ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY driving innovation in oncology

ORIGINAL ARTICLE

Improved breast cancer histological grading using deep learning

Y. Wang[1], B. Acs[2,3], S. Robertson[2,3], B. Liu[1], L. Solorzano[4], C. Wählby[4], J. Hartman[2,3,5†] & M. Rantalainen[1,5*†]

Departments of [1]Medical Epidemiology and Biostatistics; [2]Oncology-Pathology, Karolinska Institutet, Stockholm; [3]Department of Clinical Pathology and Cancer Diagnostics, Karolinska University Hospital, Stockholm; [4]Department of Information Technology and SciLifeLab, Uppsala University, Uppsala; [5]MedTechLabs, BioClinicum, Karolinska University Hospital, Solna, Sweden

Available online 29 September 2021

Background: The Nottingham histological grade (NHG) is a well-established prognostic factor for breast cancer that is broadly used in clinical decision making. However, ∼50% of patients are classified as grade 2, an intermediate risk group with low clinical value. To improve risk stratification of NHG 2 breast cancer patients, we developed and validated a novel histological grade model (DeepGrade) based on digital whole-slide histopathology images (WSIs) and deep learning.

Patients and methods: In this observational retrospective study, routine WSIs stained with haematoxylin and eosin from 1567 patients were utilised for model optimisation and validation. Model generalisability was further evaluated in an external test set with 1262 patients. NHG 2 cases were stratified into two groups, DG2-high and DG2-low, and the prognostic value was assessed. The main outcome was recurrence-free survival.

Results: DeepGrade provides independent prognostic information for stratification of NHG 2 cases in the internal test set, where DG2-high showed an increased risk for recurrence (hazard ratio [HR] 2.94, 95% confidence interval [CI] 1.24-6.97, $P = 0.015$) compared with the DG2-low group after adjusting for established risk factors (independent test data). DG2-low also shared phenotypic similarities with NHG 1, and DG2-high with NHG 3, suggesting that the model identifies morphological patterns in NHG 2 that are associated with more aggressive tumours. The prognostic value of DeepGrade was further assessed in the external test set, confirming an increased risk for recurrence in DG2-high (HR 1.91, 95% CI 1.11-3.29, $P = 0.019$).

Conclusions: The proposed model-based stratification of patients with NHG 2 tumours is prognostic and adds clinically relevant information over routine histological grading. The methodology offers a cost-effective alternative to molecular profiling to extract information relevant for clinical decisions.

Key words: breast cancer, digital pathology, deep learning, artificial intelligence, histological grade

INTRODUCTION

Breast cancer histological grade is a well-established clinical variable in breast cancer that comprises information from three aspects, namely, the degree of tubule formation, nuclear pleomorphism and mitotic counts. Compared with other widely used prognostic factors that only consider a single aspect such as age, tumour size or lymph node status, histological grading takes both morphology and proliferation into consideration, and therefore contributes with unique prognostic significance and is broadly utilised in clinical decision making.[1,2] The most broadly adopted grading classification system is the Nottingham grading system, modified by Elston and Ellis from the Bloom—Richardson grading system, and its prognostic value has been validated in studies with varied populations.

A higher Nottingham histological grade (NHG) is associated with poor prognosis, and it is an indication for more aggressive treatment, while lower grade indicates lower risk of recurrence and allows for more conservative treatment.[3,4] However, grading is conducted manually by pathologists and is associated with a substantial uncertainty indicated by large interassessor variability.[5-8] Previous studies have found higher concordance in identifying the most aggressive tumours (NHG 3), and lower concordance in distinguishing between NHG 1 and 2 tumours.[9,10]

The intermediate group (NHG 2) accounts for approximately half of the patient population,[11,12] but exhibits larger variation with regard to morphological patterns and survival outcomes[13] in comparison with NHG 1 and 3. As histological grade remains central in determining therapeutic regimens,[14] the heterogeneity in the NHG 2 group

*Correspondence to: Dr Mattias Rantalainen, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, SE-171 77 Stockholm, Sweden. Tel: +46(08)-52482465 (direct line), +46 8524-80000 (switch)

E-mail: mattias.rantalainen@ki.se (M. Rantalainen).

†Equal contribution.

poses challenges for deciding optimal treatment for individual patients. Therefore multiple approaches to address these uncertainties, and to reduce both overtreatment and undertreatment, have been proposed. Gene expression profiling has been applied to dichotomise patients with NHG 2 tumours into groups with different outcomes.[13,15] It has also been suggested that NHG 2 could be eliminated altogether.[16] Gene expression assays, including Oncotype Dx and Prosigna, can predict risk of recurrence and death in patients with intermediate risk tumours.[17-19] However, molecular diagnostics remain expensive and time-consuming compared with histopathology-based diagnostics for routine clinical applications.

The emergence of digital pathology[20] and routine acquisition of high-resolution whole-slide histopathology images (WSIs) now enables application of advanced image analysis in the standard clinical setting. Advances in artificial intelligence have also opened up new opportunities for histopathology image analysis using deep learning.[21,22] Deep convolutional neural networks (CNNs) have recently been applied successfully for detection and pathological classification across multiple cancer types[23-26] including breast cancer.[27,28] Deep CNNs offer cost-effective solutions for improved cancer diagnostics based on routine histopathology slide images.

In this study, we propose a novel deep learning-based approach, DeepGrade, for histological grading of breast cancers based on digitised haematoxylin and eosin (HE)-stained WSIs, with a particular focus on improving prognostic stratification of NHG 2 tumours. The model was developed for classification of NHG 1 and NHG 3 morphological patterns, followed with restratification of NHG 2 tumours using the learned patterns. The proposed model was validated in independent internal and external test data with respect to patient outcomes.

METHODS

Patients

The study comprises female patients with primary invasive breast cancer as primary diagnosis from four different studies and sites: ClinSeq breast cancer study (ClinSeq-BC)[15,29] ($N = 256$), TCGA breast cancer study[30] (TCGA-BC) ($N = 559$), SöS-BC-1 breast cancer cohort phase I (SöS-BC-1) ($N = 752$) and a subset of the SCAN-B study,[31] consisting of patients diagnosed in Lund (Sweden) ($N = 1262$, external test set) (Supplementary Table S1, available at https://doi.org/10.1016/j.annonc.2021.09.007). ClinSeq-BC includes patients diagnosed in Stockholm South General Hospital in 2012 or had surgery between 2001 and 2008 at the Karolinska University Hospital (Stockholm, Sweden); SöS-BC-1 is a retrospective cohort of patients diagnosed at the Stockholm South General Hospital (Stockholm, Sweden) between April 2012 and October 2014 and between October 2015 and May 2018. SCAN-B enrolled patients diagnosed as having primary invasive disease from 2010 to 2019; here we only include patients diagnosed in Lund (Sweden). Clinical and imaging data from the Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) were retrieved from the TCGA database and from pathology reports with approval by the TCGA data access committee (dbGAP projectID:5621). Clinical data for ClinSeq-BC, SöS-BC-1 and SCAN-B cohorts were retrieved from the Swedish National Breast Cancer Registry. HE-stained formalin-fixed paraffin-embedded (FFPE) histopathology slides from resected tumours were digitised in-house for ClinSeq-BC, SöS-BC-1 and SCAN-B. Available WSIs from TCGA-BC were downloaded from https://portal.gdc.cancer.gov/ in year 2019. Only WSIs that contained invasive cancer were included. WSIs from patients treated with neoadjuvant chemotherapy and WSIs scanned at 20× were excluded. Please see consort diagram for further details (Supplementary Figures S1 and S2, available at https://doi.org/10.1016/j.annonc.2021.09.007). Only one WSI image is included for each patient.

