A systematic analysis of the impact of data variation on AI-based histopathological grading of prostate cancer
Supplemental material

Patrick Fuhlert[a,b,1], Fabian Westhaeusser[a,b,1], Esther Dietrich[a,2], Maximilian Lennartz[c,2], Robin Khatri[a,2], Nico Kaiser[a,d,2], Pontus Röbeck[e,2], Roman Bülow[f], Saskia von Stillfried[f], Anja Witte[a], Sam Ladjevardi[e], Anders Drotte[b], Peter Severgardh[b], Jan Baumbach[g], Victor G. Puelles[d,h,i], Michael Häggman[e], Michael Brehler[a], Peter Boor[f], Peter Walhagen[b], Anca Dragomir[j], Christer Busch[b,e], Markus Graefen[k], Ewert Bengtsson[b,l], Guido Sauter[c,3], Marina Zimmermann[a,3], Stefan Bonn[a,b,3,*]

[a]*Institute of Medical Systems Bioinformatics, Center for Biomedical AI (bAIome), Center for Molecular Neurobiology Hamburg (ZMNH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[b]*Spearpoint Analytics AB, Stockholm, Sweden*
[c]*Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[d]*III. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[e]*Department of Urology, Uppsala University Hospital, Uppsala, Sweden*
[f]*Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany*
[g]*Institute of Computational Systems Biology, University of Hamburg, Germany*
[h]*Department of Clinical Medicine, Aarhus University, Aarhus, Denmark*
[i]*Department of Pathology, Aarhus University Hospital, Aarhus, Denmark*
[j]*Department of Pathology, Uppsala University Hospital and Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden*
[k]*Martini-Klinik Prostate Cancer Center, University Hospital Hamburg-Eppendorf, Hamburg, Germany*
[l]*Department of Information Technology, Centre for Image Analysis, Uppsala University, Uppsala, Sweden*

[*]Corresponding author.
[1]These authors contributed equally to this work.
[2]These authors contributed equally to this work.
[3]These authors contributed equally to this work.

S1. Datasets

S1.1. UKE-high-variance (UKEhv) TMAs

The UKEhv cohort provided by the University Medical Center Hamburg Eppendorf contains patients who underwent RP between 1992 and 2014 aged 63.8 ± 6.4 years at the UKE with a FU time of up to 23 years. The cohort's observed median PSA level at the time of RP is 6.9 ng/mL (IQR of 4.8 to 10.5 ng/mL). In total, 17,700 patient samples were collected in the TMA dataset, providing 69,251 images. Patients received an annual follow-up [1]. PSA values were measured following surgery and PSA recurrence was defined as a postoperative PSA of 0.2 ng/mL and increasing at subsequent measurements. Patients without any of these events are considered censored at the last follow-up date. Further, this dataset includes some patients with healthy tissue who therefore did not obtain an ISUP grading.

Building upon this rich information of 17,700 patients, a large variety of 69,251 high-quality images and spots were obtained from different protocols, which represent the foundation for building a robust prediction model in this work. ISUP grades were assigned by examining the whole prostate after RP for every individual patient. After filtering according to the aforementioned criteria, we include 8,157 unique patients and 28,236 TMA spot images as shown in fig. 1B. This extracted dataset consists of images with varying attributes, like multiple spots for the same patient, varying scanners, section thicknesses and staining times and is, to our knowledge, the largest and most variant collection of TMA spot image data paired with rich follow-up data collected to date. We divided the data into six sub-datasets as depicted in fig. S1 that were used for training (UKE-first, UKE-second, and UKE-scanner), and testing (all UKEhv datasets). The sub-datasets are described in more detail in the following. Note that all sub-datasets stem from the same patient population and a single patient can contribute images to multiple sub-datasets. Detailed patient-level information for the UKEhv sub-datasets can be found in table S1.

UKE-first. This sub-dataset, used for training and testing of the BASE and PCAI models (see section 2.5), encompasses 8,123 tissue TMA spots, each selected to represent the most characteristic spot for ISUP grading within each patient. These spots were utilized both as training and testing data. Patient-level ISUP scores were obtained as part of routine diagnostics. The

protocol for digitization followed the standard procedure of the University Medical Center Hamburg Eppendorf (UKE), where tissue samples were sectioned at a thickness of 2.5 µm, stained with Hematoxylin and Eosin for 4 minutes and 1:20 minutes, respectively, and then digitized using an Aperio scanner at a magnification of 40x (0.25 µm/pixel). For training, patient event labels were determined by combining BCR, metastasis, or prostate cancer-related death, with patients without any of these events being censored at the last FU date. The supplemental material shows that only keeping patients that experience BCR does not alter those results.

UKE-scanner. In this sub-dataset, used for training of PCAI and testing of the BASE and PCAI models, TMA images underwent scanning using a 3DHistech scanner. These images were employed both as training and testing data, with ISUP scores retrieved from routine diagnostics. Following the standard digitization protocol of the UKE, the sub-dataset contains 8,114 images scanned at 80x magnification (0.125 µm/pixel).

UKE-second. Each of the 7,156 images in the UKE-second sub-dataset, used for training of PCAI and testing of the BASE and PCAI models, represents a secondary batch of TMA spots from the cancerous area of the prostate. These TMAs were processed at a different time and underwent slight variations in the protocol. The ISUP scores were retrieved from routine diagnostics, and the digitization protocol followed the standard procedure of the UKE.

UKE-thin. The UKE-thin sub-dataset, used for testing of the BASE and PCAI models, comprises 1,602 images, each representing a different TMA spot from the cancerous area of the prostate for every patient. These images were exclusively used as testing data. ISUP scores were determined as part of routine diagnostics. Tissue samples were sectioned at 1 µm thickness, following the standard digitization protocol of the UKE.

UKE-thick. The UKE-thick sub-dataset, comprising 1,574 images and used for testing of the BASE and PCAI models, includes images representing different TMA spots from the cancerous area of the prostate for each patient. ISUP scores were obtained during routine diagnostics, and tissue samples were sectioned at a thickness of 10 µm, in line with the standard digitization protocol of the UKE.

UKE-long. In the UKE-long sub-dataset, used for testing of the BASE and PCAI models, each image represents a different TMA spot from the cancerous area of the prostate for every patient. ISUP scores were determined during routine diagnostics. Tissue samples were stained with Hematoxylin and Eosin for an extended duration of 40 minutes and 10 minutes, respectively, nearly ten times the regular staining time. This experimental sub-dataset contains 1,667 images.

S1.2. External Datasets

S1.2.1. Prostate Cancer Biorepository Network TMAs

To further understand the impact of data variation, we included two additional TMA datasets from the Prostate Cancer Biorepository Network (PCBN) [2] in the USA, collected at the New York Langone Medical Centre (NYU) and the Johns Hopkins Hospital in Baltimore (JHU) (fig. 2B). Note that these datasets were exclusively used for model testing (not for training) and that every patient received RP treatment for all PCBN datasets, similar to the UKEhv TMA dataset. As for the UKEhv TMA data, expert pathologists ISUP only graded the whole prostate instead of individual images. These datasets vary in the scanner used, slice thickness, and overall protocol, constituting an ideal external test TMA dataset for evaluating algorithmic robustness.

NYU. The TMA cohort from New York University (NYU), used for testing of the BASE and PCAI models, contains a total of 204 unique patients arranged in four TMA blocks. ISUP grading is assessed on a patient level and no additional grading details are provided. This dataset includes four TMA blocks of tissue spots (0.6 mm in diameter) from prostatectomy specimens. These spots were sectioned at 5 µm in contrast to 2.5 µm in the internal UKEhv dataset (with the notable exception of UKE-thin and UKE-thick). The TMA block images were digitized using an Aperio scanner with a magnification of 20x (0.5 µm pixel^{-1}) and cut into individual images of size 1817x1817 pixels using `QuPath` [3]. Spots showing non-neoplastic tissue were excluded. After filtering, this work integrated 515 images of 158 patients with a median of 3 images per patient.

JHU. The TMA RP samples from the Johns Hopkins University (JHU), used for testing of the BASE and PCAI models, were derived from two datasets named "Case Natural History of Prostate Cancer" (6 TMA blocks) with 235

patients and "Case PSA Progression" (16 TMA blocks) with 726 patients. ISUP grading is assessed on a patient level and no additional grading details like the number of pathologists are provided. In contrast to the other TMA datasets, the endpoint definition of this dataset in terms of event duration is only accessible in a granularity of years instead of days. These two datasets also contain rich metadata information like age, body mass index, race, local recurrence, etc. that was disregarded in this work's analysis. Moreover, we extended the aforementioned event indications by salvage treatment, leading to a censoring rate for this dataset of under 1 %. This means that this cohort can be considered to be biased towards unhealthy individuals. Also, it expresses the highest ratio of M1 (37.2 %) as well as N1 (18.6 %) patients, which is expected since the "Case PSA Progression" patients all had BCR. This is further emphasized by the highest relapse rate of JHU patients among all TMA spot datasets in the overall KM curves (data not shown). The TMAs were sectioned with a thickness of 4 µm and scanned with a `Ventana DP2005` and a `Hamamatsu NanoZoomer XP6` scanner. For integration, the 22 TMA block images were cut into individual spot images of size 3200x3200 pixels at a magnification of 40x (0.25 µm pixel^{-1}) using `QuPath` [3]. After filtering, this work integrated 3,575 TMA spot images that show prostatic adenocarcinoma from 879 patients, with a median of 4 images per patient.

UKE-sealed. The UKE-sealed TMA dataset, used for testing of the BASE and PCAI models, contains 826 patients and 4,095 images with a maximum of 10 images per patient. This dataset is special, in that it contains spot-level quantitative Gleason grading that included the percentage of Gleason 3, 4, and 5 patterns from GS, as opposed to the prostate-level annotations for spots of all other TMA datasets in this study. The information of quantitative Gleason grades was subsequently used to calculate the spot-wise IQ-Gleason, the currently best-performing clinical PCa grading system, aggregated as the mean or maximum over all images of a single patient[4]. UKE-sealed is therefore the only TMA dataset where we can objectively compare the predictive performance of our algorithm to the ISUP grading, since both utilize the exact same images and information. The name UKE-sealed stems from the fact that the access to all patient, metadata, and outcome information was and is restricted exclusively to the department of Pathology of the UKE. Also the evaluation of TMA spot predictions were conducted exclusively by the department of Pathology of the UKE. Hence, this dataset provides an objective point of performance comparison for the BASE and PCAI models.

S1.2.2. Malmö (MMX) biopsies

The MMX biopsy dataset from Malmö, Sweden, was used for the testing of the BASE and PCAI algorithms and contains 716 patients originally collected by Saemundsson et al. [5]. Whereas all TMA datasets were taken after RP, the biopsies in this dataset were derived from clinical routine, at the stage of initial diagnostics. Patient-level ISUP scores were obtained during routine diagnostics. To further allow for an image-level comparison against the BASE and PCAI models, three individual pathologists assigned image-level ISUP grades to all slides independently and blinded from any additional patient information. The biopsy-level ISUP grades provided by the three pathologists of two centers (Aachen and Uppsala) showed an inter-rater agreement Fleiss kappa of 0.199. The authors removed all patients with no or less than 2 mm of total cancer in their biopsy, missing FU information, inadequate RNA quality, and those that had already developed metastases at the time of diagnosis. Furthermore, for usage in this work, patients that were censored within the first five years as well as images with insufficient quality were removed. In total, 269 patients with 578 images were included in this work, with up to 8 images for a single patient. The time-to-event measurement begins with the biopsy date, leading to longer observed time spans in comparison to the TMA datasets, where the reference point is the date of RP. The images of this dataset were digitized using Hamamatsu and Ventana scanners at 40x magnification resulting in individual slide images with a resolution of $0.23\,\mu\mathrm{m\,pixel}^{-1}$. Image widths and heights vary but consist of up to hundreds of thousands of pixels for the long side of a biopsy.

