OPEN

Reinventing Nuclear Histo-score Utilizing Inherent Morphologic Cutoffs: Blue-brown Color H-score (BBC-HS)

Phillipe Price, MD, MSc, Usharani Ganugapati, MD,*† Zoran Gatalica, MD, PhD,‡ Archan Kakadekar, BSc, MD,* James Macpherson, MD,§ Louise Quenneville, MSc, MD,*† Henrike Rees, MD, FRCP(C),*† Elzbieta Slodkowska, MD,‖¶ Janarthanee Suresh, MD,* Darryl Yu, HON BSc MSc MD FRCP(C),*† Hyun J. Lim, PhD,# and Emina E. Torlakovic, MD, PhD*†***

Abstract: Immunohistochemistry (IHC) is a testing methodology that is widely used for large number of diagnostic, prognostic, and predictive biomarkers. Although IHC is a qualitative methodology, in addition to threshold-based stratification (positive vs. negative), the increasing levels of expression of some of these biomarkers often lead to more intense staining, which published evidence linked to specific diagnosis, prognosis, and responses to therapy. It is essential that the descriptive thresholds between positive and negative staining, as well as between frequently used graded categories of staining intensity (eg, 1+, 2+, 3+) are standardized and reproducible. Histo-score (H-score) is a frequently used scoring system that utilizes these categories. Our study introduces categorization of the cutoff points between positive and negative results and graded categories of staining intensity for nuclear IHC biomarker assays based on color interaction between hematoxylin and diaminobenzidine (DAB); the Blue-brown Color H-score (BBC-HS).

Six cases of diffuse large B-cell lymphoma were stained for a nuclear marker MUM1. The staining was assessed by H-score by 12 readers. Short tutorial and illustrated instructions were provided to readers. The novel scoring system in this study uses the interaction between DAB (DAB, brown stain) and hematoxylin (blue counterstain) to set thresholds between "0" (negative nuclei), "1+" (weakly positive nuclei), "2+" (moderately positive nuclei), and "3+" (strongly positive nuclei). The readers recorded scores for 300 cells. Krippendorff alpha (K-alpha) and intraclass correlation coefficient (ICC) were calculated. We have also assessed if reliability improved when counting the first 100 cells, first 200 cells, and for the total 300 cells using K-alpha and ICC. To assess the performance of each individual reader, the mean H-score and percent positive score (PPS) for each case was calculated, and the bias was calculated between each reader's score and the mean.

K-alpha was 0.86 for H-score and 0.76 for PPS. ICC was 0.96 for H-score and 0.92 for PPS. The biases for H-score ranged from −58 to 41, whereas for PPS it ranged from −27% to 33%. Overall, most readers showed very low bias. Two readers were consistently underscoring and 2 were consistently overscoring compared with the mean. For nuclear IHC biomarker assays, our newly proposed cutoffs provide highly reliable/reproducible results between readers for positive and negative results and graded categories of staining intensity using existing morphologic parameters. BBC-HS is easy to teach and is applicable to both human eye and image analysis. BBC-HS application should facilitate the development of new reliable/reproducible scoring schemes for IHC biomarkers.

Key Words: immunohistochemistry, Histo-score, interobserver variability, Blue-brown Color H-score (BBC-HS)

(*Appl Immunohistochem Mol Morphol* 2023;31:500–506)

Immunohistochemistry (IHC) is a testing methodology that allows for the detection of proteins in situ and is mostly applied to formalin-fixed/paraffin-embedded tissue sections. There is large and quickly growing number of diagnostic, prognostic, and predictive IHC biomarker assays that are in daily use by pathologists. Several guidelines have been developed to establish standards for testing for pre-

Received for publication October 29, 2022; accepted November 7, 2022.

From the *Department of Pathology and Laboratory Medicine, College of Medicine, University of Saskatchewan; †Department of Pathology and Laboratory Medicine, Saskatchewan Health Authority, Saskatoon; #College of Medicine, University of Saskatchewan; **Canadian Biomarker Quality Assurance, University of Saskatchewan, Saskatoon, SK; §Department of Pathology and Laboratory Medicine, Faculty of Medicine, University of Ottawa; ‖Department of Laboratory Medicine and Pathobiology, University of Toronto; ¶Department of Laboratory Medicine and Molecular Diagnostics, Sunnybrook Health Sciences Centre, Toronto, ON, Canada; and ‡Department of Pathology, Oklahoma University Medical Center, Oklahoma City, OK.

H.J.L. and E.T. are co-senior authors.

The authors declare no conflict of interest.

Reprints: Emina E. Torlakovic, MD, PhD, Department of Pathology and Laboratory Medicine, Royal University Hospital, 103 Hospital Drive, Saskatoon, Canada S7N 0W8 (e-mail: emina.torlakovic@usask.ca).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.appliedimmunohist.com.