WSI preprocessing: overview

WSIs were tiled into tiles of 598 × 598 pixels with a down-sampled resolution equivalent to 20 × (271 × 271 µm). All tiles were automatically quality controlled with respect to sharpness. Colour normalisation was applied to adjust for stain and scanner colour variability (Supplementary Figure S3, available at https://doi.org/10.1016/j.annonc.2021.09.007). Areas in slides containing invasive cancer, annotated by pathologists or predicted by model, were included in analyses (Supplementary Figure S4, available at https://doi.org/10.1016/j.annonc.2021.09.007, Supplementary Table S2, available at https://doi.org/10.1016/j.annonc.2021.09.007).

Optimisation of the DeepGrade CNN model for histological grade

Tumours of NHG 1 and NHG 3 from ClinSeq-BC, TCGA and SöS-BC-1 were randomly split into training set ($N = 844$, 70.6%), test set 1 ($N = 136$, 11.4%) and test set 2 ($N = 215$, 18.0%) on a patient level (Supplementary Table S3 and S4, available at https://doi.org/10.1016/j.annonc.2021.09.007). The training set was further split into training ($N = 674$, 56.4%) and tuning set ($N = 170$, 14.2%). For ClinSeq-BC, we used the ground truth annotated by pathologists to extract tiles within tumour regions; for TCGA-BC and SöS-BC-1, we applied the trained invasive cancer detection model to generate tumour regions and extract tiles from such regions (Supplementary Figure S5, available at https://doi.org/10.1016/j.annonc.2021.09.007). Several CNN architectures were evaluated in the tuning set (Supplementary Table S5, available at https://doi.org/10.1016/j.annonc.2021.09.007). The Inception V3 model[32] demonstrated the best prediction performance. As it is also widely adopted with overall satisfactory performance for various prediction tasks in previous studies,[24,23] it was used in all of the base models in the ensemble. Tiles in the training and tuning sets were used to optimise the DeepGrade model, an ensemble including 20 deep CNN base models (InceptionV3 model[32] with weights initialised from a model pretrained by ImageNet[33]) for classification of NHG 1 and 3. By averaging

over a number of models in the ensemble, improvements in prediction performance can be achieved by reduction of variance. AUCs for each deep CNN model were summarised in Supplementary Table S6, available at https://doi.org/10.1016/j.annonc.2021.09.007. Deep learning was performed using the Keras (2.2.4) framework with TensorFlow (1.12) backend.

Histological grade prediction by CNN models

Tile-level predicted class probabilities from the ensemble were averaged. Slide-level predictions were based on the upper quartile of the tile-level distribution for each slide. Classification performance was assessed by slide-level receiver operating characteristic (ROC) curves and the area under the curve (AUC) ('pROC' package in R). Bootstrapping (2000 samples) was performed to estimate the 95% confidence interval (CI) of AUC.[34] Optimal threshold for binary class label (NHG 1 and 3) assignment was determined by Youden's method.[35] Re-stratification of NHG 2 tumours from ClinSeq-BC and TCGA-BC was achieved by dichotomisation [DeepGrade 2 (DG2)-high or DG2-low] using the estimated classification threshold.

Survival analysis and statistical tests

We analysed and compared recurrence-free survival (RFS) rates among patients in clinically assigned histological grades (NHG), as well as between NHG 2 tumours that were predicted as high grade (DG2-high) and low grade (DG2-low) by DeepGrade. Clinical outcome data were available in ClinSeq-BC and TCGA-BC; survival analysis was performed on these two data sources together. A recurrence event was defined as having locoregional or distant relapses, contralateral tumours or death. A death without any detected metastasis was assumed to have experienced a tumour metastasis before death.[15,36] The time-to-event period was defined as the number of days between the initial diagnosis date and the date of one of the following events: local or regional relapse, distant metastasis, death or last follow-up. Kaplan–Meier curves were generated to visualise survival outcomes between groups (R-packages 'survminer' and 'survival'). Multivariate Cox proportional hazards regression models were used to estimate adjusted hazard ratios (HRs) and 95% CI (R function 'coxph' and R-package 'forestmodel'). Other risk factors included in the model were age, tumour size, lymph node status, human epidermal growth factor receptor 2 (HER2) status and estrogen receptor (ER) status. Tumour size was dichotomised into categories with tumour diameter $\geq$20 mm or $<$20 mm; lymph node status was dichotomised as having or not having lymph node metastases; HER2 status was determined using immunohistochemical staining in conjunction with FISH test (decided in clinical routine); ER status was determined based on a 10% cut-off for positively stained cells by immunohistochemical staining. Patients with missing data in any of the risk factors were excluded in the multivariate Cox regression analysis. Patient characteristics after removing missing data are summarised in Supplementary Table S7, available at https://doi.org/10.1016/j.annonc.2021.09.007.

To compare differences in distribution for categorical variables, Fisher's exact test or chi-square test was employed depending on whether the minimum value in a subgroup is smaller than 5. To test for difference in mean age, the *t*-test was applied. To test for differences in Ki67 score, the Mann–Whitney *U* test was used. The two-sided *P* value was reported and a *P* value $<$0.05 was considered as statistically significant. Analyses were performed using R (3.6.3).

Further methodological details can be found in the Supplementary Methods, available at https://doi.org/10.1016/j.annonc.2021.09.007.

RESULTS

A deep CNN ensemble model for discrimination between NHG 1 and 3 tumours

An ensemble consisting of 20 deep CNN models for binary classification of NHG 1 and NHG 3 based on routine HE WSIs was optimised (DeepGrade). The DeepGrade model was applied for re-stratification of NHG 2 tumours into two groups: DG2-high, sharing similarity with NHG 3, and DG2-low, sharing similarity with NHG 1. The prognostic performance of DeepGrade for NHG 2 stratification was evaluated based on time-to-event analysis (RFS) in independent test data (Figure 1). The number of available training tiles in ClinSeq-BC, TCGA-BC and SöS-BC-1 datasets were 1.56 million (M), 4.39 M and 3.25 M, respectively. The ROC-AUC ranged from 0.919 (95% CI 0.884-0.955) to 0.937 (95% CI 0.887-0.987) (independent test data, Supplementary Figure S6, available at https://doi.org/10.1016/j.annonc.2021.09.007, Supplementary Table S8, available at https://doi.org/10.1016/j.annonc.2021.09.007), indicating good performance in classification of NHG 1 and 3.

