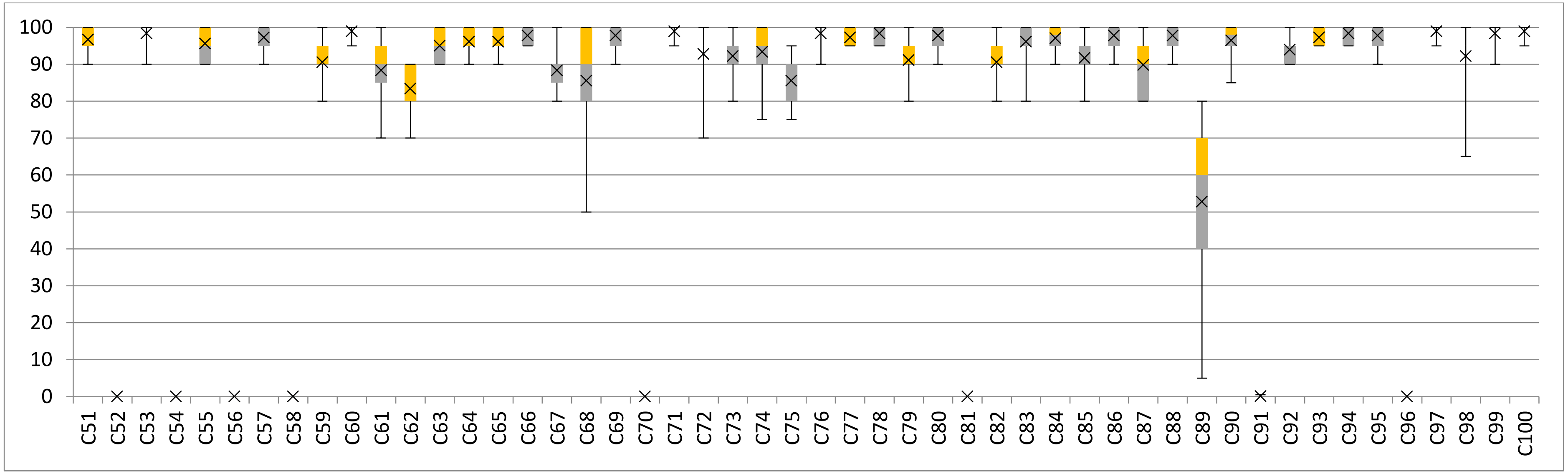
APPENDIX II: SUPPLEMENTARY FIGURES AND TABLES

Supplementary figure 1: Boxplots of ER, PR and Ki67 values as rated by the 9 observers in CNB (c1-c50) and EXC (c51-c100) cases