S1.2.3. Uppsala (UPP) biopsies

The UPP biopsy dataset from Uppsala, Sweden contains 2,611 unfiltered images of 440 patients from the SPROB20 image dataset that was enriched by patient endpoint information and was used for the testing of the BASE and PCAI algorithms [6]. Whereas all TMA datasets were taken after RP, the biopsies in this dataset were derived from clinical routine, at the stage of initial diagnostics. Since some patients in this dataset have had multiple biopsies taken, this work only considers biopsy images from the latest patient visit and excludes all earlier biopsies. ISUP scores were obtained from the pathology report of the fusion biopsies during routine diagnostics. Additionally, patients without an assigned ISUP grade, as well as patients with incomplete or conflicting treatment and FU information were excluded from this dataset. In total, 683 images of 123 patients of this dataset are

included in the evaluation of PCAI, with up to 10 images per patient at point of biopsy. The UPP biopsy slides were sectioned at a thickness of 4 µm to 5 µm and digital images were obtained from a Hamamatsu scanner on a magnification of 40x ($0.25\,\mu m\,pixel^{-1}$). Since this cohort contains patients from a pilot study for MRI-guided acquisition of prostate biopsies the number of missed biopsies may be different, higher or lower, than it would have been if the conventional procedure had been used.

S1.3. Data Splitting

The three larger sub-datasets of the UKEhv data, UKE-first, UKE-second and UKE-scanner, are split into training, validation and test set (70/15/15), and the three smaller sub-dataset UKE-thin, UKEthick and UKE-long are split into validation and test set (50/50). The data is split stratified by the binary 5-year survival indicator. Patients that contribute images to multiple sub-datasets are strictly separated across data splits to avoid leakage. Final numbers per split slightly deviate from the initial percentages since some images were excluded after assigning the split due to the image filter criteria. This work uses the training set of the UKE-first data to train the BASE model and the training sets of the UKE-first, UKE-second and UKE-scanner data to train the DA model. The remaining datasets UKE-sealed, NYU, JHU, UPP and MMX are only used for testing.

S2. PCAI model

S2.1. Preprocessing

Since histopathological images come in arbitrary shapes and sizes and contain a lot of redundant background pixels, we use a masking procedure to define the relevant tissue area in every image for usage in our network. In detail, we first create a tissue mask by separating foreground and background pixels using Otsu thresholding. In the second step, we create an anomaly mask by highlighting all foreground pixels with values outside a predefined deviation of the median of pixel values of the tissue area. This removes pen marks, blood or other undesired areas of the images, which are especially prevalent on the large biopsy images. A patch-based approach was used for our risk prediction network, as is common practice in digital pathology. For this, the images are further cut into equally sized patches of 128x128 pixels at 20x magnification based on the relevant tissue area defined by the masks. We refer to the entirety of n patches of an individual WSI as "patch bag". We then assign a binary label to each patch bag, indicating whether the patient experienced a relapse (defined as biochemical recurrence, additional treatment, metastasis or PCa-related death) in the first 5 years after examination.

S2.2. Filtering

Several filtering steps were performed to generate datasets that match our quality requirements. Figure fig. S13 shows the detailed patient- and image-level filtering steps that were performed on all datasets.

S2.3. BASE model

The baseline risk prediction network BASE is a binary classifier that assigns the probability of having a relapse in the first 5 years after examination to a bag of patches per image (fig. S2A). Since the relapse information corresponds to the full patch bag and no ground truth for individual patches is available, information of patches inside one bag needs to be aggregated. This is referred to as multiple instance learning. In detail, we use the encoder part of `EfficientNet-b0` to extract latent information of all n patches in a bag independently [5]. Next, a self-attention layer (SA), as proposed by Rymarczyk et al. [7], accounts for cross-dependencies between all patches of a bag. For every patch i, n attention weights are computed, resulting in the attention matrix $A_{SA} \in \mathbb{R}^{n \times n}$. A_{SA} contains information about the relevance of each patch i in relation to every other patch j and is multiplied with the incoming

bag feature vector after the encoder. This creates context-aware embeddings from every patch. This bag of patch embeddings is further aggregated into a single latent representation in the attention-based multiple instance learning layer (MIL), as proposed by Ilse et al. [8]. For every patch i, one attention weight is computed, resulting in the attention vector $A \in \mathbb{R}^n$. A softmax function ensures all weights sum to one. Multiplying A with the incoming patch bag yields the aggregated representation of shape $1 \times L$. This method can be seen as a learnable weighted averaging function. Finally, the risk classification head, consisting of two fully connected layers ($1280 \rightarrow 100 \rightarrow 2$ neurons) predicts the probability for both classes, using softmax activation function. The predicted probability for class 1, corresponding to having a relapse prior to five years, represents our final risk score. Figure S2 depicts a schematic of the BASE architecture.

The BASE model is derived exclusively from the UKE-first dataset. We train our network end-to-end using 100 randomly over- or undersampled patches per image with a batch size of 16, Adam optimizer and a learning rate of 2.75×10^{-6} for a maximum of 200 epochs, with early stopping on AUROC5 of the UKE-first validation split data. Dropout rate and stochastic depth of the `EfficientNet` backbone are both set to 0.34. The static number of 100 patches allowed for training with batch sizes > 1 and was chosen to be close to the median number of valid tissue patches across samples in the dataset. Patches were further randomly transformed with AugMix augmentation before input to the network to increase data variance and robustness [9]. We use class-weighted cross-entropy as our loss function. Hyperparameters were optimized for maximum AUROC5 on the UKE-first validation split data using a Bayesian search paradigm. During inference, all valid patches per image and no AugMix augmentations are used. If multiple images of any type are available for a single patient and examination, we aggregate by taking only the highest risk score predicted by our model as the final patient score.

S2.4. Mixing multiple endpoints in training

To evaluate the influence of mixing multiple endpoints during the training process, an analysis is performed where either the BASE model is trained on a combined endpoint of BCR or META or a filtered dataset that only contains BCR events as BASE-BCR. While the former approach might lead to label noise, the latter requires an additional filtering step losing 61 patients that

have documented META as the first observed event. Firstly, the prediction distribution that was described in detail in fig. S14 remains nearly unchanged for the BASE-BCR model. Secondly, the performance evaluation on the UKEhv datasets also yields only small changes and the same behavior - compared to PCAI. BASE and BASE-BCR both show the same tendency and struggle to generalize to the other sub-datasets that were analyzed (see fig. S15).

S2.5. Domain adversarial training

For the DA training, the BASE architecture is extended by a domain discriminator head, as well as a gradient reversal layer (GRL) between MIL layer and domain discriminator (fig. S2B) [9]. We extend our UKE-first training dataset by data from the UKE-second and UKE-scanner sub-datasets and assign a secondary domain label to every image, indicating from which sub-dataset it originates. We then train our DA model in a dual-task manner, where the domain discrimination head aims to correctly predict the sub-dataset a given image stems from. The key concept of domain adversarial training is then applied through the GRL, which serves identity function during the forward pass, however flips the sign of the gradient during back-propagation. This enforces adaptation of the weights of the shared network part, consisting of encoder, SA and MIL layer in the exact opposite direction of the domain discrimination loss. This leads to the desired adversarial game between an consistently improving domain discriminator head and the shared network part, which provides latent representations of the data that contain increasingly less domain-specific information. Since the main task of binary 5-year relapse classification is trained in parallel, this allows the network to provide accurate risk predictions on domain invariant features. This method is inspired by Wilm et al. [9], who prove the positive influence of DA for mitotic figure detection on histopathological images. We optimize an additional parameter λ that controls the influence of our domain adversarial loss, resulting in the overall loss function as a sum of the cross-entropy loss of the risk predictor and the cross-entropy loss of the domain discriminator as

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\text{CE}}_{\text{risk}} + \lambda \mathcal{L}^{\text{CE}}_{\text{domain}}.$$

Training procedure in the DA model is analogous to the baseline model, though here a learning rate of 9.87×10^{-7}, dropout rate of 0.5 and stochastic depth of 0.5 is used. Data from the UKE-first domain was fed twice per epoch

to put a stronger emphasis on the data containing the most representative spot per patient. We further perform early stopping as well as hyperparameter optimization on the combined 5-year AUROC of the validation splits of all UKEhv subdomains.

S2.6. Credibility estimation

To be applicable in an actual clinical setting, the predicted risk score should be accompanied with a notion of trustworthiness that quantifies how certain the model is when predicting on a given image (fig. S2C). For this we introduce the concept of credibility by computing a score for every unseen sample based on the distance to the learned distribution of the model. The underlying assumption is that samples that differ strongly from the data seen during training should receive a lower credibility score than those close to the learned distribution, independent of the actual predicted risk score.

In detail, we measure the Mahalanobis distance d_M between the latent representation of an unseen sample in the output of the MIL layer to the center of the latent representation of all training samples. To further transform the Mahalanobis distance d_M to the training center into a normalized representation of model uncertainty, ideas from the concept of conformal prediction (CP) are employed [10]. CP is a post-hoc method to measure uncertainty in pre-trained prediction models by providing sets of valid class predictions that exceed a given significance level. Here, we first define d_M as the non-conformity measure that assesses the strangeness of an unseen sample. Next, we derive a separate calibration set $S_{\text{calib}} = \{(x1, y1), \ldots, (xn, yn)\}$ with samples that stem from the same distribution as the training data but are unseen to the model. The non-conformity score d_M is computed for every sample in the calibration set. To evaluate how different an unseen sample x_u is from the training distribution, its non-conformity score $d_{M,u}$ is then compared to the non-conformity scores $d_{M,j}$ of the calibration set for both binary 5-year relapse class labels y_c, such that

$$p_c(d_{M,x}) = \frac{|\{j = 1, \ldots, n : y_j = y_c \text{ and } d_{M,j} \geq d_{M,u}\}|}{|\{j = 1, \ldots, n : y_j = y_c\}|}$$

where $p_c(d_{M,u})$ refers to the p-value (distinct from the statistical p-value) for a given class y_c. High p-values indicate high conformity with the training distribution, since most calibration examples expressed higher non-conformity

scores than x_u [11]. The maximum p-value among both 5-year relapse classes is defined as the credibility Cred_u of an unseen sample, such that

$$\mathrm{Cred}_u = \max_u \left[p_c \left(d_{M,u} \right) \right].$$

This credibility score quantifies how close a given sample is to the model's learned distribution, based on the unseen calibration dataset, and is expected to correlate with the validity of the final risk prediction.

The validation split of those UKEhv sub-domains present in the training set serves as the calibration data when applying the credibility estimation setup to PCAI and the baseline model, such that it consists of data from the UKE-first, UKE-second and UKE-scanner datasets for the former and of data from UKE-first only for the latter.

