The Lancet 2020; 395:10221: 350-360 (DOI: 10.1016/S0140-6736(19)32998-8)

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

Ole-Johan Skrede, M. Sc.[1,2,*], Sepp De Raedt, Ph. D.[1,2,*], Andreas Kleppe, Ph. D.[1,2], Tarjei S. Hveem, Ph. D.[1], Prof. Knut Liestøl, Ph. D.[1,2], John Maddison, Ph. D.[1], Hanne A. Askautrud, Ph. D.[1], Manohar Pradhan, Ph. D.[1], John Arne Nesheim, M. Sc.[1], Prof. Fritz Albregtsen, M. Sc.[1,2], Prof. Inger Nina Farstad, Ph. D.[3,4], Enric Domingo, Ph. D.[5], David N. Church, D. Phil.[6,7], Prof. Arild Nesbakken, Ph. D.[4,8,9], Prof. Neil A. Shepherd, D. M.[10], Prof. Ian Tomlinson, Ph. D.[1,11], Prof. Rachel Kerr, Ph. D.[5], Prof. Marco Novelli, Ph. D.[1,12], Prof. David J. Kerr, D. Sc.[13], Prof. Håvard E. Danielsen, Ph. D.[1,2,13**]

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, Oslo, Norway

[2]Department of Informatics, University of Oslo, Oslo, Norway

[3]Department of Pathology, Division of Laboratory Medicine, Oslo University Hospital, Oslo, Norway

[4]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[5]Department of Oncology, University of Oxford, Oxford, UK

[6]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK

[7]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[8]Department of Gastrointestinal Surgery, Oslo University Hospital, Oslo, Norway

[9]K.G. Jebsen colorectal cancer research centre, Oslo, Norway

[10]Gloucestershire Cellular Pathology Laboratory, Cheltenham General Hospital, Cheltenham, UK

[11]Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, Scotland

[12]Research Department of Pathology, University College London Medical School, London, UK

[13]Nuffield Division of Clinical Laboratory Sciences, University of Oxford, Oxford, UK

*Both authors contributed equally to this work.

**Corresponding author:

Prof Håvard E. Danielsen,

Institute for Cancer Genetics and Informatics,

Oslo University Hospital

Montebello, 0310, Oslo, Norway

Email: hdaniels@labmed.uio.no

Phone: +47 22782320

Words in abstract (not exceed 300): 297

Words in main text (up to 3500): 3889

Number of references (up to 30): 30

Number of figures: 2

Number of tables: 3

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 828 patients with distinctly good or poor disease outcome were used to train a total of 10 convolutional neural networks, purpose-built for classifying supersized heterogeneous images. A prognostic biomarker integrating the 10 networks were determined using 1645 patients with non-distinct outcome. The marker was tested on 920 patients with slides prepared in UK, and finally independently validated according to a pre-defined protocol in 1122 patients treated with single-agent capecitabine using slides prepared in Norway. The primary outcome was cancer-specific survival.

Findings: The biomarker provided a hazard ratio for poor *vs* good prognosis of 3·84 (95% confidence interval, 2·72-5·43; p<0·0001) in the primary analysis of the validation cohort, and 3·04 (95% confidence interval, 2·07-4·47; p<0·0001) after adjusting for established prognostic markers significant in univariable analyses of the same cohort; pN stage, pT stage, lymphatic invasion, and venous vascular invasion.

Interpretation: It was possible to develop a clinically useful prognostic marker 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 these 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 regimes.

Funding: The Research Council of Norway through its IKTPLUSS Lighthouse program (grant number 259204, project name DoMore!).

Research in context

Evidence before this study

Digital image analysis is one of the fields where the recent renaissance of deep learning has achieved the most impressive results. We searched PubMed on June 12, 2019 without language or time restrictions, using the terms “deep learning”, “prediction”, “survival”, “cancer”, and “histology” (full specification of the search criteria is provided in the appendix p 3). 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, e.g. 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. Perhaps even more importantly, neither of these two studies evaluated their biomarker in independent cohorts; the performance was instead estimated using cross-validation in the same cohort as utilised for training, which can easily lead to overoptimistic estimates.

Added value of the 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 stage II and III colorectal cancer patients 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 may 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

It is possible to utilise deep learning 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 analysis of a large series of patients who received consistent, modern cancer treatment.

