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

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

Abstract

[*] Corresponding author.
Email address: sbonn@uke.de (Stefan Bonn)

[1] These authors contributed equally to this work.

[2] These authors contributed equally to this work.

[3] These authors contributed equally to this work.

The histopathological evaluation of biopsies by human experts is a gold standard in clinical disease diagnosis. While recent artificial intelligence-based (AI) approaches have reached human expert-level performance, they often display shortcomings caused by variations in sample preparation, limiting clinical applicability. This study investigates the impact of data variation on AI-based histopathological grading and explores algorithmic approaches that confer prediction robustness. To evaluate the impact of data variation in histopathology, we collected a multicentric, retrospective, observational prostate cancer (PCa) trial consisting of six cohorts in 3 countries with 25,591 patients, 83,864 images. This includes a high-variance dataset of 8,157 patients and 28,236 images with variations in section thickness, staining protocol, and scanner. This unique training dataset enabled the development of an AI-based PCa grading framework by training on patient outcome, not subjective grading. It was made robust through several algorithmic adaptations, including domain adversarial training and credibility-guided color adaptation. We named the final grading framework PCAI. We compare PCAI to a BASE model and human experts on three external test cohorts, comprising 2,255 patients and 9,437 images. Variations in sample processing, particularly section thickness and staining time, significantly reduced the performance of AI-based PCa grading by up to 8.6 percentage points in the event-ordered concordance index (EOC-Index) thus highlighting serious risks for AI-based histopathological grading. Algorithmic improvements for model robustness, credibility, and training on high-variance data as well as outcome-based severity prediction give rise to robust models with grading performance surpassing experienced pathologists. We demonstrate how our algorithmic enhancements for greater robustness lead to significantly better performance, surpassing expert grading on EOC-Index and 5-year AUROC by up to 21.2 percentage points.

Keywords: Cancer Grading, Deep Learning, Digital Histopathology, Robustness

1. Introduction

Recent advancements in digital pathology, including the introduction of high throughput digital slide scanners, hold the potential to improve histopathological evaluation of various diseases [1]. The possibility to use algorithms to automatically analyze pathological samples is not only time and cost effective but offers the potential for a standardized, objective, and accurate evaluation, providing crucial insights into tumor characteristics, aiding in treatment decisions, and assessing disease progression as for example shown in the MIDOG challenge [2]. This offers the opportunity to an efficient and reproducible histopathological assessment, thereby optimizing diagnostic accuracy and streamlining workflows in diagnosis and care [3]. The biggest hurdle for automated histopathology grading, possibly, is the variation of histopathological protocols, which can compromise the robustness of algorithms [4].

Processing tissue for digitization consists of several steps, including tissue formalin fixation, paraffin embedding, sectioning, staining, and digitization with a slide scanner. Each step involves numerous parameters that can vary between clinics, research institutions, and even within the same laboratory, leading to variations in the appearance of tissue in the images. While the potential negative impact of data variation spans all areas of histopathology, these effects have been prominently observed in breast and prostate cancer. Several recent studies that use AI-based PCa grading of histopathological slides show reduced performance on external test data, which might indicate overfitting on the training data and lack of robustness to data variation [1]. The detection of PCa areas, however, seems to be robust to data variation and is already in clinical use [5, 6]. AI-guided PCa severity prediction aims to grade the severity of prostate cancer on a biopsy in accordance with the ISUP standard, which features five groups (1-5) of increasing cancer severity [7]. While recent studies showed human expert-level performance on internal validation datasets, they achieved slightly lower performance on external validation data [8, 9, 10, 11]. Bulten et al. [8], for instance, achieved a quadratic Cohen's Kappa (qk) of 0.85 on internal validation data, whereas the qk on the external validation data decreased to 0.72 and 0.716. A recent benchmark for ISUP grade prediction is the PANDA challenge [12], which featured heterogeneous biopsy data from several international clinics for training and testing 11. Out of over 1,000 participant groups, the best model achieved a qk of 0.88 on internal and 0.83 on the external validation data, rivalling the concordance between expert pathologists (0.82). Many automated PCa grading approaches are based on replicating the ISUP score, which allows for the measurement of concordance of prediction between models and human graders [8, 13, 14]. While the AI-based predictions of ISUP grades laid the necessary foundation, the obvious caveat is the perpetuation of human error and bias, which is why recent studies started concentrating on modeling the probability of relapse-free survival over time [15, 16].

First steps towards a systematic investigation of the impact of scanner variation were taken in the 2021 MIDOG challenge, using 300 histopathology images of breast cancer [2]. MIDOG comprised a training set of 200 cases digitized with four scanners and a test set of 100 cases digitized with two additional scanners. The study demonstrated that domain shifts between whole slide imaging scanners can be largely mitigated using advanced data augmentation strategies. The top-performing method exceeded human experts (maximum F1 score of 0.693) and the provided baseline approach (F1 score of 0.718) with an F1 score of 0.748 on the test set, utilizing Fourier-domain mixing for augmentation. However, the challenge was limited to selected regions of interest and variations in scanners, whereas whole slide images exhibit considerable variability in other aspects of sample preparation, including out-of-focus areas and necrosis. Given the potential data variation-induced degradation of the predictive performance of AI-based histopathology grading, concepts of model trustworthiness have recently been proposed [2, 17]. Model trustworthiness allows for the deferral of grading problematic samples to a human expert. While many promising steps towards the automated assessment of histopathological data have been taken, the challenge of model robustness to data variation requires further attention [1, 18].

Recent developments in histopathology showed the transition from CNN-based encoders to vision-transformers [19] that generate higher quality patch embeddings. Some examples of foundation models for histopathology are, among others, UNI [20], Midnight [21], or Virchow [22]. However, even though those models seem to create a better latent representation of individual patches, they show a lack in robustness [23, 24] comparable to conventional CNNs. Moreover, the aggregation of patches to slide-level, especially for biopsies that may contain thousands of patches, still needs further investigation [20].

In this work, we first systematically assess the impact of data variation on AI-based histopathology grading and subsequently show how algorithmic improvements can result in robust model predictions with state-of-the-art performance (fig. 1). As a use case, we aimed to predict PCa aggressiveness for individual images using a unique multi-centric, retrospective, observational trial that contains six cohorts with 25,591 patients with at least five years of follow-up (FU), 83,864 images from 5 different centers and 3 countries. An important part of this data is a unique high-variance cohort of 8,157 patients with 28,236 scanned tissue microarrays (TMAs) with variations in section thickness, staining protocol (i.e. differences in staining time or concentration [25]), scanner, as well as variation in patient cohort distribution. The data also contains patient relapse information for 5 years on average, which our algorithms use as an objective measure for cancer aggressiveness prediction, instead of replicating the subjective ISUP score. To enable a meaningful comparison of different relapse endpoints, the commonly used concordance index (C-Index) is extended to account for multiple possible endpoints in our newly developed event-ordered concordance index (EOC-Index). Using this dataset, we systematically evaluated the robustness of AI-based PCa grading to data variation. We observed severe performance degradation for variations in e.g. section thickness and staining time with our BASE model. We show how a select set of algorithmic improvements, including domain adversarial training and credibility-guided color adaptation, conferred robustness to data variation, leading to significant prediction improvement across all evaluated datasets. This robust model, PCAI, always outperformed ISUP grading pathologists with varying levels of experience up to expert level in both the EOC-Index and area under the 5 year receiver operating characteristic curve (AUROC5) across one external TMA and two external test cohorts.

2. Methods

2.1. Biological Background of PCa

To investigate the impact of data variation on model robustness we use a very large and diverse PCa histopathology dataset. PCa ranks among the most prevalent cancers in men, with approximately 1.4 million new cases worldwide each year with its incidence steadily increasing over the past decades [26]. Due to the wide variety in growth rates of PCa, histopathology plays a central role in the diagnosis and management of PCa. The Gleason score is the most relevant prognostic feature in prostate cancer TMAs and biopsies [27] based on Gleason grading [28] that focuses on the glandular structure of the tissue. The International Society of Urological Pathology (ISUP) has proposed a simplified system defining ISUP groups 1-5 with increasing cancer severity to predict disease aggressiveness [7], which are used to guide the urologist in treatment decisions. Unfortunately, even between expert pathologists the concordance in Gleason grading suffers from high interobserver variability, leading to possible over- or under-treatment due to the subjective nature of the visual assessment [29]. The standard for clinical diagnosis is ISUP grading of preoperative biopsies, typically obtained through transrectal ultrasound-guided biopsy. Multiple tissue samples are collected from different areas of the prostate gland to improve cancer detection rates. After biopsy, specimens are formalin-fixed and paraffin-embedded to preserve structure and stained with Hematoxylin and Eosin to enhance cellular visibility for pathologist examination. Biopsies have long edge lengths in the order of 60,000 pixels with a total of up to 10 billion pixels per image. All biopsies in this work contain image-level ISUP grades by expert pathologists.

In addition to biopsies, this work uses postoperative TMAs from radical prostatectomies. TMAs consist 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. TMA spots of this work typically have edge lengths of 3,000 to 6,000 pixels with resulting images that contain in the order of 10 million pixels and are much smaller than biopsies. TMA spots are preselected to represent a patient's cancer status. All ISUP grade annotations for TMA spots were retrieved from the routine diagnostic reports made by expert pathologists' examinations of the resected whole prostate to derive a patient-level annotation. Therefore, TMA spots might only partially capture a patient's cancer morphology, with the notable exception of the UKE-sealed dataset, which contains TMA spot-based ISUP grades for all individual images (see section 2.2).