HE-stained routine WSIs can be applied to re-stratify patients with NHG 2 tumours into two groups with significant difference in outcomes

DeepGrade was then applied to re-stratify NHG 2 cases into two groups based on the assumption that the model has capacity to capture morphological patterns related to NHG 1 and 3 also present in NHG 2. Out of 372 NHG 2 cases (independent test data), 242 (65.0%) were classified as DG2-low and 130 (35.0%) were classified as DG2-high. The prognostic performance (RFS) of DeepGrade was visualised by Kaplan–Meier curves, and independent prognostic value was evaluated by multivariable Cox proportional hazards models, adjusting for established risk factors, including age, tumour size, HER2 status, ER status and lymph node status (Figure 2). The Nottingham histological grade was prognostic in stratification of NHG 3 and NHG 1, with an estimated HR of 3.74 (95% CI 1.12-12.55, $P = 0.033$, $N = 670$, Figures 2A and 3A). The DeepGrade model (DG2-low, DG2-high) was found to be an independent prognostic factor for stratification of NHG 2 with an HR of 2.94 (95% CI 1.24-6.97, $P = 0.015$, $N = 305$; Figures 2B and 3B). We also evaluated

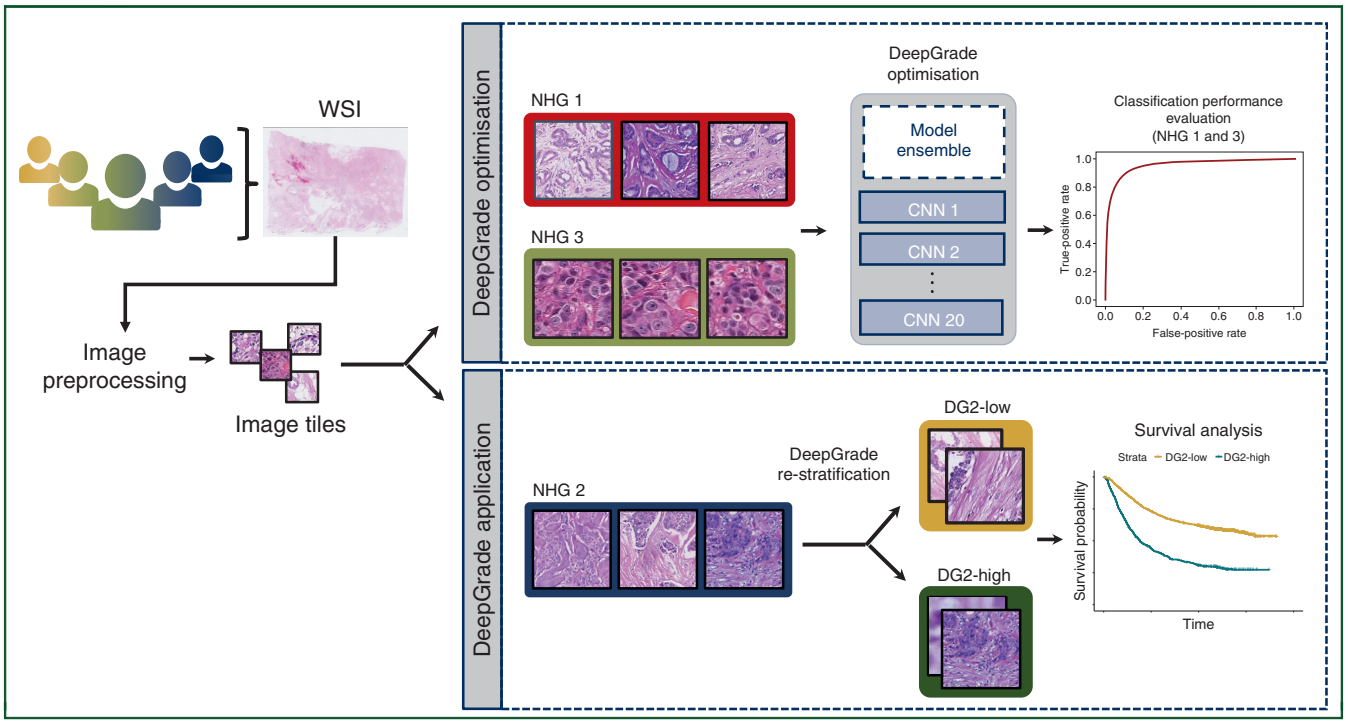

Figure 1. Schematic overview of the optimisation, application and evaluation of the DeepGrade model.
Stained histopathology slides from breast cancer surgical specimens were scanned, tumour regions were segmented and image tiles were extracted. Patients with tumours graded as Nottingham histological grade (NHG) 1 and 3 were used to optimise the DeepGrade model, a convolutional neural network (CNN) ensemble including 20 base models. The DeepGrade model was subsequently applied to re-stratify NHG 2 cases. Finally, time-to-event analysis was applied to evaluate the prognostic performance. DG, DeepGrade; WSI, whole-slide histopathology image.

prognostic performance in the NHG 2 ER-positive (ER+) subgroup (HR 3.21, 95% CI 1.32−7.79, $P = 0.010$, $N = 287$; Figures 2C and 3C), and in the smaller NHG 2 ER-positive and node-negative (ER+Node−) subgroup (HR 3.03, 95% CI 0.91-10.10, $P = 0.071$, $N = 183$; Figures 2D and 3D). Same analyses were repeated in the HER2-negative (HER2−) patients in ER+HER2− and ER+HER2−Node− subgroups, with an HR of 3.68 (95% CI 1.49-9.07, $P = 0.005$, $N = 262$, Supplementary Figure S7A and C, available at https://doi.org/10.1016/j.annonc.2021.09.007) and 3.14 (95% CI 0.95-10.41, $P = 0.061$, $N = 166$; Supplementary Figure S7B and D, available at https://doi.org/10.1016/j.annonc.2021.09.007), respectively. These results suggest that the DeepGrade model offers independent prognostic stratification of the NHG 2 group, with an HR comparable to that observed between NHG 1 and 3.

Characterisation of DG2-low and DG2-high groups

The DG2-low and DG2-high cases were characterised to ascertain potential differences in clinical variables and intrinsic molecular subtypes.[17] We observed no significant differences in Ki67 scores (Figure 4A, Supplementary Table S9, available at https://doi.org/10.1016/j.annonc.2021.09.007) between DG2-low and DG2-high ($P = 0.625$, Mann−Whitney U test). The Ki67 score was significantly higher in DG2-low compared with NHG 1 ($P = 2.80 \times 10^{-3}$), while the Ki67 score in DG2-high was significantly lower than in NHG 3 ($P = 2.94 \times 10^{-4}$, Mann−Whitney U test). The distribution of intrinsic molecular subtypes in DG2-low and NHG 1 was not significantly different, with luminal A being the dominating intrinsic subtype (Figure 4B, $P = 0.618$; Fisher's exact test). By contrast, we observed a difference in subtype distribution between DG2-high and NHG 3 ($P = 2.20 \times 10^{-16}$, Fisher's exact test), where DG2-high had a higher proportion of luminal A and lower proportion of basal-like subtypes compared with NHG 3. Luminal B was present in a higher proportion in DG2-high compared with DG2-low. These results indicate that DG2-low shares strong similarities with NHG 1 and DG2-high with NHG 3. The findings also suggest that DeepGrade identifies morphological patterns in NHG 2 associated with more aggressive tumours.