Introduction

Biomarkers are being used increasingly 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 (CRCs) for treatment with epidermal growth factor receptor inhibitors.[4] However, in the adjuvant setting for CRC, 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. If it were possible to further refine prognostic models, this could allow a more targeted approach 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 CRC using fluoropyrimidines, in combination with cytotoxic agents like oxaliplatin, have yielded an improved overall survival of around 3-5% for patients with stage II or IIIA CRC. Many patients are cured by surgery alone, while around 25% will recur despite adjuvant chemotherapy. There is likely to be a chemotherapy-associated death rate of 0·5-1%, and 20% of patients will suffer significant side-effects. The risk-benefit ratio is therefore rather marginal, but could potentially be much better if it were possible to define subgroups at higher or lower risk of recurrence and cancer-specific death.[9–12]

Although clinically validated prognostic biomarkers would facilitate adjuvant therapeutic decisions, very few have been sufficiently robustly validated for routine clinical application. A case can be made for assessment of mismatch repair (MMR) status,[13,14] as patients with MMR-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 stage II patients into very good, intermediate, and poor prognostic groups.[15] Interestingly, analysis of driver mutations and RNA signatures has shown them to be 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 a training set to learn how these representations should be generated in a manner appropriate for the given task. In contrast, traditional machine learning utilises handcrafted features to create representations of the input data that are applied to perform the task. In many applications, deep learning has been demonstrated to provide superior performance compared to other machine learning techniques, and it is a growing expectation that deep learning will transform current medical practice. Especially convolutional neural networks have excelled in many image interpretation tasks, and could therefore be hypothesised to retrieve additional information from histopathology images. The aim of the present study was to use deep learning to analyse conventional whole-slide images (WSIs) in order to develop an automatic prognostic biomarker for patients resected for primary CRC. The marker was trained using 828 patients with distinct prognosis from four cohorts, fine-tuned using 1645 other patients from the same four cohorts, and tested on slides prepared at a different laboratory from 920 patients. Finally, the marker was independently validated according to the pre-defined protocol (appendix pp 52-80) on 1122 patients analysed retrospectively from a trial (QUASAR 2) of adjuvant therapy.[16]

Methods

Training and Tuning Cohorts

Four different cohorts were utilised for training and tuning to achieve a broad patient representation and thereby improve the ability to generalise to new cohorts. Three cohorts

were consecutive series of stage I, II or III tumours from CRC patients treated at hospitals with both rural and urban catchment areas: (i) 160 patients treated 1988-2000 at Akershus University Hospital, Norway;[17] (ii) 576 patients treated 1993-2003 at Aker University Hospital, Norway;[15] and (iii) 970 patients treated in Gloucester 1988-1996 and included in the Gloucester Colorectal Cancer Study, UK.[18,19] The fourth cohort were 767 stage II or III CRC patients treated at 151 UK hospitals in 2002-2004 and included in the VICTOR trial (ISRCTN registry number ISRCTN98278138).[20] Our cohorts included only patients with resectable tumour, and a formalin-fixed, paraffin-embedded (FFPE) tumour tissue block available for analysis.

To obtain clear ground-truth, we used as training cohort the 828 patients with so-called distinct outcome, either good or poor. A patient was assigned to the good outcome group if aged less than 85 years at surgery, had more than six years follow-up after surgery, and had no record of recurrence or cancer-specific death. The poor outcome group consisted of those aged less than 85 years at surgery and suffered 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 non-distinct outcome, and these 1645 patients were used for tuning. The protocol specifies additional cohort details, and demographics are summarised in table 1.

Test Cohort

The test cohort consisted of 920 patients from the Gloucester Colorectal Cancer Study, UK.[18,19] WSIs were obtained from different FFPE tumour tissue blocks than those used in the training and tuning cohorts.

Validation Cohort

The validation cohort consisted of 1122 patients from 170 hospitals in seven countries recruited to the QUASAR 2 trial (ISRCTN registry number ISRCTN45133151).[16] Inclusion

criteria were age 18 years or older, CRC 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 five years. See protocol pp 22-25 for exclusion criteria and other details. All patients received adjuvant therapy, either capecitabine plus bevacizumab or capecitabine alone, with equal disease-free and overall survival in both trial arms.[16]

Sample Preparation

Slides in VICTOR cohort were prepared in Oxford, UK, while the other slides in the 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 a 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 three μm FFPE tissue block section with haematoxylin and eosin (H&E), and a pathologist (MP) ascertained that it contained tumour. WSIs were acquired at the highest resolution available (referred to as 40x magnification by the manufacturers) 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 (protocol pp 6-10). A WSI with the so-called 40x resolution typically contained an order of 100,000x100,000 pixels, multiple orders of magnitude larger than images currently feasible for classification by deep learning methods. To preserve prognostic information contained at high-resolution, WSIs were partitioned into multiple non-overlapping image regions called *tiles* at 10x and 40x resolutions, where each

pixel at 40x represents a physical size of approximately 0·24x0·24 μm^2. Patients without tiles were excluded.

Classification

Five networks were trained on the 634,564 10x tiles and five networks on the 11,591,555 40x tiles from the 1652 Aperio AT2 and NanoZoomer XR WSIs 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 et al[22] (figure 1). Because of spatial heterogeneity, labelling a tile with the label of its WSI might be problematic. Instead, the networks were trained on labelled collections of tiles. A collection contained tiles from a single WSI, 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, i.e. directly from image to patient outcome, and each training iteration used a batch size of 32 collections with 64 tiles each. This many tiles were possible because we utilised a novel gradient approximation technique which substantially reduce memory usage during training (appendix pp 4-6). The Noisy-AND pooling function applied a trained non-linear function on tile representation averages. This 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 WSI.