2.2. Data acquisition and endpoint definition

To the best of our knowledge, we collected the biggest and most heterogeneous histopathological PCa datasets to date, with a total of 81,572 TMA spot images and 3,388 biopsy images retrospectively collected from 25,591 patients of five different clinics with FU information of up to 23 years and a maximum of 8 images for a single patient (fig. 2). This dataset is divided into several subsets acquired with different parameters and used for training or testing the model on images with high variation. Detailed information on demographics and metadata distributions can be found in table 1. Datasets that were used to build and assess robustness of the proposed model are shown in fig. 2B. All image-level annotated, unseen datasets that were exclusively used for evaluation are depicted in fig. 2C.

The largest subset is the UKE high variance cohort provided by the University Medical Center Hamburg Eppendorf which contains 17,700 patients who underwent radical prostatectomy (RP) with a FU time up to 23 years. This unique dataset allows us to assess differences in acquisition protocol parameters and represents the foundation for building a robust prediction model in this work.

As a quantifiable measure correlating with cancer aggressiveness that does not rely on subjective human annotations, we combine the earliest reported indication of disease progression in model training to keep the maximum number of patients. Further filtering to only train on a pure endpoint did not alter the training results (see Supplemental material). For evaluation, we combine multiple endpoints with our EOC-Index. This novel index allows for the clinically meaningful comparison of different endpoints and is further explained in section 2.6.1. Possible cancer aggressiveness-related events are biochemical recurrence (BCR) based on elevated PSA levels, further unplanned treatment (TRT), developing metastasis, and PCa-related death with corresponding event times relative to the date of RP for TMA spots or to the date of the biopsy. If none of those exist, the follow-up time is used as the censoring time (FU).

We applied the same metadata and image quality-based filtering steps to all datasets (TMA and biopsy). In brief, we limited patient inclusion to patients that experienced BCR or any other indicator of disease progression (treatment, metastasis, or PCa-related death) or had at least 5 years of FU data available. All patients with a documented adjuvant treatment were removed. Additionally, we excluded images with insufficient quality (e.g. too blurry) from the analyses, as summarized in fig. 2A.

All datasets used in this study were collected in strict accordance with ethical guidelines and compliance regulations. Data collection was approved by the relevant institutional review boards or ethics committees. Informed consent was obtained from all participants involved in data collection processes or the need for informed consent was waived by the local ethics review board. Additionally, any information pertaining to participants was anonymized or de-identified prior to analysis. We will now describe the datasets used in this study in detail. The datasets are split stratified by event indicators to keep the same censoring rate across data splits and, since patients can have multiple TMA spot images, we strictly separate patients across data splits for model training and evaluation. This means that TMAs of the same patient are present in either the training or test data but never in both. An overview of the patient characteristics for the UKEhv sub-datasets can be found in table S1 while table S2 depicts the respective image characteristics.

2.3. UKE-high-variance (UKEhv) TMAs

The core of this work is the diverse UKEhv dataset, created specifically to reflect the many possible variations in the tissue processing and imaging pipeline. 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 up to 23 years. In total, 17,700 patient samples were collected in the TMA dataset, providing 69,251 images. Patients received an annual FU [27]. PSA values were measured following surgery and BCR was defined as a postoperative PSA of 0.2 ng/mL and increasing at subsequent measurements. Patients without any recorded event are considered censored at the last FU 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 unique variety of 69,251 high-quality images and spots were obtained with different protocols, which represent the foundation for assessing the impact of data variation on AI-based histopathology grading. ISUP grades were assigned by examining the whole prostate after RP for every individual patient. After filtering (see fig. 2A) according to the aforementioned criteria, we include 8,157 unique patients and 28,236 TMA spot images as shown in fig. 2B (details in Supplemental material). This extracted dataset consists of images with varying attributes, like multiple spots for the same patient, varying scanners, section thicknesses, and staining times (possibly containing further variations in the staining protocol like concentration), and is, to our knowledge, the largest and most variant collection of TMA spot image data paired with rich FU data collected to date. UKEhv contains TMAs processed with the standard protocol (UKE-first, Leica Aperio AT2 scanner, 2.5 µm thickness, stained for 4 min with hematoxylin and 1:20 min with eosin respectively), using a different lot of reagents and tissue core (UKE-second), an extended staining (UKE-long, 40 min hematoxylin and 10 min eosin respectively), thinner (1 µm, UKE-thin) and thicker (10 µm, UKE-thick) sectioning (2.5µm standard), and a different scanner vendor (UKE-scanner, 3DHistech) that are described in additional detail in the Supplemental material (fig. S1, table S2). Staining concentrations of the individual datasets are unknown. For training, only three of those sub-datasets (UKE-first, UKE-second, and UKE-scanner) are used. Note that all sub-datasets stem from the same patient population and a single patient can contribute images to multiple sub-datasets, while making sure that a single patient is always only part of either the training, validation or test set. Detailed patient-level information for the UKEhv sub-datasets can be found in table S1.

2.4. External Datasets for Evaluation

To allow for an in-depth evaluation of model robustness, this work includes TMAs and biopsies from five additional sources. Firstly, we included two TMA datasets (see fig. 2B) from the Prostate Cancer Biorepository Network from the New York Langone Medical Centre (NYU) and the Johns Hopkins Hospital in Baltimore (JHU) where we compare to patient-level ISUP annotations. We further included one TMA dataset (UKE-sealed, graded by expert pathologist Prof. Dr. med. Guido Sauter) as well as two biopsy datasets from Uppsala (UPP, additionally graded in retrospect by two experienced pathologists and one expert) and Malmö (MMX), Sweden that contain image-level ISUP annotations to fairly compare to our algorithm (See fig. 2C). A detailed analysis of these datasets can be found in the Supplemental material (section S1.2).

2.5. PCAI design rationale

The PCAI model describes our overall risk prediction algorithm, illustrated in fig. 1, with the following technical novelties:

Objective endpoint training Training labels are restricted to objective patient outcomes (event time and type), avoiding subjective grading by pathologists.

Mixed endpoint training The model is trained on patients with multiple endpoints (FU, BCR, TRT, META, PCAD), increasing dataset size and diversity (section 2.2).

Event-Ordered C-Index (EOC-Index) A novel evaluation metric enabling C-Index calculation across mixed endpoints (section S3.1).

Domain adaptation Domain adversarial training [30, 31] (DA) discourages separation of UKEhv sub-datasets in the model's latent representation (section S2.5).

Credibility-guided color adaptation Color adaptation (CA) [17] is applied when inputs deviate from the training distribution, guided by conformal prediction [32]. This modification allows a feedback loop (fig. S3) that we call credibility estimation (CE) (fig. S3 and sections S2.6 and S2.7).

Patch pre-selection for biopsies Although trained only on TMAs, PCAI can process biopsies by pre-selecting cancerous patches via our cancer indicator module (section S2.9).

The remaining part of this section describes our technical novelties in additional detail:

The subjective nature of ISUP evaluation causes shifts in the evaluation of cancer severity that depend on the human annotator. We therefore aimed to make the algorithm robust to subjective human annotations and errors thereof. We hypothesized that this is best possible if the model learns how to predict objective patient outcomes over time instead of replicating a subjective ISUP grade by experts. To this end, all datasets used in this work contain at least 5 years of follow up information for all patients. We then grade the cancer by predicting a potential disease relapse in the future. As an additional benefit of this objective end-point prediction, it is conceivable that the model's risk prediction performance can potentially exceed that of the subjective ISUP scoring system. Notably, utilizing the large number of smaller TMA spot images allowed us to train our network in an end-to-end fashion and adapt all parts of our pipeline specifically to the task at hand, without the need to rely on pre-trained models or self-supervised methods, as is often the case in histopathological deep learning applications. An additional feature of our algorithm is its interpretability, which lets human experts understand and trust model predictions. We achieve this via a cancer indicator module, which highlights and selects cancer regions of the sample with an accuracy of over 95 % as assessed on the PANDA dataset (details in Supplemental information section S2.9). This allows the algorithm to process arbitrarily sized images due to the patch-based image processing and aggregation in our model.

We first developed BASE, our cancer aggressiveness model to derive assessments of cancer risk based on individual images of arbitrary sizes. Our algorithm leverages the morphological features learned by training a CNN-based (EfficientNet[33]) neural network architecture with postoperative TMA-spot image data and corresponding patient FU information to allow for a valid risk prediction on clinically more relevant, preoperative biopsy images. It additionally utilizes a patch-level self-attention network and an attention-based patch aggregation that is finally utilized for an objective five-year relapse risk estimation. Our BASE model is exclusively trained on the single internal data domain UKE-first, containing the most representative TMA spot per patient, according to the collecting pathologist. Additional details of the BASE model can be found in the Supplemental material section S2.3.

While the BASE model features a Convolutional Neural Network backbone, extensive image augmentation, patching and patch selection, as well as outcome-based cancer aggressiveness prediction, it lacked sufficient robustness when faced with data variation (see section 3). As a leading cause for BASE model prediction errors are variations in the processing of histopathological samples, our main focus was to render it robust to these changes by implementing and developing several algorithmic adaptations, which lead to the creation of the PCAI model (fig. 1).

To increase model robustness, we utilized UKE-first, UKE-second, and UKE-scanner datasets of the UKEhv cohort and applied several techniques to increase robustness (with domain adversarial training (DA) and color adaptation (CA)), trustworthiness credibility estimation, (CE). The combination of these techniques represents our final risk prediction model PCAI (fig. 1, fig. S2B-C).

First, we hypothesized that DA training [30, 31] on our large and heterogeneous UKEhv dataset will result in a more compact latent representation and more stable predictions across unseen datasets and domains that reflect the variance encountered in everyday clinical practice (details in Supplemental Information section S2.5 and Supplemental fig. S2).

Second, even though PCAI has been trained and optimized for stable predictions across different sample processing protocols, it might still encounter histopathological slides of e.g. bad quality, for which it cannot provide a reliable grading. A relevant feature for PCAI is therefore the notion of confidence or trustworthiness via CE, not unlike a human expert that is uncertain about the grading of a particular sample and asks for a second opinion (details in Supplemental Information section S2.6 and Supplemental fig. S3).