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

dictive biomarkers in tumors of breast, lung, gastroesophagus, and colorectum.[1–5] Previous research has shown that in addition to threshold-based stratification (positive vs. negative), the increasing levels of expression of some of these biomarkers can further predict prognosis.[6–9] When evaluating an IHC-stained slide, pathologists assess the presence versus absence of staining in different cellular or extracellular compartments, and often must also assess the intensity of staining. This assessment by pathologists in IHC is a part of the analytical phase of the IHC assay, similar to the function performed by automatic readers used by other testing methodologies (eg, automated color change detection in colorimetric assays such as ELISA).[10] Although intrinsically subjective, this assessment (readout) needs to be highly reproducible between different pathologists.

The presence of staining with a chromogen is assessed by pathologists with or without the help of image analysis (IA).[11,12] Predictive biomarker assays often have specific scoring systems that need to be applied so that the results of the IHC biomarker assays are clinically meaningful. The rules for the readout are described when the use of the IHC biomarker is introduced for any given purpose. They may be incorporated in the IHC kits provided by industry, for example, in the interpretation manuals for programmed death ligand 1 (PD-L1) companion diagnostics.[13] For others, there may be guidelines published by professional organizations, for example, the ASCO/CAP guidelines for estrogen receptor (ER) and progesterone receptor testing in breast carcinoma.[1]

Depending on the IHC assay and its specific purpose, pathologists may assess the presence of staining in different cellular compartments, the intensity of staining in those compartments, the percentage of positive cells, or estimate the area containing positive cells, among other parameters. These observations could be combined in various scoring systems. One of the frequently used scoring schemes is the "Histo-score" (H-score), where the intensity of staining is multiplied by the percentage (P) of cells staining negative (0), weak (1+), moderate (2+), and strong (3+) in the following equation, giving an analytical range for 0 to 300.[14]

Although IHC is a qualitative laboratory assay and the extent of linearity of the assay is unknown, some linearity is assumed; lower staining intensity is expected to represent weaker expression of the antigen of interest, and vice versa, at least in most clinically used IHC biomarker assays. The H-score has been frequently used for pathology research and was shown also to be potentially useful for some predictive biomarkers.[15] Although the complete H-score is rarely applied to clinical practice, its components that include counting percent positive cells, as well as detecting thresholds between negative, weak positive, moderate positive, and strong positive staining of different cellular compartments are being used for large number of diagnostic, prognostic, and predictive IHC assays.[1,2,16] Automated methods have also been used for H-score and other readouts using IA of IHC-stained slides.[17] Although they are generally more precise than human readers and are becoming more popular, automated methods are still limited by cost, need for comprehensive validation, and many times may have no demonstrable improvements in readout accuracy.[18,19] It is important to emphasize that validation of IA for any given biomarker, especially for predictive and prognostic biomarkers, must include readout precision and readout accuracy, and implementation of these tools is a very complex process.[20]

Any readout of IHC staining is subject to multiple confounders, as it is a subjective task performed by human readers. Readouts that are complex, such as H-score, are even more challenging because errors at 1 level (eg, percent positive cells) are combined with errors at another level (eg, staining intensity). The potential for interobserver variability arises because there is yet no standardized rules for the interpretation of the intensity of staining, where different shades of brown (a continuous variable) is arbitrarily partitioned into discrete categories (0 vs. 1+ vs. 2+ vs. 3+). Furthermore, intensity of staining is only informative with the assumption that the IHC biomarker protocol has substantial linearity (even if the extent of linearity is unknown) where more intense staining means more expressed protein of interest.[21,22] Similarly, the lack of standardized protocols for IHC staining leads to wide differences in staining intensity of the IHC chromogen DAB and the counter stain.[23,24] This variability in staining intensity could be unrelated to protein expression and can produce clinically relevant errors.[25] The impact of hematoxylin is especially important for nuclear biomarkers such as ER because of the colocalization of the primary and counter stains.

This study was designed to assess the performance of criteria for determining cutoffs between positive and negative result as well between different intensity categories applied to nuclear H-score.

MATERIALS AND METHODS

Ethical Approval

The study used only quality assurance materials and reference samples where Research Ethics Board approval is not applicable.

Materials

Readers were provided with digital images for 6 cases of diffuse large B-cell lymphoma (DLBCL) stained for MUM1, a courtesy of Canadian Biomarker Quality Assurance Readout Proficiency Testing program (CBQAreadout.ca), a branch of Canadian Biomarker Quality Assurance—Programme Canadien d'assurance de la qualityé des biomarqueurs (CBQA-PCAB), an academic quality assurance program for predictive biomarker assays. The model of DLBCL stained for MUM1 was used because many of these lymphomas show little tissue heterogeneity in the expression of MUM1 and often cells of DLBCL will show variable intensity of nuclear staining for MUM1 (Fig. 1).

