Deep learning-based whole slide image analysis to predict sentinel node status in the INSEMA cohort

Frederik Marmé[1], Eva Krieghoff-Henning[2], Bernd Gerber[3], Max Schmitt[2], Dirk-Michael Zahm[4], Dirk Bauerschlag[5], Helmut Forstbauer[6], Guido Hildebrandt[7], Beyhan Ataseven[8], Tobias Brodkorb[1], Carsten Denkert[9], Angrit Stachs[3], David Krug[10], Jörg Heil[11], Michael Golatta[11], Thorsten Kühn[12], Valentina Nekljudova[13], Sibylle Loibl[13], Toralf Reimer[3], Titus J. Brinker[2]

[1]Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany, [2]German Cancer Research Center (DKFZ), Heidelberg, Germany, [3]Department of Obstetrics and Gynecology, University of Rostock, Germany, [4]SRH Waldklinikum Gera GmbH, Germany, [5]Universitätsfrauenklinik Aachen, Germany, [6]GOSPL – Gesellschaft für onkologische Studien, Troisdorf, Germany, [7]University Medicine Rostock, Department of Radiotherapy, Germany, [8]KEM, Evangelische Kliniken Essen Mitte, Germany, [9]Institute of Pathology, Philipps-University Marburg and University Hospital Marburg—Universitätsklinikum Marburg, Marburg, Germany, [10]Universitätsklinikum Schleswig-Holstein, Germany, [11]Uniklinikum Heidelberg, Germany, [12]Klinikum Esslingen, Germany, [13]German Breast Group, Neu-Isenburg, Germany

167P

Background

The Intergroup Sentinel Mamma (INSEMA) study as well as other ongoing studies aim at safe de-escalation of axillary surgery, which is highly desirable to reduce side effects such as lymphedema[1]. Nevertheless, it would be helpful to obtain new biomarkers that convey the same prognostic information as sentinel lymph node (SLN) status.

As known from classical histopathology, primary breast tumor tissue exhibits features such as loss of differentiation compared to the original, glandular structure and increased cell proliferation, which correlate with aggressiveness of tumor growth and might as well correlate with tumor spread into the lymph nodes.

As shown by numerous studies[2], such features can be extracted from hematoxylin and eosin (H&E)-stained breast cancer tissue sections using deep learning (DL)-based image analysis and can be used to generate digital prognostic tools.

Patients and Methods

Cohorts and patients

To train an image analysis model to predict SLN status, we used cases from the INSEMA standard arm (n=762) and a cohort from the Women's Clinic in Mannheim, Germany (n=150). For INSEMA, we used a segmentation algorithm that we had trained on part of the The Cancer Genome Atlas (TCGA) breast cancer cohort[3] to exclude slides where this algorithm did not detect enough tumor-containing tiles. The final image analysis model was tested on a holdout INSEMA set (n=381) and on the higher risk TCGA cohort[3] (n=650). Vice versa, we trained a model on TCGA whole slide images (WSIs) and tested it on the other cohorts. For the clinical classifier, we used the Ki-67 values and pT stages of the Mannheim cohort. See Table 1 for cohort characteristics (ER: estrogen receptor, PR: progesterone receptor).

Model design and training

We trained a DL image analysis model on H&E-stained WSIs of primary breast tumors. This model was based on a Resnet50 Convolutional Neural Network (CNN) architecture pre-trained with ImageNet. The entire histological images were first tessellated into smaller patches, which were processed individually. For training, the INSEMA training set was divided into folds and training was performed by 5-fold cross-validation (Figure 1). The Mannheim set was subsequently used for hyperparameter tuning. We used test time augmentation (TTA) to improve model generalization.

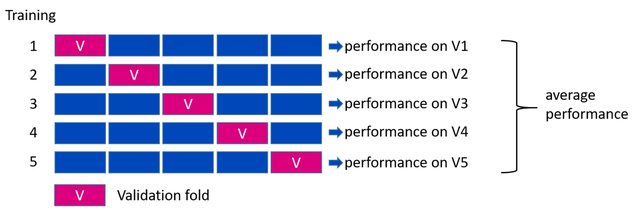

Figure 1. Cross-Validation Procedure. The INSEMA training set was divided in 5 folds and 5 models were trained using 4 folds as training data and the remaining fold as validation fold. Performances of these models were then averaged.

