Deep learning for prediction of colorectal cancer outcome: a discovery and validation study

Ole-Johan Skrede, Sepp De Raedt*, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, Inger Nina Farstad, Enric Domingo, David N Church, Arild Nesbakken, Neil A Shepherd, Ian Tomlinson, Rachel Kerr, Marco Novelli, David J Kerr, Håvard E Danielsen*

Lancet 2020; 395: 350–60

See **Comment** page 314

*Contributed equally

Institute for Cancer Genetics and Informatics (O-J Skrede MSc, S De Raedt PhD, A Kleppe PhD, T S Hveem PhD, Prof K Liestøl PhD, J Maddison PhD, H A Askautrud PhD, M Pradhan PhD, J A Nesheim MSc, Prof F Albregtsen MSc, Prof I Tomlinson PhD, Prof M Novelli PhD, Prof H E Danielsen PhD), **Department of Pathology, Division of Laboratory Medicine** (Prof I N Farstad PhD), **and Department of Gastrointestinal Surgery** (Prof A Nesbakken PhD) **Oslo University Hospital, Oslo, Norway; Department of Informatics** (O-J Skrede, S De Raedt, A Kleppe, Prof K Liestøl, Prof F Albregtsen, Prof H E Danielsen), **and Institute of Clinical Medicine** (Prof I N Farstad, Prof A Nesbakken), **University of Oslo, Oslo, Norway; Department of Oncology** (E Domingo PhD, Prof R Kerr PhD), **Wellcome Centre for Human Genetics** (D N Church DPhil), **and Nuffield Division of Clinical Laboratory Sciences** (Prof D J Kerr DSc, Prof H E Danielsen), **University of Oxford, Oxford, UK; National Institute of Health Research Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK** (D N Church); **KG Jebsen Colorectal Cancer Research Centre, Oslo, Norway** (Prof A Nesbakken); **Gloucestershire Cellular Pathology Laboratory, Cheltenham General Hospital, Cheltenham, UK** (Prof N A Shepherd DM); **Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, UK** (Prof I Tomlinson); **and Research**

Summary

Background Improved markers of prognosis are needed to stratify patients with early-stage colorectal cancer to refine selection of adjuvant therapy. The aim of the present study was to develop a biomarker of patient outcome after primary colorectal cancer resection by directly analysing scanned conventional haematoxylin and eosin stained sections using deep learning.

Methods More than 12 000 000 image tiles from patients with a distinctly good or poor disease outcome from four cohorts were used to train a total of ten convolutional neural networks, purpose-built for classifying supersized heterogeneous images. A prognostic biomarker integrating the ten networks was determined using patients with a non-distinct outcome. The marker was tested on 920 patients with slides prepared in the UK, and then independently validated according to a predefined protocol in 1122 patients treated with single-agent capecitabine using slides prepared in Norway. All cohorts included only patients with resectable tumours, and a formalin-fixed, paraffin-embedded tumour tissue block available for analysis. The primary outcome was cancer-specific survival.

Findings 828 patients from four cohorts had a distinct outcome and were used as a training cohort to obtain clear ground truth. 1645 patients had a non-distinct outcome and were used for tuning. The biomarker provided a hazard ratio for poor versus good prognosis of 3·84 (95% CI 2·72–5·43; p<0·0001) in the primary analysis of the validation cohort, and 3·04 (2·07–4·47; p<0·0001) after adjusting for established prognostic markers significant in univariable analyses of the same cohort, which were pN stage, pT stage, lymphatic invasion, and venous vascular invasion.

Interpretation A clinically useful prognostic marker was developed using deep learning allied to digital scanning of conventional haematoxylin and eosin stained tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, and gives consistent results across tumour and nodal stage. The biomarker stratified stage II and III patients into sufficiently distinct prognostic groups that potentially could be used to guide selection of adjuvant treatment by avoiding therapy in very low risk groups and identifying patients who would benefit from more intensive treatment regimes.

Funding The Research Council of Norway.

Introduction

Biomarkers are increasingly being used to match anticancer therapy to specific tumour genotypes, protein, and RNA expression profiles, usually in patients with advanced disease.[1–3] One example of this is selection of *KRAS*-wild-type colorectal cancers for treatment with epidermal growth factor receptor inhibitors.[4] However, in the adjuvant setting for colorectal cancer, the primary question is binary (whether to offer treatment at all) and subsequent selection of drugs, dose, and schedule is predominantly driven by stage rather than by companion diagnostics. Refinement of prognostic models could allow a more targeted approach to selection of adjuvant therapies by defining subgroups in which the absolute benefits of adjuvant chemotherapy are minimal relative to surgery alone and, at the other end of the spectrum, patients who might benefit from prolonged combination chemotherapy because of their poor survival rate.[5–8]

More than two decades of adjuvant trials in patients with early-stage colorectal cancer using fluoropyrimidines, in combination with cytotoxic agents such as oxaliplatin, have yielded an improved overall survival of around 3–5% for patients with stage II or IIIA colorectal cancer. Many patients are cured by surgery alone, while about 25% will recur despite adjuvant chemotherapy. The chemotherapy-associated death rate is likely to be 0·5–1%, and 20% of patients will experience substantial side-effects from treatment. The risk–benefit ratio is marginal, but could potentially be much better if subgroups could be defined as patients having a higher or lower risk of recurrence and cancer-specific death.[9–12]

Although clinically validated prognostic biomarkers would facilitate adjuvant therapeutic decisions, very few

Research in context

Evidence before this study

Digital image analysis is one of the areas in which deep learning has achieved the most important results. We searched PubMed on June 12, 2019, without language or date restrictions using the terms "deep learning", "prediction", "survival", "cancer", and "histology". We systematically reviewed the 214 search results, and found 18 original research studies which applied deep learning to predict patient outcome or related attributes using histopathology images.

In 16 studies, the patient outcome was indirectly predicted by identifying attributes known to correlate with patient outcome—eg, stromal fraction, mitotic count, or Gleason pattern. Two studies reported on direct prediction of survival, but neither presented a marker for automatic prediction of patient outcome from scanned whole-slide sections; one required manual annotation to locate interesting tissue regions, and the other classified tissue microarray spots. Notably, the two studies did not evaluate their biomarker in independent cohorts; the performance was instead estimated using cross-validation in the same cohort as was used for training, which can easily lead to overoptimistic estimates.

Added value of this study

We have applied deep learning to develop a biomarker for automatic prediction of cancer-specific survival directly from scanned haematoxylin and eosin stained, formalin-fixed, paraffin-embedded tumour tissue sections. Independent validation demonstrated that the biomarker improved prediction of cancer-specific survival by stratifying patients with stage II and III colorectal cancer into distinct prognostic groups, supplementing established prognostic markers, and outperforming most existing markers in terms of hazard ratios. The marker could potentially be used to improve selection of adjuvant treatment after resection of colorectal cancer by identifying patients at very low risk who could have been cured by surgery alone, as well as patients at high risk who are much more likely to benefit from more intensive regimes.

