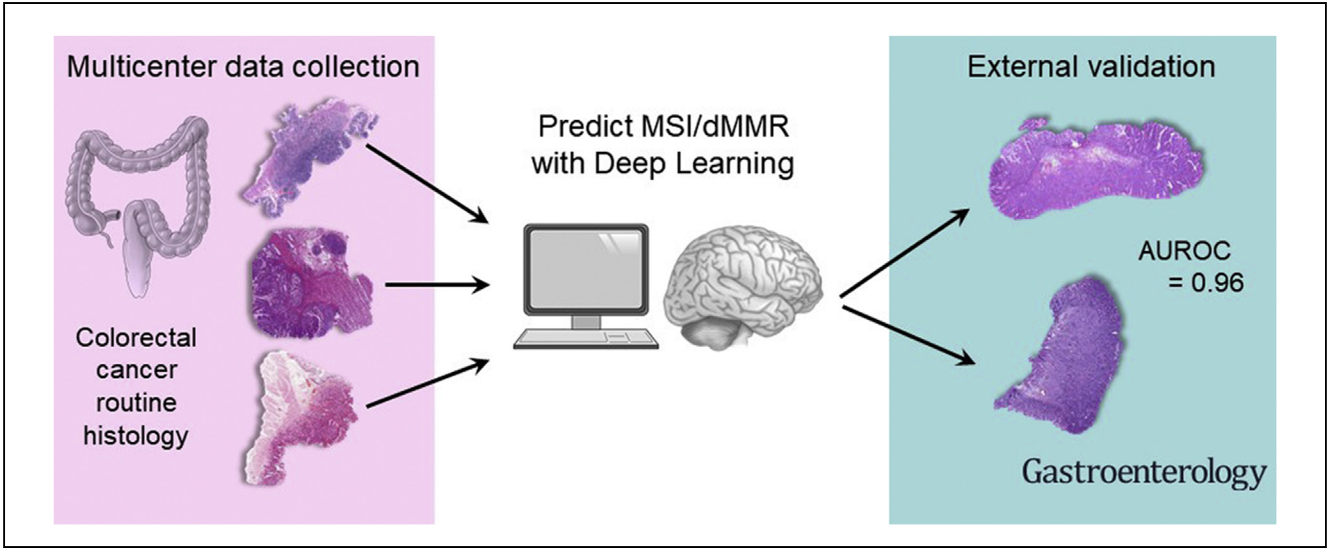
Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning

Amelie Echle,[1] Heike Irmgard Grabsch,[2,3] Philip Quirke,[3] Piet A. van den Brandt,[4] Nicholas P. West,[3] Gordon G. A. Hutchins,[3] Lara R. Heij,[5,6,7] Xiuxiang Tan,[5,6,7] Susan D. Richman,[3] Jeremias Krause,[1] Elizabeth Alwers,[8] Josien Jenniskens,[4] Kelly Offermans,[4] Richard Gray,[9] Hermann Brenner,[8,10,11] Jenny Chang-Claude,[12,13] Christian Trautwein,[1] Alexander T. Pearson,[14] Peter Boor,[7] Tom Luedde,[1,15] Nadine Therese Gaisa,[7] Michael Hoffmeister,[8] and Jakob Nikolas Kather[1,3,11,16]

[1]Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany; [2]Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands; [3]Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom; [4]Department of Epidemiology, Maastricht University Medical Center+, Maastricht, The Netherlands; [5]Visceral and Transplant Surgery, University Hospital Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany; [6]NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands; [7]Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany; [8]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany; [9]Clinical Trial Service Unit, University of Oxford, Oxford, United Kingdom; [10]Division of Preventive Oncology, German Cancer Research Center and National Center for Tumor Diseases, Heidelberg, Germany; [11]German Cancer Consortium, German Cancer Research Center, Heidelberg, Germany; [12]Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany; [13]Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [14]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois; [15]Division of Gastroenterology, Hepatology, and Hepatobiliary Oncology, Aachen, Germany; and [16]Medical Oncology, National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg, Germany


Multicenter data collection
Colorectal cancer routine histology
Predict MSI/dMMR with Deep Learning
External validation
AUROC = 0.96
Gastroenterology

See editorial on page 1235.

BACKGROUND & AIMS: Microsatellite instability (MSI) and mismatch-repair deficiency (dMMR) in colorectal tumors are used to select treatment for patients. Deep learning can detect MSI and dMMR in tumor samples on routine histology slides faster and less expensively than molecular assays. However, clinical application of this technology requires high performance and multisite validation, which have not yet been performed. **METHODS:** We collected H&E-stained slides and findings from molecular analyses for MSI and dMMR from 8836 colorectal tumors (of all stages) included in the MSI-DETECT consortium study, from Germany, the Netherlands, the United Kingdom, and the United States. Specimens with dMMR were identified by immunohistochemistry analyses of tissue microarrays for loss of MLH1, MSH2, MSH6, and/or PMS2. Specimens with MSI were identified by genetic analyses. We trained a deep-learning detector to identify samples with MSI from these slides; performance was assessed by cross-validation (N = 6406 specimens) and validated in an external cohort (n = 771 specimens). Prespecified endpoints were area under the receiver operating characteristic (AUROC) curve and area under the precision-recall curve

(AUPRC). **RESULTS:** The deep-learning detector identified specimens with dMMR or MSI with a mean AUROC curve of 0.92 (lower bound, 0.91; upper bound, 0.93) and an AUPRC of 0.63 (range, 0.59–0.65), or 67% specificity and 95% sensitivity, in the cross-validation development cohort. In the validation cohort, the classifier identified samples with dMMR with an AUROC of 0.95 (range, 0.92–0.96) without image preprocessing and an AUROC of 0.96 (range, 0.93–0.98) after color normalization. **CONCLUSIONS:** We developed a deep-learning system that detects colorectal cancer specimens with dMMR or MSI using H&E-stained slides; it detected tissues with dMMR with an AUROC of 0.96 in a large, international validation cohort. This system might be used for high-throughput, low-cost evaluation of colorectal tissue specimens.

Keywords: biomarker; cancer immunotherapy; Lynch syndrome; mutation.

Mismatch repair deficiency (dMMR) is observed in 10% to 20% of patients with colorectal cancer (CRC) and indicates a biologically distinct type of CRC with broad prognostic, predictive, and therapeutic relevance.[1] In CRC and other cancer types, dMMR causes microsatellite instability (MSI), a specific DNA damage pattern. MSI and dMMR are associated with lack of chemotherapy response in intermediate stage CRC (pT3–4 N0–2), a reduced incidence of locoregional metastases, and hence the opportunity of cure by local excision in early-stage disease and a reduced requirement for adjuvant chemotherapy in stage II disease. In late-stage disease, MSI and dMMR are predictive of response to immune checkpoint inhibition and constitute the only clinically approved pancancer biomarker for checkpoint inhibition in the United States.[2] Furthermore, MSI and dMMR are the genetic mechanism driving carcinogenesis in Lynch syndrome (LS), the most common hereditary condition leading to CRC.[3] Because of this broad clinical importance, MSI or dMMR testing is recommended for all patients with CRC by national and international guidelines such as the British National Institute for Health and Care Excellence (NICE) guideline[4] and the European Society for Medical Oncology guidelines.[5] However, in clinical practice, only a subset of patients with CRC is investigated for presence of MSI or dMMR because of the high costs associated with universal testing. This lack of testing potentially leads to overtreatment with adjuvant chemotherapy; underdiagnosis of LS; reduced opportunities to consider local excision instead of extensive surgery, with related risks; and morbidity and failure to identify candidates for cancer immunotherapy.

Current laboratory assays for MSI and dMMR testing involve a multiplex polymerase chain reaction (PCR) assay or a multiplex immunohistochemistry (IHC) panel. Specifically, MSI can be tested by the Bethesda panel PCR,[6] whereas a 4-plex IHC can show absence of 1 of 4 mismatch-repair enzymes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*).[7] However, both assays for MSI or dMMR incur cost,[8] require additional sections of tumor tissue in addition to routine H&E histology,[9] and yield imperfect results. The sensitivity and specificity of these tests have been evaluated in numerous population-based studies, which are summarized in current clinical guidelines.[10] In these reference studies, test performance of molecular assays is reported with a sensitivity of 100% and specificity of 61.1%[11] or a higher specificity of 92.5% with a lower sensitivity of 66.7%[12] for MSI testing. Similarly, for IHC-based tests, sensitivity is reported as 85.7% with a 91.9% specificity in a key study,[13] whereas other international guidelines estimate that IHC testing has a sensitivity of 94% and a specificity of 88%.[5] This variable performance of clinical criterion standard tests indicates that there is need for improvement. In addition, all available tests incur a substantial cost and require specialized molecular pathology laboratories. This highlights the need for new robust, low-cost, and ubiquitously applicable diagnostic assays for MSI or dMMR detection in patients with CRC.

In routine H&E histologic images, MSI and dMMR tumors are characterized by distinct morphologic patterns such as tumor-infiltrating lymphocytes, mucinous differentiation, heterogeneous morphology, and a poor differentiation.[14] Although these patterns are well known to pathologists,

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Microsatellite instability (MSI) and mismatch-repair deficiency (dMMR) in colorectal tumors are used to select treatment for patients. Deep learning can detect MSI and dMMR in tumor samples on routine histology slides faster and cheaper than molecular assays.

NEW FINDINGS

We developed a deep-learning system that detects colorectal tumor specimens with MSI using hematoxylin and eosin-stained slides; it detected tissues with MSI with an area under the receiver operating characteristic curve of 0.96 in a large, international validation cohort.

LIMITATIONS

This system requires further validation before it can be used routinely in the clinic.

IMPACT

This system might be used for high-throughput, low-cost evaluation of colorectal tissue specimens.

Abbreviations used in this paper: **AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating curve; CRC, colorectal cancer; DACHS, Darmkrebs: Chancen der Verhütung durch Screening; dMMR, deficient mismatch repair; IHC, immunohistochemistry; LS, Lynch syndrome; MSI, microsatellite instability; NLCS, Netherlands Cohort Study; PCR, polymerase chain reaction; pMMR, proficient mismatch repair; QUASAR, Quick and Simple and Reliable; TCGA, The Cancer Genome Atlas Network; YCR-BCIP, Yorkshire Cancer Research Bowel Cancer Improvement Programme.**

Most current article

0016-5085
https://doi.org/10.1053/j.gastro.2020.06.021

manual quantification of these features by experts is not reliable enough for clinical diagnosis and, therefore, is not feasible in routine clinical practice.[15] In contrast, computer-based image analysis by deep learning has enabled robust detection of MSI and dMMR status directly from routine H&E histology: we recently presented[16] and later refined[17] such a deep learning assay, which was independently validated by 2 other groups.[18,19] However, all of these studies used a few hundred patients with CRC at most, but clinical implementation of a deep learning–based diagnostic assay requires enhanced sensitivity and specificity to those previously reported and large-scale validation across multiple populations in different countries.

To address this, we formed the MSIDETECT consortium: a group of multiple academic medical centers across and beyond Europe (http://www.msidetect.eu). In this not-for-profit consortium, we collected tumor samples from more than 8000 patients with molecular annotation. The prespecified intent was to train and externally validate a deep learning system for MSI and dMMR detection in CRC. The primary endpoint was diagnostic accuracy measured by area under the receiver operating curve (AUROC), area under the precision-recall curve (AUPRC), and, correspondingly, specificity at multiple sensitivity levels (99%, 98%, and 95%).

Materials and Methods

Ethics Statement and Patient Cohorts

We retrospectively collected anonymized H&E-stained tissue slides of patients with colorectal adenocarcinoma from multiple previous studies and population registers. For each patient, at least 1 histologic slide was available, and MSI status or MMR status was known. We included patients from the following 4 previous studies with the intent of retraining a previously described deep learning system.[16,17] First, we used the publicly available Cancer Genome Atlas (TCGA) ($n = 616$ patients) (Supplementary Figure 1), a multicenter study with patients with stage I–IV disease, mainly from the United States.[20] All images and data from the TCGA study are publicly available at https://portal.gdc.cancer.gov. Second, we used Darmkrebs: Chancen der Verhütung durch Screening (DACHS) ($n = 2292$) (Supplementary Figure 2), a population-based study of patients with stage I–IV CRC from southwestern Germany.[21] Tissue samples from the DACHS study were provided by the Tissue Bank of the National Center for Tumor Diseases (Heidelberg, Germany) in accordance with the regulations of the tissue bank and the approval of the ethics committee of Heidelberg University.[21,22] Third, we used samples from the Quick and Simple and Reliable trial (QUASAR) ($n = 2206$) (Supplementary Figure 3), which originally aimed to determine survival benefit from adjuvant chemotherapy in patients from the United Kingdom with mainly stage II tumors.[23] Finally, the Netherlands Cohort Study (NLCS) ($n = 2197$) (Supplementary Figure 4)[24,25] collected tissue samples as part of the Rainbow-Tissue Microarray consortium, and like DACHS, this study included patients with any tumor stage. All studies were cleared by the institutional ethics board of the respective institutions, as described before (for QUASAR,[23] DACHS,[22] and NLCS[25]).