Furthermore, we assessed if the NHG subcomponent scores (mitotic count, nuclear pleomorphism and tubular formation) in the NHG 2 group were associated with DG2-low and DG2-high. Only the score for the mitotic count subcomponent ($P = 6.54 \times 10^{-3}$, Fisher's exact test) was statistically different in DG2-high compared with DG2-low (Figure 4C, Supplementary Table S10, available at https://doi.org/10.1016/j.annonc.2021.09.007). Finally, we investigated if the sum of NHG subcomponent scores (Supplementary Methods, available at https://doi.org/10.1016/j.annonc.2021.09.007) provided prognostic stratification of NHG 2, as a potential alternative to DeepGrade. NHG 2 tumours were dichotomised into ScoreSum-low (≤ 6) and ScoreSum-high (≥ 7). No significant difference in RFS could be detected between these two groups (multivariable Cox proportional hazard model; HR 1.19, 95% CI 0.46-3.10, $P = 0.715$; Supplementary Figure S8A-C, available at

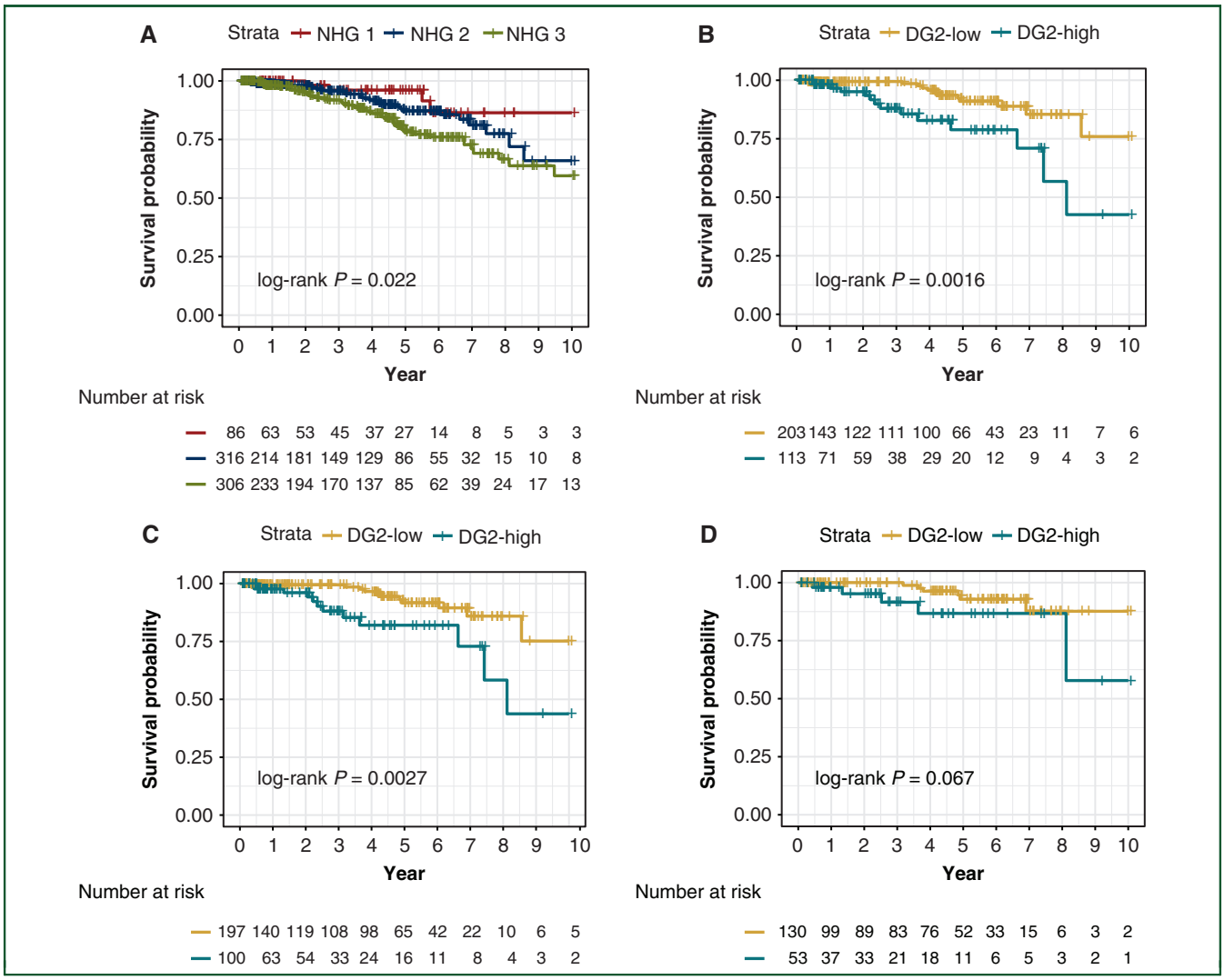

Figure 2. Recurrence-free survival outcomes for breast cancer patients by Nottingham histological grade, and by DeepGrade-re-stratified Nottingham histological grade (NHG) 2 patients.

(A) Kaplan—Meier curves for patients stratified by NHG 1-3. NHG 2 had an intermediate survival rate, whereas the NHG 3 had the worst prognosis. (B) Kaplan—Meier curves for DeepGrade-re-stratified NHG 2 cases. Worse prognosis was observed in the DG2-high group. (C) Kaplan—Meier curves for the DeepGrade-re-stratified NHG 2 ER-positive subgroup. (D) Kaplan—Meier curves for the DeepGrade-re-stratified NHG 2 ER-positive and node-negative subgroup. DG, DeepGrade; ER, estrogen receptor.

https://doi.org/10.1016/j.annonc.2021.09.007), indicating that prognostic stratification of NHG 2 cannot be achieved based on the NHG score sum. This suggests that DeepGrade captures independent prognostic information for stratification of NHG 2 cases, and more importantly, these features are not substitutable by factors included in routine pathological assessment.

Validation of DeepGrade in an external cohort

To evaluate the generalisability of the DeepGrade model, we analysed WSIs from 1262 patients in the SCAN-B Lund cohort, representing a completely independent external study material. The cohort is population representative and the digitised WSIs were only used for the purpose of testing the model performance.

We first applied the DeepGrade model to classify NHG 1 and NHG 3 patients, and the resulting AUC was 0.907 (95% CI 0.885-0.930; Supplementary Figure S9A, available at https://doi.org/10.1016/j.annonc.2021.09.007), which demonstrated that the model was able to provide classification performance comparable with the internal test data. In addition, the RFS between DeepGrade-classified NHG 1 and 3 patients was similar to that with clinically assigned NHG 1 and 3 (Supplementary Figure S9B-E, available at https://doi.org/10.1016/j.annonc.2021.09.007).

The DeepGrade model was applied to re-stratify the NHG 2 group ($N = 608$) and to evaluate prognostic performance (Figures 5 and 6). A total of 376 (61.8%) patients were classified as DG2-low, and 232 (38.2%) patients as DG2-high.

The DeepGrade model (DG2-low, DG2-high) provided significant prognostic value for stratification of NHG 2 ($P = 0.0045$, log-rank test; Figure 5B) with an HR of 1.91 (95% CI 1.11-3.29, $P = 0.019$, $N = 583$; Figure 6B). In comparison, the prognostic value comparing NHG 3 and NHG 1 was found to have an HR of 1.59 (95% CI 0.78-3.24, $P = 0.201$, $N = 1191$; Figure 6A).