Implications of all the available evidence

Deep learning can be used to develop biomarkers for automatic prediction of patient outcome directly from conventional histopathology images. In colorectal cancer, the marker was found to be a clinically useful prognostic marker in the analysis of a large series of patients who received consistent, modern cancer treatment.

Department of Pathology, University College London Medical School, London, UK (Prof M Novelli)

Correspondence to:
Prof Håvard E Danielsen, Institute for Cancer Genetics and Informatics, Oslo University Hospital, NO-0424 Oslo, Norway
hdaniels@labmed.uio.no

have been sufficiently robustly validated for routine clinical application. A case can be made for assessment of mismatch repair status,[13,14] as patients with mismatch repair-deficient tumours tend to have a good prognosis. We have recently reported that measurement of tumour cellular DNA content (ploidy) in combination with stromal fraction can stratify patients with stage II tumours into very good, intermediate, and poor prognostic groups.[15] An analysis has shown that driver mutations and RNA signatures are individually weak prognostic markers and unable to guide clinical decision making.[8,14]

Deep learning refers to the class of machine learning methods that make use of successively more abstract representations of the input data to perform a specific task. These methods use training data to learn how these representations should be generated in a manner appropriate for the given task. By contrast, traditional machine learning uses handcrafted features to create representations of the input data that are applied to perform the task. In many applications, deep learning has been shown to be superior to other machine learning techniques, and is expected to transform current medical practice. Convolutional neural networks have excelled in many image interpretation tasks and could be hypothesised to retrieve additional information from histopathology images. The aim of this study was to use deep learning to analyse conventional whole-slide images to develop an automatic prognostic biomarker for patients resected for primary colorectal cancer.

Methods

Training and tuning cohorts

Four different cohorts were used for training and tuning to achieve a broad patient representation and thereby improve the ability to generalise results to other cohorts. Three cohorts were consecutive series of stage I, II, or III tumours from patients with colorectal cancer treated at hospitals with both rural and urban catchment areas: patients treated between 1988 and 2000 at Akershus University Hospital (Ahus), Norway;[16] patients treated between 1993 and 2003 at Aker University Hospital, Norway;[15] and patients treated in Gloucester, UK, between 1988 and 1996 and included in the Gloucester Colorectal Cancer Study, UK.[17,18] The fourth cohort consisted of patients with stage II or III colorectal cancer treated at 151 UK hospitals in 2002–04 and included in the VICTOR trial (ISRCTN registry, ISRCTN98278138).[19] Common inclusion criteria for the four cohorts were resectable stage I, II, or III non-synchronous colorectal cancer, slides with haematoxylin and eosin (H&E) stained tumour tissue section of adequate quality, and at least one tile within the automatically segmented tumour region (appendix pp 52–62).

See Online for appendix

To obtain clear ground truth, patients with a so-called distinct outcome, either good or poor, were used as a training cohort. A patient was assigned to the good outcome group if they were younger than 85 years at surgery, had more than 6 years follow-up after surgery, and had no record of recurrence or cancer-specific death. The poor outcome group consisted of patients younger than 85 years at surgery with cancer-specific death

between 100 days (inclusive) and 2·5 years (exclusive) after surgery. Patients not satisfying either of these group criteria were defined as having a non-distinct outcome, and these patients were used for tuning. The protocol specifies additional cohort details (appendix pp 52–56).

Test cohort

The test cohort consisted of patients from the Gloucester Colorectal Cancer Study, UK.[17,18] Whole-slide images were obtained from different formalin-fixed paraffin-embedded (FFPE) tumour tissue blocks than those used for the training and tuning cohorts, and the slides were also prepared at different laboratories.

Validation cohort

The marker was independently validated according to the predefined protocol (appendix pp 52–80). The validation cohort consisted of patients from 170 hospitals in seven countries recruited to the QUASAR 2 trial (ISRCTN registry, ISRCTN45133151).[20] Inclusion criteria were age 18 years or older, colorectal adenocarcinoma histologically proven to be R0 M0 stage III or high-risk stage II, primary resection 4–10 weeks before randomisation, WHO performance status score 0 or 1, and life expectancy (with comorbidities, but excluding cancer risk) of at least 5 years. Exclusion criteria and other details are presented in the appendix (pp 73–76). All patients received adjuvant therapy, either capecitabine plus bevacizumab, or capecitabine alone, with equal disease-free and overall survival in both trial groups.[20]

Sample preparation

Slides from the VICTOR cohort were prepared in Oxford, UK, while the slides in the other three training and tuning cohorts were prepared at the Institute for Cancer Genetics and Informatics (ICGI), Norway. Introducing this variation in the development phase was hypothesised to increase the robustness and generalisability of the trained marker. Slides in the test cohort were prepared as part of the routine histopathological examination in Cheltenham, UK, and the performance in this cohort should thus indicate the prognostic ability when the marker is assayed at a different laboratory using original slides. Slides in the validation cohort were prepared at ICGI. All slides were made by staining a 3 µm FFPE tissue block section with H&E, and a pathologist (MP) ascertained that it contained tumour tissue. Whole-slide images were acquired at the highest resolution available (40× magnification) on two scanners, an Aperio AT2 (Leica Biosystems, Germany) and a NanoZoomer XR (Hamamatsu Photonics, Japan).

Areas with high tumour content were identified using a segmentation network that was trained on a subset of the training and tuning cohorts (appendix pp 57–61). A whole-slide image with the 40× resolution typically contained an order of 100 000×100 000 pixels, multiple orders of magnitude larger than images currently feasible for classification by deep learning methods. To preserve prognostic information contained at high resolution, whole-slide images were partitioned into multiple non-overlapping image regions called tiles at 10× and 40× resolutions, and each pixel at 40× represents a physical size of approximately 0·24×0·24 µm².

Classification

Five networks were trained on the 634 564 tiles at 10× resolution and five networks on the 11 591 555 tiles at 40× resolution from the 1652 Aperio AT2 and NanoZoomer XR whole-slide images in the training cohort with the patients' distinct outcomes as ground truth. All networks were DoMore v1 networks, which we designed for classifying supersized heterogeneous images. The DoMore v1 network was built around multiple instance learning and comprised of a MobileNetV2[21] representation network, a Noisy-AND pooling function,[22] and a fully connected classification network similar to the one used by Kraus and colleagues[22] (figure 1). Because of spatial heterogeneity, labelling a tile with the label of its whole-slide image might be problematic. Instead, the networks were trained on labelled collections of tiles. A collection contained tiles from a single whole-slide image, which label it inherits. Collections of tiles were processed by the representation network before the resulting tile representations were pooled and classified. The entire network was trained end-to-end (ie, directly from image to patient outcome), and each training iteration used a batch size of 32 collections with 64 tiles each. The use of this many tiles was possible because we used a novel gradient approximation technique that substantially reduces memory usage during training (appendix pp 4–6). The Noisy-AND pooling function applied a trained non-linear function on tile representation averages. This function enhances robustness against tiles not representing the ground truth and, together with the large number of tiles, alleviates the issues of spatial heterogeneity. During inference, the network processed all tiles in the whole-slide image.