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 (protocol pp 11-20). 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 10x and the 40x ensemble scores were determined by evaluations in the tuning cohort to define the ensemble classifiers (protocol pp 20-22). Furthermore, evaluations in the test cohort indicated that combining 10x and 40x markers might be desirable, and two such markers were defined, one continuous and one categorical. The continuous DoMore-v1-CRC score was defined as the average of the 10x and the 40x ensemble scores. The categorical DoMore-v1-CRC classifier assigned to good prognosis if both ensemble classifiers predicted good outcome, uncertain if the ensemble classifiers predicted differently, and poor prognosis if both predicted poor outcome. 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 (protocol pp 11-22), and tested as a secondary analysis (protocol p 27). While the DoMore-v1-CRC marker was trained using multiple instance learning, each single tile was labelled with the label of its WSI in training 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 (protocol pp 15-16).

Statistical Analysis

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

The pre-defined primary analysis for each scanner was univariable cancer-specific survival (CSS) analysis of the DoMore-v1-CRC classifier; for simplicity, we first present results for the Aperio AT2 scanner and in a separate paragraph address scanner differences. The classifier was included as the only variable in a Cox model to compute the hazard ratio (HR)

with 95% confidence interval (CI) of patients with uncertain and poor prognosis relative to patients with good prognosis. The proportional hazards assumption was found satisfactory fulfilled using log-log plots (appendix p 26). The Mantel-Cox log-rank test was used to assess whether the classifier predicted CSS.

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). To calculate classification metrics for 3-year CSS, patients without event and less than 3-year follow-up were excluded and events after 3 years were ignored. Category-free net reclassification improvement (NRI) was computed using the Kaplan-Meier estimates of five-year CSS. Two-sided $p<0{\cdot}05$ was considered statistically significant. The confidence level of CIs is 95%. The bias-corrected and accelerated bootstrap CI were computed for NRIs, c-indices and areas under the curves (AUCs) using 10,000 bootstrap replicates and an acceleration constant estimated using leave-one-out cross-validation. Time to CSS in the validation cohort was calculated from date of randomisation to date of cancer-specific death or loss to follow-up. Survival analyses were carried out in Stata/SE 15·1 (StataCorp, TX).

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

The DoMore-v1-CRC classifier was a strong predictor of CSS in the primary analysis of the validation cohort (HR for uncertain *vs* good prognosis, 1·89; CI, 1·14-3·15; HR for poor *vs* good prognosis, 3·84; CI, 2·72-5·43; figure 2A). The classifier remained strong in

multivariable analysis (HR for uncertain *vs* good prognosis, 1·56; CI, 0·92-2·65; HR for poor *vs* good prognosis, 3·04; CI, 2·07-4·47; 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% (CI, 41%-63%), specificity 78% (CI, 75%-81%), positive predictive value 19% (CI, 14%-25%), negative predictive value 94% (CI, 92%-96%), and correct classification rate 76% (CI, 73%-79%) when comparing 3-year CSS to good prognosis *vs* uncertain and poor prognosis. Compared to good and uncertain prognosis *vs* poor prognosis, the sensitivity was 69% (CI, 58%-78%), specificity 66% (CI, 63%-69%), positive predictive value 17% (CI, 13%-21%), negative predictive value 96% (CI, 94%-97%), and correct classification rate 67% (CI, 63%-69%).

The constituents of the DoMore-v1-CRC classifier, the 10x and the 40x 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 associated strongly with the patient outcome (appendix p 30), and provided a c-index of 0·674 (CI, 0·624-0·719; appendix p 16) in all validation patients and an AUC of 0·713 (CI, 0·624-0·789; appendix p 31) in patients with distinct outcome. The c-index and AUC of the 10x ensemble score were similar to the ones obtained for the DoMore-v1-CRC score (appendix pp 16 and 31).

The DoMore-v1-CRC classifier was a significant predictor of CSS in stage II (HR for poor *vs* good prognosis, 2·71; CI, 1·25-5·86; figure 2C) and stage III (HR for poor *vs* good prognosis, 4·09; CI, 2·77-6·03; figure 2D), and this was confirmed in multivariable analysis (table 2) and for the continuous score (appendix pp 9 and 13). The categorical marker identified patient

groups with substantially different CSS in stage IIIB and IIIC (appendix p 32), and was also significant in 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 five-year CSS was 61·6% (CI, 43·5%-79·3%); the event-NRI was 3·2% (CI, -13·2%-20·0%), and the non-event-NRI was 58·3% (CI, 52·7%-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). Of special interest is the relation to the histopathological grading into well, moderately, and poorly differentiated tumours. This was further studied in the test cohort where all gradings were centrally reviewed by one highly experienced pathologist (NAS).[18,19] Among 133 tumours characterised as well differentiated, the DoMore-v1-CRC classifier assigned 101 as good prognosis, 18 as uncertain and 14 as poor prognosis (appendix p 17). The moderately differentiated tumours were distributed fairly evenly over the DoMore-v1-CRC classes, while among 292 poorly differentiated tumours, the marker assigned 223 as poor prognosis, 36 as uncertain, and 33 as good prognosis. Thus, the DoMore-v1-CRC class was clearly associated to tumour differentiation. The large proportion of tumours classified as moderately differentiated (e.g. 53% [489 of 920] in the test cohort and 75% [846 of 1122] in the validation cohort) restricts the usefulness of this grading system, but also these patients could be risk stratified by the DoMore-v1-CRC marker (appendix p 34).

