Research Article

Outcome and Biomarker Supervised Deep Learning for Survival Prediction in Two Multicenter Breast Cancer Series

Dmitrii Bychkov[1,2], Heikki Joensuu[2,3], Stig Nordling[4], Aleksei Tiulpin[5,6,7], Hakan Kücükel[1,2], Mikael Lundin[1], Harri Sihto[4], Jorma Isola[8], Tiina Lehtimäki[9], Pirkko-Liisa Kellokumpu-Lehtinen[10], Karl von Smitten[11], Johan Lundin[1,2,12], Nina Linder[1,2,13]

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland, [2]iCAN Digital Precision Cancer Medicine Program, [3]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, [4]Department of Pathology, Medicum, University of Helsinki, Helsinki, Finland, [5]Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland, [6]Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland, [7]Ailean Technologies Oy, Oulu, Finland, [8]Department of Cancer Biology, BioMediTech, University of Tampere, Tampere, Finland, [9]Helsinki University Hospital, Helsinki, Finland, [10]Department of Oncology, Tampere University Hospital, Tampere, Finland, [11]Eira Hospital, Helsinki, Finland, [12]Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden, [13]Department of Women's and Children's Health, International Maternal and Child Health, Uppsala University, Uppsala, Sweden

Submitted: 22-Apr-2021 Revised: 10-Jun-2021 Accepted: 20-Jun-2021 Published: 18-Jan-2022

Abstract

Background: Prediction of clinical outcomes for individual cancer patients is an important step in the disease diagnosis and subsequently guides the treatment and patient counseling. In this work, we develop and evaluate a joint outcome and biomarker supervised (estrogen receptor expression and *ERBB2* expression and gene amplification) multitask deep learning model for prediction of outcome in breast cancer patients in two nation-wide multicenter studies in Finland (the FinProg and FinHer studies). Our approach combines deep learning with expert knowledge to provide more accurate, robust, and integrated prediction of breast cancer outcomes. **Materials and Methods:** Using deep learning, we trained convolutional neural networks (CNNs) with digitized tissue microarray (TMA) samples of primary hematoxylin-eosin-stained breast cancer specimens from 693 patients in the FinProg series as input and breast cancer-specific survival as the endpoint. The trained algorithms were tested on 354 TMA patient samples in the same series. An independent set of whole-slide (WS) tumor samples from 674 patients in another multicenter study (FinHer) was used to validate and verify the generalization of the outcome prediction based on CNN models by Cox survival regression and concordance index (c-index). Visual cancer tissue characterization, i.e., number of mitoses, tubules, nuclear pleomorphism, tumor-infiltrating lymphocytes, and necrosis was performed on TMA samples in the FinProg test set by a pathologist and combined with deep learning-based outcome prediction in a multitask algorithm. **Results:** The multitask algorithm achieved a hazard ratio (HR) of 2.0 (95% confidence interval [CI] 1.30–3.00), $P < 0.001$, c-index of 0.59 on the 354 test set of FinProg patients, and an HR of 1.7 (95% CI 1.2–2.6), $P = 0.003$, c-index 0.57 on the WS tumor samples from 674 patients in the independent FinHer series. The multitask CNN remained a statistically independent predictor of survival in both test sets when adjusted for histological grade, tumor size, and axillary lymph node status in a multivariate Cox analyses. An improved accuracy (c-index 0.66) was achieved when deep learning was combined with the tissue characteristics assessed visually by a pathologist. **Conclusions:** A multitask deep learning algorithm supervised by both patient outcome and biomarker status learned features in basic tissue morphology predictive of survival in a nationwide, multicenter series of patients with breast cancer. The algorithms generalized to another independent multicenter patient series and whole-slide breast cancer samples and provide prognostic information complementary to that of a comprehensive series of established prognostic factors.

Keywords: Breast cancer, convolutional neural networks, digital pathology, ERBB2 gene, estrogen receptor, multitask deep learning, outcome prediction

Introduction

In this study, we suggest a novel approach for extraction of cancer outcome-related information[1-4] from tissue morphology by joint outcome and biomarker supervised deep learning with convolutional neural networks (CNNs). This technique is known as multitask learning[5] and has not been previously applied

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_29_21

Address for correspondence: Mr. Dmitrii Bychkov, Institute for Molecular Medicine Finland FIMM, Nordic EMBL Partnership for Molecular Medicine, P. O. Box: 20, FI-00014 University of Helsinki, Helsinki, Finland.
Biomedicum Helsinki 2U, Tukholmankatu 8, 00290 Helsinki, Finland.
E-mail: dmitrii.bychkov@helsinki.fi

How to cite this article: Bychkov D, Joensuu H, Nordling S, Tiulpin A, Kücükel H, Lundin M, *et al*. Outcome and biomarker supervised deep learning for survival prediction in two multicenter breast cancer series. J Pathol Inform 2022;13:9.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2022/13/1/9/335947

to outcome prediction in breast cancer using conventional hematoxylin-eosin (H and E) stained digitized tissue specimens. It has been demonstrated that a multitask approach can improve the accuracy of classification of breast tissue samples according to histologic type and grade of differentiation[6] as well as diagnosis of breast cancer in mammograms.[7]

In breast cancer, the expression of estrogen receptors (ER), as well as the expression and gene amplification of *ERBB2* (erb-b2 receptor tyrosine kinase 2, also known as HER2) guide the selection of treatment. Previous studies demonstrate that the ER and *ERBB2* status can be predicted directly from HE stained breast cancer tissue samples[8-10] and that the tissue morphological features predictive of the *ERBB2* status also predict patient outcome.[11] In addition, it has been shown that CNNs can be trained to predict survival in breast cancer directly from the tissue morphology, supervised by patient outcome.[3] Therefore, we hypothesized that a combination of biomarker and outcome-supervised training with a multitask approach could improve the accuracy of outcome prediction in breast cancer.

To assess if outcome and biomarker supervised multitask CNNs can learn tissue-based features that are independent of established prognostic factors, a series of tissue characteristics including histological grade (with subfactors mitotic figures, nuclear pleomorphism, tubule formation),[12] tumor necrosis, axillary lymph node status and tumor size[13] were included in multivariate prognostic models. We evaluated how these characteristics, which were determined visually by a human expert, and the prognostic information extracted by CNN-based multitask learning could complement each other in breast cancer outcome prediction. In the current study, we trained the algorithms with images from tissue microarray (TMA) samples from a nationwide patient series and then validated the results on whole slide tissue specimens from another multicenter trial. Our aim was to validate the generalization of the deep learning algorithms for outcome prediction when applied to breast cancer from an independent patient series.

