

WILEY

2021 INSIGHTS IMMUNOLOGY

Learn about the latest trends in immunology from today's influencers. Discover how they leverage high-dimensional biology to drive new discoveries and research. Earn CE credit.

Register here



**CURRENT
PROTOCOLS**
A Wiley Brand



The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice

Anna C S Bodén,^{1,2}  Jesper Molin,³ Stina Garvin,¹ Rebecca A West,^{4,5} Claes Lundström^{2,3} & Darren Treanor^{1,2,4,6}

¹Department of Clinical Pathology, Department of Biomedical and Clinical Sciences, Linköping University, Linköping,

²Centre for Medical Image Science and Visualization, Linköping University, Linköping, ³Sectra AB, Linköping, Sweden,

⁴Leeds Teaching Hospitals NHS Trust, Leeds, ⁵Department of Histopathology, Dewsbury and District Hospital, Dewsbury, and ⁶Pathology and Data Analytics, University of Leeds, Leeds, UK

Date of submission 5 December 2020

Accepted for publication 14 February 2021

Published online Article Accepted 15 February 2021

Bodén ACS, Molin J, Garvin S, West RA, Lundström C & Treanor D.
(2021) *Histopathology* 79, 210–218. <https://doi.org/10.1111/his.14356>

The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice

Aims: One of the major drivers of the adoption of digital pathology in clinical practice is the possibility of introducing digital image analysis (DIA) to assist with diagnostic tasks. This offers potential increases in accuracy, reproducibility, and efficiency. Whereas stand-alone DIA has great potential benefit for research, little is known about the effect of DIA assistance in clinical use. The aim of this study was to investigate the clinical use characteristics of a DIA application for Ki67 proliferation assessment. Specifically, the human-in-the-loop interplay between DIA and pathologists was studied.

Methods and results: We retrospectively investigated breast cancer Ki67 areas assessed with human-in-the-loop DIA and compared them with visual and automatic approaches. The results, expressed as

standard deviation of the error in the Ki67 index, showed that visual estimation ('eyeballing') (14.9 percentage points) performed significantly worse ($P < 0.05$) than DIA alone (7.2 percentage points) and DIA with human-in-the-loop corrections (6.9 percentage points). At the overall level, no improvement resulting from the addition of human-in-the-loop corrections to the automatic DIA results could be seen. For individual cases, however, human-in-the-loop corrections could address major DIA errors in terms of poor thresholding of faint staining and incorrect tumour–stroma separation.

Conclusion: The findings indicate that the primary value of human-in-the-loop corrections is to address major weaknesses of a DIA application, rather than fine-tuning the DIA quantifications.

Keywords: artificial intelligence, breast cancer, computational pathology, digital image analysis (DIA), digital pathology, human-in-the-loop, Ki67, machine learning

Introduction

As more pathology departments implement digital review, artificial intelligence (AI) systems for digital

image analysis (DIA) will also become more commonplace.^{1–4} It is important to recognize accurate quantification, facilitated through AI for prognostic and predictive scoring, as a diagnostic companion to pathologists. A popular candidate for AI use is Ki67 scoring in breast cancer (BC), for which studies have shown that DIA is equal to or better than manual scoring by pathologists.^{5–9} There is, however,

Address for correspondence: A C S Bodén, Klinisk patologi, Region Ostergötland, 58185, Linköping, Sweden. e-mail: anna.c.boden@regionostergotland.se

relatively little evidence to show how these algorithms perform in a clinical setting. It is known that DIA can fail because of, for instance, poor slide quality,^{8,10} which could be a recurring challenge in the clinic. Moreover, AI systems are known to regress in performance when applied to data from sources that are not represented in the training data, because of so-called domain shift.¹¹

Recent studies have investigated human-in-the-loop (HITL) AI systems, in which pathologists interact with DIA systems, and have shown that an HITL approach can yield a performance improvement.^{12–15} In order to reach the full potential of the HITL approach, a deeper understanding of how these systems should be designed is needed.^{13,16}

In this study, we evaluated an existing HITL AI system in clinical use. The pathology department in Linköping, Sweden was an early adopter of digital pathology,¹⁷ and has, since 2015, used DIA applied to Ki67 for immunohistochemistry (IHC)-based intrinsic subtyping of luminal HER2-negative BCs; there have been >1600 such cases to date. The role of Ki67 in BC as a prognostic and predictive factor is debated, but it has been proven to correlate with a poor prognosis in primary BC, and is included in the St Gallen surrogate IHC classification for intrinsic subtypes.^{18–22} Ki67 is included in the Swedish guidelines for BC pathology, with a floating cut-off that removes the local bias component caused by differences in staining intensity and interpretation consensus.²³

To explore the effect that an AI system, with or without HITL interaction, has on diagnostic accuracy, we performed a retrospective study based on Ki67 DIA applied to hotspot areas on whole slide images of invasive BC by pathologists at the laboratory at Linköping University Hospital.