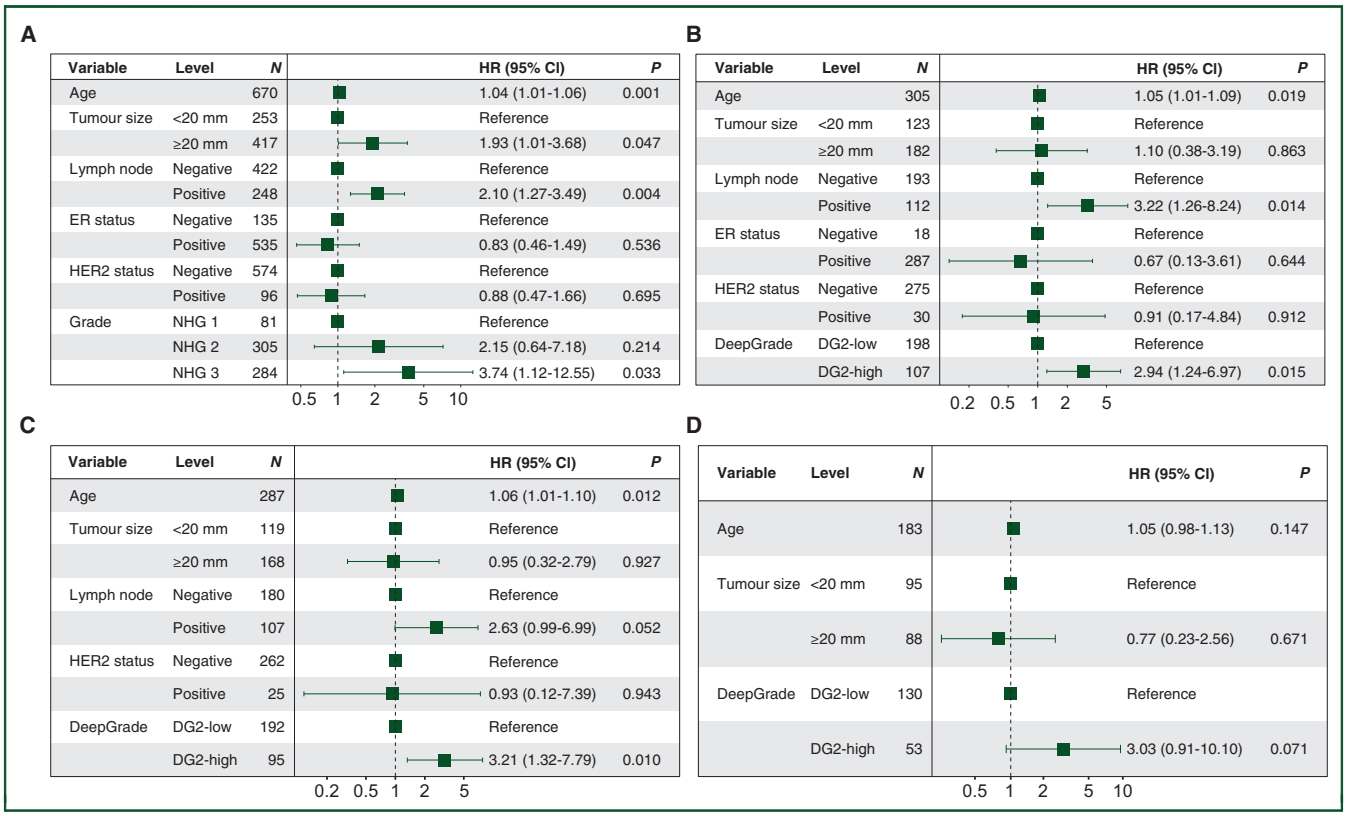

Figure 3. Forest plots from multivariable Cox proportional hazard regression.
(A) Results from multivariable Cox proportional hazard regression analysis of patients stratified by Nottingham histological grade (NHG) 1-3. NHG 2 was not significantly different from NHG 1, while the hazard ratio (HR) between NHG 1 and 3 was 3.74 (95% CI 1.12-12.55, $P = 0.033$). (B) Results from multivariable Cox proportional hazard regression analyses of DeepGrade-re-stratified NHG 2 cases. The estimated HR between DG2-low and DG2-high was 2.94 (95% CI 1.24-6.97, $P = 0.015$). (C) Results from Cox proportional hazard regression in the DeepGrade-re-stratified NHG 2 ER-positive subgroup (HR 3.21, 95% CI 1.32-7.79, $P = 0.010$). (D) Results from Cox proportional hazard regression of the DeepGrade-re-stratified NHG 2 ER-positive and node-negative subgroup (HR 3.03; 95% CI 0.91-10.10, $P = 0.071$). All Cox proportional hazard models were adjusted for age, tumour size, lymph node metastases, ER status and HER2 status. CI, confidence interval; DG, DeepGrade; ER, estrogen receptor; HR, hazard ratio.

In the ER+ subgroup and the ER+Node− subgroup the HR was 1.66 (95% CI 0.96-2.88, $P = 0.070$, $N = 567$; Figure 6C) and 1.68 (95% CI 0.87-3.22, $P = 0.119$, $N = 405$; Figure 6D), respectively. The prognostic performance was further evaluated in the ER+HER2− subgroup (HR 1.79, 95% CI 1.02-3.12, $P = 0.041$, $N = 538$), and the ER+HER2− Node− subgroup (HR 1.90, 95% CI 0.98-3.70, $P = 0.058$, $N = 383$; Supplementary Figure S10, available at https://doi.org/10.1016/j.annonc.2021.09.007). The Ki67 score distribution was found to be different between NHG 1 and DG2-low ($P = 3.16 \times 10^{-11}$, Mann−Whitney U test), and between DG2-low and DG2-high ($P = 4.24 \times 10^{-05}$), as well as between DG2-high and NHG 3 ($P = 1.55 \times 10^{-37}$; Supplementary Figure S11, available at https://doi.org/10.1016/j.annonc.2021.09.007).

DISCUSSION

In this study, we developed and validated a novel method, DeepGrade, for histological grading of breast tumours, focused on re-stratification of NHG 2 cases. We demonstrated that DeepGrade-based stratification of NHG 2 (intermediate risk) cases provides independent prognostic information of a magnitude comparable to that observed between NHG 1 and 3. We also showed that the DG2-low and DG2-high groups share clinical phenotype characteristics with NHG 1 and 3, while none of the routine clinical variables provided information that could be used for prognostic stratification of NHG 2. The morphological characteristics of NHG 1 and 3 are usually clearly distinguished by pathologists, whereas NHG 2 has more vague and variable characteristics. DeepGrade captures morphological features in NHG 2 tumours that are shared with the more well-defined NHG 1 and 3 groups, and exploits these to enable consistent and precise stratification of NHG 2. Because the NHG-associated morphologies encompass a continuous spectrum, rather than distinct defined groups, the application of a computer model to capture NHG-related patterns is most likely required to consistently distinguish subtle grade-related morphological differences. These results were further verified in a completely independent external test set.

We found that DeepGrade provided independent prognostic value when applied for stratification of the NHG 2 group (HR 2.94), which was of a comparable effect size to previously reported studies applying gene expression profiling. Wang et al.[15] proposed a model based on a 34-gene panel (RNA sequencing) that dichotomised NHG 2 tumours into a high- and a low-risk group with an HR of 2.43. In another study, a 97-gene signature [Genomic Grade