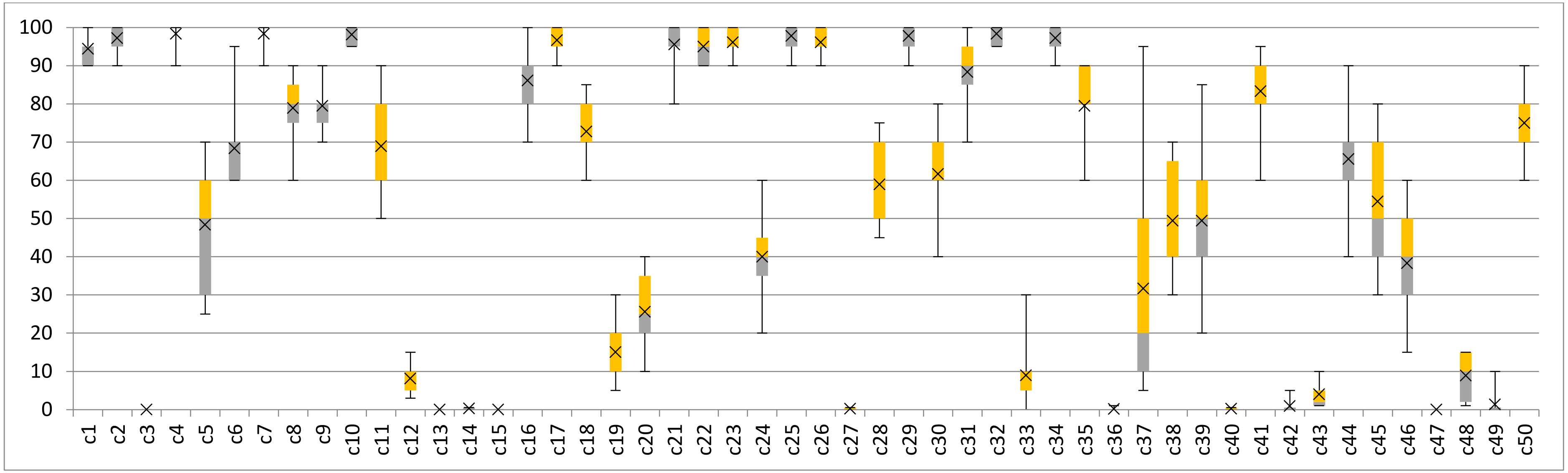
The boxes have an upper yellow part with the top representing the median value of the upper (2nd) half of the data (3rd quartile, Q3), the lower grey part with the bottom representing the median value of the lower (1st) half of the data (1st quartile, Q1) and the transition between the two parts representing the median value of all data. The box itself gives the interquartile range (IQR=Q3-Q1). The x symbols refer to the mathematical mean (average). Top whiskers refer to maximum and bottom whiskers to minimum values. For some cases, parts of the boxplot are overlapping, and therefore, are not visualized separately.

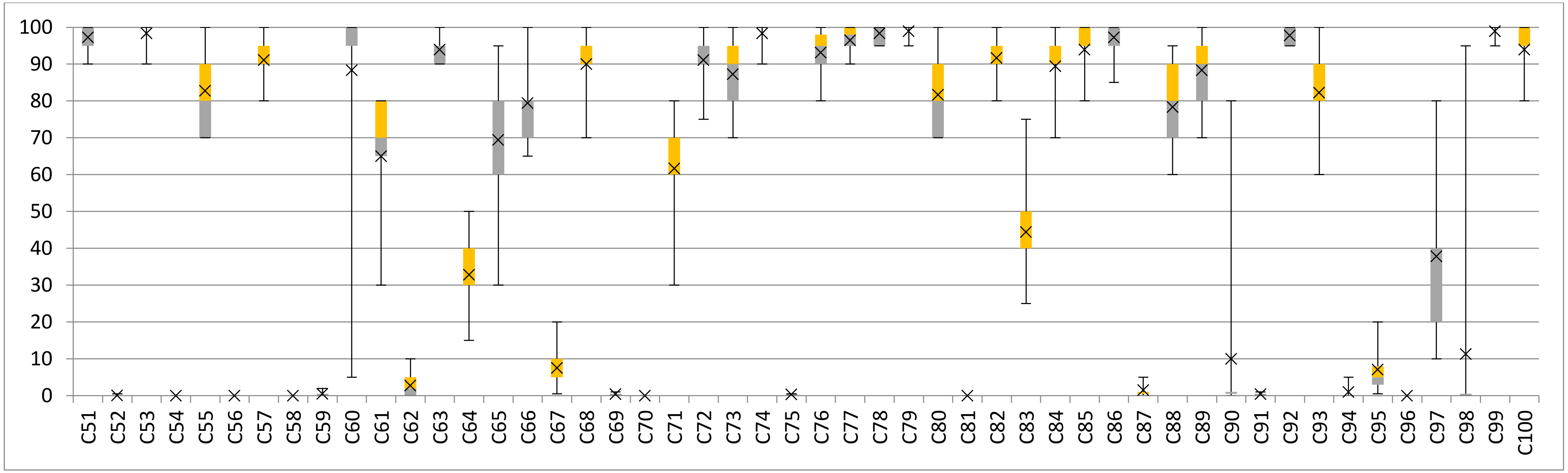
ER


100
90
80
70
60
50
40
30
20
10
0
c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29 c30 c31 c32 c33 c34 c35 c36 c37 c38 c39 c40 c41 c42 c43 c44 c45 c46 c47 c48 c49 c50