With the intent of external validation of the deep learning system, we collected H&E slides from the population-based Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR-BCIP)[26] cohort, where routine National Health Service diagnosis of dMMR was undertaken with further *BRAF* mutation and/or *hMLH1* methylation screening to identify patients at high risk of having LS. The primary validation cohort from YCR-BCIP contained $n = 771$ patients with standard histology after surgical resection (YCR-BCIP-RESECT) (Supplementary Figure 5). For an additional exploratory analysis, we also acquired a nonoverlapping set of $n = 1531$ patients from YCR-BCIP with endoscopic biopsy samples (YCR-BCIP-BIOPSY) (Supplementary Figure 6). A set of $n = 128$ polypectomy samples from the YCR-BCIP study (YCR-BCIP-BIOPSY) contained only $n = 4$ MSI or dMMR patients and was not used for further analyses because AUROC and AUCPR values are not meaningful for such low prevalence features. For all patient samples in YCR-BCIP,[26] a fully anonymized, single scanned image of a representative H&E slide for each patient was used as a service evaluation study with no access to tissue or patient data aside from mismatch repair status.

Available clinicopathologic characteristics of all cases in each cohort are summarized in Supplementary Table 1. MSI status in the TCGA study was determined genetically as described before.[20] MSI status in the DACHS study was determined genetically with a 3-plex panel as described before.[27] In the QUASAR, NLCS, and YCR-BCIP cohorts, mismatch-repair deficiency (dMMR) or proficiency (pMMR) was determined with a standard immunohistochemistry assays on tissue microarrays as described before (2-plex for *MLH1* and *MSH2* in NLCS and QUASAR, 4-plex for *MLH1*, *MSH2*, *MSH6* and *PMS2* for YCR-BCIP).[23] This study complies with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement as shown in Supplementary Table 2.

Image Preprocessing and Deep Learning

All slides were individually, manually reviewed by trained observers supervised by expert pathologists to ensure that tumor tissue was present on the slide and the slide had diagnostic quality. Observers and supervisors were blinded regarding MSI status and any other clinical information. Tumor tissue was manually outlined in each slide. A small number of cases were excluded because of insufficient quality, technical issues, absence of tumor tissue on the observed slide, or lack of molecular information (Supplementary Figures 1–6). Tumor regions were tessellated into square tiles of 256-μm edge length and saved at a resolution of 0.5 μm per pixel using QuPath, version 0.1.2.[28] Initially, the method pipeline was kept as simple as possible, and color normalization was not used to preprocess the images. In a slight variation of the initial experiments, all image tiles were color normalized with the Macenko method[29] as described previously.[30] A modified shufflenet deep learning system with a 512 × 512 input layer was trained on these image tiles in MATLAB R2019a (MathWorks, Natick, MA) with the hyperparameters listed in Supplementary Table 3, as described before.[17] Tile-level predictions were averaged on the patient level, with the proportion of predicted MSI or dMMR tiles (positive threshold) being the free parameter for the receiver operating characteristic analysis. All confidence intervals were obtained by 10-fold bootstrapping. No image tiles or slides from the same patient were ever part of the training set and test set. All trained deep learning classifiers were assigned a unique identifier as listed in

Supplementary Table 4. All classifiers can be downloaded at https://dx.doi.org/10.5281/zenodo.3627523. Source codes are publicly available at https://github.com/jnkather/DeepHistology.

Experimental Design

All deep learning experiments (training and test runs) were prespecified and are listed in Supplementary Table 5. All patients from TCGA, DACHS, QUASAR, and NLCS were combined and served as the training set (the international cohort). To assess the magnitude of batch effects, we trained a deep learning system on each subcohort in this international training cohort, assessing intercohort and intracohort performance, the latter being estimated by 3-fold cross-validation (experiment 1). In addition, we performed a 3-fold cross-validation on the full international cohort without (experiment 2) and with color normalization (experiment 2N), which was used for a detailed subgroup analysis according to predefined clinicopathologic and molecular subgroups. To identify the optimal number of patients needed for training, we used the international cohort, randomly set aside n = 906 patients for testing, and trained on increasing proportions of the remaining n = 5500 patients (experiment 3). To evaluate the deep learning system in an independent, external, population-based cohort, we trained on the international cohort and tested on YCR-BCIP-RESECT (experiment 4; this was the primary objective of our study). This experiment was repeated with color-normalized image tiles (experiment 4N). YCR-BCIP-RESECT was regarded as the "holy" test set and was not used for any other purpose than to evaluate the final classifier. Exploratively, we also evaluated the final classifier on YCR-BCIP-BIOPSY (experiment 5). Furthermore, to investigate the performance "train-on-biopsy, test-on-biopsy," we exploratively trained a 3-fold cross-validated classifier on YCR-BCIP-BIOPSY (experiment 6).

Results

Deep Learning Consistently Predicts Microsatellite Instability in Multiple Patient Cohorts

In the MSIDETECT consortium, a deep learning system was trained to predict MSI or dMMR status from digitized routine H&E whole slide images alone, with ground truth labels according to local standard procedures (PCR testing for MSI or IHC testing for dMMR). First, we investigated deep learning classifier performance in patients of the TCGA, DACHS, QUASAR, and NLCS cohorts alone. We found that training the deep learning system on individual cohorts yielded an intracohort AUROC of 0.74 (0.66–0.80) in the TCGA cohort (n = 426), 0.89 (0.86–0.91) in the QUASAR cohort (n = 1770), 0.92 (0.91–0.94) in the DACHS cohort (n = 2013), and 0.89 (0.88–0.92) in the NLCS cohort (n = 2197 patients) (Supplementary Table 6). This high intracohort performance dropped in some intercohort experiments (Table 1 and experiment 1 in Supplementary Table 5). Together, these data show that deep learning systems attain high diagnostic accuracy in single-center cohorts but do not necessarily generalize to other patient cohorts.

Increasing Patient Number Compensates for Batch Effects and Improves Performance

In the intracohort experiments (Table 1), training on larger cohorts generally yielded higher performance, corroborating the theoretical assumption that training on larger data sets yields more robust classifiers. To quantify this effect, we merged all patients from TCGA, DACHS, QUASAR, and NLCS into a large international cohort (N = 6406 patients) (Figure 1*A*). From these digitized whole slide histology images, we created a library of image tiles for training deep learning classifiers (Figure 1*B*). Thus, we increased the patient number as well as the data heterogeneity due to different preanalytic pipelines in the respective medical centers. We set aside a randomly chosen proportion of n = 906 of these patients and retrained deep learning classifiers on 500, 1000, 1500, and so on, up to 5500 patients of the international cohort. In this experiment, we found that AUROC (Figure 1*C*) and AUPRC (Supplementary Figure 7) on the test set initially increased as the number of patients in the training set increased. However, each increase in patient number yielded diminishing performance returns, and AUROC and AUPRC plateaued at approximately 5000 patients (Figure 1*D*). The top performance was achieved by training on 5500 patients and testing on the fixed test set of n = 906 patients, with an AUROC of 0.92 (0.90–0.93) (compared to a baseline of 0.5 by a random model) (Figure 1*C*), an AUPRC of 0.59 (0.4–0.63) (compared to a baseline of 0.12 in a random model) (Supplementary Figure 7 and experiment 3 in Supplementary Table 5), translating to a specificity of 52% at a sensitivity of 98%. To ensure that this performance was

Table 1. Estimating Batch Effects by Analyzing Intracohort and Intercohort Performance in all Subcohorts in the International Cohort

	Train on TCGA n = 426 15% MSI	Train on QUASAR n = 1770 14% dMMR	Train on DACHS n = 2013 14% MSI	Train on NLCS n = 2197 10% dMMR
Test on TCGA (United States)	0.74 (0.66–0.80)	0.76 (0.70–0.79)	0.77 (0.73–0.79)	0.72 (0.71–0.78)
Test on QUASAR (United Kingdom)	0.67 (0.64–0.68)	0.89 (0.86–0.91)	0.71 (0.68–0.75)	0.76 (0.73–0.78)
Test on DACHS (Germany)	0.81 (0.79–0.83)	0.68 (0.65–0.72)	0.92 (0.91–0.94)	0.80 (0.78–0.82)
Test on NLCS (The Netherlands)	0.77 (0.74–0.79)	0.80 (0.78–0.81)	0.82 (0.79–0.83)	0.90 (0.89–0.91)

NOTE. Main performance measure AUROC, shown as mean with lower and upper bounds in a 10-fold bootstrapped experiment. Intracohort performance was estimated by 3-fold cross-validation.

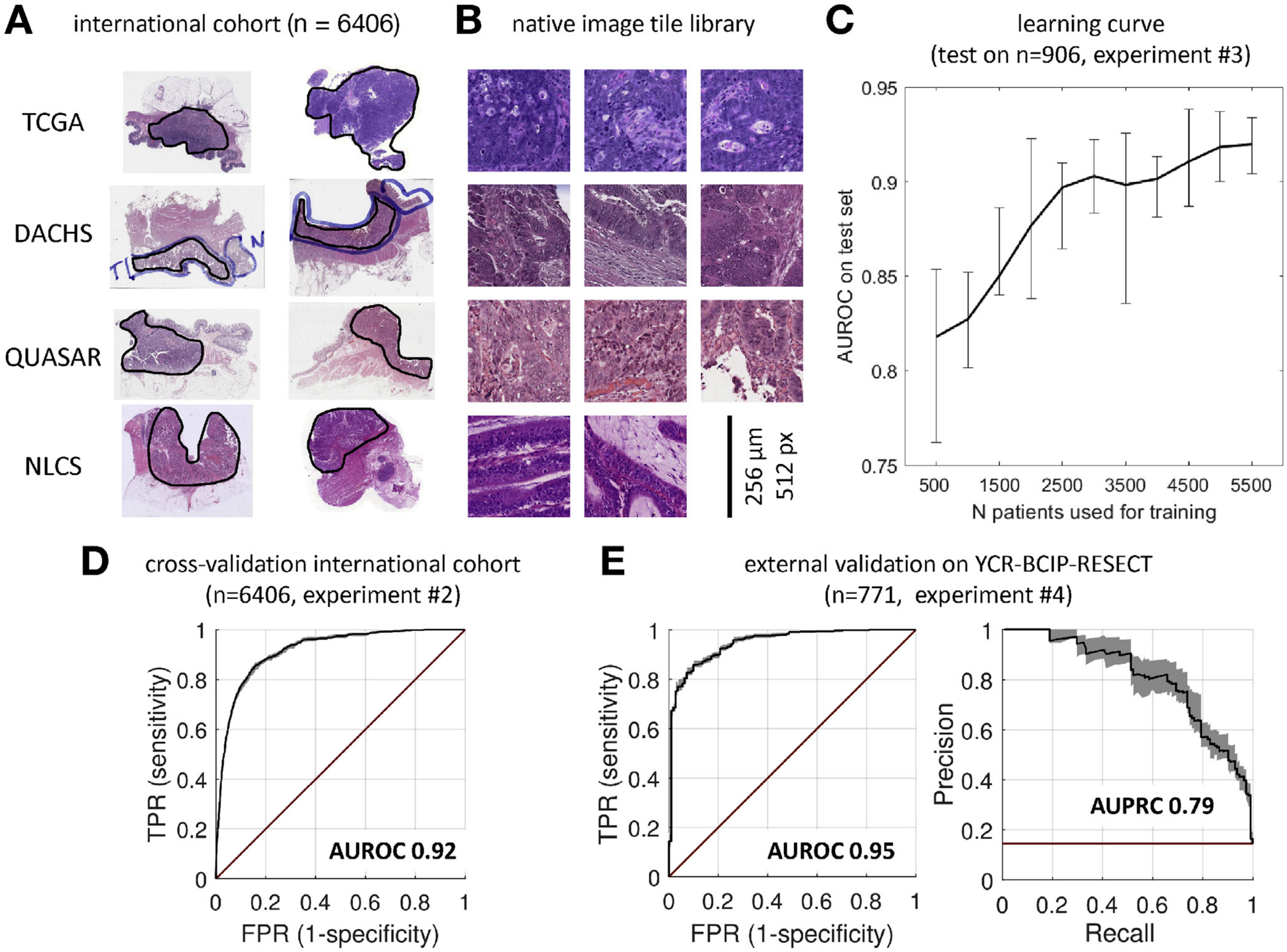


Figure 1. Deep learning workflow and learning curves. (*A*) Histologic routine images were collected from 4 large patient cohorts. All slides were manually quality checked to ensure the presence of tumor tissue (*outlined in black*). (*B*) Tumor regions were automatically tessellated, and a library of millions of nonnormalized (native) image tiles was created. (*C*) The deep learning system was trained on increasing numbers of patients and evaluated on a random subset (n = 906 patients). Performance initially increased by adding more patients to the training set but reached a plateau at approximately 5000 patients. (*D*) Cross-validated experiment on the full international cohort (comprising TCGA, DACHS, QUASAR, and NLCS). Receiver operating characteristic (ROC) with true positive rate shown against false positive rate AUROC is shown on top. (*E*) ROC curve (*left*) and precision-recall curve (*right*) of the same classifier applied to a large external data set. High test performance was maintained in this data set, and thus, the classifier generalized well beyond the training cohorts. The black line indicates average performance, the shaded area indicates bootstrapped confidence interval, and the red line indicates random model (no skill). FPR, false positive rate; TPR, true positive rate;

not due to the random selection of the internal test set, we performed a patient-level 3-fold cross-validation on the full international cohort (N = 6406), reaching a similar mean AUROC of 0.92 (0.91–0.93((Figure 1*D* and experiment 2 in Supplementary Table 5). Together, these data show that approximately 5000 patients are necessary and sufficient to train a high-quality deep learning detector of MSI and dMMR.

