



Deep learning supported mitoses counting on whole slide images: A pilot study for validating breast cancer grading in the clinical workflow

Stijn A. van Bergeijk^{a,1}, Nikolas Stathonikos^{a,1}, Natalie D. ter Hoeve^a, Maxime W. Lafarge^{b,c}, Tri Q. Nguyen^a, Paul J. van Diest^{a,*}, Mitko Veta^b

^a Department of Pathology, University Medical Center Utrecht, Postal Box 85500, 3508 GA Utrecht, The Netherlands

^b Medical Image Analysis Group (IMAG/e), Eindhoven University of Technology, Eindhoven, The Netherlands

^c Computational and Translational Pathology Group, Department of Pathology and Molecular Pathology, University Hospital and University of Zürich, Schmelzbergstrasse 12, 8091 Zurich, Switzerland

ARTICLE INFO

Keywords:

Breast cancer
Artificial intelligence
Digital pathology
Light microscopy
Bloom & Richardson grade
Mitotic index

ABSTRACT

Introduction: Breast cancer (BC) prognosis is largely influenced by histopathological grade, assessed according to the Nottingham modification of Bloom-Richardson (BR). Mitotic count (MC) is a component of histopathological grading but is prone to subjectivity. This study investigated whether mitoses counting in BC using digital whole slide images (WSI) compares better to light microscopy (LM) when assisted by artificial intelligence (AI), and to which extent differences in digital MC (AI assisted or not) result in BR grade variations.

Methods: Fifty BC patients with paired core biopsies and resections were randomly selected. Component scores for BR grade were extracted from pathology reports. MC was assessed using LM, WSI, and AI. Different modalities (LM-MC, WSI-MC, and AI-MC) were analyzed for correlation with scatterplots and linear regression, and for agreement in final BR with Cohen's κ .

Results: MC modalities strongly correlated in both biopsies and resections: LM-MC and WSI-MC (R^2 0.85 and 0.83, respectively), LM-MC and AI-MC (R^2 0.85 and 0.95), and WSI-MC and AI-MC (R^2 0.77 and 0.83). Agreement in BR between modalities was high in both biopsies and resections: LM-MC and WSI-MC (κ 0.93 and 0.83, respectively), LM-MC and AI-MC (κ 0.89 and 0.83), and WSI-MC and AI-MC (κ 0.96 and 0.73).

Conclusion: This first validation study shows that WSI-MC may compare better to LM-MC when using AI. Agreement between BR grade based on the different mitoses counting modalities was high. These results suggest that mitoses counting on WSI can well be done, and validate the presented AI algorithm for pathologist supervised use in daily practice. Further research is required to advance our knowledge of AI-MC, but it appears at least non-inferior to LM-MC.

Introduction

The yearly worldwide breast cancer (BC) incidence is over 2 million, which makes it the most diagnosed cancer. Female BC currently occupies the fifth place in cancer mortality worldwide, and incidence keeps rising.¹ However, when diagnosed in an early stage, the prognosis of BC can be good.^{1,2} One of the strongest factors to determine BC prognosis is histological grade, usually assessed according to the Nottingham modification of Bloom-Richardson (BR) grade.^{3,4} BR requires the pathologist to score 3 features: tubule formation, nuclear pleomorphism, and mitotic count (MC). Each category gets a score from 1 to 3. Scores 3–5 define grade 1, 6–7 grade 2, and 8–9 make up grade 3 BC. Grade 1 cancers have a significantly better survival than grade 2 or 3 cancers.^{3,5,6} Studies have shown histological grading, tumor size, and lymph node status to be of equal importance

for the prognosis of BC.^{5,6} Furthermore, histological grade proved to be decisive in up to a third of treatment decisions.⁷

MC is, as a marker of tumor proliferation, the strongest constituent of BR grade, and a high MC is associated with poor prognosis.^{8–10} Several studies have shown a moderate to good reproducibility for BR.^{11–13} When focusing solely on MC, reproducibility also ranges from moderate to high.^{14,15} However, concerns for reproducibility still exist as 1 recent study again found substantial inter- and intra-laboratory variations in BR in more than 33 000 patients.⁷ Because of these variations and the importance of MC for the prognosis of BC, higher reproducibility is required.

With the development of digital whole slide imaging (WSI), breast cancer diagnostics have increasingly been performed digitally as WSI have been validated for diagnostic purposes.^{16,17} It has been argued that standard WSI has limitations for reliable histologic grading, as the quality of the images

* Corresponding author at: Department of Pathology, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands.

E-mail address: p.j.vandiest@umcutrecht.nl (P.J. van Diest).

¹ These authors contributed equally.