IHC Scoring

A short tutorial and written illustrated instructions were provided to readers. A novel scoring system was de-

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc.

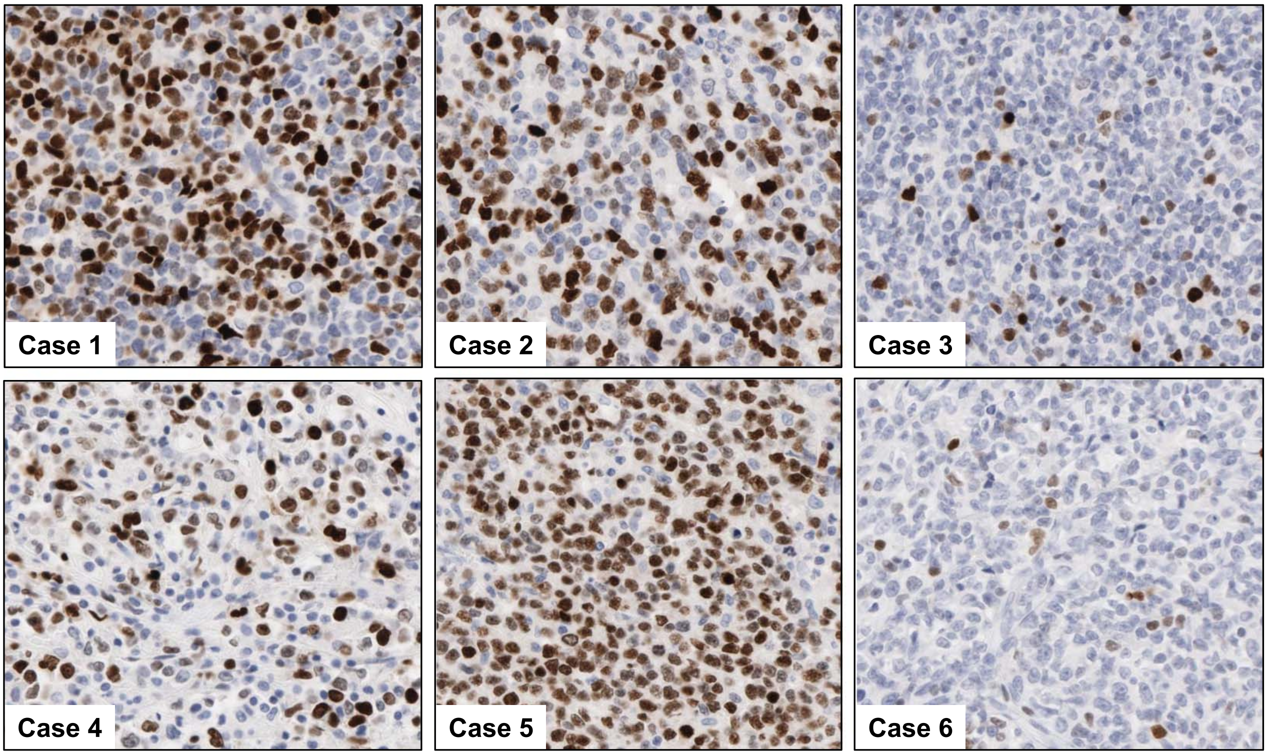


FIGURE 1. Cases distributed to the readers for H-score and percent positive score. The slides show lymph node tissue stained with immunohistochemistry against MUM1. Cases 3 and 6 have a low percentage of positive cells, cases 2 and 4 have moderate percentage, and cases 1 and 5 have high percentage.

signed using the interaction between DAB (brown chromogen) and hematoxylin (blue counterstain) to set thresholds between "0" (negative nuclei), "1+" (weakly positive nuclei), "2+" (moderately positive nuclei), and "3+" (strongly positive nuclei). This is detailed in Figures 2 and 3. Some nuclei have heterogeneous staining; readers were instructed

