Application of computer-generated images to train pattern recognition used in semiquantitative immunohistochemistry scoring

GLEN KRISTIANSEN\textsuperscript{1} and MATTHIAS SCHMID\textsuperscript{2}

\textsuperscript{1}Institute of Pathology, University Hospital Bonn; \textsuperscript{2}Institute of Medical Biometry, Informatics and Epidemiology (IMBIE), University Hospital Bonn, Bonn, Germany

This study aimed to clarify whether the pattern recognition involved in scoring proliferation fractions can be trained by abstract computerized images of virtual tissues. Twenty computer-generated images with randomly distributed blue or red dots were scored by 12 probands (all co-workers or collaborators of the Institute of Pathology, University of Bonn). Afterward, the probands underwent a training phase during which they received an immediate feedback on the actual rate of positivity after each image. Finally, the initial testing series was rescored. In a second round with 15 different probands, 20 Ki-67 immunohistochemistry images of tonsil tissue were scored, followed by the same training phase with computer-generated images, before the immunohistochemistry slides were scored again. Paired t-tests were used to compare the differences in mean rates pre- and post-training. Concerning computerized images, untrained probands scored the percentages of positive dots with a mean deviation from the true rates of 8.2%. Following training, the same testing series was scored significantly better with a mean deviation of 4.9% (mean improvement 3.3%, $p < 0.001$). Scoring real immunohistochemistry slides, the training with computerized images also improved correct estimations, albeit to a lesser degree (mean improvement 1%, $p = 0.03$). Abstract computerized images of virtual tissues may be a useful tool to train and improve the accuracy of pattern recognition involved in semiquantitative scoring of immunohistochemistry slides. As a side result, this study highlights the value of computer-generated images to verify the performance of image-analysis software.

Key words: Immunohistochemistry scoring; computer-aided learning; tissue models.

Glen Kristiansen, Institute of Pathology, University Hospital Bonn, Sigmund-Freud-Strasse 25, 53105 Bonn, Germany. e-mail: glen.kristiansen@ukbonn.de

Immunohistochemistry has become the most important ancillary technique for surgical pathologists [1]. Next to qualitative diagnostic markers that aid in tumor typing, a growing number of markers demands a quantitation of expression either in intensity (e.g., Her2) or number of positive cells (Ki-67, hormone receptors) [2–4]. Hence, a recurrent task of pathologists is quantitation of immunohistochemistry slides. Accurate determination of these scores is often crucial for further therapy planning. For instance, the grade of neuroendocrine tumors is defined by either the mitotic count or the proliferative fraction (Ki-67) [5]. Clinicians may consider the proliferative activity of a tumor also in other tumor entities to better estimate the prognosis or to tailor adjuvant therapy [6–9]. With the growing application of personalized approaches in oncology, an increasing number of tests (e.g., PD-L1, tumoral T-cell infiltrate) will necessitate quantitation or semiquantitation by pathologists.

Immunohistochemistry itself is strictly not a quantitative technique; hence, it shows marked variability of its results. These may depend on pre-analytical or analytical factors, but probably more important is the influence of the observer and the assessment method and many studies have noted significant rates of variability [10–14]. It is widely believed that computerized analyses of immunohistochemistry images will eventually standardize immunohistochemistry scores and a variety of software solutions for quantitation are already available. Presently, the majority of surgical pathologists that do not work in a digitized laboratory will adhere to the conventional analog techniques of a)
either counting negative and positive cells (‘accountant approach’), which is cumbersome and tedious or b) simply estimating the rate of positive cells (‘eyeballing’ or ‘cowboy approach’) which is faster but may be inaccurate [15]. Due to the increasingly pressing work loads of contemporary pathologists, most will practice a mixed approach of both, counting a small area and eyeballing the overall case. Clearly, estimating rates of positivity by eyeballing can be a very efficient way to read an immunostain.

The authors wondered whether the pattern recognition required for reading, for example, a proliferation fraction could be trained by computer-simulated patterns, by presenting randomly generated grids with negative and positive dots. Also, we addressed the question whether training with computer-generated images would result in better estimates in scoring real immunohistochemistry slides. This study presents and validates a very simple computer program to evaluate and train recognition of positivity rates.

