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)

ONEST CALCULATOR DEVELOPED FOR THE PH.D. THESIS

Cserni B. ONEST Calculator. https://github.com/csernib/onest

1. INTRODUCTION

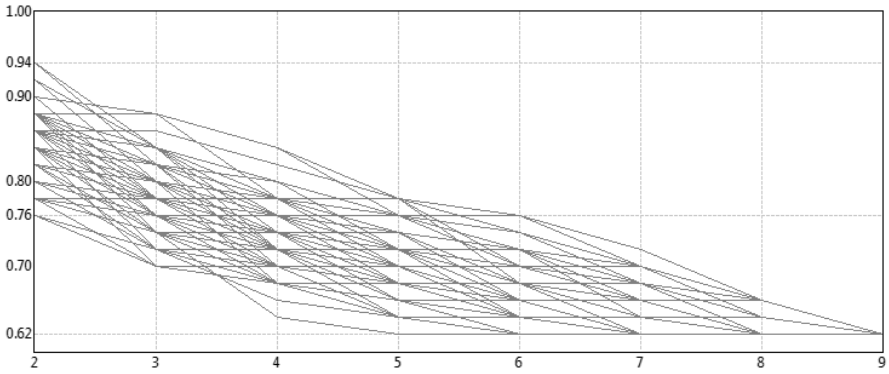
Excluding skin malignancies, breast cancer is the most common malignant tumor among women in Europe and worldwide.

Many prognostic factors of breast cancer 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.

Of the classifications of breast cancer, one of the most important is the segregation of carcinomas into estrogen receptor (ER)+ and ER- groups, of which only the first is likely to benefit from endocrine treatments. Currently, ER status is universally determined by IHC, and the judgement of what constitutes an ER+ and ER- status is somewhat arbitrary. Progesterone receptors (PR) also influence endocrine responsiveness. The evaluation of PR and its interpretation is similar to that of ER. 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. ER, PR and Ki67 assessment by microscopy requires the quantification of nuclei that stain with the relevant antibodies. A common method of doing this is by eyeballing, i.e., having a look at the slide and estimating the amount of tumor cells staining.

The grade of differentiation is a prognostic parameter reflecting the biology of the tumor. It is determined by the sum of 3 subscores reflecting glandular differentiation (“tubule formation”), nuclear pleomorphism and mitotic activity. Despite the recognized prognostic impact of histological grade, issues about the less than perfect reproducibility of grading have been the subject of several publications.

As a new approach, we have used a recently developed method, ONEST (Observers Needed to Evaluate Subjective Tests) to characterize these prognostic/predictive tests. ONEST is based on plotting the OPA (overall percent agreement; 0-1, i.e., 0-100% agreement) values 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. Well reproducible tests have high values of OPA(n) (i.e., OPA for all observers) with low numbers of raters to reach a plateau (small ONEST value) and small difference between the best and worst agreement of two raters (small bandwidth).



Example of an ONEST plot: Ki67 on CNB (core needle biopsy) with theoretical <1%, 1-10% and >10% categorization; axis x: number of observers, axis y: OPA values. 100 overlapping OPACs form the plot, the bottom values reach a plateau at OPA=0.62 (which is also OPA(9) – 62%overall agreement) at 5 observers (ONEST value); bandwidth is 0.94-0.76=0,18; i.e., 18% maximum difference in classification between 2 observers.

Based on its first description, the author created a tool to help ONEST calculations.

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 ER, PR and Ki67 with IHC (Study I), and the determination of histological grade and its components on HE stained sections (Study II).

To have an insight into the results of previous reproducibility studies in the light of the results gained by ONEST.

3. MATERIALS AND METHODS

From the archives of the Bács-Kiskun County Teaching Hospital, 100 breast cancer cases with routine IHC determination of ER, PR and Ki67 were selected. The cases included 50 core biopsy samples (CNB) and 50 samples from unrelated resected tumor specimens (EXC).

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%). 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%. 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; (2-5) with cut-offs suggested by the 2009, 2011, 2013 and 2015 St Gallen Consensus Conferences, respectively. Rating reliability was analyzed by the

intraclass correlation coefficient (ICC(2,1); two-way random effects, absolute agreement, single rater/measurement).

In parallel (Study II), 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 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 were used.

ONEST, as initially described by Reisenbichler et al (*Mod Pathol 2020; 33: 1746-1752*), was calculated for a randomly selected 100 permutations of the 362,880 (=9!) possible permutations of ranked pathologists. 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.

In the light of our findings, previous reproducibility studies of histological grading were looked at, and their results analyzed on the basis of their statistical approaches, and the number of observers involved in generating the figures.

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.

According to the ICC values for the evaluated parameters, most classifications relating to the ER and PR status of the tumors (percentages and Allred scores) 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 (ICC: 0.909-0.996) vs Ki67 (ICC: 0.625, EXC – 0.673, CNB) is striking, whereas the difference in ICC values of different Ki67 categorizations (all ≤ 0.760) is less prominent. There were no major or consistent differences between the ICC values of CNB and EXC specimens.

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.

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.

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 ONEST values (i.e., the number of observers required for the reliable estimation of reproducibility) were 2 for ER and 3 for PR categorization for both CNB and EXC, and ranged between 4-6 for the various Ki67 categorizations.

For histological grade, the kappa and ICC values reflect that the reproducibility of histological grading is 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 was somewhat better on excision specimens. Pleomorphism was the least reproducibly scored component of histological grade. In general, the middle categories (scores) were less reproducible than the extremes.

Importantly, ONEST suggests 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.

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.

Previous studies on histological grading on the basis of kappa values and OPA for all observers were also investigated. The results suggest that the reproducibility figures gained with less than 4 observers (i.e., the ONEST value) or by pairwise comparisons (virtually) reflect better agreement.

5. DISCUSSION

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.

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. 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).

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, 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). 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). 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.

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. 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. Our ONEST 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 examination of the literature highlights that OPA figures from studies with less than 4 to 6 raters are somewhat better than those gained with more observers. Studies reporting kappa statistics reflect the same trend. Many studies on the reproducibility of grading have used Cohen's kappa, which is devised for 2 observers, therefore pairwise comparisons were made, and the range or average was reported, but these basically reflect data derived from 2 observers, which may mirror a better performance than what the ONEST analysis implies.

Some further considerations are worth to be mentioned. 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.

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 in our studies I and II, 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.

6. CONCLUSIONS

In summary, we have first applied ONEST for characterizing the reproducibility of three biomarkers, ER, PR and Ki67, all evaluated by estimating the proportion of immunostained nuclei. 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 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, 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.