Materials and Methods

Patient series

The study was based on cancer tissue samples, clinicopathological data, and follow-up data from two independent breast cancer series: The FinProg series (which consists of the original FinProg series[14] and the FinProg validation series[15]) and the FinHer clinical trial series (ISRCTN76560285).[16] The original FinProg patient series with data from 2,936 patients, is a nationwide series that includes approximately 50% of all women diagnosed with breast cancer in Finland in 1991 or 1992[17] and covers most (93%) of the patients with breast cancer diagnosed within five selected geographical regions [Supplementary Figure 1]. The FinProg validation series consists of 565 patients diagnosed mainly in the Helsinki metropolitan region who were treated at the Departments of Surgery and Oncology, Helsinki University Hospital, from 1987 to 1990.[3] The outcome and cause of death data (breast cancer-specific survival [BCSS]) were retrieved from the Finnish Cancer Registry and Statistics Finland. Corresponding clinical information and pathologic tumor characteristics, including cancer histological grade, tumor size in centimeters, and axillary lymph node status were available from the hospital records. Tumour TMAs were prepared from each patient's representative formalin-fixed paraffin-embedded breast cancer samples. Amplification of the *ERBB2* gene was quantified by chromogenic *in situ* hybridization (CISH) on TMA core sections as described previously,[14] and ER expression was determined by immunohistochemistry.[14] A total of 1047 FinProg patients with one TMA image per patient were split into a training and tuning set (n = 693) and an internal test set (n = 354) [Table 1 and Supplementary Figure 1]. The median follow-up time of patients included in the training and tuning set was 15.5 years.

The FinHer trial (ISRCTN76560285) was an open-label multicenter randomized trial that included 1010 patients in Finland in 2000–2003.[18] Eligible women were ≤65 years of age, had undergone breast surgery with axillary nodal dissection, and had either axillary lymph node-positive or high-risk node-negative cancer [Supplementary Figure 2]. Breast cancer ER and *ERBB2* expression were determined by immunohistochemistry according to institutional guidelines.[16] For patient samples considered positive for *EBBR2* expression by immunohistochemistry (either 2+ or 3+ on a scale from 0 to 3+), *ERBB2* gene amplification status was determined by CISH.[16] Breast cancers with ≥6 gene copies were considered *ERBB2*-positive. Patients were randomly assigned to receive three cycles of docetaxel or vinorelbine, followed in both groups by three cycles of fluorouracil, epirubicin, and cyclophosphamide. The 232 (23.0%) patients with *ERBB2*-positive cancer underwent a second randomization either to receive concomitant intravenous trastuzumab for 9 weeks or to not receive trastuzumab.[16] One patient with overt distant metastases at the time of random assignment was excluded from survival analyses. The primary endpoint for the FinHer trial participants was distant disease-free survival (DDFS), defined as the time from randomization to the detection of distant metastasis.[16] The median follow-up time was 5.2 years after random assignment.[16] A total of 712 HE-stained whole-slide images (WSIs), one slide per FinHer patient were used as an external test set, not used for training or tuning the algorithms [Table 1 and Supplementary Figure 2].

Ethics approval

The use of the FinProg patient series and the clinical data was approved by the operative Ethics Committee of the Hospital District of Helsinki and Uusimaa (94/13/03/02/2012), and the National Supervisory Authority for Welfare and Health (Valvira) approved the use of human tissues (7717/06.01.03.01/2015). Profiling of tumors from the FinHer patient series was approved by the institutional review board of the Helsinki University Hospital (HUS 177/13/03/02/2011).

Table 1: Biological characteristics of breast cancers and patient survival in the FinHer and FinProg series

	FinProg Patient Series (original and validation)						FinHer Patient Series					
	Training and tuning (n=693)		Internal test set (n=354)		Included patients (n=1047)		Total (n=1299)		External test set (n=712)		Total (n=1009)	
Variables:	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Histological grade	98	14.1	68	19.2	166	15.9	226	17	95	13.3	150	14.9
1												
2	244	35.2	127	35.9	371	35.4	450	35	276	38.8	397	39.3
3	168	24.2	68	19.2	236	22.5	273	21	303	42.6	414	41.0
NA	183	26.4	91	25.7	274	26.2	350	27	38	5.3	48	4.8
ERBB2 (CISH)												
Negative	557	80.4	288	81.4	845	80.7	944	73	548	77.0	776	76.9
Positive	136	19.6	66	18.6	202	19.3	216	17	164	23.0	233	23.1
NA							139	10				
ER												
Positive	472	68.1	243	68.6	715	68.3	812	63	501	70.4	729	72.2
Negative	221	31.9	111	31.4	332	31.7	364	28	211	29.6	280	27.8
NA							123	9				
Survival*												
Censored	483	69.7	254	71.8	737	70.4	979	75	593	83.3	846	83.8
Uncensored	210	30.3	100	28.2	310	29.6	205	16	119	16.7	163	16.2

* FinProg – Breast cancer-specific survival; FinHer – Distant disease-free survival (DDFS); CISH – chromogenic *in situ* hybridization; NA – not available.

Annotation of tissue images

Mitotic figures, nuclear pleomorphism, and tubule formation were assessed by a pathologist (S.N.) on 354 TMA spot images from the FinProg test set. These expert-derived features were further combined into a TMA-based histological grade according to a modification of the established breast cancer grading system[12,19] [Supplementary Table 1]. Scores 3–5, 6–7, and 8–9 formed grades I, II, and III, respectively. Tissue necrosis and tumor-infiltrating lymphocytes (TILs) were also assessed on the same set of FinProg TMA images. Further, a visual risk score (VRS) was determined by a pathologist (S.N.), such that the patients were assigned into a low-risk or a high-risk group, based on the morphology of the corresponding TMA samples.