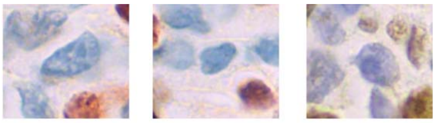
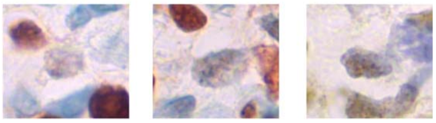
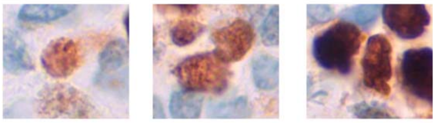
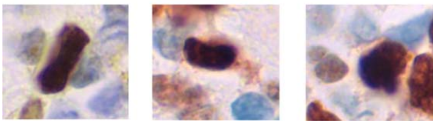
Intensity Score and Pattern [1]	Examples	Comment
0: Blue in > 50% of nuclear surface, chromatin pattern visible		Hematoxylin staining and no detectable DAB stain
1+: Gray[2,3,] in > 50% of nuclear surface, chromatin pattern visible		There is DAB stain in the nuclei, but it is balanced; neither hematoxylin nor DAB predominate
2+: Brown and chromatin pattern visible in > 50% of nuclear surface		Although DAB stain predominates over hematoxylin, nuclear texture consistent with chromatin pattern is still visible
3+: Dark brown, chromatin pattern not visible in > 50% of nuclear surface		DAB stain saturates the nucleus so almost no light can pass through, obscuring the chromatin pattern

FIGURE 2. Novel scoring system designed in this study. Training tutorials were performed using the following rules, and they were distributed to the readers for continued use while scoring the cases. DAB indicates diaminobenzidine. Cell in the center of the image is representative.

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc.

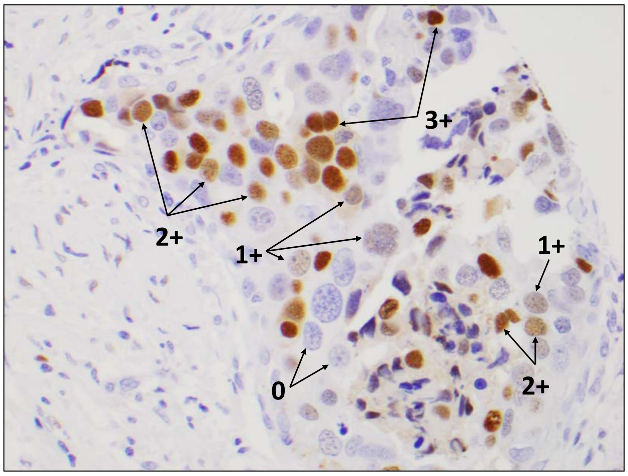


FIGURE 3. Sample image distributed to the readers as a visual aid for scoring immunohistochemistry staining intensity. 0 = negative or borderline staining, 1+ = weak staining, 2+ = moderate staining, and 3+ = strong staining. See Figure 2 for definitions.

to score based on the predominant pattern (covering > 50% of the nucleus).

H-score and Percent Positive Score

The readers recorded their scores for 300 cells by entering the sequence of numbers (from 0 to 3) into a word processor (supplemental material, Supplemental Digital Content 1, http://links.lww.com/AIMM/A373). The strings of digits were input into a simple code (Python) that counted the number of occurrences of each digit, the percentage of each digit, and the H-score (supplemental material, Supplemental Digital Content 1, http://links.lww.com/AIMM/A373). The H-score is calculated using the following formula using the percentage (P) of cells (0% to 100%), where P_0 is the proportion of negative (0) cells, P_{1+} is the percentage of 1+ cells, P_{2+} is the percentage of 2+ cells, and P_{3+} is the percentage of 3+ cells, which are summed; H-score $= 0 \times P_0 + 1 \times P_{1+} + 2 \times P_{2+} + 3 \times P_{3+}$. This gives an analytical range for H-score from 0 to 300. The Percent Positive Score (PPS) is simply the percentage of "positive" cells. This allowed for using the same string of digits used for H-score to calculate the PPS. The analytical range for PPS is 0% to 100%, where 0% is no cells staining, and 100% is all cells staining.

Statistical Reliability

To illustrate the performance of each reader, their scores were compared with the average of all readers for each case. The average was calculated by taking the mean of all readers' scores for each case. The analytical bias[26] was calculated as the difference between one reader's score for a given case and the average that case. Two statistical tests of reliability were chosen, Krippendorff alpha (K-alpha) and intraclass correlation coefficient (ICC). K-alpha is a reliability coefficient for continuous data, which measures between-reader agreement. If $\alpha \geq 0.80$, we can say there is strong interobserver reliability, and $\alpha = 0.67$ is the lowest limit deemed as an acceptable conclusion. Conclusions drawn base on $0.67 > \alpha < 0.80$ are described as "tentative."[27,28] ICC is another test of continuous data but uses the variance of a group of data points. An ICC coefficient <0.40 is poor, 0.40 to 0.59 is fair, 0.60 to 0.74 is good, and 0.75 to 1.00 is excellent.[29] The reliability coefficient given by each test cannot be directly compared, but instead has to follow the above rules of interpretation.[27,30,31]

Reliability for Number of Cells Counted

Counting and scoring the staining intensity of cells is a manual task that can be laborious. It is not feasible to ask human readers to count and score every cell on a slide, so a compromise is made to decide on a representative sample of the whole. We hypothesized that a higher number of counted cells will be more reliable than smaller numbers. Therefore, we assessed whether there was a difference in reliability between the first 100 cells, first 200 cells, and for the total 300 cells using K-alpha and ICC. Statistical analysis was performed to assess the reliability of H-score and PPS.

