End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study

Xiaofeng Jiang, Michael Hoffmeister, Hermann Brenner, Hannah Sophie Muti, Tanwei Yuan, Sebastian Foersch, Nicholas P West, Alexander Brobeil, Jitendra Jonnagaddala, Nicholas Hawkins, Robyn L Ward, Titus J Brinker, Oliver Lester Saldanha, Jia Ke, Wolfram Müller, Heike I Grabsch, Philip Quirke, Daniel Truhn, Jakob Nikolas Kather

Summary

Background Precise prognosis prediction in patients with colorectal cancer (ie, forecasting survival) is pivotal for individualised treatment and care. Histopathological tissue slides of colorectal cancer specimens contain rich prognostically relevant information. However, existing studies do not have multicentre external validation with real-world sample processing protocols, and algorithms are not yet widely used in clinical routine.

Methods In this retrospective, multicentre study, we collected tissue samples from four groups of patients with resected colorectal cancer from Australia, Germany, and the USA. We developed and externally validated a deep learning-based prognostic-stratification system for automatic prediction of overall and cancer-specific survival in patients with resected colorectal cancer. We used the model-predicted risk scores to stratify patients into different risk groups and compared survival outcomes between these groups. Additionally, we evaluated the prognostic value of these risk groups after adjusting for established prognostic variables.

Findings We trained and validated our model on a total of 4428 patients. We found that patients could be divided into high-risk and low-risk groups on the basis of the deep learning-based risk score. On the internal test set, the group with a high-risk score had a worse prognosis than the group with a low-risk score, as reflected by a hazard ratio (HR) of 4·50 (95% CI 3·33–6·09) for overall survival and 8·35 (5·06–13·78) for disease-specific survival (DSS). We found consistent performance across three large external test sets. In a test set of 1395 patients, the high-risk group had a lower DSS than the low-risk group, with an HR of 3·08 (2·44–3·89). In two additional test sets, the HRs for DSS were 2·23 (1·23–4·04) and 3·07 (1·78–5·3). We showed that the prognostic value of the deep learning-based risk score is independent of established clinical risk factors.

Interpretation Our findings indicate that attention-based self-supervised deep learning can robustly offer a prognosis on clinical outcomes in patients with colorectal cancer, generalising across different populations and serving as a potentially new prognostic tool in clinical decision making for colorectal cancer management. We release all source codes and trained models under an open-source licence, allowing other researchers to reuse and build upon our work.

Funding The German Federal Ministry of Health, the Max-Eder-Programme of German Cancer Aid, the German Federal Ministry of Education and Research, the German Academic Exchange Service, and the EU.

Lancet Digit Health 2024; 6: e33–43

Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany (X Jiang MMed, H S Muti MD, O L Saldanha MSc, Prof J N Kather MD); **Department of Medicine III** (X Jiang, O L Saldanha, Prof J N Kather) **and Department of Diagnostic and Interventional Radiology** (D Truhn MD), **University Hospital Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany; Division of Clinical Epidemiology and Ageing Research** (Prof M Hoffmeister PhD, Prof H Brenner MD, T Yuan MSc), **German Cancer Consortium** (Prof H Brenner), **and Digital Biomarkers for Oncology Group** (T J Brinker MD), **German Cancer Research Center, Heidelberg, Germany; Division of Preventive Oncology, German Cancer Research Center and National Center for Tumour Diseases, Heidelberg, Germany** (Prof H Brenner); **Institute of Pathology, University Medical Center Mainz, Mainz, Germany** (S Foersch MD); **Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK** (N P West PhD, Prof H I Grabsch PhD, Prof P Quirke PhD, Prof J N Kather); **Institute of Pathology** (A Brobeil MD) **and Tissue Bank** (A Brobeil) **and Medical Oncology** (Prof J N Kather), **National Center for Tumour Diseases, University Hospital Heidelberg, Heidelberg, Germany; School of Population Health** (J Jonnagaddala PhD) **and School of Medical Sciences** (Prof N Hawkins PhD, Prof R L Ward PhD), **Faculty of Medicine and Health, University of New South Wales Sydney, Kensington, NSW, Australia; Faculty of Medicine**

Introduction

Prediction of individual prognostic profiles is of exceptional importance for patients with colorectal cancer.[1] In particular, for patients with localised stage 2–3 colorectal cancer, accurate prognostication is important to decide whether they would benefit from adjuvant chemotherapy[2] and to establish the frequency of follow-up examinations after tumour resection. However, the current prognostic system, TNM staging, does not consider the large heterogeneity observed in histopathological tissue slides of colorectal cancer.

The histopathological phenotype of colorectal cancer contains a large amount of prognostically important information, including features such as tumour budding and lymphovascular infiltration, which are associated with prognosis.[3,4] Manual quantification of these features by pathologists often has the drawback of inter-observer and intra-observer variability.[4] Several studies have proposed machine learning systems to automate quantification; for example, automatic methods to count immune cells in pathology slides are widely used.[5,6] One of these methods is the immunoscore, in which an image analysis algorithm is used to count lymphocytes in tumour tissue.[7] The original immunoscore has been recreated in an open-source implementation[8] and improved with a more general artificial intelligence method[9] by more recent studies.

However, the focus on a predefined set of morphological structures of interest, such as lymphocytes, is a limitation of these approaches. A more general solution is to

and Health, The University of Sydney, Camperdown, NSW, Australia (Prof R L Ward); Department of General Surgery (Colorectal Surgery), Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, and Biomedical Innovation Center, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China (J Ke MD); Gemeinschaftspraxis Pathologie, Starnberg, Germany (W Müller MD); Department of Pathology, GROW School for Oncology and Reproduction, Maastricht University Medical Center, Maastricht, Netherlands (Prof H I Grabsch); Department of Medicine I, University Hospital Dresden, Dresden, Germany (Prof J N Kather)

Correspondence to: Prof Jakob Nikolas Kather, Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany
jakob-nikolas.kather@alumni.dkfz.de

Research in context

Evidence before this study

Deep learning can extract prognostic markers from routine pathology slides of colorectal cancer tissue stained with haematoxylin and eosin. Previous proof-of-concept studies did not use the latest deep learning technology, have not been validated in large-scale cohorts processed at multiple centres, and are closed source. We searched PubMed on Jan 13, 2023, without language or date restrictions from database inception using the user query ("survival" OR "prognosis" OR "prognostication" OR "risk stratification" OR "prediction model" OR "decision support") AND "(colorectal OR colon OR rectal)" AND "(cancer OR carcinoma)" AND ("deep learning" OR "artificial intelligence"). We systematically reviewed the 285 search results and identified 26 original research studies that applied deep learning using histopathology images. Of these 26 studies, 17 used established histopathological features that are known to be prognostically relevant to indirectly predict patient outcomes. Only nine studies applied an end-to-end approach to directly predict outcomes, with two using tissue microarray and six requiring tumour segmentation. Only one study predicted the prognosis of colorectal cancer directly from the whole-slide image (WSI); however, the authors have not made their source code or model open source.