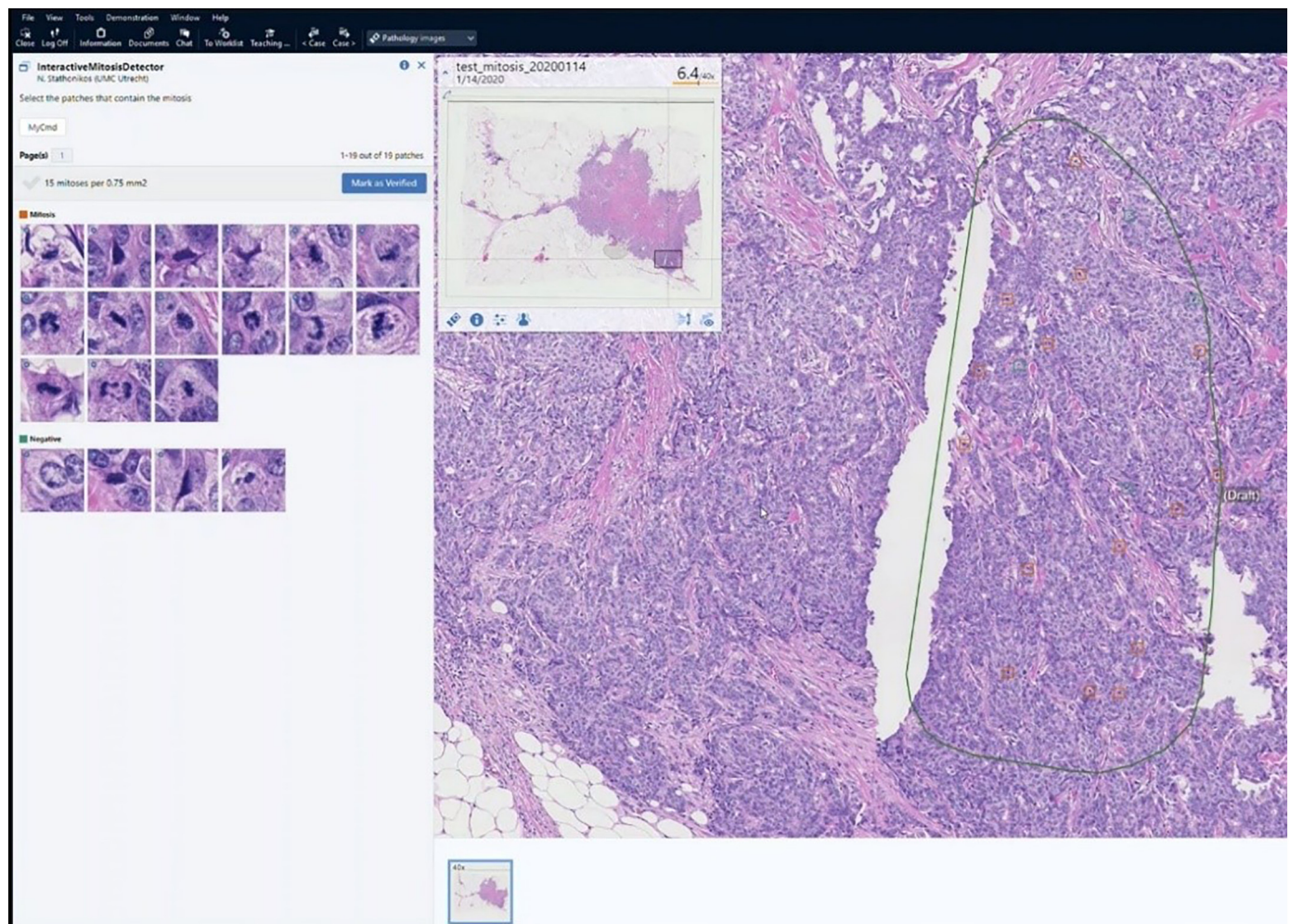


Fig. 1. Screenshot of the Sectra PACS where an area of interest has interactively been drawn on the right-hand side, after which an AI algorithm has found candidate mitoses and mitosis-like objects, which are displayed in the galleries in the upper left-hand side of the screen. By clicking on a thumbnail in either of the galleries, the PACS displays the candidate object in the center on the right for review, and false positives can be dragged to the negative gallery and vice versa, after which a final AI supported MC is established.

may not be high enough for properly assessing the MC in all cases due to lack of a z-axis (i.e. fine-tuning of the focal length), which pathologists often use when microscopically assessing MC. Pathologist familiarity with WSI in the clinical workflow might also be limiting factor. Also, a change in ergonomics is required when using a computer mouse instead of a microscope which might further influence pathologist opinion on WSI. Two studies have shown that MCs in WSI and traditional light microscopy (LM) show comparable results.^{18,19} However, other studies suggest that although the inter-observer agreement on WSI is similar to LM, MC tends to be systematically lower on WSI.^{16,17,20,21}

The increased usage of WSI has stimulated the rise of artificial intelligence (AI) algorithms in pathology. Several of these have been developed for assisting the pathologist in performing MC, expecting to improve the reproducibility of MC, often tested in validation cohorts.^{5,19,22–27} The next step is to test AI algorithms in a clinical setting. The present study validates an in-house developed AI algorithm for mitoses counting in BC on digital WSI by comparing AI supported MC to light microscopic MC and evaluating influence of putative differences of these MC modalities on BR grade in breast cancer.

Methods

Study design and population

Fifty BC patients with paired core biopsies and resections were randomly selected from the workflow of the Department of Pathology at the

UMC Utrecht between December 2018 and February 2020. For each patient, tubular differentiation (scored 1, 2, or 3) and nuclear polymorphism scores (1, 2, or 3) according to Elston and Ellis³ were taken from the original pathology report (14 grade 1, 28 grade 2, and 8 grade 3). An approval from our Institutional Review Board was requested and granted under the application number TCBio-20-777.