MATERIAL AND METHODS

Computer platform and program to generate artificial images

In his youth, the first author had used a Sinclair ZX Spectrum (48k), one of the most popular home computers in the UK and Europe in the early to mid 80ties of last century for hobby programming [16]. For this project, the ZX Spectrum emulator FUSE (http://fuse-emulator.sourceforge.net) that runs on an Apple Macbook pro (2015) was used to write a program that generates different screens with randomly assigned positive and negative dots, reminiscent of cells with different proliferation fractions.

The program was written in Sinclair Basic (listed in Fig. 1). First, a small solid ball to represent a nucleus or a cell is added to the user-defined graphics (subroutine lines 2000–2070); then, the user is asked to define the range (minimum, maximum) of positivity rates to evaluate. Afterward, the program establishes a three-dimensional array to encompass 10 screens with randomly generated positivity rates in the user-defined range, distributed over the screen at random again (subroutine lines 2200–2330). The program ensures that the exact number of positive cells calculated will be displayed. Finally, the main routine of this program (lines 130–170) will present the user a succession of these 10 screens and individually ask for the estimated percentage of positive ‘cells’ (Fig. 2A). Afterward, the program presents a line-wise comparison of real and estimated numbers (Fig. 2B). The program code is provided as a snapshot file upon request.

Probands

Twelve voluntary probands consisting of seven histopathology residents in training, three basic scientists, and two visitors with a non-medical background were analyzed. In the second round, another 15 voluntary probands consisting of 12 beginners (<3 years of histopathology training) and 3 histopathology consultants participated. No ethics approval was necessary to conduct this study.

Validation of training effect in correctly estimating positivity rates of computer images

The testing range of the program was set to 0 to 100% positive cells. To speed up testing and training,
screenshots were compiled in a PowerPoint presentation (File S1). The 20 test screens displayed various rates of positive cells, ranging from 2 to 93% (mean 40%). The proband’s results of the first 20 cases were recorded in SPSS (v25), followed by a training phase of evaluating 30 cases, which was not recorded. In the training phase, the correct percentage of red cells was displayed after each individual image, upon pressing a key to give the proband an immediate feedback. Eventually, the first 20 cases were scored again and the results were recorded.

Validation of training effect in correctly estimating positivity rates of real Ki-67 immunohistochemistry images

To check, if training semiquantitative image analysis with computer-generated images does result in improved estimations in real-world immunohistochemistry images, a set of 10 computer-generated images were evaluated by the software packages Qupath and Ki-67 quantifier. To analyze the images with Qupath, the colors for DAB and hematoxylin were adjusted in the image data to allow correct detection (DAB: 0.048, 0.706, 0.706; hematoxylin 0.996, 0.065, 0.067). Then, the positivity rate was measured by analyzing fast cell counts. For the image analysis with Ki-67 quantifier, which is a software package optimized for the detection of DAB-stained nuclei, the colors of the images were optimized as well (red turned to brown, cyan to blue).

Verification of positivity rates by computerized image analysis using the software packages Qupath and Ki-67 quantifier

To verify, if even primitive computer-generated images could be used to validate image analysis software, a set of 10 computer-generated images were evaluated by the software packages Qupath and Ki-67 quantifier. For the image analysis with Qupath, the colors for DAB and hematoxylin were adjusted in the image data to allow correct detection (DAB: 0.048, 0.706, 0.706; hematoxylin 0.996, 0.065, 0.067). Then, the positivity rate was measured by analyzing fast cell counts. For the image analysis with Ki-67 quantifier, which is a software package optimized for the detection of DAB-stained nuclei, the colors of the images were optimized as well (red turned to brown, cyan to blue).

Statistical analysis

Paired t-tests were used to compare the proliferation rates estimated from the pre- and post-training data of all participants. Nominal p-values were used to evaluate statistical significance, defined as p < 0.05. Due to the exploratory nature of the experiments, no procedures were applied to correct p-values for multiple comparisons. All statistics were calculated with SPSS v25 (Armonk, NY: IBM Corp., USA) and the R Language and Environment for Statistical Computing, version 4.1.0 (R Core Team 2021).