Median processing time per patient for the entire classification pipeline, i.e. from scan to predicted patient outcome, was 2·8 minutes (interquartile range, 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 CSS with only slightly worse performance than the DoMore-v1-CRC classifier (appendix pp 16 and 35-36).

In the test cohort with slides 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 using another scanner (NanoZoomer XR), the DoMore-v1-CRC score tended towards slightly higher values compared to when evaluated using 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; CI, 0·951-0·961), and the classifier provided similar prognostic information with both scanners (see appendix pp 9, 16, 18-25, and 39-51 for results with NanoZoomer XR). Thus, the classifier was also a strong predictor of CSS in the primary analysis of the validation cohort when evaluated on NanoZoomer XR slide images (HR for uncertain *vs* good prognosis, 2·42; CI, 1·45-4·03; HR for poor *vs* good prognosis, 3·39; CI, 2·36-4·87; 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 CRC which directly analyse standard H&E stained histological sections. 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, the good and poor prognosis groups included nearly 90% of the patients and differed about 4 times in HR for CSS in univariable analysis and about 3 times in 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] Recent studies have

suggested that deep learning could be used to develop markers which potentially utilise basic morphology to predict the outcome of cancer patients, but these findings have not been validated in independent cohorts.[27,28] We have not yet seen independently validated markers for directly predicting the outcome of cancer patients based on histological images.

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 WSI, while the DoMore-v1-CRC marker was developed using multiple instance learning to allow training on tile collections labelled with the label of its WSI. Both markers were strong predictors of CSS, but the DoMore-v1-CRC marker performed slightly better and was the marker pre-selected for independent validation in the QUASAR 2 cohort.

Automatic prognostication procedures reduce human intervention, and has the potential to increase reproducibility of biomarkers. New procedures like the DoMore-v1-CRC markers may initially be performed as services carried out at specialised laboratories with a high degree of standardisation of procedure to avoid disparities in sample handling, including the 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 may include standardisation modules and facilitate a more rapid spread of the automatic procedures. Moreover, supplemented by increased robotisation of wet-lab procedures, the higher analytic throughput will allow decisions based on multiple samples from a tumour. This may reduce the challenge of tumour heterogeneity, which may 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 most other markers in terms of HRs in stage-specific multivariable

analyses, on a par with pN staging. As opposed to the grading system, the classifier had few patients in the intermediate "uncertain" group.

The DoMore-v1-CRC classifier is technically simple to apply and can be delivered at pathology laboratories everywhere. 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 CRC on the pros and cons of different adjuvant treatment options. Although the number of drugs used in the adjuvant setting is limited to fluoropyrimidines ± oxaliplatin, recent data demonstrate that three months treatment achieves approximately the same survival outcomes as six months for the majority of stage III patients, while high risk patients (pT4 and pN2) might benefit from prolonged therapy.[29,30] It would be reasonable to hypothesise that stage III patients 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 + oxaliplatin + irinotecan as their high risk of cancer-specific death should positively skew the risk-benefit ratio of more aggressive treatments (figures 2D and F). At the other end, stage III patients with DoMore-v1-CRC good prognosis, the great majority of whom are pN1, have very good survival with single-agent capecitabine (figure 2E), and good prognosis stage II patients 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 randomising patients into observation, low intensity and high intensity regimes depending on relative risk score. However, the currently available data may 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 CRC following adjuvant treatment is remarkably consistent at 20%

across most well-designed clinical trials, thus translating into quite different absolute survival improvements for low and high risk subgroups.

Limitation of this study include 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 there are still challenges related to standardisation, as illustrated by the differences between the tested scanners. Differences between laboratories may also be seen for sample handling procedures, and this is why the introduction into the clinic is suggested to be through services performed at specialised laboratories. A well-known disadvantage of deep learning is its black-box nature. The DoMore-v1-CRC marker is related to histological grading, but the marker is still using small-scale features of the histological images with unknown biological correlates.

In summary, it has been possible to develop a clinically useful prognostic marker 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 over adjuvant treatment choices.

Contributors

OJS, 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 clinical/pathological data and interpretations. OJS, SDR, and JM performed the machine learning. AK performed the statistical analyses. OJS, SDR, AK, TSH, KL, DJK, and HED interpreted the data and analyses. All authors vouch for

the data, analyses, and interpretations. OJS, 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

OJS, 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.