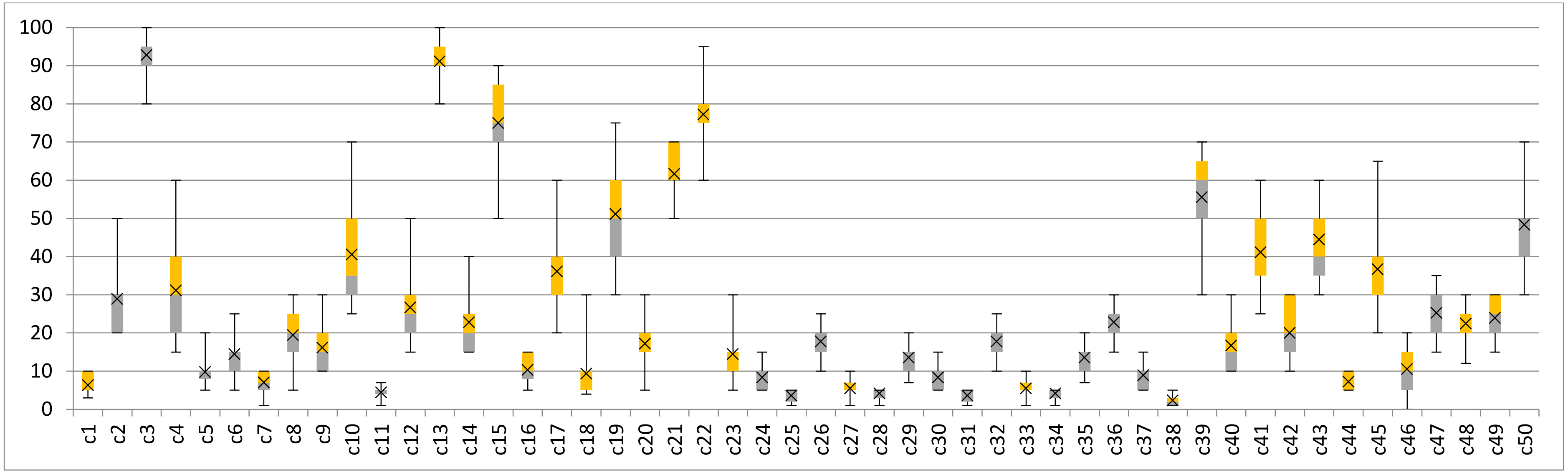

100
90
80
70
60
50
40
30
20
10
0
C51 C52 C53 C54 C55 C56 C57 C58 C59 C60 C61 C62 C63 C64 C65 C66 C67 C68 C69 C70 C71 C72 C73 C74 C75 C76 C77 C78 C79 C80 C81 C82 C83 C84 C85 C86 C87 C88 C89 C90 C91 C92 C93 C94 C95 C96 C97 C98 C99 C100