Clinical-Grade Performance in an External Test Cohort

Deep learning systems are prone to overfit to the data set they were trained on and, thus, must be validated in external test sets. Correspondingly, the prespecified primary endpoint of this study was the test performance in a completely independent set of patients. This set of patients was intended to be population-based, that is, to mirror the clinicopathologic characteristics of a real-world screening population. It was used for no other purpose than to validate the final classifier, which was previously trained on the international cohort. The test set comprised routine H&E slides from the population-based YCR-BCIP study (YCR-BCIP-RESECT, n = 771 patients, 1 slide per patient). In this population, we found a high classification performance with a mean AUROC of 0.95 and (0.92–0.96) lower and upper bootstrapped confidence bounds, respectively (Figure 1*E* and Supplementary Table 6, experiment 4). Because the target feature MSI and dMMR are unbalanced in real-world populations such as YCR-BCIP-RESECT, we also assessed the precision-recall characteristics of this test,

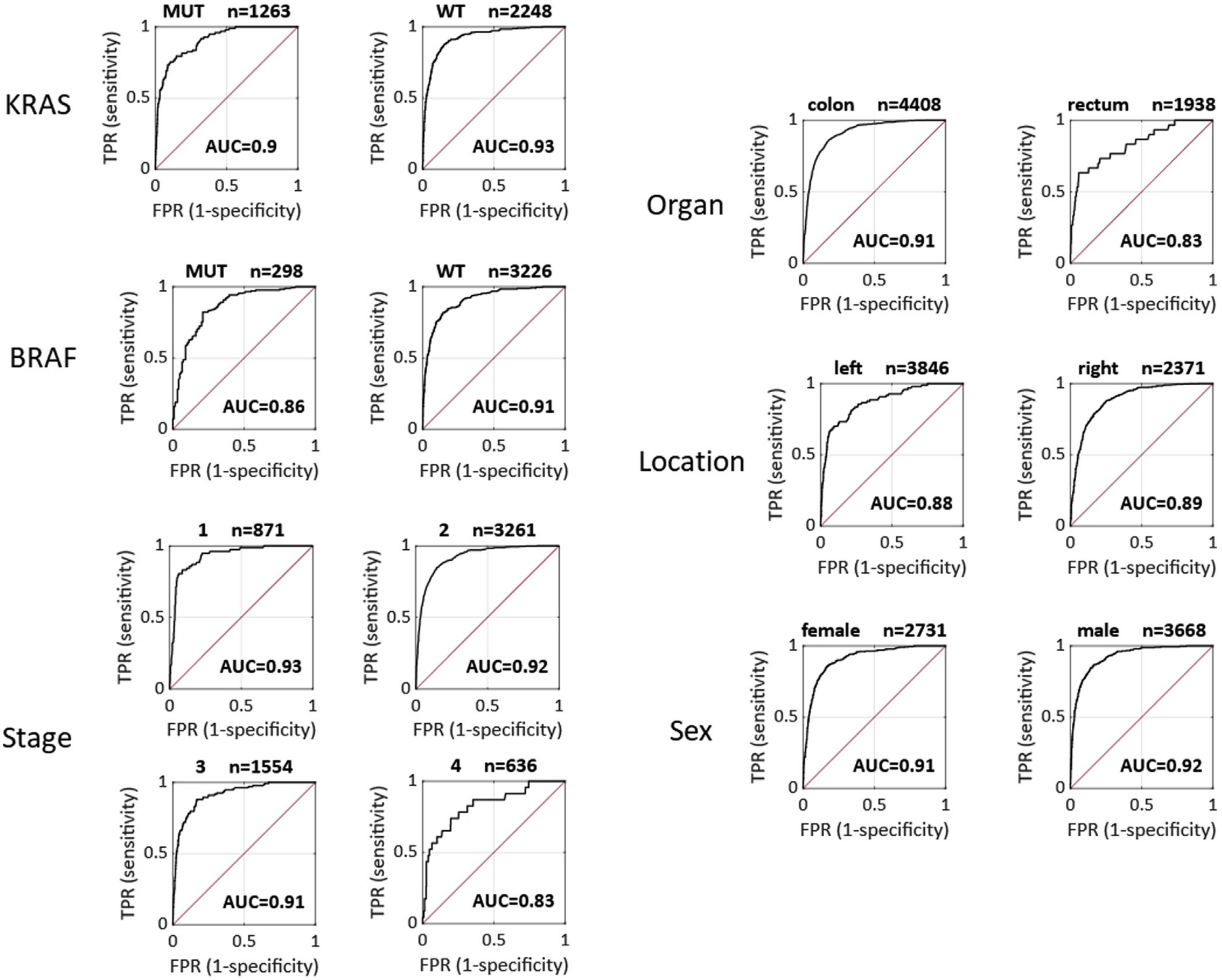


Figure 2. Cross-validated subgroup analysis for the detection of MSI and dMMR in the international cohort (N = 6406 patients). AUC, area under the receiver operating curve as shown in the image; FPR, false positive rate; MUT, mutated; TPR, true positive rate; WT, wild type.

showing a very high AUPRC of 0.79 (0.74–0.86), compared to the baseline AUPRC of 0.14 of the null model in this cohort. These data show that a deep learning system trained on a large and heterogeneous international training cohort generalizes well beyond the training set and thus constitutes a tool of potential clinical applicability.

Prediction Performance Is Robust in Clinicopathologic and Molecular Subgroups

CRC comprises a number of anatomically and biologically distinct molecular subgroups, including right- and left-sided colon cancer, rectal cancer, and *BRAF*-driven and *RAS*-driven tumors, among others. This is especially relevant because these features are partially dependent on each other; for example, *BRAF* mutations and right-sidedness are associated with MSI status.[31,32] To assess if image-based MSI prediction is robust across these heterogeneous subgroups, we used the cross-validated deep learning system (experiment 2 in Supplementary Table 5) and compared AUROC and AUPRC across subgroups (Figure 2 and Supplementary Figure 8). We found some variation in classifier performance regarding anatomic location: the AUROC was 0.89 for right-sided cancer (n = 2371 patients), 0.88 for left-sided cancer (n = 3846), 0.91 for colon cancer overall (n = 4408), and 0.83 for rectal cancer (n = 1938). Little variation was observed in classifier performance according to molecular features: AUROC was 0.86 in *BRAF* mutants (n = 298) and 0.91 in *BRAF* wild type (n = 3226); also, AUROC was 0.90 in *KRAS* mutants (n = 1263) and 0.93 in *KRAS* wild-type tumors (n = 2248). Finally, we analyzed the robustness of MSI predictions for different Union for International Cancer Control stages, showing stable performance with an AUROC of 0.93 in stage I (n = 871), 0.92 in stage II (n = 3261), and 0.91 in stage III (n = 1554) tumors and a minor reduction of performance in patients with stage IV tumors (n = 636), reaching an AUROC of 0.83. In addition, histologic grading (Supplementary Figure 9) did not influence classification performance. Next, we asked if this robust performance across subgroups was maintained in

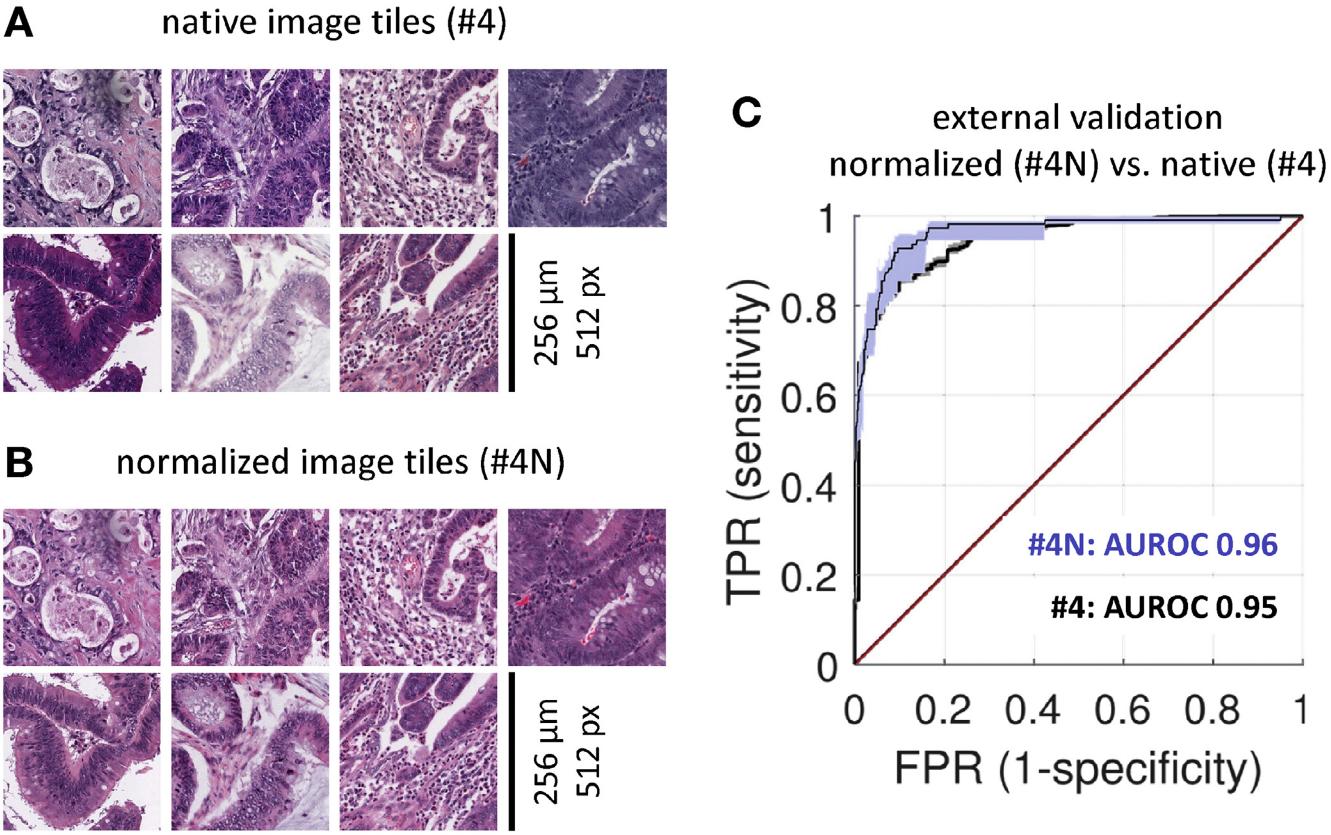


Figure 3. Effect of color normalization on classifier performance. (*A*) A representative set of tiles from the MSIDETECT study. (*B*) The same tiles after color normalization. (*C*) Classifier performance on an external test set (YCR-BCIP-RESECT, n = 771 patients) improves after color normalizing the training and test sets. Experiment 4N is with color normalization; experiment 4 is without color normalization. FPR, false positive rate; TPR, true positive rate.

the external test cohort (YCR-BCIP-RESECT, n = 771 patients). Again, in this cohort, we did not find any relevant loss in performance with regard to the following subgroups: tumor stage, organ, anatomical location, and sex (Supplementary Figures 10 and 11). In summary, this analysis shows and quantifies variations in performance according to CRC subgroups but demonstrates that overall, MSI and dMMR detection performance is robust.

Application of the Deep Learning System to Biopsy Samples

As additional exploratory endpoints, we tested if a deep learning system trained on histologic images from surgical resections can predict MSI and dMMR status of images from endoscopic biopsy tissue. Biopsy samples include technical artifacts (fragmented tissue and small tissue area) (Supplementary Figure 12*A*) as well as biological artifacts (sampled from the luminal portions of the tumor only). We acquired endoscopic biopsy samples from n = 1557 patients in the YCR-BCIP-BIOPSY study and tested the resection-trained classifier (experiment 5 in Supplementary Table 6). We found that AUROC was reduced to 0.78 (0.75–0.81) (Supplementary Figure 12*B*) in this experiment. In a 3-fold cross-validated experiment on all n = 1531 patients in the YCR-BCIP-BIOPSY cohort, MSI and dMMR detection performance was restored to an AUROC of 0.89 (0.88–0.91) (experiment 6 in Supplementary Table 5). These data suggest that MSI and dMMR testing on biopsy samples requires a classifier trained on biopsy samples.