RESULTS

Performance of the BASIC program

Even though the screen presentation is very schematic, the random presentation of blue or red cells on this matrix of 640 cells is sufficiently complex to discourage exact counting and to promote `eyeballing’ estimates (Fig. 2A). The program itself may be used to evaluate 10 screens and then presents a line-wise comparison of real and estimated data (Fig. 2B). The Microsoft PowerPoint presentation with the 80 images (testing/training/re-testing) was accomplished by all participants in about half an hour.
Estimates of percentages on computer images (pre- and post-training)

For the twelve participants of the first round, the maximum deviation of the estimate from the real value was 36 percentage points, and the mean deviation (overall 20 images) ranged from 6.15 to 12.90 percentage points (mean 8.2). The participants reported that estimating the extremes was subjectively easier than the middle region from 10 to 90%. This is confirmed by the statistics, showing a deviation of 2.57 percentage points for images with <10% or >90% positive dots, compared with 10.05% points for the middle range (10–90%).

Following training, the overall estimation performance was markedly improved. The maximum deviation was 29 percentage points, and the mean deviation ranged from 3.6 to 6.5 percentage points (mean 4.9). The difference of means (pre- vs post-training) was highly significant (Fig. 4A, p < 0.001).

The individual analysis of the participants’ estimations revealed differences in their learning success. The paired t-test demonstrated significant (p < 0.05) differences of mean estimations pre- and post-training for four individuals, three more reached p-values ranging between 0.05 and 0.1, and five still had closer estimates post-training but had p-values > 0.1.

Measurements and estimates of percentages on real immunohistochemistry images (pre- and post-training with computer images)

Measured over a series of 37 Ki-67 images of tonsil tissue, both algorithms (QuPath, Ki-67 Quantifier) showed highly concordant, yet minimally differing values (intraclass correlation coefficient 0.98, 95% confidence interval [0.95; 0.99]), with QuPath measuring a mean of 23.6% and Ki-67 Quantifier a mean 20.7%. This mean difference of 2.9% was significant though in a paired t test (p < 0.001). The case with the maximum difference (QuPath 31.8%, Ki-67 Quantifier 18%) is shown in Fig. 3A, and a case with nearly identical rates (QuPath 39.1%, Ki-67 Quantifier 38.9%) is shown in Fig. 3B. To define the ground truth for further analysis of proband data, the mean proliferation rate of both algorithms was used. Thus, the proliferation rates of the 20 immunohistochemistry images ranged from 2 to 81% (median 15%).

For the 15 participants, the maximum deviation of their estimates from the real values was 32 percentage points, and the mean deviation (overall 20 images) ranged from 3.6 to 10.3 percentage points (mean 6.5). Following training, the overall performance of estimations was slightly improved with lower deviation rates in 11 participants. The mean deviation ranged from 2.7 to 9.4 percentage points (mean 5.5), which was significantly lower than pre-training (Fig. 4B, p = 0.034). In the group of residents, more variation, but also more marked improvements were seen than in the group of consultants, whose ratings were more stable (Fig. 4B, see p13-15).

Verification of positivity of computer-generated images using the software packages Qupath and Ki-67 Quantifier

In all ten computer-generated images, the absolute correct number of cells (640) and positive ‘cells’ was detected by QuPath. The Ki-67 quantifier detected between 639 and 643 cells per image (mean 640.2), a minimal variation from the ground truth of 640 that did not impact the correct percentual rate of positive cells. In all ten computer-generated images, the correct percentual rate of positive ‘cells’ was detected by both programs.
DISCUSSION

Determining fractions of positive cells in immunohistochemistry slides is a daily task of surgical pathologists, and different approaches ranging from accurate counts to rough estimates by eyeballing are being practiced. This study employed computer-generated images displaying positive and negative dots to evaluate and train the capacity of observers to estimate the correct fraction of positive dots. To our knowledge, this new concept of training pattern recognition of pathologists using computer emulated images as a model of tissues has not been established so far but ought to be of general interest to surgical pathologists.