Materials and methods

Ethical approval was given by the Regional Ethical Review Board, in Linköping (Approval no. 2017/276-31).

SAMPLES AND IMAGE DATA

A total of 200 analysed areas, containing 200 tumour cells each, in Ki67-stained slides from different BC cases, produced by the Department of Clinical Pathology, Linköping, Sweden, were extracted from two time periods, starting in May 2015 and February 2017. One hundred and eighty-four of these cases were primary invasive BCs, and the remaining 16 cases were

local metastases or recurrences, distant metastases, or only suspected BC. Both the patient data and data on the reviewing pathologists were anonymised at the time of the extraction. The extracted areas corresponded to two different versions (denoted V1 and V2; same user interface) of the underlying nuclear detection algorithm, V1 from May 2015 and V2 from February 2017. From each time period, 100 consecutive BC cases were selected by the use of search tools in the pathology picture archiving and communication system (PACS) (Sectra AB, Linköping, Sweden) based on the combination of the local SNOMED code 'breast cancer' and existing 200-cell Ki67 annotations. Relevant clinical metadata were also extracted.

The Ki67 index in the annotations was checked for consistency with the value in the signed pathology report. With the local cut-off level of 41% positive cells in the hotspot, the Ki67 index was low in 140 specimens (71%). With a cut-off of 20%, the Ki67 index was low in 64 specimens (33%). Twenty-two cases were reported as HER2-positive (11%). Nottingham grading had been performed on 147 of the cases; 19 were grade 1, 78 were grade 2, and 50 were grade 3.

IHC

No additional staining procedure was performed for the study. All BC specimens had been processed with the established in-house clinical procedures, including the NordiQC external quality programme. Fixation times were semicontrolled within the range of 72 h. The routine protocol consisted of cutting one 4-µm-thick section from each formalin-fixed paraffin-embedded tumour specimen, and then mounting the section on a precoated slide for IHC analysis with anti-Ki67 rabbit monoclonal antibody (MIB1 30-9; Ventana/Roche, Tuscon, AZ, USA) by use of the Intellipath autostainer (Biocare Medical, Concord, CA, USA). Diaminobenzidine was used as a chromogen substrate for all specimens. All sections were counterstained with haematoxylin.

DIGITALISATION AND DIA

No additional scanning was performed for the study. All slides were originally scanned in routine clinical practice with either Aperio Scanscope AT Turbo scanners (Leica Biosystems, Buffalo Grove, IL, USA) or the Hamamatsu NanoZoomer XR (Hamamatsu, Photonics, Hamamatsu City, Japan), with application of the ×20 scanning resolution mode (□0.5 µm/pixel, JPEG Q70-80) in a single focus plane automatically

identified by the scanner. Quality control was performed manually.

The workstations for the original diagnostic work were equipped with an 8-megapixel colour-calibrated screen (Radiforce RX850; Eizo Corporation, Hakusan, Japan). All pathologists responsible for the cases included in this study had, at the time, >6 months of experience with the digital pathology system. At the time of the original diagnostic review, the pathologists did not know that the cases would be retrospectively investigated.

The nuclear detection algorithm used in the system was based on machine learning, with pixel annotations in V1²⁴ and a larger number of point annotations in V2.²⁵ The cell detection F1 scores were 0.79 for V1 and 0.83 for V2, as determined from an image dataset described by Molin *et al.*²⁶ The computer-aided tool was fully integrated in the viewer.

The Ki67 tool uses an HITL workflow in four steps (Figure 1). (1) The pathologist selects the area for analysis by placing a circle with a fixed radius on a hotspot. (2) The tool processes the area, detecting positive and negative cells. (3) The pathologist is presented with the result as an interactive visualisation, whereby it is possible to add or remove cells and change the positivity of cells, with corresponding updating of the Ki67 index. (4) When finished, the pathologist marks the result as verified.

GROUND TRUTH

The ground truth was established independently by a panel of three experienced breast pathologists (A.B., S.G., and R.W.), who manually marked every positive and negative cell in each area by using a web-based tool custom-built for this study. The image was interpreted without access to the full whole slide image or the corresponding haematoxylin and eosin-stained slide. An initial session established a joint view on cell categorisation by use of a separate training set. The median Ki67 indexes from the three panellists were used as the ground truth; one pathologist had previously been involved in the original HITL scoring. Fleiss' κ and standard deviation were calculated to enable the quality of the ground truth to be judged.