Color Normalization Improves External Test Performance

Because previous studies have pointed to a benefit of color normalizing histology images before quantitative analysis,[29] the main experiments in this study were repeated on color-normalized image tiles. Native (non-normalized) image tiles (Figure 3*A*) were subjectively more diverse in terms of staining hue and intensity than normalized tiles (Figure 3*B*). Repeating MSI and dMMR prediction by 3-fold cross-validation on the full international cohort with color-normalized tiles (experiment 2N in Supplementary Table 5), we found that color normalization modestly improves specificity at predefined sensitivity levels: specificity was 57% at 99% sensitivity in experiment 2N, as opposed to a specificity of 38% at 99% sensitivity in the corresponding nonnormalized experiment (2). However, this increase in specificity did not result in a higher AUROC overall (Supplementary Table 5). To test if color normalization improves the external test performance of MSI and dMMR predictors, we repeated experiment 4 (training on full

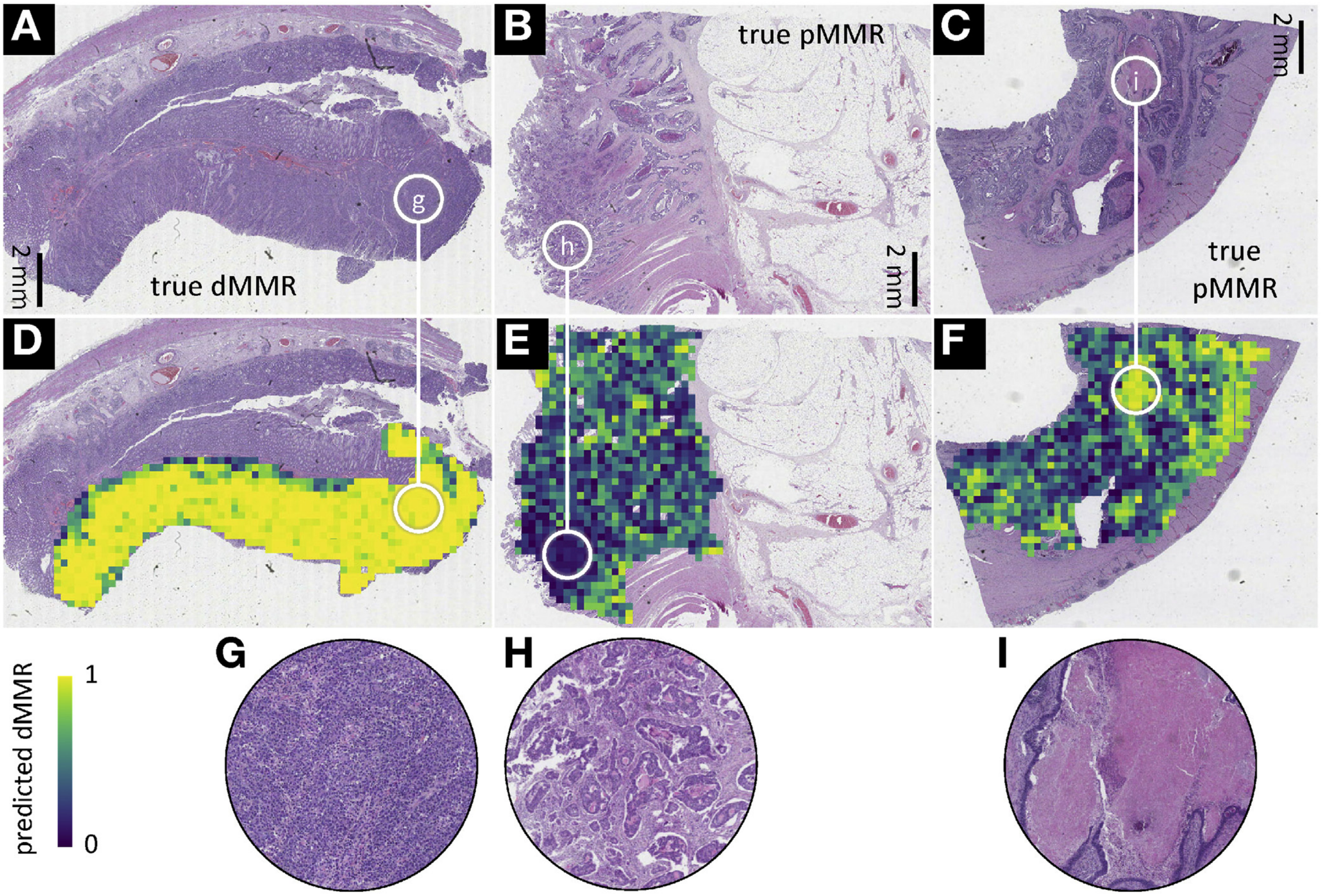


Figure 4. Prediction map in the external test cohort YCR-BCIP-RESECT. (*A–C*) Representative images from the YCR-BCIP-RESECT test cohort labeled with immunohistochemically defined mismatch repair (MMR) status. (*D–F*) Corresponding deep learning prediction maps. The edge length of each prediction tile is 256 μm. (*G–I*) Higher magnification of the regions highlighted in *A-E*. True MSI or dMMR patients were strongly and homogeneously predicted to be MSI or dMMR (such as the patient shown in *A*). True MSS or pMMR patients were overall predicted to be MSS or pMMR (such as the patients in *B* and *C*), but a pronounced heterogeneity was observed in necrotic areas, poorly differentiated areas, and immune-infiltrated tumor areas at the invasive edge.

international cohort, external testing on YCR-BCIP-RESECT) after color normalization (experiment 4N). In this case, AUROC did improve (no normalization in 4: AUROC, 0.95 [0.92–0.96]; color normalization in 4N: AUROC, 0.96 [0.93–0.98]). This slight increase in AUROC translated into a higher specificity at predefined sensitivity levels, reaching 58% specificity at 99% sensitivity (Supplementary Table 5). These data show that color normalization can further improve classifier performance and improve generalizability of deep learning–based inference of MSI and dMMR status.

Discussion

A Clinical-Grade Deep Learning–Based Molecular Biomarker in Cancer

Analyzing more than 8000 patients with CRC in an international consortium, we show that deep learning can reliably detect MSI and dMMR tumors based on routine H&E histology alone. In an external validation cohort, the deep learning MSI and dMMR detector performed with similar characteristics to criterion standard tests,[12] reaching clinical-grade performance. As shown in previous studies,[16] it can be assumed that this deep learning–based method can be less expensive and faster than routine laboratory assays and therefore has the potential to improve clinical diagnostic workflows. Our data show that classifier performance in surgical specimens remains robust even when the classifier is applied to external cohorts, but performance is lower in biopsy samples where tissue areas are much smaller than those of surgically resected specimens. This highlights the need to perform thorough large-scale evaluation of deep learning–based biomarkers in each intended use case. Deep learning histology biomarkers such as the MSI and dMMR detection system can be made understandable by visualization of prediction maps (Figure 4*A–I*) or by visualizing highly scoring image tiles (Supplementary Figure 13*A* and *B*). Together, these approaches show that the deep learning system yielded plausible predictions. For example, high MSI or dMMR scores were assigned to poorly differentiated tumor tissue (Supplementary Figure 13*A*), whereas high microsatellite stable or pMMR scores were assigned to well-differentiated areas. Interestingly, the spatial patterns of tile-level predictions showed varying degrees of heterogeneity: in all analyzed true positive MSI and dMMR cases in the YCR-BCIP-RESECT validation cohort, we found a homogeneously strong prediction of MSI and dMMR, as shown in Figure 4*A* and *D*. In contrast, predictions in true MSS and pMMR cases were more heterogeneous. Necrotic, poorly differentiated, or immune-infiltrated areas tended to be falsely predicted to be MSI or dMMR (Figure 4*C* and *F*). However, because patient-level predictions reflected overall scores in the full tumor area, most true MSS and pMMR patients were correctly predicted after pooling tile-level predictions, despite some degree of tile-level heterogeneity.

Clinical Application: Prescreening or Definitive Testing

In this study, diagnostic performance was stable across multiple clinically relevant subgroups, except for lower-than-average performance in patients with rectal cancer, possibly due to neoadjuvant pretreatment of some of these patients. In summary, this study defines a thoroughly validated deep learning system for genotyping CRC based on histology images alone, which could be used in clinical settings after regulatory approval. By varying the operating threshold, sensitivity and specificity of this test can be changed according to the clinical workflow this test is embedded in: high-sensitivity deep learning assays could be used to prescreen patients and could trigger additional genetic testing in the case of positive predictions. Even with imperfect specificity, such classifiers could speed up the diagnostic workflow and provide immediate cost savings, especially in the context of universal MSI and dMMR testing, as recommended by clinical guidelines. Recent discussions and calculations on the cost effectiveness of systematic MSI or dMMR testing in patients with CRC[33] should incorporate deep-learning–based assays among the other strategies in the future.

Alternatively, deep learning biomarkers such as the method presented in this study could be used for definitive testing in the clinic, especially in health care settings where limited resources are currently prohibitive for universal molecular biology tests. Further studies are needed to determine the optimal operating thresholds for specific patient populations and clinical settings. In addition, clinical deployment will require prospective validation and regulatory approval. Ultimately, this method should rapidly identify MSS and pMMR cases with high certainty and identify high risk MSI, dMMR, and possible LS cases for confirmation by other tests. This could substantially reduce molecular testing load in clinical workflows and enable direct, universal, low-cost MSI and dMMR testing from ubiquitously available routine material.

Technical improvements could conceivably further improve performance and open up new clinical applications. In this study, we explored color normalization as a way of reducing heterogeneity in staining intensity and hue between patient cohorts. This intervention (experiment 4N in Supplementary Table 5) modestly improved performance, increasing specificity from 51% to 58% at 99% sensitivity in an external validation cohort. The deep learning system and the source codes used in this study have been publicly released, enabling other researchers to independently validate and, potentially, further improve its performance.

Limitations

A limitation to our experimental workflow is that the ground truth labels used to train the deep learning system are imperfect. In the MSIDETECT group, clinical routine assays were used to assess MSI or dMMR status, and these assays have a nonzero error rate. Correspondingly, classifier performance could suffer from noisy labels in the training data. On the other hand, test cases flagged as *false positive* could be true MSI or dMMR cases that were missed by the clinical criterion standard test. Ultimately, it is conceivable that deep learning assays can outperform classical genetic or molecular tests in terms of predictive and prognostic performance, but testing this hypothesis would require large cohorts with clinical endpoint data and/or deep genetic characterization. In particular, the deep learning classifier could potentially detect rare genetic aberrations with MSI-like morphology, but again, lack of large training cohorts for these rare features currently precludes deeper investigation of this aspect. Another potential limitation of this study is the performance in patient groups of potential clinical interest that were not analyzed in the subgroup analysis, such as hereditary vs sporadic MSI and dMMR cases or different ethnic backgrounds. This is due to the lack of available clinical data in the utilized patient cohorts, and future studies are needed to investigate the stability of deep learning–based prediction in these and further subpopulations.

Interestingly, when we analyzed the per-patient predictions of MSI status in the external test set (YCR-BCIP-RESECT), we found an outlier among the false negative predictions: patient 441999 had a very low “predicted MSI probability” of less than 15%, whereas all other “true MSI” patients had MSI probability scores of more than 40%. We went back to the original histology slide of patient 441999 and noticed that a technical artifact had resulted in a blurred image, which was visible at only high magnification and had thus gone undetected in the manual quality check. This shows that an improved quality control at multiple magnification levels could increase the sensitivity of the deep learning assay, maintaining a high specificity.

Finally, a possible practical challenge in further validation and future integration of the deep learning methods in a clinical workflow is the current lack of regular installation of slide scanners in hospitals. However, in the United Kingdom and other countries, large academic consortia are currently implementing nationwide digital pathology workflows. This trend can be expected to further accelerate and will be supported by clinically useful applications of deep learning technology, especially after regulatory approval of such tools.[34] Still, initially it is probably more realistic to establish central testing facilities that are equipped with slide scanners and the further hardware needed for deep learning applications. In this setting, smaller hospitals and medical centers would not be confronted with high fixed costs but only with expenses and work that come with the distribution of H&E glass slides to central testing facilities.

Context: Multicenter Validation of Deep Learning Biomarkers

Recent years have seen a surge of deep learning methods in digital pathology, but previous large-scale studies are limited to simple image analysis tasks such as tumor detection[35] and do not extend to scenarios of molecular biomarker detection. Smaller proof-of-concept studies have shown that deep learning can detect a range of molecular biomarkers directly from routine histology, including multiple clinically relevant oncogenes.[17–19] However, these classifiers were not validated in large multicenter cohorts and cannot be readily generalized beyond the training set. To our knowledge, the present study is the first international collaborative effort to validate such a deep learning–based molecular biomarker. It identifies the need for very large series; training on a variety of sample types, such as resection and biopsy; and different populations. The high performance in this particular use case yields a tool of immediate clinical applicability and provides a blueprint for the emerging class of deep learning–based molecular tests in oncology, with the potential to broadly improve workflows in precision oncology worldwide.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2020.06.021.