Acknowledgements

We thank Akershus University Hospital for access to their patient material, National Institute for Health Research for funding support to Marco Novelli 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 last, but not least all participating patients for making this study possible.

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. QUASAR Collaborative Group. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet* 2007; **370**: 2020–29.

11. Andre 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. Andre 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. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2018: 4510–20.

22. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 2016; **32**: i52–i59.

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 U S A* 2018; **115**: E2970–E79.

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.

Figure Legends

Figure 1: **Pipeline of DoMore-v1-CRC classification**

Top: A whole-slide image (WSI) is segmented, and the segmented regions tiled at 40x resolution and 10x 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 10x and one for 40x. If the ensemble score is above a certain threshold, the WSI is classified as poor prognosis. The DoMore-v1-CRC class is determined by the agreement between the two ensemble classifications. Bottom: 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 WSI 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.

Figure 2: **Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in the QUASAR 2 validation cohort**

(A) The primary analysis; all patients evaluated with the pre-defined DoMore-v1-CRC classifier. (B) A post-hoc analysis; all patients evaluated with the DoMore-v1-CRC classifier variant with five categories. (C) A secondary analysis; stage II (equivalent to pN0) patients evaluated with the pre-defined DoMore-v1-CRC classifier. (D) A secondary analysis; stage III patients evaluated with the pre-defined DoMore-v1-CRC classifier. (E) A post-hoc analysis; pN1 patients evaluated with the pre-defined DoMore-v1-CRC classifier. (F) A post-hoc analysis; pN2 patients evaluated with the pre-defined DoMore-v1-CRC classifier.

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

	Group	Training cohort	Tuning cohort	Test cohort	Validation cohort
		(N=828)	**(N=1645)**	**(N=920)**	**(N=1122)**
Age, years		69 (61-75)	70 (61-77)	71 (64-78)	65 (59-71)
Sex					
	Female	402 (51%)	689 (42%)	421 (46%)	477 (43%)
	Male	426 (49%)	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 (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 (0%)	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 (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 (0%)	53 (5%)
Adjuvant treatment					
	No	467 (56%)	826 (50%)	538 (58%)	0 (0%)
	Chemotherapy	173 (21%)	397 (24%)	51 (6%)	1122 (100%)
	Radiotherapy	11 (1%)	6 (0%)	14 (2%)	0 (0%)
	Chemo- and radiotherapy	3 (0%)	9 (1%)	3 (0%)	0 (0%)
	Missing	174 (21%)	407 (25%)	314 (34%)	0 (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 number (%). IQR=interquartile range.

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

	Group	**Stage II and III**		**Stage II**		**Stage III**	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
DoMore-v1-CRC			<0·0001		0·028		0·0001
	Good prognosis	ref.		ref.		ref.	
	Uncertain	1·56 (0·92-2·65)		1·22 (0·35-4·24)		2·14 (1·15-3·99)	
	Poor prognosis	3·04 (2·07-4·47)		2·71 (1·25-5·86)		2·95 (1·81-4·82)	
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·84 (1·13-2·98)				ref.	
	pN2	5·94 (3·71-9·52)				3·31 (2·14-5·13)	
pT stage			0·0058				0·014
	pT1	NA				NA	
	pT2	1·86 (0·90-3·86)				1·68 (0·64-4·45)	
	pT3	ref.				ref.	
	pT4	1·75 (1·22-2·51)				2·07 (1·33-3·22)	
Lymphatic invasion	Yes	1·66 (1·07-2·56)	0·023			1·98 (1·20-3·28)	0·0079
Venous vascular invasion	Yes	1·07 (0·76-1·51)	0·71			0·98 (0·64-1·52)	0·94
Sidedness	Right					1·09 (0·70-1·70)	0·69
BRAF	Mutated					1·39 (0·81-2·40)	0·24