An experienced Pathologist Assistant (PA) trained in breast microscopy first determined the most cellular and proliferative area of the tumor using LM without prior knowledge of the BR grade and MC. The MC was reassessed using LM (LM-MC) in 2 mm² of adjacent fields.¹⁴ After getting the exact count, MC was scored as 1, 2, or 3 points, for respectively ≤ 7 , 8–12, and ≥ 13 mitoses. After a washout period of at least 2 months, MC was assessed digitally using WSI (WSI-MC), and after another 2 months washout period, MC was assessed supported by the AI algorithm (AI-MC).

Prior to start using the AI algorithm, a standard operation procedure document (SOP) was made for the AI tool and the PA was trained on the usage of the tool on the test PACS environment.

Digital pathology and AI

Slides had routinely been scanned within the workflow of the UMC Utrecht at 40 \times magnification (resolution of 0.22 μ m per pixel) with a Nanozoomer 2.0-XR (Hamamatsu, Japan). All WSI were viewed using standard high-resolution 4k computer screens in the Sectra PACS (Linköping, Sweden).

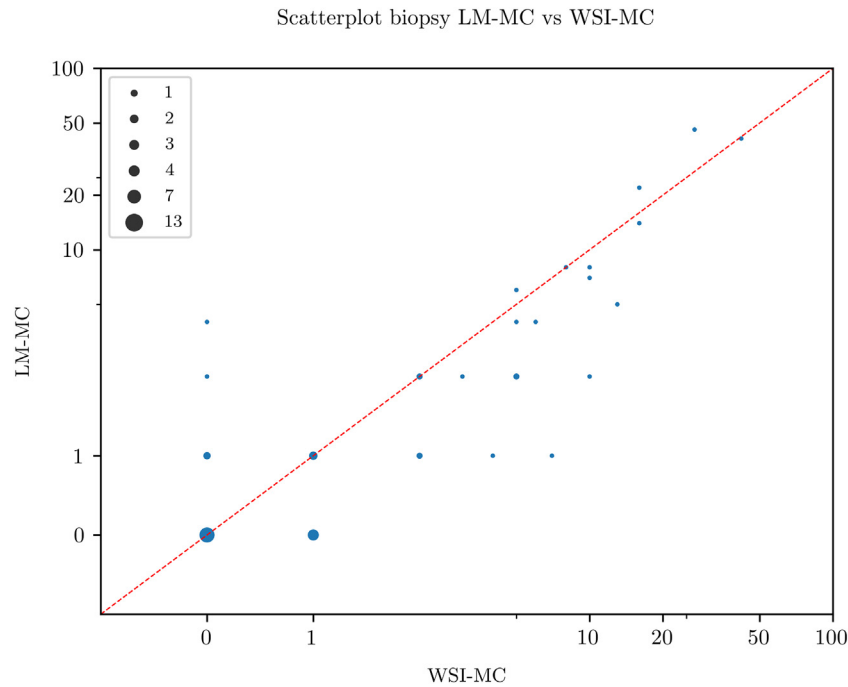


Fig. 2. Scatterplot showing a high concordance between whole slide image-based digital mitotic count (WSI-MC) and light microscopic MC (LM-MC) in 50 breast cancer biopsies.

The automated mitosis detection system was developed internally based on the methodology introduced by Cireřan et al.²⁸ and the improvements upon this work by Lafarge et al.²⁹ The model was trained using Tensorflow 1.12 on python 2.7 and is based on rotation invariant group convolutional neural networks. We used the TUPAC16 and AMIDA13 grand challenge (GC) dataset to train the network as well as a smaller annotated dataset containing mostly hard-negatives and ink artifacts to improve robustness. Most GC datasets include examples from within the tumor and rarely from the periphery of the slide—here the most ink artifacts and other

mimics are found—which can lead to performance degradation when whole slide inference is performed.

The model is a 6-layer group CNN, the architecture is extensively described in Lafarge et al.²⁹ In short, we used a patch size of 68×68 pixels with a batch size of 64 and it was trained on NVIDIA K80 and NVIDIA V100 hardware. We evaluated the performance of the model on test sets of the GC datasets and used the F2-score threshold for the clinical implementation. The F2-score threshold gives more weight to recall than precision in contrast to F1-score which gives equal weight to both. This