The networks were trained beyond apparent convergence using TensorFlow 1.10, and a model was selected from each network training using the performance in the tuning cohort with the c-index as metric, resulting in five models for each resolution (appendix pp 62–71). Each of the five models provides a score reflecting the probability of poor outcome, and the average was defined as the ensemble score. For use in categorical markers, suitable thresholds for the 10× and the 40× ensemble scores were determined by evaluations in the tuning cohort to define the ensemble classifiers (appendix pp 71–73). Furthermore, evaluations in the development phase indicated that combining 10× and 40× markers might be desirable, and two such markers were defined, one continuous and one categorical. The continuous DoMore-v1-colorectal cancer (DoMore-v1-CRC) score was defined as the average of the 10× and the 40× ensemble

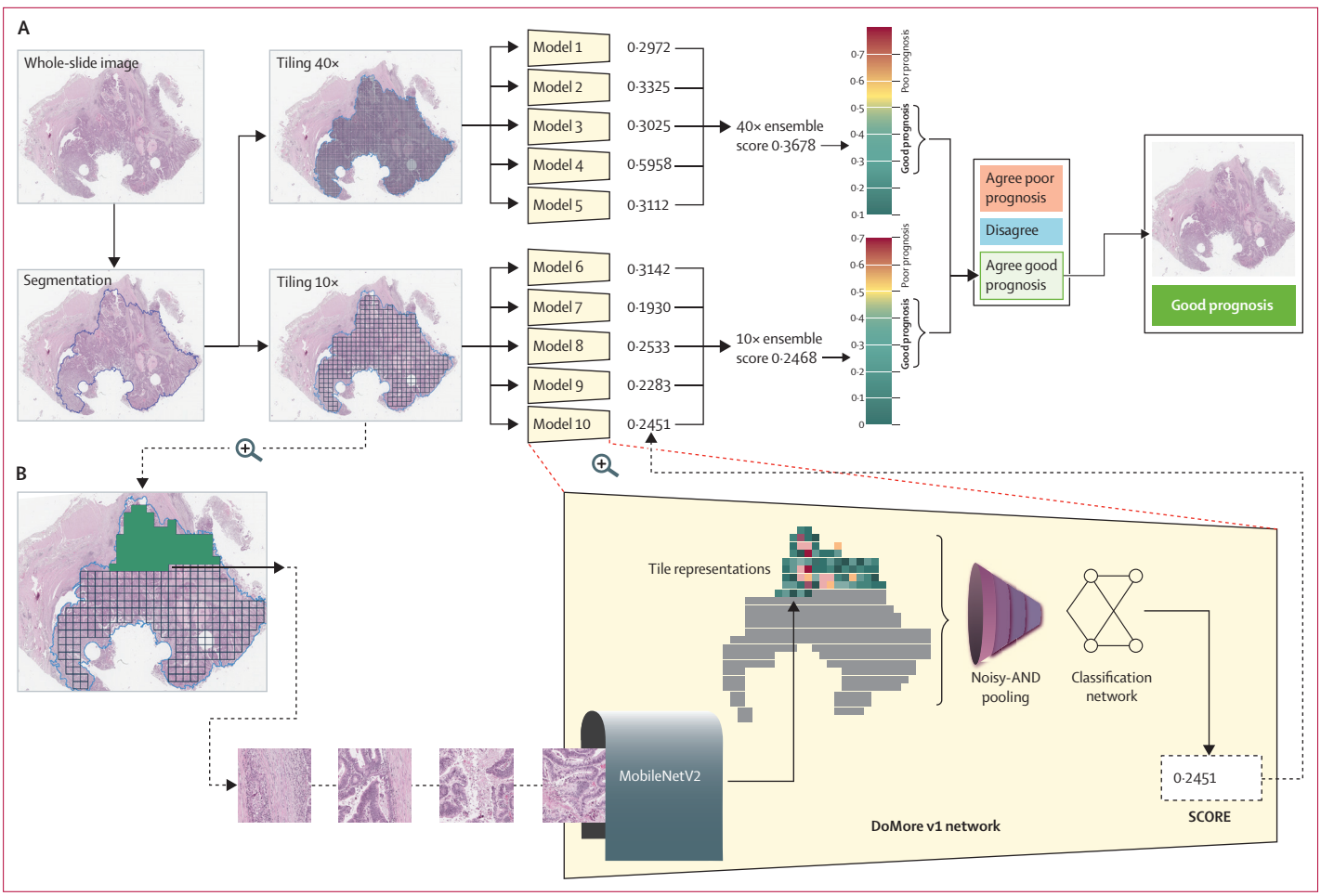

Figure 1: **Pipeline of DoMore-v1-CRC classification**
(A) A whole-slide image is segmented, and the segmented regions tiled at 40× resolution and 10× resolution. For each resolution, the five trained models each produce one score reflecting the probability of poor outcome. The average of those scores is the ensemble score, one for 10× and another for 40×. If the ensemble score is above a certain threshold, the whole-slide image is classified as poor prognosis. The DoMore-v1-CRC class is determined by the agreement between the two ensemble classifications. (B) The DoMore v1 network is comprised of a representation network (MobileNetV2),[21] a pooling function (Noisy-AND),[22] and a simple fully connected classification network. All components of the DoMore v1 network involve trainable parameters, and the entire network is trained end-to-end. All tiles from a whole-slide image are processed by the representation network one by one, resulting in a collection of tile representations. The pooling function reduces the representations into two numbers, which are then processed by the classification network to produce the score outputted by the model. CRC=colorectal cancer.

scores. The categorical DoMore-v1-CRC classifier assigned patients to good prognosis if both ensemble classifiers predicted a good outcome, to uncertain prognosis if the ensemble classifiers predicted differently, and to poor prognosis if both ensemble classifiers predicted a poor outcome. The appendix video visually presents how the DoMore-v1-CRC classifier was trained, tuned, tested, and independently validated. In a post-hoc analysis, the continuous DoMore-v1-CRC score was categorised into five risk groups (appendix p 6).

Inception v3, a state-of-the-art convolutional neural network, was trained, tuned, and evaluated with the same study setup as the DoMore v1 network (appendix pp 62–73), and tested as a secondary analysis in the QUASAR 2 validation cohort (appendix p 78). Although the DoMore-v1-CRC marker was trained using multiple instance learning, each single tile was labelled with the label of its whole-slide image in training of the Inception v3 marker. The image distortion algorithm and network hyperparameters were determined independently of the DoMore v1 network in the discovery phase, resulting in slightly different choices for the Inception v3 network (appendix pp 66–67).