References

1. Luchini C, Bibeau F, Ligtenberg MJL, et al. ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. Ann Oncol 2019;30:1232–1243.
2. Kather JN, Halama N, Jaeger D. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. Semin Cancer Biol 2018;52:189–197.
3. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology 2010;138:2073–2087.
4. Molecular testing strategies for Lynch syndrome in people with colorectal cancer: recommendations. NICE Pathways. https://www.nice.org.uk/guidance/dg27/chapter/1-Recommendations. Accessed November 13, 2019.
5. Stjepanovic N, Moreira L, Carneiro F, et al. Hereditary gastrointestinal cancers: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 2019;30:1558–1571.
6. Boland CR, Thibodeau SN, Hamilton SR, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248–5257.
7. Kawakami H, Zaanan A, Sinicrope FA. Microsatellite instability testing and its role in the management of colorectal cancer. Curr Treat Options Oncol 2015;16(7):30.
8. Snowsill T, Coelho H, Huxley N, et al. Molecular testing for Lynch syndrome in people with colorectal cancer: systematic reviews and economic evaluation. Health Technol Assess 2017;21:1–238.
9. Evrard C, Tachon G, Randrian V, et al. Microsatellite instability: diagnosis, heterogeneity, discordance, and clinical impact in colorectal cancer. Cancers 2019;11(10):1567.
10. Molecular testing strategies for Lynch syndrome in people with colorectal cancer: evidence. NICE Pathways. https://www.nice.org.uk/guidance/dg27/chapter/4-Evidence. Accessed April 30, 2020.
11. Poynter JN, Siegmund KD, Weisenberger DJ, et al. Molecular characterization of MSI-H colorectal cancer by *MLHI* promoter methylation, immunohistochemistry, and mismatch repair germline mutation screening. Cancer Epidemiol Biomarkers Prev 2008;17:3208–3215.
12. Barnetson RA, Tenesa A, Farrington SM, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. N Engl J Med 2006;354:2751–2763.
13. Limburg PJ, Harmsen WS, Chen HH, et al. Prevalence of alterations in DNA mismatch repair genes in patients with young-onset colorectal cancer. Clin Gastroenterol Hepatol 2011;9:497–502.
14. De Smedt L, Lemahieu J, Palmans S, et al. Microsatellite instable vs stable colon carcinomas: analysis of tumour heterogeneity, inflammation and angiogenesis. Br J Cancer 2015;113:500–509.
15. Greenson JK, Huang S-C, Herron C, et al. Pathologic predictors of microsatellite instability in colorectal cancer. Am J Surg Pathol 2009;33:126–133.
16. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25:1054–1056.
17. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer 2020. https://doi.org/10.1038/s43018-020-0087-6. Accessed September 6, 2020.
18. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. bioRχiv 2019:813543. https://www.biorxiv.org/content/10.1101/813543v1. Accessed October 27, 2019.
19. Schmauch B, Romagnoni A, Pronier E, et al. Transcriptomic learning for digital pathology. bioRχiv

2019:760173. https://www.biorxiv.org/content/10.1101/760173v1. Accessed September 11, 2019.

20. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487(7407):330–337.

21. Amitay EL, Carr PR, Jansen L, et al. Association of aspirin and nonsteroidal anti-inflammatory drugs with colorectal cancer risk by molecular subtypes. J Natl Cancer Inst 2019;111:475–483.

22. Brenner H, Chang-Claude J, Seiler CM, et al. Does a negative screening colonoscopy ever need to be repeated? Gut 2006;55:1145–1150.

23. QUASAR Collaborative Group. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. Lancet 2007;370(9604):2020–2029.

24. van den Brandt PA. Molecular pathological epidemiology of lifestyle factors and colorectal and renal cell cancer risk. J Pathol 2018;246(Suppl 1):S1–S46.

25. van den Brandt PA, Goldbohm RA, van 't Veer P, et al. A large-scale prospective cohort study on diet and cancer in The Netherlands. J Clin Epidemiol 1990;43:285–295.

26. Taylor J, Wright P, Rossington H, et al. Regional multidisciplinary team intervention programme to improve colorectal cancer outcomes: study protocol for the Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR BCIP). BMJ Open 2019;9:e030618.

27. Hoffmeister M, Bläker H, Kloor M, et al. Body mass index and microsatellite instability in colorectal cancer: a population-based study. Cancer Epidemiol Biomarkers Prev 2013;22:2303–2311.

28. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7:16878.

29. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. Boston, MA: IEEE, 2009:1107–1110.

30. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med 2019;16(1):e1002730.

31. Salem ME, Weinberg BA, Xiu J, et al. Comparative molecular analyses of left-sided colon, right-sided colon, and rectal cancers. Oncotarget 2017;8:86356–86368.

32. Lochhead P, Kuchiba A, Imamura Y, et al. Microsatellite instability and *BRAF* mutation testing in colorectal cancer prognostication. J Natl Cancer Inst 2013;105:1151–1156.

33. Kang Y-J, Killen J, Caruana M, et al. The predicted impact and cost-effectiveness of systematic testing of people with incident colorectal cancer for Lynch syndrome. Med J Aust 2020;212:72–81.

34. Paige. https://paige.ai/. Accessed April 2, 2020.

35. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–1309.

36. Liu Y, Sethi NS, Hinoue T, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. Cancer Cell 2018;33:721–735.

Received March 27, 2020. Revised May 22, 2020. Accepted June 8, 2020.

Correspondence

Address correspondence to: Jakob N. Kather, MD, MSc, Department of Gastroenterology, Hepatology and Medical Intensive Care, RWTH University Hospital Aachen, Pauwelsstr 30, 52074 Aachen, Germany. e-mail: jkather@ukaachen.de.

Acknowledgments

The authors are grateful to all investigators and contributing pathologists from the TCGA study (more information on *http://portal.gdc.cancer.gov*), the Rainbow-TMA consortium in the Netherlands (listed in Supplementary Table 6), the DACHS consortium in Germany, the QUASAR consortium, and the YCR-BCIP consortium in the United Kingdom. Collection and testing of the YCR-BCIP cases was funded by Yorkshire Cancer Research L386 and L394 as part of previous studies. Philip Quirke is an NIHR Senior Investigator. We thank the Yorkshire and Humber Histopathologists who facilitated the collection of the YCR BCIP cases.

CRediT Authorship Contributions

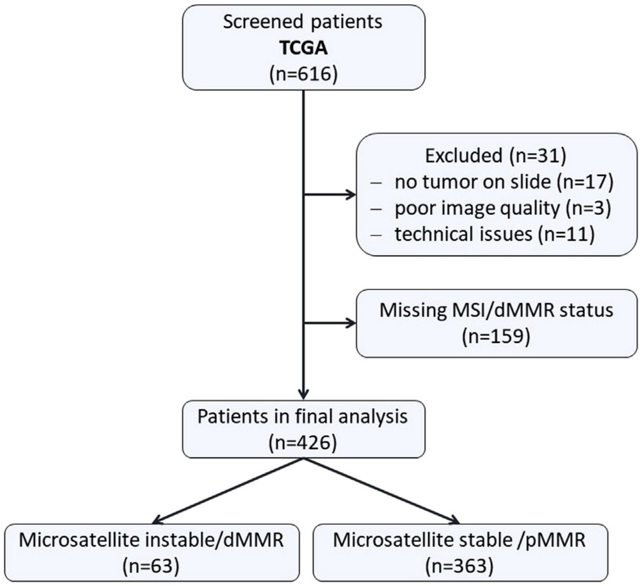
Amelie Echle (Conceptualization: Equal; Formal analysis: Equal; Methodology: Equal; Writing – original draft: Lead); Heike Grabsch (Conceptualization: Equal; Supervision: Equal; Validation: Equal; Writing – review & editing: Equal); Philip Quirke (Conceptualization: Equal; Supervision: Equal; Validation: Equal; Writing – review & editing: Equal); Piet A. van den Brandt (Conceptualization: Equal; Resources: Equal; Validation: Equal; Writing – review & editing: Equal); Nicholas P. West (Data curation: Equal; Supervision: Equal; Validation: Equal; Writing – review & editing: Equal); Gordon G.A. Hutchins (Data curation: Equal; Validation: Equal; Writing – review & editing: Equal); Lara R. Heij (Investigation: Equal; Supervision: Equal; Writing – review & editing: Equal); Xiuxiang Tan (Formal analysis: Equal; Writing – review & editing: Equal); Susan D. Richman (Data curation: Equal; Validation: Equal; Writing – review & editing: Equal); Jeremias Krause (Formal analysis: Equal; Methodology: Equal; Writing – review & editing: Equal); Elizabeth Alwers (Formal analysis: Equal; Methodology: Equal; Writing – review & editing: Equal); Josien Jenniskens (Formal analysis: Equal; Methodology: Equal; Writing – review & editing: Equal); Kelly Offermans (Formal analysis: Equal; Methodology: Equal; Writing – review & editing: Equal); Richard Gray (Resources: Equal; Supervision: Equal; Writing – review & editing: Equal); Hermann Brenner (Resources: Equal; Supervision: Equal; Writing – review & editing: Equal); Jenny Chang-Claude (Resources: Equal; Supervision: Equal; Writing – review & editing: Equal); Christian Trautwein (Resources: Equal; Supervision: Equal; Writing – review & editing: Equal); Alexander T. Pearson, MD, PhD (Conceptualization: Equal; Methodology: Equal; Supervision: Equal; Writing – review & editing: Equal); Peter Boor (Methodology: Equal; Supervision: Equal; Writing – review & editing: Equal); Tom Luedde, MD, PhD (Methodology: Equal; Supervision: Equal; Validation: Equal; Writing – review & editing: Equal); Nadine Therese Gaisa (Investigation: Equal; Supervision: Equal; Writing – review & editing: Equal); Michael Hoffmeister (Data curation: Equal; Resources: Equal; Validation: Equal; Writing – review & editing: Equal); Jakob Nikolas Kather, MD, MSc (Conceptualization: Lead; Formal analysis: Lead; Investigation: Lead; Methodology: Lead; Software: Lead; Visualization: Lead; Writing – review & editing: Lead).

Conflicts of interest

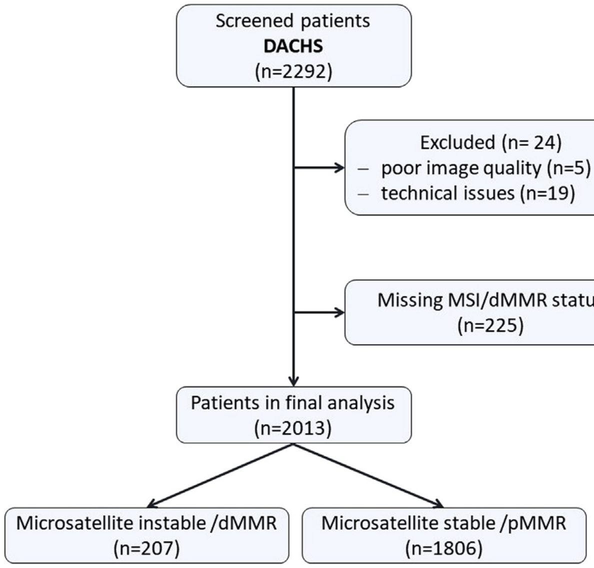
The authors disclose the following: Jakob N. Kather has an informal, unpaid advisory role at Pathomix (Heidelberg, Germany) that does not relate to this research. The remaining authors disclose no conflicts

Funding

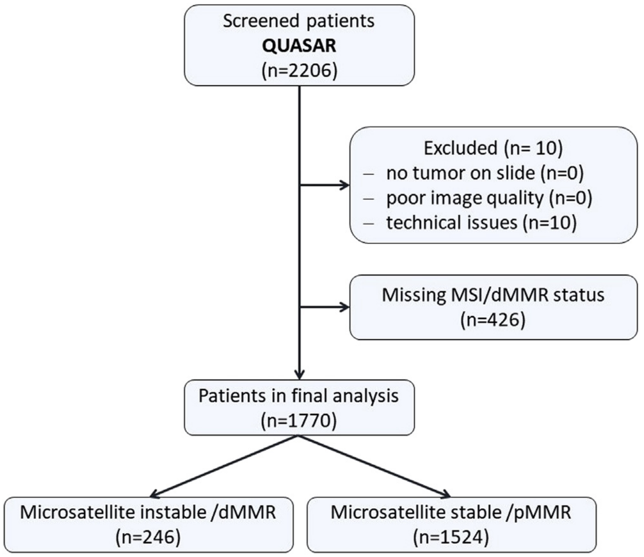
This study was primarily funded by the authors' academic institutions. These authors are supported by additional grants: Peter Boor: German Research Foundation (SFB/TRR57, SFB/TRR219, BO3755/3-1, and BO3755/6-1), the German Federal Ministry of Education and Research (BMBF: STOP-FSGS-01GM1901A), and the German Federal Ministry of Economic Affairs and Energy (BMWi: EMPAIA project). Alexander T. Pearson: National Institutes of Health/National Institute of Dental and Craniofacial Research (K08-DE026500), Institutional Research Grant (IRG-16-222-56) from the American Cancer Society, Cancer Research Foundation Research Grant, and the University of Chicago Medicine Comprehensive Cancer Center Support Grant (P30-CA14599). Tom Luedde: Horizon 2020 through the European Research Council Consolidator Grant PhaseControl (771083), a Mildred Scheel–Endowed Professorship from the German Cancer Aid (Deutsche Krebshilfe), the German Research Foundation (SFB CRC1382/P01, SFB-TRR57/P06, LU 1360/3-1), the Ernst-Jung-Foundation Hamburg, and the Interdisciplinary Center of Clinical Research) at RWTH Aachen. Jakob Nikolas Kather: RWTH University Aachen (START 2018-691906), Max-Eder-Programme of the German Cancer Aid (Deutsche Krebshilfe, 70113864).