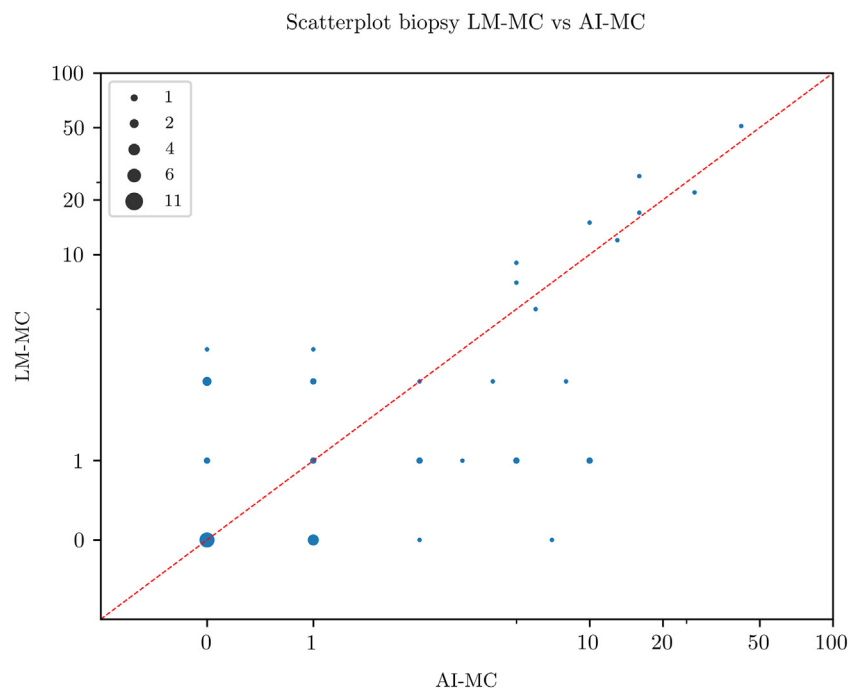


Fig. 3. Scatterplot showing a high concordance between artificial intelligence-based mitotic count (AI-MC) and light microscopic MC (LM-MC) in 50 breast cancer biopsies.

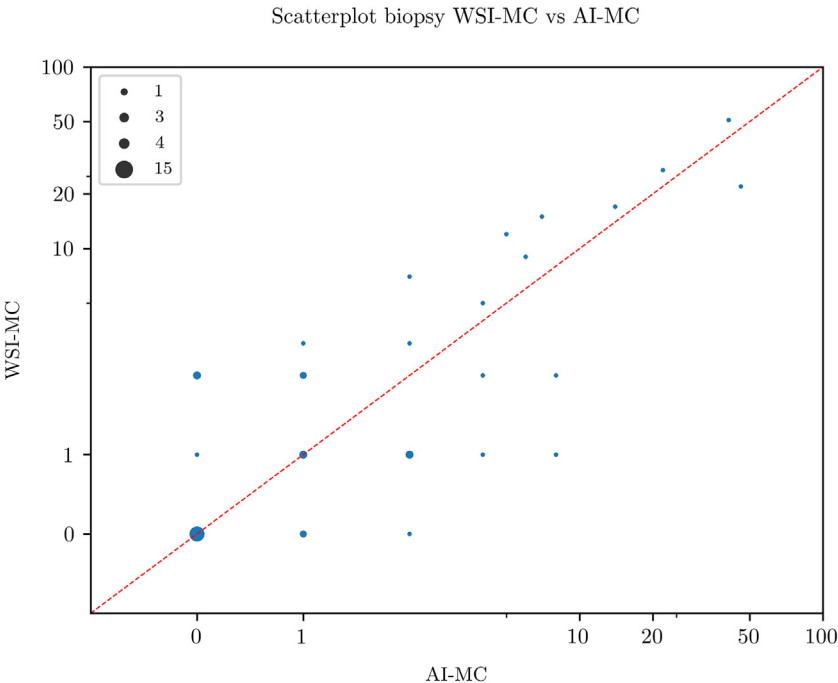


Fig. 4. Scatterplot showing a high concordance between artificial intelligence-based mitotic count (AI-MC) and whole slide image-based digital MC (WSI-MC) in 50 breast cancer biopsies.

threshold allows the pathologist to review more objects while not overwhelming them with too many objects to review.

The model takes large image patch of $40\times$ resolution and generates a probability map of that patch. Then by using local-maxima extraction, it gets the positions of mitosis on that patch. The MC AI algorithm (both model and integration with PACS) was in-house developed. In the Sectra PACS, an area of interest of the appropriate size of 2 mm^2 (as described for LM-MC) is interactively drawn, after which the algorithm automatically identifies candidate mitoses and mitoses-like objects and displays them in 2 galleries. Objects are interactively reviewed and dragged to the correct gallery, resulting in a final AI MC per 2 mm^2 (Fig. 1).

Data analysis

Using the MC from the 3 modalities, 3 BR grades were composed for each biopsy and resection as usual by summing up the scores from tubular differentiation, nuclear polymorphism, and MC, total score 3–5 defining grade 1, scores 6–7 grade 2, and scores 8–9 grade 3. Data for biopsies and resections were separately analyzed. MC data were pairwise displayed in logarithmic scatterplots with reference lines between the different MC modalities and R^2 was calculated to detect systematic differences. To assess the concordance in BR resulting from the different MC modalities, crosstabs were created, using Cohen’s κ to assess BR agreement between the different MC modalities.³⁰ Scores of 0 meant no agreement, 0.01–0.20 none to slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.8 substantial, and 0.81–1.00 almost perfect agreement.³² All statistics were done using Python version 3.8.5. and scikit-learn 1.0.2 and pingouin 0.5.2 python packages.

Results

Biopsies

Scatterplots for pairwise comparison between the 3 MC modalities are shown in Figs 2, 3 and 4. All MC modalities were strongly correlated: R^2 between LM-MC and WSI-MC was 0.85, 0.85 between LM-MC and AI-MC, and 0.77 between WSI-MC and AI-MC.

Table 1

Crosstab between Bloom & Richardson (BR) grade based on light microscopic mitotic count (LM-MC) and artificial intelligence supported MC (AI-MC) in 50 breast cancer biopsies ($\kappa=0.894$, 95% CI 0.78–1.01).

LM-MC-based grade	AI-MC-based grade			Total
	1	2	3	
1	17	0	0	17
2	1	26	1	28
3	0	1	4	5
Total	18	27	5	50

The crosstabs for the BR grades resulting from the different MC modalities are shown in Table 1, 2 and 3, all showing high κ values: 0.93 for LM-MC versus WSI-MC-based BR, 0.89 for LM-MC versus AI-MC-based BR, and 0.96 for WSI-MC versus AI-MC-based BR.