See **Online** for video

Statistical analysis

This study conformed to the REMARK guideline[23] and relevant aspects of the guideline proposed by Luo and colleagues[24] (appendix pp 7–8). Primary and secondary analyses were planned before the evaluations in the validation cohort and are described in the protocol.

The predefined primary analysis for each scanner was univariable cancer-specific survival analysis of the DoMore-v1-CRC classifier; for simplicity, we first present results for the Aperio AT2 scanner and then address

	Training cohort (n=828)	Tuning cohort (n=1645)	Test cohort (n=920)	Validation cohort (n=1122)
Age, years	69 (61–75)	70 (61–77)	71 (64–78)	65 (59–71)
Sex				
Female	402 (49%)	689 (42%)	421 (46%)	477 (43%)
Male	426 (51%)	956 (58%)	499 (54%)	645 (57%)
Stage				
I	101 (12%)	102 (6%)	70 (8%)	..
II	317 (38%)	797 (48%)	354 (38%)	402 (36%)
III	410 (50%)	746 (45%)	496 (54%)	720 (64%)
pN stage				
pN0	415 (50%)	891 (54%)	425 (46%)	402 (36%)
pN1	241 (29%)	492 (30%)	258 (28%)	508 (45%)
pN2	167 (20%)	239 (15%)	237 (26%)	183 (16%)
Missing	5 (1%)	23 (1%)	0	29 (3%)
pT stage				
pT1	26 (3%)	30 (2%)	6 (1%)	17 (2%)
pT2	110 (13%)	137 (8%)	65 (7%)	71 (6%)
pT3	464 (56%)	1034 (63%)	411 (45%)	582 (52%)
pT4	223 (27%)	423 (26%)	437 (48%)	404 (36%)
Missing	5 (1%)	21 (1%)	1 (<1%)	48 (4%)
Histological grade				
1	77 (9%)	196 (12%)	134 (15%)	45 (4%)
2	568 (69%)	1151 (70%)	489 (53%)	846 (75%)
3	178 (21%)	280 (17%)	297 (32%)	168 (15%)
Missing	5 (1%)	18 (1%)	0	63 (6%)
Location				
Rectum	222 (27%)	457 (28%)	311 (34%)	165 (15%)
Distal colon	262 (32%)	533 (32%)	280 (30%)	451 (40%)
Proximal colon	307 (37%)	505 (31%)	329 (36%)	453 (40%)
Missing	37 (4%)	150 (9%)	0	53 (5%)
Adjuvant treatment				
No	467 (56%)	826 (50%)	538 (58%)	0
Chemotherapy	173 (21%)	397 (24%)	51 (6%)	1122 (100%)
Radiotherapy	11 (1%)	6 (<1%)	14 (2%)	0
Chemotherapy and radiotherapy	3 (<1%)	9 (1%)	3 (<1%)	0
Missing	174 (21%)	407 (25%)	314 (34%)	0
Follow-up time, years	6·4 (1·7–8·2)	4·0 (2·2–5·2)	2·4 (1·0–4·6)	4·6 (3·3–5·1)

Data are median (IQR) or n (%).

Table 1: **Baseline characteristics in the training, tuning, test, and validation cohorts**

scanner differences. The classifier was included as the only variable in a Cox model to compute the hazard ratio (HR) with 95% CI of patients with uncertain and poor prognosis relative to patients with good prognosis. The proportional hazards assumption was found to be satisfactorily fulfilled using log–log plots (appendix p 26). The Mantel-Cox log-rank test was used to assess whether the classifier predicted cancer-specific survival.

Both the classifier and the continuous score were evaluated in multivariable Cox models as secondary and post-hoc analyses, including markers available at the time of analysis (patients with at least one missing value were excluded). The p values in the multivariable analyses were calculated using the Wald χ^2 test, both when testing the difference between a specific category of a marker and its reference category, and when testing the overall difference between the categories of a marker. Associations between the classifier and other markers were evaluated with Spearman's correlation coefficients. To calculate classification metrics for 3-year cancer-specific survival, patients without an event and with less than a 3-year follow-up were excluded, and events after 3 years were not included in the analysis. Category-free net reclassification improvement (NRI) was computed with the Kaplan-Meier estimates of 5-year cancer-specific survival. A two-sided p value of less than 0·05 was considered significant, and the confidence level of CIs is 95%. The bias-corrected and accelerated bootstrap CIs were computed for NRIs, c-indices, and areas under the curves (AUCs) using 10 000 bootstrap replicates and an acceleration constant was estimated using leave-one-out cross-validation. Time to cancer-specific survival in the validation cohort was calculated from date of randomisation to date of cancer-specific death or loss to follow-up. Survival analyses were done with Stata/SE 15.1.

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, writing the report, or the decision to submit the paper for publication. The corresponding author had full access to all data and the final responsibility to submit for publication.

Results

Four cohorts were used for training and tuning, 160 patients were included from the Ahus cohort, 576 patients from the Aker cohort, 970 patients from the Gloucester cohort, and 767 patients from the VICTOR trial cohort (appendix pp 52–56). 828 patients from these four cohorts had a distinct outcome (good or poor) and were used as a training cohort to obtain clear ground truth. 1645 patients had a non-distinct outcome and were used for tuning. Patient demographics are summarised in table 1.

The DoMore-v1-CRC classifier was a strong predictor of cancer-specific survival in the primary analysis of the 1122 patients in the validation cohort (for uncertain *vs* good prognosis, HR 1·89, 95% CI 1·14–3·15; for poor *vs* good prognosis 3·84, 2·72–5·43; p<0·0001; figure 2). The classifier remained strong in multivariable analysis (for uncertain *vs* good prognosis, HR 1·56, 0·92–2·65, p=0·10; for poor *vs* good prognosis, 3·04, 2·07–4·47, p<0·0001; table 2) adjusting for established prognostic markers significant in univariable analyses: pN stage, pT stage, lymphatic invasion, and venous vascular invasion (appendix p 9).