The statistical evaluation of a dozen participants who scored 20 screens, underwent training, and rescored the initial series of 20 screens clearly validates a positive training effect, allowing them to improve their accuracy of estimating percentages by an average of 3.3 percentage points. Also, the maximum deviations from the true values decreased after training. Considering this finding, a recurring and obvious question is, if this observation, based on highly artificial and crude computer graphics, has any value in the sense that it translates into improved estimations of positivity rates of real immunohistochemistry slides. Therefore, a second round with a different set of scorers was conducted, who evaluated individually a series of 20 real Ki-67 stainings of lymphatic tissue, then underwent a training phase with computerized images, and finally rescored the initial IHC slides. Again, following the training phase with computer images, the second round of Ki-67 estimates was on average closer to the correct values that had been determined before using two different

![Histogram of results. Percentual deviations (estimates - real numbers) pre (blue) and post (red) training) tested on (A) computer-generated images, and (B) real immunohistochemistry images (Ki-67, tonsil).](image-url)
images analysis programs. This verifies that this type of training with artificial images had indeed increased the accuracy of estimates even on real immunohistochemistry slides, which has, to our knowledge, not been shown before. We are not entirely surprised by this finding, as the principles of pattern recognition either applied to abstract color dots on the computer screen or stained cells under the microscope ought to be identical. We are confident that this novel approach provides a simple way for pathologists to train their accuracy in these estimations. It is also obvious that the rate of improvement in real IHC slides is lower (1%) than in scoring computerized slides of the first series (3.3%). We assume that this is in fact due to a ‘loss in translation’ from crude computer graphics to real-life IHC images, which are more complex to read: cells differ in size, are unevenly distributed, partially overlap, the staining intensity varies–all these factors contribute to the difficulty of estimating slides quickly and correctly. Future versions of this training program will have to simulate human tissues more realistically, and it will be of interest to see, if this further increases the training success, as verified on real slides.

These percentual differences (pre- and post-training) may appear small, but may in fact impact patient care, for example in neuroendocrine tumors, in which the proliferation rates are crucial to determine the tumor grade [5]. Although the number of participants in this study is rather small, which precludes more detailed stratified analyses, it has already proven that pattern recognition can be trained in an abstract form to improve correct percentual estimation of ‘positive’ items.

The data set generated from the abstract images also allowed to identify the problematic zone, as estimations in the extremes (<10% or >90%) were significantly closer to the real numbers as in the intermediate zone, which is also reported for real slides of Ki-67 stains [15].

As intra- and interobserver variability is a proven obstacle in semiquantitative immunohistochemistry, every tool to increase accuracy of IHC reading ought to be welcome [19,20]. Even though the graphical user interface provided by this program running on an emulated historical home computer is rather crude and obviously over-simplified, it is already sufficient to allow a successful training of correct estimates of positive cells. The participants also had fun to score the screens provided and particularly liked the simple charme of the outdated graphics, which reminded them of video games of the 80ies. So far, very few examples of medical software for the ZX Spectrum have been published and this program is a rather late addition to this stock [21,22].

With the increasingly widespread introduction of digital pathology, and applications of artificial intelligence to image analysis, many believe that tasks of quantifications by humans will soon be history [23]. This may, or may not be so. Studies comparing human and computer-based expression analysis demonstrate the general feasibility of automated analysis, yet also provide reasons for caution, as results may differ and supervision by an experienced pathologist is still warranted [24–26]. In the interpretation of biomarker expression, the correct selection of the regions of interest (ROI) is crucial, especially in heterogeneously expressed biomarkers and future computer algorithms may have to accommodate this and provide not only values of general percentual positivity but also detect hot spot regions [27].