For inference, a probability score was assigned to each tile of a slide. This score was considered SLN positive if the CNN output was higher than 0.5 and the predictions for all tiles were averaged to obtain a slide level prediction. Training and inference were implemented in Python. To generate the clinical classifier, we fitted a logistic regression, where we could also integrate the model output as an additional variable.

Statistics

We report the mean Area under the Receiver Operating Characteristic (AUROC) curve as metric. 95% confidence intervals (95%CIs) were generated by bootstrapping (1000x). Calculations were performed in Python 3.7.7 extended with the library SciPy.

Results

Table 1 shows a comparison of relevant tumor characteristics across the different data sets used for training and testing. INSEMA and Mannheim were similar, whereas TCGA was a higher-risk cohort.

characteristic n (%)	INSEMA training (n=762)	INSEMA hold-out (n=381)	Mannheim (n=150)	TCGA (n=650)
ER/PR status				
ER/PR positive	752 (98.7)	374 (98.16)	150 (100)	467 (71.85)
ER/PR negative	10 (1.31)	7 (1.84)	0 (0)	139 (21.38)
unclear	0 (0)	0 (0)	0 (0)	44 (6.77)
HER2 status				
HER2 positive	37 (4.86)	18 (4.72)	0 (0)	108 (16.62)
HER2 negative	725 (95.14)	363 (95.28)	150 (100)	454 (69.85)
unclear	0 (0)	0 (0)	0 (0)	88 (13.54)
grading				
G1	270 (35.43)	139 (36.48)	0 (0)	n.a.
G2	461 (60.50)	233 (61.15)	150 (100)	n.a.
G3	31 (4.07)	9 (2.36)	0 (0)	n.a.
pT stage				
pT0	0 (0)	0 (0)	2 (1.33)	0 (0)
pT1	590 (77.43)	301 (79.00)	84 (56)	182 (28.00)
pT2	168 (22.05)	77 (20.21)	64 (42.67)	360 (55.38)
pT3	4 (0.52)	2 (0.52)	0 (0)	88 (12.54)
pT4	0 (0)	1 (0.26)	0 (0)	20 (3.08)
SLN positive	99 (12.99)	50 (13.12)	22 (14.67)	357 (54.92)

Table 1. Descriptive characteristics of the 4 data sets from 3 independent cohorts employed in the study.

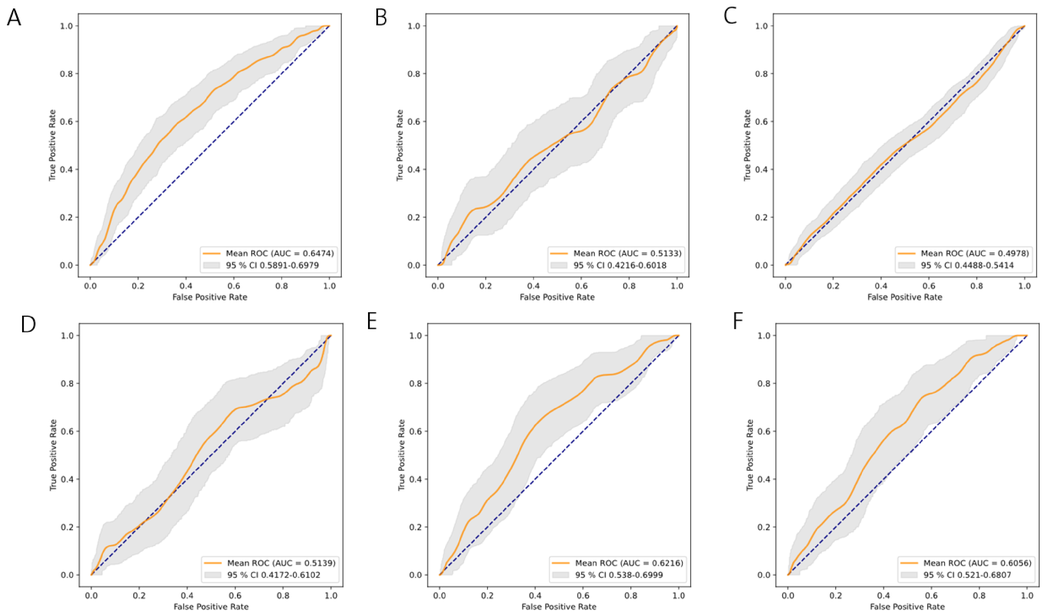

Figure 2. Internal and external performance of the generated models. A Internal cross-validation performance of the INSEMA-trained DL model. B Performance of the INSEMA-trained DL model on the INSEMA hold-out test set (blinded analysis). C Performance of the INSEMA-trained DL model on the TCGA cohort. D Performance of a TCGA-trained DL model on the INSEMA test set. E Performance of the clinical model on INSEMA. F Performance of the combined model on the INSEMA hold-out set. Mean ROC curves are shown in orange, corresponding 95% CIs in grey.