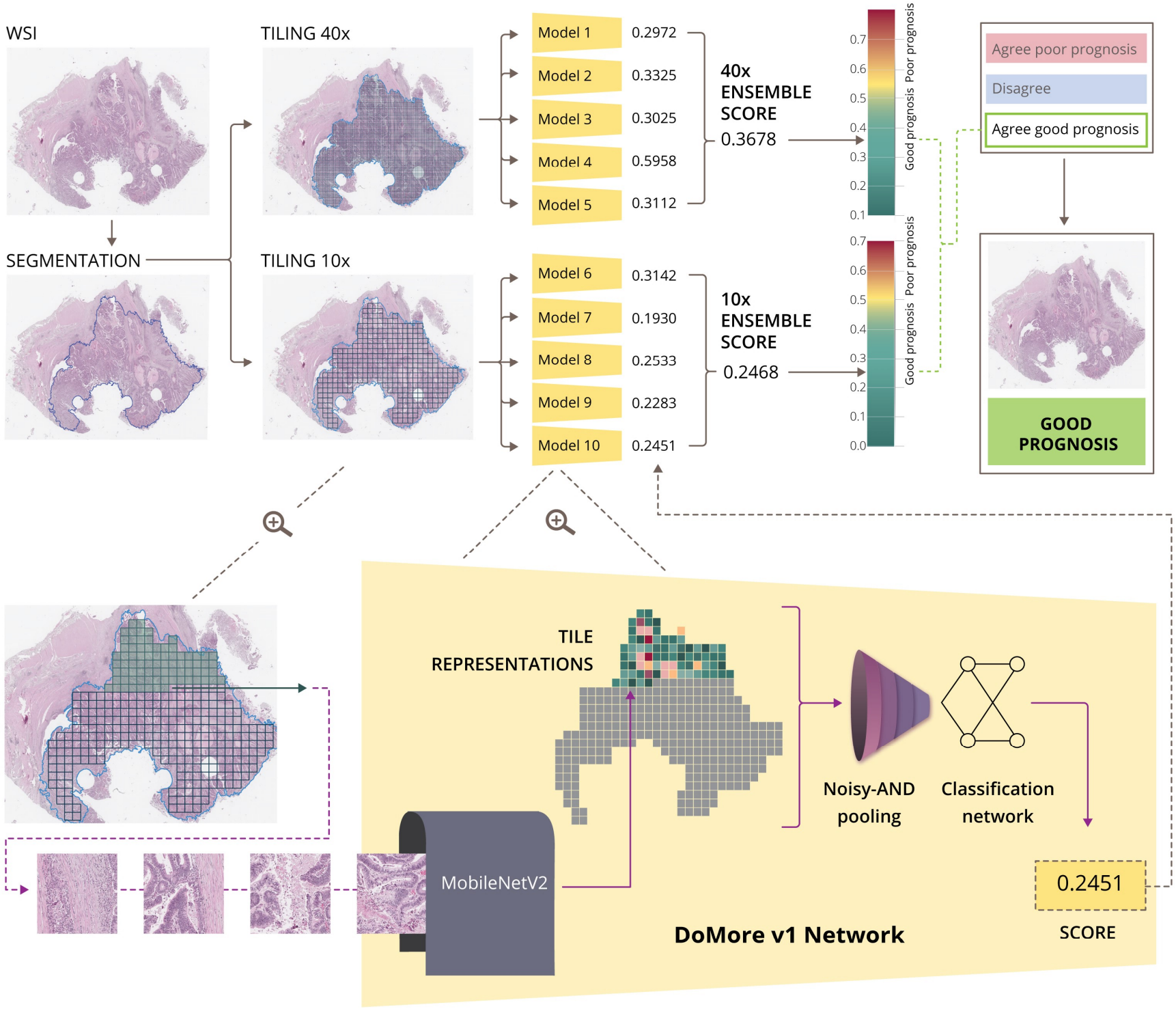
Ref.=reference; NA=not available

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

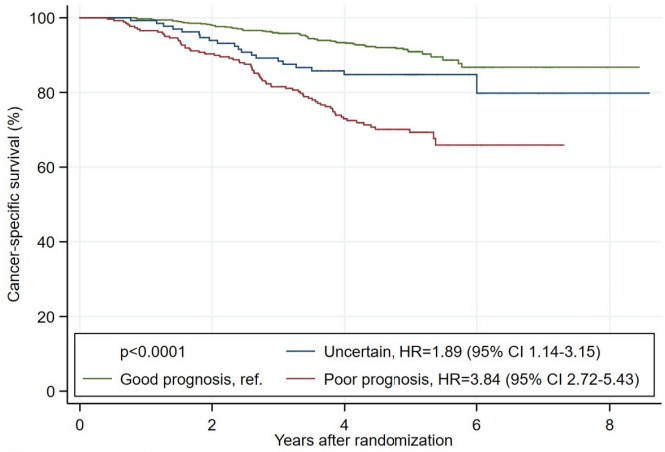
	Group	DoMore-v1-CRC good prognosis	DoMore-v1-CRC uncertain	DoMore-v1-CRC poor prognosis	Spearman's correlation	
		(N=704)	(N=136)	(N=270)	ρ (95% CI)	p
Age (continuous), years		64 (58-71)	65 (60-71)	66 (60-72)	0·07 (0·01 to 0·13)	0·024
Age (dichotomous), years					0·03 (-0·03 to 0·09)	0·38
	≤72	568 (81%)	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 (21%)	19 (14%)	28 (11%)		
	IIB	110 (16%)	27 (20%)	54 (21%)		
	IIIA	67 (10%)	2 (2%)	6 (2%)		
	IIIB	269 (40%)	51 (38%)	104 (41%)		
	IIIC	83 (12%)	34 (26%)	64 (25%)		
pN stage					0·10 (0·04 to 0·16)	0·0008
	pN0	261 (38%)	48 (36%)	88 (33%)		
	pN1	339 (50%)	53 (39%)	111 (42%)		
	pN2	83 (12%)	34 (25%)	64 (24%)		
pT stage					0·26 (0·21 to 0·32)	<0·0001
	pT1	15 (2%)	0 (0%)	2 (1%)		
	pT2	61 (9%)	3 (2%)	6 (2%)		
	pT3	402 (60%)	75 (56%)	100 (39%)		
	pT4	194 (29%)	56 (42%)	148 (58%)		
Lymphatic invasion					0·04 (-0·02 to 0·10)	0·20
	No	599 (91%)	122 (92%)	220 (87%)		
	Yes	62 (9%)	10 (8%)	33 (13%)		
Venous vascular invasion					0·05 (-0·01 to 0·11)	0·11
	No	409 (61%)	74 (56%)	145 (56%)		
	Yes	257 (39%)	58 (44%)	112 (44%)		
Histological grade					0·14 (0·08 to 0·20)	<0·0001
	1	27 (4%)	7 (6%)	8 (3%)		
	2	565 (85%)	88 (69%)	186 (74%)		
	3	76 (11%)	32 (25%)	59 (23%)		
Location					0·15 (0·09 to 0·21)	<0·0001
	Rectum	118 (18%)	21 (16%)	23 (9%)		
	Distal colon	301 (45%)	46 (35%)	100 (38%)		
	Proximal colon	246 (37%)	64 (49%)	138 (53%)		
Sidedness					0·14 (0·08 to 0·20)	<0·0001
	Left	419 (63%)	67 (51%)	123 (47%)		
	Right	246 (37%)	64 (49%)	138 (53%)		
KRAS					-0·06 (-0·12 to 0·00)	0·069
	Wild-type	410 (65%)	86 (73%)	169 (70%)		
	Mutated	224 (35%)	32 (27%)	73 (30%)		
BRAF					0·22 (0·16 to 0·28)	<0·0001
	Wild-type	588 (93%)	89 (75%)	190 (77%)		