Third, with the aim to stabilize PCAI on images where it shows low credibility, we further introduce a novel CE-guided CA procedure that maps the color scheme of low-credible samples to that of the model's training distribution. As we later show in fig. 3, this additionally improves overall performance and robustness. If probes are still problematic after CA, the grading of them can be deferred to the pathologist. (details in Supplemental Information section S2.7 and Supplemental fig. S3).

Further details of the full algorithm, including deep-learning architecture, hyperparameter and training details are provided in the Supplemental Information section S2.

2.6. Performance evaluation and statistics

Due to the differences in cohort composition and metrics, this work only draws comparisons within and not across patient cohorts. As is standard in all of our analyses, the training and test datasets did not contain any overlapping patients to avoid overestimation of model performance. Further, to estimate performance variation, significance, and provide confidence intervals, each dataset is evaluated 1,000 times using bootstrapping. To determine result significance between the predictions, pairwise t-testing for significantly different means in the analyzed metrics (EOC-Index, AUROC5) was performed. For additional information and p-values of all performed comparisons, refer to the Supplemental material tables S3 and S4.

2.6.1. Evaluation Measures

The algorithms of this work are evaluated and compared to human annotations on datasets that contain patients with multiple possible endpoints in contrast to the mostly homogeneous endpoint of BCR (or FU) that was observed in the training data. For the datasets of this work, a patient's event type is one of FU, BCR, TRT, META or PCAD in that order of severity. To allow for the meaningful comparison of patients with different endpoint severity, we introduce the novel EOC-Index. As shown in the supplemental section S3.1 and fig. S4, we remove comparisons in the C-Index such that an event type with lower severity, but shorter survival time is not compared to a corresponding event type with higher severity and longer survival time, making such a comparison irrelevant for the evaluation. As an example, a patient that encounters BCR with a shorter event time than a patient that is indicated as developing metastasis (META) should not be compared to each other (see Supplemental fig. S4F). Further, fig. S5 illustrates that our EOC-Index only removes a small fraction of comparisons per analyzed dataset (with a maximum of 1.76 % for NYU) except for JHU where 14.47 % of comparisons are removed. Consequently, fig. S6 shows what percent of comparisons are then performed for all event types by our EOC-Index for each individual dataset. In addition to the EOC-Index, we use the AUROC5 to measure human and model performance.

3. Results

3.1. The effect of data variation on the BASE model

Data variation can significantly reduce the performance of AI-based decision systems. This problem exists across biomedical domains, ranging from genomics to image processing and beyond [2, 31, 34, 35]. To assess the influence of different variations of WSIs on predictive performance, this work utilizes a unique dataset of histopathological slides of PCa patients. In contrast to the dataset of the 2021 MIDOG challenge[2], which focused exclusively on the impact of scanner variation, this work analyses multiple sources of data variation such as formalin fixation and paraffin embedding of tissue, sectioning, staining, and slide scanning.

In this work, we aimed to systematically investigate the effect of data variation on AI-based algorithms, employing our extensive and heterogeneous PCa dataset as a use case. As an example, we evaluate PCa grading performance, using the UKEhv dataset that contains TMAs processed with variations to the standard protocol (see Supplemental material section S1.1). Notably, the dataset contains all protocol variations for a subset of the same 1,537 patients, taken from the same RP sample, constituting an ideal basis for the evaluation of data variation on model performance. As can be observed in fig. 4A, the UKE high-variance datasets show marked differences in the visual appearance, which is confirmed through further analysis in fig. S7.

First, we trained the BASE model (fig. S2A) on the UKE-first training set and evaluated its performance on all six UKEhv test datasets. For simplicity, multiple endpoints that indicate disease progression are combined to prevent further filtering of the dataset during training (see Supplemental material section S2.3). Although the BASE model is trained on the large UKE-first training dataset of 8,123 images and patients, a significant decrease in prediction performance is seen across all data variations. While the BASE model achieves a EOC-Index of 0.653 on the UKE-first dataset, the performance drops significantly by 5.0 (UKE-scanner) to 8.6 (on UKE-thin) percentage points on the same patient population for other TMAs in the UKEhv dataset (fig. 4B).

These results strongly indicate that AI-based models, even when trained on large datasets and using image augmentation[36], have significant difficulties with data variations they were not trained on. Standard operating procedures for histopathology vary with the location but also over time at the same medical center, which constitutes a central problem for the robustness and fidelity of AI-based histopathological grading systems.

3.2. Conferring algorithmic robustness to data variation

An algorithm can never be trained on all current and future data variations. It is therefore pivotal to stabilize AI-based PCa grading by using algorithmic modifications such as domain adversarial training (fig. 1). Even when stabilized algorithmically and trained on large datasets, a model might still encounter probes that are difficult for it to grade. A relevant feature for a model is therefore the notion of prediction confidence or trustworthiness, not unlike a human expert that is uncertain about the grading of a particular sample and asks for a second opinion. A model that assesses the confidence and estimates the credibility of its predictions is able to fix problematic probes using color adaptation or defer them to a pathologist (fig. S3A-B). With the aim to enable valid predictions even on images where PCAI shows a low credibility, a color adaptation setup to map the color of those images to the color scheme of the training distribution is established (fig. S2C, fig. S3A, fig. S7). fig. S7 shows that color is a strong separator between the datasets used in this work. We included this novel concept of credibility-guided color adaptation in our algorithm by assigning a credibility along with the cancer risk prediction for each individual input image. If a sample shows too little credibility, we use color adaptation and repredict with the intention to move that image closer to our training data. In summary, we implemented algorithmic modifications for robustness, credibility, and interpretability into the BASE model, naming it PCAI (fig. 1, fig. S2C), and again assessed the performance on the UKEhv dataset. We trained PCAI on the sub-datasets UKE-first, UKE-second, and UKE-scanner, consisting of 19,883 images and 6,937 patients, and evaluated the performance on the same test datasets as the BASE model.

PCAI showed significantly increased grading performance in the EOC-Index across all UKEhv data variations (fig. 4). In brief, the performance of PCAI on the UKEhv test datasets increased by 2.1 percentage points on UKE-first (0.653 to 0.674), 6.2 on UKE-second (0.581 to 0.643), 5.1 on UKE-scanner (0.602 to 0.653), 5.5 on UKE-thin (0.566 to 0.622), 3.6 on UKE-thick (0.580 to 0.616), and 4.2 on UKE-long (0.590 to 0.632) over BASE (fig. 4B). We next evaluated the performance of the BASE and PCAI models on the external NYU and JHU datasets. Again the mean performance of PCAI increased by 5.3 percentage points on NYU (0.643 to 0.696, $p < 1e-4$) and 1 percentage point for JHU (0.577 to 0.587, $p < 1e-4$) over BASE (fig. 4C-D). Finally, it is worth noting that PCAI marginally but significantly outperformed ISUP grading in the EOC-Index on the JHU dataset (0.576 vs 0.573, $p < 1e-4$), while performing worse than ISUP on NYU dataset (0.696 vs 0.763, $p < 1e-4$). The JHU results are surprising as the ISUP grades for all TMA datasets are patient-based and reflect the status of the whole prostate after RP, while PCAI relies exclusively on a single TMA spot for its prediction.

3.3. Assessing the impact of credibility-guided color adaptation

To obtain more detailed insights into which algorithmic components confer robustness to data variation, we conducted an ablation analysis of PCAI. fig. 3A illustrates what percentage of credibility-guided images per dataset can be significantly decreased using PCAI's robustness extensions compared to BASE to provide a more focal correction of problematic images. While almost all images are color adapted for BASE outside UKE-first, PCAI adapts significantly less and only targets 4 % of images in UKE-long, for instance. Note that the values for the external datasets of 20 % for NYU and 67 % for JHU fall in between the percentage of color adapted images for the internal UKEhv datasets (min for UKE-second with 3 %, max for UKE-scanner with 68 %).

Next, we ablated the credibility-guided color adaptation to understand its impact on the grading performance on our internal and external TMA test datasets regarding AUROC5. We compare adapting all images (ALL) to our credibility estimated approach (CE). fig. 3B shows that selective color adaptation for BASE does have a positive impact on UKE-first (-1.7 percentage points for ALL; $-.5$ percentage points for selective color adaptation). Further, it does not significantly alter the results compared to ALL in terms of AUROC5 on the other TMA datasets, with only a slight positive impact on UKE-thick (0.5 percentage points increase) and UKE-long (0.1 percentage points increase), while remaining unchanged for all other datasets. Conversely, the impact of credibility-guided color adaptation for PCAI resulted in an increase in seven out of eight datasets (fig. 3C). Only UKE-thick (-0.3 percentage points) showed a marginal decrease in AUROC5 after selective adaptation. Importantly, while adapting all images leads to an overall increase of 3.0 percentage points, credibility-guided color adaptation leads to an overall gain of 7.9 percentage points. Comparable results can additionally be observed regarding EOC-Index in supplementary fig. S8.

In summary, these results strongly suggest that algorithmic modifications for robustness and credibility, in conjunction with a high quality training dataset, can improve AI-based algorithms to be more robust for a broad, and previously unseen, spectrum of data variations. Our novel approach of credibility-guided color adaptation proved both effective and quick in adapting problematic images.

3.4. Human interpretation of PCAI's predictions

It is pivotal that AI-based clinical decisions support systems are interpretable by human experts. Interpretability allows the expert to trust or ignore a model prediction. PCAI delivers at least two highly interpretable results. First, PCAI's cancer indicator provides visual cues on images as to the location of cancerous areas (fig. S9A-B). Second, PCAI can distinguish 7 different risk groups in order to transform the continuous cancer grade score into categories that may be more appreciated for clinical interpretation than continuous data (fig. S9C-D).