S2.7. Color adaptation (CA)

We propose a cluster-based histogram matching procedure, which Dietrich [12] found to improve over matching randomly to a training domain image. For this, we first derive 8 k-means clusters from the histograms of the training data in the HSV space, using Wasserstein distance as the distance measure. This clustering approach smoothes the effect of outliers while preserving inherent type differences inside the dataset. Using the CE setup described above, we define a threshold on the credibility scores such that 75 % of the calibration set (i.e. the validation data of the training domains UKE-first, UKE-second and UKE-scanner) expresses higher credibility scores. During inference of the PCAI model, we then match the histograms of samples of the test set that express credibility scores below the defined threshold with the histogram of the closest cluster in the training data and feed those adapted samples through the deep learning network again (fig. S3A). The 75 % threshold as well as the number of 8 clusters was chosen by optimizing the increase in AUROC5 on the validation sets of the internal UKEhv sub-domains. Performance metrics of PCAI reported in this manuscript are calculated on the predictions of the resulting combination of raw and color adapted (CA samples, based on their credibility. fig. S3B shows histograms of an exemplary sample from the MMX dataset before and after CA.

Further, we ablated the credibility-guided color adaptation to understand its impact on the grading performance on our internal and external TMA

datasets regarding EOC-Index. Looking into the individual credibility for all datasets in fig. S16, the median credibility of unseen datasets for BASE is almost non-existent with at most 6 % except for UKE-first, PCAI can boost this value to significantly higher median values of 55 % for NYU 33 % for JHU. Figure S8A shows that selective color adaptation for BASE does not significantly alter the results compared to ALL in terms of EOC-Index, with only a slight positive impact on UKE-thick (0.3 percentage points) and UKE-long (0.2 percentage points), while remaining unchanged for all other datasets. Conversely, the impact of color-guided credibility adaptation on PCAI resulted in an EOC-Index increase in six out of eight datasets (fig. S8B). These results further suggest that our algorithmic modifications for robustness and credibility, in conjunction with a high quality training dataset, can improve our AI-based algorithm in terms of robustness for several data variations.

S2.8. PCAI risk groups

To enhance interpretability of our PCAI risk score, we can stratify the patients into risk groups by k-means clustering on risk scores by taking the risk $r_i \in [0, 1]$ for each individual i and perform a 1-dimensional k-means clustering algorithm to obtain k distinct groups of patients [13]. To estimate the maximum number of groups that are statistically significant in terms of outcome, a Fleming Harrington-weighted pairwise log-rank test was used on a separate validation set as suggested by Li et al. [14]. P-values for the pairwise logrank test can be found in fig. S10.

S2.9. Cancer indicator

To indicate cancer-containing regions for our biopsy datasets, the cancer indicator is trained on patch-wise cancer vs non-cancer labels extracted from segmentation masks of the PANDA dataset [15]. Figure S17 shows an exemplary slide of the PANDA dataset with mask overlay in green for healthy tissue and red for cancerous regions with exemplary patches with a side length of 256 pixels with healthy, cancerous, or rejected labels. A total of 4,459,674 training and 504,027 test set patches were extracted from the dataset. The cancer indicator model consists of a CNN-encoder, specifically the `Efficientnet-b0` architecture [16], and a subsequent fully connected classification layer. The cancer indicator is trained using patch-wise labels extracted from segmentation annotations provided by expert (uro-) pathologists in the PANDA challenge. With this, it achieves an AUROC of 0.94 on the PANDA test set patches of previously unseen slides. In the overall

PCAI model, cancer indicator is utilized to reduce noise and redundancy in our risk prediction on biopsies. It is used to predict cancer heatmaps on our biopsy datasets and select the 100 patches with the highest predicted cancer during inference into our PCAI model.

S2.9.1. PANDA biopsy dataset

The PANDA dataset contains biopsy slides and corresponding tissue and cancer annotations of pathologists. This dataset was only used to train the cancer indicator model to find cancerous areas of biopsy images and not for endpoint prediction. PANDA is one of the largest publicly available whole slide image (WSI) datasets in the world with 10,616 provided biopsy slides with slide-level primary and secondary Gleason score as well as ISUP annotation from 2,113 patients. It was published in the Prostate Cancer Grade Assessment challenge and a part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2020 [15]. The training and validation data for this challenge is provided by two centers, the Karolinska Institute in Stockholm, Sweden (5,456 WSIs) and the Radboud University Medical Center in Nijmengen, Netherlands (5,160 WSIs). The PANDA dataset is used to train the cancer indicator based on 9,554 training- and 1,062 test WSIs with 3.94 and 0.51 million extracted patches respectively, which utilizes the expert annotations to classify extracted tissue patches into cancer containing or benign tissue (fig. S17). The main purpose of the cancer indicator is to function as a patch selector for the PCAI model to identify the most relevant patches for risk assessment from up to tens of thousands of patches per biopsy. In addition, the cancer indicator increases the interpretability of PCAI's results by highlighting decision-relevant cancerous areas.

S3. Statistical Evaluation

To evaluate statistical significance of the results, pairwise t-tests are performed for each metric on $n = 1000$ bootstrapped datasets using the `statannotations` package [17]. The null hypothesis of no difference in predictive performance between the approaches was analyzed, using a paired t-test in which either estimator performs better or worse. In detail, the t-statistic is obtained by dividing the mean difference between the metric of interest for the two predictors by the corresponding standard deviation. All instances for an individual are included in the analysis. If multiple instances

per individual are present, they are aggregated using the maximum prediction to ensure that only one prediction for each individual is used.

S3.1. Event Ordered C-Index

To evaluate the algorithms and human annotations on datasets that show more than one type of endpoint, the commonly used C-Index [18] is modified to incorporate for event types of varying severity. Exemplary cases of this modification can be found in fig. S4. To avoid combining those endpoints or filtering out for a single endpoint, a custom C-index evaluation is performed that only evaluates comparable pairs where the former event time shows an event type of higher or equal severity and is not censored. The Event Ordered C-Index provides the probability that a randomly drawn pair is correctly ordered from all possible pairs where the severity of the event type for the individual with a shorter event time is at least not censored and as severe as the event type of the latter event type. Further, Figure fig. S6 shows what kinds of comparisons contribute to the now modified Event Ordered C-Index in our datasets while fig. S5 visualizes what percentage of comparisons is discarded by combining all events with a binary censoring indicator compared to the commonly used C-Index using our modification.

S4. Figures

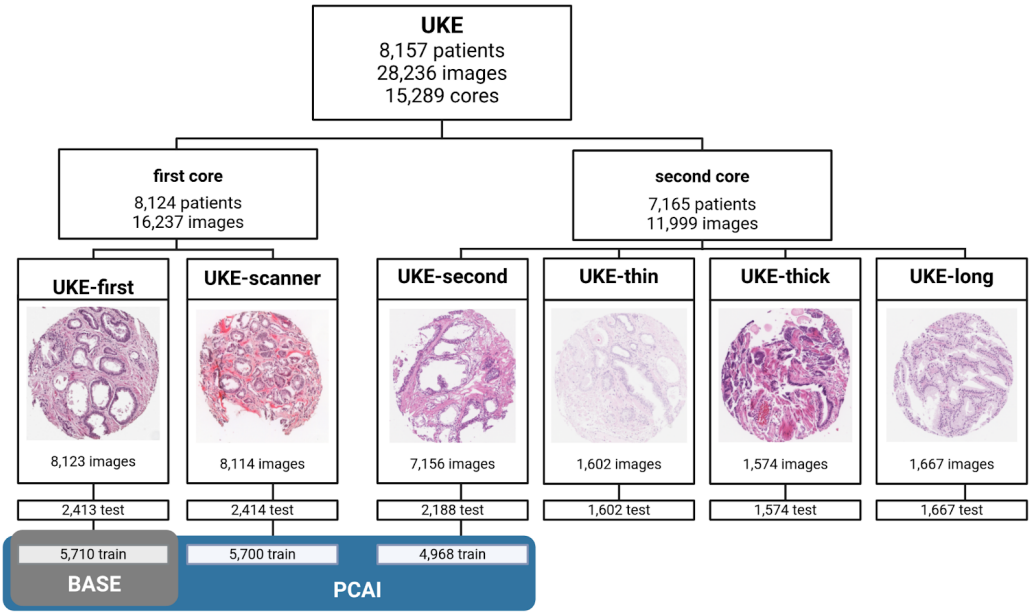


Figure S1: Number of patients and images of the UKEhv sub-datasets per train and test split. The BASE model is trained on TMAs of UKE-first and PCAI on UKE-first, -scanner, and UKE-second training datasets.

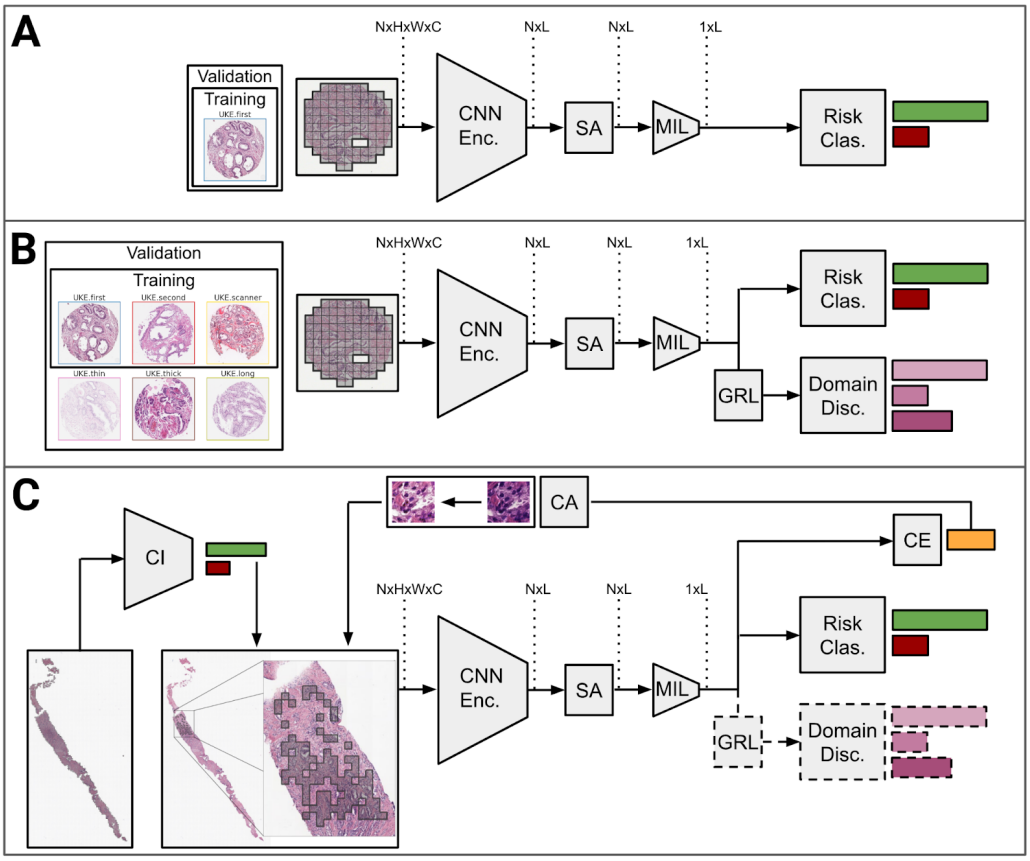


Figure S2: Overview of the architecture and training regime. **A** shows the BASE risk prediction network, **B** the BASE model including the DA module, and **C** PCAI with added cancer indicator based patch sampling, CE, and CA. CNN Enc. = Convolutional neural network encoder; SA = Self-attention layer; MIL = Attention-based multiple instance learning layer; GRL = Gradient reversal layer; Risk Clas. = Risk classifier; Domain Disc. = Domain discriminator; DA = Domain adversarial; CI = Cancer indicator; CE = Credibility estimation; CA = Color adaptation.