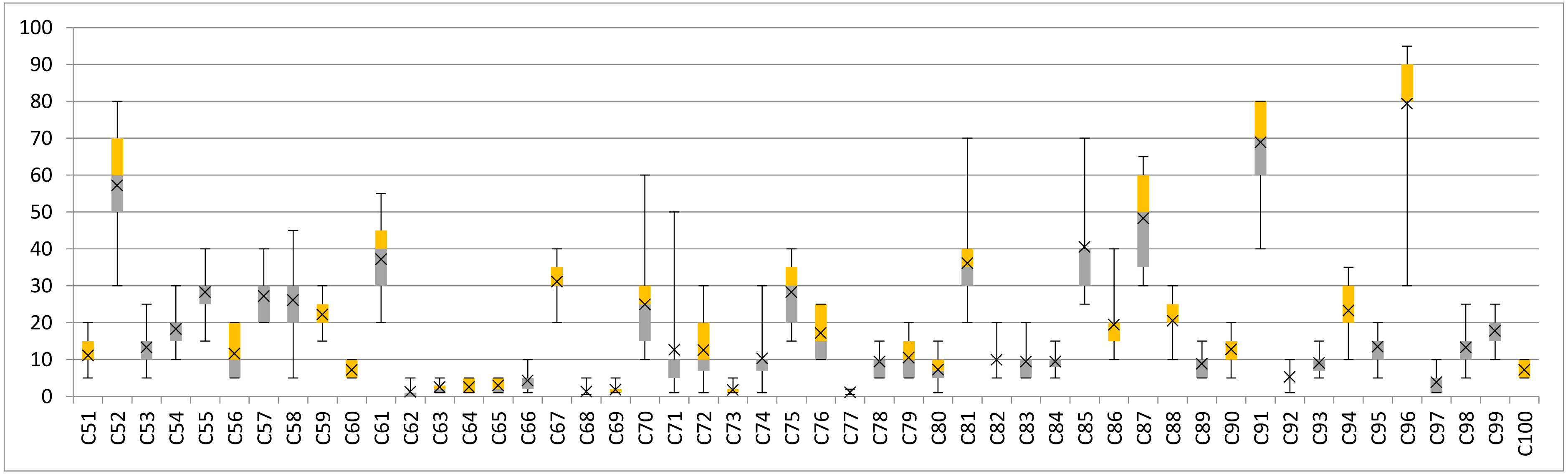
PR


100
90
80
70
60
50
40
30
20
10
0
c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29 c30 c31 c32 c33 c34 c35 c36 c37 c38 c39 c40 c41 c42 c43 c44 c45 c46 c47 c48 c49 c50


100
90
80
70
60
50
40
30
20
10
0
C51 C52 C53 C54 C55 C56 C57 C58 C59 C60 C61 C62 C63 C64 C65 C66 C67 C68 C69 C70 C71 C72 C73 C74 C75 C76 C77 C78 C79 C80 C81 C82 C83 C84 C85 C86 C87 C88 C89 C90 C91 C92 C93 C94 C95 C96 C97 C98 C99 C100

Ki67


100
90
80
70
60
50
40
30
20
10
0
c1
c2
c3
c4
c5
c6
c7
c8
c9
c10
c11
c12
c13
c14
c15
c16
c17
c18
c19
c20
c21
c22
c23
c24
c25
c26
c27
c28
c29
c30
c31
c32
c33
c34
c35
c36
c37
c38
c39
c40
c41
c42
c43
c44
c45
c46
c47
c48
c49
c50


100
90
80
70
60
50
40
30
20
10
0
C51
C52
C53
C54
C55
C56
C57
C58
C59
C60
C61
C62
C63
C64
C65
C66
C67
C68
C69
C70
C71
C72
C73
C74
C75
C76
C77
C78
C79
C80
C81
C82
C83
C84
C85
C86
C87
C88
C89
C90
C91
C92
C93
C94
C95
C96
C97
C98
C99
C100

Supplementary Table 1

Agreement on the three-category-classification (positive, weakly positive and negative) of the steroid receptor statuses and analogous Ki67 distribution