In detail, the cancer indicator allows for patch-wise prediction of cancer probability. The cancer indicator was trained on the PANDA dataset and achieves an AUROC of 0.94 on the PANDA test data (fig. S9A). Notably, it can be observed that the cancer indicator achieves a slightly lower AUROC (0.869) in ISUP group 1 as compared to all other ISUP groups (AUROC> 0.93). The cancer indicator enables us to create cancer probability heatmaps on all images, which provide visual means of interpretability that can also be used as an automatic quantitative indicator of how much cancerous tissue is present in the images (fig. S9B). Especially on the biopsy images, where we use the cancer indicator to focus PCAI on the relevant cancerous regions, this transparently highlights the amount of cancerous tissue on any given biopsy and consequently the salient regions that contributed to the final risk score.

PCAI's cancer grading is a continuous score from 0 (low risk) to 1 (high risk), which poses the challenge to define the appropriate clinical decisions for a given score. To potentially enhance the interpretability of the predicted risk score, we derive 7 statistically distinct risk groups from the UKEhv training data by using k-means clustering (fig. S10). The Kaplan-Meier curves of the UKE-first test data show a clear separation of the patients in our proposed risk groups (fig. S9D, fig. S11).

In summary, we equipped our risk prediction model PCAI with the necessary tools to provide robust (domain adversarial training, color adaptation), trustworthy (credibility estimation) and interpretable (cancer indication, risk groups) predictions, laying the foundation for actual clinical application. We now set out to evaluate if PCAI can rival or exceed the current clinical standard of ISUP ratings, on one TMA and two external biopsy test datasets with significant data variation.

For the following sections, we test for statistical significance in terms of performance between models and human pathologists using pairwise t-tests on bootstrapped (n=1000) results as described in section 2.6.

3.5. PCAI surpasses expert ISUP grading on TMAs

A main aim of the patient outcome-based PCa grading prediction of the PCAI model lies in its potential ability to exceed the current five tier ISUP grading, if sufficiently robust to data variation. We therefore assessed the performance of the BASE and PCAI models on the UKE-sealed dataset, which is the only TMA dataset that contains spot-level ISUP grading from UKE pathologists. 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 data analysts were blinded to all patient, metadata, and outcome information, which was exclusively handled by the department of Pathology of the UKE.

In addition, the UKE-sealed TMA spots were graded using the IQ Gleason (GIQ) grading system. The GIQ is currently one of the best performing grading systems for PCa histopathology [37]. In the case of multiple TMAs for a single patient in the UKE-sealed dataset the average GIQ was calculated as the patient-level prediction. Similarly, the BASE and PCAI model were evaluated taking the mean aggregated image-wise predictions and compared to the worse ISUP grade. On UKE-sealed, PCAI significantly exceeds BASE performance by 3.3 percentage points (EOC-Index of 0.713 to 0.746, fig. 5A, $p < 1e - 4$). Moreover, PCAI significantly outperformed the five-tier ISUP grading by 3.3 percentage points in EOC-Index (fig. 5A, $p < 1e - 4$). Comparably, PCAI significantly exceeds BASE performance by 3.3 percentage points regarding AUROC5 (0.790 to 0.757, fig. 5B, $p < 1e - 4$). PCAI significantly outperformed ISUP grading by 2.6 percentage points in AUROC5 (fig. 5B). Importantly, PCAI performed comparably to the GIQ grading (0.2 percentage points worse in EOC-Index, $p = 0.67$, and 0.5 percentage points worse in AUROC5, $p = 0.11$). In this context it is interesting to note that the BASE model almost always performed worse than ISUP and GIQ grading.

These results indicate that the algorithmic adaptations in PCAI resulted in above ISUP grading in EOC-Index and AUROC5 on TMAs.

3.6. PCAI surpasses expert ISUP grading on biopsies

The litmus test for PCAI is whether its robustness and performance extends to clinical, preoperative biopsy PCa samples. In this final proof of concept, the model has to be robust to considerable protocol variation (scanner, staining, thickness, clinical endpoints) as well as data variation stemming from significantly larger whole slide images (WSI) as compared to the TMAs the model was trained on (see Data acquisition and endpoint definition section). To this end, we evaluate PCAI on the two biopsy cohorts from two clinical centers in Sweden, UPP (123 patients and 683 images) and MMX (269 patients and 578 images). To focus PCAIs risk prediction on relevant tissue areas, we preselect the 100 most likely cancerous patches per whole slide image using PCAI's cancer indicator module. If multiple WSIs per patient were available for a single biopsy, the maximum risk score across images was used as the patient-level prediction.

PCAI achieved an EOC-Index of 0.604 on the UPP dataset, which is 0.7 percentage points higher than ISUP (0.597, $p = 0.0092$) and 1.9 percentage points higher than BASE (0.585, $p < 1e - 4$) (fig. 5C). Similarly, PCAI reached an AUROC5 of 0.672 on the UPP dataset, which exceeds ISUP (0.660, $p < 1e - 4$) by 1.2 and BASE (0.663, $p = 0.0003$) by 0.9 percentage points (fig. 5D).

On the MMX data, PCAI achieved an EOC-Index of 0.865, 4.7 percentage points higher than ISUP (0.818) and 8.5 percentage points higher than BASE (0.780, $p < 1e - 4$). Using AUROC5 as a measure of performance, PCAI (0.869) exceeded ISUP (0.813) predictions by 5.6 and BASE (0.832, $p < 1e - 4$) predictions by 3.3 percentage points.

To obtain a notion of human inter-rater variability and further assess significance, we also compared model performance to the image-wise ISUP grading of three highly skilled pathologists from Germany and Sweden. In EOC-Index, PCAI (0.865) significantly exceeded the performance of expert ISUP grading (A1: 0.839, A2: 0.834, A3: 0.641, mean: 0.771, $p < 1e - 4$ for all comparisons) by 9.4 (mean) percentage points (fig. 5E). This holds also true for AUROC5, where PCAI (0.869) significantly surpasses expert ISUP

(A1: 0.827, A2: 0.826, A3: 0.657, mean: 0.770, $p < 1e - 4$ for all comparisons) grading by 9.9 (mean) percentage points (fig. 5F). Notably, even though PCAI is trained on 5 year relapse, it scores consistently high for every relapse cutoff point in time (fig. S12).

Taken together, PCAI significantly outperforms ISUP grading on all biopsy datasets and outscores individual experts in each comparison we conducted. These results on biopsy-derived whole slide images substantiate our findings on UKE-sealed TMAs, highlighting the importance of our algorithmic adaptations to exceed the five-tier ISUP-level cancer grading on multiple, highly variable external test datasets.

4. Discussion

This work systematically analyzes the influence of data variation on the robustness of AI-based model predictions. To evaluate the impact of data variation in histopathology, we collected one of the largest multicentric, retrospective, and observational PCa trials. We show how data variation significantly diminishes the grading performance of a state-of-the-art BASE model, even when trained on a large dataset of 8,123. By improving the BASE model with algorithmic extensions for robustness, we developed a model that is stable across five external test datasets and reaches above expert grading performance, naming it PCAI.

Our adaptations to introduce more robustness to the BASE model, namely joint domain adversarial training and credibility-guided color adaptation, together with the outcome-based endpoint prediction, showed a consistent and significant improvement in predictive performance across all datasets. Especially on the JHU, UPP and MMX cohort, these adaptations provided the decisive edge to surpass the predictive performance of ISUP grading. The high separability of the PCAI risk groups derived from the UKE data, which is also visible in the respective Kaplan-Meier curves, proves a strong correlation of PCAI's output score with the actual patient endpoints. Interpretability is further enhanced by the reported credibility score for every sample, which provides a notion of trustworthiness of our model's prediction.

To evaluate our algorithm on datasets with heterogeneous endpoints, another source of data variation that can reduce model robustness, we developed the

EOC-Index, which takes the severity of an endpoint into consideration and focuses the C-Index on medically valid comparisons between individual patients, while making use of the full dataset. We believe that the concept of the EOC-Index can easily be extended to other endpoint-based comparisons, well beyond the histopathology use case described here.

Before decision support systems such as PCAI can be generally adapted in clinical routine, they have to prove their ability to handle the large data variability and give an accurate measurement of their confidence providing more robust results and therefore strengthening the trustworthiness in the system. Although the first support systems have been approved by regulatory agencies, their application is still limited by the data they were trained on. Our work is unique, because we systematically evaluated the robustness of AI-based histopathological PCa prediction when faced with variations in data acquisition and processing. Even when trained on thousands of patients and tens of thousands of images, the models show significantly diminished predictive performance when subjected to variations in slide thickness and staining time. The introduction of domain adversarial training and credibility-guided color adaptation, in conjunction with training data of higher variability, resulted in robust model performance in EOC-Index and AUROC5 across multiple high-variability and three external test datasets. It is important to note that PCAI was exclusively trained on TMA data and generalized with state-of-the-art performance to the PCa grading of external test biopsies, which contain considerable data variation in size, scanner, staining, and endpoint composition.

These results, however, do not imply that PCAI cannot fail on yet unseen data, which underscores the importance of credibility estimation by conformal prediction. In this work we introduce the concept of credibility-guided color adaptation, an algorithm that uses credibility estimates to fix low-credible samples by adapting their color to the training distribution. By utilizing a measure of trustworthiness, PCAI is able to first adapt difficult samples and subsequently either grade them, or confer biopsies that remain problematic after adaptation to the pathologist. It is interesting to observe that credibility-guided color adaptation clearly outperforms color adaptation of all samples, while it also reduces the workload of the algorithm and leaves room for the deferral of problematic samples to the clinical expert. This concept can be easily extended to other areas of AI research. This case study shows how training a DL model can benefit from robustness extensions to perform reasonably well on previously unseen datasets. Taking a step back, this strategy is not unique for the usage of AI in prostate cancer or histopathology images. It is the rule and not the exception that the noisy process of image acquisition leads to differences in datasets. As shown in the 2021 MIDOG challenge[2], different scanners can have a huge influence on model performance, for instance. This covariate shift can either stem from digitization artifacts introduced by the scanner, or from other alterations in the sample processing workflow. While our results present solutions on how robustness to data variation and credibility estimation might be achieved, they also highlight a lingering weakness of AI-based systems, in general. The negative impact of data variation on AI-based clinical decision support systems requires further attention and future solutions should include collaborative data sharing strategies, the establishment of robust data standards, the development of further algorithmic strategies, and the leveraging of larger and more heterogeneous training datasets.