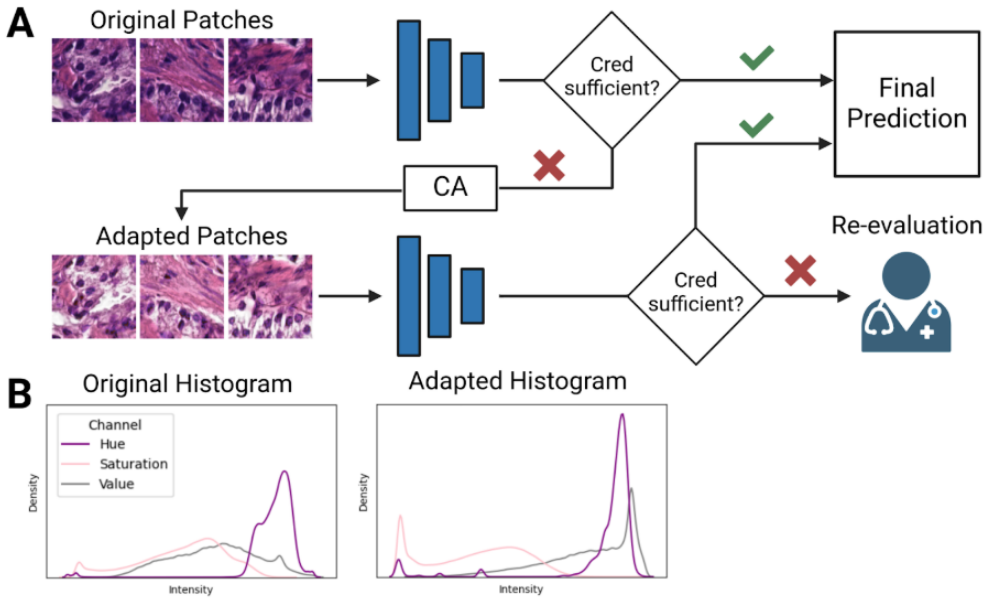


Figure S3: Credibility-guided color adaptation in PCAI. **A** Feedback loop of the credibility-guided color adaptation procedure. If during initial processing of the image in the deep learning network (blue) sufficient credibility is not reached, the color of the problematic sample is adapted by matching its histogram with the training distribution. If sufficient credibility is still not reached, grading of the images can be conferred to the pathologist. **B** Exemplary HSV histograms of a sample before and after applying credibility-guided color adaptation.

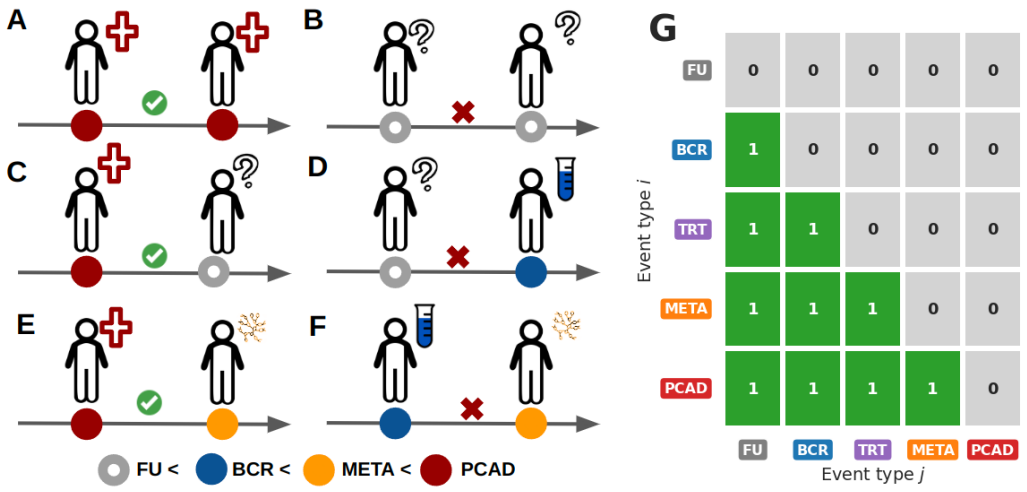


Figure S4: Comparing pairs of patients with varying event types (FU in gray, BCR in blue, TRT in purple, META in orange, PCAD in red - in this order of severity). For the C-Index, comparing a patient with an earlier event time and a non-censored event can be compared to another patient with an observed event of the same severity **A** or a censored individual **C**. Patients with an earlier censoring time (FU) are not compared to any other patient group (**B** and **D**). This work modifies the definition by labeling comparisons as valid between patients that show an earlier event time only to other patients with a longer event time and an event type of lesser severity **E**. If the patient with the earlier event time has a recorded event type that is less severe **F**, the comparison is considered non-valid. **G** shows the remaining allowed comparisons (with a one on green background) between patient i and j with corresponding event times t_i and t_j where $t_i < t_j$ is only included if the event type of patient i is at least as severe as the event type of patient j.

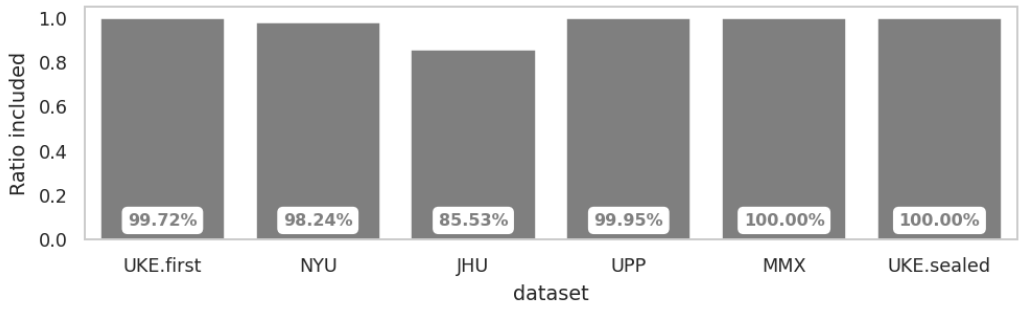


Figure S5: Ratio of comparisons using the EOC-Index compared to the ordinary C-Index. It can be observed that in most datasets only a few comparisons are removed. For JHU which shows the most heterogeneous event type distribution, 14.5 % of comparisons are discarded. The other UKEhv datasets with a similar ratio to UKE-first are excluded to avoid cluttering.

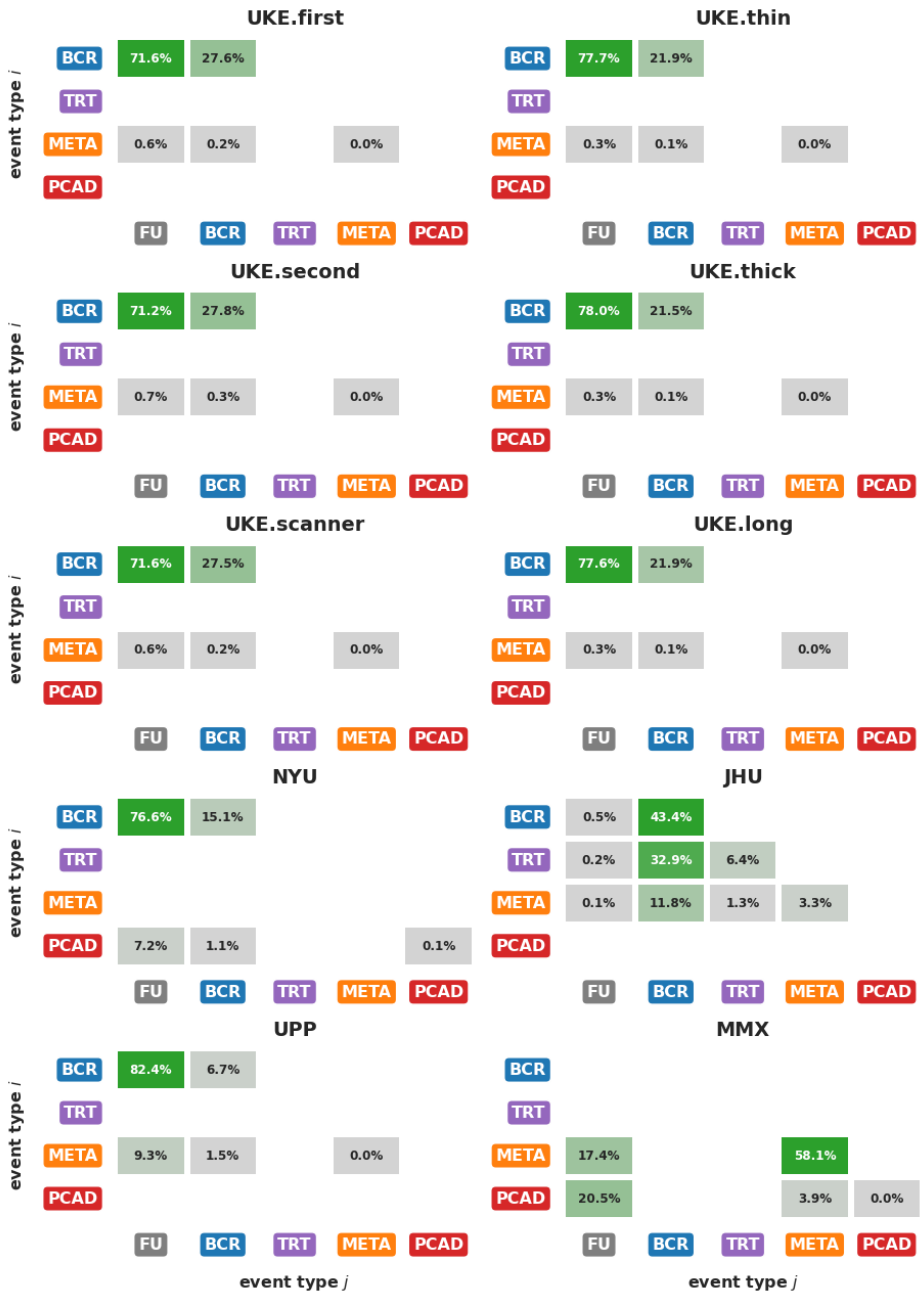


Figure S6: Valid comparisons ratio per event type combination in the event specific C-Index per dataset. The type of the individual i with a shorter event time is visualized on the y axis while the x axis represents the event type for individual j meaning $t_i < t_j$.