Image preprocessing and augmentation

Images of TMA samples from the FinProg series (average size 3500 × 3500 pixels) were available in a Portable Network Graphics format extracted from WSIs scanned with a whole slide scanner (Pannoramic 250 FLASH, 3DHISTECH Ltd, Budapest, Hungary) and the FinHer samples as original whole-slide image files (MRXS) digitized with the same scanner [Supplementary Material]. Tiles of 950 × 950 pixels [209 × 209 μm with 0.22 μm pixel size, Supplementary Material] were extracted from the FinHer WSIs and saved in a JPEG format. Both the FinProg TMA images and FinHer image tiles were color-normalized[20] to adjust for HE staining variation across the tissue samples.

During training on FinProg images, we extracted square crops from a random location in the TMA spot images. One crop of size 950 × 950 pixels per TMA spot was extracted at each epoch. Thus, at every epoch, the networks were supplied with a different set of crops that originated from various locations of the TMA spots included in the training set. On the fly, data augmentation was applied to the FinProg TMA images during training. Image up/down-scale (0%–30%), rotation (±90°), shear (0%–20%), and gamma correction (0%–30%) were randomly applied to the TMA crops [Supplementary Material].

Network architecture and training

We built the deep learning model around a ResNet[21] CNN backbone. The backbone constitutes a stack of convolutional layers and outputs three-dimensional arrays, i.e., feature maps. These feature maps are globally average pooled to produce a feature vector of a fixed size. Thereby, global average pooling allows to input images of arbitrary size into the outcome prediction pipeline. Finally, the feature vector is passed through a fully connected layer to predict a corresponding continuous-value risk score, associated with the input image. The GuanRank,[22] a nonparametric ranking-based technique was used to transform time-to-event data into a linear space of hazard ranks representing BCSS for each patient. Thereby, the outcome prediction was turned into a regression task with the mean squared error loss. This transformation was applied only at the training phase. Regarding the application of the algorithm to the samples in the test and validation sets the algorithm output was a continuous-value risk score. Breast cancer outcome in the form of follow-up time and censor status were used as the ground truth.

In addition to predicting the main endpoint, i.e., BCSS, the ER and *ERBB2* status of the tumor samples was used as auxiliary endpoints in the training. Predicting multiple endpoints at the same time is referred to as multitask learning[7] and it has been shown[6] to improve learning efficacy and prediction accuracy by introducing additional regularization to the network.

Deep learning architectures were implemented using an opensource machine learning library (*PyTorch,* Facebook's AI Research lab-FAIR).[23] The networks were trained on the FinProg TMA images using a five-fold cross-validation and then evaluated on the FinHer WSIs. We used Adam[24] – an adaptive learning rate optimization algorithm to train the models. During the first three epochs, only the weights of fully connected layers were updated. Starting from the fourth epoch, the last three convolutional layers on the CNN backbone were released and trained for 100 more epochs together with the fully connected layers. Mean squared error loss was used to penalize risk score prediction and focal loss (alpha = 0.25, gamma = 2)[25] to penalize binary auxiliary endpoints, i.e., ER and *ERBB2* status in the multi-task setups. We used an initial learning rate of 1*e*–4 and dropped it by a factor of 10 at epoch 10 and 50. The L2 regularization term was added to the loss function with a weight decay parameter set to 1*e*–3. A dropout layer (P = 0.3) was introduced before the fully connected blocks. Finally, the convolutional backbones were fine-tuned starting from the ImageNet pretrained weights[26] whereas the fully connected blocks were initialized with random weights.

Inference procedure

To evaluate the generalization of the models trained on the FinProg TMA sample images we employed two independent test sets: The FinProg test-set patients that we refer to as the internal test set and the FinHer patient series that we did not use for training at all. In both sets, we averaged outputs from the five models trained in cross-validation to reduce the variance of the CNNs and boost the prediction accuracy.

Statistical analysis

The concordance between the predicted risk score (CNN output) and the actual time-to-event data (follow-up time and censor status) was estimated with the concordance index (c-index) in the patients included in the test sets. We applied Cox Proportional Hazards (PH) univariate survival regression to derive hazard ratios (HR) (effect size) associated with the risk score predicted by the CNNs and other clinicopathological variables. In addition, Cox PH multivariate regression was performed to check the independence of the variables in prediction of the risk score. The log-rank test was used to compare survival distributions between two patient subgroups.

Results

Multitask learning and outcome prediction accuracy in the FinProg series

We trained CNNs to extract prognostic information from the breast cancer TMA samples in the FinProg series [Figure 1]. We used TMA images from 693 FinProg patients to train the algorithm in a five-fold cross-validation and then applied the trained models to a test set of 354 FinProg patients. The "Solo" models that were supervised with outcome data only (i.e., the GuanRank value) achieved an HR of 1.7 (95% confidence interval [CI] 1.10–2.60) in a univariate Cox PH regression, P = 0.009 and concordance index (c-index) of 0.57 [Table 2].

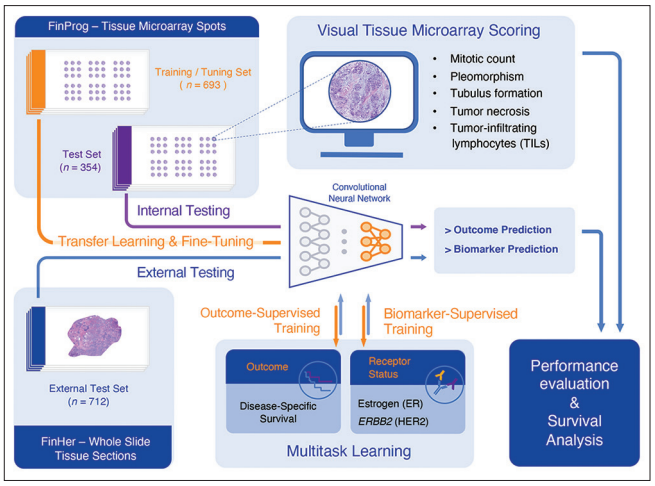


Figure 1: Deep convolutional neural networks were trained on images of hematoxylin and eosin-stained tumor tissue microarray spots from a nationwide breast cancer series (FinProg) to predict risk scores of breast cancer-specific survival. The training was performed using a transfer learning approach with ImageNet pretrained weights. The multitask approach combined outcome-supervised and biomarker-supervised feature learning. At the test phase, the networks generate a risk score for each patient in the test sets which consisted of FinProg test set patients and patients from the FinHer series. Additionally, conventional tissue entities in the tissue microarray spot images in the FinProg test set were assessed by a pathologist, i.e., mitoses, nuclear pleomorphism, tubules, tissue necrosis and tumor-infiltrating lymphocytes. Finally, a survival analysis on expert-derived and deep learning-based features was performed using Cox Proportional Hazards method.