	CNB	CNB	CNB	EXC	EXC	EXC
	ER status	PR status	Ki67 "status"	ER status	PR status	Ki67 "status"
Majority opinion: positive (>10%)						
100% agreement	44	33	23	41	30	12
<100% agreement (% range)	3 (78-89%)	2 (67-78%)	11 (67-89%)	1 (89%)	2 (89%)	11 (56-89%)
Majority opinion: low positive (1-10%)						
100% agreement	0	1	8	0	0	10
<100% agreement (% range)	0	4 (56-78%)	8 (56-89%)	0	2+2x0.5 (44-89%)*	17 (56-89%)
Majority opinion: negative (<1%)						
100% agreement	3	7	0	8	8	0
<100% agreement (% range)	0	3 (89%)	0	0	6+2x0.5 (44-89%)*	0

CNB: core needle biopsy samples, EXC: excision samples; * There were 2 cases with equal (4 and 4) ratings into negative and low positive with an additional one into positive.

Supplementary Table 2

Agreement on the different Ki67 categorization according to consecutive St Gallen consensus conference recommendations

CNB	St Gallen 2009	St Gallen 2011	St Gallen 2013	St Gallen 2015
Cut-offs	(>30%, 16-30%, <=15%)	(>=14%, <14%)	(>=20%, <20%)	(>=25%, 6-24%, <=5%)
Ki67 High				
100% agreement	5	22	10	12
<100% agreement (% range)	7 (66-89%)	12 (67-89%)	11 (56-89%)	9 (56-89%)
Ki67 Intermediate				
100% agreement	0	NA	NA	2
<100% agreement (% range)	13 (56-89%)	NA	NA	19 (56-78%)
Ki67 Low				
100% agreement	11	8	11	3
<100% agreement (% range)	14 (44-89%)	8 (56-89%)	13 (56-89%)	5 (56-89%)
EXC				
Ki67 High				
100% agreement	1	12	9	5
<100% agreement (% range)	6 (56-89%)	11 (67-89%)	9 (56-89%)	8 (67-89%)
Ki67 Intermediate				
100% agreement	0	NA	NA	8
<100% agreement (% range)	10 (56-89%)	NA	NA	24 (56-89%)
Ki67 Low				
100% agreement	18	14	19	0
<100% agreement (% range)	15 (44-89%)	13 (56-89%)	13 (56-89%)	5 (56-89%)

NA: not applicable

Supplementary Table 3 – Minimum, maximum and median of the OPAs for 100 random permutations of 9 pathologists

A. ER, PR and Ki67 (<1%, 1-10% and >10% categories) on CNB and EXC specimens

Observers	Minimum	Maximum	Median	Minimum	Maximum	Median
	ER CNB			ER EXC		
2	0.96	1.00	0.98	0.98	1.00	1.00
3	0.96	1.00	0.98	0.98	1.00	1.00
4	0.96	1.00	0.96	0.98	1.00	1.00
5	0.96	1.00	0.96	0.98	1.00	0.99
6	0.96	0.98	0.96	0.98	1.00	0.98
7	0.96	0.98	0.96	0.98	1.00	0.98
8	0.96	0.96	0.96	0.98	1.00	0.98
9	0.96	0.96	0.96	0.98	0.98	0.98
	PR CNB			PR EXC		
2	0.86	0.98	0.92	0.80	0.98	0.92
3	0.82	0.94	0.88	0.76	0.94	0.86
4	0.82	0.94	0.86	0.76	0..94	0.82
5	0.82	0.90	0.84	0.76	0.88	0.80
6	0.82	0.88	0.84	0.76	0.84	0.78
7	0.82	0.86	0.82	0.76	0.82	0.78
8	0.82	0.84	0.82	0.76	0.78	0.76
9	0.82	0.82	0.82	0.76	0.76	0.76
	Ki67 CNB			Ki67 EXC		
2	0.76	0.94	0.86	0.64	0.90	0.80
3	0.70	0.88	0.78	0.52	0.82	0.68
4	0.64	0.84	0.72	0.48	0.74	0.60
5	0.62	0.78	0.70	0.44	0.70	0.58
6	0.62	0.76	0.68	0.44	0.66	0.52
7	0.62	0.72	0.66	0.44	0.60	0.50
8	0.62	0.66	0.64	0.44	0.52	0.46
9	0.62	0.62	0.62	0.44	0.44	0.44