Resections

Scatterplots for pairwise comparison between the 3 MC modalities are shown in Figs 5, 6, and 7. All MC modalities were strongly correlated: R^2 between LM-MC and WSI-MC was 0.83, 0.95 between LM-MC and AI-MC and 0.83 between WSI-MC and AI-MC.

The crosstabs for the BR grades resulting from the different MC modalities are shown in Table 4, 5 and 6, all showing high κ values: 0.83

Table 2

Crosstab between Bloom & Richardson (BR) grade based on light microscopic mitotic count (LM-MC) and whole slide image-based digital MC (WSI-MC) in 50 breast cancer biopsies ($\kappa=0.928$, 95% CI 0.83–1.01).

LM-MC-based grade	WSI-MC-based grade			Total
	1	2	3	
1	17	0	0	17
2	1	27	0	28
3	0	1	4	5
Total	18	28	4	50

Table 3

Crosstab between Bloom & Richardson (BR) grade based on whole slide image-based digital mitotic count (WSI-MC) and artificial intelligence supported MC (AI-MC) in 50 breast cancer biopsies ($\kappa = 0.964$, 95% CI 0.90–1.03).

WSI-MC-based grade	AI-MC-based grade			Total
	1	2	3	
1	17	0	0	17
2	1	26	1	28
3	0	1	4	5
Total	18	27	5	50

for LM-MC-based BR versus WSI-MC, 0.83 for LM-MC versus AI-MC-based BR, and 0.73 for WSI-MC versus AI-MC-based BR.

Discussion

In this study, we investigated whether mitoses counting in BC using digital WSI compares better to LM-MC when assisted by AI, and to which extent differences in digital MC (AI assisted or not) result in BR grade variations.

For biopsies, LM-MC and AI-MC showed an equal R^2 when compared with LM-MC and WSI-MC, despite the latter already correlating well. For resections, R^2 of LM-MC and AI-MC even surpassed that of LM-MC and WSI-MC. This data suggests that not only does AI correlate as well as WSI with LM for mitotic count, but also might perhaps compare better to LM.

It was noted that AI-MC resulted in systematically slightly lower MC values compared to LM-MC and WSI-MC. This indicates that the AI algorithm may miss some mitoses and needs further improvement. However, as the observer checked the results, the observer may not have been critical enough when reviewing mitoses which the AI classified as mitoses-like objects. This could lower AI-MC compared to LM-MC and WSI-MC and underlines the importance of careful human supervision of the output of algorithms when AI is used in daily practice.

Several other studies showed similar results regarding the comparability between LM-MC and WSI-MC.^{16,20,31,32} Noted differences between LM-MC and WSI-MC were perceived to be within the range of

inter-observer differences in LM-MC. Also, studies which used $40\times$ magnification for scanning and high-resolution displays noted that differences between WSI and LM tended to get smaller, suggesting that a certain standard of technology is required for proper mitoses counting on WSI. As to AI, a recent study applying AI to select a mitoses hotspot in which to count showed improved inter-observer agreement in interactive mitoses counting on WSI, with similar inter-observer κ values for LM-MC and AI-MC.¹⁹ However, one study demonstrated higher inter-observer agreement for AI-MC compared to LM-MC, and a substantial saving in time.³³ So, different studies seem to point at least to non-inferiority of AI-MC compared to LM-MC in BC. The potential to save time is another reason to further explore the possibilities of AI.

Both biopsies and resections showed near perfect agreement in BR between different modalities, although the κ for WSI-MC versus AI-MC-based BR in the resection group was slightly lower. This indicates that differences in MC between different modalities hardly influence BR grade.

One study compared BR based on LM and WSI in over 1600 cases, showing a strong association (Cramer's V: 0.58) between both modalities.¹⁶ Another study focusing on inter-observer differences in BR when using WSI, showed the concordance to be similar to inter-observer differences in BR using LM.²¹ These studies substantiate our results. To the best of our knowledge, no previous study has been conducted that compares agreement of BR using LM-MC or WSI-MC and AI-MC. The high agreement in BR in this study is probably related to 2 factors. Firstly, WSI-MC and AI-MC were performed on the exact same slide as LM-MC, whereas larger tumors may be heterogeneous across different tissue blocks. Secondly, grading in different modalities was assessed by the same observer, causing the criteria for mitotic figures to be interpreted singularly and increasing the chance of selecting the same hotspot.

This study has some limitations. First, the gold-standard is LM-MC assessed by a single observer. Due to significant inter-observer differences for LM-MC, a study with multiple observers may provide a more realistic view on the added value of AI. Another option would be to use Phosphohistone H3 immunohistochemistry, which enhances recognition of mitotic figures and may make LM-MC (and perhaps even AI-MC) more reproducible.³⁴ Secondly, this study has a relatively small number of cases.