Models trained in a multitask fashion, i.e., predicting ER and *ERBB2* status together with outcome achieved an HR of 2.0 (95% CI 1.30–3.00), $P < 0.001$, and an accuracy as measured by the c-index of 0.59 [Table 2]. Examples of high-risk and low-risk FinProg patient TMA samples are presented in Supplementary Figure 3.

Morphological characteristics of tumors assessed on tissue microarray samples predict patient survival

We examined whether the subcomponents of histological grade, i.e., mitotic figures, nuclear pleomorphism, and tubule formation predict survival of patients in the FinProg series when assessed by a pathologist viewing the TMA images. Univariate Cox PH regression showed that all three features were predictive of BCSS. Marked nuclear pleomorphism had an HR of 3.00 (95% CI 1.34–6.70), P = 0.008, c-index of 0.59; low tubulus formation had an HR of 2.20 (95% CI 1.10–4.60); high mitotic count reached an HR of 2.00 (95% CI 1.10–3.60) [Table 2]. The TMA-based grading had an HR of 3.00 (95% CI 1.50–6.10), P = 0.002, and a c-index of 0.60 on the FinProg test set [Table 2]. The original histological grading assessed on whole-slides (WS grade) by pathologists at the time of diagnosis demonstrated an HR of 4.00 (95% CI 2.00–8.30), $P < 0.001$, and a c-index of 0.64. The presence of necrotic tissue was associated with an HR of 5.00 (95% CI 2.40–10.00), $P < 0.001$, whereas a higher number of TILs was not a statistically significant predictor of survival in a univariate

Table 2: Univariate Cox proportional hazards analysis of tissue characteristics assessed on tissue microarrays within the FinProg test set

	n	HR	95% CI	*P*	c-index
Mitotic count (TMA)					
Low	256		Reference		0.57
Moderate	43	1.50	0.88-2.70	0.132	
High	31	2.00	1.10-3.60	≤ 0.05*	
Pleomorphism (TMA)					
Minimal	45		Reference		0.59
Moderate	193	1.90	0.86-4.20	0.11	
Marked	92	3.00	1.34-6.70	≤ 0.01**	
Tubulus formation (TMA)					
High	49		Reference		0.54
Low	281	2.20	1.10-4.60	≤ 0.05*	
Histological grade (TMA)*					
I	74		Reference		0.60
II	194	2.1	1.10-3.80	≤ 0.05*	
III	62	3.0	1.50-6.10	≤ 0.01**	
Histological grade (WS)					
I	64		Reference		0.64
II	119	2.70	1.30-5.30	≤ 0.01**	
III	61	4.00	2.00-8.30	≤ 0.001***	
Tumor necrosis (TMA)					
Absent	320		Reference		0.54
Present	11	5.00	2.40-10.00	<0.001***	
Tumor-infiltrating lymphocytes (TMA)					
Low	289		Reference		0.54
High	50	1.60	0.94-2.60	0.083	
Visual risk (TMA)					
Low risk	213		Reference		0.58
High risk	114	1.80	1.20-2.70	≤ 0.01**	
Axillary lymph node status					
Negative	200		Reference		0.62
Positive	128	2.40	1.60-3.60	≤ 0.001***	
Tumor size (cm)	336	1.50	1.30-1.70	≤ 0.001***	0.71
"Solo" CNN (TMA)					
Low risk	177		Reference		0.57
High risk	177	1.70	1.10-2.60	≤ 0.01**	
Multitask CNN (TMA)					
Low risk	177		Reference		0.59
High risk	177	2.00	1.30-3.00	≤ 0.001***	

*Supplementary Table 1. Association of the variables with breast cancer-specific survival is reported as effect size (HR) and a c-index. Prognostic performance of the "solo" and multitask models is compared to tissue characteristics assessed by a pathologist, as well as to the tumor size and lymph node status. HR: Hazard ratio, c-index: Concordance index, CI: Confidence interval, TMA: Tissue microarrays, WS: Whole-slides, CNN: Convolutional neural networks

Cox PH regression [Table 2]. The VRS reached an HR of 1.80 (95% CI 1.20–2.70), $P = 0.004$, and a c-index of 0.58.

Deep learning combined with expert visual assessment of tissue samples

To evaluate how the deep learning-based outcome prediction can complement visual tissue assessment, we first combined "solo" and multitask CNN models with visual TMA-based histological grading. The multivariate (TMA grade + CNN) Cox PH regression showed that the multitask CNN was an independent predictor of BCSS when adjusted for the visual TMA-based histological grade with an HR of 1.7 (95% CI 1.10–2.70), a $P = 0.017$, and a c-index of 0.63. A similar c-index (0.63) was observed when the multitask CNN was combined with the VRS. Importantly, the "solo" CNN was not a statistically significant predictor of BCSS when adjusted for the TMA-based histological grade and for the VRS.

We then expanded the analysis by including TMA histological grade, necrosis, and TILs in the multivariate Cox PH regression together with the CNN predictor. Again, we observed that the multitask CNN remained an independent and statistically significant predictor of BCSS with an HR of 1.70 (95% CI 1.06–2.70), $P = 0.029$, and a c-index of 0.66.

Conventional histological grading of the WS tissue samples was available for the FinProg patient's tumors and we evaluated the prognostic value of the outcome supervised CNN when combined with WS histological grade. The multitask CNN remained independent of WS histological grade, whereas the "solo" model was not a significant predictor of BCSS. The compound model (multitask CNN + WS histological grade) had a c-index of 0.66, the same that was achieved with the TMA level features (histological grade, necrosis, and TILs) only. Tumor size and axillary lymph node status were also included in the multivariate Cox PH regression together with the multitask CNN model, which reached an HR of 1.70 (95% CI 1.10–2.50), $P = 0.022$, and a c-index of 0.73 after adjustment for size and lymph node status [Figure 2].