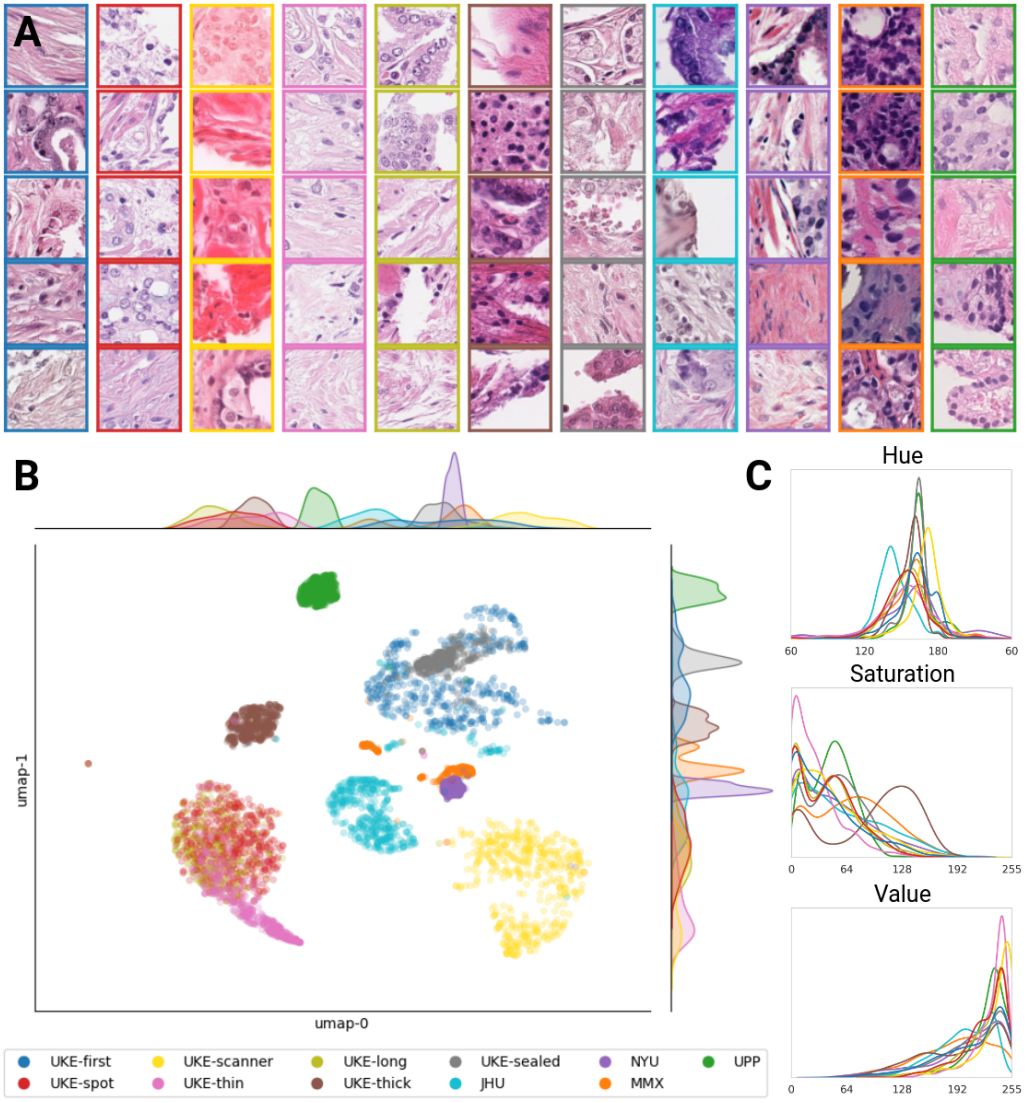


Figure S7: Overview of color variations across all datasets. **A** Example patches of all datasets used in this study. Color-coded margins depict data origin. **B** UMAP of the HSV histograms. **C** Aggregated hue, saturation, and value histograms of all valid foreground pixels of all images per dataset.

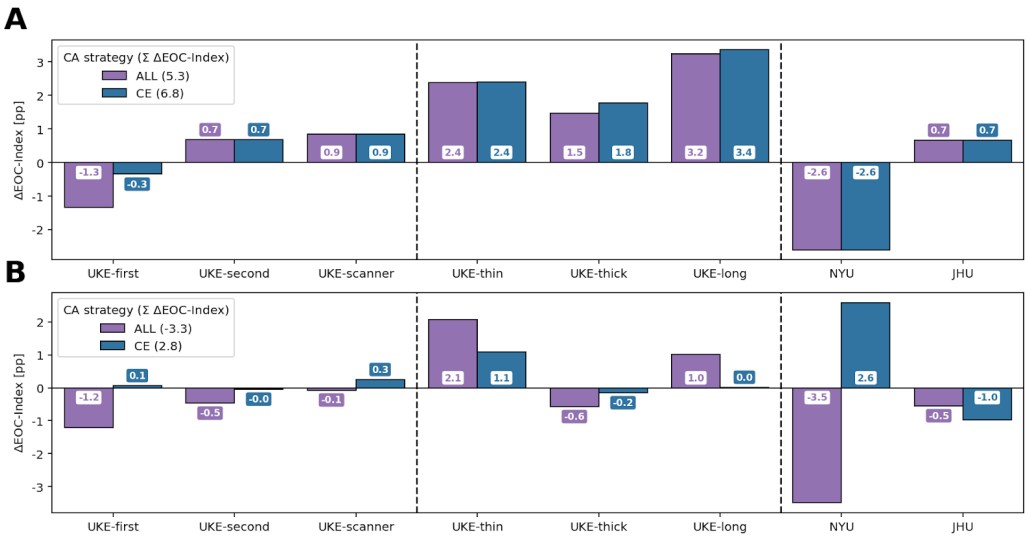


Figure S8: Comparing BASE (**A**) and PCAI (**B**) with respect to credibility-guided color adaptation on the TMA test datasets. with the overall sum of gain or loss in EOC-Index compared to no color adaptation shown in the legend. Vertical dashed lines divide domains used for training, the rest of the internal UKEhv domains, and external domains.

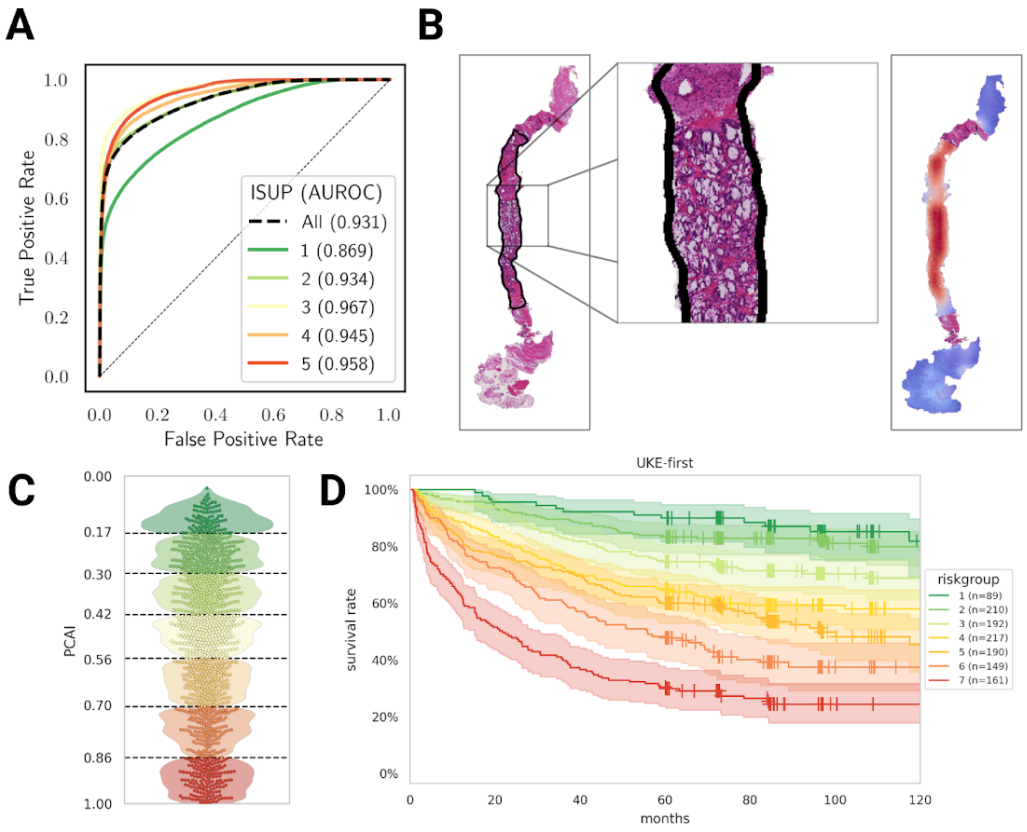


Figure S9: Interpretation of PCAI. **A** Overall and per ISUP patch-wise cancer classification ROC of the cancer indicator on the PANDA test data. **B** Human annotated outline (left) and cancer-indicator-generated heatmap (right) of an exemplary PANDA biopsy sample where blue shaded regions show healthy tissue and red shaded regions contain cancer. **C** PCAI prediction distribution and risk groups for the UKE-first test data. **D** KM curve for stratified predictions of UKE-first that shows clear separation among the risk groups.

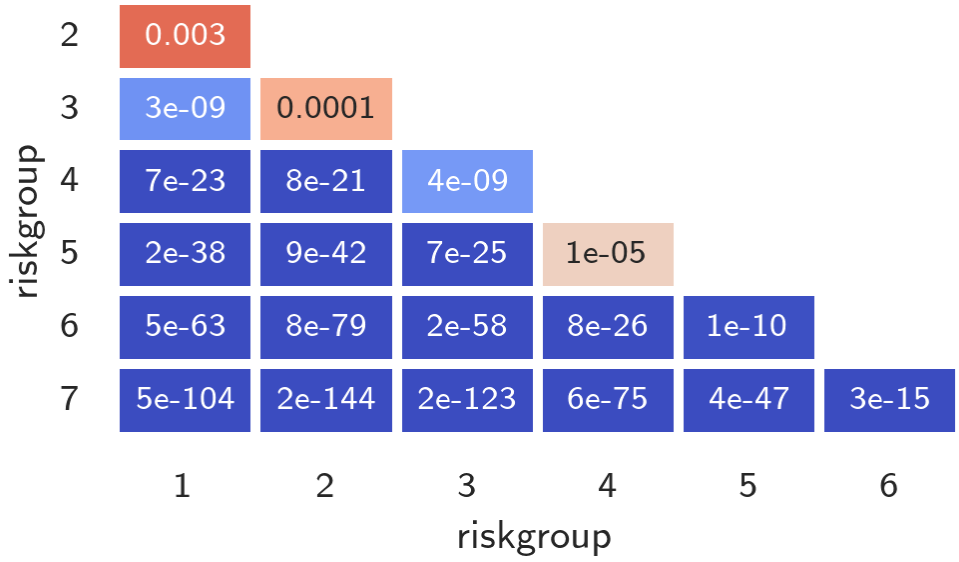
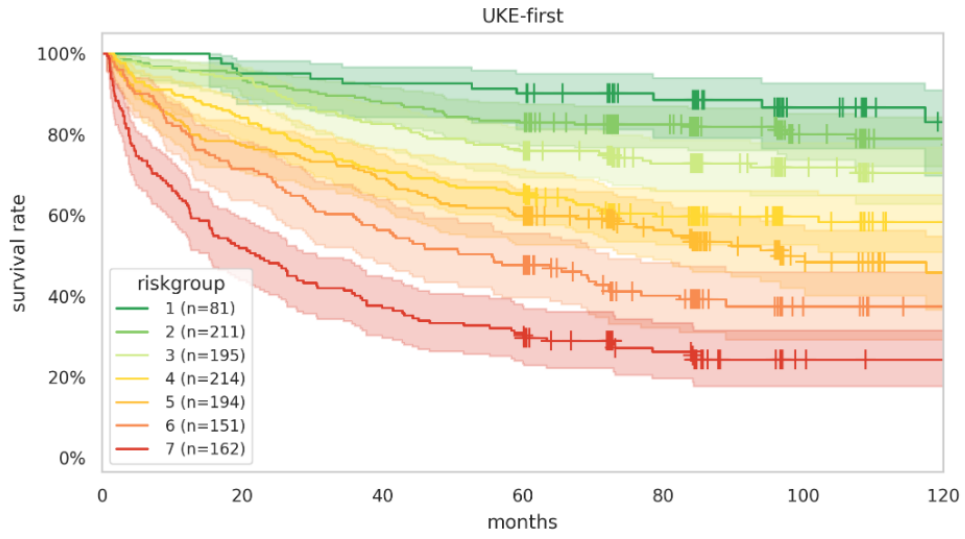


Figure S10: Results of the pairwise log-rank test for the UKEhv datasets based on PCAI predictions. Values for $p < 0.05$ were interpreted as statistically significant.