B. ER and PR intensity and Allred scores for CNB and EXC specimens

Observers	Minimum	Maximum	Median	Minimum	Maximum	Median
	ER intensity (0-1-2-3) CNB			ER intensity (0-1-2-3) EXC		
2	0.58	0.90	0.78	0.50	0.86	0.74
3	0.54	0.82	0.68	0.42	0.72	0.56
4	0.52	0.76	0.62	0.40	0.64	0.50
5	0.50	0.68	0.56	0.38	0.58	0.44
6	0.48	0.64	0.53	0.38	0.52	0.42
7	0.48	0.58	0.50	0.38	0.50	0.40
8	0.48	0.54	0.50	0.38	0.42	0.38
9	0.48	0.48	0.48	0.38	0.38	0.38
	ER Allred scores (0,2; 3-4; 5-6; 7-8) CNB			ER Allred scores (0,2; 3-4; 5-6; 7-8) EXC		
2	0.82	0.94	0.88	0.90	1.00	0.96
3	0.76	0.90	0.82	0.90	0.98	0.94
4	0.72	0.86	0.80	0.90	0.96	0.92
5	0.72	0.84	0.78	0.90	0.96	0.92
6	0.72	0.82	0.74	0.90	0.94	0.90
7	0.72	0.80	0.72	0.90	0.94	0.90
8	0.72	0.78	0.72	0.90	0.92	0.90
9	0.72	0.72	0.72	0.90	0.90	0.90
	PR intensity (0-1-2-3) CNB			PR intensity (0-1-2-3) EXC		
2	0.52	0.88	0.66	0.50	0.92	0.68
3	0.42	0.70	0.54	0.44	0.74	0.54
4	0.40	0.64	0.48	0.40	0.64	0.48
5	0.38	0.58	0.44	0.38	0.52	0.44
6	0.38	0.50	0.40	0.38	0.46	0.40
7	0.38	0.44	0.40	0.36	0.44	0.38
8	0.38	0.40	0.38	0.36	0.40	0.38
9	0.38	0.38	0.38	0.36	0.36	0.36
	PR Allred scores (0,2; 3-4; 5-6; 7-8) CNB			PR Allred scores (0,2; 3-4; 5-6; 7-8) EXC		
2	0.64	0.86	0.76	0.70	0.90	0.80
3	0.54	0.74	0.64	0.60	0.82	0.74
4	0.52	0.66	0.58	0.58	0.76	0.70
5	0.50	0.62	0.54	0.58	0.72	0.66
6	0.48	0.58	0.52	0.58	0.68	0.62
7	0.48	0.56	0.50	0.58	0.64	0.60
8	0.48	0.52	0.48	0.58	0.62	0.58
9	0.48	0.48	0.48	0.58	0.58	0.58