Generalization to independent series whole slide samples

To evaluate generalization of the proposed approach, the CNNs trained on the TMA samples from the FinProg patient series were applied to WSIs from the independent FinHer patient series. Univariate Cox PH regression showed that both multitask, and "solo" CNN models were statistically significant predictors of DDFS in patients from the FinHer series ($n = 674$). The "solo" model reached an HR of 1.8 (95% CI 1.3–2.7), a $P = 0.002$ and a c-index of 0.57. The multitask model achieved an HR of 1.7 (95% CI 1.2–2.6), $P = 0.003$ and a c-index 0.57. We then evaluated both of the models in a multivariate Cox PH regression adjusted for the WS histological grade and observed that both of the models were statistically significant predictors of survival, independent of histological grade on WSs [Table 3]. The "solo" model reached an HR of 1.7 (95% CI 1.1–2.5), a $P = 0.009$ and a c-index of 0.60 in a multivariate Cox PH analysis, whereas the multitask CNN reached an HR of 1.5 (95% CI 1.0–2.3), a $P = 0.033$ and a c-index of 0.59 [Table 3].

Conclusions

Our study demonstrates the feasibility of breast cancer outcome prediction using a multitask deep learning approach across two multicenter patient series. We show that the algorithms trained on one patient series (FinProg) can generalize to an independent patient series (FinHer). Although several studies have shown that outcome supervised deep learning can extract significant prognostic information from tumor morphology in breast cancer, they are constrained to the analysis of single-center series.[3,4] To our knowledge, this work is the first to explore generalization of the method when applied to whole-slide breast cancer tissue images from an independent multicenter patient series.

With images of H and E-stained tumor tissue samples as the input, we applied both outcome and biomarker supervised learning to extract predictive information encoded in the tumor morphology. Our best multitask algorithm achieved an HR of 2.0 and a c-index of 0.59 in predicting BCSS in the FinProg test set patients. Moreover, we demonstrated that the multitask approach allows extraction of image features that remain independent of the pathologist-derived features such as mitoses, nuclear pleomorphism, tubules, and necrosis. In contrast to the "solo" training, the multitask deep learning-based risk score was

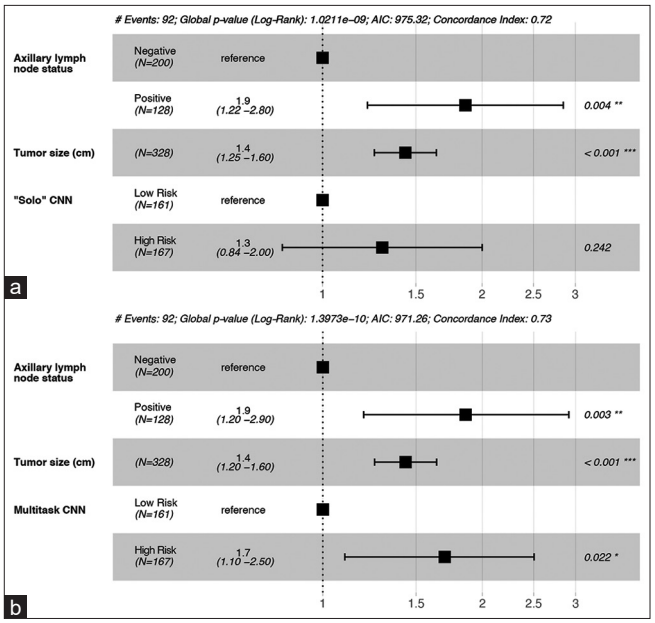


Figure 2: Multivariate Cox Proportional Hazards analysis of deep learning models together with prognostic factors related to the extent of disease in breast cancer, i.e., spread of the cancer to axillary lymph nodes and size of the primary tumor in the FinProg test set. The results indicate that multitask training (b) was an independent predictor of survival as compared to outcome supervised training only (a)

Table 3: Multivariate Cox proportional hazards regression of deep learning-based outcome predictions adjusted for tumor histological grade on the independent FinHer ($n = 674$) patient series

		"Solo" CNN			Multitask CNN		
	n	HR	95% CI	*P*	HR	95% CI	*P*
CNN risk score							
Low risk	337		Reference			Reference	
High risk	337	1.70	1.10-2.50	0.009	1.50	1.00-2.30	0.033
Histological grade (WS)							
Low (I and II)	371		Reference			Reference	
High (III)	303	1.60	1.10-2.30	0.022	1.50	1.00-2.20	0.037
c-index, Log-rank *P*			0.60, <0.001			0.59, 0.001	

WS: Whole-slides, CNN: Convolutional neural networks, HR: Hazard ratio, c-index: Concordance index, CI: Confidence interval

a significant predictor of breast cancer-specific survival after adjustment for tumor size and axillary lymph nodes status in the FinProg series. Interestingly, we observed that the information extracted through visual assessment of TMA images by a pathologist and by the CNNs together could ultimately increase prognostic accuracy to a c-index of 0.66. We recognize that the multitask approach did not demonstrate an increased accuracy in the external FinHer WS samples, as compared to the "solo" model. Potential reasons could be different endpoints used in the FinProg and the FinHer, a relatively short follow-up time, and significant heterogeneity introduced through analysis of whole slide tissue samples within the FinHer series. In the current study, we did not explore the tile size sampling effect on the performance of the models, since the training set comprised images of TMAs. Training with tiles smaller than 950 × 950 pixels (209 × 209 μm) would limit the contextual information further and likely lead to a reduced performance. On the other hand, larger tiles would lead to increased morphological heterogeneity and inclusion of non-tumor tissue areas if extended to WSIs, and likely require even larger datasets than the sample series available in the current study if training is done with sample-level labels i.e., weakly supervised learning. Systematic evaluation of sampling strategies has to be studied separately. Taken together, a deep learning model trained on TMA samples stained for basic morphology (HE) and supervised by outcome and biomarker status based complemented visual tissue assessment of established tissue entities by a pathologist in the prediction of patient outcome.