Added value of this study

In this study, we developed a survival prediction model on the basis of two of the latest and most robust deep-learning methods, which comprised self-supervised learning and attention-based multiple-instance learning. We built a computational pipeline which can directly predict the prognosis of colorectal cancer from WSI, without using any manual intermediary steps. We trained and validated the model on 4428 patients from across the world, of whom 2157 patients were used for external validation. Our results show that the model demonstrates good generalisability and can be applied to diverse studies prepared and digitised by different institutions. The model can stratify patients into different risk groups with better performance than most available markers. Furthermore, we explored the pathobiological mechanisms associated with the predicted risk scores, providing biological interpretability of the predicted results. To promote the reproducibility and dissemination of our research, we have released all the source code and trained models under open-source licence, allowing other researchers to reuse and build upon our work.

Implications of all the available evidence

The deep learning-based risk score extracted from WSIs of colorectal cancer can complement existing clinical prognostic factors and could be expected to identify patients who benefit from adjuvant therapy, thus aiding patient management and clinical decision making.

directly use deep learning to analyse histopathology image data, without manually predefining structures of interest.[10,11] The main advantage of deep learning-based assessment of pathology slides compared with human experts is that the deep learning systems are not constrained to predefined image features a priori. They can assess any histopathological pattern, place it in context with other coexisting patterns, and derive a risk score. Several studies have used this approach, including Wulczyn and colleagues,[12] who used an in-house database to train a survival prediction model, and the DoMore Diagnostics team, who have shown that deep-learning algorithms could outperform current risk-stratification systems.[13,14]

Nevertheless, these studies have several important limitations that might affect their performance and applicability in clinical settings. First, the external validation of the algorithms used in these studies is limited, raising concerns about their generalisability and reliability. Second, the closed-source nature of the employed algorithms makes it challenging for other researchers to examine or modify them. Third, the focus on tumour regions as model inputs might overlook the crucial interactions between normal tissue and the tumour. Lastly, not all of these studies use the latest methodological advances, such as attention-based multiple-instance learning (attMIL) and self-supervised learning (SSL), which outperform standard deep-learning pipelines in many computational pathology applications.[15–17]

To address these limitations, we developed an open-source deep-learning survival prediction model using attMIL and SSL. We externally validated this model in three large international cohorts. Additionally, we used The Cancer Genome Atlas Colorectal Cancer (TCGA-CRC) dataset to explore the pathobiological mechanisms associated with the predicted risk scores, which aimed to provide a biological interpretability of the predictions.

Methods

Ethics statement

This retrospective study adhered to the Declaration of Helsinki[18] and we obtained ethical approval from each contributing centre (appendix p 4). Data collection and analysis were done in an anonymised manner. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (checklist in appendix p 29).[19] Because of the retrospective nature of the study and the use of anonymised data, the ethics board at the Medical Faculty of TU Dresden waived the requirement for a formal ethics vote.

See Online for appendix

Patient cohorts

In this retrospective, multicentre study, we used anonymised haematoxylin and eosin-stained colorectal adenocarcinoma slides from four large cohorts in Australia, Germany, and the USA for model training and validation. Digitised tumour-bearing tissue slides from the Darmkrebs: Chancen der Verhütung durch Screening (DACHS) study,[20,21] a large population-based case-control and patient cohort study on colorectal cancer from southwestern Germany, were used for the training of the network.

Three cohorts were used as independent external validation cohorts. First, the Molecular and Cellular Oncology study (MCO; Australia, n=1395),[22,23] a prospective study on more than 1500 participants undergoing curative resection for colorectal cancer from 1994 to 2010, in which clinical and pathological data were collected for all patients. Overall survival was followed up once per year for up to 5 years. Second, TCGA (USA, n=565) public repository, which includes colorectal cancer tissue samples of all stages with the primary intent of genomic characterisation. Third, the Marien-Hospital in Düsseldorf, Germany cohort (DUSSEL, n=197), a case series of colorectal cancer specimens resected with curative intent and collected at the Marien-Hospital in Düsseldorf, Germany, between January, 1990 and December, 1995, with the intent of doing research studies.[24]

The DACHS dataset contains the outcomes of overall survival, disease-free survival (DFS), and disease-specific survival (DSS). Overall survival and DSS were available for TCGA[25] and MCO. The outcomes of DFS and DSS were available for DUSSEL. Inclusion criteria were patients with colorectal cancer who underwent surgical resection in each cohort, and those who did not have complete pathological examination and follow-up information were excluded. Detailed inclusion and exclusion criteria can be found in the appendix (p 12).

Experimental design and statistics

The model generated patient-level risk scores, and C indexes and the areas under the time-dependent receiver operating characteristic curves (AUC) were calculated to assess the ability of the model to predict risk for overall survival, DFS, and DSS. Overall survival represented the time between surgery and death from any cause or the date of the last follow-up. DFS indicated the interval between surgery and recurrence, metastasis, death from any cause, or last follow-up. DSS denoted the time from surgery to death from colorectal cancer or last follow-up.

Patients were stratified into a high-risk group (higher than or equal to the threshold) and a low-risk group (lower than the threshold) by using the median deep learning-based risk score of the training set as a threshold. We then did the following analysis on the MCO cohort: first, we investigated whether the deep learning-based risk score could provide additional prognostic value in patients with stage 2 colorectal cancer, a group in which the availability of adjuvant chemotherapy remains controversial. A new triple stratification was created by merging deep-learning risk stratification with T stage, classifying patients with T3 and deep learning-predicted low risk as having a low-risk prognosis, those with T4 and deep learning-predicted high risk as having a high-risk prognosis, and the remaining patients as having a medium-risk prognosis. In addition, to make our analysis comparable with a study in 2022,[14] we adopted the protocol used in previous studies, dividing patients into three groups, those who had low risk, those who had medium risk, and those who had high risk, on the basis of the 25% and 75% cutoffs of the risk score of the training set. A Kaplan-Meier analysis and log-rank test were used to compare the survival differences between the groups. A Cox proportional hazard model was used on the basis of this grouping.

Statistical analyses were done using R version 3.4.0. The survcomp package was used to calculate the C index and its 95% CI. The C index, or concordance index, ranges from 0 to 1, and assesses the ability of prognostic models to correctly rank patient survival times. The timeROC package was used for the time-dependent receiver-operating characteristic analysis. The Kaplan-Meier method and the log-rank test were done with survival and survminer packages. Multivariable analyses were done using a Cox proportional hazard model of the survival package, in which the multivariable analysis was adjusted for confounding factors, including age, sex, and pathological T stage, N stage, and M stage. All statistical tests were two-sided and $p<0·05$ was considered to indicate a statistically significant result.