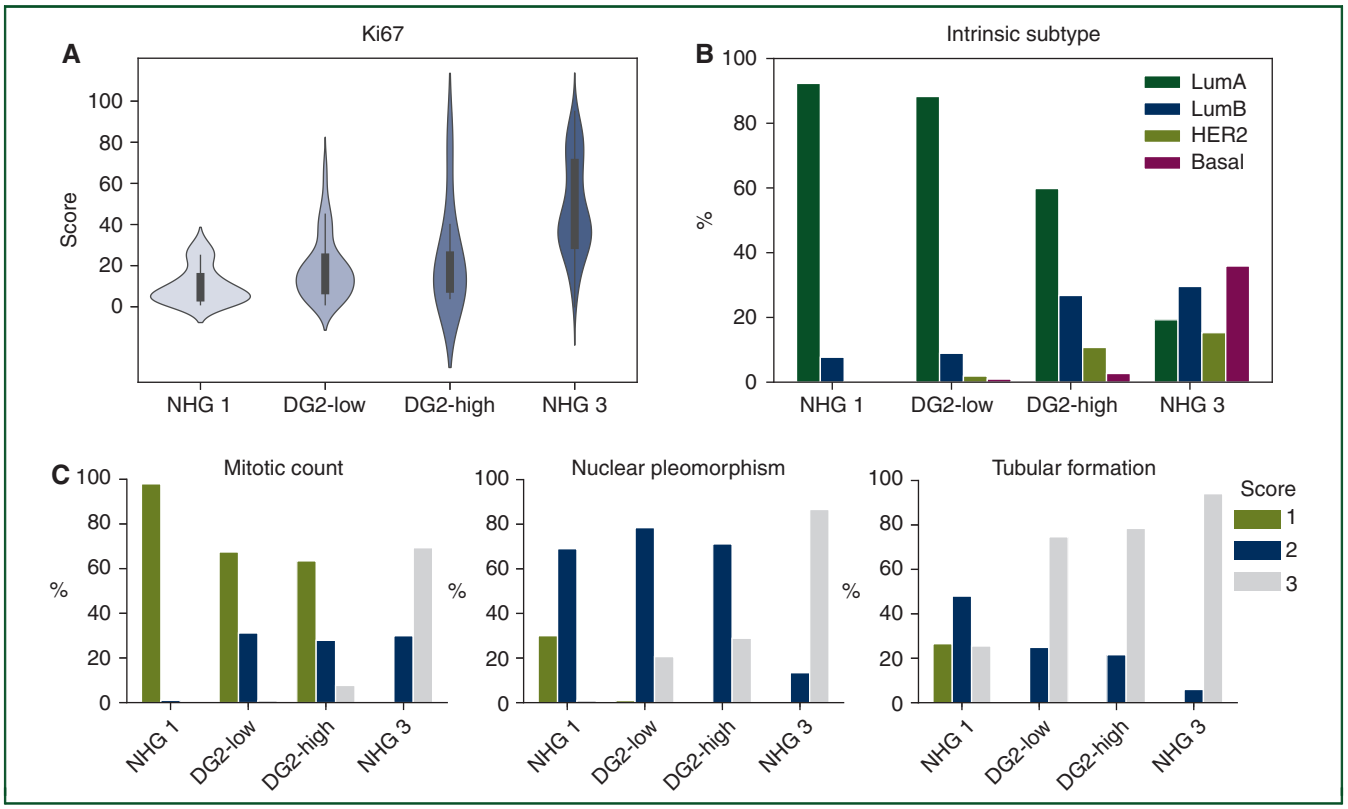

Figure 4. Ki67 score, intrinsic subtype distribution and NHG subcomponent score distribution across NHG 1, DG2-low and DG2-high, and NHG 3 patient groups. (A) Violin plot showing distribution of Ki67 (data only available in ClinSeq-BC). The distribution was different between NHG 1 and DG2-low ($P = 2.80 \times 10^{-3}$, Mann—Whitney *U* test), and different between DG2-high and NHG 3 ($P = 2.94 \times 10^{-4}$, Mann—Whitney *U* test). No significant difference between DG2-low and DG2-high was observed ($P = 0.625$, Mann—Whitney *U* test). (B) Distribution of intrinsic subtypes. DG2-low was similar to NHG 1 with the majority being luminal A ($P = 0.618$, Fisher's exact test). DG2-high has a larger proportion of HER2 and basal type compared with DG2-low. The subtype distribution for NHG 3 is significantly different with DG2-high ($P = 2.20 \times 10^{-16}$, Fisher's exact test). (C) Distribution of three NHG subcomponent scores with respect to mitotic count, nuclear polymorphism and tubular formation. Only the score for mitotic count was found to be significantly different between DG2-low and DG2-high ($P = 6.54 \times 10^{-3}$, Fisher's exact test). Basal, basal-like; DG, DeepGrade; Her2, Her2-enriched; LumA, luminal A; LumB, luminal B; NHG, Nottingham histological grade.

Index (GGI)] dichotomised NHG 2 cases into two groups with an estimated HR of 3.61.[13] Another RNA-seq based method (EndoPredict) for ER-positive, HER2-negative tamoxifen-treated patients has a reported HR of 5.07 for NHG 2 cases.[37] These studies have consistently reported that NHG 2 tumours can be further stratified with independent prognostic value. In contrast to molecular-based assays, DeepGrade only requires HE-stained FFPE sections, which are part of the routine diagnostic work-up, thus providing a rapid and cost-effective alternative that could increase access to improved diagnostics.

Furthermore, we found that the subcomponent scores of nuclear pleomorphism and tubule formation were not significantly different between DG2-low and DG2-high, and that stratification of NHG 2 could not be achieved by dichotomisation based on NHG score sum 6 and 7. The results suggest that the stratification provided by the DeepGrade model is not solely based on the NHG subcomponents, rather, the representations of morphological patterns that the model captures is, at least to some extent, independent of the established NHG criteria. Other risk factors including tumour size and lymph node stage were also similar between DG2-low and DG2-high, indicating that DeepGrade could discern more subtle features, and provide improved capability for tumour grading.

Despite the similarity in morphologies, we identified notable differences in the distribution of intrinsic subtypes. DG2-low shared a similar molecular subtype distribution with NHG 1. The DG2-high group shared a similar distribution of HER2-enriched and luminal B subtypes as the NHG 3 group, whereas the frequency of the basal-like subtype was lower in the DG2-high group compared with the NHG 3 group. Basal-like tumours are typically associated with a high Ki67 score,[38,39] and we also observed a significant difference in Ki67 score between DG2-high and NHG 3. It should be noted that basal-like tumours have distinct morphological features including severe nuclear atypia and expression of basal cytokeratins.[40] The luminal B subtype, by contrast, is defined by more subtle morphological features, and cannot always be easily distinguished from the luminal A subtype by microscopic assessment. As luminal A tumours are associated with lower grades and good prognosis whereas HER2-enriched and basal-like tumours typically have higher grade and poorer prognosis,[41] these results further support that DeepGrade identifies aggressive morphological patterns. High-grade morphological features are considered as indications for adjuvant systemic chemotherapy,[14,42] suggesting a potential for DeepGrade to strengthen information relevant for therapeutic selection.