In one of the first studies[27] to address breast cancer outcome prediction with machine learning applied to basic tissue morphology, the authors used regularized logistic regression and image features from breast cancer epithelium and stroma. This approach reached HRs of 1.54–1.78 in two patient populations as estimated by a multivariate Cox PH regression in prediction of overall survival. These effect sizes are roughly at the same level as the HRs of 1.5–2.0 that were measured in prediction of BCSS and DDFS in the current patient series. Another study[3] used a deep learning approach to predict BCSS in one of the series also used in the current study (the FinProg series). The machine learning-based predictor reached an HR of 2.04 in a test set of 431 patients, but the approach was not validated on independent data. In a study that addressed morphology-based cancer survival prediction in multiple cancer types,[4] the authors trained a deep learning model on 488 WS breast cancer samples from the Cancer Genome Atlas[28] (TCGA) project. An HR of 2.86 was achieved in a multivariate Cox PH analysis on 250 heldout patients from the same TCGA patient cohort without cross-validation. Results so far suggest that significant prognostic information can be extracted from basic tissue morphology by the use of machine learning, but that effect sizes do not yet exceed those for some of the established prognostic tissue features currently assessed visually by experts. It remains to be established if the prognostic accuracy can be further improved by training and validating algorithms based on WSIs that better represent tissue heterogeneity as compared to TMAs.

Limitations related to our study include that BCSS was used in training of the algorithm on the FinProg data whereas DDFS was used as an endpoint for evaluation on the FinHer series. Although a strong correlation has been shown between disease-free and overall survival in studies on early breast cancer,[29] the strength of correlation between BCSS and DDFS remains to be established. Additionally, the tissue samples used in our study were centrally scanned using the same instrument. Thus, possible image variations due to the scanning hardware were eliminated but the generalizability of the method to samples digitized with other similar whole slide scanners have to be addressed in future studies.

In future research, the prognostic accuracy and generalization of the deep learning models can be further improved by exposing deep learning algorithms to datasets that cover an even larger spectrum of variations of tissue morphologies, including training on WSIs. Quantification of conventional prognostic features using machine learning algorithms instead of visual assessment as in the current study could further improve accuracy, consistency, and reproducibility of outcome prediction. Previous studies have demonstrated a good performance of machine learning algorithms in counting mitosis,[30] quantifying tumor-infiltrating immune cell,[31-33] assessing the grade of tumor differentiation,[34] and tissue necrosis.[35,36] A combination of computationally quantified conventional prognostic features with features learned through end-to-end oucome supervised learning should be addressed in future studies.

Our findings indicate that outcome and biomarker supervised deep learning models for breast cancer outcome prediction generalize to patient samples from an independent multicenter series. Integrative techniques such as multitask deep learning can extract image features that remain statistically independent of established prognostic factors in breast cancer. Hence, established prognostic features and features learned through machine learning approaches can complement each other and lead to more accurate and interpretable tumor tissue analysis for patient cancer outcome prediction.

Acknowledgments

We would like to thank the Digital Microscopy and Molecular Pathology unit at Institute for Molecular Medicine Finland FIMM, University of Helsinki, supported by the Helsinki Institute of Life Science and Biocenter Finland for providing slide scanning services.

Financial support and sponsorship

The study was supported by the Sigrid Jusélius Foundation, the Biomedicum Helsinki Foundation, the Orion-Pharmos Research Foundation, Finska Läkaresällskapet, Medicinska Understödsföreningen Liv och Hälsa, Stiftelsen Dorothea Olivia, Karl Walter och Jarl Walter Perkléns minne, K. Albin Johanssons Stiftelse, iCAN Digital Precision Cancer Medicine Flagship, and HiLIFE Helsinki Institute of Life Sciences.

Conflicts of interest

Johan Lundin and Mikael Lundin are the founders and co-owners of Aiforia Technologies Oy, Helsinki, Finland.

Heikki Joensuu is employed by Orion Pharma, serves as the Chairman of the Advisory Board of Neutron Therapeutics, has received funds from Neutron Therapeutics and owns stocks of Orion Pharma and Sartar Therapeutics. Aleksei Tiulpin is a co-founder, shareholder, and CTO of Ailean Technologies Oy. The other authors have no conflicts of interest.

Availability of data and materials

The data that support the findings of this study were used under a license for the current study, and some restrictions apply to their availability. The data are available from the authors upon reasonable request and with permission from the University of Helsinki.