While it is interesting to observe the limitations that data variation poses to state-of-the-art AI algorithms, it is perplexing to realize that the systematic evaluation of how human experts deal with data variation is largely lacking. To the best of our knowledge, an evaluation of how expert graders handle differences in staining, thickness, and image resolution of the same histopathological samples, for instance, is missing.

In addition to the domain adversarial training and credibility-guided color adaptation, we included objective label utilization to provide more robust model predictions. The recent progress in AI-based PCa grading is considerable, resulting in several cancer detection and grading systems that have been approved for clinical use. Most systems, however, have been trained on human annotations for cancerous regions or cancer grade according to the ISUP and Gleason systems. This poses two limitations to the algorithms, as they are bound by subjective expert decisions and the quality of the grading system they mimic. Furthermore, these human-made grading systems change over time and are in part differentially applied across clinics, resulting in label bias and data variation. In this work we provide evidence that an AI-based model can consistently outperform expert-based ISUP grading in EOC-Index and AUROC5 when trained on objective patient outcome. Most notably, PCAI outperforms human ISUP PCa grading on average and in every one-to-one comparison on TMA and biopsy external test datasets. To the best of our knowledge, PCAI is the first system that performs systematically better than human experts on external test data that differs significantly from the training data. On the UKE-sealed TMA test dataset, PCAI outperforms ISUP grading and yields comparable results to the GIQ score. In addition, PCAI outperforms ISUP grading on two PCa biopsy cohorts from two clinical centers in Sweden (UPP and MMX), even though it was only trained on TMA samples of RPs from the UKE. These results are quite remarkable as the biopsy datasets vary substantially in size, protocol, and scanner, all of which are completely novel to the PCAI algorithm. On the UPP dataset, PCAI outperforms the ISUP grade in terms of EOC-Index and AUROC5. To further demonstrate the predictive performance of PCAI over ISUP grading, we incorporated image-wise ISUP annotations of three experienced pathologists along with the ISUP score that was collected during clinical routine from different clinics on the MMX dataset and evaluated our risk score against those. PCAI outperforms ISUP grading significantly on the mean prediction performance in EOC-Index and AUROC5 and exceeds ISUP predictions irrespective of the individual pathologist in every individual comparison, reaching an AUROC5 of 0.87. Several large-scale studies from our group have demonstrated that the traditional Gleason grading can be markedly improved by taking the relative distribution of Gleason 4 and Gleason 5 patterns in a tumor into account [27, 37]. The fact that PCAI achieved a comparable performance to the IQ Gleason on 4,095 TMA samples may mean that the performance of our PCAI reaches the best possible use of Gleason pattern evaluation. As digital image analysis is not limited to the gland architecture but can also take into account other potentially prognostic parameters such as nuclear features and stroma composition, further improvements of AI-based prognosis assessment models may be possible.

PCAI achieves state-of-the-art performance on PCa biopsies while it is exclusively trained on TMA spot images after RP. As RP treatment might shift recurring events (outcomes) to later time points, it is conceivable that PCAI's performance might be further increased when trained directly on biopsy datasets. This could be done via fine tuning of (parts of) the model on biopsy data, which could also reduce potential problems with domain shift [38].

5. Conclusion

This work highlighted some salient problems with AI-based grading of histopathological images and offered some solutions to further increase the quality and robustness of existing algorithms. We believe that many technical concepts we introduce in this work, including credibility-guided color adaptation, outcome-based prediction, and the EOC-Index, might improve the robustness and fidelity of AI-based algorithms well beyond our PCa histopathology use case.

Computational hardware and software

The project was implemented using `python 3.10` with `pytorch-lightning 1.8.2`. The PCAI, baseline and cancer indicator models were trained on a NVIDIA Quadro RTX 8000 with 48GB GPU memory and an Intel(R) Xeon(R) Silver 4214 CPU. A distilled version of the code is available at https://github.com/imsb-uke/pcai/.

Data availability

Data from Prostate Cancer Biorepository Network (PCBN), namely the JHU and NYU datasets in this work are available upon request at https://prostatebiorepository.org/. The PANDA dataset with corresponding GT segmentation masks is available on the challenge website at https://panda.grand-challenge.org/data/. The UPP dataset images are publicly available under the name SPROB20 at https://datahub.aida.scilifelab.se/10.23698/aida/sprob20. However, the public version is anonymized and does not provide metadata such as endpoint information for the individual biopsies and patients. The MMX dataset is not publicly available. The UKEhv and UKE-sealed datasets can be obtained upon reasonable request to the department of Pathology at the UKE.

Author contributions

SB and GS initialized this work. SB, MZ, GS, EB, and PW conceptualized the project, algorithm, and computational analyses. SL, ADro, PS, VGP, MH, MB, JB, PB, PW, ADra, CB, MG, EB, GS, MZ, and SB supervised the work. GS, CB, PR, RB, and SvS helped grading some of the cohorts. GS and ML collected and aggregated the UKE cohort. PR, SL, MH, and ADra collected and annotated the UPP cohort. FW, PF, ED, RK, NK, PW, AW, and MZ were responsible for implementing and running the algorithms and computational analyses. SB, FW, PF, MB, and MZ wrote the manuscript. PB, GS, and all other authors critically read and amended the manuscript.

Acknowledgements

We would like to thank the IT of the Institute for Medical Systems Biology and the bAIome Center for Biomedical AI, Sven Heins, Sergio Oller, and Vadim Ustinov. This work is supported by the Department of Defense Prostate Cancer Research Program, DOD Award No W81XWH-18-2-0013, W81XWH-18-2-0015, W81XWH-18-2-0016, W81XWH-18-2-0017, W81XWH-18-2-0018 and W81XWH-18-2-0019 PCRP Prostate Cancer Biorepository Network (PCBN)

JHU 726: We acknowledge the contributions of Dr. Angelo M. De Marzo and Helen L. Fedor from the Brady Urological Research Institute Prostate Specimen Repository at the Johns Hopkins University School of Medicine and Dr. Elizabeth A. Platz from the Department of Epidemiology at the Johns Hopkins Bloomberg School of Public Health for the generation of the 'Case PSA Progression' TMAs for the prostate cancer recurrence nested case-control study, funded in part by the Prostate SPORE Pathology Core (P50 CA58236), Oncology Tissue and Imaging Services at the Sidney Kimmel Comprehensive Cancer Center (P30 CA006973), and a DOD grant (DAMD 17-03-0273).

JHU 235: We acknowledge Drs. Patrick C. Walsh, Alan Partin and Misop Han of The Johns Hopkins University School of Medicine Brady Urological Research Institute for dataset curation for the 235 patient 'Natural History of Prostate Cancer' TMA. PB is supported by the German Research Foundation (DFG, Project IDs 322900939, 432698239 & 445703531), European Research Council (ERC Consolidator Grant No 101001791), and the Federal Ministry of Education and Research (BMBF, STOP-FSGS-01GM2202C).

RK was supported by DFG FOR5068 P9 and the 3R initiative of the UKE, PF and MB by DFG SFB1286 SP02, and SB by EU E-rare MAXOMOD, the M3I excellence initiative of the UKE, and DFG SFB1192 B8 and C3. NK and MZ were supported by DFG SFB1192 B8 and B9, and AW, MZ, and

JB by CDL FLIGHT of the University of Hamburg. ED was supported by DFG KFO306 and FW by DFG KFO296.

5.1. Competing interests

PF, FW, ADro, PS, PW, CB, EB, and SB work part time for Spearpoint Analytics AB, a company developing AI-based digital pathology solutions. The authors declare no other conflicts of interest.