Image preprocessing and deep-learning procedures

All glass slides were scanned with Leica Aperio scanners at their respective institution. The digital images in this study were preprocessed under the Aachen deep-learning histopathology protocol.[26] All whole-slide images (WSIs) were segmented into image tiles with an edge length of 256 µm and saved with 224×224 pixels, yielding an effective magnification of 1·14 µm per pixel. During the process, all tiles with an average number of Canny edges lower than a threshold of 2 (ie, image tiles containing background or blur) were removed from the dataset. All image tiles were subsequently colour normalised using the Macenko method.[27]

We used our open-source pipeline, Marugoto, to train and validate deep-learning models. This pipeline employs a self-supervised pretrained histology-specific encoder, RetCCL, which translates each image tile into a 2048-dimensional feature vector.[17] All of the feature vectors obtained from the tiles of each slide image were then combined into a bag for patient-level risk-score prediction using attMIL.[15,28] This process involves projecting the feature vectors onto a length-256 feature space through a linear encoder and applying an attention module to compute an attention score for each tile. The bag-level feature vector was calculated by weighting the feature

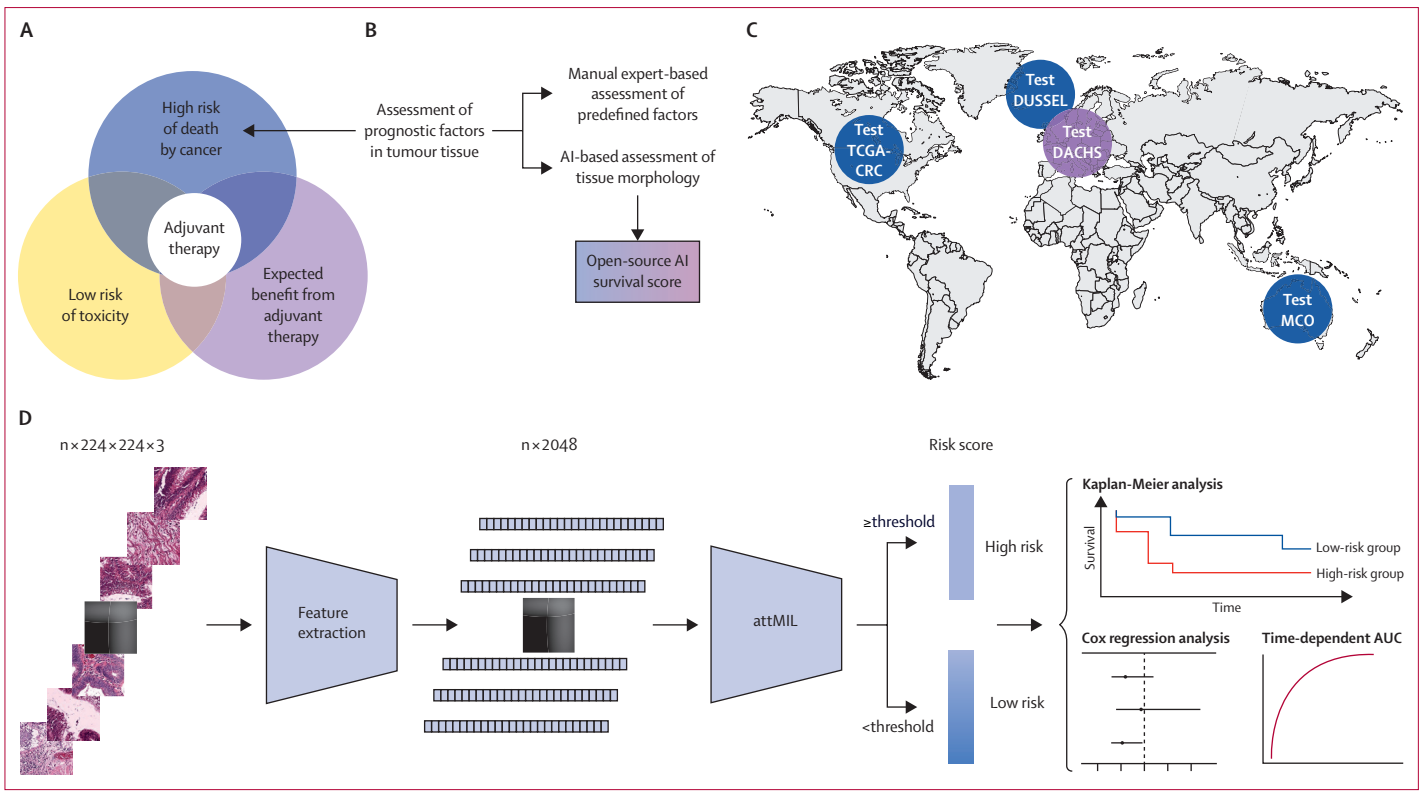

Figure 1: **Clinical need and outline of the study**
(A) There is a clinical need to extract prognostic factors from tumour tissue to assist clinical decision making. (B) Established methods are based on manual extraction of predefined prognostic features; an open-source end-to-end AI-based survival score would help with clinical roll-out and validation. (C) The DACHS cohort from Germany was used for training and internal testing. The MCO cohort from Australia, the TCGA-CRC cohort from the USA, and the DUSSEL cohort from Germany were used for external validation. (D) All tiles from one whole-slide image were processed through a self-supervised learning network to extract features. These features were then used by an attMIL network to generate patient-level risk scores. The median risk score of the training set was used as the threshold to classify patients into high-risk and low-risk groups, which were subsequently used for survival analysis. AI=artificial intelligence. attMIL=attention-based multiple-instance learning. AUC=area under the curve. DACHS=Darmkrebs: Chancen der Verhütung durch Screening study. DUSSEL=The Marien-Hospital in Duesseldorf, Germany. MCO=Molecular and Cellular Oncology study. TCGA-CRC=The Cancer Genome Atlas Colorectal Cancer study.

projection with attention scores, and the risk score output was generated by an additional fully connected layer.

In each training epoch, a set of 512 tiles was randomly drawn from the WSI. Because the training runs for 50 epochs, effectively all tiles from the WSI are seen by the network during training. For deployment, we used all of the tiles in the slides and set the batch size to 1 to handle different bag sizes. All analysis was done at the patient level, where the tiles from all slides of a patient were combined into a single bag for use in subsequent training or validation processes. We randomly divided the DACHS cohort into a training set, validation set, and test set in a 4:4:2 ratio on the patient level. The best model checkpoint based on the validation set was saved and then validated on the test set and on three external validation sets. Further training details are available in the appendix (p 1).

Visualisation and explainability

To better understand the predictive patterns of the risk score and the model internals, we generated WSI heatmaps showing the spatial distribution of attention and predicted scores and the tiles with the highest attention-weighted predicted score (top tiles). To gain insights into the biological basis of the predictions of the model, we generated attention heatmaps, weight-score heatmaps, and top tiles for six representative patients with the highest deep learning-based risk scores in the MCO cohort. The attention heatmap offers a visualisation of the regions of the image that the model attended to when making the predictions, whereas the weight-score heatmap shows the relative importance of different regions within the image. The tiles with the highest weight scores indicate the specific areas of the image that had the greatest effect on the prognostication of the model. To explore the biological characteristics of the risk score, we analysed the RNA-sequence data from the TCGA cohort. Differentially expressed genes (DEGs) between the two risk groups were identified using the edgeR package in R. Gene Set Enrichment Analysis (GSEA) was done for these DEGs. To further understand the correlation between risk score and immune infiltration, the CIBERSORT algorithm and ssGSEA method were used to calculate the fractions of tumour-infiltrating immune cells (appendix p 2).