References

1. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018;115:E2970-9.
2. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 2018;8:3395.
3. Turkki R, Byckhov D, Lundin M, Isola J, Nordling S, Kovanen PE, *et al.* Breast cancer outcome prediction with tumour tissue images and machine learning. Breast Cancer Res Treat 2019;177:41-52.
4. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, *et al.* Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS One 2020;15:e0233678.
5. Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. Adv Neural Inf Process Syst 2007;19:41-8.
6. Xipeng P, Li L, Yang H, Liu Z, He Y, Li Z, *et al.* Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. Multimed Tools Appl 2020;79:14509-8.
7. Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol 2017;62:8894-908.
8. Shamai G, Binenbaum Y, Slossberg R, Duek I, Gil Z, Kimmel R. Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. JAMA Netw Open 2019;2:e197700.
9. Rawat RR, Ortega I, Roy P, Sha F, Shibata D, Ruderman D, *et al*. Deep learned tissue 'fingerprints' classify breast cancers by ER/PR/Her2 status from H and E images. Springer Science and Business Media (LLC). Sci Rep 2020;10. [doi = 10.1038/s41598-020-64156-4].
10. Naik N, Madani A, Esteva A, Keskar NS, Press MF, Ruderman D, *et al.* Deep learning-enabled breast cancer hormonal receptor status determination from base-level H and E stains. Nat Commun 2020;11:5727.
11. Bychkov D, Linder N, Tiulpin A, Kücükel H, Lundin M, Nordling S, *et al.* Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. Sci Rep 2021;11:4037.
12. 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;19:403-10.
13. WHO Classificationof Tumours Editorial Board. WHO Classification of Breast Tumours: WHO Classification of Tumours. 2nd ed., Vol. 2.: World Health Organization; 2019.
14. Joensuu H, Isola J, Lundin M, Salminen T, Holli K, Kataja V, *et al.* Amplification of erbB2 and erbB2 expression are superior to estrogen receptor status as risk factors for distant recurrence in pT1N0M0 breast cancer: A nationwide population-based study. Clin Cancer Res 2003;9:923-30.
15. Lundin J, Lundin M, Isola J, Joensuu H. A web-based system for individualised survival estimation in breast cancer. BMJ 2003;326:29.
16. Joensuu H, Kellokumpu-Lehtinen PL, Bono P, Alanko T, Kataja V, Asola R, *et al.* Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. N Engl J Med 2006;354:809-20.
17. Joensuu H, Tiina L, Kaija H, Liisa E, Taina TH, Vesa K, *et al.* Risk for distant recurrence of breast cancer detected by mammography screening or other methods. JAMA 2004;292:1064-73.
18. Joensuu H, Petri B, Vesa K, Tuomo A, Riitta K, Raija A, *et al*. Fluorouracil, epirubicin, and cyclophosphamide with either docetaxel or vinorelbine, with or without trastuzumab, as adjuvant treatments of breast cancer: Final results of the FinHer trial. J Clin Oncol 2009;27:5685-92.
19. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 1957;11:359-77.
20. Macenko M, Niethammer M, Marron J, Borland D, Woosley J, Guan X, *et al.* "A Method for Normalizing Histology Slides for Quantitative Analysis," In Proceedings of the Sixth IEEE International Conference on Symposium on Biomedical Imaging: From Nano to Macro; 2009. p. 1107-10.
21. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv e-prints, p. arXiv: 1512.03385; 2015. Available from: http://arxiv.org/abs/1512.03385. [Last access date 2021 Jun 01].
22. Huang Z, Zhang H, Boss J, Goutman SA, Mukherjee B, Dinov ID, *et al.* Complete hazard ranking to analyze right-censored data: An ALS survival study. PLoS Comput Biol 2017;13:e1005887.
23. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, *et al.* PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024-35.
24. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv e-prints, p. arXiv: 1412.6980; 2017.
25. Lin TY, Goyal P, Girshick RB, He K, Dollár P. Focal loss for dense object detection. 2017 IEEE Int Conf Comput Vis 2017;abs/1708.02002:2999-3007.
26. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, *et al.* ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.
27. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3:108ra113.
28. Chang K, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, *et al.* The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45:1113-20.
29. Saad ED, Squifflet P, Burzykowski T, Quinaux E, Delaloge S, Mavroudis D, *et al.* Disease-free survival as a surrogate for overall survival in patients with HER2-positive, early breast cancer in trials of adjuvant trastuzumab for up to 1 year: A systematic review and meta-analysis. Lancet Oncol 2019;20:361-70.
30. Dif N, Elberrichi Z. Deep learning methods for mitosis detection in breast cancer histopathological images: A comprehensive review BT. In: Holzinger A, Goebel R, Mengel M, Müller H, editors. Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges. Cham: Springer International Publishing; 2020. p. 279-306.
31. Stenman S, Bychkov D, Kucukel H, Linder N, Haglund C, Arola J, *et al.* Antibody supervised training of a deep learning based algorithm for leukocyte segmentation in papillary thyroid carcinoma. IEEE J Biomed Health Inform 2021;25:422-8.
32. Turkki R, Linder N, Kovanen P, Pellinen T, Lundin J. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. J Pathol Inform 2016;7:38.
33. Linder N, Taylor JC, Colling R, Pell R, Alveyn E, Joseph J, *et al.* Deep learning for detecting tumour-infiltrating lymphocytes in testicular germ cell tumours. J Clin Pathol 2019;72:157-64.
34. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, *et al.* Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 2018;4:30.
35. Arunachalam HB, Mishra R, Daescu O, Cederberg K, Rakheja D, Sengupta A, *et al.* Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. PLoS One 2019;14:e0210706.
36. Turkki R, Linder N, Holopainen T, Wang Y, Grote A, Lundin M, *et al.* Assessment of tumour viability in human lung cancer xenografts with texture-based image analysis. J Clin Pathol 2015;68:614-21.

Supplementary Material

Image Data Preparation and Preprocessing

A batch of 16 random crops constituted one training iteration, which corresponded to input tensors of size (950, 950, 3, 16) (height, width, color channels, batch size). Original size of each TMA spot was 3500 × 3500 pixels on average. All input tensors were normalized with mean and standard deviation, as estimated on the training data: mean – (0.8198558, 0.78990823, 0.91205645), std – (0.1421396, 0.15343277, 0.07634846) for RGB channels accordingly. After image normalization, we performed on-the-fly training image augmentations using SOLT data augmentation library (https://github.com/MIPT-Oulu/solt) with the following parameters:

- Random scaling with 0.5 probability and 0.3 scale range
- Random rotation with 0.5 probability and ± 90-degree range
- Random shear with 0.5 probability and 0.2 shear range
- Random gamma correction with 0.5 probability and 0.3 gamma range.

For internal testing on 354 FinProg TMA spots we used a center crop of size 2100 × 2100 pixels and applied no image augmentation.

For evaluations on the external test set – the FinHer series, we extracted non-overlapping (step size 950 pixels) tiles of 950 × 950 pixels from each of the whole slide tissue images. We then applied HistoQC quality control tool[4] to eliminated tiles that contain artefacts such as out of focus regions, tissue folding etc. Each tile that passed the quality check was processed by five models trained in cross-validation on the FinProg training set. The predictions were averaged to obtain a single tile-level Risk Scores. No test-time image augmentation was performed on the FinHer tiles.