RESULTS

Interobserver Reliability and Readers' Bias for Blue-brown Color H-score

Reliability testing for H-score gives a K-alpha of 0.86 (CI = 0.84 to 0.88) and ICC of 0.96 (CI = 0.92 to 0.97), which is strong and excellent, respectively. For PPS, the ICC coefficient is excellent (0.92, CI = 0.83 to 0.95); however, the K-alpha is "tentative" (0.76, CI = 0.73 to 0.79) (Table 1).

Bias data are plotted in Figure 4 for each reader (A to L), where zero indicates no difference from the average, and the horizontal lines represent the range of scores relative to the average. This shows which readers routinely scored above or below the average, and the overall variability of a particular reader. A reader can be said to be more reliable if they have a narrower range, and more accurate if they are closer to the average. Overall, about half of the readers had ranges of <20%. There was a trend showing that readers who underscored or overscored also had wider ranges.

Reliability and Number of Cells Counted

Reliability analysis was repeated using Krippendorff alpha and ICC to see whether the statistical reliability improved by scoring more cells. Table 2 shows that

TABLE 1. Krippendorff Alpha and Intraclass Correlation Coefficient Results

		95% CI			95% CI	
Scoring System	K-alpha	Lower	Upper	ICC	Lower	Upper
H-score	0.86	0.84	0.88	0.96	0.92	0.97
PPS	0.76	0.73	0.79	0.92	0.83	0.95

ICC indicates intraclass correlation coefficient; PPS percent positive score.

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc.

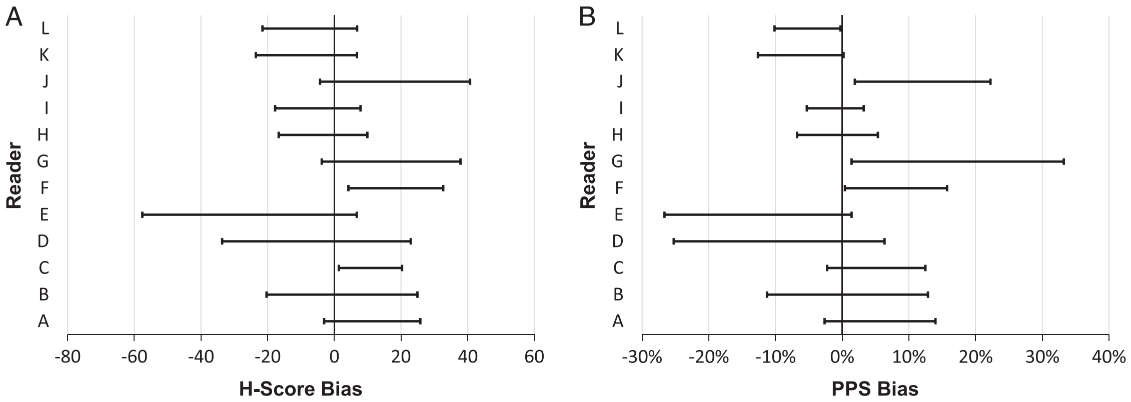


FIGURE 4. Bias was calculated for H-score (A) and PPS (B) for all readers (A-L). The bars show the variability of readers relative to the average. A bias of zero indicates that the readers score was equal to the average. PPS indicates percent positive score.

whether scoring 100 cells, 200 cells, or 300 cells, the reliability coefficients were almost identical.

DISCUSSION

We have developed new criteria to help readers to distinguish the thresholds between negative (0), weak (1+), moderate (2+), and strong (3+) nuclear staining. The new criteria utilize inherent interactions between the DAB brown chromogen and the counterstaining with blue hematoxylin. Although this interaction between DAB and hematoxylin is impacting the results of IHC staining, it is generally ignored in routine clinical practice. This approach does not require alteration of staining protocols, nor does it add extra work or cost in preparing the IHC slide for readout. This new approach for assessment of the nuclear staining could be used to generate the complete H-score and, even more importantly, it could be applied to define a single relevant threshold between positive and negative cells or between different categories for staining intensity. Previous methods rely on the intensity of DAB staining alone represented by shades of brown; a purely continuous variable. This is both arbitrary and poorly reproducible because there is no way to standardize one reader's thresholds and another's. Previous studies have shown highly variable reliability coefficients for many different staining-intensity-scoring systems.[15,19,32–36] For instance, McClelland et al[19] had readers from multiple centers perform H-scores and PPS for ER and progesterone receptor and showed wide variations for some cases; the worst example showed PPS ranging from 15% to 90% and H-score from 50 to 200 for the same case. In their study, however, there was no training for the readers. Introducing a new scoring system without training might not reflect the true potential of the scoring system.[37,38]