	DACHS training set (n=908)	DACHS validation set (n=908)	DACHS test set (n=455)	MCO (n=1395)	DUSSEL (n=197)	TCGA-CRC (n=565)
Number of slides per patient						
One	889 (98%)	901 (99%)	450 (99%)	1324 (95%)	197 (100%)	557 (99%)
Two	9 (1%)	7 (1%)	4 (1%)	63 (5%)	..	7 (1%)
Three	..	..	1 (<1%)	6 (<1%)	..	1 (<1%)
Four	..	..	..	2 (<1%)	..	..
Age						
≤65 years	300 (33%)	342 (38%)	167 (37%)	546 (39%)	65 (33%)	186 (33%)
>65 years	608 (67%)	566 (62%)	288 (63%)	849 (61%)	132 (67%)	225 (40%)
Missing	..	..	..	..	..	154 (27%)
Sex						
Male	523 (58%)	533 (59%)	276 (61%)	767 (55%)	113 (57%)	207 (37%)
Female	385 (42%)	375 (41%)	179 (39%)	628 (45%)	84 (43%)	205 (36%)
Missing	..	..	..	..	..	153 (27%)
Location						
Left	256 (28%)	253 (28%)	116 (25%)	343 (25%)	65 (33%)	164 (29%)
Right	314 (35%)	303 (33%)	172 (38%)	521 (37%)	64 (32%)	189 (33%)
Rectum	337 (37%)	350 (39%)	167 (37%)	527 (38%)	68 (35%)	57 (10%)
Missing	1 (<1%)	2 (<1%)	..	4 (<1%)	..	155 (27%)
Stage						
1	172 (19%)	160 (18%)	82 (18%)	269 (19%)	46 (23%)	69 (12%)
2	322 (35%)	303 (33%)	150 (33%)	499 (36%)	78 (40%)	142 (25%)
3	294 (32%)	309 (34%)	157 (35%)	451 (32%)	70 (36%)	133 (24%)
4	120 (13%)	135 (15%)	66 (15%)	176 (13%)	3 (2%)	56 (10%)
Missing	..	1 (<1%)	..	..	..	165 (29%)
Pathological T stage						
T1	57 (6%)	43 (5%)	27 (6%)	110 (8%)	20 (10%)	13 (2%)
T2	150 (17%)	160 (18%)	77 (13%)	232 (17%)	34 (17%)	72 (13%)
T3	569 (63%)	570 (63%)	279 (47%)	726 (52%)	125 (63%)	282 (50%)
T4	109 (12%)	116 (13%)	62 (10%)	327 (23%)	18 (9%)	43 (8%)
Missing	23 (3%)	19 (2%)	10 (2%)	..	..	155 (27%)
Pathological N stage						
N0	499 (55%)	473 (52%)	229 (50%)	796 (57%)	126 (64%)	224 (40%)
N+	386 (43%)	419 (46%)	210 (46%)	599 (43%)	71 (36%)	185 (33%)
Missing	23 (3%)	16 (2%)	16 (4%)	..	..	156 (28%)
Metastasis						
M0	788 (87%)	772 (85%)	389 (85%)	1219 (87%)	194 (98%)	344 (61%)
M+	120 (13%)	135 (15%)	66 (15%)	176 (13%)	3 (2%)	56 (10%)
Missing	..	1 (<1%)	..	..	..	165 (29%)
Microsatellite status						
Microsatellite instability	81 (9%)	80 (9%)	46 (10%)	205 (15%)	30 (15%)	58 (10%)
Microsatellite stability	723 (80%)	733 (81%)	351 (77%)	1183 (85%)	166 (84%)	352 (62%)
Missing	104 (11%)	95 (10%)	58 (13%)	7 (1%)	1 (1%)	155 (27%)

Data are n (%). DACHS=Darmkrebs: Chancen der Verhütung durch Screening study. DUSSEL=The Marien-Hospital in Duesseldorf, Germany. MCO=Molecular and Cellular Oncology study. TCGA-CRC=The Cancer Genome Atlas Colorectal Cancer study.

Table: **Patient characteristics in the training, validation, internal test, and external test sets**

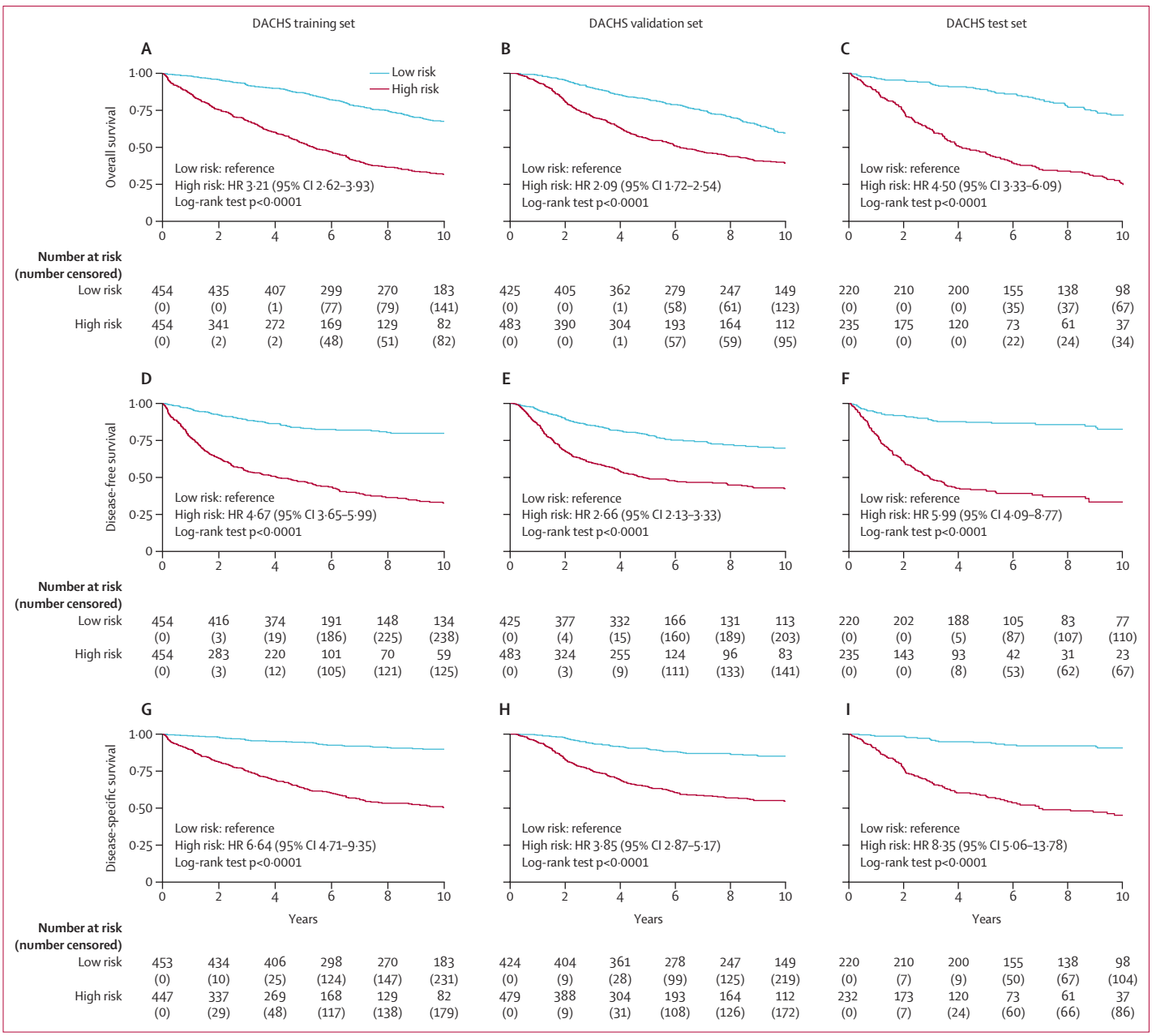

Figure 2: **Kaplan-Meier analysis by deep learning-based risk score in the DACHS cohort**
Patients were stratified by deep-learning model-based score as high risk (red line) or low risk (blue line). The median deep learning-based risk score of the training set was used as the cutoff (–0·135). Kaplan-Meier curves for overall survival are presented in the training set (A), validation set (B), and test set (C). Kaplan-Meier curves for disease-free survival are presented in the training set (D), validation set (E), and test set (F). (G-I) Kaplan-Meier curves for disease-specific survival are presented in the training set (G), validation set (H), and test set (I). DACHS=Darmkrebs: Chancen der Verhütung durch Screening study. DSS=disease-specific survival. HR=hazard ratio.