	Mutated	47 (7%)	29 (25%)	56 (23%)		
Microsatellite instability					-0·10 (-0·16 to -0·04)	0·0018
	Yes	66 (10%)	26 (21%)	40 (16%)		
	No	595 (90%)	99 (79%)	213 (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)	0·0006

Data are median (IQR) or number (%). IQR=interquartile range.


WSI
TILING 40x
SEGMENTATION
TILING 10x
Model 1
0.2972
Model 2
0.3325
Model 3
0.3025
Model 4
0.5958
Model 5
0.3112
40x ENSEMBLE SCORE
0.3678
Model 6
0.3142
Model 7
0.1930
Model 8
0.2533
Model 9
0.2283
Model 10
0.2451
10x ENSEMBLE SCORE
0.2468
Poor prognosis
Good prognosis
Agree poor prognosis
Disagree
Agree good prognosis
GOOD PROGNOSIS
TILE REPRESENTATIONS
MobileNetV2
Noisy-AND pooling
Classification network
0.2451
SCORE
DoMore v1 Network

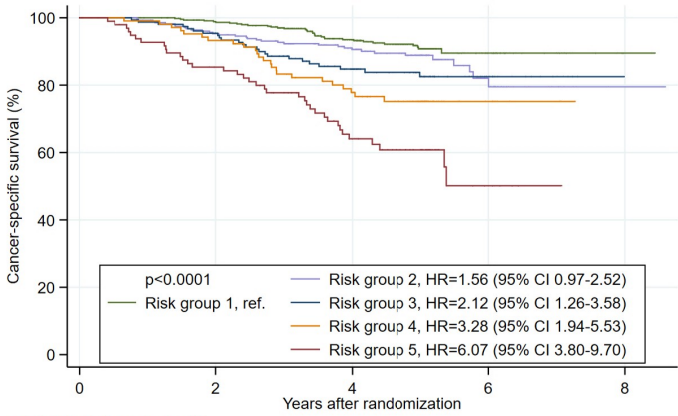
A



Number at risk (number of events)

Good prognosis	704	(13)	659	(29)	492	(15)	71	(0)	1
Uncertain	136	(8)	121	(11)	87	(0)	17	(1)	1
Poor prognosis	270	(25)	228	(41)	151	(8)	26	(0)	0

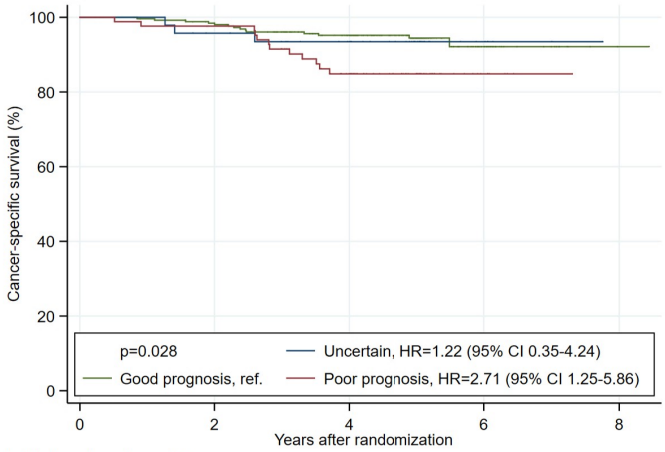
B



Number at risk (number of events)