A



B

PCAI riskgroup		time						
		0	20	40	60	80	100	120
1 (n=81)	At risk	81	77	75	73	57	34	22
	Censored	0	0	0	0	15	37	48
	Events	0	4	6	8	9	10	11
2 (n=211)	At risk	211	198	185	175	139	73	56
	Censored	0	0	0	0	35	98	114
	Events	0	13	26	36	37	40	41
3 (n=195)	At risk	195	183	161	148	110	51	30
	Censored	0	0	0	0	33	91	111
	Events	0	12	34	47	52	53	54
4 (n=214)	At risk	214	180	152	140	85	42	28
	Censored	0	0	0	0	45	88	101
	Events	0	34	62	74	84	84	85
5 (n=194)	At risk	194	150	134	116	78	32	18
	Censored	0	0	0	0	33	72	84
	Events	0	44	60	78	83	90	92
6 (n=151)	At risk	151	108	85	72	41	13	7
	Censored	0	0	0	0	22	48	54
	Events	0	43	66	79	88	90	90
7 (n=162)	At risk	162	84	60	50	29	9	7
	Censored	0	0	0	0	15	33	35
	Events	0	78	102	112	118	120	120

Figure S11: KM curves in **A** with the corresponding at-risk table in **B** for the UKEhv test dataset to visualize the discriminative performance of the PCAI risk grouping.

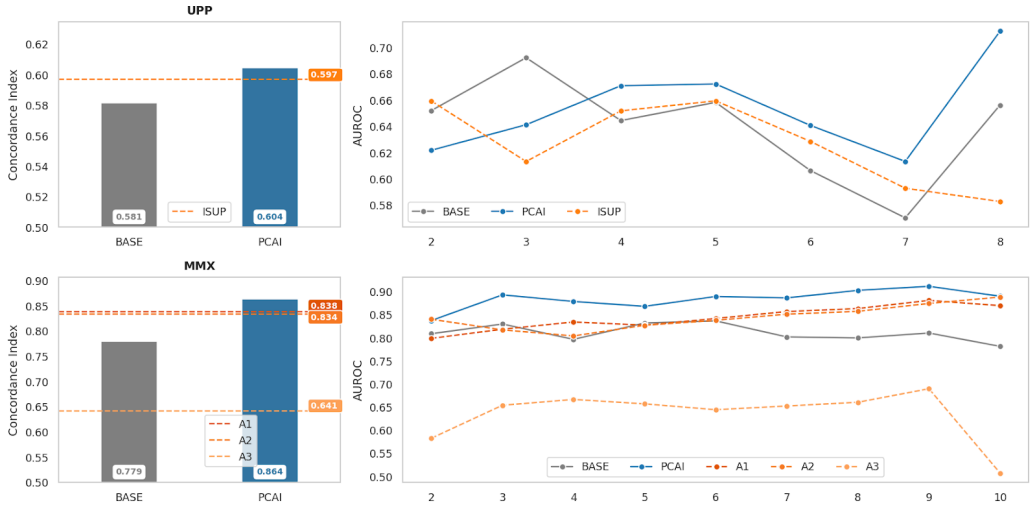


Figure S12: BASE and PCAI performance compared to human annotators for the UPP and MMX biopsy datasets. The 2-10 year AUROC is shown for each prediction (gray for BASE, blue for PCAI) and human annotation (orange shades). It is interesting to observe that for almost every temporal threshold (year) the PCAI model performs best.

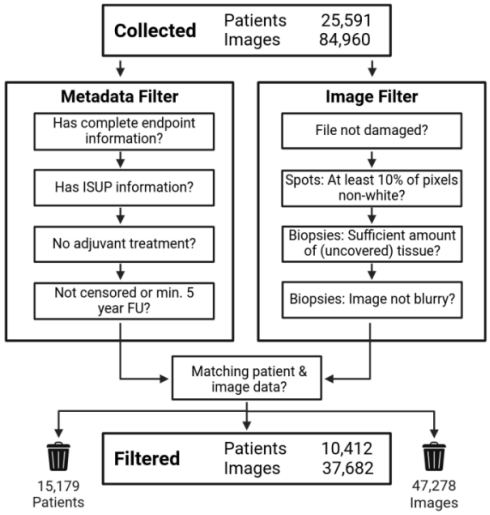


Figure S13: Overview of data preprocessing for the development TMA datasets. Filtering was performed on the individual patient's metadata (sufficient endpoint information, minimum 5 years of follow-up (FU) duration or any observed event, no adjuvant treatment, ISUP information) and image quality (file not damaged, enough tissue on slide, slide not blurry).

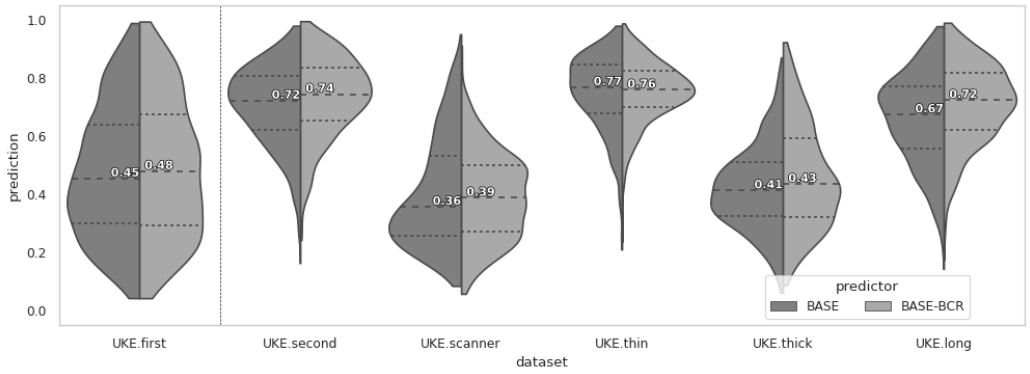


Figure S14: Prediction distribution of BASE and BASE-BCR for the UKE-high-variance sub-datasets on the $n = 1537$ overlapping patients.

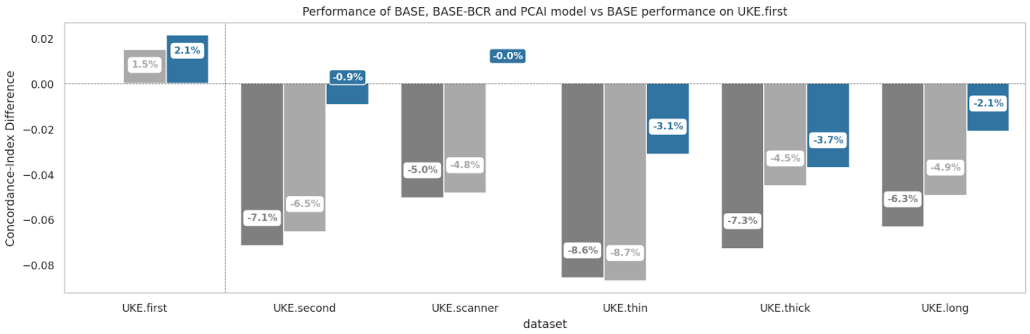


Figure S15: Difference in EOC-Index of BASE (dark gray), BASE-BCR (light gray), and PCAI (blue) compared to that of BASE on UKE-first for the same overlapping $n = 1537$ patients with one image in each sub-dataset.

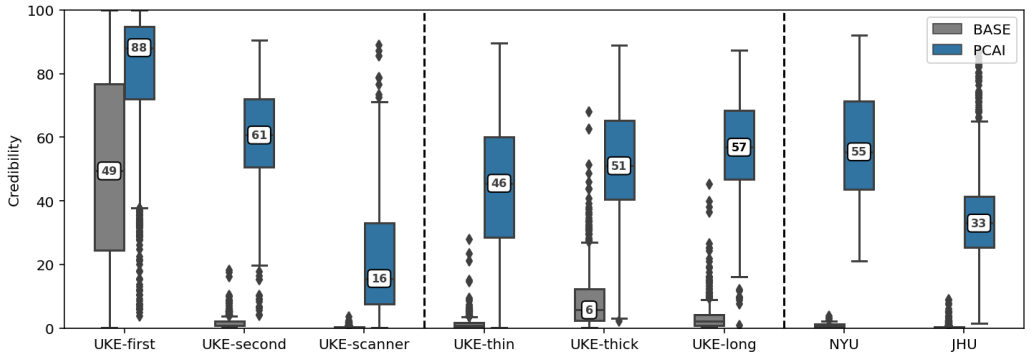


Figure S16: Box plots for BASE (gray) and PCAI (blue) assigned credibilities for all images in the corresponding test datasets.

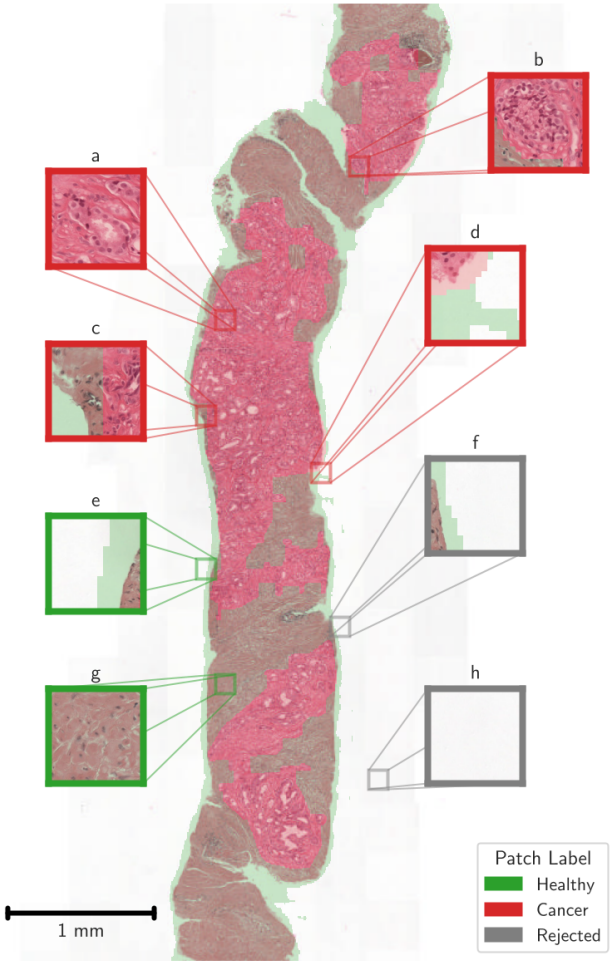


Figure S17: Exemplary slide with cancer and tissue mask of the PANDA dataset. **A**-**H** Visualization of potentially extracted patches from the slide. Gray patches are rejected, red patches are labeled as cancerous, and green patches as healthy. Note that the segmentation masks extend not only in the tissue but also in the background in patches in **D**-**F**.