Role of the funding source
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results
We developed an end-to-end attMIL deep-learning model for pathology-based survival prediction (figure 1). We trained and validated our model on 4428 patients, of whom 2157 were used for external validation. The clinicopathological characteristics of the patients in each dataset have been summarised (table). We reported the differences in the distribution of clinical features between each dataset (appendix p 5). Median overall follow-up time was 6·3 years in the DACHS cohort (IQR 3·7; follow-up time, 10·1 years), 4·9 years in the MCO cohort

(IQR 2·8; follow-up time, 4·9 years), 1·8 years in the TCGA cohort (IQR 1·1; follow-up time, 3·0 years), and 3·9 years in the DUSSEL cohort (IQR 2·0; follow-up time, 6·2 years). The model accurately predicted survival from unannotated histology slides with high performance, with C-indexes of 0·76 (95% CI 0·69–0·82) for predicting overall survival, 0·82 (0·73–0·88) for predicting DFS, and 0·78 (0·70–0·84) for predicting DSS in the internal test set. For non-metastatic cases, the deep learning-based risk score exhibited consistent C-indexes of 0·74 (95% CI 0·65–0·81) for overall survival, 0·80 (0·68–0·88) for DFS, and 0·77 (0·67–0·84) for DSS in the internal test set (appendix p 6). The AUCs for the internal test set were 0·79 (95% CI 0·71–0·86) for 1 year overall survival, 0·84 (0·80–0·88) for 3 year overall survival, and 0·83 (0·79–0·87) for 5 year overall survival (appendix p 13; the detailed performance of other metrics is in appendix p 7). Patients were divided into high-risk and low-risk groups by median deep learning-based risk score (–0·135) of the training cohort. Patients in the high-risk DACHS group had worse outcomes than patients in the low-risk group in the internal test set (figure 2). In summary, these data suggest that deep learning can be helpful to predict the outcome of patients with colorectal cancer from pathology slides.

We validated the prognostication performance in three large external validation cohorts. For the MCO cohort consisting of 1395 patients from Australia, the C-index for overall survival reached 0·65 (95% CI 0·60–0·70), and for DSS was 0·70 (0·64–0·75). Similarly, the C-index for overall survival was 0·64 (0·52–0·74) and for DSS was 0·69 (0·54–0·81) in the public TCGA-CRC cohort. Even in DUSSEL, a relatively small cohort, the model maintained a fair performance with a C-index of 0·62 (0·47–0·75; appendix p 6) for DSS. AUCs for each dataset at different survival times are shown (appendix p 13). In the Kaplan-Meier analysis, the deep-learning risk score demonstrated significant risk stratification in all external validation cohorts (figure 3). Together, these results show that the model reaches stable performance even with a relatively shorter follow-up and a smaller sample size in external validation sets.

We further did a multivariable Cox regression analysis to evaluate the prognostic value of the deep learning-based risk group with adjustment of established prognostic variables. The adjusted hazard ratios of the

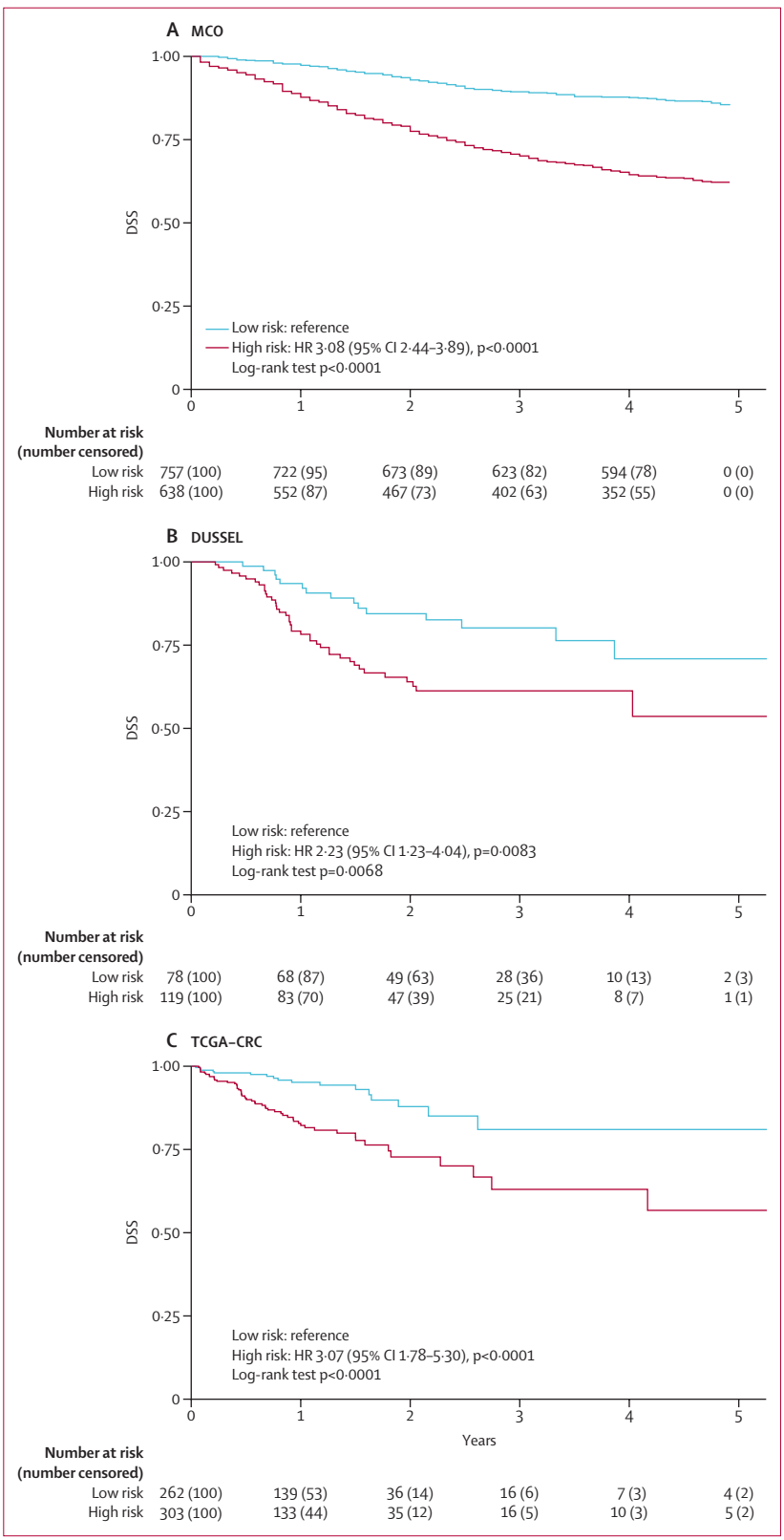

Figure 3: **Kaplan-Meier analysis by deep learning-based risk score in the external validation cohort**

Patients were stratified by deep learning model-based score as high risk (red line) or low risk (blue line). The median deep learning-based risk score of the training set was used as the cutoff (–0·135). (A) Kaplan-Meier curves for DSS in the MCO cohort. (B) Kaplan-Meier curves for DSS in the DUSSEL cohort. (C) Kaplan-Meier curves for DSS in the TCGA-CRC cohort. DSS=disease-specific survival. DUSSEL=the Marien-Hospital in Duesseldorf, Germany. HR=hazard ratio. MCO=Molecular and Cellular Oncology study. TCGA-CRC=The Cancer Genome Atlas Colorectal Cancer study.