Figures

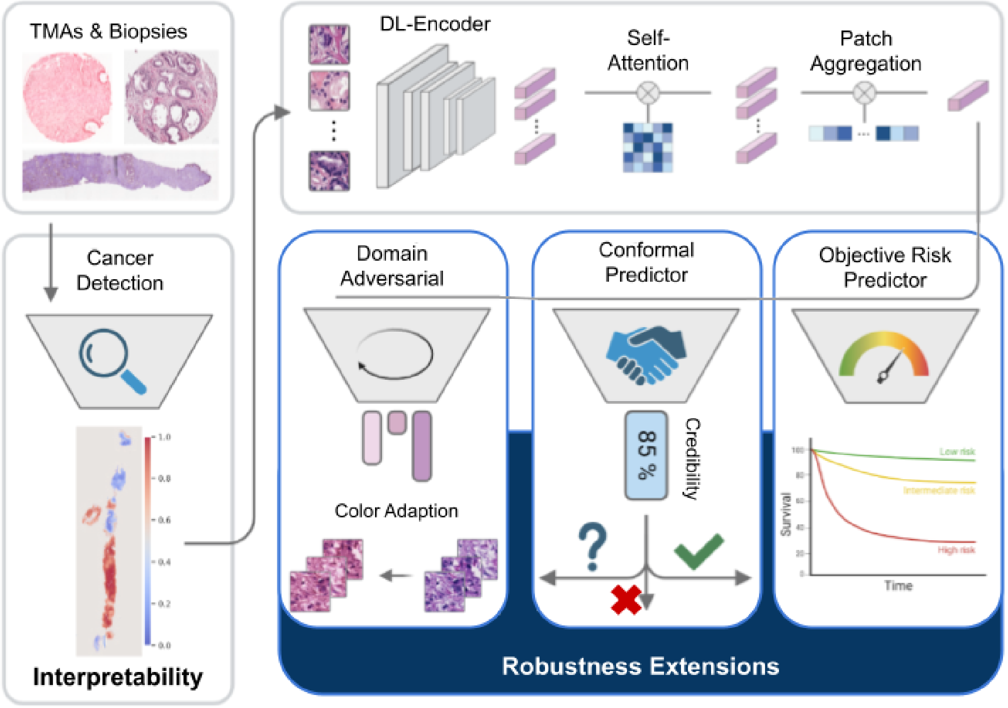
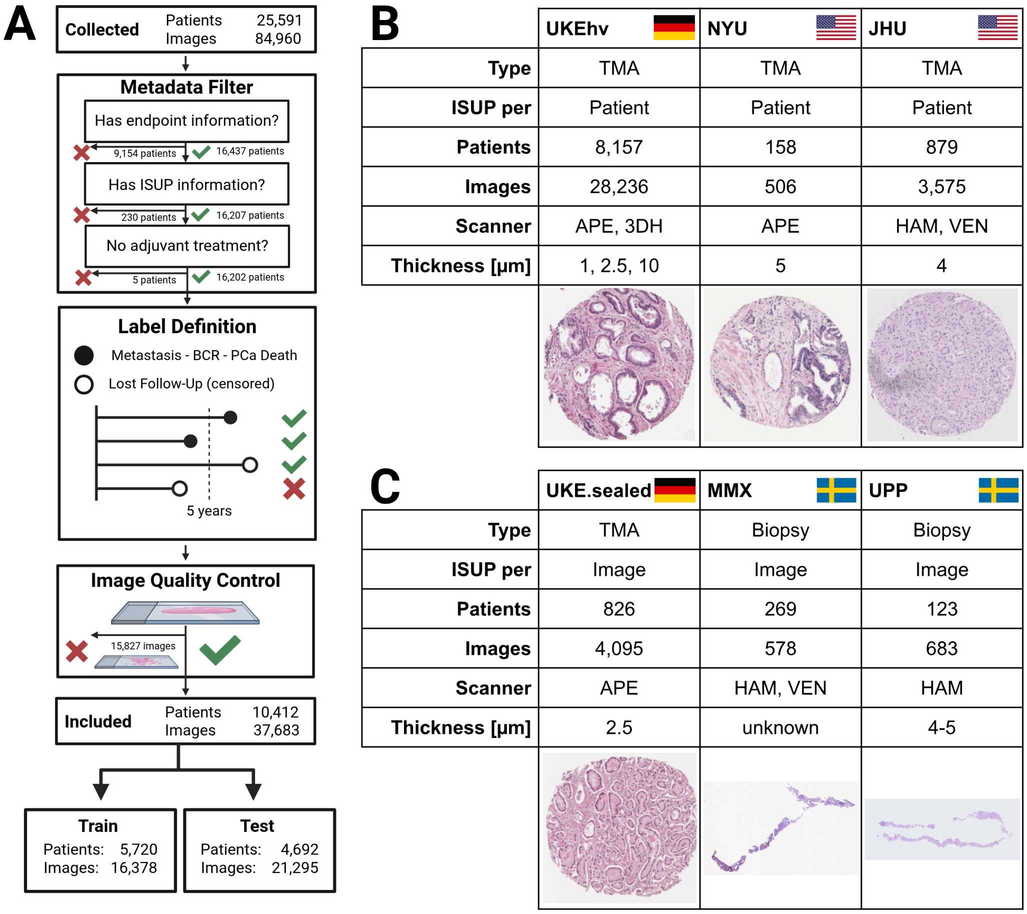


Figure 1: Overview of the PCAI PCa cancer grading algorithm. It can process arbitrarily sized TMA-spot and biopsy images. A separate cancer indication network highlights cancerous regions on the input images, providing visual interpretability. The most relevant patches are then processed in the deep-learning network and used for cancer grading that exceeds stratification based on expert-assigned ISUP compared to objective ground-truth survival information. A credibility estimation setup outputs a credibility score with every risk prediction that adds a quantifiable measure of trustworthiness. Finally, a domain adversarial training regime as well as a credibility-guided color adaptation setup contribute to the model's robustness across five unseen TMA-spot and biopsy datasets.



	UKEhv	NYU	JHU
Type	TMA	TMA	TMA
ISUP per	Patient	Patient	Patient
Patients	8,157	158	879
Images	28,236	506	3,575
Scanner	APE, 3DH	APE	HAM, VEN
Thickness [µm]	1, 2.5, 10	5	4

	UKE.sealed	MMX	UPP
Type	TMA	Biopsy	Biopsy
ISUP per	Image	Image	Image
Patients	826	269	123
Images	4,095	578	683
Scanner	APE	HAM, VEN	HAM
Thickness [µm]	2.5	unknown	4-5

Figure 2: Overview of data preprocessing and the TMA and biopsy datasets. In total, data from 5 different clinics across three countries were collected and integrated. Filtering was performed on the individual patient's metadata (sufficient endpoint information, minimum 5 years of FU duration or any observed event, ISUP information) and image quality (enough tissue on slide, slide not blurry). **A** shows the preprocessing and filtering of all datasets that are divided into training and test sets. **B** depicts the datasets used to build and assess the robustness of the deep learning algorithm. From the UKEhv dataset, which includes highly variant data from six different domains, three are used for training PCAI. The remaining three domains as well as the external JHU and NYU dataset are used to evaluate robust performance on unseen data that expresses a covariate shift. **C** depicts the data used for benchmarking our algorithm against human annotated IQ Gleason and ISUP on a per-image level. Predictive performance is assessed on one unseen TMA-spot (UKE-sealed) and two external biopsy datasets (MMX, UPP). Scanner vendors: APE=Aperio, 3DH=3DHistech, HAM=Hamamatsu, VEN=Ventana.

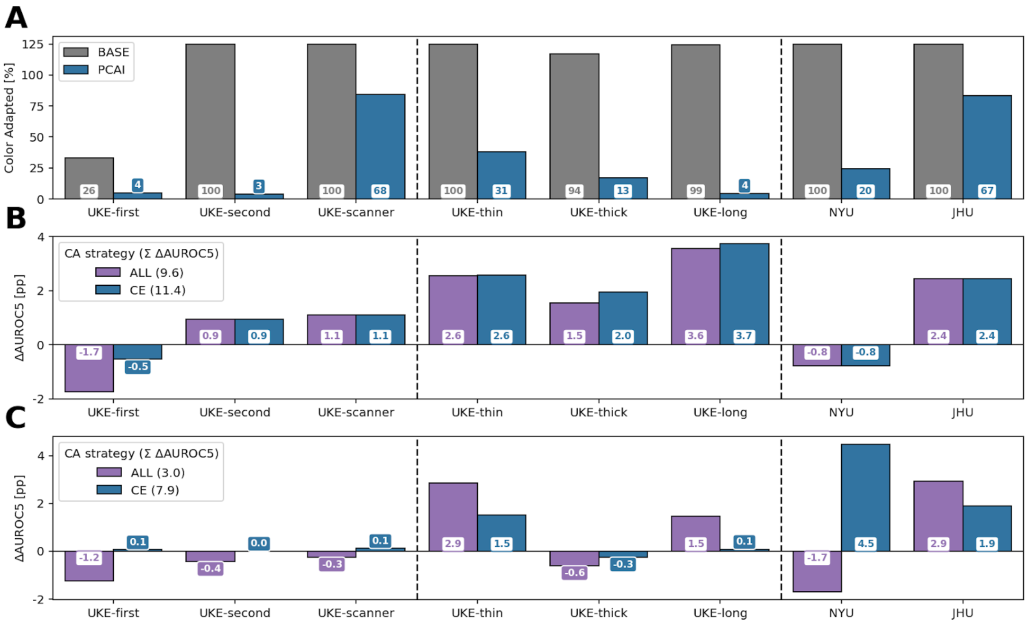


Figure 3: Comparing BASE (gray) and PCAI (blue) with respect to credibility-guided color adaptation on the TMA datasets respective test splits. A shows what percentage of images get color adapted by our credibility-guided approach. The following plots analyze the change in AUROC5 compared to no color adaptation if either all (ALL, purple) or only credibility-guided (CE, blue) samples are color adapted for BASE (**B**) and PCAI (**C**) with the overall sum of gain or loss in AUROC5 shown in the legend. Vertical dashed lines divide domains used for training, the rest of the internal UKEhv domains, and external domains.

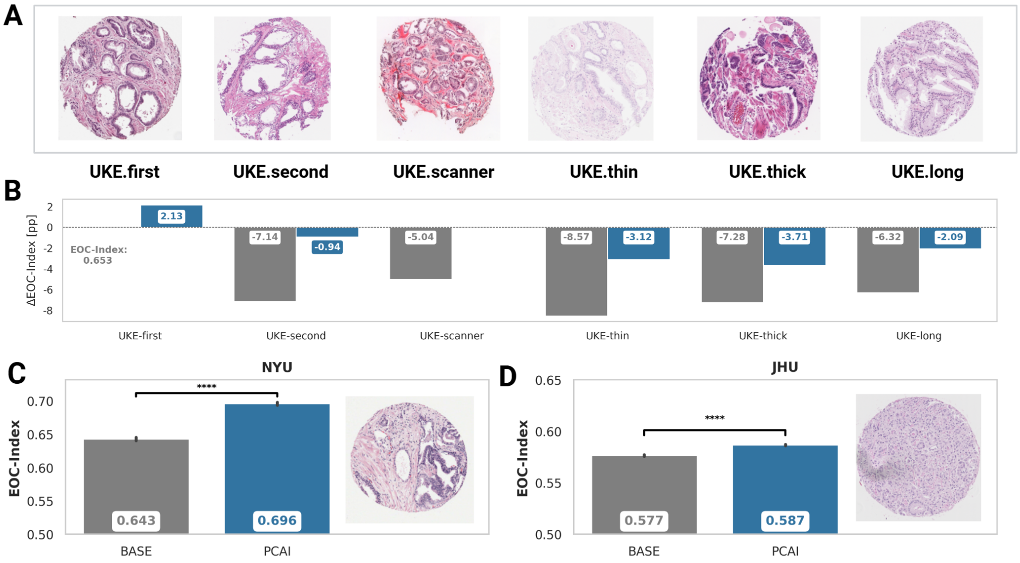


Figure 4: The effect of data variance on model performance. **A** Exemplary images for each of the UKEhv sub-datasets **B** Difference in EOC-Index of PCAI (blue) and BASE (gray) compared to that of BASE on UKE-first (0.692) for the same overlapping patients (n=1537) with one image in each sub-dataset. **C** and **D** show barplots of bootstrapped results with the mean EOC-Index and 95 % confidence intervals of PCAI and BASE on the JHU and NYU datasets along with two exemplary TMAs for the respective datasets. Significance was evaluated using pairwise t-tests that compare PCAI vs all other predictors (*: $p < 0.05$, **: $p < 0.05$, ***: $p < 1e-3$, ****: $p < 1e-4$).