C. Ki67 values by different St Gallen recommendations in CNB and EXC specimens

Observers	Minimum	Maximum	Median	Minimum	Maximum	Median
	Ki67 (as ER and PR: <1%, 1-10%, >10%) CNB			Ki67 (as ER and PR: <1%, 1-10%, >10%) EXC		
2	0.76	0.94	0.86	0.64	0.90	0.80
3	0.68	0.88	0.78	0.52	0.82	0.68
4	0.64	0.84	0.72	0.48	0.74	0.60
5	0.62	0.80	0.70	0.44	0.70	0.58
6	0.62	0.76	0.68	0.44	0.66	0.52
7	0.62	0.72	0.66	0.44	0.60	0.50
8	0.62	0.66	0.64	0.44	0.52	0.46
9	0.62	0.62	0.62	0.44	0.44	0.44
	Ki67 (StGallen 2009: <16%,16-30%,>30%) CNB			Ki67 (StGallen 2009: <16%,16-30%,>30%) EXC		
2	0.54	0.84	0.74	0.60	0.88	0.74
3	0.38	0.72	0.56	0.46	0.78	0.62
4	0.32	0.64	0.48	0.40	0.66	0.56
5	0.32	0.56	0.44	0.38	0.64	0.50
6	0.32	0.48	0.40	0.38	0.60	0.46
7	0.32	0.46	0.38	0.38	0.52	0.44
8	0.32	0.38	0.32	0.38	0.46	0.38
9	0.32	0.32	0.32	0.38	0.38	0.38
	Ki67 (StGallen 2011: <14%. >=14%) CNB			Ki67 (StGallen 2011: <14%. >=14%) EXC		
2	0.76	0.94	0.86	0.66	0.90	0.82
3	0.70	0.88	0.78	0.56	0.84	0.72
4	0.64	0.84	0.72	0.52	0.76	0.64
5	0.62	0.78	0.69	0.50	0.74	0.62
6	0.60	0.74	0.66	0.50	0.70	0.58
7	0.60	0.70	0.64	0.50	0.66	0.56
8	0.60	0.64	0.62	0.50	0.58	0.52
9	0.60	0.60	0.60	0.50	0.50	0.50
	Ki67 (StGallen 2013: <20%. >=20%) CNB			Ki67 (StGallen 2013: <20%. >=20%) EXC		
2	0.68	0.90	0.82	0.66	0.92	0.84
3	0.60	0.82	0.72	0.60	0.88	0.76
4	0.56	0.78	0.66	0.60	0.80	0.68
5	0.54	0.70	0.62	0.54	0.80	0.66
6	0.52	0.66	0.58	0.54	0.74	0.64
7	0.52	0.62	0.56	0.54	0.68	0.60
8	0.52	0.58	0.54	0.54	0.60	0.58
9	0.52	0.52	0.52	0.54	0.54	0.54
	Ki67 (StGallen 2015: <6%. 6-24%. >24%) CNB			Ki67 (StGallen 2015: <6%. 6-24%. >24%) EXC		
2	0.54	0.84	0.68	0.46	0.8	0.70
3	0.40	0.66	0.54	0.36	0.66	0.54
4	0.34	0.58	0.44	0.3	0.56	0.44
5	0.32	0.54	0.39	0.26	0.5	0.38
6	0.32	0.42	0.36	0.26	0.44	0.32
7	0.32	0.40	0.34	0.26	0.36	0.30
8	0.32	0.34	0.34	0.26	0.3	0.28
9	0.32	0.32	0.32	0.26	0.26	0.26

APPENDIX III: MAGYAR NYELVŰ ÖSSZEFOGLALÓ

A tézis alapját adó két közleményben egy viszonylag új, ONEST (Observers Needed to Evaluate Subjective Tests; Szubjektív Tesztek Kiértékeléséhez Szükséges Vizsgálók; *Reisenbichler ES et al. Mod Pathol 2020;33:1746-1752*) nevű statisztikai módszert alkalmaztunk az ösztrogén receptor (ER), progeszteron receptor (PR), Ki67, a hisztológiai grade, illetve a grade egyes komponenseinek szubjektív értékelésekor meghatározott besorolás reprodukálhatóságának vizsgálatára. Az értékelést 9 különböző tapasztalattal rendelkező patológus végezte, akik 50 hengerbiopsziás (CNB), illetve 50 excíziós (EXC) mintán, az odavonatkozó nemzetközi ajánlások alapján pontozták az ER, PR és Ki67 értékeket. A grade és komponenseinek értékelése hasonló módon történt, szintén 9 patológussal, szintén 50 CNB, illetve 50 EXC mintán.