Screened patients
TCGA
(n=616)
Excluded (n=31)
– no tumor on slide (n=17)
– poor image quality (n=3)
– technical issues (n=11)
Missing MSI/dMMR status
(n=159)
Patients in final analysis
(n=426)
Microsatellite instable/dMMR
(n=63)
Microsatellite stable /pMMR
(n=363)

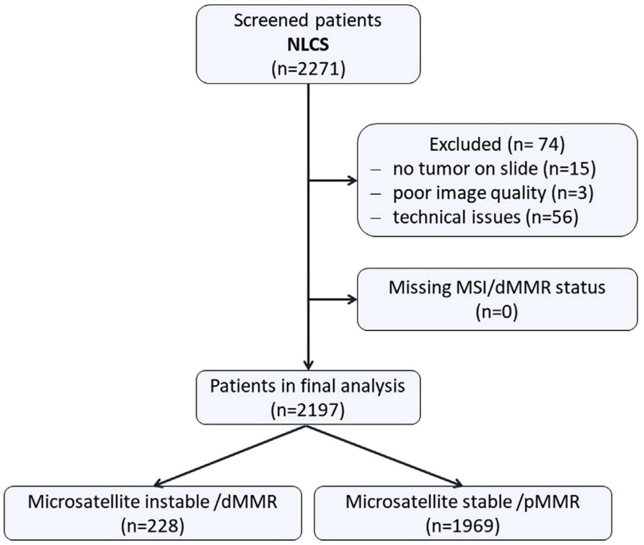
Supplementary Figure 1. Sample flowchart for the TCGA cohort.


Screened patients
DACHS
(n=2292)
Excluded (n= 24)
– poor image quality (n=5)
– technical issues (n=19)
Missing MSI/dMMR status
(n=225)
Patients in final analysis
(n=2013)
Microsatellite instable /dMMR
(n=207)
Microsatellite stable /pMMR
(n=1806)

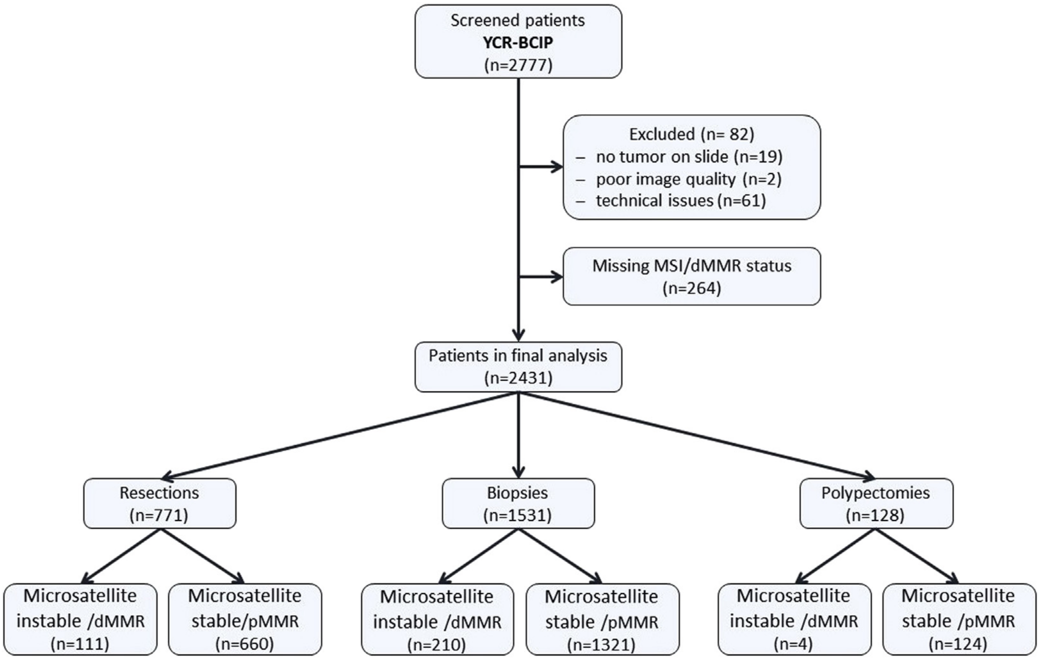
Supplementary Figure 2. Sample flowchart for the DACHS cohort.


Screened patients
QUASAR
(n=2206)
Excluded (n= 10)
– no tumor on slide (n=0)
– poor image quality (n=0)
– technical issues (n=10)
Missing MSI/dMMR status
(n=426)
Patients in final analysis
(n=1770)
Microsatellite instable /dMMR
(n=246)
Microsatellite stable /pMMR
(n=1524)

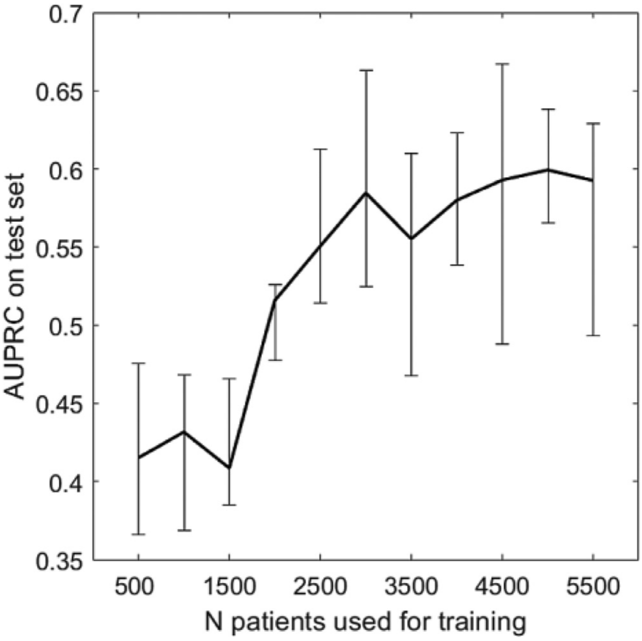
Supplementary Figure 3. Sample flowchart for the QUASAR cohort.


Screened patients
NLCS
(n=2271)
Excluded (n= 74)
– no tumor on slide (n=15)
– poor image quality (n=3)
– technical issues (n=56)
Missing MSI/dMMR status
(n=0)
Patients in final analysis
(n=2197)
Microsatellite instable /dMMR
(n=228)
Microsatellite stable /pMMR
(n=1969)

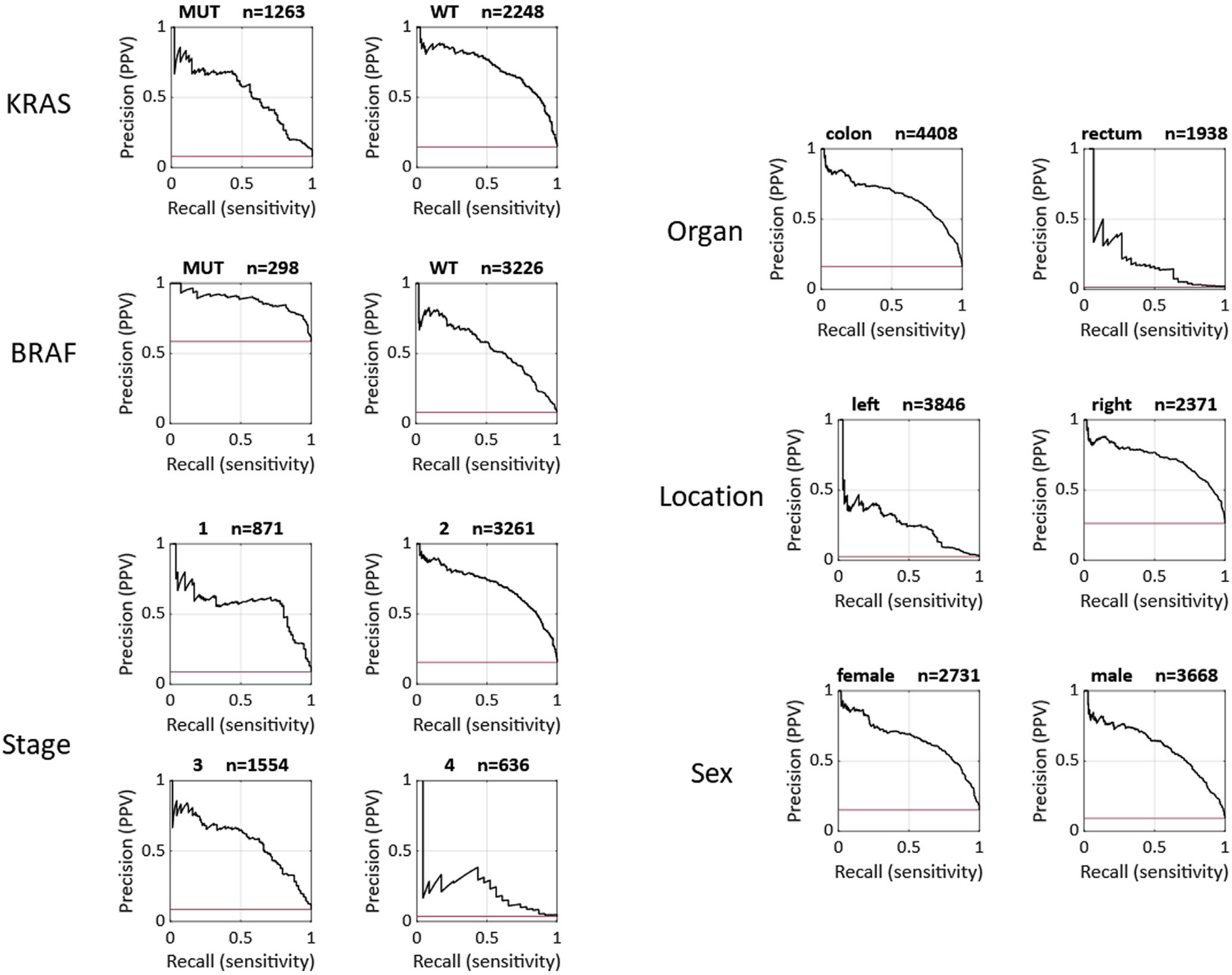
Supplementary Figure 4. Sample flowchart for the NLCS cohort.


Screened patients
YCR-BCIP
(n=2777)
Excluded (n= 82)
– no tumor on slide (n=19)
– poor image quality (n=2)
– technical issues (n=61)
Missing MSI/dMMR status
(n=264)
Patients in final analysis
(n=2431)
Resections
(n=771)
Biopsies
(n=1531)
Polypectomies
(n=128)
Microsatellite instable /dMMR (n=111)
Microsatellite stable/pMMR (n=660)
Microsatellite instable /dMMR (n=210)
Microsatellite stable /pMMR (n=1321)
Microsatellite instable /dMMR (n=4)
Microsatellite stable /pMMR (n=124)

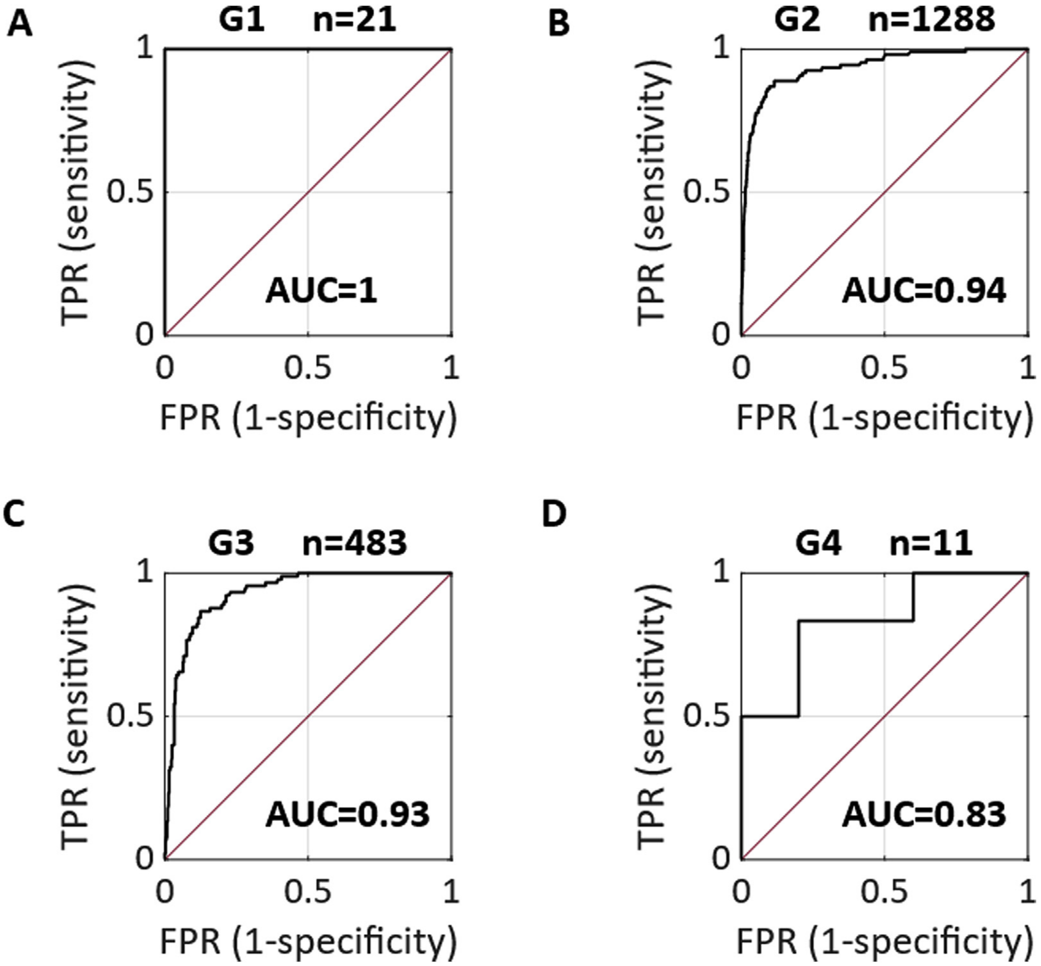
Supplementary Figure 5. Sample flowchart for the external test cohort YCR-BCIP. The primary intention of our study was to validate the MSI detection classifier in the surgical resection samples. As an explorative analysis, we validated the classifier in endoscopic biopsy samples. Polypectomy samples were not assessed because of a low relative proportion and absolute number of positive cases.