S5. Tables

		UKE-first	UKE-second	UKE-scanner	UKE-thin	UKE-thick	UKE-long
Patients (images)		8123	7156	8114	1602	1574	1667
Age [years], mean ± SD		63.5 ± 6.1	63.6 ± 6.1	63.5 ± 6.1	63.2 ± 6	63.2 ± 5.9	63.2 ± 6
Censoring [%]		61.3	60.9	61.3	67.4	67.6	67.4
Median survival [years]		1.6	1.6	1.6	2.4	2.4	2.4
Median follow-up [years]		8	8	8	7.2	7.2	7.7
ISUP	0	407 (5.01%)	370 (5.17%)	405 (4.99%)	98 (6.12%)	96 (6.10%)	103 (6.18%)
	1	1792 (22.06%)	1557 (21.76%)	1789 (22.05%)	305 (19.04%)	304 (19.31%)	322 (19.32%)
	2	4001 (49.26%)	3512 (49.08%)	4001 (49.31%)	879 (54.87%)	864 (54.89%)	911 (54.65%)
	3	1366 (16.82%)	1223 (17.09%)	1366 (16.83%)	253 (15.79%)	243 (15.44%)	262 (15.72%)
	4	109 (1.34%)	94 (1.31%)	109 (1.34%)	19 (1.19%)	19 (1.21%)	21 (1.26%)
	5	448 (5.52%)	400 (5.59%)	444 (5.47%)	48 (2.99%)	48 (3.05%)	48 (2.88%)
Event type	BCR	3084 (37.97%)	2745 (38.36%)	3071 (37.87%)	518 (32.33%)	506 (32.15%)	539 (32.33%)
	FU	4978 (61.28%)	4355 (60.88%)	4972 (61.28%)	1080 (67.42%)	1064 (67.60%)	1123 (67.37%)
	META	61 (0.75%)	56 (0.78%)	61 (0.75%)	4 (0.25%)	4 (0.25%)	5 (0.30%)
T-stage	$\leq$ T1	2 (0.02%)	2 (0.03%)	2 (0.02%)	–	–	–
	T2	4940 (60.81%)	4300 (60.09%)	4932 (60.78%)	976 (60.92%)	958 (60.86%)	1021 (61.25%)
	T3	3120 (38.41%)	2857 (39.91%)	3120 (38.40%)	610 (38.08%)	601 (38.19%)	628 (37.67%)
	T4	61 (0.75%)	57 (0.80%)	60 (0.74%)	16 (1.00%)	15 (0.95%)	18 (1.08%)
N-stage	N0	4290 (86.39%)	3716 (86.37%)	4266 (86.37%)	923 (90.22%)	907 (90.25%)	917 (90.49%)
	N1	676 (13.61%)	585 (13.63%)	674 (13.63%)	100 (9.78%)	98 (9.75%)	96 (9.51%)
M-stage	M0	6306 (78.44%)	5499 (77.89%)	6378 (78.43%)	1260 (78.59%)	1237 (78.89%)	1313 (79.05%)
	M1	1733 (21.56%)	1558 (22.33%)	1736 (21.57%)	336 (20.95%)	331 (21.11%)	348 (20.95%)

Table S1: Cohort characteristics across different UKEhv sub-datasets.

(sub-) dataset	#pixels long edge	#pixels short edge	scanner(s)	mag.	µm/pixel
UKE-first	2900 ± 200	2900 ± 200	APE	40x	0.25
UKE-second	2900	2900	APE	40x	0.25
UKE-scanner	6100	6100	3DH	80x	0.125
UKE-thin	2900	2900 ± 100	APE	40x	0.25
UKE-thick	2900	2900	APE	40x	0.25
UKE-long	2900	2900	APE	40x	0.25
UKE-sealed	3100 ± 200	3100 ± 200	APE	40x	0.25
NYU	1800	1800	APE	20x	0.5
JHU	3600	3600	HAM, VEN	40x	0.23
UPP	67100 ± 16300	28800 ± 8100	APE	40x	0.25
MMX	64900 ± 22000	30200 ± 17400	HAM, VEN	40x	0.23
PANDA	26100 ± 8600	15900 ± 8900	APE, 3DH, HAM	20x	0.486

Table S2: Basic image properties of this work's image datasets showing the dataset's tissue type (TMA=T or biopsy=B), mean ± std of the number of pixels on the long and short edge of each image, used scanner vendor (APE=Leica Aperio, 3DH=3DHistech, HAM=Hamamatsu, VEN=Ventana), mag.=maximum magnification level and the resulting physical resolution in µm per pixel.

dataset	predictor 1	predictor 2	p-value	t-statistic
UKE-sealed	BASE	GIQ	1.58e-06	-6.84e+00
UKE-sealed	BASE	ISUP	9.86e-01	-1.79e-02
UKE-sealed	BASE	PCAI	6.44e-07	-7.29e+00
UKE-sealed	ISUP	GIQ	6.96e-12	-1.48e+01
UKE-sealed	PCAI	GIQ	7.33e-01	-3.46e-01
UKE-sealed	PCAI	ISUP	3.85e-06	6.40e+00
JHU	BASE	ISUP	2.72e-19	-9.17e+00
JHU	BASE	PCAI	3.71e-85	-2.16e+01
JHU	PCAI	ISUP	2.09e-27	1.12e+01
NYU	BASE	ISUP	0.00e+00	-7.53e+01
NYU	BASE	PCAI	7.75e-155	-6.19e+01
NYU	PCAI	ISUP	4.86e-277	-5.05e+01
UPP	BASE	ISUP	2.05e-05	-4.28e+00
UPP	BASE	PCAI	6.37e-13	-7.29e+00
UPP	PCAI	ISUP	3.62e-03	2.92e+00
MMX	A1	A2	1.62e-08	5.70e+00
MMX	A1	A3	0.00e+00	1.65e+02
MMX	A2	A3	0.00e+00	1.66e+02
MMX	BASE	A1	1.17e-245	-4.55e+01
MMX	BASE	A2	1.53e-228	-4.29e+01
MMX	BASE	A3	0.00e+00	-5.48e+01
MMX	BASE	ISUP	1.13e-107	-2.50e+01
MMX	BASE	PCAI	0.00e+00	-6.43e+01
MMX	ISUP	A1	2.37e-58	-1.72e+01
MMX	ISUP	A2	1.57e-37	-1.34e+01
MMX	ISUP	A3	0.00e+00	1.25e+02
MMX	PCAI	A1	1.69e-113	2.59e+01
MMX	PCAI	A2	3.11e-169	3.40e+01
MMX	PCAI	A3	0.00e+00	1.78e+02
MMX	PCAI	ISUP	2.14e-184	3.63e+01

Table S3: P-value and t-statistic for EOC-Index comparisons drawn in this work. Each dataset is evaluated $n = 1000$ times using bootstrapping. Statistical comparison was performed with a paired t-test of metric difference between the predictors.

dataset	predictor 1	predictor 2	p-value	t-statistic
UKE-sealed	BASE	GIQ	3.42e-06	-6.46e+00
UKE-sealed	BASE	ISUP	2.27e-01	-1.25e+00
UKE-sealed	BASE	PCAI	1.42e-05	-5.78e+00
UKE-sealed	ISUP	GIQ	2.17e-11	-1.39e+01
UKE-sealed	PCAI	GIQ	3.72e-01	-9.14e-01
UKE-sealed	PCAI	ISUP	7.86e-04	3.99e+00
JHU	BASE	ISUP	8.67e-282	-5.91e+01
JHU	BASE	PCAI	8.31e-186	-2.38e+01
JHU	PCAI	ISUP	2.88e-76	-2.02e+01
NYU	BASE	ISUP	0.00e+00	-6.47e+01
NYU	BASE	PCAI	1.76e-171	-3.44e+01
NYU	PCAI	ISUP	2.34e-156	-3.22e+01
UPP	BASE	ISUP	4.27e-01	7.94e-01
UPP	BASE	PCAI	9.34e-04	-3.32e+00
UPP	PCAI	ISUP	2.41e-06	4.74e+00
MMX	A1	A2	4.86e-01	6.96e-01
MMX	A1	A3	0.00e+00	1.15e+02
MMX	A2	A3	0.00e+00	1.18e+02
MMX	BASE	A1	1.47e-04	3.81e+00
MMX	BASE	A2	2.91e-06	4.70e+00
MMX	BASE	A3	0.00e+00	1.10e+02
MMX	BASE	ISUP	3.21e-25	1.12e+01
MMX	BASE	PCAI	3.86e-111	-2.55e+01
MMX	ISUP	A1	9.26e-16	-8.17e+00
MMX	ISUP	A2	1.92e-14	-7.77e+00
MMX	ISUP	A3	0.00e+00	8.27e+01
MMX	PCAI	A1	5.39e-157	2.33e+01
MMX	PCAI	A2	3.16e-179	3.55e+01
MMX	PCAI	A3	0.00e+00	1.33e+02
MMX	PCAI	ISUP	7.72e-152	3.15e+01

Table S4: P-value and t-statistic for AUROC5 comparisons drawn in this work. Each dataset is evaluated $n = 1000$ times using bootstrapping. Statistical comparison was performed with a paired t-test of metric difference between the predictors.