A 9 patológus értékeléséből származó eredményeken elvégeztük az ONEST elemzést, melyhez egy saját fejlesztésű számítógépes programot használtunk. Az ER, PR és Ki67 értékek esetében az eredeti ONEST 100 véletlenszerű permutációval számolt változatán túl megvizsgáltuk azt is, hogy milyen eredményeket kapnánk, ha az ONEST-et mind a 9! (9 faktoriális; 362880) permutációt figyelembe véve alkalmaznánk. Megállapítottuk, hogy a 100, illetve az összes permutációból származtatott ONEST között nincs szignifikáns eltérés.

A munkánk során definiáltunk három ONEST-ből származtatott értéket (sávszélesség, OPA(n), ONEST érték), melyeket összehasonlítottunk az ER, PR és Ki67, a grade és komponensei, valamint a CNB és EXC minták között. Ez alapján megállapítottuk, hogy az ER és PR kategorizálásának reprodukálhatósági elemzése kevés (2-3) vizsgálóval már megbízható, míg a Ki67 kategorizálásánál 4-6 vizsgáló szükséges a reprodukálhatóság jó becsléséhez. Ki67 esetén – az elvárásokkal összhangban – nagyobb vizsgálószám szükséges a három-, mint a kétkategóriás besorolás esetén. Szintén megállapítottuk, hogy ER, PR és Ki67 esetén a reprodukálhatóság általánosságban nem függ attól, hogy CNB vagy EXC mintákról van-e szó.

A hisztológiai grade ONEST elemzése során megállapítottuk, hogy 6 (CNB) vagy 7 (EXC) vizsgáló szükséges a reprodukálhatóság reális meghatározásához. Az ONEST-ből származó OPA(9) (overall percent agreement, azaz a teljes egyezés aránya mind a 9 vizsgáló esetén) értékeket összehasonlítottuk Kappa statisztikából, illetve osztályon belüli korrelációs együtthatóból (intraclass correlation coefficient, ICC) kapott eredményekkel is, melyekkel azokat összhangban találtuk. Eredményeink tükrözik azt, hogy a grade mérsékelten reprodukálható, illetve, hogy a grade komponensei közül a magpleomorfizmus kategorizálása a legnehezebb.

Megvizsgáltuk, hogy az ONEST által javasolt minimális szükséges vizsgálószám hogyan van összhangban az irodalomban fellelhető korábbi grade reprodukálhatósági vizsgálatok Kappa és OPA eredményeivel, és megállapítottuk, hogy azok a vizsgálatok, melyeket az ONEST értéknél kevesebb vizsgálószámmal végeztek, vagy ahol a Kappát páronként vett vizsgálók alapján számolták, ott a reprodukálhatóság jobbnak tűnik, mint a nagyobb vizsgálószám alapján számított értékeknél.

A vizsgálataink során az ONEST elemzések néhány korlátjára is rámutattunk. Akárcsak más reprodukálhatósági mértéket, ezt is több tényező befolyásolja, mint például a besorolási kategóriák száma, a vizsgált paraméter szélsőséges értékekhez közeli vagy távoli eloszlása, a paraméter homogenitása vagy heterogenitása, a vizsgálók tapasztalata és száma, az általános besorolástól következetesen eltérő vizsgálók jelenléte… stb. Ezek miatt az ONEST elemzések eredményét is kellő fenntartásokkal, becslés jellegűnek kell tekinteni.

APPENDIX IV: PRINTED ARTICLES

I. **Cserni B**, Bori R, Csörgő E, Oláh-Németh O, Pancsa T, Sejben A, Sejben I, Vörös A, Zombori T, Nyári T, Cserni G. 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(6):1101-1109. doi: 10.1007/s00428-021-03172-9

IF(2021): 4.535 (Scimago journal ranking: Q1)

II. **Cserni B**, Bori R, Csörgő E, Oláh-Németh O, Pancsa T, Sejben A, Sejben I, Vörös A, Zombori T, Nyári T, Cserni G. 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* 2022;229:153718. doi: 10.1016/j.prp.2021.153718.

IF(2021): 3.309 (Scimago journal ranking: Q2)