TABLE 2. Reliability Analysis for the First 100, 200, and 300 Cells Counted Measured by K-alpha and ICC

		95% CI			95% CI	
Scoring system	K-alpha	Lower	Upper	ICC	Lower	Upper
100 cells	0.861	0.838	0.883	0.961	0.936	0.974
200 cells	0.859	0.837	0.881	0.957	0.925	0.973
300 cells	0.835	0.803	0.863	0.963	0.925	0.978

ICC indicates intraclass correlation coefficient.

Although 2 different variables are combined to produce H-score (percent positive cells and staining intensity), our results showed that interobserver reliability was as good for H-score if not slightly superior to PPS, as shown by the statistical analysis. One explanation might be that differences between readers for low positive cases have a lesser impact on the total H-score than differences in strongly positive cases because the former percent is multiplied by 1 and the latter by 3. The drawback, however, is that H-score is not relevant for many IHC biomarkers and is also more laborious, especially if pathologists use a click counter and the calculation is performed manually. Our study provided a simple tool that is available free to all pathologists where no click counter or manual calculations are required. This tool could easily be adapted into an application or an online calculator that could streamline the process. This approach could make H-score more accessible for clinical and research applications where potentially relevant. Furthermore, H-score is already being adapted for use with artificial intelligence image recognition software.[17] This, in combination with the increased availability of whole-slide imaging and microscopic photography, could further improve the speed and convenience of H-score. However, the greater value of the new proposed scoring system is the defined cutoffs between 0, 1+, 2+, and 3+ because they are not arbitrarily set and already exist when the IHC slides are stained.

In the study of reliability, there is no agreement on which statistical test to use. Krippendorff alpha and ICC are tests that can accept continuous and ordinal data, which cannot be handled by ordinal tests of reliability such as Fleiss' kappa. K-alpha measures agreement between readers and has the advantage of high flexibility regarding the measurement scale (ie, number of categories in a scoring system) and the number of readers, and can

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc.

handle missing values. Thus, K-alpha emerged in content analysis but is widely applicable wherever 2 or more methods of generating data (readers) are applied to the same set of objects and the question is how much the resulting data can be trusted to represent something real.[27,28] ICC measures the "the proportion of variance of an observation due to between-subject variability in the true scores."[39] The advantage of this test is that it is intuitive and measures the reliability of groups rather then paired data.[29] The main advantage of K-alpha and ICC is that they do not require the arbitrary splitting of continuous data to generate categorical data. This is an important point for the study or interobserver reliability with a new scoring system, where we need to assess interobserver reliability for continuous data directly. Lastly, many studies have used linear regression such as Pearson or Spearman rho to determine reliability.[36,40,41] This seems to be effective at measuring agreement for continuous or ordinal data. However, Pearson or Spearman correlation is intrinsically misleading when assessing the interobserver variability because assessment of a single variable by 2 different readers is expected to produce significant correlation because the same variable is being assessed. In this way, Pearson correlation or Spearman correlation are inappropriately used in many pathology studies.

Although the number of readers is low for definitive conclusions, our study also showed a trend that readers who consistently underscored or overscored also had much wider range of bias. This was not dependent on the level of training because half of them were pathology residents and half were senior pathologists (data not shown). Also shown was that the number of cells counted did not impact reliability. Reliability coefficients remained high and varied little when comparing counts of 100, 200, or 300 cells. This would suggest that a 100-cell count would be sufficient at least in our study model that used a model with low tissue heterogeneity of expression for the chosen nuclear biomarker (MUM1).

The results of our study also highlight the importance of standardized counterstaining with hematoxylin for the readout of IHC assays. Whether the interaction between hematoxylin and DAB is used on purpose to distinguish between different intensities of staining in the IHC assays or not, this interaction happens on every slide where these 2 stains are present and colocalize; this has been mostly ignored and unexplored. Hematoxylin staining is partly standardized in IHC protocols that include automated (so-called "on board") staining. Most instruments used for automated IHC staining offer this possibility with a very small number of options for different hematoxylin types and different incubation times. Although it is uncertain to what degree even these limited differences may impact the intensity of nuclear staining, the larger issue is where laboratories perform manual hematoxylin staining or where pathologists personally prefer and demand stronger/darker hematoxylin staining than recommended by the manufacturer. It may be time for proficiency testing programs for IHC to put greater emphasis on the quality of hematoxylin staining to provide necessary stimulus for its further standardization.