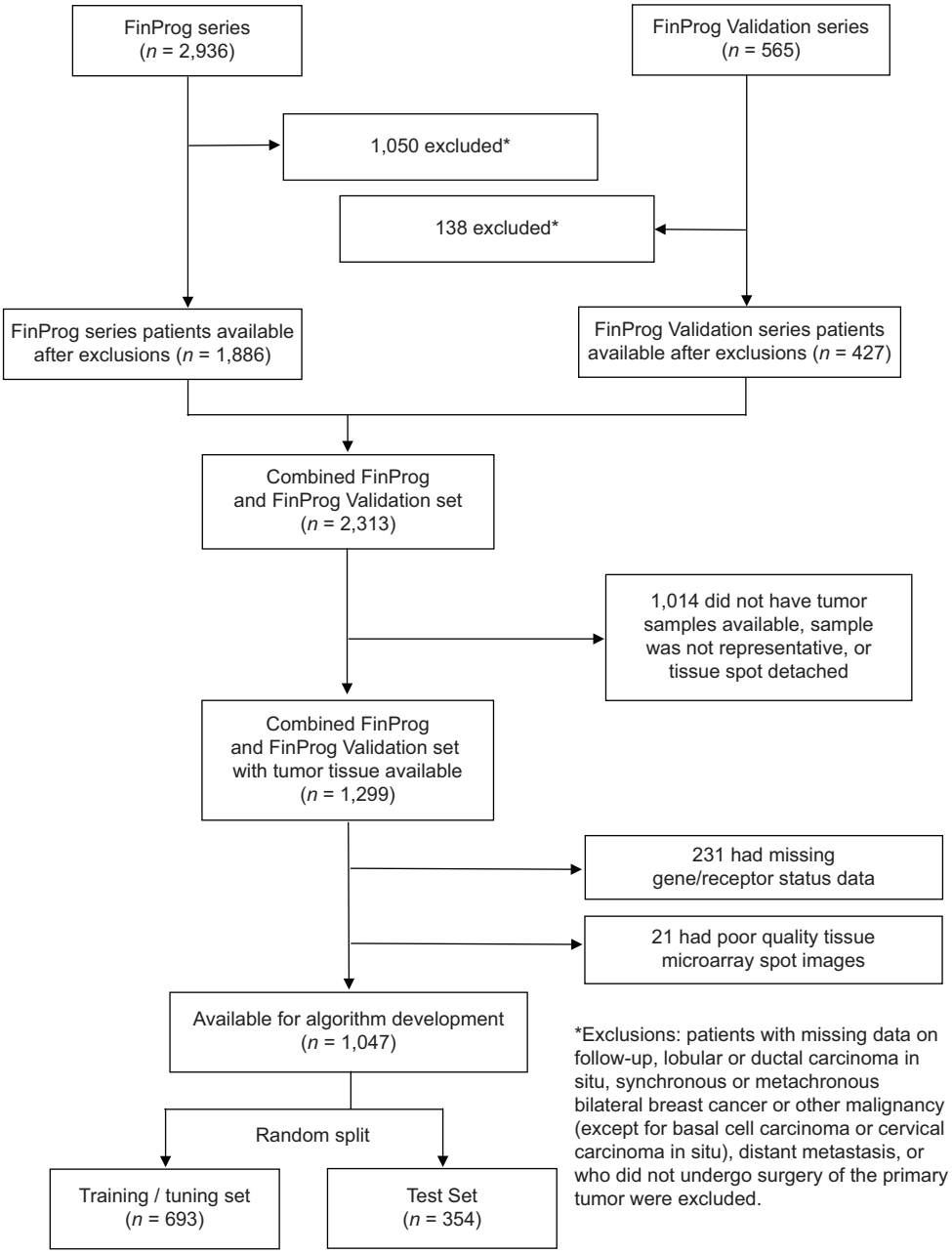
Image Acquisition

Five-micrometer thick sections were cut from the TMA blocks, stained with hematoxylin and eosin, and digitized with a whole-slide scanner (Pannoramic 250 FLASH, 3DHISTECH Ltd., Budapest, Hungary) equipped with a 20 × objective (numerical aperture 0.80) and a 1 × adapter, and a progressive scan color camera with three separate charge-coupled devices with 1 618 × 1 236 pixels sized 4.40 µm × 4.40 µm (CIS_VCC_F52U25CL, CIS Corporation, Tokyo, Japan). This resulted in an image where one pixel represents an area of 0.22 µm × 0.22 µm. The images were stored in a whole-slide image format (MRX, 3DHISTECH Ltd., Budapest, Hungary), and were further compressed to a wavelet file format (Enhanced Compressed Wavelet, ECW, ER Mapper, Intergraph, Atlanta, GA) with a compression ratio of 1:10. The compressed virtual slides were uploaded to a whole-slide image management server (WebMicroscope, Aiforia Technologies Oy, Helsinki, Finland), where the individual images of the TMA spots were segmented from the whole-slide TMA image.

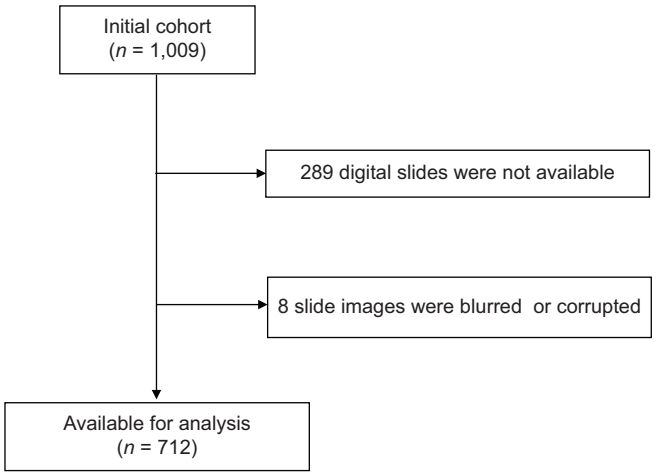
Supplementary Table 1: Tissue microarray histological scoring

Feature	Category	Score
Mitoses	0 per HPF	1
	1 per HPF	2
	>1 per HPF	3
Nuclear pleomorphism	Minimal	1
	Moderate	2
	Marked	3
Tubules	>75%	1
	10%-75%	2
	<10%	3

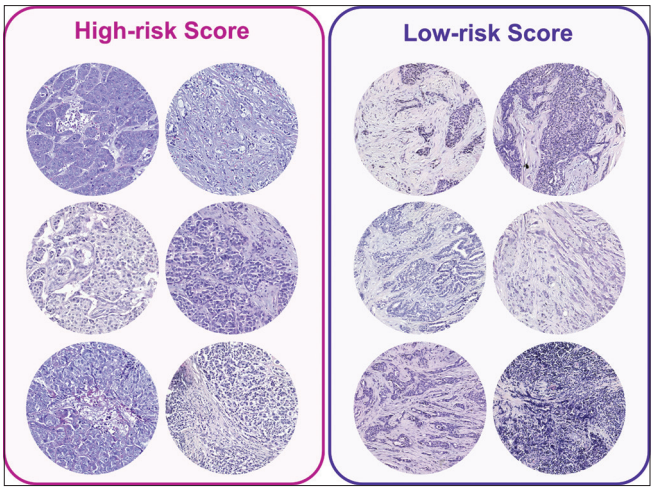
*HPF: High-power field



Supplementary Figure 1: FinProg CONSORT Diagram



Supplementary Figure 2: FinHer CONSORT Diagram



Supplementary Figure 3: Examples of high-risk and low-risk patient tissue microarray spots as predicted by the multitask model