https://doi.org/10.1016/j.annonc.2021.09.007

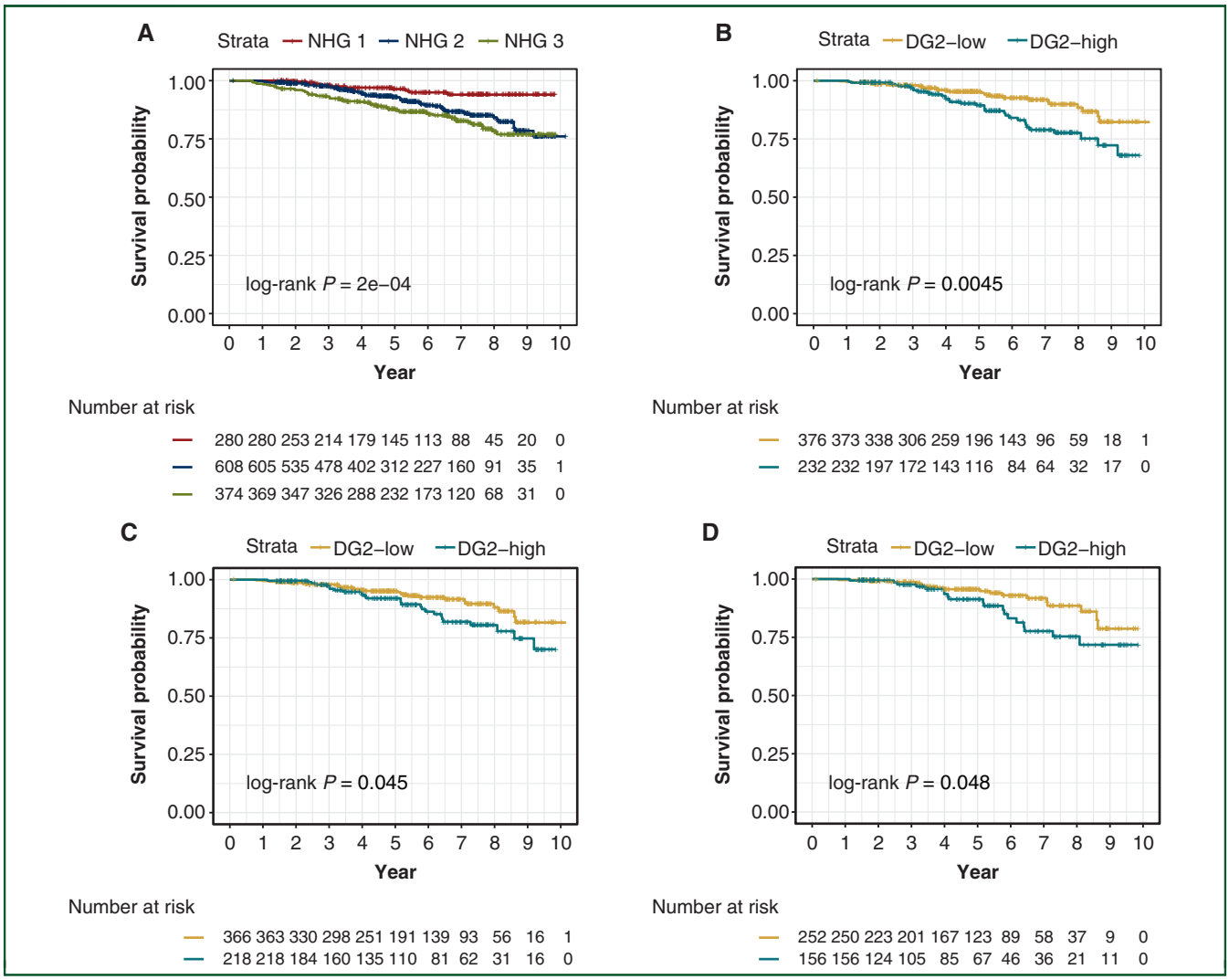

Figure 5. Recurrence-free survival outcomes for breast cancer patients from the external test set by Nottingham histological grade, and by DeepGrade re-stratified NHG 2 patients.

(A) Kaplan—Meier curves for patients stratified by NHG 1-3. NHG 2 had an intermediate survival rate, whereas the NHG 3 had the worst prognosis. (B) Kaplan—Meier curves for DeepGrade-re-stratified NHG 2 cases. DG2-high displayed significantly worse prognosis compared with the DG2-low group. (C) Kaplan—Meier curves for the DeepGrade-re-stratified NHG 2 ER-positive subgroup from the external test set. (D) Kaplan—Meier curves for the DeepGrade re-stratified NHG 2 ER-positive and node-negative subgroup from the external test set. DG, DeepGrade; ER, estrogen receptor; NHG, Nottingham histological grade.

The prognostic performance of DeepGrade was further validated in an external cohort, which confirmed independent prognostic value between re-stratified NHG 2 (DG2-low and DG2-high) patients (HR 1.91).

A central objective in modern breast oncology is to de-escalate adjuvant therapy for patients with ER-positive/HER2-negative tumours with few or no lymph node metastases. This comprises the absolutely largest group of breast cancer patients. In a population-based register study of pathology data with over 45 000 patients in Sweden,[11] >60% of the ER+ HER2− patients were reported as NHG2. This clearly shows that methods to identify high- and low-risk patients in this patient group could contribute to clinical decision making of adjuvant chemotherapy. Over the past decade, gene expression profiling assays have become available in clinical practice for stratification of the same patient group into a low- and high-risk group, to supplement therapeutic decision making.[43] The majority of tumours considered for gene expression profiling in routine healthcare are NHG 2. Gene-expression profiling takes 1-2 weeks in the clinical setting and is associated with significant costs.

The DeepGrade model provides prognostic stratification for NHG 2 patients, and related subgroups, in a relatively short time.

Timewise, the proposed workflow required, on average, 1.24 and 16.4 min to predict tumour and carry out risk stratification on one GPU card, respectively. Significant gain in speed can be achieved by parallelising the computation with multiple GPUs. The direct compute costs are estimated to be <€1 in a public cloud infrastructure. Hence, it offers a potentially cost-effective alternative to gene expression profiling. Although the process can be considered time-consuming compared with pathologists, we believe that the additional stratification provides pathological information that is needed for clinical decision making in general, and for de-escalation of chemotherapy in particular.