In summary, our study provides evidence that the introduction of new rules for assessing the intensity of staining for either H-score or determining specific thresholds between positive and negative cells for PPS provides for a scoring system with excellent interobserver reliability. The main limitation of this study is that we included only cases with low tissue heterogeneity. However, this was done to control for factors that would introduce noise when trying to assess interobserver reliability for the new scoring system. Therefore, the interobserver reliability for tumors and biomarkers, which are known to have significant tissue heterogeneity, needs to be done in future studies. Lastly, it is important to note that reliability does not imply validity. It is possible to train readers and standardize IHC staining protocols to a point where readers have excellent reliability, but it still must be determined how these new rules for assessing the intensity of staining affect prognosis and predict response to therapy.

REFERENCES

1. Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP Guideline update. *J Clin Oncol*. 2020;38:1346–1366.
2. Wolff AC, Hammond MEH, Allison KH, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline focused update. *J Clin Oncol*. 2018;36:2105–2122.
3. Kalemkerian GP, Narula N, Kennedy EB, et al. Molecular testing guideline for the selection of patients with lung cancer for treatment with targeted tyrosine kinase inhibitors: American Society of Clinical Oncology Endorsement of the College of American Pathologists/International Association for the Study of Lung Cancer/Association for Molecular Pathology Clinical Practice Guideline Update. *J Clin Oncol*. 2018;36:911–919.
4. Sepulveda AR, Hamilton SR, Allegra CJ, et al. Molecular biomarkers for the evaluation of colorectal cancer: guideline from the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology. *J Clin Oncol*. 2017;35:1453–1486.
5. Bartley AN, Washington MK, Colasacco C, et al. HER2 testing and clinical decision making in gastroesophageal adenocarcinoma: guideline from the College of American Pathologists, American Society for Clinical Pathology, and the American Society of Clinical Oncology. *J Clin Oncol*. 2017;35:446–464.
6. Chen Y, Mu C-Y, Huang J-A. Clinical significance of programmed death-1 ligand-1 expression in patients with non-small cell lung cancer: a 5-year-follow-up study. *Tumori*. 2012;98:751–755.
7. Ma H, Lu Y, Marchbanks PA, et al. Quantitative measures of estrogen receptor expression in relation to breast cancer-specific mortality risk among white women and black women. *Breast Cancer Res BCR*. 2013;15:R90.
8. Morgan DAL, Refalo NA, Cheung KL. Strength of ER-positivity in relation to survival in ER-positive breast cancer treated by adjuvant tamoxifen as sole systemic therapy. *Breast Edinb Scotl*. 2011;20:215–219.
9. Elledge RM, Green S, Pugh R, et al. Estrogen receptor (ER) and progesterone receptor (PgR), by ligand-binding assay compared with ER, PgR and pS2, by immuno-histochemistry in predicting response to tamoxifen in metastatic breast cancer: a Southwest Oncology Group Study. *Int J Cancer*. 2000;89:111–117.
10. Cheung CC, D'Arrigo C, Dietel M, et al. Evolution of quality assurance for clinical immunohistochemistry in the era of precision medicine: part 1: fit-for-purpose approach to classification of clinical

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc.

immunohistochemistry biomarkers. *Appl Immunohistochem Mol Morphol AIMM*. 2017;25:4–11.