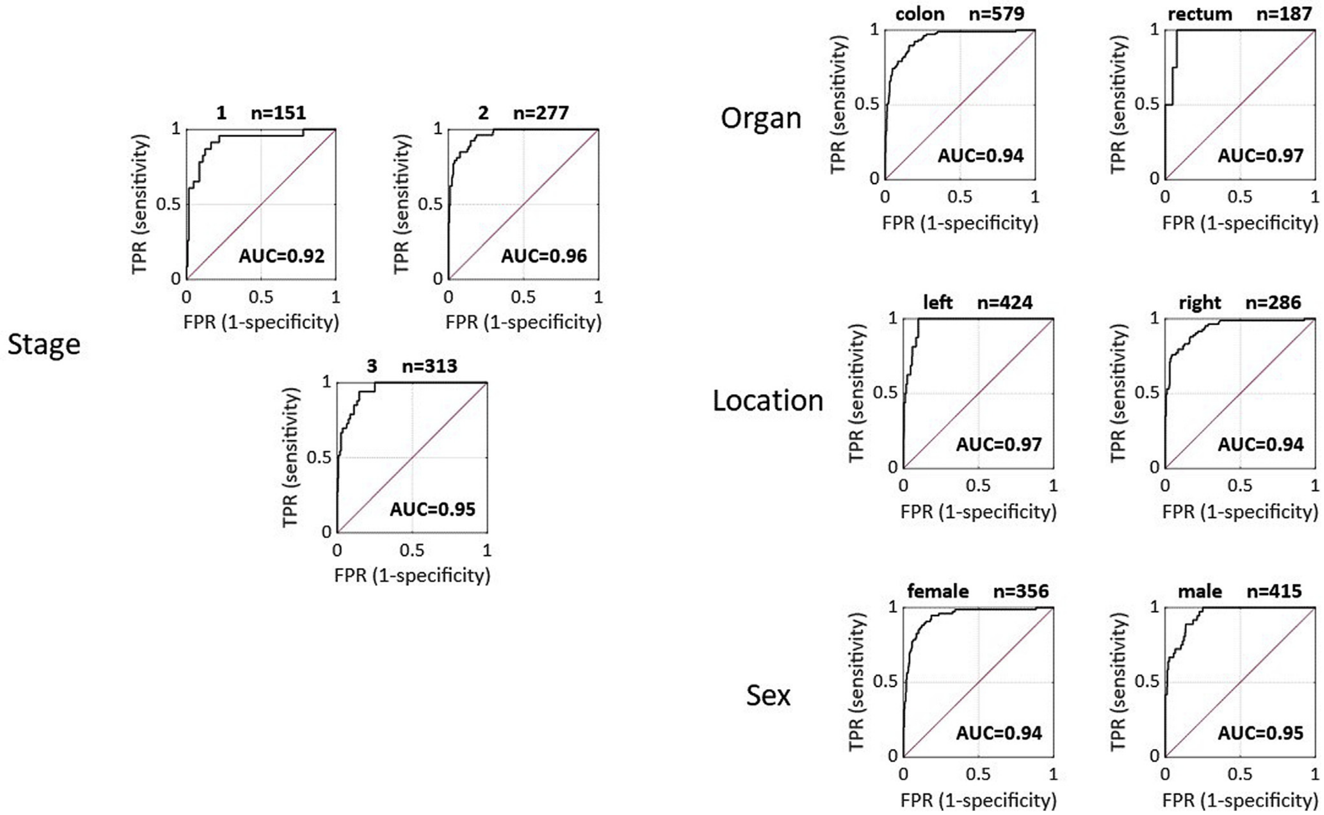
Supplementary Figure 6. AUPRC for the learning curve experiment. Includes experiment 3, related to Figure 1*C*.



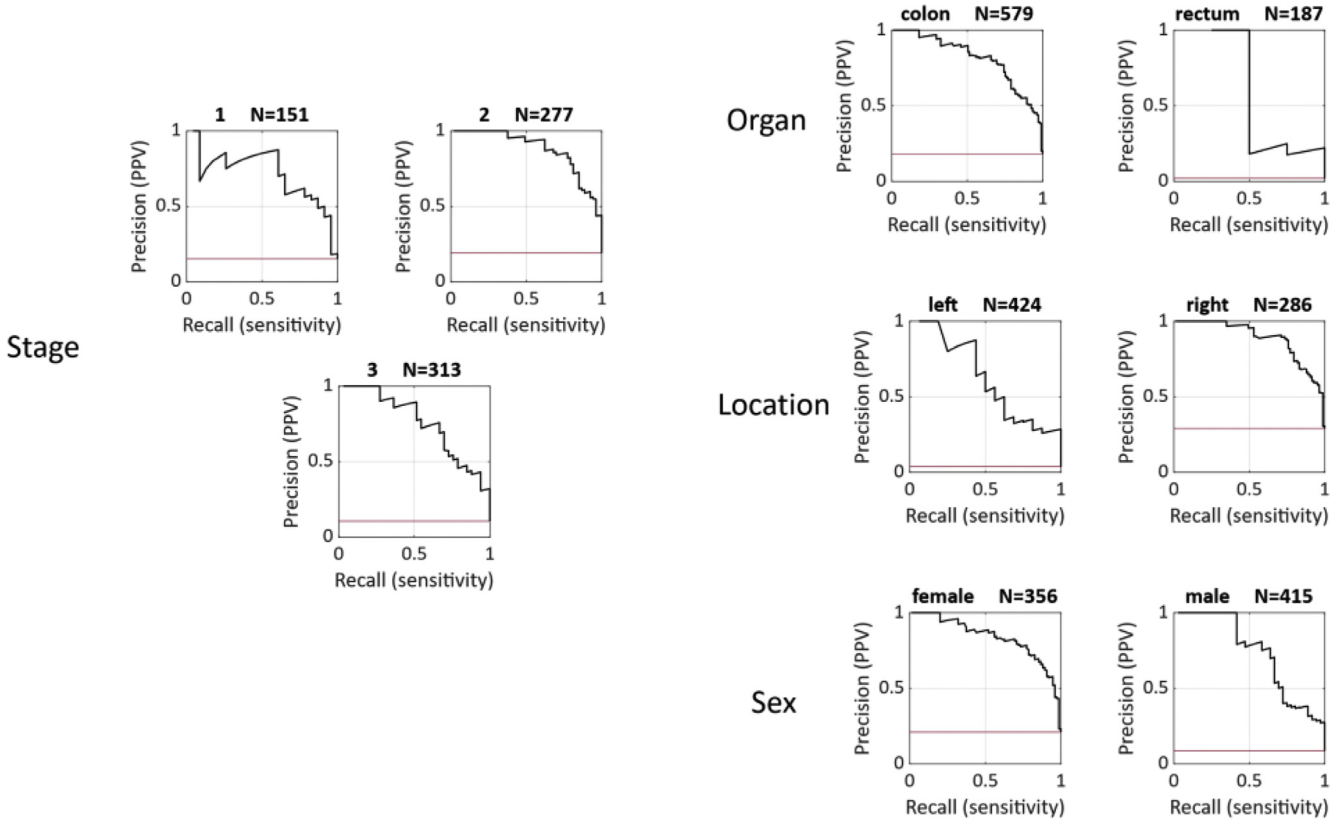
Supplementary Figure 7. Cross-validated subgroup analysis in TCGA, DACHS, and QUASAR: precision-recall curves. Related to Figure 2. The red line indicates baseline (random) model with no skill. MUT, mutated; PPV, positive predictive value; WT, wild type.



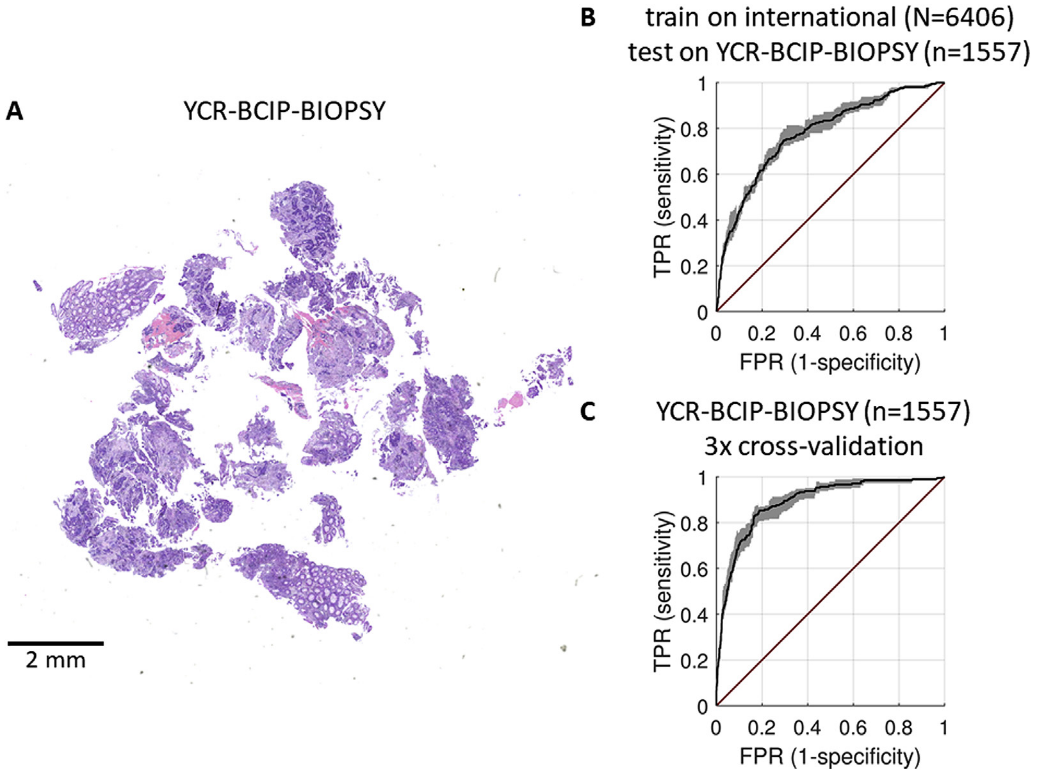
Supplementary Figure 8. Subgroup analysis for MSI and dMMR detection according to histologic grading. (*A–D*) Receiver operating characteristic analysis for subgroups of patients stratified by histologic grading G1–G4. Related to experiment 2. Grading information was available only for patients in the DACHS cohort. FPR, false positive rate; TPR, true positive rate.



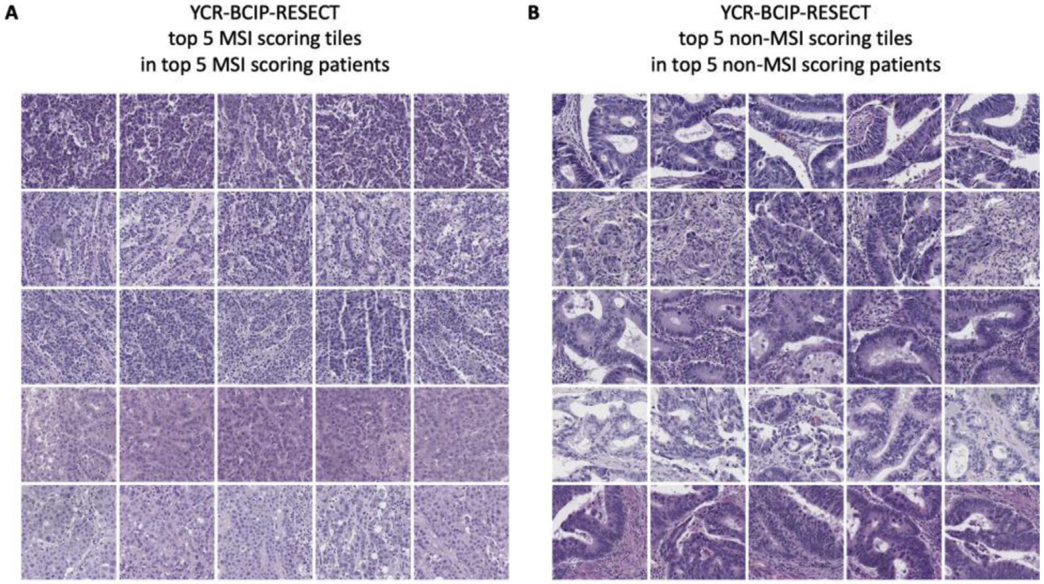
Supplementary Figure 9. Subgroup analysis of MSI detection performance in the test set (YCR-BCIP-RESECT), Receiver operating characteristic curves.



Supplementary Figure 10. Subgroup analysis of MSI detection performance in the test set (YCR-BCIP-RESECT): precision-recall curves.