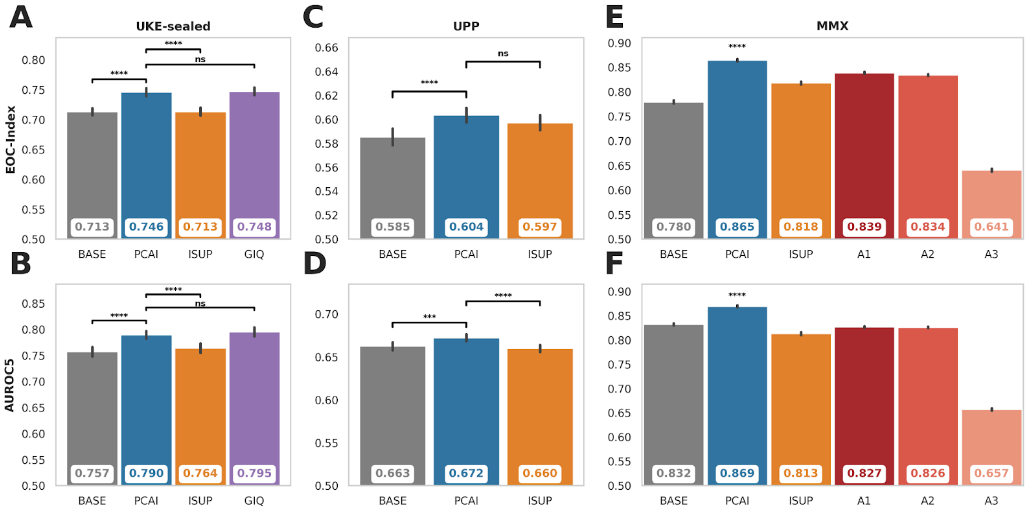


Figure 5: Bootstrapped results of PCa grading performance of human experts and the BASE and PCAI models showing mean and an errorbar of the corresponding 95 % confidence intervals. Performance of BASE (gray) and PCAI (blue) in terms of EOC-Index and AUROC5 on the UKE-sealed TMA spot dataset (**A**, **B**) as well as the UPP (**C**, **D**) and MMX (**E**, **F**) biopsy datasets are compared to human pathologists, image-wise annotations (ISUP from clinical routine in orange; A1-A3 are two experienced and one expert pathologist in red, GIQ by expert pathologist Guido Sauter in purple). To aggregate image-level predictions to patient-level, mean aggregation is performed on UKE-sealed while max aggregation is performed on the biopsy datasets (UPP and MMX). For visual clarity, only the pairwise t-tests that compare PCAI vs all other predictors are shown (*: $p < 0.05$, **: $p < 1e-2$, ***: $p < 1e-3$, ****: $p < 1e-4$). In **E** and **F**, **** denotes that all comparisons to PCAI are statistically significant with $p < 1e-4$.

Tables

		UKE	NYU	JHU	UPP	MMX
	patients	8,157	158	879	123	269
	image type	TMA	TMA	TMA	Biopsy	Biopsy
	age [years] ± SD	63.5 ± 6.1	60.9 ± 7	59.2 ± 6.3	-	67.6 ± 8.9
	censoring [%]	61.4	70.3	0.3	83.7	88.5
	median survival	1.6	3.9	2	2.1	4.3
	median followup	8	17.8	16	7	9.1
ISUP	0	410 (5.03%)	-	-	-	-
	1	1,806 (22.14%)	49 (31.01%)	133 (15.13%)	9 (7.32%)	15 (5.58%)
	2	4,016 (49.23%)	67 (42.41%)	337 (38.34%)	66 (53.66%)	82 (30.48%)
	3	1,367 (16.76%)	16 (10.13%)	184 (20.93%)	27 (21.95%)	80 (29.74%)
	4	109 (1.34%)	11 (6.96%)	123 (13.99%)	12 (9.76%)	36 (13.38%)
	5	449 (5.50%)	15 (9.49%)	102 (11.60%)	9 (7.32%)	56 (20.82%)
Gleason Score	≤ 3+3	2,216 (27.17%)	49 (31.01%)	134 (15.28%)	9 (7.32%)	15 (5.58%)
	3+4	4,016 (49.23%)	67 (42.41%)	334 (38.08%)	66 (53.66%)	82 (30.48%)
	3+5	37 (0.45%)	7 (4.43%)	28 (3.19%)	1 (0.81%)	10 (3.72%)
	4+3	1,367 (16.76%)	16 (10.13%)	185 (21.09%)	27 (21.95%)	80 (29.74%)
	4+4	55 (0.67%)	3 (1.90%)	84 (9.58%)	11 (8.94%)	26 (9.67%)
	4+5	366 (4.49%)	10 (6.33%)	85 (9.69%)	8 (6.50%)	46 (17.10%)
	5+3	17 (0.21%)	1 (0.63%)	10 (1.14%)	-	-
	5+4	82 (1.01%)	5 (3.16%)	15 (1.71%)	1 (0.81%)	8 (2.97%)
	5+5	1 (0.01%)	-	2 (0.23%)	-	2 (0.74%)
Event Type	BCR	3,089 (37.87%)	43 (27.22%)	521 (59.27%)	18 (14.63%)	
	FU	5,007 (61.38%)	111 (70.25%)	3 (0.34%)	103 (83.74%)	226 (84.01%)
	META	61 (0.75%)	-	142 (16.15%)	2 (1.63%)	42 (15.61%)
	PCAD	-	4 (2.53%)	-	-	1 (0.37%)
	TRT	-	-	213 (24.23%)	-	-
T-stage	≤ T1	2 (0.02%)	-	-	95 (78.51%)	122 (45.35%)
	T2	4,966 (60.88%)	104 (65.82%)	134 (15.42%)	26 (21.49%)	90 (33.46%)
	T3	3,128 (38.35%)	52 (32.91%)	735 (84.58%)	-	54 (20.07%)
	T4	61 (0.75%)	2 (1.27%)	-	-	3 (1.12%)
N-stage	N0	4,306 (86.41%)	56 (35.44%)	700 (80.18%)	-	-
	N1	677 (13.59%)	1 (0.63%)	163 (18.67%)	-	-
	N2	-	-	2 (0.23%)	-	-
	NX	-	101 (63.92%)	8 (0.92%)	-	-
M-stage	M0	6,335 (78.47%)	-	509 (60.89%)	7 (5.69%)	79 (29.48%)
	M1	1,738 (21.53%)	-	327 (39.11%)	7 (5.69%)	-
	MX	-	-	-	109 (88.62%)	189 (70.52%)

Table 1: Basic patient characteristics of all experiments showing number of unique patients, number of images and image type, age, PSA level at diagnosis (NYU, JHU), biopsy (MMX) or RP (Other), censoring rate, median survival and FU time in years, the event type classification (BCR=biochemical recurrence, META=metastasis, TRT=any additional treatment, FU=lost to follow-up, PCAD=PCa death), primary and secondary Gleason score, ISUP, pathological (TMA), and clinical (biopsy) T-, N- and M-stage.

Glossary

AUROC5 Area Under the Receiver Operating Characteristic Curve. We evaluate at 5 years of survival time. 5, 12, 13, 16, 18, 19, 21–23, 29, 31

BASE The baseline model is a CNN-based predictive framework trained exclusively on a single internal data domain, UKE-first. 5, 11, 14–16, 18–20, 29–31

BCR Possible endpoint, refers to the rise in PSA levels in the blood following radical treatment or radiation for PCa patients, indicating recurrence. 7, 8, 10, 13, 32

C-Index A metric to measure the concordance of a risk prediction with patient survival. 5, 10, 13, 21

CA Color Adaptation 10–12

CE Credibility Estimation 10–12

CNN Convolutional Neural Network 4

DA Domain Adversarial Training 10–12

EOC-Index To enable a meaningful comparison of different endpoints, the concept of an EOC-Index is introduced. 5, 7, 10, 12–16, 18, 19, 21–24, 30, 31

FU An endpoint for a PCa patient. Lost to follow-up means he did not experience any relapse. 5–8, 10, 11, 13, 28, 32

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. 18, 23, 31

Gleason Most relevant prognostic feature in PCa biopsies and TMAs, based on Gleason grading. Evaluates PCa aggressiveness, by assessing architectural patterns of prostate glands by a (uro-)pathologist. Higher scores indicate more aggressive cancer. 5, 6, 22, 23, 28, 32

ISUP Simplified PCa grading system defining groups 1–5 with increasing cancer severity to predict disease aggressiveness. 3, 5, 6, 8–10, 15, 17–20, 22, 23, 27, 28, 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. 9, 13, 15, 16, 20, 24, 25, 28, 30, 32