The sensitivity was 52% (95% CI 41–63), specificity 78% (75–81), positive predictive value 19% (14–25), negative predictive value 94% (92–96), and the proportion

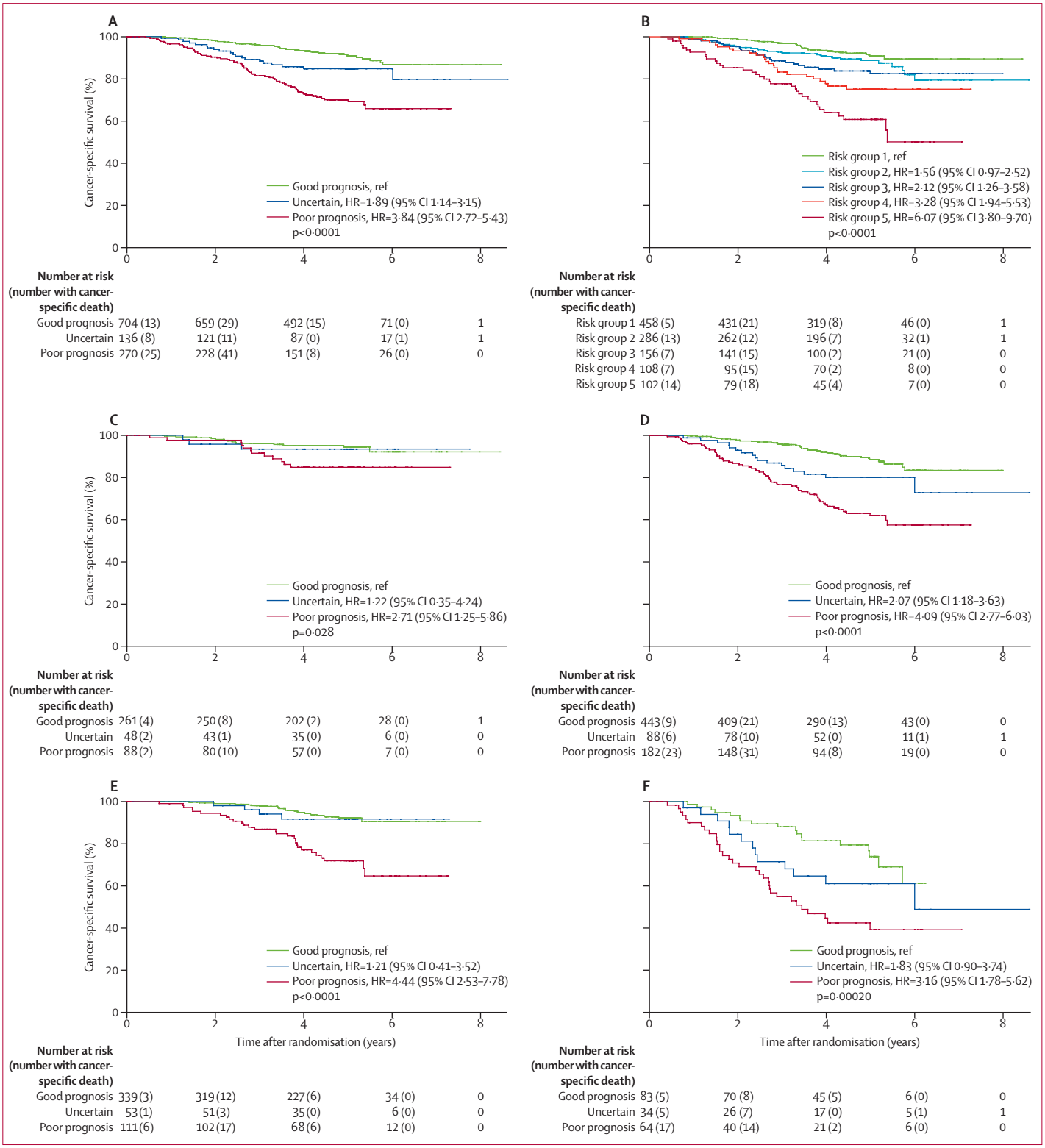

Figure 2: **Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC classifier, evaluated on Aperio AT2 slide images in the QUASAR 2 validation cohort**
(A) All patients were evaluated with the predefined DoMore-v1-CRC classifier for the primary analysis. (B) All patients were evaluated with the DoMore-v1-CRC classifier variant with five categories in a post-hoc analysis. (C) Patients with stage II (equivalent to pN0) cancer were evaluated with the predefined DoMore-v1-CRC classifier in a secondary analysis. (D) Patients with stage III cancer were evaluated with the predefined DoMore-v1-CRC classifier in a secondary analysis. (E) pN1 patients were evaluated with the predefined DoMore-v1-CRC classifier in a post-hoc analysis. (F) pN2 patients were evaluated with the predefined DoMore-v1-CRC classifier in a post-hoc analysis. CRC=colorectal cancer. HR=hazard ratio.

	Stage II and III		Stage II		Stage III	
	HR (95% CI)	p value	HR (95% CI)	p value	HR (95% CI)	p value
DoMore-v1-CRC	..	<0·0001*	..	0·028*	..	<0·0001*
Good prognosis	1 (ref)	..	1 (ref)	..	1 (ref)	..
Uncertain	1·56 (0·92–2·65)	0·10	1·22 (0·35–4·24)	0·76	2·14 (1·15–3·99)	0·017
Poor prognosis	3·04 (2·07–4·47)	<0·0001	2·71 (1·25–5·86)	0·011	2·95 (1·81–4·82)	<0·0001
pN stage	..	<0·0001*	..	..	..	..
pN0	1 (ref)	..	..	..	..	..
pN1	1·84 (1·13–2·98)	0·014	..	..	1 (ref)	..
pN2	5·94 (3·71–9·52)	<0·0001	..	..	3·31 (2·14–5·13)	<0·0001
pT stage	..	0·0058*	..	..	..	0·014*
pT1	0 (0–∞)	1	..	..	0 (0–∞)	1
pT2	1·86 (0·90–3·86)	0·096	..	..	1·68 (0·64–4·45)	0·29
pT3	1 (ref)	..	..	..	1 (ref)	..
pT4	1·75 (1·22–2·51)	0·0024	..	..	2·07 (1·33–3·22)	0·0013
Lymphatic invasion						
No	1 (ref)	..	..	..	1 (ref)	..
Yes	1·66 (1·07–2·56)	0·023	..	..	1·98 (1·20–3·28)	0·0079
Venous vascular invasion						
No	1 (ref)	..	..	..	1 (ref)	..
Yes	1·07 (0·76–1·51)	0·71	..	..	0·98 (0·64–1·52)	0·94
Sidedness						
Left	..	..	..	..	1 (ref)	..
Right	..	..	..	..	1·09 (0·70–1·70)	0·69
BRAF						
Wild type	..	..	..	..	1 (ref)	..
Mutated	..	..	..	..	1·39 (0·81–2·40)	0·24

The multivariable model included the DoMore-v1-CRC class evaluated on Aperio AT2 slide images and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort. HR=hazard ratio. CRC=colorectal cancer. *Wald test of difference between categories of the variable.

Table 2: **Multivariable cancer-specific survival analyses in the validation cohort**

of correctly classified patients (accuracy) was 76% (73–79) when comparing 3-year cancer-specific survival for the good prognosis group of the DoMore-v1-CRC classifier with the uncertain and poor prognosis groups. When comparing good and uncertain prognosis with poor prognosis, the sensitivity was 69% (95% CI 58–78), specificity 66% (63–69), positive predictive value 17% (13–21), negative predictive value 96% (94–97), and accuracy 67% (63–69).