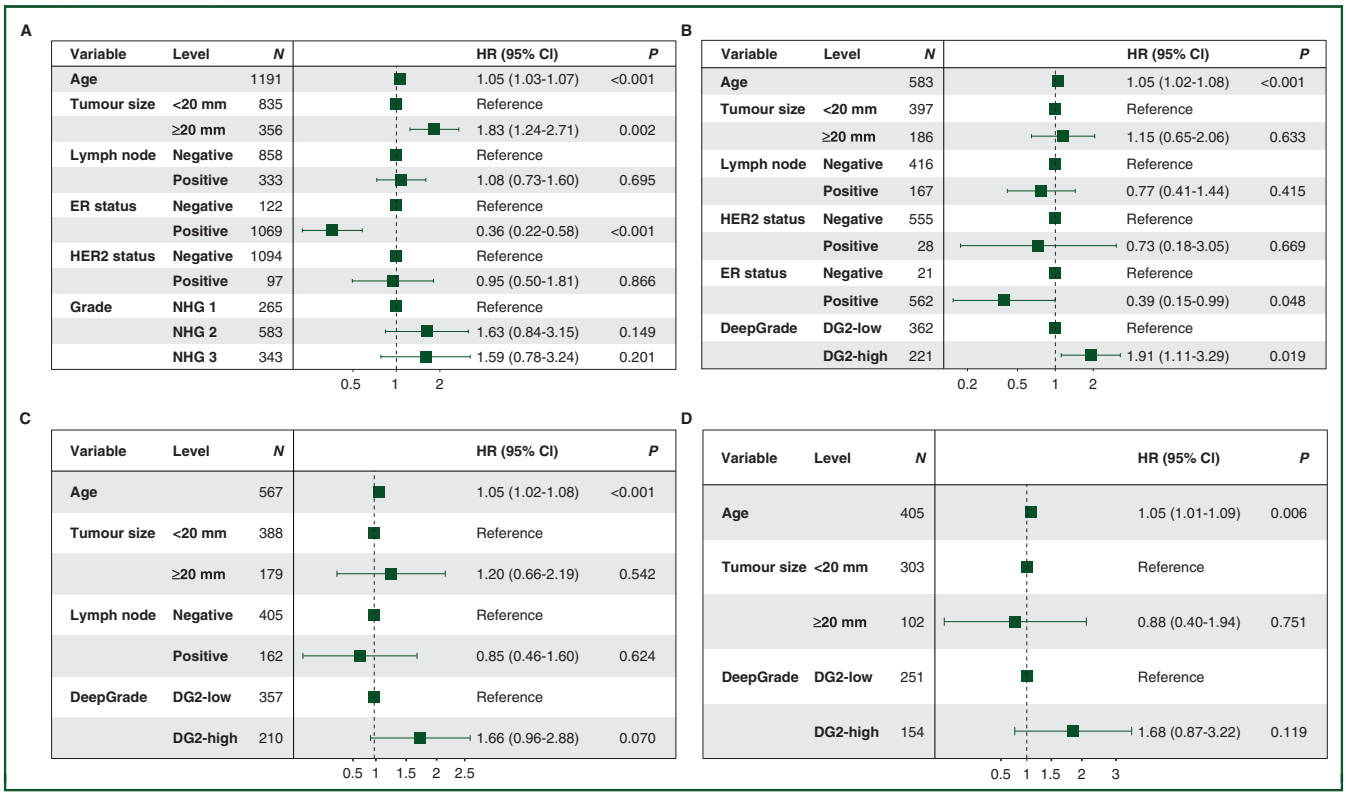

Figure 6. Forest plots from multivariable Cox proportional hazard regression analysis in the external test set.
(A) Stratification of all patients by routine NHG 1-3. (B) Stratification of NHG 2 cases by DeepGrade. (C) Stratification of patients in the NHG 2 and ER-positive subgroups by DeepGrade. (D) Stratification of patients in the NHG 2, ER-positive and node-negative subgroups by DeepGrade. DG, DeepGrade; ER, estrogen receptor; NHG, Nottingham histological grade.

This study has limitations. The study is based on retrospective materials because prospective studies are challenging due to the need for long follow-up times. Ductal carcinoma *in situ* annotations in slide images were not available, which could cause a conservative underestimation of the prognostic performance of DeepGrade. Only ClinSeq-BC has pathologist annotation of invasive cancer areas, which were used to exclude benign regions in subsequent analysis. An invasive detection model was optimised based on the annotated ClinSeq-BC study and applied to WSIs in all other studies to outline tumour regions. Hence, it is not impossible that the DeepGrade model could achieve better performance if we would perform manual annotations of cancer regions on all WSIs, as these are expected to be more precise. The DeepGrade model performance has not been possible to test systematically across different types of slide scanners, although images from multiple scanners were included in the study.

In conclusion, improved stratification of intermediate risk breast cancer patients has the potential to reduce overtreatment and undertreatment by adjuvant chemotherapy. The DeepGrade model provides independent prognostic stratification for NHG 2, and offers a potential cost-effective alternative to gene expression profiling, which could increase access to pathological information needed for clinical decision making in general, and for de-escalation of chemotherapy in particular.

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 acknowledge help and support from Dr Johan Vallon-Christersson at Lund University (Sweden) with preparation of clinicopathological information for the SCAN-B study. We thank the Hungarian Society of Senology, for supporting Balazs Acs.

FUNDING

This work was supported by funding from the Swedish Research Council (no grant number), Swedish Cancer Society (no grant number), Karolinska Institutet (no grant number), ERA PerMed (ERAPERMED2019-224-ABCAP), ERC (ERC2015CoG 682 810), MedTechLabs (no grant number), Swedish e-science Research Centre (SeRC) - eCPC, Stockholm Region (no grant number), Stockholm Cancer Society (no grant number) and Swedish Breast Cancer Association (no grant number).

DISCLOSURE

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. YW has received personal fees from Stratipath AB outside the submitted work. All other authors have declared no conflicts of interest.

REFERENCES

1. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat*. 1992;22(3):207-219.
2. Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. *Br J Cancer.* 1982;45(3):361-366.
3. Sundquist M, Thorstenson S, Brudin L, Nordenskjöld B. Applying the Nottingham Prognostic Index to a Swedish breast cancer population. South East Swedish Breast Cancer Study Group. *Breast Cancer Res Treat*. 1999;53(1):1-8.
4. Early Breast Cancer Trialists' Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. *Lancet*. 2000;355(9217):1757-1770.
5. Ellis IO, Coleman D, Wells C, et al. Impact of a national external quality assessment scheme for breast pathology in the UK. *J Clin Pathol*. 2006;59(2):138-145.
6. Dalton LW, Pinder SE, Elston CE, et al. Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Mod Pathol*. 2000;13(7):730-735.
7. Zhang R, Chen H-J, Wei B, 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):1976-1982.
8. Page DL. Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. *Mod Pathol*. 2007;18(1):67.
9. van Dooijeweert C, van Diest PJ, Willems SM, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: a nationwide study of 33,043 patients in the Netherlands. *Int J Cancer.* 2020;146(3):769-780.
10. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom—Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Mod Pathol*. 2005;18(8):1067-1078.
11. Acs B, Fredriksson I, Rönnlund C, et al. Variability in breast cancer biomarker assessment and the effect on oncological treatment decisions: a nationwide 5-year population-based study. *Cancers*. 2021;13:1166.
12. Balslev I, Axelsson CK, Zedeler K, et al. The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). *Breast Cancer Res Treat*. 1994;32(3):281-290.
13. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*. 2006;98(4):262-272.
14. Curigliano G, Burstein HJ, Winer EP, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. *Ann Oncol*. 2019;30(7):1181.
15. Wang M, Klevebring D, Lindberg J, et al. Determining breast cancer histological grade from RNA-sequencing data. *Breast Cancer Res*. 2016;18(1):48.
16. Ivshina AV, George J, Senko O, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*. 2006;66(21):10292-10301.
17. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167.
18. Nielsen TO, Parker JS, Leung S, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin Cancer Res*. 2010;16(21):5222-5232.
19. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351(27):2817-2826.
20. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology.* 2017;70(1):134-145.
21. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. *J Intern Med*. 2020;288:62-81.
22. Bera K, Schalper KA, Rimm DL, et al. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715.
23. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559-1567.
24. Ström P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol*. 2020;21:222-232.
25. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21:233-241.
26. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309.
27. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-2210.
28. Couture HD, Williams LA, Geradts J, 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.
29. Rantalainen M, Klevebring D, Lindberg J, et al. Sequencing-based breast cancer diagnostics as an alternative to routine biomarkers. *Sci Rep*. 2016;6:38037.
30. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61-70.
31. Vallon-Christersson J, Häkkinen J, Hegardt C, et al. Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. *Sci Rep*. 2019;9:12184.
32. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. *arXiv* 1512.00567v3;2015.
33. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211-252.
34. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
35. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32-35.
36. Early Breast Cancer Trialists' Collaborative Group. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. *Lancet*. 2011;378(9804):1707-1716.
37. Dubsky P, Filipits M, Jakesz R, et al. EndoPredict improves the prognostic classification derived from common clinical guidelines in ER-positive, HER2-negative early breast cancer. *Ann Oncol*. 2013;24(3):640-647.
38. Healey MA, Hirko KA, Beck AH, et al. Assessment of Ki67 expression for breast cancer subtype classification and prognosis in the Nurses' Health Study. *Breast Cancer Res Treat*. 2017;166(2):613-622.
39. Rhee J, Han S-W, Oh D-Y, et al. The clinicopathologic characteristics and prognostic significance of triple-negativity in node-negative breast cancer. *BMC Cancer.* 2008;8:307.
40. Rakha EA, Reis-Filho JS, Ellis IO. Basal-like breast cancer: a critical review. *J Clin Oncol*. 2008;26(15):2568-2581.
41. Eroles P, Bosch A, Pérez-Fidalgo JA, Lluch A. Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer Treat Rev.* 2012;38(6):698-707.
42. Aebi S, Davidson T, Gruber G, Castiglione M. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2010;21:v9-v14.
43. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. *Nat Rev Clin Oncol*. 2017;14(10):595-610.