SCORING METHODS

The study evaluated the difference in accuracy between three different scoring methods: eyeballing, automatic scoring, and HITL scoring.

Eyeballing

Eyeballing was performed by three pathologists (A.B., S.G., and R.W.) before the ground truth elicitation. Eyeballing was performed by estimating the percentage, without decimals (e.g. 17%), of Ki67-positive tumour cells by visual assessment of the digital image, avoiding actual counting. A single digit was preferred over brackets because the use of brackets would introduce additional variability and therefore create an unfair advantage for the other scoring methods. A non-enforced timing guideline of 20 s for each area estimate was used.

Automatic scoring

The automatic scoring value is the output provided by the nuclear detection algorithm, retrieved from the pathology PACS, corresponding to the situation that would exist if the pathologist directly accepted the Ki67 index without modification after the processing step in Figure 1.

HITL scoring

The HITL scoring corresponds to the Ki67 index from the clinical review, retrieved from the pathology PACS, corresponding to the output of step 4 in Figure 1. The pathologist could either correct the algorithmic result or leave it unchanged.

STATISTICS

The statistical accuracy was calculated at three levels of detail: Ki67 status, Ki67 index, and individual cells. The Ki67 status refers to the binary assessment of being below or above the predetermined cut-off value (41%), as compared with ground truth. For the Ki67 index level, the accuracy is measured as the percentage point difference from the ground truth index value. The cell level compares the positivity for each individual cell with ground truth.

The statistical analysis was performed with PANDAS (0.25.2), SCIPY (1.3.1), and STATSMODELS (0.10.1). Statistical testing for the overall accuracy comparison was performed on the variability error on the Ki67 index in the V2 period by the use of Levene's test with Holm correction. Estimates of other variables are also reported, but no hypothesis testing was performed. The statistical testing plan was derived *a priori* from a smaller pilot dataset extracted in a similar manner without any cases overlapping with the cases included in the final study.

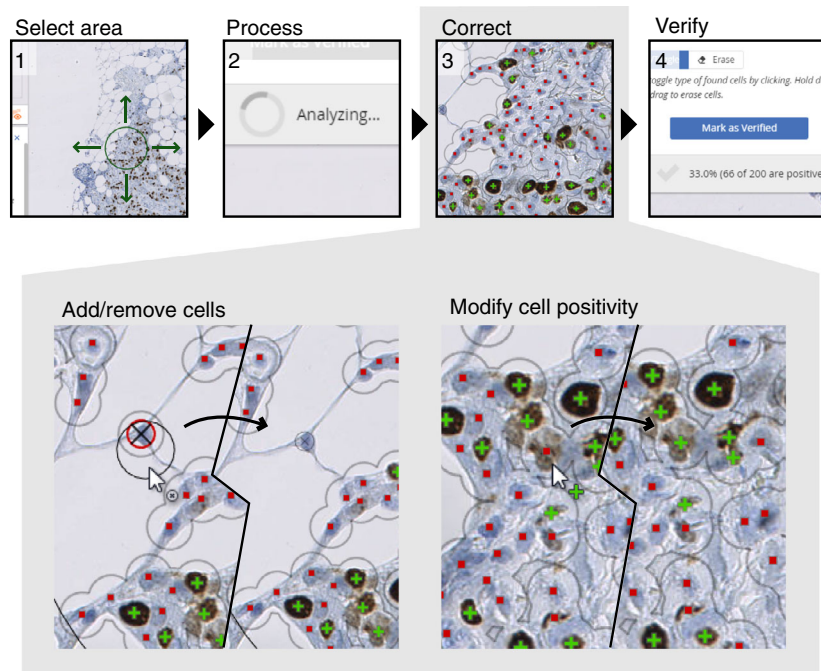


Figure 1. The Ki67 tool uses a human-in-the-loop workflow in four steps. (1) Pathologist selects an area. (2) Algorithmic processing. (3) Pathologist corrects cell status by adding/removing cells and/or modifying cell positivity (optional). (4) Pathologist verifies the result.

Results

For the ground truth assessments, Fleiss' κ between raters showed the Ki67 status-level variability to be 0.87. The estimated standard error of the ground truth Ki67 index was 2.9.

In the study, eight pathologists in total used the system, six in each period and four in both. The duration of tool use was measured from the start of processing of an area until the area had been marked as verified (steps 2–4 in Figure 1). The median durations were 113 s [interquartile range (IQR) 79–176] in V1 and 73 s (IQR 47–122) in V2.

ACCURACY OF SCORING METHODS

Ki67 status agreement, Ki67 index error and cell-level error are shown in Table 1. Automatic scoring improved the accuracy with both algorithm versions. The Ki67 index error for eyeballing (14.9) was significantly larger ($P < 0.05$) than those for automatic scoring (7.2) and HITL scoring (6.9), on comparison of the results in the V2 period. In general, HITL corrections did not result in an overall improvement in the Ki67 index as compared with automatic scoring. The overall correlation patterns for the different scoring methods are shown in Figure 2.

The HITL and automatic methods had lower variability, as shown in Figure 3, whereas they showed a slight statistical bias in that the observations clustered on the lower half, and the mean error deviated from zero as opposed to eyeballing. To further investigate possible clinical impacts of Ki67 status, discordances were mapped as status error. The performance of the algorithm dropped in the clinical environment; this was a domain shift, as reflected by an F1 score for V2 of 0.68 as compared with the formerly reported 0.83.²⁵

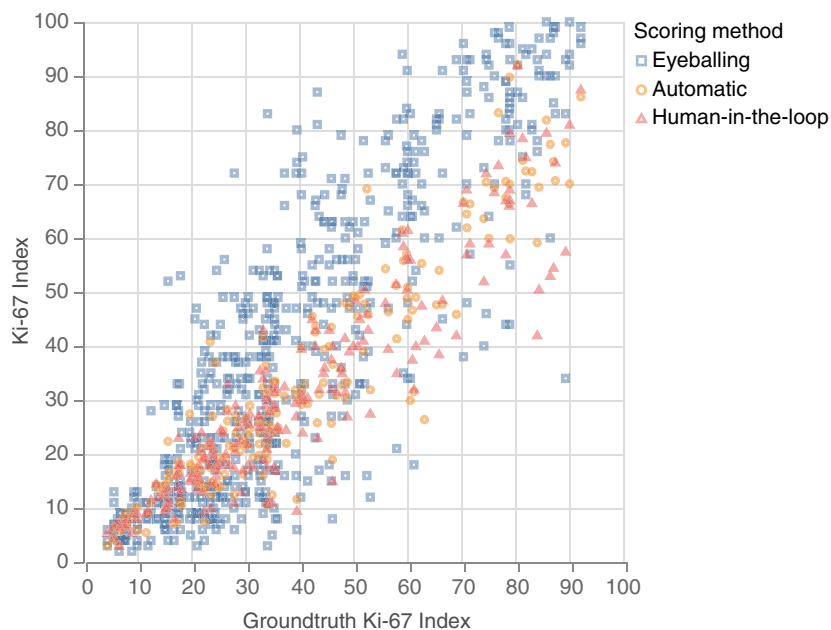
HITL

The effect of the pathologists' corrections, with regard to improving or worsening the automatic result in terms of Ki67 index error, is shown in Figure 4. A large number of cases had an error difference close to zero, showing that the corrections had little effect. In V1, corrections worsened the Ki67 index by, on average, −3.1 percentage points, and in V2 the corrections caused an improvement of, on average, 0.9 percentage points. This small effect does not mean that few corrections were made. Areas with an index error difference below 5 had, on average, 33 corrections. There was a slight difference in the number of corrections close to the cut-off; within 10 percentage

Table 1. Three levels of accuracy for each method as compared with ground truth; eyeballing, automatic, and human-in-the-loop, with algorithm versions (V1 and V2) separated

	Status agreement		Ki67 index error				Cell-level error	
	V1	V2	V1		V2		V1	V2
	Cohen's κ	Cohen's κ	Variability (SD)	Bias (mean)	Variability (SD)	Bias (mean)	F1 score	F1 score
Eyeballing	0.66	0.62	14.6	0.56	14.9	0.76	NA	NA
Automatic	0.73	0.84	8.4	−8.2	7.2	−6.0	0.66	0.68
Human-in-the-loop	0.66	0.76	9.1	−11.2	6.9	−5.4	0.68	0.68

SD, Standard deviation.

**Figure 2.** Overall correlation between the ground truth Ki67 index and the Ki67 index determined with each investigated scoring method. Note that there were three eyeballing sessions per area, but only one session per area for automatic and HITL, respectively.

points of the cut-off, the users performed, on average, 37.6 (± 5.5) corrections as compared with 42.1 (± 4.5) outside.

The annotations of all 200 cases, before and after corrections, were also visually inspected by one pathologist (A.B.). Some causes of suboptimal results were then found, and are shown in Figure 5. There were examples of user error in V1, as shown in Figure 5A. For many cases in which HITL caused deterioration, however, there was no obvious explanation other than the challenges of the diagnostic task as such. The visual inspection also revealed issues when the DIA system had failed and had been corrected by HITL. Root causes

included overlapping positive nuclei, as shown in Figure 5B–E, misidentification of tumour and stroma morphology, staining quality, or a combination of different factors.

Discussion

This is one of very few studies that have evaluated how an HITL AI system is used in clinical practice.

Our results, at all three levels analysed, are in support of the safe use of DIA. Agreement for Ki67 status was good, both in automatic and in assisted modes (automatic κ 0.84 and HITL κ 0.76). This is on a par

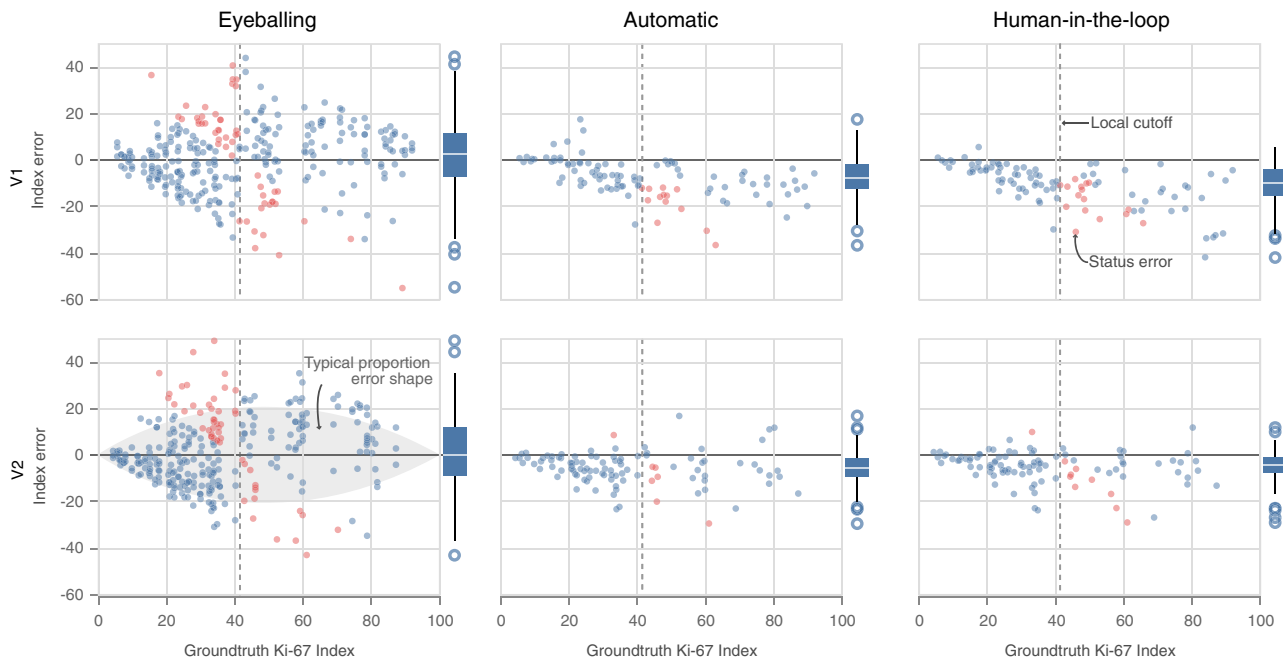


Figure 3. The Ki67 index error by scoring method and study periods. Red indicates when observations showed a discordance in Ki67 status. The eyeballing scoring method had high variability but was unbiased. The automatic and HITL scoring methods had lower variability but a small bias.

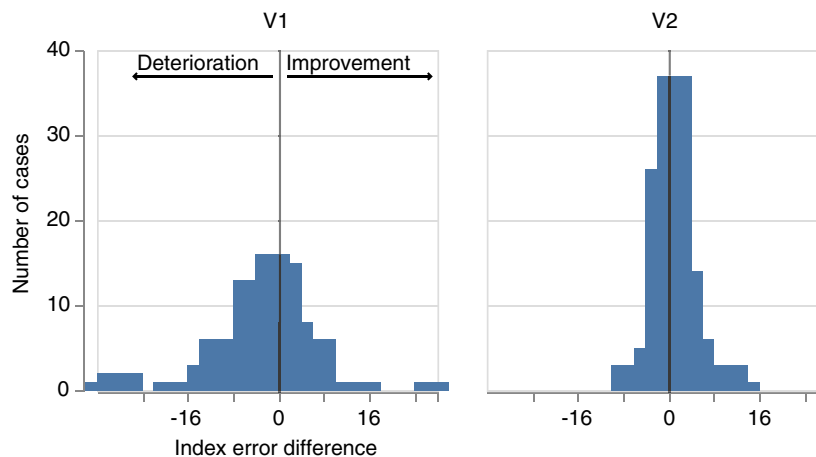


Figure 4. A histogram of case distribution illustrating the difference in Ki67 index error before and after human-in-the-loop correction. A negative value corresponds to a case in which corrections shifted the Ki67 index further away from the ground truth index, and a positive value corresponds to a case in which corrections shifted the Ki67 index towards the ground truth index.

with the Ki67 agreement previously found in the Swedish setting (κ 0.77).²⁷ With regard to the Ki67 index, which was the second level of evaluation, HITL and automatic DIA were significantly more accurate, with lower deviation of the standard error.

With respect to potential error sources in DIA use, our results are in line with those of Kwon *et al.*⁸ The

following causes of discordance between visual assessment and DIA regarding the Ki67 index in BC were identified: (i) tumour heterogeneity; (ii) visual assessment error; (iii) misidentification of tumour cells; (iv) poor immunostaining or slide quality; and (v) estimation of non-tumour cells. Additionally, we propose that the fourth category should also include

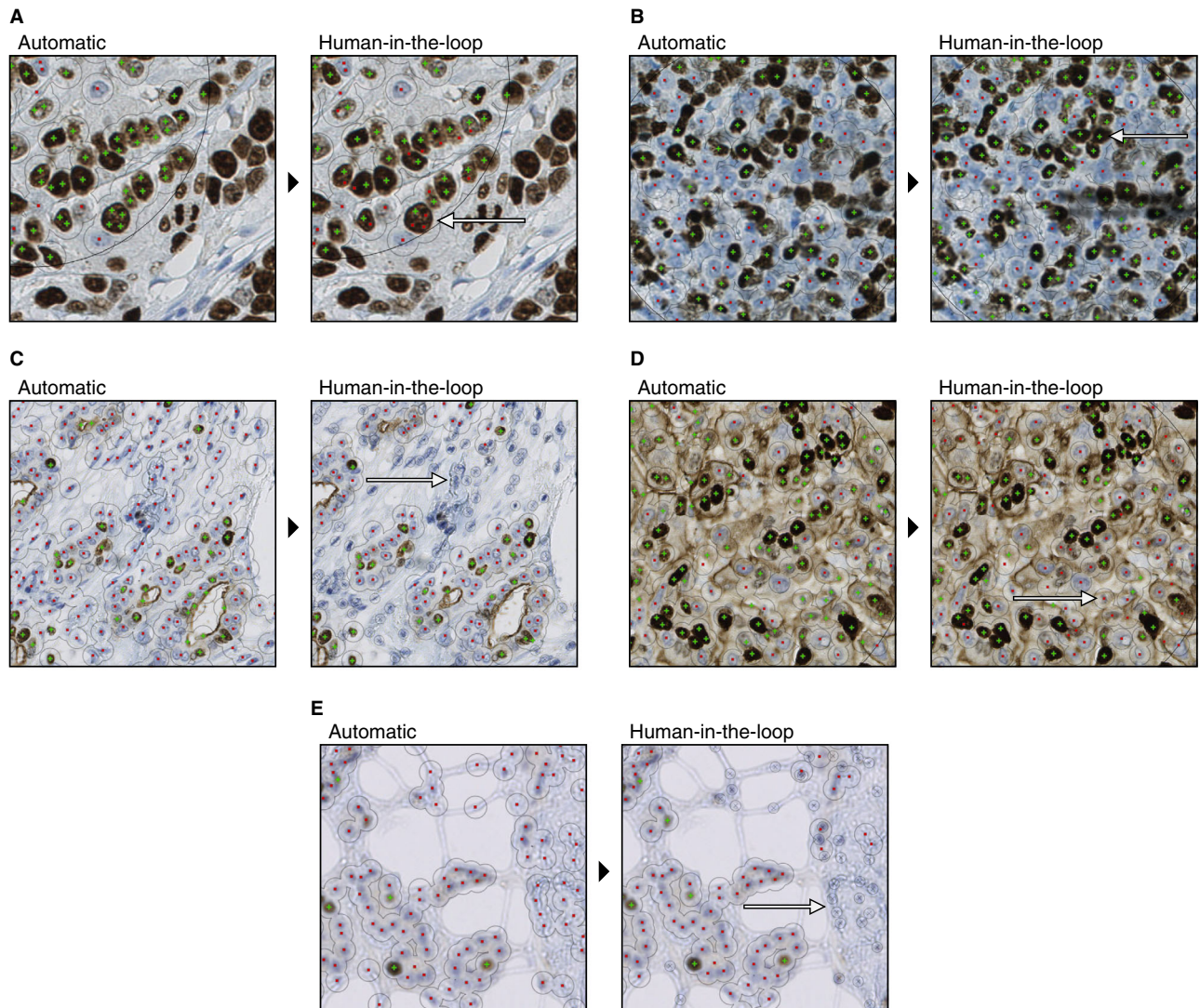


Figure 5. Manual inspection of the analysed areas (A–E) before and after correction revealed causes of suboptimal results. **A**, The pathologist has changed falsely detected positive nuclei into negative nuclei (arrow). It looks as though the intention must have been to remove these detections rather than modifying the positivity. **B**, The digital image analysis (DIA) system has failed to detect overlapping positive cells, and the user has correctly added the missing detections (arrow). **C**, The DIA system has poor performance when separating tumour from non-tumour cells. Here, the pathologist has removed excessive detections (arrow). **D**, Poor staining quality causes the DIA system to mark nuclei as positive even though the staining in this case is unspecific. The pathologist has correctly changed many nuclei from positive to negative (arrow). **E**, An out-of-focus artefact causes the algorithm to detect nuclei even though there are no nuclei at all. These have been removed by the pathologist (arrow).

digitisation quality, and that a sixth category should be introduced: user handling error.

A main question is whether the HITL intervention of the pathologists is justified in comparison with automatic AI in the analysis of Ki67 in routine BC diagnostics. Our findings are not as categorical as those of recently published studies of AI assistance within pathology, which have indicated either improvement or failure.^{13,28,29} Our results reveal individual cases in which poor DIA results can be

remedied by pathologist HITL intervention, but human adjustments can also worsen the results.

Justifying pathologist interaction is also dependent on the time spent on performing the HITL task. In our study, the time spent was substantially improved in V2, being 40 s lower. The improvement can partly be attributed to the higher accuracy of the nuclear detection algorithm, resulting in fewer corrections, but other causes cannot be excluded. The diagnostic review in pathology is a complex mixture of

quantitative and qualitative measures, making it challenging to evaluate results in terms of added value in the diagnostic process. One weakness of this study is that we did not investigate the pathologists' experience regarding the usage and perceived benefit of the tool. One clear indication is, however, that the tool is frequently and consistently used among breast pathologists in everyday practice in Linköping.

Even though our study shows that, for these 200 cases, the pathologist could have been removed from the loop and time could have been saved without substantially worsening the results, it is uncertain whether this result is generalisable. The HITL approach provides an important safety mechanism for detecting and correcting algorithmic errors that may occur. Retaining human oversight of analysis outputs is often a recommendation for quality control and safety in digital image analysis^{9,30}. Key issues for successful implementation of AI assistance in clinical routine will be to determine where the limited time of pathologists is best used and to identify the value added by AI in the diagnostic process, focusing on the resulting benefit for patient care in terms of accurate quantification (as facilitated through AI for predictive scoring).

Summary

We have shown that DIA applied in real-world clinical routine, both in automated and in HITL settings, contributes to more accurate scoring of Ki67 in patients with BC. The main finding is that the primary value of HITL correction is to detect major weaknesses of the DIA algorithm rather than fine-tuning by analysing individual cells.

Financial support

This work is supported by an ALF grant from Region Östergötland.

Acknowledgements

This work was supported by grants from ALF Grants, Region Östergötland. The authors are grateful for the pathologists who contributed substantially by using the Ki67 aid tool in clinical practice.

Author contributions

A. C. S. Bodén, J. Molin, and D. Treanor: study concept, design, and planning. A. C. S. Bodén and J. Molin: data

gathering and case collection, review, anonymisation, and analysis. A. C. S. Bodén, S. Garvin, and R. A. West: evaluation of ground truth and eyeballing. J. Molin and A. C. S. Bodén: statistical analysis and figure design. A. C. S. Bodén, J. Molin, C. Lundström, and D. Treanor: manuscript preparation and writing.

Conflict of interest

J. Molin and C. Lundström are with Sectra AB, Linköping, Sweden.

References

1. Goacher E, Randell R, Williams B, Treanor D. The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. *Arch. Pathol. Lab. Med.* 2017; **141**: 151–161.
2. Snead DRJ, Tsang Y-W, Meskiri A *et al.* Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016; **68**: 1063–1072.
3. Mukhopadhyay S, Feldman MD, Abels E *et al.* Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). *Am. J. Surg. Pathol.* 2018; **42**: 39–52.
4. Cheng CL, Azhar R, Sng SH *et al.* Enabling digital pathology in the diagnostic setting: navigating through the implementation journey in an academic Medical Centre. *J. Clin. Pathol.* 2016; **69**: 784–792.
5. Gudlaugsson E, Skaland I, Janssen EAM *et al.* Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology* 2012; **61**(6): 1134–1144.
6. Stålhammar G, Martinez NF, Lippert M *et al.* Digital image analysis outperforms manual biomarker assessment in breast cancer. *Mod. Pathol.* 2016; **29**: 318–329.
7. Acs 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.
8. Kwon A-Y, Park HY, Hyeon J *et al.* Practical approaches to automated digital image analysis of Ki-67 labeling index in 997 breast carcinomas and causes of discordance with visual assessment. *PLoS One* 2019; **14**: 1–14.
9. Riber-Hansen R, Vainer B, Steiniche T. Digital image analysis: a review of reproducibility, stability and basic requirements for optimal results. *APMIS* 2012; **120**(4): 276–289.
10. Liu Y, Kohlberger T, Norouzi M *et al.* Artificial intelligence-based breast cancer nodal metastasis detection insights into the black box for pathologists. *Arch Pathol Lab Med* 2019; **143**: 859–868.
11. Stacke K, Eilertsen G, Unger J, Lundström C. A Closer Look at Domain Shift for Deep Learning in Histopathology. 2019.
12. Gavrielides MA, Miller M, Hagemann IS *et al.* Clinical decision support for ovarian carcinoma subtype classification: a pilot observer study with pathology trainees. *Arch. Pathol. Lab. Med.* 2019. <https://doi.org/10.5858/arpa.2019-0390-0a>.
13. Steiner DF, MacDonald R, Liu Y *et al.* Impact of deep learning assistance on the histopathologic review of lymph nodes for

- metastatic breast cancer. *Am. J. Surg. Pathol.* 2018; **42**(12); 1636–1646.
14. Bulten W, Balkenhol M, Belinga J-JA *et al.* Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod. Pathol.* 2021; **34**(3); 660–671.
 15. Kiani A, Uyumazturk B, Rajpurkar P *et al.* Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit. Med.* 2020; **3**; 1–8.
 16. Lindvall M, Molin J, Löwgren J. From machine learning to machine teaching: the importance of UX. *Interactions* 2018; **25** (6); 52–57.
 17. Thorstenson S, Molin J, Lundström C. Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: digital pathology experiences 2006–2013. *J. Pathol. Inform.* 2014; **5**; 14.
 18. 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.
 19. Curigliano G, Burstein HJ, Winer EP *et al.* De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. *Ann. Oncol.* 2017; **28**; 1700–1712.
 20. 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**(22); 1656–1664.
 21. Denkert C, Budczies J, von Minckwitz G *et al.* Strategies for developing Ki67 as a useful biomarker in breast cancer. *Breast* 2015; **24**; S67–S72.
 22. Duffy MJ, Harbeck N, Nap M *et al.* Clinical use of biomarkers in breast cancer: updated guidelines from the European Group on Tumor Markers (EGTM). *Eur. J. Cancer* 2017; **75**; 284–298.
 23. Romero Q, Bendahl P-O, Fernö M, Grabau D, Borgquist S. A novel model for Ki67 assessment in breast cancer. *Diagn. Pathol.* 2014; **9**; 118.
 24. Kost H, Homeyer A, Bult P *et al.* A generic nuclei detection method for histopathological breast images. *Digit. Pathol. Proc.-SPIE* 2016; **9791**; 97911E.
 25. Kost H, Homeyer A, Molin J, Lundström C, Karl HH. Training nuclei detection algorithms with simple annotations. *J. Pathol. Informatics* 2017. https://doi.org/10.4103/jpi.jpi_3_17.
 26. Molin J, Bodén A, Treanor D, Fjeld M, Lundström C. Scale Stain: Multi-Resolution Feature Enhancement in Pathology Visualization. 2016. *arXiv1610.04141v*. <https://arxiv.org/abs/1610.04141> 1–14.
 27. Ekholm M, Grabau D, Bendahl P-O *et al.* Highly reproducible results of breast cancer biomarkers when analysed in accordance with national guidelines – a Swedish survey with central re-assessment. *Acta Oncol.* 2015; **54**(7); 1040–1048.
 28. Uyumazturk B, Kiani A, Rajpurkar P *et al.* Deep Learning for the Digital Pathologic Diagnosis of Cholangiocarcinoma and Hepatocellular Carcinoma: Evaluating the Impact of a Web-based Diagnostic Assistant. *ArXiv191107372 Eess*. 2019.
 29. Bulten W, Balkenhol M, Belinga J-J *et al.* Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod. Pathol.* 2021; **34**(3); 660–671.
 30. Aeffner F, Zarella MD, Buchbinder N *et al.* Introduction to digital image analysis in whole-slide imaging: a white paper from the Digital Pathology Association. *J. Pathol. Inform.* 2019; **10**; 9.