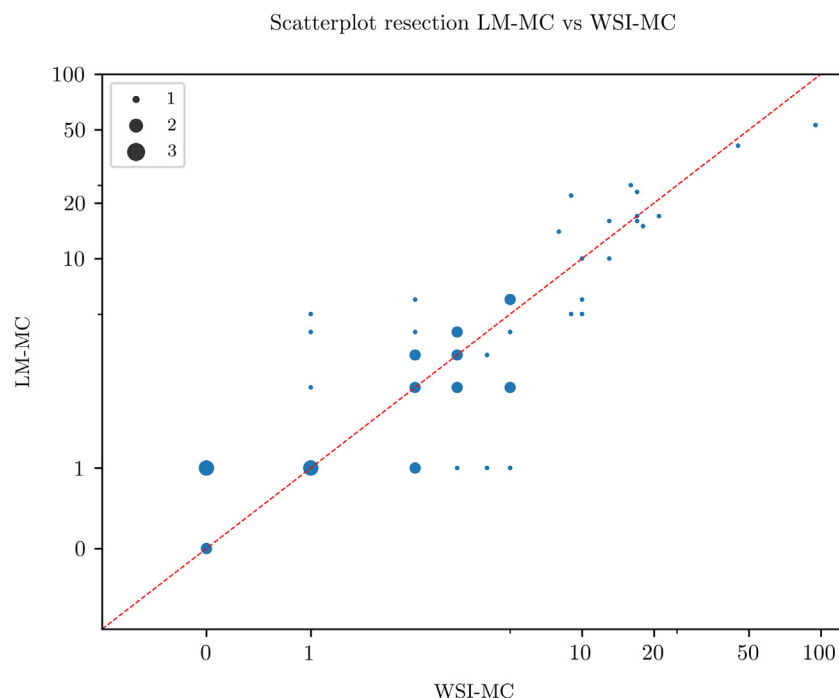


Fig. 5. Scatterplot showing a high concordance between whole slide image-based digital mitotic count (WSI-MC) and light microscopic MC (LM-MC) in 50 breast cancer resections.

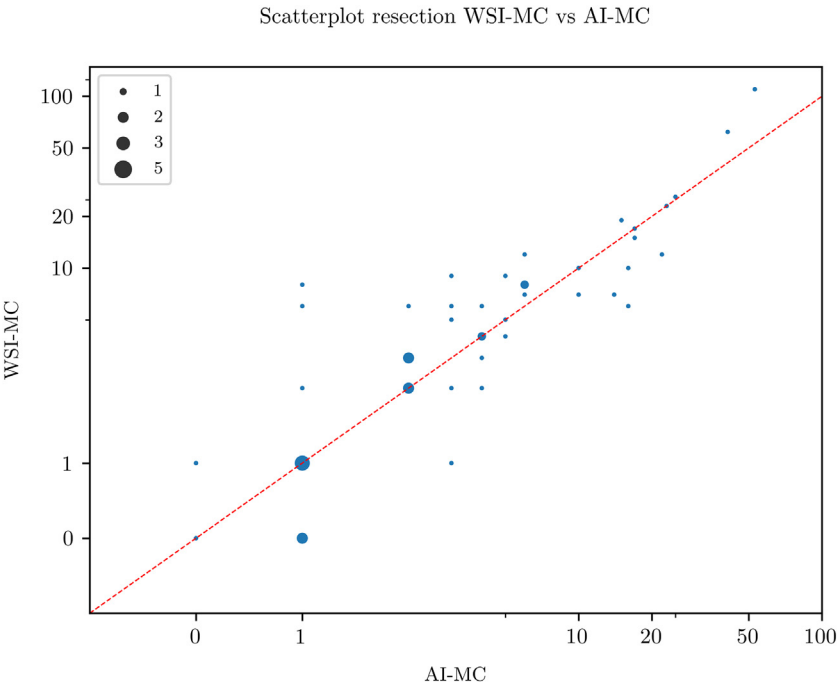


Fig. 6. Scatterplot showing a high concordance between artificial intelligence-based mitotic count (AI-MC) and light microscopic MC (LM-MC) in 50 breast cancer resections.

In daily pathology practice, digital WSI is increasingly used worldwide. This study, in combination with previous studies in this field, shows WSI-MC to be suitable for grading BC. Especially pathology laboratories which have a digital workflow could thereby incorporate WSI-MC in their daily practice of grading BC.

In general, AI algorithms show great promise in improving pathology practice. This study demonstrates that mitoses counting in BC can not

only be performed by an AI algorithm, but also might compare better to LM than WSI. We expect the next generation algorithms to be improved even further.³⁵ These algorithms may also save valuable interaction time for the pathologist, especially when algorithms run in the background on WSI, providing the pathologist with mitotic hotspots.

In conclusion, this first validation study shows that WSI-MC might compare better to LM-MC by using AI. Agreement between different modalities