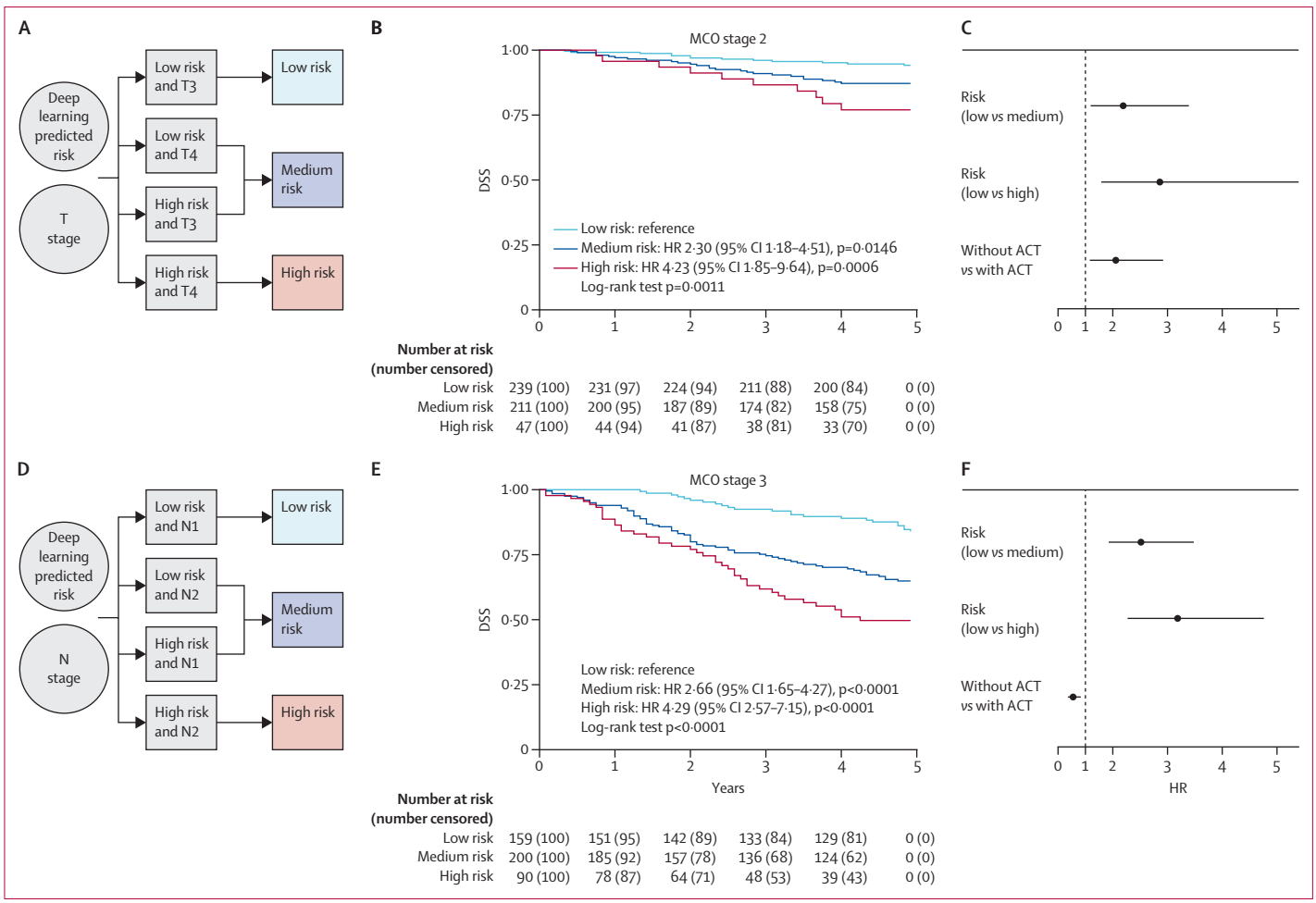

Figure 4: **Integrating deep-learning risk score with T stage and N stage for stages 2–3 risk stratification in the MCO cohort**
(A) Patients with stage 2 colorectal cancer were stratified into low-risk, medium-risk, and high-risk groups on the basis of a combination of deep learning predicted risk and T stage for risk assessment. (B) Kaplan-Meier curves in patients stratified into low-risk, medium-risk, and high-risk groups on the basis of a combination of deep learning-predicted risk and T stage. (C) Multivariate Cox regression analysis between deep learning-predicted risk and T stage regrouping and adjuvant chemotherapy. (D) Patients with stage 3 colorectal cancer were stratified into low-risk, medium-risk, and high-risk groups on the basis of a combination of deep learning-predicted risk and N stage for risk assessment. (E) Kaplan-Meier curves in patients stratified into low-risk, medium-risk, and high-risk groups on the basis of a combination of deep learning-predicted risk and N stage. (F) Multivariate Cox regression analysis between deep learning-predicted risk and N stage regrouping and adjuvant therapy. DSS=disease-specific survival. HR=hazard ratio. MCO=Molecular and Cellular Oncology study.

deep learning-based risk group for DSS were 1·40 (95% CI 1·09–1·81; p=0·0091) in the MCO cohort and 2·30 (1·21–4·35; p=0·011) in the DUSSEL cohort, and were significantly higher or similar to those of the pathological T stage or N stage (appendix p 14). The deep learning-based risk score remained robust in all cohorts, except in the TCGA-CRC cohort, in which the results of the multivariable analysis were not significant for either in the overall survival or DSS.

In the MCO dataset, we further validated whether the deep learning-based risk score was able to identify patients with an unfavourable prognosis and with stage 2–3 colorectal cancer. Kaplan-Meier analysis showed that the restratified high-risk group had a worse prognosis than the restratified low-risk group (figure 4B). The performance of the new triple stratification compared with the deep-learning risk stratification and T stage and corresponding Kaplan-Meier curves are shown in the appendix (p 15). Multivariable Cox analysis showed that the new stratification was a predictor of prognosis independent of adjuvant chemotherapy (figure 4C). The same analysis was applied to patients with stage 3 colorectal cancer, with a focus on N stage (figure 4D). Similar results were observed, with patients reclassified as being in the high-risk group exhibiting poorer prognosis (figure 4E), and the new stratification being an independent predictor of prognosis (figure 4F). The Kaplan-Meier curves demonstrated superior performance of the three-group classification across all cohorts, except for the DUSSEL cohort (appendix pp 16–17). However, in the multivariable Cox analysis results, which incorporated other clinical factors, the deep-learning risk stratification exhibited more consistent and reliable performance across both grouping approaches (appendix p 8).