11. Laurinavicius A, Plancoulaine B, Laurinaviciene A, et al. A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast Cancer Res BCR*. 2014;16:R35.
12. Brügmann A, Eld M, Lelkaitis G, et al. Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. *Breast Cancer Res Treat*. 2012;132:41–49.
13. Agilent. PD-L1 IHC 22C3 pharmDx Interpreation Manual - NSCLC. Accessed October 28, 2022. https://www.agilent.com/cs/library/usermanuals/public/29158_pd-l1-ihc-22C3-pharmdx-nsclc-interpretation-manual.pdf.
14. McCarty KS, Szabo E, Flowers JL, et al. Use of a monoclonal anti-estrogen receptor antibody in the immunohistochemical evaluation of human tumors. *Cancer Res*. 1986;46:4244s–4248s.
15. Avilés-Salas A, Muñiz-Hernández S, Maldonado-Martínez HA, et al. Reproducibility of the EGFR immunohistochemistry scores for tumor samples from patients with advanced non-small cell lung cancer. *Oncol Lett*. 2017;13:912–920.
16. College of American Pathologists. PD-L1 Testing of Patients With Lung Cancer for Immunooncology Therapies. Accessed October 26, 2022. https://www.cap.org/protocols-and-guidelines/upcoming-cap-guidelines/pd-l1-testing-of-patients-with-lung-cancer-for-immunooncology-therapies.
17. Ram S, Vizcarra P, Whalen P, et al. Pixelwise H-score: a novel digital image analysis-based metric to quantify membrane biomarker expression from immunohistochemistry images. *PLoS One*. 2021;16:e0245638.
18. Røge R, Nielsen S, Riber-Hansen R, et al. Ki-67 Proliferation Index in breast cancer as a function of assessment method: a NordiQC experience. *Appl Immunohistochem Mol Morphol AIMM*. 2021;29:99–104.
19. McClelland RA, Wilson D, Leake R, et al. A multicentre study into the reliability of steroid receptor immunocytochemical assay quantification. *Eur J Cancer Clin Oncol*. 1991;27:711–715.
20. Lara H, Li Z, Abels E, et al. Quantitative image analysis for tissue biomarker use: a white paper from the digital pathology association. *Appl Immunohistochem Mol Morphol*. 2021;29:479–493.
21. Kraus JA, Dabbs DJ, Beriwal S, et al. Semi-quantitative immunohistochemical assay versus oncotype DX(®) qRT-PCR assay for estrogen and progesterone receptors: an independent quality assurance study. *Mod Pathol Off J U S Can Acad Pathol Inc*. 2012;25:869–876.
22. Hirsch FR, Varella-Garcia M, Bunn PA, et al. Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis. *J Clin Oncol Off J Am Soc Clin Oncol*. 2003;21:3798–3807.
23. Vyberg M, Nielsen S. Proficiency testing in immunohistochemistry–experiences from Nordic Immunohistochemical Quality Control (NordiQC). *Virchows Arch Int J Pathol*. 2016;468:19–29.
24. Vyberg M, Torlakovic E, Seidal T, et al. Nordic immunohistochemical quality control. *Croat Med J*. 2005;46:368–371.
25. Cheung CC, Lim HJ, Garratt J, et al. Diagnostic accuracy in fit-for-purpose PD-L1 testing. *Appl Immunohistochem Mol Morphol AIMM*. 2019;27:251–257.
26. Rifai N, Horvath AR, Wittwer CT. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics.* , 6th ed. Elsevier Canada; 2022.
27. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. *Educ Psychol Meas*. 1970;30:61–70.
28. Zapf A, Castell S, Morawietz L, et al. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol*. 2016;16:93.
29. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284–290.
30. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
32. Hsia Y, Huang H-C, Chang W-C. Comparison of different hepatocyte nuclear factor 4α clones for invasive mucinous adenocarcinoma of the lung. *Appl Immunohistochem Mol Morphol*. 2022;30:383–388.
33. Ali A, Bell S, Bilsland A, et al. Investigating various thresholds as immunohistochemistry cutoffs for observer agreement. *Appl Immunohistochem Mol Morphol AIMM*. 2017;25:599–608.
34. Koens L, van de Ven PM, Hijmering NJ, et al. Interobserver variation in CD30 immunohistochemistry interpretation; consequences for patient selection for targeted treatment. *Histopathology*. 2018;73:473–482.
35. Jaraj SJ, Camparo P, Boyle H, et al. Intra- and interobserver reproducibility of interpretation of immunohistochemical stains of prostate cancer. *Virchows Arch Int J Pathol*. 2009;455:375–381.
36. Cohen DA, Dabbs DJ, Cooper KL, et al. Interobserver agreement among pathologists for semiquantitative hormone receptor scoring in breast carcinoma. *Am J Clin Pathol*. 2012;138:796–802.
37. Van Bockstal MR, Cooks M, Nederlof I, et al. Interobserver Agreement of PD-L1/SP142 immunohistochemistry and tumor-infiltrating lymphocytes (TILs) in distant metastases of triple-negative breast cancer: a proof-of-concept study. A report on behalf of the International Immuno-Oncology Biomarker Working Group. *Cancers*. 2021;13:4910.
38. 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 Off J U S Can Acad Pathol Inc*. 2020;33:1746–1752.
39. Everitt BS. *Making Sense of Statistics in Psychology: A Second-level Course*. Oxford University Press; 1996. xii, 350.
40. Sinclair W, Kobalka P, Ren R, et al. Interobserver agreement in programmed cell death-ligand 1 immunohistochemistry scoring in nonsmall cell lung carcinoma cytologic specimens. *Diagn Cytopathol*. 2021;49:219–225.
41. Tan WCC, Nerurkar SN, Cai HY, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun Lond Engl*. 2020;40:135–153.

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc.