Breast Cancer Research

RESEARCH

Open Access

Development and prognostic validation of a three-level NHG-like deep learning-based model for histological grading of breast cancer

Abhinav Sharma[1], Philippe Weitz[1], Yinxi Wang[1], Bojing Liu[1,2†], Johan Vallon-Christersson[6], Johan Hartman[3,4,5†] and Mattias Rantalainen[1,5*†]

Abstract

Background Histological grade is a well-known prognostic factor that is routinely assessed in breast tumours. However, manual assessment of Nottingham Histological Grade (NHG) has high inter-assessor and inter-laboratory variability, causing uncertainty in grade assignments. To address this challenge, we developed and validated a three-level NHG-like deep learning-based histological grade model (predGrade). The primary performance evaluation focuses on prognostic performance.

Methods This observational study is based on two patient cohorts (SöS-BC-4, $N=2421$ (training and internal test); SCAN-B-Lund, $N=1262$ (test)) that include routine histological whole-slide images (WSIs) together with patient outcomes. A deep convolutional neural network (CNN) model with an attention mechanism was optimised for the classification of the three-level histological grading (NHG) from haematoxylin and eosin-stained WSIs. The prognostic performance was evaluated by time-to-event analysis of recurrence-free survival and compared to clinical NHG grade assignments in the internal test set as well as in the fully independent external test cohort.

Results We observed effect sizes (hazard ratio) for grade 3 versus 1, for the conventional NHG method (HR = 2.60 (1.18–5.70 95%CI, p-value = 0.017)) and the deep learning model (HR = 2.27, 95%CI 1.07–4.82, p-value = 0.033) on the internal test set after adjusting for established clinicopathological risk factors. In the external test set, the unadjusted HR for clinical NHG 2 versus 1 was estimated to be 2.59 (p-value = 0.004) and clinical NHG 3 versus 1 was estimated to be 3.58 (p-value < 0.001). For predGrade, the unadjusted HR for predGrade 2 versus 1 HR = 2.52 (p-value = 0.030), and 4.07 (p-value = 0.001) for preGrade 3 versus 1 was observed in the independent external test set. In multivariable analysis, HR estimates for neither clinical NHG nor predGrade were found to be significant (p-value > 0.05). We tested for differences in HR estimates between NHG and predGrade in the independent test set and found no significant difference between the two classification models (p-value > 0.05), confirming similar prognostic performance between conventional NHG and predGrade.

Conclusion Routine histopathology assessment of NHG has a high degree of inter-assessor variability, motivating the development of model-based decision support to improve reproducibility in histological grading. We found

†Bojing Liu, Johan Hartman and Mattias Rantalainen equally contributed to the work.

*Correspondence:
Mattias Rantalainen
mattias.rantalainen@ki.se
Full list of author information is available at the end of the article

that the proposed model (predGrade) provides a similar prognostic performance as clinical NHG. The results indicate that deep CNN-based models can be applied for breast cancer histological grading.

Keywords Breast cancer, Pathology, Deep learning, Image analysis, Clinical decision support

Background

Histological grading is a well-established prognostic factor for breast cancer and is associated with the aggressiveness of the tumour [1]. An assessment of three morphological features determines the histological grade of breast tumours. These features include tubular formation (glandular differentiation), nuclear pleomorphism, and mitotic counts, and each component is given a score from I to III. The sum of the sub-component scores enables the assignment of tumours into three grades (Grade 1–3), referred to as Nottingham Histological Grade (NHG), where Grade 1 is associated with a good prognosis and Grade 3 is associated with a poor prognosis [2]. It provides prognostic information for clinically relevant subgroups (like estrogen receptor (ER)-positive and human epidermal growth factor receptor 2 (HER2)-negative patients) to determine the plan for adjuvant chemotherapy [3].

However, the assessment of histological grading has a high inter-observer variability including the assessments of individual subcomponents of histological grading [4–6]. A recent nationwide study in Sweden reported significant inter-laboratory variabilities for histological grading across different pathology laboratories [7]. Such variabilities indicate an intrinsic uncertainty in routine NHG assessment and potential errors, which can cause both under and over-treatment of breast cancer.

Recent advances in high-resolution digital whole-slide images (WSIs) have greatly enhanced the computer-based pathology workflow, paving the way to novel digital decision support solutions. Recently, deep learning-based analyses on WSIs have shown promising results in a multitude of tasks, including cancer classification, grading, and predictions of genetic mutations in prostate and lung cancers [8–10].

Deep learning, especially deep convolutional neural networks (CNNs), has been proven to be effective for modelling of WSI data, including in the application of breast cancer histological grading. Previously models for the classification of grades 1 and 2 (together) versus grade 3 have been implemented for breast cancer [11, 12]. Jaroensri et al. implemented a model that classified the sub-components, and the sub-component score, for breast cancer histological grading and the prognostic performance, was compared against routine classification [13]. Wang et al. developed a model based on histological grade morphology in breast cancer that was applied to improve risk stratification of intermediate-risk patients (histological grade 2) [14].

To our knowledge, this is the first study focussing on the development of a deep-learning-based breast cancer histological grade classification with a three-level grading system resembling the routine NHG in breast cancer with prognostic evaluation. We evaluate the proposed model (predGrade) from the perspective of prognostic performance (time-to-event) in both internal test data and a fully independent external test cohort and compare it with the routine clinical grade assignment.

Methods

Study materials

The patients in this study were from two Swedish cohorts, SöS BC-4 ($n=2421$), and the SCAN-B cohort ($n=1262$). SöS BC-4 is a retrospective observational study that included patients diagnosed at Södersjukhuset (South General Hospital) in Stockholm between 2012 and 2018 that had archived histological slides available and also available histological grade information. Patients that had received neoadjuvant therapy were excluded ($n=56$). The SCAN-B cohort, which we used as an independent external test set, includes a subset of patients ($n=1262$) enrolled in the prospective SCAN-B study [15], diagnosed between 2010 and 2019 in Lund, Sweden. Both cohorts consist of patients diagnosed with invasive breast cancer. Patients' clinical information (i.e. clinical NHG, ER status, epidermal growth factor receptor 2 (HER2) status, tumour size, and lymph node status) was retrieved from the Swedish National Registry for Breast Cancer (NKBC) (Additional file 1: Table S1). For CONSORT diagram, see Fig. 1.

WSIs were generated (40X magnification) using Hamamatsu NanoZoomer histopathology slide scanners (S360 or XR) from clinical routine Haematoxylin & Eosin (H&E)-stained, formalin-fixed paraffin-embedded (FFPE) resected tumour slides. We included one H&E WSI per patient, which was either the established primary diagnostic fraction or otherwise the H&E WSI with the largest predicted tumour area.

Image pre-processing and deep learning modelling methods

WSIs were pre-processed and quality controlled in a standardised processing pipeline, followed by model

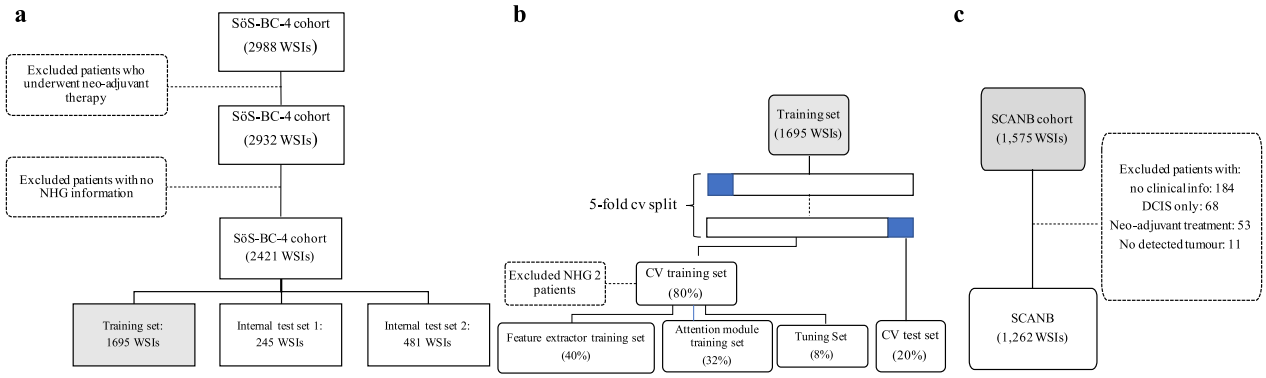

Fig. 1 SöS-BC-4 and SCANB cohort descriptions and splitting criteria. **a.** The SöS cohort was first split into the training, internal test set 1, and internal test set 2 on the patient level. The split was stratified by clinical histological grading (NHG), ER status, HER2 status, and Ki-67 status. **b.** A five-fold cross validation (CV) split was further generated on the patient level within the training set ($n=1695$ WSIs). Each CV fold consisted of a CV training set (80%) and a CV test set (20%) balanced on clinical NHG. The CV training set is further sub-split into the feature extractor training set (40%), the attention module (32%), and the tuning set (8%). **c.** SCANB cohort was used as the independent external test set

optimisation, and performance validation of the system (Fig. 2).

WSI preprocessing

The WSI pre-processing pipeline has been previously described in detail in [14]. A brief overview of the pre-processing steps is shown in Fig. 2a. First, we generated tissue masks excluding most of the backgrounds from the WSIs. We added a maximum value of 25 on the Otsu threshold in order to reduce the removal of the tissue regions in some cases due to the high threshold value on the transformed saturation channel. The tissue regions were divided into image patches (i.e. tiles) of size 1196×1196 pixels. The image tiles were down-sampled by a factor of two from the original scanning resolution (40X) to 20X resolution (598×598 pixels; 271×271 µm). Next, we applied the Laplacian filter (OpenCV package version 3.4.2) on all the image tiles and computed the variance of the filtered tiles. Tiles with a variance lower than 500 units were considered blurry and excluded from further analyses [16]. To mitigate stain colour variability, the colour normalization method described by Macenko et al. [17] was applied, with a modification to enable WSI-level colour correction, as previously described in [14]. Lastly, we applied a pre-trained CNN model developed in [14] to detect invasive cancer in our current study population. Only tiles predicted as invasive cancer from the pre-trained model were considered as regions of interest and thereafter included in further analyses. The median number of invasive cancer tiles for the SöS-BC4 training set and SCANB cohort was 911 and 3069 per WSI (Additional file 1: Fig. S1).

Image analysis using deep learning

The SöS cohort was used for model development and internal validation. The cohort was split into the training set ($n=1695$), internal test set 1 ($n=245$ WSIs), and internal test set 2 ($n=481$ WSIs) as shown in Fig. 1a). The training and internal test sets were split on the patient level and stratified by histological grading (NHG), estrogen receptor (ER) status, epidermal growth factor receptor 2 (HER2), and Ki-67 status.

The training and optimisation of the feature extractor and attention module were performed on the training set ($n=1695$) using five-fold cross validation (CV). For each CV fold, the training set was split into a CV training set (80%) and a CV test set (20%) stratified by histological grading (NHG) as shown in Fig. 1b. The CV training set was further sub-split into the feature extractor training set (50%), the attention module training set (40%), and the tuning set (10%). Both the feature extractor (Resnet-18 CNN) model and the attention module were optimised against binary class labels (NHG 1 and 3) to ensure that the model learns high- and low-grade patterns despite substantial label noise (reflected by high inter-assessor variability in NHG grade label assignments). The proposed approach implicitly assumes that the NHG grading follows a continuum of morphological changes (1–3), and that NHG2 is the intermediate group with the highest assessment uncertainty and inter-rater variabilities. We, therefore, excluded NHG 2 from the model optimisation. These three sub-splits were stratified on the clinical NHG (Fig. 1b).

The attention-based Multiple Instance Learning (MIL) model was considered as the CNN modelling architecture inspired by Lu et al. [18]. It consisted of two separate trainable modules: the feature extractor

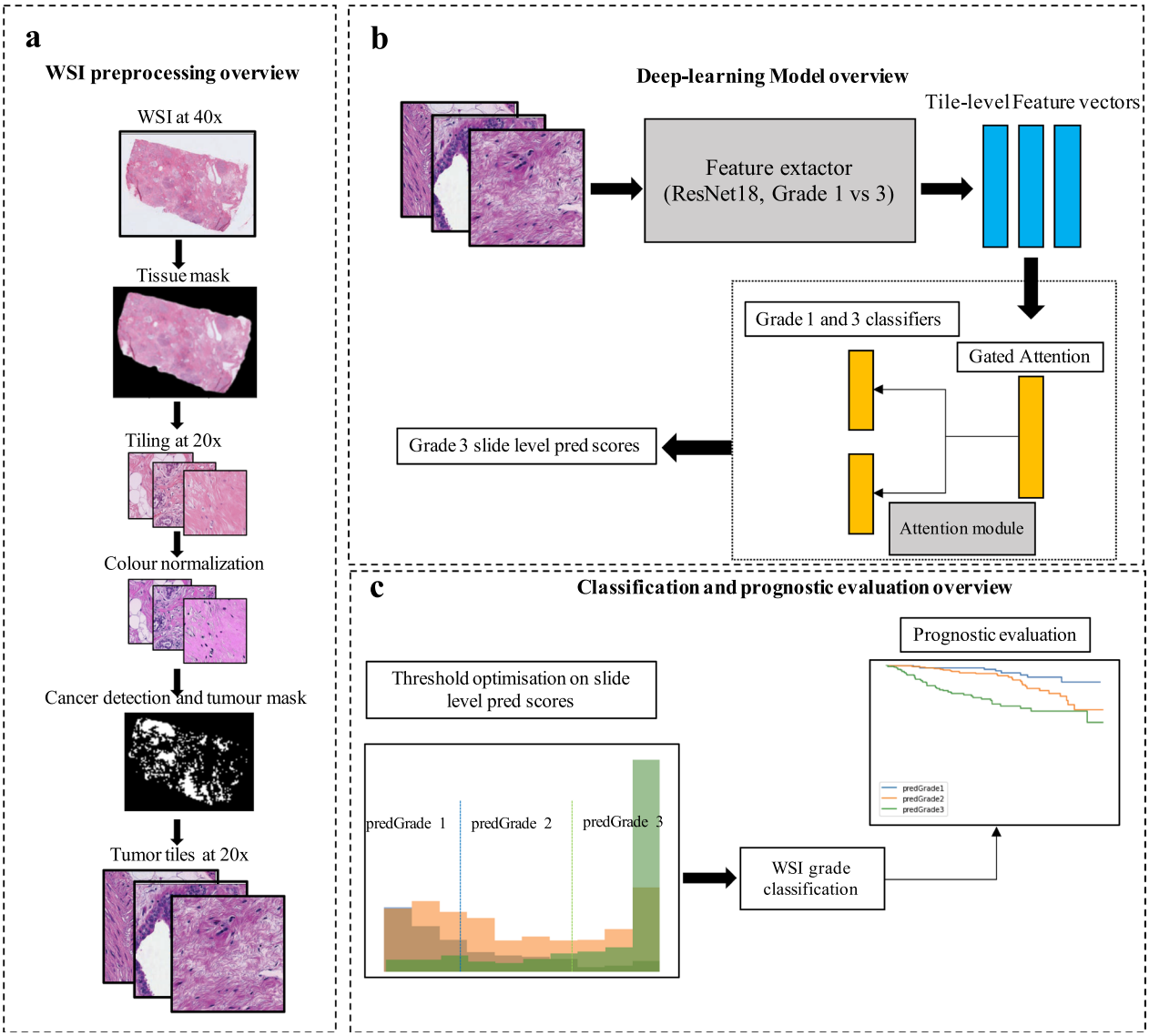

Fig. 2 Overview of the image pre-processing, model optimisation and performance evaluation. **a.** Standardised WSI preprocessing pipeline from retrieval of WSI at 40 × magnification to the cancer-detected tumour tiles from the WSI. **b.** Schematic overview of the image modelling strategy, including the deep CNN feature extractor and attention module. Model optimisation, hyperparameter tuning, and model selection were performed by cross validation (CV). In each CV training round, the feature extractor and attention module were trained from cancer tiles in the CV training set. In each CV validation round, the features extractor and attention model were re-optimised and subsequently, the CV validation set was evaluated. **c.** Two cut-offs were further derived from the slide-level prediction scores, which categorised the prediction scores into three-level predicted grades. The cut-offs were optimised by maximising the agreement between the predicted grade and clinical NHG. We further evaluated the prognostic performance of the predicted grade on recurrence-free survival

and the attention module. The feature extractor was trained to learn breast cancer domain-specific tile-level representations, and the attention module was trained to aggregate these tile-level representations to whole-slide-level prediction scores. Importantly, we specifically used different sub-splits of the training set for feature extractor optimisation and attention module optimisation, respectively,

Feature extractor module

The feature extractor was a binary weakly supervised learning model [9, 14]. We applied the Resnet-18 [19] CNN architecture initialised with weights pre-trained from Imagenet [20]. In order to reduce overfitting, we included a dropout layer with a probability of 0.2 after the global average pooling layer. Furthermore, a fully connected layer of 1024 hidden units followed by rectified

linear unit (ReLU) activation was added before the final output layer to increase the depth of the architecture. This model was trained on binary labels of NHG 1 versus NHG 3 with cross-entropy loss. We used stochastic gradient descent (SGD) optimiser [21] with a learning rate of 1e-5 and a momentum value of 0.9. At each training partial epoch end, we used the tuning set to validate the training performance and save the best model according to the lowest validation loss from the tuning set. We applied an early stopping to terminate the training when the validation loss showed no improvement after 50 consecutive partial epochs.

Attention module

We used the feature extractor to extract a 512-dimensional feature vector from the average pooling layer for each image tile in the attention module training set and the tuning set (Fig. 2b). These learned features were further used to train the attention module. The attention module consisted of an attention backbone and a classification layer with two output neurons, one for each class [18]. The attention backbone assigns and optimises the weights for each tile-level feature vector from each WSI and these derived weights sum up to one in order to be invariant to the number of tiles in each slide. The tile-to-slide feature aggregation was facilitated by the weighted average feature vectors from all image tiles in each slide [22]. The attention module was trained as a binary classification task to predict NHG 1 versus NHG 3 tumours using the cross-entropy loss. We used SGD optimiser with a learning rate 1e-5 and a momentum of 0.9. At each training epoch, we used the batch size of one single slide including all image tiles in it, based on our previous work [23].

Assignment of predicted histological grade (i.e. predGrade)

We obtained the slide-level predicted scores (i.e. $P[\text{class}=\text{NHG3}|\text{WSI}_i]$) for the entire training set ($N=1695$ WSIs) from the five-fold CV. We further optimised the two thresholds θ_1, θ_2 on $P(\text{class}=\text{NHG3}|\text{WSI}_i)$ to generate a three-level predicted grade (i.e. predGrade 1, 2, and 3). The thresholds θ_1, θ_2 were established through an exhaustive search by maximising the agreement between the clinical NHG and the predicted grades (i.e. predGrade 1, 2, and 3) using Cohen's Kappa score (κ). The thresholds were optimised on the training set using five-fold CV.

Assessment of model performance

Performance of predGrade was evaluated in both five-fold CV and in the independent external test set (SCAN-B cohort, $n=1262$). In performance evaluation in the SCAN-B cohort, the five CV models were treated as base models in an ensemble model, where the five predicted scores of $P(\text{class}=\text{NHG3}|\text{WSI}_i)$ were aggregated using the median across all base-model predictions. Next, we applied the thresholds θ_1, θ_2(see above and Fig. 2c) to map predictions to predGrade 1, 2, and 3.

First, we assessed the agreement between the predGrade and the clinical NHG in the independent external test using confusion matrices. Since the clinical NHG has high inter-rater variability, we utilise patient outcome (recurrence-free survival (RFS)), as our primary evaluation metric. We compared the prognostic performance of predGrade and the clinical NHG grade. The RFS defined recurrence (i.e. local or distant metastasis, detection of contralateral tumours) or death as the event outcome. Patients were followed from the initial diagnosis to the date of death/recurrence, emigration, or the last registration date, whichever occurred first. Kaplan–Meier (KM) curves for the predGrade and the clinical NHG on RFS using time since the initial diagnosis as the underlying timescale were used for visualisation purposes. Differences in survival probability among clinical NHG and predGrade subgroups were tested using the log-rank test. We assessed the associations between predGrade and RFS as well as clinical NHG and RFS separately by estimating hazard ratios (HRs) with 95% confidence intervals (CIs) using the Cox proportional hazard (PH) models. We used the time since the initial diagnosis as the underlying timescale. First, we fitted univariate Cox models for the predGrade and clinical NHG, respectively. Next, we fitted multivariable Cox models additionally adjusting for the well-established clinicopathological factors including tumour size, ER status, HER2 status, lymph node status, and age at the diagnosis. Tumour size was dichotomized as ≥ 20 mm or < 20 mm. ER status was positive if the immunohistochemical (IHC) staining indicated the presence of more than 10% ER positively stained cells. HER2 status was determined using IHC staining and FISH or SISH assay. Lymph node status denoted the presence of lymph node metastasis. Cases with missingness in the outcome or in any covariate were excluded from analyses. Two-sided alpha of 0.05 was used for all the statistical tests. Further, the c-index was computed for categorical predGrade, clinical NHG, and continuous slide-level predGrade score. Bootstrap resampling ($n=1000$) was used to calculate the confidence intervals (CI).

Software packages used for computer vision and statistical analyses

WSIs were read using the package openslide (v.3.4.1) [24]. Tissue masking, tiling, and colour normalisation steps were implemented using the packages scikit-image (v.0.16.2) [25], OpenCV (v.3.4.2) [26], SciPy (v.1.5.0) [27], pillow (v.7.2.0) [28], pandas (v.1.0.5) [29], and NumPy

(v.1.18.5) [30]. Further, TensorFlow (v.1.12.0) [31] was used to implement the pretrained invasive cancer detection model. The training, optimisation, and validation of the feature extractor and attention module were performed using PyTorch (v.1.7.1) [32]. The hyperparameter optimisation of the learning rate was performed using the tune library [33] from the package ray (v.0.8.6) [34]. All the WSI preprocessing and deep learning analyses were performed in Python v.3.6.10. Statistical testing for differences in hazard ratio estimates between the grade models was performed using the hr.comp2 function in the survcomp package (v.1.48.0) in R [35]. C-index was computed using the concordance function in the *survival* package (v.3.5.0) in R [36], and bootstrapping estimates of CI were calculated using the bootstrap function in the sjstats package (v.0.18.2) in R [37].

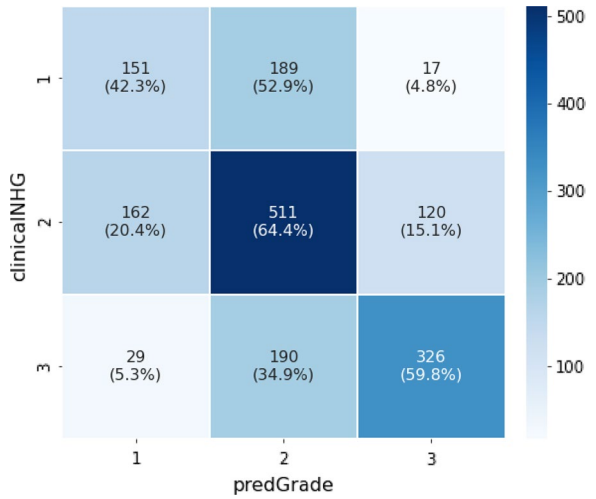

Fig. 3 Confusion matrix shows the agreement between the predGrade and clinicalNHG in the five-fold CV

Results

Model performance in the five-fold CV of SöS-BC-4 training set

Classification performance

Classification performance (Fig. 3) of predGrade in comparison with clinical NHG was first assessed using CV (see Methods section), indicating fair agreement between predGrade and clinical NHG (Cohen's $\kappa = 0.33$) [38]. 4.8% of clinical NHG 1 was classified as predGrade 3, and 5.3% of clinical NHG 3 was classified as predGrade 1. We have added an example of tiles from the correctly classified predGrade 1, 2, and 3 in (Additional file 1: Table S2).

Prognostic performance

Subsequently we evaluated the prognostic performance through time-to-event analysis. Figure 4 shows the KM curves comparing the risk stratification on RFS by clinical and predGrade. We observed similar stratification effects by the predGrade and clinicalNHG. Patients with clinicalNHG 1 or predGrade 1 showed the best survival, while patients assigned the clinical NHG 3 or predGrade 3 showed the worst survival (Fig. 4).

In the univariate Cox models, we observed similar effect sizes between the predGrade or clinical NHG and RFS (Fig. 5a and b). The predGrade 3 (HR = 3.12, 95%CI = 1.69–5.76, p-value < 0.001) and clinicalNHG 3 (HR = 3.19, 95% CI = 1.74–5.85, p-value < 0.001) showed approximately two times higher risk of an event as compared to predGrade 1 and clinicalNHG 1, respectively. Neither HR estimates for predGrade 2 nor clinicalNHG 2 were found to be significant (p-value < 0.05) (Fig. 5a and b).

In the multivariable Cox PH models, adjusting for tumour size, lymph node, ER, and HER2 status, the predGrade3 remained associated with a higher risk of death/recurrence (HR = 2.27, 95%CI = 1.07–4.82,

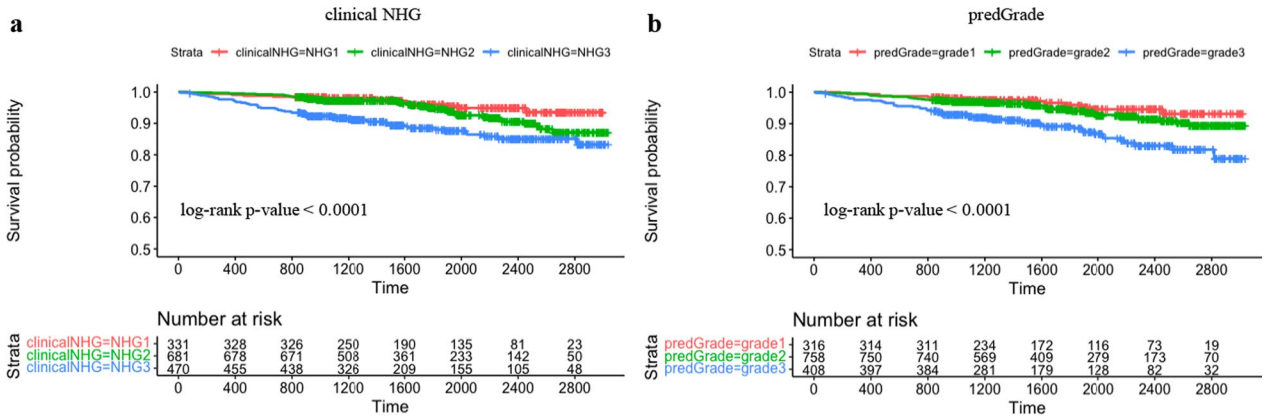

Fig. 4 Kaplan–Meier (KM) curves on recurrence-free survival from five-fold CV in the SöS-BC-4 training set. **a.** KM curve stratified by clinical NHG and **b.** KM curve stratified by predGrade

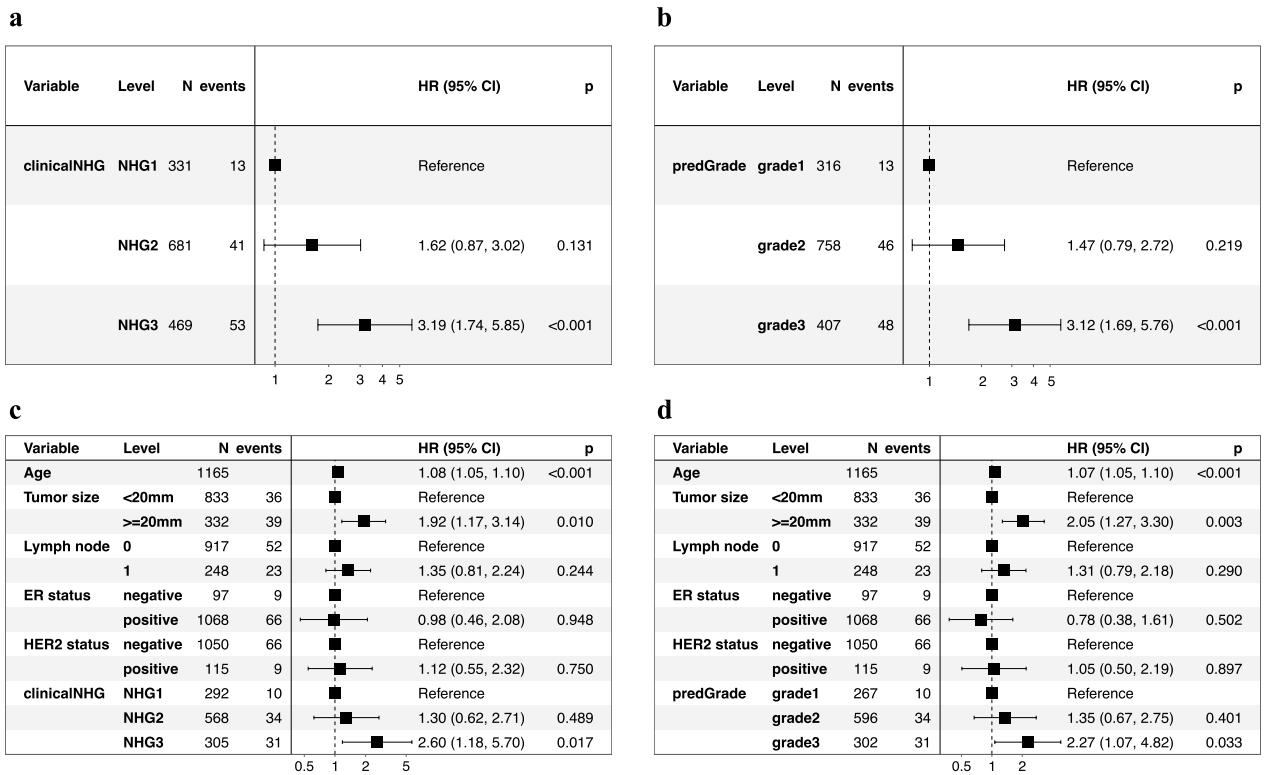

Fig. 5 Evaluation of the prognostic performance on recurrence-free survival (RFS) in five-fold CV in the SöS-BC-4 training set. **a.** Univariate Cox PH regression analysis between the clinical NHG and RFS; **b.** univariate Cox PH model between the predGrade and RFS; **c.** multivariable Cox PH model between the clinical NHG and RFS adjusting for age, tumour size, lymph node, ER, and HER2 status; **d.** multivariable Cox PH model between the predGrade and RFS adjusting for age, tumour size, lymph node, ER, and HER2 status

p-value = 0.033) (Fig. 5d). A similar association was also noted for the clinicalNHG 3 (HR = 2.60, 95% CI = 1.18–5.70, p-value = 0.017) (Fig. 5c). Neither predGrade2 nor clinicalNHG2 was found to be significantly (p-value < 0.05) associated with RFS (Figs. 5c and d). In addition, we observed a higher risk of death/recurrence linked to older age and tumour size equal to or larger than 20 mm, while the number of lymph nodes, ER status, and HER2 status was not related to RFS (Fig. 5c and d).

Model performance in the independent external test set (SCANB cohort)

Classification performance

The classification performance, as assessed by the confusion matrix (Fig. 6) and estimation of Cohen's $\kappa = 0.33$, was found to be consistent with CV results.

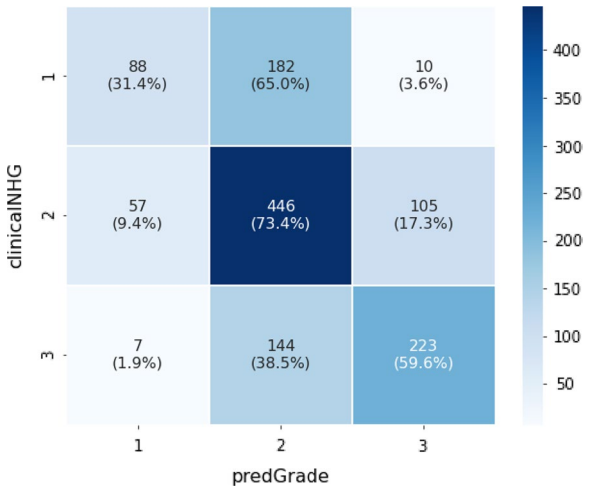

Fig. 6 Confusion matrix shows the agreement between the predGrade and clinical NHG in the independent external test set

Prognostic performance

In the independent external test set, KM curves showed similar risk stratification on RFS by the predGrade (log-rank p-value = 0.00049) as compared to the clinical NHG (log-rank p-value = 0.0002) (Fig. 7).

In the univariate Cox PH model, we observed similar effect sizes in the associations between clinical NHG, predGrade, and RFS (Fig. 8a and b). Patients

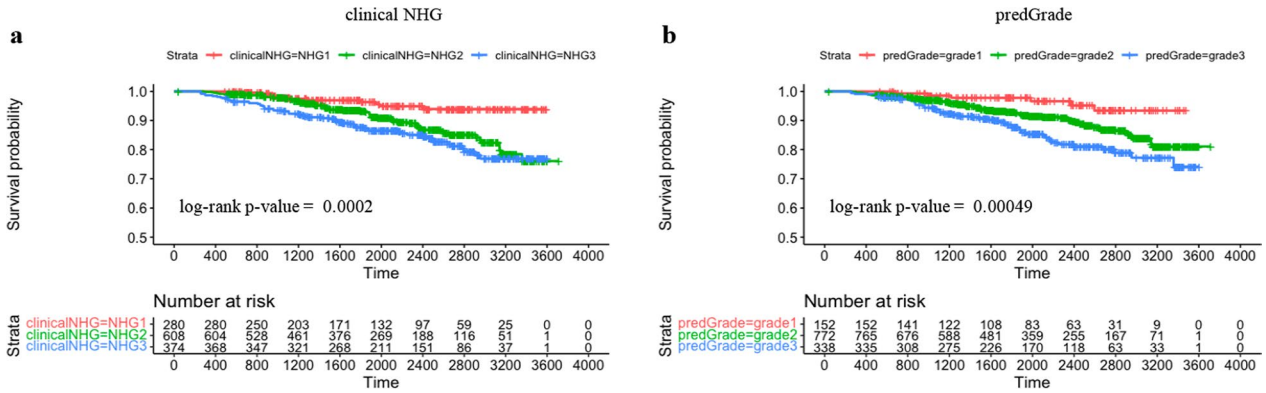

Fig. 7 Kaplan–Meier (KM) curves on recurrence-free survival stratified by clinical NHG and predGrade in the independent external test set. **a.** KM curve stratified by the clinical NHG and **b.** KM curve stratified by the predGrade

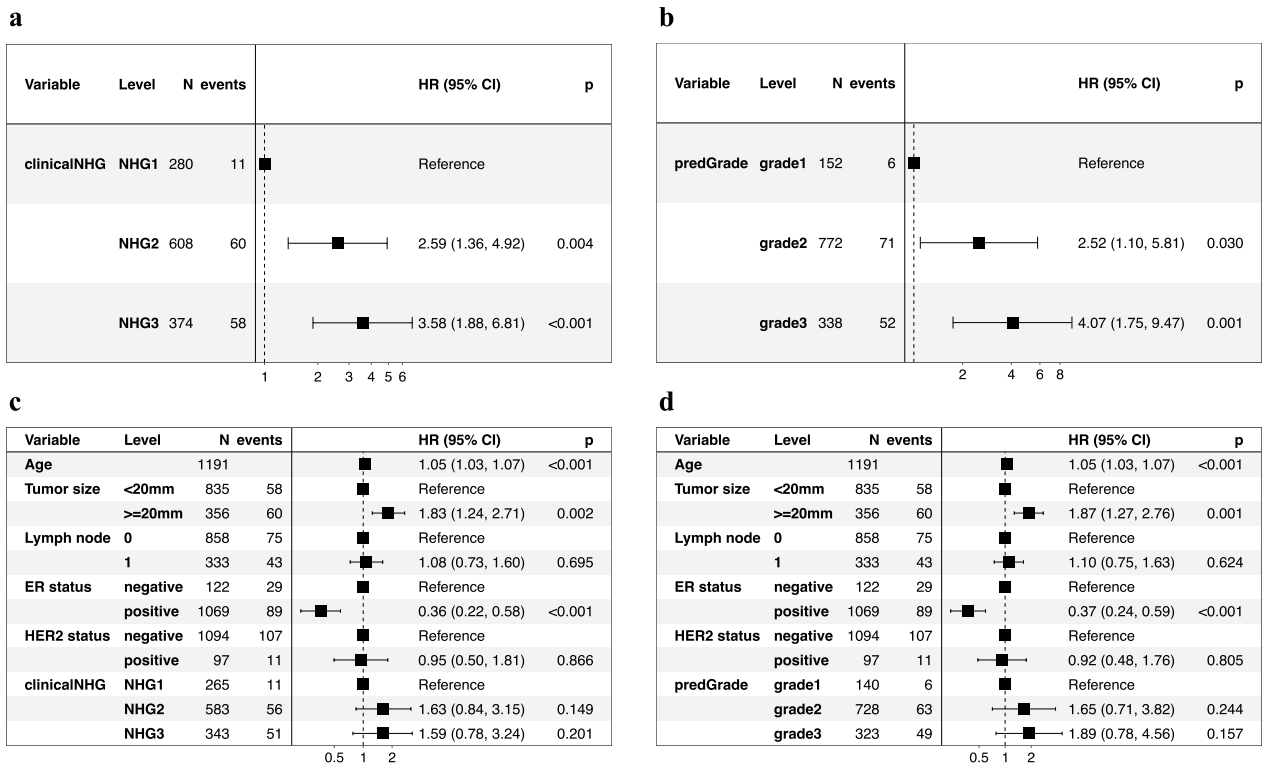

Fig. 8 Evaluation of the prognostic performance on recurrence-free survival (RFS) in the independent external test set. **a.** univariate Cox model between the clinical NHG and RFS; **b.** univariate Cox model between the predGrade and RFS; **c.** multivariable Cox model between the clinical NHG and RFS adjusting for age, tumour size, lymph node, ER, and HER2 status; and **d.** multivariable Cox model between the predGrade and RFS adjusting for age, tumour size, lymph node, ER, and HER2 status

with clinical NHG 3 (HR = 3.58, 95% CI 1.88–6.81, p-value < 0.001) or predGrade3 (HR = 4.07, 95% CI 1.75–9.47, p-value < 0.001) had three-to-four fold increased risks of death/recurrence (Fig. 8a and b) as compared to those with clinical NHG 1 or predGrade 1. On the other hand, the clinical NHG 2 (HR = 2.59, 95% CI 1.36–4.92, p-value = 0.004) and predGrade 2 (HR = 2.52, 95% CI 1.10–5.81, p-value = 0.030) were linked to around 2.5-fold increased risk of death/recurrence (Fig. 8a and b).

In multivariable Cox PH analysis, adjusting for tumour size, lymph node, ER, and HER2 status, the associations between the predGrade and the clinical NHG, with RFS were no longer found to be statistically significant (Fig. 8c and d), while the effect size estimate was in the same

direction as for the univariate analysis. In addition, we noted that older age at diagnosis and larger tumour size was linked to a higher risk of death/recurrence, while ER positive was related to a lower risk of death/recurrence (Fig. 8c and d). The number of lymph nodes and HER2 status was not related to RFS (Fig. 8c and d).

Next, we tested for the difference in hazard ratio estimates for clinical NHG 2 versus 1 and predGrade 2 versus 1, indicating no statistically significant difference (p-value > 0.05). We also tested for the difference in hazard ratio estimates for clinical NHG 3 versus 1 and predGrade 3 versus 1, revealing no significant difference (p-value > 0.05). The hazard ratios in this analysis were calculated from the multivariate Cox PH model after adjusting for the established covariates.

Lastly, we evaluated the c-index of the categorical predGrade and clinical NHG in the univariate Cox PH model. We observed the c-index of 0.62 (95% CI 0.57–0.67) and 0.64 (95% CI 0.59 – 0.69) for predGrade and clinical NHG, respectively. Further, we observed the c-index 0.62 (95% CI 0.56 – 0.68) for the continuous predGrade predicted slide score.

Subgroup analysis restricting to ER (+ ve) or ER(+ ve)/HER2(− ve) groups in the independent test set

We plotted KM curves on RFS stratified by clinical NHG and predGrade among ER (+ ve) patients (Fig. 9a and b) and ER (+ ve)/HER2(-ve) patients (Fig. 9c and d).

In the univariate Cox model restricted to ER (+ ve) patients, we observed increased risks of death/recurrence associated with clinical NHG 3 (HR = 2.63, 95% CI 1.32–5.24, p-value = 0.006) or predGrade 3 (HR = 3.27, (95% CI1.37–7.82, p-value = 0.008) (Fig. 10a and b). The clinicalNHG 2 had a 2.42-fold increased risk of death/recurrence, while the predGrade2 was not related to RFS albeit a similar point estimate HR of 2.07 (95% CI 0.86–4.64, p-value = 0.091) (Fig. 10a and b). In the multivariable analysis, the association between clinical NHG and

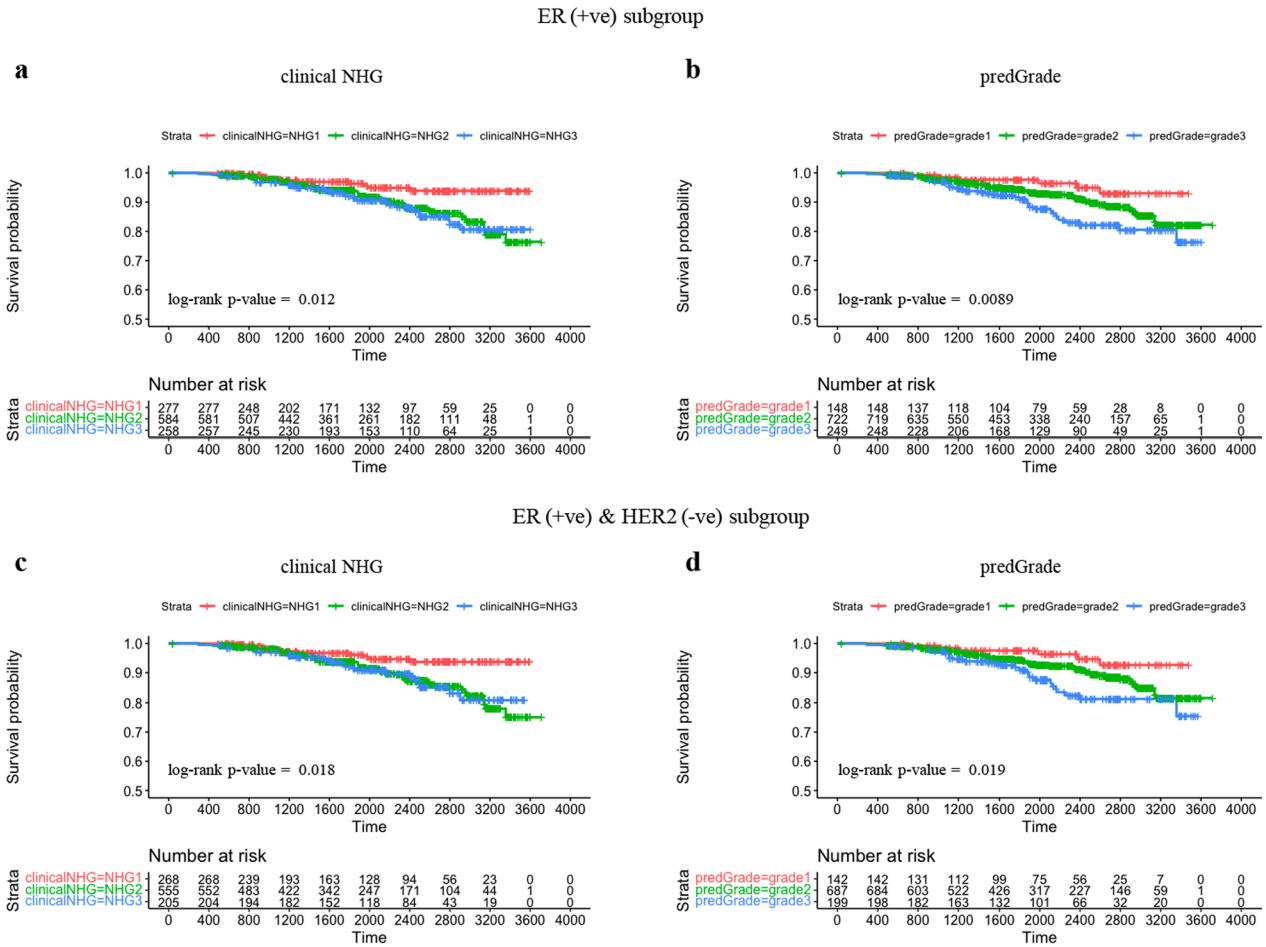

Fig. 9 Subgroup analysis: Kaplan–Meier (KM) curves on recurrence-free survival (RFS) in the independent external test set within ER(+ ve) or ER(+ ve)/HER2(− ve) groups. **a.** KM stratified by clinical NHG in ER(+ ve) patients; **b.** KM stratified by predGrade in ER(+ ve) patients; **c.** KM stratified by clinical NHG in ER(+ ve)/HER2(− ve) patients. **d.** KM stratified by predGrade in ER(+ ve)/HER2(-ve) patients

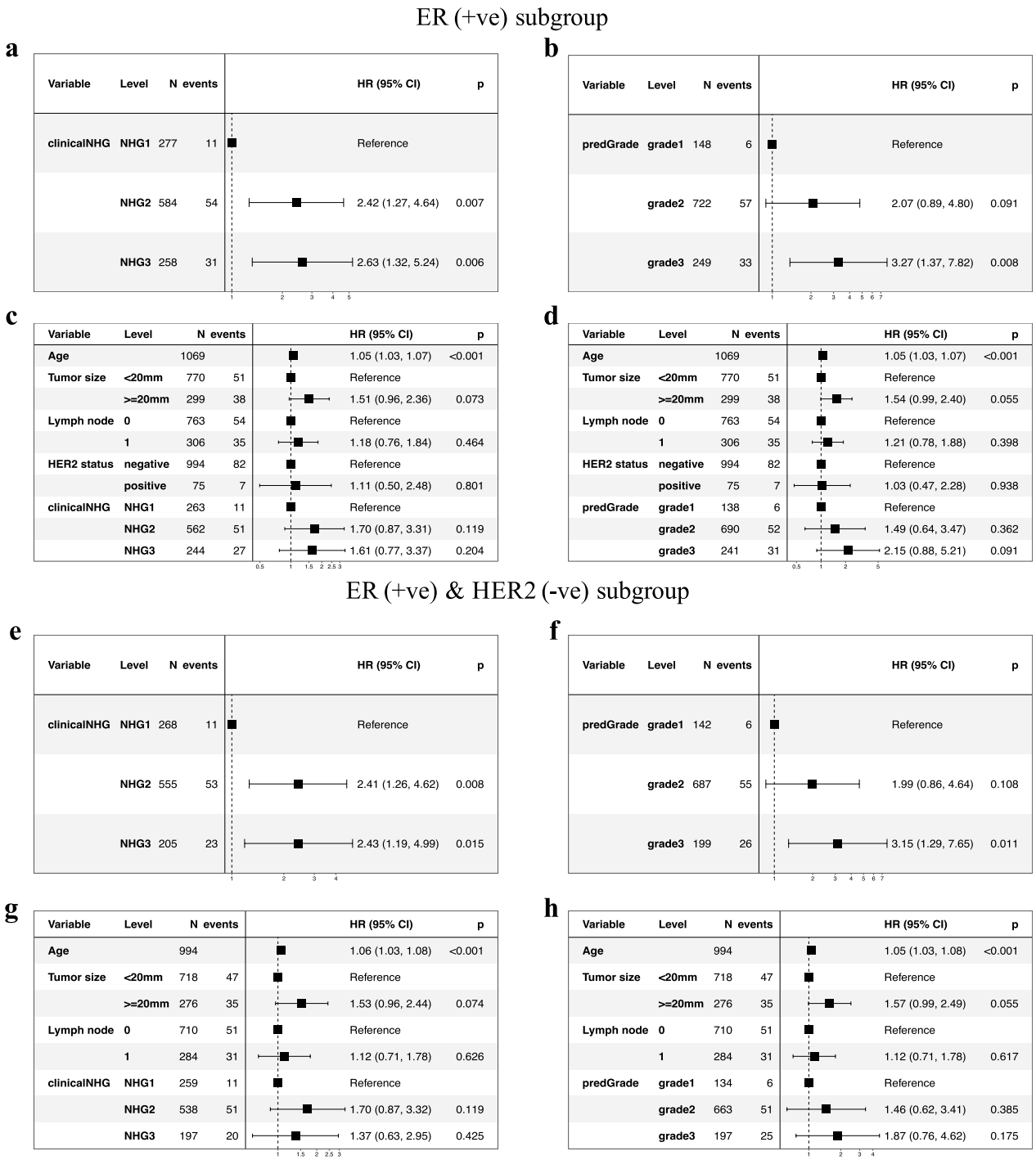

Fig. 10 Evaluation of the prognostic performance (RFS) of predGrade on in the independent external test cohort within ER(+ve) or ER(+ve)/HER2(−ve) groups. Univariate Cox PH model for **a.** clinical NHG and **b.** predGrade on RFS among ER(+ve) patients, **c.** multivariable Cox PH models between c) clinical NHG, **d.** predGrade and FRS among ER(+ve) patients adjusting for age, tumour size, and lymph node. Univariate Cox PH model between **e.** clinical NHG, **f.** predGrade and RFS among ER(+ve)/HER2(−ve) patients. Multivariable Cox model between **g.** clinical NHG, **h.** predGrade and RFS among ER(+ve)/HER2(−ve) patients adjusting for age, tumour size, and lymph node

RFS as well as the association between predGrade and RFS diminished and was no longer statistically significant (Fig. 10c and d).

Among the ER(+ve)/HER2(−ve) patients, we observed HR = 2.43 (95% CI 1.19–4.99, p-value = 0.015) for clinical NHG 3 and HR = 3.15 (95% CI 1.29–7.65, p-value = 0.011) for predGrade 3 (Fig. 10e and f). The clinical NHG 2 was associated with the RFS (HR = 2.41, 95%: CI 1.26–4.62, p-value = 0.008), while the predGrade2 was not related to the RFS (Figs. 10e and f). In the multivariable Cox PH models among ER(+ve)/HER2(−ve) patients, neither clinical NHG nor predGrade was related to RFS (Fig. 10g and h), while older age was linked to poor RFS in the analysis for predGrade (Figure h).

Again, we tested for the difference in HR estimates between NHG and predGrade in the subgroup analyses, both for grade 1 versus 2 and grade 1 versus 3 and found that neither was significantly different (p-value > 0.05), indicating that the prognostic performance was similar between NHG and the predGrade model. Further, we evaluated the c-index of the categorical clinical NHG (0.59 (95% CI 0.55–0.64)) and predGrade (0.61 (95% CI 0.56–0.65)) in the univariate Cox PH analysis. Lastly, we observed the c-index of 0.62 (95% CI 0.56–0.67) for the continuous predGrade predicted slide score.

Discussion

In this study, we developed a deep learning model to reproduce clinical NHG breast cancer patients. The proposed model was first evaluated using CV, followed by validation in a fully independent external test set. Histological grading of breast tumours is routinely assessed in the clinical setting and remains an important prognostic factor contributing to clinical decision making, especially for ER (+ve)/HER2(−ve) patients. However, it is well known that NHG suffers from substantial inter-assessor and inter-laboratory variability, which motivates the development of decision support solutions that can improve quality and consistency in the assessment.

Our proposed model, predGrade, exhibited a fair label agreement with the clinical NHG ($\kappa = 0.33$). This imperfect agreement is likely driven by the ground truth labelling noise, both during training and validation, mostly from the intermediate NHG 2, given the high inter-rater variability observed in NHG 2 [4]. However, interestingly we observed similar prognostic performances (RFS) for predGrade compared with the clinicalNHG. We also noted similar prognostic performance between predGrade and clinical NHG when restricted to the clinically relevant subgroups of ER(+ve) or ER(+ve) and HER2(−ve) patients. Our results suggest that the deep learning-based predGrade provides similar prognostic performance (RFS) of the clinical NHG, which is a key consideration since the conventional NHG grade is primarily used in clinical settings as a prognostic factor [39]. This indicates that deep learning-based solutions can provide decision support based on the same principles of histological grading while offering the benefits of being objective and consistent. The model has the potential to reduce inter-assessor variability between pathologists and systematic variability between pathology laboratories, which has recently been shown to result in [7] unequal diagnostic quality for patients.

Previous studies have focussed on classifying NHG1 and 2 (low-intermediate) combined versus NHG3 (high) [11, 12]. Wetstein et al. reported a 37% increased risk of recurrence associated with a high grade compared to a low–intermediate grade. Wang et al. on the other hand demonstrated that a deep learning model optimised to discriminate NHG 3 versus 1 can [14] further stratify the intermediate NHG 2 into NHG2-low and NHG2-high, enabling improved prognostic stratification of the NHG 2 group of patients.

Jaroensri et al. mainly focussed on the development of the composite NHG score by developing predictive models for the individual subcomponent scores and validated them against the panel of pathologists [13]. Despite a difference in survival endpoints, we observed similar c-index of 0.62 in comparison with 0.60 (95% CI 0.55–0.65) reported by Jaroensri et al. for the combined deep learning-based NHG grade. We believe that the NHG score, rather than the individual subcomponents, has a higher clinical relevance, at least with respect to prognostic applications.

In our modelling strategy, we make some key assumptions. Due to the inter-observer and inter-laboratory variability present in the clinical NHG with relatively higher variability in NHG 2, thus we decided to optimise our model for the classification of NHG 3 and 1 where there is less label noise. As stated earlier, this modelling strategy is based on the assumption that the tumour grading exists in a continuum instead of the discrete labels at the morphological level with a spectrum ranging from low to high grade, apart from assuming more reliable ground truth for low and high NHG. Such variabilities in the ground truth labels are one of the important challenges in developing deep learning-based clinical decision support tools, especially the development of weakly supervised learning-based models where the label is only available at the WSI level. An alternative approach to the modelling problem would be to attempt to reduce label noise, which could be achieved by e.g. utilising consensus labels assigned by a set of assessors as performed in [13]. However, such attempts remain challenging due to the number of resources required and the shortage of pathologists available in most parts of the world.

Conclusion

In this study we developed and validated a deep learning-based model for breast cancer histological grading, providing a similar three-group grade assignment as the well-established Nottingham Histological Grading system. We found that despite the relatively low concordance of grade labels with clinical NHG, the proposed model provides equivalent prognostic stratification of breast cancer patients. The proposed model has the potential to provide objective and consistent decision support for histological grading, reducing previously observed inter-assessor and systematic inter-laboratory variability in breast cancer histological grading, and with the benefit of increased equality for patients and reduced risk for over- and under-treatment.

Abbreviations

NHG	Nottingham Histological Grade
CNN	Convolutional neural network
RFS	Recurrence-free survival
HR	Hazard ratio
ER	Estrogen receptor
HER2	Human epidermal growth factor receptor 2
WSI	Whole-slide image
CV	Cross-validation
H&E	Haematoxylin & Eosin
FFPE	Formalin-fixed paraffin-embedded
MIL	Multiple instance learning
ReLU	Rectified linear unit
SGD	Stochastic gradient descent
KM	Kaplan–Meier
CI	Confidence interval
PH	Proportional hazard
IHC	Immunohistochemical

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-024-01770-4.

Additional file 1. Table1: Baseline characteristics of the patients in SöS-BC-4 training set and the SCANB cohort. **Figure 1:** Boxplots of the distribution of number of tiles for each WSI in the two study cohorts. **Table 2:** Visualization of tiles from the SöS-BC4 training set.

Acknowledgements

The authors acknowledge patients, clinicians, and hospital staff participating in the SCAN-B study, the staff at the central SCAN-B laboratory at Division of Oncology, Lund University, the Swedish National Breast Cancer Quality Registry (NKBC), Regional Cancer Center South and the South Swedish Breast Cancer Group (SSBCG). We also further acknowledge Marike Gabrielson, Senior Research Specialist, at Karolinska Institutet (Sweden) for the reviews and feedbacks on the pre-final draft of the manuscript. SCAN-B was funded by the Swedish Cancer Society, the Mrs. Berta Kamprad Foundation, the Lund-Lausanne L2-Bridge/Biltema Foundation, the Mats Paulsson Foundation, and Swedish governmental funding (ALF).

Author contributions

AS contributed to data preparation, methods development, software implementation, statistical analysis and computing, visualization, preparing draft of manuscript, editing manuscript. PW contributed in data preparation, software implementation, contribution to preparing, editing and approving manuscript. YW contributed to data preparation, software implementation, visualization, editing and approval of manuscript. BL contributed in supervision, preparing of draft manuscript, editing and approval of manuscript. JVC contributed in study materials, preparation of clinicopathological information for the SCAN-B study, editing and approval of manuscript. JH contributed in conceptualization, resources, supervision, directed the study, editing and approval of manuscript. MR conceived the study, contributed to conceptualization, resources, supervision, methodology, preparing of draft manuscript, editing and approval of manuscript, and directed the study.

Funding

Open access funding provided by Karolinska Institute. This work was supported by funding from the Swedish Research Council, Swedish Cancer Society, Karolinska Institutet, ERA PerMed (ERAPERMED2019-224-ABCAP), VINNOVA (SwAIPP project), MedTechLabs, Swedish e-science Research Centre (SeRC)—eCPC, Stockholm Region, Stockholm Cancer Society, and Swedish Breast Cancer Association.

Availability of data and materials

Due to legal constraints, data in the present study cannot be made openly available without constraints. Reasonable access requests to the corresponding author will be considered. All analyses in this study are based on publicly available software packages (see Methods).

Declarations

Ethics approval and consent to participate

The study has approval by the regional ethics review board (Stockholm, Sweden).

Consent for publication

Not applicable.

Competing interests

JH has obtained speaker's honoraria or advisory board remunerations from Roche, Novartis, AstraZeneca, Eli Lilly and MSD and has received institutional research grants from Cepheid and Novartis. MR and JH are shareholders of Stratipath AB. All other authors have declared no conflicts of interest.

Author details

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [2]Division of Precision Medicine, Department of Medicine, NYU Grossman School of Medicine, New York, NY, USA. [3]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. [4]Department of Clinical Pathology and Cancer Diagnostics, Karolinska University Hospital, Stockholm, Sweden. [5]MedTechLabs, BioClinicum, Karolinska University Hospital, Solna, Sweden. [6]Division of Oncology, Department of Clinical Sciences, Lund University, Lund, Sweden.

Received: 24 April 2023 Accepted: 15 January 2024
Published online: 29 January 2024

References

1. Rakha EA, El-Sayed ME, Lee AHS, Elston CW, Grainge MJ, Hodi Z, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26(19):3153–8.
2. Ellis IO, Elston CW. The value of histological grade in breast-cancer-experience from a large study with long-term follow-up. In: Journal of pathology. Wiley Baffins Lane Chichester, W Sussex, England PO19 1UD; 1990. p. A358–A358.
3. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010;12(4):207.
4. Ginter PS, Idress R, D'Alfonso TM, Fineberg S, Jaffer S, Sattar AK, et al. Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy. Mod Pathol. 2021;34(4):701–9.
5. Zhang R, Chen H-J, Wei B, Zhang H-Y, Pang Z-G, Zhu H, et al. Reproducibility of the Nottingham modification of the Scarff-Bloom-Richardson

histological grading system and the complementary value of Ki-67 to this system. Chin Med J. 2010;123(15):1976–82.
6. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers CCHJ, van der Wall E, Overbeek LIH, 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(3):769–80.
7. Acs B, Fredriksson I, Rönnlund C, Hagerling C, Ehinger A, Kovács A, et al. Variability in breast cancer biomarker assessment and the effect on oncological treatment decisions: a nationwide 5-year population-based study. Cancers. 2021. https://doi.org/10.3390/cancers13051166.
8. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020;21(2):222–32.
9. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–9.
10. Coudray N, Moreira AL, Sakellaropoulos T, Fenyö D, Razavian N, Tsirigos A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning [Internet]. medRxiv. bioRxiv; 2017. https://www.nature.com/articles/s41591-018-0177-5?sf197831152=1
11. 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;3(4):30.
12. Wetstein SC, de Jong VMT, Stathonikos N, Opdam M, Dackus GMHE, Pluim JPW, et al. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. Sci Rep. 2022;12(1):15102.
13. Jaroensri R, Wulczyn E, Hegde N, Brown T, Flament-Auvigne I, Tan F, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. NPJ Breast Cancer. 2022;8(1):113.
14. Wang Y, Acs B, Robertson S, Liu B, Solorzano L, Wählby C, et al. Improved breast cancer histological grading using deep learning. Ann Oncol. 2022;33(1):89–98.
15. Vallon-Christersson J, Häkkinen J, Hegardt C, Saal LH, Larsson C, Ehinger A, et al. Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. Sci Rep. 2019;9(1):12184.
16. Pech-Pacheco JL, Cristobal G, Chamorro-Martinez J, Fernandez-Valdivia J. Diatom autofocusing in brightfield microscopy: a comparative study. In: Proceedings 15th international conference on pattern recognition ICPR-2000. 2000. p. 314–7 vol.3.
17. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International symposium on biomedical imaging: from nano to macro. 2009. p. 1107–10.
18. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555–70.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 770–8.
20. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. 2009. p. 248–55.
21. Bottou L. Stochastic gradient learning in neural networks. 1991 [cited 2022 Nov 14]; https://www.semanticscholar.org/paper/82eec4af1475de9a7e876bcbaddb4a0c4a1dc187
22. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Dy J, Krause A (Eds). 2018;80:2127–36.
23. Weitz P, Wang Y, Hartman J, Rantalainen M. An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives. In: 2021 IEEE/CVF international conference on computer vision workshops (ICCVW). IEEE; 2021. p. 611–9.
24. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: a vendor-neutral software foundation for digital pathology. J Pathol Inform. 2013;27(4):27.
25. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-image: image processing in python. PeerJ. 2014;19(2):e453.
26. Bradski G, Kaehler A. Learning OpenCV: computer vision with the OpenCV Library. "O'Reilly Media, Inc."; 2008. 580 p.
27. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72.
28. Clark A. Pillow (PIL Fork) Documentation [Internet]. readthedocs; 2015. https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf
29. The pandas development team. pandas-dev/pandas: Pandas [Internet]. Zenodo; 2023. https://doi.org/10.5281/zenodo.3509134
30. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62.
31. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C et al. TensorFlow: large-scale machine learning on heterogeneous systems [Internet]. 2015. https://www.tensorflow.org/
32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style high-performance deep learning library. In: Advances in neural information processing systems. Curran Associates Inc.; 2019. p. 8024–35.
33. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: a research platform for distributed model selection and training [Internet]. arXiv [cs.LG]. 2018. http://arxiv.org/abs/1807.05118
34. Moritz P, Nishihara R, Wang S, Tumanov A, Liaw R, Liang E, et al. Ray: A Distributed Framework for Emerging AI Applications [Internet]. arXiv [cs.DC]. 2017. http://arxiv.org/abs/1712.05889
35. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011;27(22):3206–8.
36. Therneau TM. A package for survival analysis in R [Internet]. 2023. https://CRAN.R-project.org/package=survival
37. Lüdecke D. Sjstats: statistical functions for regression models (Version 0.18.2) [Internet]. 2022. https://CRAN.R-project.org/package=sjstats
38. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
39. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham prognostic index in primary breast cancer. Breast Cancer Res Treat. 1992;22(3):207–19.

Publisher's Note