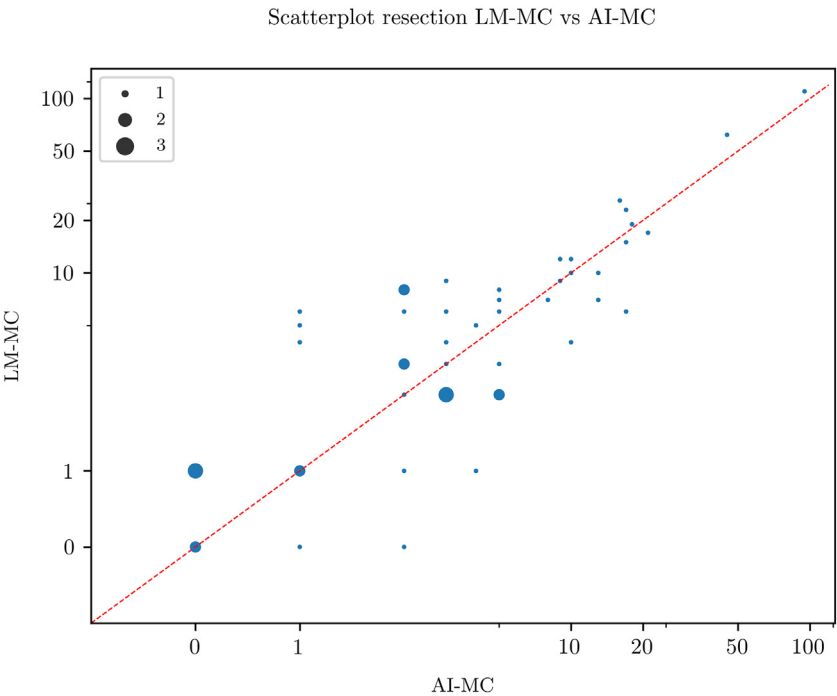


Fig. 7. Scatterplot showing a high concordance between artificial intelligence-based mitotic count (AI-MC) and whole slide image-based digital MC (WSI-MC) in 50 breast cancer resections.

Table 4

Crosstab between Bloom & Richardson (BR) grade based on light microscopic mitotic count (LM-MC) and whole slide image-based digital MC (WSI-MC) in 50 breast cancer resections ($\kappa = 0.834$, 95% CI 0.70–0.97).

LM-MC-based grade	WSI-MC-based grade			Total
	1	2	3	
1	13	1	0	14
2	2	24	2	28
3	0	0	8	8
Total	15	25	10	50

Table 5

Crosstab between Bloom & Richardson (BR) grade based on light microscopic mitotic count (LM-MC) and artificial intelligence supported MC (AI-MC) in 50 breast cancer resections ($\kappa = 0.825$, 95% CI 0.68–0.97).

LM-MC-based grade	AI-MC-based grade			Total
	1	2	3	
1	13	1	0	14
2	2	26	0	28
3	0	2	6	8
Total	15	29	6	50

Table 6

Crosstab between Bloom & Richardson (BR) grade based on whole slide image-based digital mitotic count (WSI-MC) and artificial intelligence supported MC (AI-MC) in 50 breast cancer resections ($\kappa = 0.732$, 95% CI 0.56–0.90).

WSI-MC-based grade	AI-MC-based grade			Total
	1	2	3	
1	13	2	0	15
2	2	23	0	25
3	0	4	6	10
Total	15	29	6	50

for BR was high. WSI-MC appears as a viable alternative to LM-MC. Further research is required to advance our knowledge of AI-MC, but it appears at least non-inferior to LM-MC and has the potential to save time.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;71:209–249. <https://doi.org/10.3322/caac.21660>.

2. Ahmad A. Breast cancer statistics: recent trends. In: *Ahmad A, ed. Breast Cancer Metastasis and Drug Resistance: Challenges and Progress. Advances in Experimental Medicine and Biology. Springer International Publishing; 2019. p. 1–7.*

3. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991;19:403–410. <https://doi.org/10.1111/j.1365-2559.1991.tb00229.x>.

4. Genestie C, Zafrani B, Asselain B, et al. Comparison of the prognostic value of Scarff-Bloom-Richardson and Nottingham histological grades in a series of 825 cases of breast cancer: major importance of the mitotic count as a component of both grading systems. *Anticancer Res* 1998;18:571–576. <https://doi.org/10.1038/modpathol.3800161>.

5. van Doijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. *Virchows Archiv* 2021;1:1–11. <https://doi.org/10.1007/s00428-021-03141-2>.

6. Rakha EA, Reis-Filho JS, Baehner F, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* 2010;12(4). <https://doi.org/10.1186/bcr2607>.

7. van Doijeweert C, van Diest PJ, Willems SM, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands. *Int J Cancer* 2020;146:769–780. <https://doi.org/10.1002/ijc.32330>.

8. van Diest PJ, van der Wall E, Baak JPA. Prognostic value of proliferation in invasive breast cancer: a review. *J Clin Pathol* 2004;57:675. <https://doi.org/10.1136/jcp.2003.010777>.

9. Baak JPA, van Diest PJ, Voorhorst FJ, et al. Prospective multicenter validation of the independent prognostic value of the mitotic activity index in lymph node-negative breast cancer patients younger than 55 years. *J Clin Oncol* 2005;23:5993–6001. <https://doi.org/10.1200/JCO.2005.05.511>.

10. Klintman M, Strand C, Ahlin C, et al. The prognostic value of mitotic activity index (MAI), phosphohistone H3 (PPH3), cyclin B1, cyclin A, and Ki67, alone and in combinations, in node-negative premenopausal breast cancer. *PLoS One* 2013;8, e81902. <https://doi.org/10.1371/journal.pone.0081902>.

11. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern Pathol* 2005;18:1067–1078. <https://doi.org/10.1038/modpathol.3800388>.

12. Robbins P, Pinder S, de Klerk N, et al. Histological grading of breast carcinomas: a study of interobserver agreement. *Human Pathol* 1995;26:873–879. [https://doi.org/10.1016/0046-8177\(95\)90010-1](https://doi.org/10.1016/0046-8177(95)90010-1).