Risk group 1	458	(5)	431	(21)	319	(8)	46	(0)	1
Risk group 2	286	(13)	262	(12)	196	(7)	32	(1)	1
Risk group 3	156	(7)	141	(15)	100	(2)	21	(0)	0
Risk group 4	108	(7)	95	(15)	70	(2)	8	(0)	0
Risk group 5	102	(14)	79	(18)	45	(4)	7	(0)	0

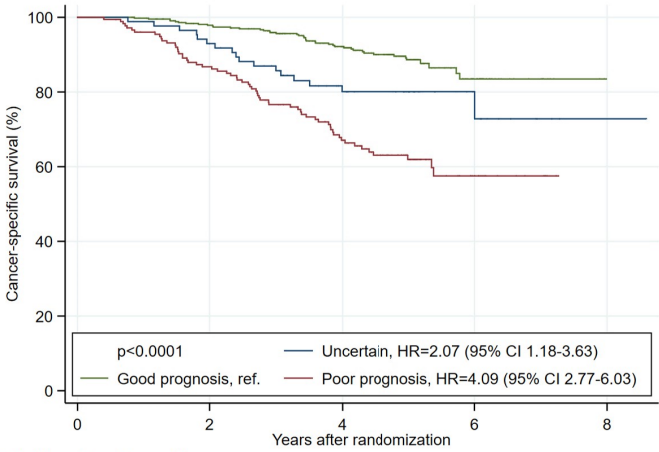
C



Number at risk (number of events)

Good prognosis	261	(4)	250	(8)	202	(2)	28	(0)	1
Uncertain	48	(2)	43	(1)	35	(0)	6	(0)	0
Poor prognosis	88	(2)	80	(10)	57	(0)	7	(0)	0

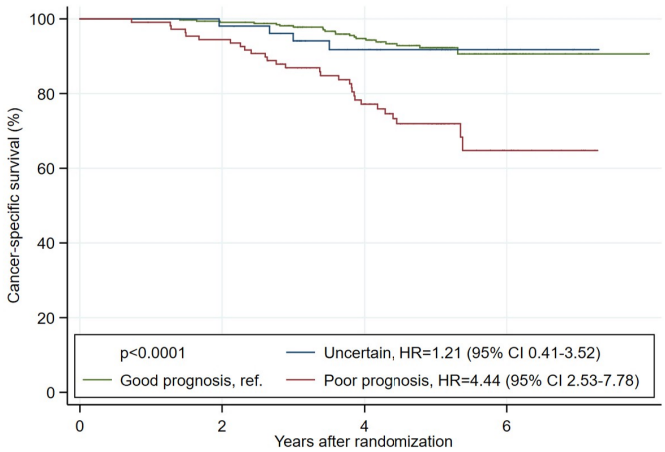
D



Number at risk (number of events)

Good prognosis	443	(9)	409	(21)	290	(13)	43	(0)	0
Uncertain	88	(6)	78	(10)	52	(0)	11	(1)	1
Poor prognosis	182	(23)	148	(31)	94	(8)	19	(0)	0

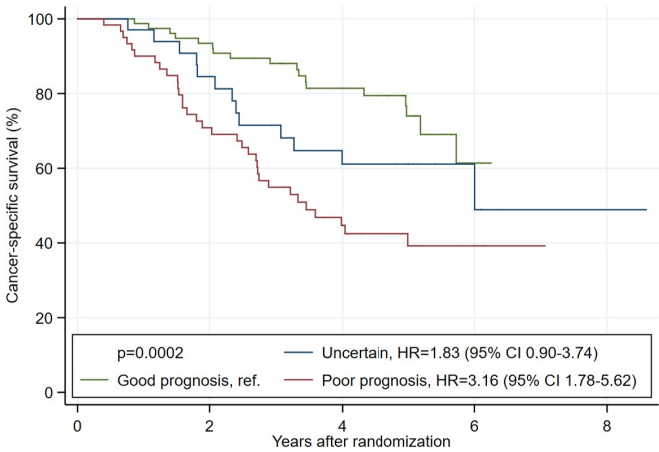
E



Number at risk (number of events)

Good prognosis	339	(3)	319	(12)	227	(6)	34
Uncertain	53	(1)	51	(3)	35	(0)	6
Poor prognosis	111	(6)	102	(17)	68	(6)	12

F



Number at risk (number of events)

Good prognosis	83	(5)	70	(8)	45	(5)	6	(0)	0
Uncertain	34	(5)	26	(7)	17	(0)	5	(1)	1
Poor prognosis	64	(17)	40	(14)	21	(2)	6	(0)	0