For the time being, pathologists have to read biomarkers from glass slides in their daily routine work and any training tool to help them accomplish this task better should be welcome. Also, a couple of images analysis applications have been developed, for example, for Ki-67, to date and the first benchmark studies have been carried out to validate these on human tissues [28,29]. These studies demonstrate a high rate of concordance, which we can confirm, but we also found a minimal systemic difference between the two platforms used here. This may possibly be adjusted by changing detection parameters. But a remaining problem in these comparative studies is the missing gold standard, so the resulting data from different algorithms are usually only correlated with itself and possibly clinical outcome data. Images of virtual tissues with an exactly predefined rate of positivity, as exemplified in this prototypical study, may be a valuable tool to train and validate algorithms for tissue analysis. It is assuring to see that all ten computer-generated images read by the widely used academic quantitative software package Qupath and by the Ki-67 Quantifier have been correctly quantified.

CONCLUSIONS

In summary, this study exemplifies that the pattern recognition capabilities needed to score rates of positivity that are so crucial in evaluating actual immunohistochemistry slides can be trained and improved by computer-generated images of tissues. Further improvements of this program on a contemporary computer platform are under development to display a more natural variability of cell distribution, shape, and size and may also incorporate variation of color intensity as is seen with real
immunohistochemistry stains. Also, other staining qualities as membranous or cytoplasmic staining may be addressed in future versions. Other common diagnostic problems, as discrimination of tumor cells from inflammatory, or stroma cells may also be implemented in future emulations, to allow for a more realistic training of life-like diagnostic situations.

We are grateful to the histopathology residents and consultants of the Institute of Pathology of the University of Bonn for their time and enthusiasm for this project and for being voluntary probands. We thank Kai Saeger (VMScope, Berlin) for providing us with Ki-67 measurements of the immunohistochemistry images. This research did not receive grants from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICTS OF INTEREST

We declare to have no conflict of interests.