The constituents of the DoMore-v1-CRC classifier, the 10× and the 40× ensemble classifiers, were strong predictors in univariable (appendix p 27) and multivariable analyses (appendix pp 10–11). The ensemble classifiers performed similarly as the best classifiers based on one of the ten individual models that constituted the ensemble models (appendix pp 12 and 28–29). The continuous ensemble scores were also strong predictors in univariable (appendix p 9) and multivariable analyses (appendix pp 13–15). The DoMore-v1-CRC score was strongly associated with the patient outcome (appendix p 30), and provided a c-index of 0·674 (95% CI 0·624–0·719; appendix p 16) in all validation patients and an AUC of 0·713 (0·624–0·789; appendix p 31) in patients with a distinct outcome. The c-index and AUC of the 10× ensemble score were similar to the values obtained for the DoMore-v1-CRC score (appendix pp 16 and 31).

The DoMore-v1-CRC classifier was a significant predictor of cancer-specific survival in stage II (for poor *vs* good prognosis, HR 2·71, 95% CI 1·25 to 5·86, p=0·011; figure 2C) and stage III (poor *vs* good prognosis, 4·09, 2·77 to 6·03, p<0·0001; figure 2D), and this was confirmed in multivariable analysis (table 2) and for the continuous score (appendix pp 9, 13). The categorical marker identified patient groups with substantially different cancer-specific survival periods in stage IIIB and IIIC (appendix p 32), and also identified significant differences within pN stages (figures 2C, E, and F) and pT stages (pT1–3 *vs* pT4; appendix p 33). The category-free NRI of supplementing substage with the DoMore-v1-CRC class for prediction of 5-year cancer-specific survival was 61·6% (95% CI 43·5 to 79·3); the event-NRI was 3·2% (–13·2 to 20·0), and the non-event-NRI was 58·3% (52·7 to 63·8).

The DoMore-v1-CRC classifier correlated with a number of factors such as age, pN stage, pT stage, histological grade, location, tumour sidedness, *BRAF* mutation, and microsatellite instability (table 3). The association

between the DoMore-v1-CRC classifier and histological grading was further studied in the test cohort, in which all gradings were centrally reviewed by a highly experienced pathologist (NAS).[17,18] Among 133 tumours characterised as well differentiated, the DoMore-v1-CRC classifier assigned 101 tumours as good prognosis, 18 as uncertain, and 14 as poor prognosis (appendix p 17). The moderately differentiated tumours were evenly distributed among the DoMore-v1-CRC classes. Of 292 poorly differentiated tumours, the marker

	DoMore-v1-CRC classification			Spearman's correlation	
	Good prognosis (n=704)	Uncertain (n=136)	Poor prognosis (n=270)	ρ (95% CI)	p value
Age (continuous), years	64 (58–71)	65 (60–71)	66 (60–72)	0·07 (0·01 to 0·13)	0·024
Age (dichotomous), years	..	..	..	0·03 (–0·03 to 0·09)	0·38
≤72	568 (81%)	112 (82%)	209 (77%)	..	..
>72	136 (19%)	24 (18%)	61 (23%)	..	..
Sex	..	..	..	–0·02 (–0·08 to 0·04)	0·59
Female	297 (42%)	53 (39%)	122 (45%)	..	..
Male	407 (58%)	83 (61%)	148 (55%)	..	..
Stage	..	..	..	0·04 (–0·02 to 0·10)	0·20
II	261 (37%)	48 (35%)	88 (33%)	..	..
III	443 (63%)	88 (65%)	182 (67%)	..	..
Stage with substage	..	..	..	0·15 (0·09 to 0·21)	<0·0001
IIA	143/672 (21%)	19/133 (14%)	28/256 (11%)	..	..
IIB	110/672 (16%)	27/133 (20%)	54/256 (21%)	..	..
IIIA	67/672 (10%)	2/133 (2%)	6/256 (2%)	..	..
IIIB	269/672 (40%)	51/133 (38%)	104/256 (41%)	..	..
IIIC	83/672 (12%)	34/133 (26%)	64/256 (25%)	..	..
pN stage	..	..	..	0·10 (0·04 to 0·16)	<0·0001
pN0	261/683 (38%)	48/135 (36%)	88/263 (33%)	..	..
pN1	339/683 (50%)	53/135 (39%)	111/263 (42%)	..	..
pN2	83/683 (12%)	34/135 (25%)	64/263 (24%)	..	..
pT stage	..	..	..	0·26 (0·21 to 0·32)	<0·0001
pT1	15/672 (2%)	0	2/256 (1%)	..	..
pT2	61/672 (9%)	3/134 (2%)	6/256 (2%)	..	..
pT3	402/672 (60%)	75/134 (56%)	100/256 (39%)	..	..
pT4	194/672 (29%)	56/134 (42%)	148/256 (58%)	..	..
Lymphatic invasion	..	..	..	0·04 (–0·02 to 0·10)	0·20
No	599/661 (91%)	122/132 (92%)	220/253 (87%)	..	..
Yes	62/661 (9%)	10/132 (8%)	33/253 (13%)	..	..
Venous vascular invasion	..	..	..	0·05 (–0·01 to 0·11)	0·11
No	409/666 (61%)	74/132 (56%)	145/257 (56%)	..	..
Yes	257/666 (39%)	58/132 (44%)	112/257 (44%)	..	..
Histological grade	..	..	..	0·14 (0·08 to 0·20)	<0·0001
1	27/668 (4%)	7/127 (6%)	8/253 (3%)	..	..
2	565/668 (85%)	88/127 (69%)	186/253 (74%)	..	..
3	76/668 (11%)	32/127 (25%)	59/253 (23%)	..	..
Location	..	..	..	0·15 (0·09 to 0·21)	<0·0001
Rectum	118/665 (18%)	21/131 (16%)	23/261 (9%)	..	..
Distal colon	301/665 (45%)	46/131 (35%)	100/261 (38%)	..	..
Proximal colon	246/665 (37%)	64/131 (49%)	138/261 (53%)	..	..
Sidedness	..	..	..	0·14 (0·08 to 0·20)	<0·0001
Left	419/665 (63%)	67/131 (51%)	123/261 (47%)	..	..
Right	246/665 (37%)	64/131 (49%)	138/261 (53%)	..	..
KRAS	..	..	..	–0·06 (–0·12 to 0·00)	0·069
Wild type	410/634 (65%)	86/118 (73%)	169/242 (70%)	..	..
Mutated	224/634 (35%)	32/118 (27%)	73/242 (30%)	..	..