Analyses based on other clinical subgroups are shown in the appendix (pp 18–24). In conclusion, our deep-learning risk stratification reached a high performance across several clinical subgroups and in combination with existing clinical stratification provides a more accurate prognostic stratification.

These heatmaps were generated from WSIs that were not annotated with tumour regions. For the high-risk group (appendix p 25), the heatmaps showed that the model was able to focus on tumour regions, with the top tiles comprising mainly poorly differentiated tumours and fat-infiltrated tumour regions.

By performing a bioinformatics analysis on the TCGA-CRC cohort, we identified 113 DEGs between the high-risk and low-risk groups (appendix p 26; the details of the DEGs are listed in the appendix p 9). GSEA revealed that the DEGs were significantly enriched in seven pathways, including hallmark oxidative phosphorylation, hallmark E2F targets, hallmark myogenesis, and hallmark epithelial mesenchymal transition (appendix p 26). Immune cell type-specific analysis revealed a higher proportion of CD8 T cells (p=0·022), CD4 memory T cells (p=0·025), and M1 macrophages (p=0·027), and a lower proportion of M0 macrophages in the low-risk group (p=0·0057; appendix p 26). The calculated scores of immune cells based on ssGSEA, shown as a heatmap, demonstrate a similar trend of immune infiltration (appendix p 26). The correlation between the deep learning-based risk score and the tumour immune-infiltrating cells is shown in the appendix (p 27). Despite differences in immune infiltration between high-risk and low-risk groups, there is no significant correlation between the risk score and immune cells. No statistically significant differences in other immune-related cells were observed between the low-risk and high-risk groups. These findings demonstrate a correlation between the predicted groups of the model and the previous prognostic cellular information and tissue information, providing model interpretability.

Discussion

In this study, we developed a deep learning-based prognostic-stratification system for automatic prediction of overall and tumour-specific survival in patients with resected colorectal cancer. Our SSL-based attMIL open-source analysis pipeline reached promising predictive performance on a large cohort from Germany and was consistently validated in other cross-international cohorts, as measured by the hazard ratio in Cox-proportional hazard models. In addition, deep learning-based risk score was a predictor independent of clinical features such as TNM staging and MSI.

In the development of the model, we used RetCCL,[17] an SSL pretrained model in histopathology, to extract feature vectors from image tiles. Compared with previous studies that used ImageNet-based pretraining models, the SSL weights within the domain help to extract intrinsic tissue features in greater depth. Previous studies have shown that SSL with attMIL achieves a good performance on classification tasks and the present study extends this evidence to survival prediction.[15–17] Our approach could therefore potentially provide a new paradigm for processing time-to-event data in computational pathology and potentially in other medical image modalities.

Our model has been validated to quantify several datasets and has exhibited consistent performance. However, in the multivariable Cox analysis, the binary risk stratification was not independent of TNM staging in the TCGA dataset. On the one hand, the short follow-up period of TCGA compared with other cohorts might not reveal the predictive efficacy of the model for long-term prognosis. In the TCGA cohort, the highest time-dependent AUCs were found within 2 years and then decreased over time, which is also consistent with its median follow-up time distribution. On the other hand, this finding might be related to the heterogeneity of sample preparation across centres; the challenges associated with standardisation still require attention.

Currently, there is controversy over whether adjuvant chemotherapy is necessary for patients with stage 2 colorectal cancer, and accurate outcome predictions might help inform clinical decision making. In this study, by combining deep learning-predicted risk with clinical staging, we were able to provide more precise stratification for this patient subgroup. For patients with stage 2 colorectal cancer classified as low risk, higher DSS rates suggest that follow-up could replace adjuvant therapy. By contrast, adjuvant therapy is already the standard treatment for patients with stage 3 colorectal cancer and for those identified as having high-risk colorectal cancer, more aggressive treatment strategies and closer monitoring could be beneficial.

Despite achieving excellent performance, the interpretability of deep-learning models still poses a challenge for their clinical application. To address this issue, we did a visualisation analysis of the regions of interest in our model. Our findings suggest that for patients with high-risk colorectal cancer, the model is more attentive to the adipose tissue surrounding the tumour, which is in line with previous research.[12] Furthermore, we did an in-depth exploration of the biological associations of our model by using bio-molecular information from the TCGA dataset. We observed significant differences in immune infiltration between the high-risk and low-risk groups. Specifically, the low-risk group exhibited a significantly higher degree of anti-tumour immune infiltration, particularly in CD8 T cells, CD4 memory T cells, and M1 macrophages. This enhanced anti-tumour immunity has been shown to resist tumour progression and improve prognosis.[29] Moreover, anti-tumour immunity suggests that patients in the low-risk group might be more responsive to immunotherapy.[30] In addition, the GSEA analysis revealed that DEGs between two groups

were enriched in pathways related to metabolism, immune dysregulation, and epithelial mesenchymal transition. This finding suggests that tumours in the high-risk group were more proliferative and aggressive. Furthermore, the enrichment results indicate that the cell cycle might have a crucial role in the prognosis of colon cancer, as demonstrated by the enriched gene sets for E2F targets, MYC targets, and G2M checkpoints. These results highlight the potential of targeting the cell cycle as a therapeutic strategy for the treatment of colorectal cancer.

Our study has several limitations. Although we validated the model in a cross-national cohort, all cases were collected retrospectively, which might introduce inherent bias and hidden confounders into the study. Specifically, compared with the DACHS cohort, the TCGA cohort had a shorter follow-up period and younger patients, which limits the assessment of disease progression and introduces age bias, affecting the evaluation of model performance. Additionally, because of the scarcity of the data, important factors such as tumour margins and postoperative complications, which substantially reduce survival time, were not included in the analysis, potentially resulting in an overestimation of the efficacy of the model. Therefore, a larger prospective cohort is needed to validate our results.

In conclusion, we developed and validated a deep learning-based model that uses attMIL and SSL to directly predict survival from WSIs in patients with colorectal cancer. The output risk score of the model is an independent prognostic factor that can serve as a tool for surgeons and oncologists for postoperative risk stratification of localised and advanced colorectal cancer. We have open-sourced our model, making it easy for other research groups to reuse, reproduce, or extend our approach. Like any biomarker, our proposed digital biomarker should be further evaluated in prospective clinical trials. Particularly, the predictive value as a decision aid for whether an individual patient should receive adjuvant chemotherapy on the basis of output of the model requires additional prospective clinical evidence.

Contributors

XJ and JNK conceived the study. XJ wrote the manuscript and did the data processing, experiment, and analysis. MH, HB, HSM, TY, SF, NPW, AB, JJ, NH, RLW, TJB, OLS, JK, WM, HIG, PQ, and DT contributed materials and clinical expertise. DT and JNK supervised the work. All authors contributed to the experimental design, the interpretation of the results, and editing of the final manuscript. All authors had full access to the data, and XJ, HSM, OLS, and JNK directly accessed and verified the data in the study. All authors accept the final responsibility to submit for publication and take responsibility for the contents of the manuscript.