During training, the image and the clinical model yielded AUROCs of approximately 0.65 on INSEMA training cohort and of 0.62 on the Mannheim cohort, respectively (Fig. 2A and data not shown). However, performance of the image model was random on the INSEMA (determined by blinded assessment) and TCGA BRCA test sets (Fig. 2 B, C). The image model trained on TCGA also yielded random performance on the INSEMA and Mannheim cohorts (Figure 2D and data not shown). The clinical classifier, whose performance was dominated by tumor size, retained an AUROC of about 0.62 on the INSEMA set. Inclusion of the image classifier output in the logistic regression did not improve performance on INSEMA and yielded a numerically worse performance (AUROC of approx. 0.61, Fig. 2 E,F).

Conclusions

In contrast to known clinical risk factors for lymph node positivity such as pathological tumor stage and Ki-67, our image analysis algorithms trained on H&E stains of the primary tumors from INSEMA or TCGA were unable to predict sentinel status, although the technique was previously employed successfully for other tasks. This may suggest a lack of detectable systematic histological differences by lymph node status in these cohorts.

As in real life, both cohorts are dominated by ER-positive breast tumors, and the INSEMA cohort in particular is fairly homogenous also with respect to tumor grading. Still, even for INSEMA, pathological tumor size and cell proliferation were useful factors to predict SLN status. Of note, using our current pipeline, tumor size is not detected in the image analysis model, although high cell proliferation, which may in turn lead to increased tumor sizes, might be seen.

The finding that our image analysis algorithm failed to properly predict lymph node status, together with the observation that tumor size was the best predictor of SLN status, may argue that tumor spread into the lymph nodes is mostly a stochastic process driven by the total number and local spread of cancer cells in these cohorts.

One limitation of our approach may be, however, that by averaging probability scores across all tiles generated from a tumor, we don't fully take into account that tumors may be heterogeneous and may contain small areas with a high propensity for tumor cell spread. Attention-based methods could be tested to address this problem. However, considering the negative results so far, in our experience, it is unlikely that this would be sufficient to yield an accurate predictor of SLN status. Moreover, in other studies where we employed very similar approaches, we managed to predict lymph node status for prostate and colorectal cancer, demonstrating that this is feasible in principle[4,5].

Thus, DL-based WSI analysis may not be a good strategy to replace sentinel node assessment for breast cancer patients, especially in low- to intermediate-risk, hormone receptor-positive breast cancer.

References

1. Reimer T et al. (2022). Patient-reported outcomes for the Intergroup Sentinel Mamma study (INSEMA): A randomised trial with persistent impact of axillary surgery on arm and breast symptoms in patients with early breast cancer. EClinicalMedicine 55:101756. doi: 10.1016/j.eclinm.2022.101756.
2. Duggento A et al. (2021). Deep computational pathology in breast cancer. Semin Cancer Biol. 72:226-237. doi: 10.1016/j.semcancer.2020.08.006.
3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61-70. doi: 10.1038/nature11412.
4. Wessels F et al. (2021). Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer. BJU Int. 128(3):352-360. doi: 10.1111/bju.15386.
5. Kiehl L , Kuntz S et al. (2021). Deep learning can predict lymph node status directly from histology in colorectal cancer. Eur J Cancer. 157:464-473. doi: 10.1016/j.ejca.2021.08.039.

Presented at: ESMO Breast Cancer Congress 2023 (May 11-13), Berlin, Germany
GBG Forschungs GmbH received funding for research from Abbvie, AstraZeneca, BMS, Daiichi-Sankyo, Gilead, Novartis, Pfizer and Roche; other (non-financial/medical writing) from Daiichi-Sankyo, Gilead, Novartis, Pfizer, Roche and Seagen; and has the following royalties/patents: EP14153692.0, EP21152186.9, EP15702464.7, EP19808852.8 and VM Scope GmbH.

INSEMA trial is supported by German Cancer Aid (Deutsche Krebshilfe, Bonn, Germany), Grant No. 110580 and Grant No. 70110580 to University Medicine Rostock. The TPI project was funded by the German Federal Ministry of Health (2519DAT712) to Titus J. Brinker.