(Table 3 continues on next page)

	DoMore-v1-CRC classification			Spearman's correlation	
	Good prognosis (n=704)	Uncertain (n=136)	Poor prognosis (n=270)	ρ (95% CI)	p value
(Continued from previous page)					
BRAF	..	..	..	0·22 (0·16 to 0·28)	<0·0001
Wild type	588/635 (93%)	89/118 (75%)	190/246 (77%)	..	..
Mutated	47/635 (7%)	29/118 (25%)	56/246 (23%)	..	..
Microsatellite instability	..	..	..	–0·10 (–0·16 to –0·04)	0·0018
Yes	66/661 (10%)	26/125 (21%)	40/253 (16%)	..	..
No	595/661 (90%)	99/125 (79%)	213/253 (84%)	..	..
Follow-up time, years	4·8 (3·7–5·1)	4·9 (3·1–5·1)	4·1 (2·8–5·1)	–0·10 (–0·16 to –0·04)	<0·0001

Data are n (%), n/N (%), or median (IQR). CRC=colorectal cancer.

Table 3: **Associations between the DoMore-v1-CRC class evaluated on Aperio AT2 slide images and different patient characteristics in the validation cohort**

assigned 223 as poor prognosis, 36 as uncertain, and 33 as good prognosis. Thus, the DoMore-v1-CRC class was clearly associated with tumour differentiation. The large proportion of tumours classified as moderately differentiated (489 [53%] of 920 in the test cohort and 846 [75%] of 1122 in the validation cohort) restricts the usefulness of this grading system, but these patients could also be risk stratified by the DoMore-v1-CRC marker (appendix p 34).

Median processing time per patient for the entire classification pipeline (ie, from scan to predicted patient outcome) was 2·8 min (IQR 1·8–3·9) in the validation cohort on a computer with an NVIDIA GeForce RTX 2080 Ti and an Intel Core i7–7700K.

Inception v3 provided a marker of cancer-specific survival with a slightly worse performance than the DoMore-v1-CRC classifier (appendix pp 16, 35–36).

In the test cohort with slides from 920 patients prepared at a different hospital, the classifier provided similar HRs (appendix p 37) as in the validation cohort (figure 2), supporting that it is robust against inter-laboratory differences in tissue preparation and staining.

When evaluated with another scanner (NanoZoomer XR), the DoMore-v1-CRC score tended towards slightly higher values than it did when evaluated with the Aperio AT2 scanner, resulting in a higher DoMore-v1-CRC class for some patients near the classification thresholds (appendix p 38). However, the scores correlated strongly (Pearson's r=0·956; 95% CI 0·951–0·961), and the classifier provided similar prognostic information with both scanners (appendix pp 9, 16, 18–25, 39–51). Thus, the classifier was also a strong predictor of cancer-specific survival in the primary analysis of the validation cohort when evaluated on NanoZoomer XR slide images (for uncertain *vs* good prognosis, HR 2·42, 95% CI 1·45–4·03; for poor *vs* good prognosis, 3·39, 2·36–4·87; p<0·0001; appendix p 39).

Discussion

Building on recent developments in machine learning, we have developed a biomarker for automatic prediction of the outcome of a patient resected for early-stage colorectal cancer, which directly analyses standard histological sections stained with H&E. To assay the biomarker, one convolutional neural network first automatically outlines cancerous tissue, and then a second convolutional neural network stratifies the patients into prognostic categories. In the validation cohort, the good and poor prognosis groups included nearly 90% of the patients and HR for cancer-specific survival was about four times higher in the univariable analysis and about three times higher in the multivariable analysis. The multivariable result indicated that the new biomarker will be a useful supplement to the established markers and improve risk stratification.

Deep learning has already been shown to be suitable for detection and delineation of some tumour types,[25] and various cancer classifications have been reported.[26] Previous studies have suggested that deep learning could be used to develop markers that potentially use basic morphology to predict the outcome of patients with cancer, but these findings have not been validated in independent cohorts.[27,28]

We derived two markers using the same study setup, but different deep learning techniques. In training the Inception v3 marker, each tile was labelled with the label of its whole-slide image, while the DoMore-v1-CRC marker was developed using multiple instance learning to allow training on tile collections labelled with the label of its whole-slide image. Both markers were strong predictors of cancer-specific survival, but the DoMore-v1-CRC marker performed slightly better and was the marker preselected for independent validation in the QUASAR 2 cohort.

Automatic prognostication procedures reduce human intervention and have the potential to increase reproducibility of biomarkers. New procedures such as the DoMore-v1-CRC markers might initially be performed as services carried out at specialised laboratories with a high degree of standardisation to avoid disparities in sample handling, including staining and scanning. Such centralised processing will also facilitate the collection of information on new procedures and enable improvements in the decision support to pathologists and clinicians.

As an increasing number of laboratories are becoming digitalised, accompanying decision support systems could include standardisation modules and facilitate a more rapid spread of the automatic procedures. Moreover, supplemented by increased robotisation of wet-lab procedures, the higher analytical throughput will allow decisions based on multiple samples from a tumour. The potential to base decisions on multiple samples from a single tumour could reduce the challenge of tumour heterogeneity, which could be a key to improved accuracy of prognosis.

The DoMore-v1-CRC biomarker correlated with several recognised prognostic factors, including the histological grading carried out by a specialised pathologist. The classifier performed better than did most other markers in terms of HRs in stage-specific multivariable analyses, and similar to the pN staging system. By contrast to the grading system, the classifier categorised only a few patients into the intermediate uncertain group.

The DoMore-v1-CRC classifier is technically simple to apply and can be delivered at pathology laboratories in a variety of settings. Although training the networks was resource demanding, new patients can be assayed in a few minutes using consumer hardware.

Clinically, the marker will inform discussion with patients with stage II and III colorectal cancer on the risks and benefits of different adjuvant treatment options. Although the drugs used in the adjuvant setting are limited to fluoropyrimidines with or without oxaliplatin, recent data demonstrate that 3 months' treatment results in approximately the same survival outcomes as 6 months' treatment for the majority of patients with stage III cancer, and suggest that high-risk patients (pT4 and pN2) might benefit from prolonged therapy.[29,30] Hypothesising that patients with stage III cancer identified as poor prognosis by the DoMore-v1-CRC classifier could benefit from prolonged combination chemotherapy with oxaliplatin, or even consider experimental therapy combining fluoropyrimidine with oxaliplatin and irinotecan might be reasonable, because their high risk of cancer-specific death should positively skew the risk–benefit ratio of more aggressive treatments. However, patients with stage III cancer who were classified as good prognosis—the majority of whom are pN1—have very good survival with single-agent capecitabine, and good prognosis patients with stage II cancer have a very high chance of surgical cure, potentially eliminating the need for adjuvant treatment.

We plan to undertake prospective adjuvant trials stratifying patients into different prognostic groups using the DoMore-v1-CRC biomarker and randomly assigning patients into observation, low intensity, and high intensity regimes, depending on relative risk score. However, the currently available data could also be used by clinicians and patients to make joint and more informed decisions on adjuvant chemotherapy choices, as the proportional reduction in the HRs for recurrence and death from colorectal cancer following adjuvant treatment is remarkably consistent at 20% across most well designed clinical trials, thus translating into quite different absolute survival improvements for low-risk and high-risk subgroups.