13. Theissig F, Kunze KD, Haroske G, et al. Histological grading of breast cancer: interobserver, reproducibility and prognostic significance. *Pathol Res Pract* 1990;186:732–736. [https://doi.org/10.1016/S0344-0338\(11\)80263-3](https://doi.org/10.1016/S0344-0338(11)80263-3).

14. van Diest PJ, Baak JPA, Matze-Cok P, et al. Reproducibility of mitosis counting in 2,469 breast cancer specimens: Results from the Multicenter Morphometric Mammary Carcinoma Project. *Human Pathol* 1992;23:603–607. [https://doi.org/10.1016/0046-8177\(92\)90313-r](https://doi.org/10.1016/0046-8177(92)90313-r).

15. Boiesen P, Bendahl PO, Anagnostaki L, et al. Histologic grading in breast cancer—reproducibility between seven pathologic departments. *South Sweden Breast Cancer Group. Acta Oncol* 2000;39(1):41–45. <https://doi.org/10.1080/028418600430950>.

16. Rakha EA, Aleskandarani M, Toss MS, et al. Breast cancer histologic grading using digital microscopy: concordance and outcome association. *J Clin Pathol* 2018;71:680–686. <https://doi.org/10.1136/jclinpath-2017-204979>.

17. Williams B, Hanby A, Millican-Slater R, et al. Digital pathology for primary diagnosis of screen-detected breast lesions - experimental data, validation and experience from four centres. *Histopathology* 2020;76:968–975. <https://doi.org/10.1111/his.14079>.

18. Al-Janabi S, van Slooten HJ, Visser M, et al. Evaluation of mitotic activity index in breast cancer using whole slide digital images. *PLoS One* 2013;8(12). <https://doi.org/10.1371/journal.pone.0082576>.

19. Balkenhol MCA, Tellez D, Vreuls W, et al. Deep learning assisted mitotic counting for breast cancer. *Lab Invest* 2019;99:1596–1606. <https://doi.org/10.1038/s41374-019-0275-0>.

20. Lashen A, Ibrahim A, Katayama A, et al. Visual assessment of mitotic figures in breast cancer: a comparative study between light microscopy and whole slide images. *Histopathology* 2021;79:913–925. <https://doi.org/10.1111/his.14543>.

21. Ginter PS, Idress R, D'Alfonso TM, et al. Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy. *Modern Pathol* 2021;34:701–709. <https://doi.org/10.1038/s41379-020-00698-2>.

22. Malon C, Brachtel E, Cosatto E, et al. Mitotic figure recognition: agreement among pathologists and computerized detector. *Anal Cell Pathol (Amsterdam)* 2012;35(2):97. <https://doi.org/10.3233/ACP-2011-0029>.

23. Veta M, van Diest PJ, Willems SM, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med Image Anal* 2015;20(1):237–248. <https://doi.org/10.1016/j.media.2014.11.010>.

24. Roux L, Racoceanu D, Loménie N, et al. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *J Pathol Inform* 2013;4(1):8. <https://doi.org/10.4103/2153-3539.112693>.

25. Nateghi R, Danyali H, Helfroush MS. A deep learning approach for mitosis detection: application in tumor proliferation prediction from whole slide images. *Artif Intel Med* 2021;114, 102048. <https://doi.org/10.1016/j.artmed.2021.102048>.

26. Li C, Wang X, Liu W, et al. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Med Image Anal* 2019;53:165–178. <https://pubmed.ncbi.nlm.nih.gov/30798116/>.

27. Bertram CA, Aubreville M, Donovan TA, et al. Computer-assisted mitotic count using a deep learning-based algorithm improves interobserver reproducibility and accuracy. *Vet Pathol* 2022;59:211–226. <https://doi.org/10.1177/03009858211067478>.

28. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin, Heidelberg: Springer; 2013, September. p. 411–418. https://doi.org/10.1007/978-3-642-40763-5_51.*

29. Lafarge MW, Bekkers EJ, Pluim JP, et al. Roto-translation equivariant convolutional networks: application to histopathology image analysis. *Med Image Anal* 2021;68, 101849. <https://doi.org/10.1016/j.media.2020.101849>.

30. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22:276–282. <https://doi.org/10.11613/BM.2012.031>.

31. Wei BR, Halsey CH, Hoover SB, et al. Agreement in histological assessment of mitotic activity between microscopy and digital whole slide images informs conversion for clinical diagnosis. *Acad Pathol* 2019;6. <https://doi.org/10.1177/2374289519859841.2374289519859841>.

32. Shaw EC, Hanby AM, Wheeler K, et al. Observer agreement comparing the use of virtual slides with glass slides in the pathology review component of the POSH breast cancer cohort study. *J Clin Pathol* 2012;65:403–408. <https://doi.org/10.1136/jclinpath-2011-200369>.
33. Pantanowitz L, Hartman D, Qi Y, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagn Pathol* 2020;15:80. <https://doi.org/10.1186/s13000-020-00995-z>.
34. van Steenhoven JEC, Kuijer A, Kornegoer R, et al. Assessment of tumour proliferation by use of the mitotic activity index, and Ki67 and phosphohistone H3 expression, in early-stage luminal breast cancer. *Histopathology* 2020;77:579–587. <https://doi.org/10.1111/his.14185>.
35. Auberville M, Stathonikos N, Bertram CA, et al. Mitosis domain generalization in histopathology images - the MIDOG challenge. arXiv:2204.03742 [eess.IV]. <https://arxiv.org/pdf/2204.03742.pdf>. Preprint.