Glossary

AUROC5 Area Under the Receiver Operating Characteristic Curve. We evaluate at 5 years of survival time. 9, 12, 32

BASE The baseline model is a CNN-based predictive framework trained exclusively on a single internal data domain, UKE-first. 2–10, 13, 16, 17, 22, 26, 27, 31, 32

BCR Possible endpoint, refers to the rise in PSA levels in the blood following radical treatment or radiation for PCa patients, indicating recurrence. 3, 5, 9, 10, 19, 27, 29

C-Index A metric to measure the concordance of a risk prediction with patient survival. 15, 19, 20

CA Color Adaptation 12

CE Credibility Estimation 12

CNN Convolutional Neural Network 13

DA Domain Adversarial Training 10

EOC-Index To enable a meaningful comparison of different endpoints, the concept of an EOC-Index is introduced. 13, 19, 22, 27, 31

FU An endpoint for a PCa patient. Lost to follow-up means he did not experience any relapse. 2, 3, 6, 19, 26, 29

GIQ Extended Gleason score that provides a continuous numerical score to better integrate tertiary Gleason patterns. GIQ is currently one of the best performing grading systems for PCa histopathology. 31, 32

IQR Inter-quartile range. 2

ISUP Simplified PCa grading system defining groups 1–5 with increasing cancer severity to predict disease aggressiveness. 2–6, 14, 23, 26, 29, 31, 32

JHU TMA dataset from the Prostate Cancer Biorepository Network, collected at the Johns Hopkins Hospital in Baltimore, USA, exclusively used for model testing. 4, 5, 7, 13, 19, 30–32

META An endpoint for a PCa patient that developed metastases. 9, 10, 19, 29

MMX Biopsy dataset from Malmö, Sweden, exclusively used for model testing. 6, 7, 12, 26, 30–32

NYU TMA dataset from the Prostate Cancer Biorepository Network, collected at the New York Langone Medical Centre, USA, exclusively used for model testing. 4, 7, 13, 30–32

PANDA Prostate cANcer graDe Assessment (PANDA) Challenge dataset with 10,616 biopsies (2,113 patients) from the Karolinska Institute in Stockholm, Sweden and the Radboud University Medical Center in Nijmegen, Netherlands. 13, 14, 23, 28, 30

PCa Prostate cancer. 5, 8

PCAD An endpoint for a PCa patient that died from the disease. 19

PCAI AI-based PCa detection and grading framework that contains several algorithmic adaptations to increase prediction robustness over the BASE model. Employing domain adversarial training and credibility-guided color adaptation, making it robust to data variation, interpretable, and adding a measure of credibility. 2–7, 10, 12–14, 16–18, 22–27, 31, 32

PSA Protein produced primarily in the prostate gland. It is commonly measured in the blood as a biomarker for prostate health. Elevated PSA levels can indicate (recurring) PCa. 2, 5

RP Surgical procedure performed to treat localized PCa by removing the entire prostate gland along with surrounding tissues. 2, 4, 6

TMA Technique for tissue analysis, consisting of many small cylindrical representative samples, termed spots, that are extracted from paraffin-embedded tissue and are widely used in biomarker discovery and validation studies. 2–6, 12, 16, 22, 26, 30

TRT An endpoint for a PCa patient indicating disease progression by any additional treatment. 19

UKE-first Sub-dataset of UKEhv that includes 8,123 TMA spots following the standard procedure of the University Medical Center Hamburg Eppendorf for tissue digitization, where tissue samples were sectioned at a thickness of 2.5 µm, stained with Hematoxylin and Eosin for 4 minutes and 1:20 minutes, respectively, and then digitized using an Aperio scanner at a magnification of 40x (0.25 µm/pixel). 2, 7, 9, 10, 12, 13, 16, 19, 23, 27, 29, 30

UKE-long Sub-dataset of UKEhv contains TMA spots with nearly ten times the regular staining time when compared to UKE-first. 4, 7, 13, 29, 30

UKE-scanner Sub-dataset of UKEhv scanned with an alternative 3DHistech scanner compared to UKE-first. 2, 3, 7, 10, 12, 29, 30

UKE-sealed Unique TMA dataset used for testing the BASE and PCAI models. Unlike other TMA datasets, UKE-sealed provides spot-level quantitative Gleason grading. Access to patient data and outcomes is restricted to the Department of Pathology at the University Medical Center Hamburg-Eppendorf, and the evaluation of TMA spot predictions is also conducted solely by this department. 5, 7, 30–32

UKE-second Sub-dataset of UKEhv representing a secondary batch of cancerous prostate areas with slight variations in processing protocol compared to UKE-first. 2, 3, 7, 10, 12, 16, 29, 30

UKE-thick Sub-dataset of UKEhv contains TMA spots with a thicker sectioning of 10µm compared to UKE-first. 3, 4, 13, 29, 30

UKE-thin Sub-dataset of UKEhv contains TMA spots with a thinner sectioning of 1µm compared to UKE-first. 3, 4, 7, 29, 30

UKEhv One-of-a-kind dataset of 28,236 PCa histopathological images with variations in section thickness, staining protocol, and scanner, allowing for the systematic evaluation and optimization of model robustness. 2, 4, 7, 10–12, 16, 19, 22, 24, 25, 29

UPP Biopsy dataset from Uppsala, Sweden, exclusively used for model testing. 6, 7, 26, 30–32

WSI Digital high resolution scans of entire tissue slides for histopathological analysis. 8, 14

References

[1] G. Sauter, S. Steurer, S. Clauditz, T. Krech, C. Wittmer, F. Lutz, M. Lennartz, T. Janssen, N. Hakimi, R. Simon, M. V. Petersdorff-Campen, F. Jacobsen, K. V. Loga, W. Wilczak, S. Minner, M. C. Tsourlakis, V. Chirico, A. Haese, H. Heinzer, B. Beyer, M. Graefen, U. Michl, G. Salomon, T. Steuber, L. H. Budäus, E. Hekeler, J. Malsy-Mink, S. Kutzera, C. Fraune, C. Göbel, H. Huland, T. Schlomm, Clinical utility of quantitative gleason grading in prostate biopsies and prostatectomy specimens, European urology 69 (2016) 592–598. URL: http://dx.doi.org/10.1016/j.eururo.2015.10.029. doi:10.1016/j.eururo.2015.10.029.

[2] J. Melamed, N. Y. U. S. of Medicine, Prostate cancer biorepository network (pcbn) (2019).

[3] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-Tellez, P. W. Hamilton, Qupath: Open source software for digital pathology image analysis, Scientific Reports 2017 7:1 7 (2017) 1–7. URL: https://www.nature.com/articles/s41598-017-17204-5. doi:10.1038/s41598-017-17204-5.

[4] G. Sauter, T. Clauditz, S. Steurer, C. Wittmer, F. Büscheck, T. Krech, F. Lutz, M. Lennartz, L. Harms, L. Lawrenz, C. Möller-Koop, R. Simon, F. Jacobsen, W. Wilczak, S. Minner, M. C. Tsourlakis, V. Chirico, S. Weidemann, A. Haese, T. Steuber, G. Salomon, M. Matiu, E. Vettorazzi, U. Michl, L. Budäus, D. Tilki, I. Thederan, D. Pehrke, B. Beyer, C. Fraune, C. Göbel, M. Heinrich, M. Juhnke, K. Möller, A. A. A. Bawahab, R. Uhlig, H. Huland, H. Heinzer, M. Graefen, T. Schlomm, Integrating tertiary gleason 5 patterns into quantitative gleason grading in prostate biopsies and prostatectomy specimens, European Urology 73 (2018) 674–683. doi:10.1016/J.EURURO.2017.01.015.

[5] A. Saemundsson, L. D. Xu, F. Meisgen, R. Cao, G. Ahlgren, Validation of the prognostic value of a three-gene signature and clinical parameters-based risk score in prostate cancer patients, The Prostate 83 (2023) 1133–1140. URL: https://pubmed.ncbi.nlm.nih.gov/36988135/. doi:10.1002/PROS.24530.

[6] P. Walhagen, R. Pontus, B. Ewert, B. Christer, H. Michael, Spear prostate biopsy 2020 (sprob20), 2020. URL: https://datahub.aida.scilifelab.se/10.23698/aida/sprob20.

[7] D. Rymarczyk, A. Borowa, J. Tabor, B. Zielinski, Kernel self-attention for weakly-supervised image classification using deep multiple instance learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1721–1730. doi:10.1109/WACV48630.2021.00176.

[8] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, 35th International Conference on Machine Learning, ICML 2018 5 (2018) 3376–3391. URL: https://arxiv.org/abs/1802.04712v4.

[9] F. Wilm, C. Marzahl, K. Breininger, M. Aubreville, Domain adversarial retinanet as a reference algorithm for the mitosis domain generalization challenge, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 5–13.

[10] V. Vovk, A. Gammerman, G. Shafer, Conformal prediction: General case and regression, Algorithmic Learning in a Random World (2022) 19–69. URL: https://link.springer.com/chapter/10.1007/978-3-031-06649-8_2. doi:10.1007/978-3-031-06649-8_2.

[11] T. Pereira, S. Cardoso, D. Silva, A. de Mendonça, M. Guerreiro, S. C. Madeira, Towards trustworthy predictions of conversion from mild cognitive impairment to dementia: A conformal prediction approach, Advances in Intelligent Systems and Computing 616 (2017) 155–163. URL: https://link.springer.com/chapter/10.1007/978-3-319-60816-7_19. doi:10.1007/978-3-319-60816-7_19/FIGURES/2.

[12] E. Dietrich, Deep learning-based discrete-time survival prediction on prostate cancer histopathology images (2022). URL: https://ediss.sub.uni-hamburg.de/handle/ediss/10251.

[13] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, Oakland, CA, USA, 1967, pp. 281–297.

[14] H. Li, D. Han, Y. Hou, H. Chen, Z. Chen, Statistical inference methods for two crossing survival curves: a comparison of methods, PLoS One 10 (2015) e0116774. doi:10.1371/journal.pone.0116774.

[15] W. Bulten, K. Kartasalo, P. H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, C. H. van de Kaa, J. van der Laak, M. B. Amin, A. J. Evans, T. van der Kwast, R. Allan, P. A. Humphrey, H. Grönberg, H. Samaratunga, B. Delahunt, T. Tsuzuki, T. Häkkinen, L. Egevad, M. Demkin, S. Dane, F. Tan, M. Valkonen, G. S. Corrado, L. Peng, C. H. Mermel, P. Ruusuvuori, G. Litjens, M. Eklund, A. Brilhante, A. Çakır, X. Farré, K. Geronatsiou, V. Molinié, G. Pereira, P. Roy, G. Saile, P. G. Salles, E. Schaafsma, J. Tschui, J. Billoch-Lima, E. M. Pereira, M. Zhou, S. He, S. Song, Q. Sun, H. Yoshihara, T. Yamaguchi, K. Ono, T. Shen, J. Ji, A. Roussel, K. Zhou, T. Chai, N. Weng, D. Grechka, M. V. Shugaev, R. Kiminya, V. Kovalev, D. Voynov, V. Malyshev, E. Lapo, M. Campos, N. Ota, S. Yamaoka, Y. Fujimoto, K. Yoshioka, J. Juvonen, M. Tukiainen, A. Karlsson, R. Guo, C. L. Hsieh, I. Zubarev, H. S. Bukhar, W. Li, J. Li, W. Speier, C. Arnold, K. Kim, B. Bae, Y. W. Kim, H. S. Lee, J. Park, Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge, Nature Medicine 2022 28:1 28 (2022) 154–163. URL: https://www.nature.com/articles/s41591-021-01620-2. doi:10.1038/s41591-021-01620-2.

[16] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 36th International Conference on Machine Learning, ICML 2019 2019-June (2019) 10691–10700. URL: https://arxiv.org/abs/1905.11946v5.

[17] F. Charlier, M. Weber, D. Izak, E. Harkin, M. Magnus, J. Lalli, L. Fres-

nais, M. Chan, N. Markov, O. Amsalem, et al., getzze, repplinger s (2022) statannotations (v0. 6). zenodo, 2022.

[18] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the yield of medical tests, 1982. URL: https://jamanetwork.com/journals/jama/fullarticle/372568https://jamanetwork.com/. doi:10.1001/JAMA.1982.03320430047030.