Supplementary Figure 11. Classifier performance in biopsy samples (YCR-BCIP biopsy). (*A*) Representative whole slide image; brightness and contrast have been linearly increased for better visibility (+20%). This examples shows that endoscopic biopsy tissue is usually fragmented. (*B*) Receiver operating characteristic curve related to experiment 5 in Supplementary Table 5. (*C*) Receiver operating characteristic curve related to experiment 6 in Supplementary Table 5.


A
YCR-BCIP-RESECT
top 5 MSI scoring tiles
in top 5 MSI scoring patients
B
YCR-BCIP-RESECT
top 5 non-MSI scoring tiles
in top 5 non-MSI scoring patients

Supplementary Figure 12. Highly scoring tiles by MSI and MMR status. (*A*) The 5 highest scoring image tiles in the 5 highest scoring MSI or dMMR patients in the YCR-BCIP-RESECT cohort from experiment 4 in Supplementary Table 5. (*B*) Correspondingly, the highest scoring non-MSI tiles in the highest scoring non-MSI patients.

Supplementary Table 1. Clinicopathologic Features of Each Cohort

	DACHS	QUASAR	TCGA	NLCS	YCR-BCIP-RESECT	YCR-BCIP-BIOPSY	YCR-BCIP-POLYP
Number of patients	2013	1770	426	2197	771	1531	128
Region	Germany	United Kingdom	United States	The Netherlands	United Kingdom	United Kingdom	United Kingdom
WSI format	SVS	SVS	SVS	MRXS	SVS	SVS	SVS
MSI or dMMR	PCR 3-plex	IHC 2-plex	Genetic[a]	IHC 2-plex	IHC 4-plex	IHC 4-plex	IHC 4-plex
Mean age at diagnosis, *y*	68.8	62.2	65.6	73.8	70.5	71.9	67.8
MSI positive, n (%)	207 (10.3)	246 (13.9)	63 (14.8)	228 (10.4)	111 (14.39)	210 (13.72)	4 (3.13)
Stage I, n (%)	369 (18.3)	0 (0.0)	67 (15.7)	435 (19.8)	151 (19.6)	Unknown	Unknown
Stage II, n (%)	674 (33.5)	1608 (90.8)	154 (36.2)	825 (37.6)	277 (35.9)	Unknown	Unknown
Stage III, n (%)	690 (34.3)	156 (8.8)	133 (31.2)	575 (26.2)	313 (40.6)	Unknown	Unknown
Stage IV, n (%)	280 (13.9)	0 (0.0)	59 (13.8)	297 (13.5)	0 (0.0)	Unknown	Unknown
Male, n (%)	1161 (57.7)	1073 (60.6)	211 (49.5)	1223 (55.7)	415 (53.8)	912 (59.6)	89 (69.5)
Female, n (%)	852 (42.3)	692 (39.1)	213 (50.0)	974 (44.3)	356 (46.2)	615 (40.2)	38 (29.7)
Colon cancer, n (%)	1260 (62.6)	1274 (72.0)	321 (75.4)	1553 (70.7)	579 (75.1)	858 (56.0)	69 (53.9)
Rectal cancer, n (%)	753 (37.4)	436 (24.6)	105 (24.6)	644 (29.3)	187 (24.3)	658 (43.0)	59 (46.1)
BRAF mutant, n (%)	138 (6.9)	104 (5.9)	56 (13.1)	Unknown	Unknown	Unknown	Unknown
BRAF wild type, n (%)	1711 (85.0)	1145 (64.7)	370 (86.9)	Unknown	Unknown	Unknown	Unknown
KRAS mutant, n (%)	606 (30.1)	465 (26.3)	192 (45.1)	Unknown	Unknown	Unknown	Unknown
KRAS wild type, n (%)	1264 (62.8)	750 (42.4)	234 (54.9)	Unknown	Unknown	Unknown	Unknown

MRXS, Mirax Digital Slide Format; SVS, Aperio SVS file format; WSI, whole slide image.

[a] Liu Y, Sethi NS, Hinoue T, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. Cancer Cell 2018;33:721–735.

Supplementary Table 2.TRIPOD checklist

Section/Topic	Item		Checklist Item
Title and abstract			
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction			
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.
Methods			
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	D;V	Describe eligibility criteria for participants.
	5c	D;V	Give details of treatments received, if relevant.
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D;V	Explain how the study size was arrived at.
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	V	For validation, describe how the predictions were calculated.
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	D;V	Provide details on how risk groups were created, if done.
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Results			
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).

Supplementary Table 2. Continued

Section/Topic	Item		Checklist Item
Model development	14a	D	Specify the number of participants and outcome events in each analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	D	Explain how to the use the prediction model.
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).
Discussion			
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.
Other information			
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.

Supplementary Table 3. Description and Values of All Hyperparameters and Properties of the Deep Learning Workflow

Hyperparameter	Description	Value
Network architecture	Deep neural network layout	shufflenet
Tile size	Size of image tiles on the whole slide image	256 μm
Tile magnification	Image magnification of the tiles	0.5 μm/pixel
Effective tile size	Size of image tile files for training	512 px
Tiles per slide	Limit the number of randomly picked tiles	1000 (single cohort) 1000 (learning curve) 500 (combined)
Maximum epochs	Maximum number each image tile is shown to the network during training	4
Trainable layers	Number of network layers with nonzero learning rate, counted from the end	30
Validation fraction	Use some training tiles as a validation set to mitigate overfitting	0% (single cohort) 5% (learning curve) 5% (combined)
Graphics processing unit	Graphics processing unit hardware	NVidia RTX6000
Initial learn rate	Network learning rate during training	5 × 10–5
L2 regularization	L2 regularization to mitigate overfitting	0.0001
Mini batch size	Number of image tiles processes in parallel	512
Solver	Optimizer to update weights and biases	adam

NOTE. Single-cohort models were trained on 1 study population only (eg, QUASAR), whereas combined cohort experiments were trained on patients across cohorts. Parameters for the learning curve experiment refer to Figure 1*D*.

Supplementary Table 4. Unique Identifiers of Downloadable Models

Unique model identifier	Trained on	Additional information
HWSLMTNSRQSY	TCGA (n = 426, 15% MSI)	Used to quantify batch effect 1000 tiles per slide
QWAFEDQLFICH	QUASAR (n = 1770, 14% dMMR)	Used to quantify batch effect 1000 tiles per slide
YGREQLWMWSLR	DACHS (n = 2013, 14% MSI)	Used to quantify batch effect 1000 tiles per slide
VKTALAYFITSN	NLCS (n = 2197, 10% dMMR)	Used to quantify batch effect 1000 tiles per slide
HLVDDREQHWQK	International cohort, nonnormalized (TCGA + QUASAR + DACHS + NLCS)	This is the final model deployed on the external test cohorts 500 tiles per slide.
AWAMMGTGLNAF	International cohort, color normalized (TCGA + QUASAR + DACHS + NLCS)	This is the final model deployed on the external test cohorts 500 tiles per slide.

NOTE. All deep learning models trained in this study are freely available for academic reuse.

Supplementary Table 5. Performance Statistics for All Experiments Described in This Article

ID	Experiment description	Result statistics
#1	Estimate intracohort and intercohort performance for all subcohorts in the international cohort (Total N = 6406, 12% MSI or dMMR)	See Table 1
#2	3-fold cross-validation on full international cohort (Total N = 6406, 12% MSI or dMMR)	AUROC: 0.92 (0.91–0.93) AUPRC: 0.63 (0.59–0.65) Sensitivity: 99%, specificity: 38% Sensitivity: 98%, specificity: 48% Sensitivity: 95%, specificity: 67%
#2N	3-fold cross-validation on full international cohort (with color normalization)	AUROC: 0.92 (0.91–0.93) AUPRC: 0.63 (0.61–0.66) Sensitivity: 99%, specificity: 57% Sensitivity: 98%, specificity: 78% Sensitivity: 95%, specificity: 83%
#3	Train on n = 5500 international patients Test on n = 906 international patients (Test set 12% MSI or dMMR, last part of learning curve)	AUROC: 0.92 (0.90–0.93) AUPRC: 0.59 (0.49–0.63) Sensitivity: 99%, specificity: 49% Sensitivity: 98%, specificity: 52% Sensitivity: 95%, specificity: 68%
#4	Train on the full international cohort (N = 6406, 12% MSI or dMMR, model ID: HLVDDREQHWQK) External test on YCR-BCIP-RESECT (test set n = 771, 14% dMMR)	AUROC: 0.95 (0.92–0.96) AUPRC: 0.79 (0.74–0.86) Sensitivity: 99%, specificity: 51% Sensitivity: 98%, specificity: 66% Sensitivity: 95%, specificity: 74%
#4N	Train on the full international cohort (model ID: AWAMMGTGLNAF) External test on YCR-BCIP-RESECT (with color normalization)	AUROC: 0.96 (0.93–0.98) AUPRC: 0.85 (0.829–0.90) Sensitivity: 99%, specificity: 58% Sensitivity: 98%, specificity: 79% Sensitivity: 95%, specificity: 86%
#5	Train on the full international cohort (N = 6406, 12% MSI or dMMR, model ID: HLVDDREQHWQK) External test on YCR-BCIP-BIOPSY (test set n = 1531, 14% dMMR)	AUROC: 0.78 (0.75–0.81) AUPRC: 0.37 (0.32–0.43) Sensitivity: 99%, specificity: 19% Sensitivity: 98%, specificity: 20% Sensitivity: 95%, specificity: 25%
#6	3-fold cross-validation on YCR-BCIP-BIOPSY (Concatenate test partitions, total n = 1531, 14% dMMR)	AUROC: 0.89 (0.88–0.91]) AUPRC: 0.58 (0.56–0.61) Sensitivity: 99%, specificity: 35% Sensitivity: 98%, specificity: 38% Sensitivity: 95%, specificity: 56%

NOTE. Detailed performance statistics, corresponding to Figure 2. No patient in a training set was ever part of a test set in the same experiment. Experiment 4 was the prespecified primary endpoint of this study.

Supplementary Table 6. The Rainbow-TMA Consortium Associated With the NLCS Study

Rainbow-TMA project group	P.A. van den Brandt, A. zur Hausen, H. Grabsch, M. van Engeland, L.J. Schouten, J. Beckervordersandforth (Maastricht University Medical Center, Maastricht, Netherlands); P.H.M. Peeters, P.J. van Diest, H.B. Bueno de Mesquita (University Medical Center Utrecht, Utrecht, Netherlands); J. van Krieken, I. Nagtegaal, B. Siebers, B. Kiemeney (Radboud University Medical Center, Nijmegen, Netherlands); F.J. van Kemenade, C. Steegers, D. Boomsma, G.A. Meijer (VU University Medical Center, Amsterdam, Netherlands); F.J. van Kemenade, B. Stricker (Erasmus University Medical Center, Rotterdam, Netherlands); L. Overbeek, A. Gijsbers (PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive, Houten, Netherlands)
Rainbow-TMA collaborating pathologists, among others	A. de Bruïne (VieCuri Medical Center, Venlo); J.C. Beckervordersandforth (Maastricht University Medical Center, Maastricht); J. van Krieken, I. Nagtegaal (Radboud University Medical Center, Nijmegen); W. Timens (University Medical Center Groningen, Groningen); F.J. van Kemenade (Erasmus University Medical Center, Rotterdam); M.C.H. Hogenes (Laboratory for Pathology Oost-Nederland, Hengelo); P.J. van Diest (University Medical Center Utrecht, Utrecht); R.E. Kibbelaar (Pathology Friesland, Leeuwarden); A.F. Hamel (Stichting Samenwerkende Ziekenhuizen Oost-Groningen, Winschoten); A.T.M.G. Tiebosch (Martini Hospital, Groningen); C. Meijers (Reinier de Graaf Gasthuis/ S.S.D.Z., Delft); R. Natté (Haga Hospital Leyenburg, The Hague); G.A. Meijer (VU University Medical Center, Amsterdam); J.J.T.H. Roelofs (Academic Medical Center, Amsterdam); R.F. Hoedemaeker (Pathology Laboratory Pathan, Rotterdam); S. Sastrowijoto (Orbis Medical Center, Sittard); M. Nap (Atrium Medical Center, Heerlen); H.T. Shirango (Deventer Hospital, Deventer); H. Doornewaard (Gelre Hospital, Apeldoorn); J.E. Boers (Isala Hospital, Zwolle); J.C. van der Linden (Jeroen Bosch Hospital, Den Bosch); G. Burger (Symbiant Pathology Center, Alkmaar); R.W. Rouse (Meander Medical Center, Amersfoort); P.C. de Bruin (St. Antonius Hospital, Nieuwegein); P. Drillenburg (Onze Lieve Vrouwe Gasthuis, Amsterdam); C. van Krimpen (Kennemer Gasthuis, Haarlem); J.F. Graadt van Roggen (Diaconessenhuis, Leiden); S.A.J. Loyson (Bronovo Hospital, The Hague); J.D. Rupa (Laurentius Hospital, Roermond); H. Kliffen (Maasstad Hospital, Rotterdam); H.M. Hazelbag (Medical Center Haaglanden, The Hague); K. Schelfout (Stichting Pathologisch en Cytologisch Laboratorium West-Brabant, Bergen op Zoom); J. Stavast (Laboratorium Klinische Pathologie Centraal Brabant, Tilburg); I. van Lijjnschoten (PAMM Laboratory for Pathology and Medical Microbiology, Eindhoven); K. Duthoi (Amphia Hospital, Breda)