Declaration of interests

JNK has received consulting fees from Owkin; DoMore Diagnostics; Panakeia; and Histofy; furthermore, JNK holds shares and holds a leadership role in StratifAI and has received honoraria for lectures by Bayer, Eisai, Merck Sharp & Dohme (MSD), Bristol-Myers Squibb (BMS), Roche, Pfizer, and Fresenius and has participated on a Data Safety Monitoring Board or Advisory Board for Bayer, Eisai, MSD, BMS, Roche, and Pfizer. PQ and NPW declare research funding from Roche and PQ consulting and speaker services for Roche. JJ declares consulting fees from WHO (Development of Digital Health Solution) and CARE International, Papua New Guinea (Development of Survey Instruments). SF declares grants and contracts from Bundesministerium für Bildung und Forschung, Deutsche Forschungsgemeinschaft, and German Cancer Aid, and payment or honoraria from BMS, MSD, European Society for Medical Oncology, and Deutsche Gesellschaft für Pathologie. DT holds shares in StratifAI. WM is a shareholder of Gemeinschaftspraxis Pathologie Starnberg, a private pathology practice in Germany. All other authors declare no competing interests.

Data sharing

The Molecular and Cellular Oncology study dataset is available through the Secure Research Environment for Digital Health Consortium (www.sredhconsortium.org) and was used with approvals. Data from The Cancer Genome Atlas Colorectal Cancer study are publicly available at https://portal.gdc.cancer.gov/. All codes are open source and available at https://github.com/KatherLab/marugoto and https://github.com/KatherLab/deepmed. Trained models are available at https://github.com/KatherLab/crc-models-2022. The remaining data were provided by the corresponding study Principal Investigators, and specific data sharing policies can be found in the corresponding original publications.

Acknowledgments

We extend our gratitude to the tissue bank of the National Center for Tumor Diseases at the Institute of Pathology at University Hospital Heidelberg, Germany for providing access to the biobank data. We also acknowledge the assistance of the SREDH Consortium's (www.sredhconsortium.org) Translational Cancer Bioinformatics working group in obtaining access to the Molecular and Cellular Oncology colorectal cancer dataset. JNK is supported by the German Federal Ministry of Health (Deep Liver, ZMVI1-2520DAT111), the Max-Eder-Programme of German Cancer Aid (grant 70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; Camino, 01EO2101; SWAG, 01KD2215A; Transform Liver, 031L0312A and Tangerine, 01KT2302 through ERA-NET Transcan), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (Transplant.KI, 01VSF21048) and the EU's Horizon Europe innovation programme (Odelia, 101057091; Genial, 101096312). JNK, PQ, and NPW are supported by the National Institute for Health and Care Research (NIHR; NIHR213331) Leeds Biomedical Research Centre. XJ is supported by the programme of the China Scholarships Council (202106380048). The Darmkrebs: Chancen der Verhütung durch Screening study (HB and MH) was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HO 5117/2-2, HE 5998/2-1, HE 5998/2-2, KL 2354/3-1, KL 2354/3-2, RO 2270/8-1, RO 2270/8-2, BR 1704/17-1, and BR 1704/17-2), the Interdisciplinary Research Program of the National Centre for Tumour Diseases (Germany), and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, and 01ER1505B). PQ and NPW are supported by Yorkshire Cancer Research Programme grants L386 (QUASAR series) and L394 (YCR BCIP series). PQ is a National Institute of Health Research senior investigator. JJ was funded by the Australian National Health and Medical Research Council (GNT1192469) and was also supported by Google through the 2022 research innovator and cloud research credits programme (GCP19980904) and the Research Technology Services at University of New South Wales Sydney, and NVIDIA Academic Hardware grant programmes.

References

1 Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin* 2020; **70:** 145–64.

2 Argilés G, Tabernero J, Labianca R, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2020; **31:** 1291–305.

3 Martin B, Schäfer E, Jakubowicz E, et al. Interobserver variability in the H&E-based assessment of tumor budding in pT3/4 colon cancer: does it affect the prognostic relevance? *Virchows Arch* 2018; **473:** 189–97.

4 Harris EI, Lewin DN, Wang HL, et al. Lymphovascular invasion in colorectal cancer: an interobserver variability study. *Am J Surg Pathol* 2008; **32:** 1816–21.

5 Pagès F, Mlecnik B, Marliot F, et al. International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* 2018; **391:** 2128–39.

6 Pai RK, Banerjee I, Shivji S, et al. Quantitative pathologic analysis of digitized images of colorectal carcinoma improves prediction of recurrence-free survival. *Gastroenterology* 2022; **163:** 1531–46.e8.

7 Galon J, Costes A, Sanchez-Cabo F, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 2006; **313:** 1960–64.

8 Alwers E, Kather JN, Kloor M, et al. Validation of the prognostic value of CD3 and CD8 cell densities analogous to the Immunoscore by stage and location of colorectal cancer: an independent patient cohort study. *J Pathol Clin Res* 2023; **9:** 129–36.

9 Foersch S, Glasner C, Woerl A-C, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med* 2023; **29:** 430–39

10 Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer* 2022; **3:** 1026–38.

11 Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology: new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019; **16:** 703–15.

12 Wulczyn E, Steiner DF, Moran M, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit Med* 2021; **4:** 71.

13 Skrede O-J, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* 2020; **395:** 350–60.

14 Kleppe A, Skrede O-J, De Raedt S, et al. A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol* 2022; **23:** 1221–32.

15 Niehues JM, Quirke P, West NP, et al. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: a retrospective multi-centric study. *Cell Rep Med* 2023; **1:** 100980.

16 Saldanha OL, Loeffler CML, Niehues JM, et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ Precis Oncol* 2023; **7:** 35.

17 Wang X, Du Y, Yang S, et al. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. *Med Image Anal* 2022; **83:** 102645.

18 World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013; **310:** 2191–94.

19 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg* 2015; **102:** 148–58.

20 Hoffmeister M, Jansen L, Rudolph A, et al. Statin use and survival after colorectal cancer: the importance of comprehensive confounder adjustment. *J Natl Cancer Inst* 2015; **107:** djv045.

21 Brenner H, Chang-Claude J, Seiler CM, Hoffmeister M. Long-term risk of colorectal cancer after negative colonoscopy. *J Clin Oncol* 2011; **29:** 3761–67.

22 Hawkins N. MCO study tumour collection. 2011. https://researchdata.edu.au/mco-study-tumour-collection/1957427 (accessed July 21, 2022).

23 Jonnagaddala J, Croucher JL, Jue TR, et al. Integration and analysis of heterogeneous colorectal cancer data for translational research. *Stud Health Technol Inform* 2016; **225:** 387–91.

24 Grabsch H, Dattani M, Barker L, et al. Expression of DNA double-strand break repair proteins ATM and BRCA1 predicts survival in colorectal cancer. *Clin Cancer Res* 2006; **12:** 1494–500.

25 Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 2018; **173:** 400–16.e11.

26 Muti HS, Loeffler C, Echle A, et al. The Aachen protocol for deep learning histopathology: a hands-on guide for data preprocessing. *Zenodo* 2020; published online March 3. DOI:10.5281/zenodo.3694994.

27 Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Boston, MA: IEEE, 2009: 1107–10.

28 Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. *arXiv* 2018; **80:** 2127–36 (preprint).

29 Pagès F, Berger A, Camus M, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer.- *N Engl J Med* 2005; **353:** 2654–66.

30 Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* 2012; **12:** 298–306.