REFERENCES

1. Taylor CR. Quantitative in situ proteomics; a proposed pathway for quantification of immunohistochemistry at the light-microscopic level. Cell Tissue Res. 2015;360:109–20.
2. Taylor CR, Cooper CL, Kurman RJ, Goebelsmann U, Markland FS Jr. Detection of estrogen receptor in breast and endometrial carcinoma by the immunoperoxidase technique. Cancer. 1981;47:2634–40.
3. Ratcliffe N, Wells W, Wheeler K, Memoli V. The combination of in situ hybridization and immunohistochemical analysis: an evaluation of Her2/neu expression in paraffin-embedded breast carcinomas and adjacent normal-appearing breast epithelium. Mod Pathol. 1997;10:1247–52.
4. Denkert C, Budezies J, von Minckwitz G, Wienert S, Loibl S, Klauschen F. Strategies for developing Ki67 as a useful biomarker in breast cancer. Breast. 2015;24 (Suppl 2):S67–72.
5. Philips P, Kooby DA, Maithel S, Merchant NB, Weber SM, Winslow ER, et al. Grading using Ki-67 index and mitotic rate increases the prognostic accuracy of pancreatic neuroendocrine tumors. Pancreas. 2018;47:326–31.
6. Wong E, Nahar N, Hau E, Varikatt W, Gebski V, Ng T, et al. Cut-point for Ki-67 proliferation index as a prognostic marker for glioblastoma. Asia Pacific J Clin Oncol. 2018;15:5–9.
7. Garrity MM, Burgart LJ, Mahoney MR, Windschil HE, Salim M, Wiesenfeld M, et al. Prognostic value of proliferation, apoptosis, defective DNA mismatch repair, and p53 overexpression in patients with resected Dukes’ B2 or C colon cancer: a North Central Cancer Treatment Group Study. J Clin Oncol. 2004;22:1572–82.
8. Kitson S, Sivalingam VN, Bolton J, McVey R, Nickhoo-Amiry M, Powell ME. Ki-67 in endometrial cancer: scoring optimization and prognostic relevance for window studies. Mod Pathol. 2017;30:459–68.
9. Pujol JL, Simony J, Jolimoy G, Jaffuel D, Demoly P, Quantin X, et al. Hypodiploidy, Ki-67 growth fraction and prognosis of surgically resected lung cancers. Br J Cancer. 1996;74:964–70.
10. Taylor CR, Levenson RM. Quantification of immunohistochemistry–issues concerning methods, utility and semiquantitative assessment II. Histopathology. 2006;49:411–24.
11. Walker RA. Quantification of immunohistochemistry–issues concerning methods, utility and semiquantitative assessment I. Histopathology. 2006;49:406–10.
12. Kristiansen G. Markers of clinical utility in the differential diagnosis and prognosis of prostate cancer. Mod Pathol. 2018;31:143–55.
13. Røge R, Nielsen S, Riber-Hansen R, Vyberg M. Impact of primary antibody clone, format, and stainer platform on Ki67 proliferation indices in breast carcinomas. Appl Immunohistochem Mol Morphol. 2019;2017:732–9.
14. Røge R, Nielsen S, Riber-Hansen R, Vyberg M. Ki-67 proliferation index in breast cancer as a function of assessment method: a NordiQC experience. Appl Immunohistochem Mol Morphol. 2021;29:99–104.
15. Varga Z, Diebold J, Dommann-Scherrer C, Frick H, Kaup D, Noske A, et al. How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists. PLoS One. 2012;7: e37379.
16. Kristiansen G, Prahm A. 30 tolle Maschinencodieroutinen für den ZX Spectrum. Happy Comput. 1985;5:98–102.
17. Klauschen F, Wienert S, Schmitt WD, Loibl S, Gerber B, Blohmer JU, et al. Standardized Ki67 diagnostics using automated scoring-clinical validation in the GeparTrio breast cancer study. Clin Cancer Res. 2015;21:3651–7.
18. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017;7:16878.
19. Lawrie CH, Ballabio E, Soilleux E, Sington J, Hatton HE, Salim M, Wiesenfeld M, et al. Prognostic value of assessment method: a NordiQC experience. Appl Immunohistochem Mol Morphol. 2019;2017:732–9.
20. Kristiansen G, Prahm A. 30 tolle Maschinencodieroutinen für den ZX Spectrum. Happy Comput. 1985;5:98–102.
21. de Paz P, Barrio JP. Stereological parameters from window studies. Mod Pathol. 2017;30:143–55.
22. Lass P, Nowak J, Banecki B, Slominski JM. The CVA database as an example of the use of a ZX SPECTRUM microcomputer for recording data on Cold War developments. Comput Biol Med. 1985;15:153–8.
neurologic patients for the purpose of scientific research. Wiad Lek. 1988;41:604–8.
23. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703–15.
24. Mohammed ZM, McMillan DC, Elsberger B, Going JJ, Orange C, Mallon E, et al. Comparison of visual and automated assessment of Ki-67 proliferative activity and their impact on outcome in primary operable invasive ductal breast cancer. Br J Cancer. 2012;106:383–8.
25. Yamazaki Y, Nakamura Y, Shibahara Y, Konosu-Fukaya S, Sato N, Kubota-Nakayama F, et al. Comparison of the methods for measuring the Ki-67 labeling index in adrenocortical carcinoma: manual versus digital image analysis. Hum Pathol. 2016;53:41–50.
26. González-González R, Molina-Frechero N, Carreón-Burciaga RG, López-Verdin S, Robles-Bonilla C, Pereira-Prado V, et al. Comparison between manual and automated methods for Ki-67 immunostaining quantification in ameloblastomas. Anal Cell Pathol (Amst). 2016;2016:7486989.
27. Wessel Lindberg AS, Conradsen K, Larsen R, Friis Lippert M, Røge R, Vyberg M. Quantitative tumor heterogeneity assessment on a nuclear population basis. Cytometry A. 2017;91:574–84.
28. Acs B, Pelekanou V, Bai Y, Martinez-Morilla S, Toki M, Leung SCY, et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. Lab Invest. 2019;99:107–17.
29. Koopman T, Buikema HJ, Hollema H, de Bock GH, van der Vegt B. Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. Breast Cancer Res Treat. 2018;169:33–42.
30. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2021. https://www.R-project.org/

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

File S1. Powerpoint presentation for testing, training and retesting of positivity rates using computerized images only.
File S2. Powerpoint presentation for testing, training and retesting, using Ki-67 immunohistochemistry images for testing/retesting and computerized images for training.