A limitation of this study was that the DoMore-v1-CRC marker has not yet been tested prospectively in clinical settings and, although we are planning a clinical trial with randomisation, we at present only know the outcome of thorough retrospective testing. The test and validation indicate good transferability between populations, but challenges associated with standardisation remain, as shown by the differences between the tested scanners. Differences between laboratories might also be seen for sample handling procedures, and therefore introduction into the clinic is suggested to be through services provided by specialised laboratories. A well known disadvantage of deep learning is its black-box nature. The DoMore-v1-CRC marker is associated with histological grading, but the marker is still using small-scale features of the histological images with unknown biological correlates.

In summary, a clinically useful prognostic marker has been developed using deep learning allied to digital scanning of conventional H&E stained, FFPE tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, gives consistent results across tumour and nodal stage, and can potentially be used by clinicians to improve decision making regarding adjuvant treatment choices.

Contributors

O-JS, SDR, AK, TSH, KL, FA, DJK, and HED designed the study. HAA, JAN, AN, NAS, IT, RK, MN, and DJK collected the samples and acquired the image data. MP, INF, ED, DNC, AN, NAS, IT, RK, MN, and DJK provided the clinical and pathological data and interpretations. O-JS, SDR, and JM performed the machine learning. AK did the statistical analyses. O-JS, SDR, AK, TSH, KL, DJK, and HED interpreted the data and analyses. All authors vouch for the data, analyses, and interpretations. O-JS, SDR, AK, TSH, KL, DJK, and HED wrote the first draft of the manuscript, and all authors reviewed, contributed to, and approved the manuscript.

Declaration of interests

O-JS, TSH, KL, JM, and HED report filing of a patent application entitled Histological image analysis with International Patent Application Number PCT/EP2018/080828. The University of Oxford (to DJK) received educational grants from Roche to support the QUASAR 2 trial and from Merck to support the VICTOR trial. All other authors declare no competing interests.

Acknowledgments

This study was funded by The Research Council of Norway through its IKTPLUSS Lighthouse program (grant number 259204, project name DoMore!). We thank Akershus University Hospital for access to their patient material, National Institute for Health Research for funding support to MN through Biomedical Research Centres, Paul Callaghan for animating the appendix video, Marian Seiergren for creating figure 1 and assembling figure 2, the laboratory and technical personnel at the Institute for Cancer Genetics and Informatics for assistance, and the reviewers for valuable suggestions. We also would like to thank the participating centres in the VICTOR and QUASAR 2 trials as well as the staff at Akershus University Hospital, Aker University Hospital,

and the Gloucestershire hospitals contributing to the Gloucester Colorectal Cancer Study, and all participating patients.

References

1 La Thangue NB, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nat Rev Clin Oncol* 2011; **8:** 587–96.

2 Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014; **20:** 682–88.

3 Moscow JA, Fojo T, Schilsky RL. The evidence framework for precision cancer medicine. *Nat Rev Clin Oncol* 2018; **15:** 183–92.

4 Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* 2008; **359:** 1757–65.

5 Kerr DJ, Shi Y. Biological markers: tailoring treatment and trials to prognosis. *Nat Rev Clin Oncol* 2013; **10:** 429–30.

6 Hutchins G, Southward K, Handley K, et al. Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *J Clin Oncol* 2011; **29:** 1261–70.

7 Salazar R, Roepman P, Capella G, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29:** 17–24.

8 Gray RG, Quirke P, Handley K, et al. Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. *J Clin Oncol* 2011; **29:** 4611–19.

9 QUASAR Collaborative Group. Comparison of fluorouracil with additional levamisole, higher-dose folinic acid, or both, as adjuvant chemotherapy for colorectal cancer: a randomised trial. *Lancet* 2000; **355:** 1588–96.

10 Gray R, Barnwell J, McConkey C, Hills RK, Williams NS, Kerr DJ. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet* 2007; **370:** 2020–29.

11 André T, Boni C, Navarro M, et al. Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *J Clin Oncol* 2009; **27:** 3109–16.

12 André T, de Gramont A, Vernerey D, et al. Adjuvant fluorouracil, leucovorin, and oxaliplatin in stage II to III colon cancer: updated 10-year survival and outcomes according to BRAF mutation and mismatch repair status of the MOSAIC study. *J Clin Oncol* 2015; **33:** 4176–87.

13 Sinicrope FA. DNA mismatch repair and adjuvant chemotherapy in sporadic colon cancer. *Nat Rev Clin Oncol* 2010; **7:** 174–77.

14 Mouradov D, Domingo E, Gibbs P, et al. Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. *Am J Gastroenterol* 2013; **108:** 1785–93.

15 Danielsen HE, Hveem TS, Domingo E, et al. Prognostic markers for colorectal cancer: estimating ploidy and stroma. *Ann Oncol* 2018; **29:** 616–23.

16 Bondi J, Husdal A, Bukholm G, Nesland JM, Bakka A, Bukholm IR. Expression and gene amplification of primary (A, B1, D1, D3, and E) and secondary (C and H) cyclins in colon adenocarcinomas and correlation with patient outcome. *J Clin Pathol* 2005; **58:** 509–14.

17 Petersen VC, Baxter KJ, Love SB, Shepherd NA. Identification of objective pathological prognostic determinants and models of prognosis in Dukes' B colon cancer. *Gut* 2002; **51:** 65–69.

18 Mitchard JR, Love SB, Baxter KJ, Shepherd NA. How important is peritoneal involvement in rectal cancer? A prospective study of 331 cases. *Histopathology* 2010; **57:** 671–79.

19 Midgley RS, McConkey CC, Johnstone EC, et al. Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *J Clin Oncol* 2010; **28:** 4575–80.

20 Kerr RS, Love S, Segelov E, et al. Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial. *Lancet Oncol* 2016; **17:** 1543–57.

21 Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: inverted residuals and linear bottlenecks. 2018 Institute of Electrical and Electronics Engineers/Computer Vision Foundation (IEEE/CVF) Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; June 18–23, 2018.

22 Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 2016; **32:** i52–59.

23 Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Med* 2012; **10:** 51.

24 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; **18:** e323.

25 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:** 2199–210.

26 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:** 1559–67.

27 Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep* 2018; **8:** 3395.

28 Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci USA* 2018; **115:** e2970–79.

29 Grothey A, Sobrero AF, Shields AF, et al. Duration of adjuvant chemotherapy for stage III colon cancer. *N Engl J Med* 2018; **378:** 1177–88.

30 Iveson TJ, Kerr RS, Saunders MP, et al. 3 versus 6 months of adjuvant oxaliplatin-fluoropyrimidine combination therapy for colorectal cancer (SCOT): an international, randomised, phase 3, non-inferiority trial. *Lancet Oncol* 2018; **19:** 562–78.