META An endpoint for a PCa patient that developed metastases. 10, 13, 32

MMX Biopsy dataset from Malmö, Sweden, exclusively used for model testing. 9, 19, 20, 23, 24, 28, 31, 32

NYU TMA dataset from the Prostate Cancer Biorepository Network, collected at the New York Langone Medical Centre, USA, exclusively used for model testing. 9, 13, 15, 16, 24, 28, 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. 3, 11, 17, 24

PCa Prostate cancer. 3–7, 13, 14, 18–24, 27, 31, 32

PCAD An endpoint for a PCa patient that died from the disease. 10, 13, 32

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. 5, 10–12, 15–24, 27–31

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. 7, 8, 25, 32

RP Surgical procedure performed to treat localized PCa by removing the entire prostate gland along with surrounding tissues. 7, 8, 14, 15, 23, 32

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. 5–9, 11, 14–21, 23, 25, 27–32

TRT An endpoint for a PCa patient indicating disease progression by any additional treatment. 7, 10, 13, 32

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). 8, 9, 11, 14–17, 30

UKE-long Sub-dataset of UKEhv contains TMA spots with nearly ten times the regular staining time when compared to UKE-first. 9, 15, 16

UKE-scanner Sub-dataset of UKEhv scanned with an alternative 3DHistech scanner compared to UKE-first. 9, 11, 14–16

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. 6, 9, 18, 20, 23, 24, 28, 31

UKE-second Sub-dataset of UKEhv representing a secondary batch of cancerous prostate areas with slight variations in processing protocol compared to UKE-first. 9, 11, 15, 16

UKE-thick Sub-dataset of UKEhv contains TMA spots with a thicker sectioning of 10µm compared to UKE-first. 9, 15, 16

UKE-thin Sub-dataset of UKEhv contains TMA spots with a thinner sectioning of 1µm compared to UKE-first. 9, 14, 15

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. 8–12, 14–17, 24, 28–30

UPP Biopsy dataset from Uppsala, Sweden, exclusively used for model testing. 9, 19, 20, 23–25, 28, 31, 32

WSI Digital high resolution scans of entire tissue slides for histopathological analysis. 13, 19

References

[1] M. Brehler, P. Walhagen, C. Busch, S. Bonn, E. Bengtsson, Difficulties and recommendations for ai-based prediction of prostate cancer aggressiveness in digital pathology, Medical Research Archives 11 (2023). URL: https://esmed.org/MRA/mra/article/view/4586. doi:10.18103/MRA.V11I11.4586.

[2] M. Aubreville, N. Stathonikos, C. A. Bertram, R. Klopfleisch, N. ter Hoeve, F. Ciompi, F. Wilm, C. Marzahl, T. A. Donovan, A. Maier, J. Breen, N. Ravikumar, Y. Chung, J. Park, R. Nateghi, F. Pourakpour, R. H. Fick, S. B. Hadj, M. Jahanifar, A. Shephard, J. Dexl, T. Wittenberg, S. Kondo, M. W. Lafarge, V. H. Koelzer, J. Liang, Y. Wang, X. Long, J. Liu, S. Razavi, A. Khademi, S. Yang, X. Wang, R. Erber, A. Klang, K. Lipnik, P. Bolfa, M. J. Dark, G. Wasinger, M. Veta, K. Breininger, Mitosis domain generalization in histopathology images — the midog challenge, Medical Image Analysis 84 (2023) 102699. doi:10.1016/J.MEDIA.2022.102699.

[3] A. Zhang, L. Xing, J. Zou, J. C. Wu, Shifting machine learning for healthcare from development to deployment and from models to data, Nature Biomedical Engineering 6 (2022) 1330–1345. doi:10.1038/s41551-022-00898-y.

[4] K. Stacke, G. Eilertsen, J. Unger, C. Lundström, Measuring domain shift for deep learning in histopathology, IEEE journal of biomedical and health informatics 25 (2020) 325–336. doi:10.1109/JBHI.2020.3032060.

[5] L. Pantanowitz, G. M. Quiroga-Garza, L. Bien, R. Heled, D. Laifenfeld, C. Linhart, J. Sandbank, A. A. Shach, V. Shalev, M. Vecsler, P. Michelow, S. Hazelhurst, R. Dhir, An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study, The Lancet Digital Health 2 (2020) e407–e416. doi:10.1016/S2589-7500(20)30159-X.

[6] K. Sandeman, S. Blom, V. Koponen, A. Manninen, J. Juhila, A. Rannikko, T. Ropponen, T. Mirtti, Ai model for prostate biopsies predicts cancer survival, Diagnostics 12 (2022) 1031. doi:10.3390/diagnostics12051031.

[7] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, P. A. Humphrey, T. Al-Hussain, F. Algaba, M. Aron, D. Berman, D. Berney, F. Brimo, D. Cao, J. Cheville, D. Clouston, M. Colecchia, E. Comperat, I. W. D. Cunha, A. D. Marzo, D. Ertoy, S. Fine, C. Foster, D. Grignon, N. Gupta, R. Gupta, J. Kench, G. Kristiansen, L. Kunju, K. R. M. Leite, M. Loda, A. Lopez-Beltran, T. Lotan, M. S. Lucia, C. Magi-Galluzzi, R. Montironi, J. McKenney, J. Merrimen, G. Netto, R. Orozco, G. Paner, A. Parwani, G. Pizov, V. Reuter, J. Ro, H. Samaratunga, L. Schultz, J. Shanks, I. Sesterhenn, S. Shen, J. Simko, S. Suzigan, M. Suryavanshi, P. H. Tan, H. Takahashi, S. Tomlins, K. Trpkov, P. Troncoso, L. True, T. Tsuzuki, T. V. D. Kwast, M. Varma, A. Warren, T. Wheeler, X. Yang, M. Zhou, P. Kantoff, M. Eisenberger, W. Stadler, G. Andriole, E. Klein, M. Benson, F. Montorsi, D. Crawford, S. Loeb, J. Catto, E. Schaeffer, J. N. Nacey, T. DeWeese, H. Sandler, A. Zietman, A. Pollack, G. Rodrigues, The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system, American Journal of Surgical Pathology 40 (2016) 244–252. doi:10.1097/PAS.0000000000000530.

[8] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. H. van de Kaa, G. Litjens, Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study, The Lancet Oncology 21 (2020) 233–241. doi:10.1016/S1470-2045(19)30739-9.

[9] Y. Li, M. Huang, Y. Zhang, J. Chen, H. Xu, G. Wang, W. Feng, Automated gleason grading and gleason pattern region segmentation based on deep learning for pathological images of prostate cancer, IEEE Access 8 (2020) 117714–117725.

[10] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rueschoff, M. Claassen, Automated gleason grading of prostate cancer tissue microarrays via deep learning, Scientific reports 8 (2018) 12054.

[11] T. H. Nguyen, S. Sridharan, V. Macias, A. Kajdacsy-Balla, J. Melamed, M. N. Do, G. Popescu, Automatic gleason grading of prostate can-

cer using quantitative phase imaging and machine learning, Journal of biomedical optics 22 (2017) 036015–036015.

[12] 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`.

[13] K. Nagpal, D. Foote, Y. Liu, P.-H. C. Chen, E. Wulczyn, F. Tan, N. Olson, J. L. Smith, A. Mohtashamian, J. H. Wren, et al., Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer, NPJ digital medicine 2 (2019) 48. doi:`10.1001/jamaoncol.2020.2485`.

[14] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Medicine 2019 25:8 25 (2019) 1301–1309. URL: `https://www.nature.com/articles/s41591-019-0508-1`. doi:`10.1038/s41591-019-0508-1`.

[15] E. Dietrich, P. Fuhlert, A. Ernst, G. Sauter, M. Lennartz, H. S. Stiehl, M. Zimmermann, S. Bonn, Towards explainable end-to-end prostate

cancer relapse prediction from h&e images combining self-attention multiple instance learning with a recurrent neural network, in: Machine Learning for Health, PMLR, 2021, pp. 38–53.

[16] P. Walhagen, E. Bengtsson, M. Lennartz, G. Sauter, C. Busch, Ai-based prostate analysis system trained without human supervision to predict patient outcome from tissue samples, Journal of Pathology Informatics 13 (2022) 100137. doi:10.1016/j.jpi.2022.100137.

[17] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga, B. Delahunt, C. Lindskog, E. A. Janssen, A. Blilie, L. Egevad, O. Spjuth, M. Eklund, Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction, Nature Communications 2022 13:1 13 (2022) 1–10. URL: https://www.nature.com/articles/s41467-022-34945-8. doi:10.1038/s41467-022-34945-8.

[18] M. Sikaroudi, M. Hosseini, R. Gonzalez, S. Rahnamayan, H. Tizhoosh, Generalization of vision pre-trained models for histopathology, Scientific reports 13 (2023) 6065.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR 2021 - 9th International Conference on Learning Representations (2020). URL: https://arxiv.org/pdf/2010.11929.

[20] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, F. Mahmood, Towards a general-purpose foundation model for computational pathology, Nature medicine 30 (2024) 850. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11403354/. doi:10.1038/S41591-024-02857-3.

[21] M. Karasikov, J. van Doorn, N. Känzig, M. E. Cesur, H. M. Horlings, R. Berke, F. Tang, S. Otálora, Training state-of-the-art pathology foundation models with orders of magnitude less data (2025). URL: https://arxiv.org/pdf/2504.05186.

[22] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, S. Liu, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, et al., Virchow: A million-slide digital pathology foundation model, arXiv preprint arXiv:2309.07778 (2023).

[23] E. D. D. Jong, E. Marcus, J. Teuwen, Current pathology foundation models are unrobust to medical center differences (2025).

[24] J. Kömen, E. D. D. Jong, J. Hense, H. Marienwald, J. Dippel, P. Naumann, E. Marcus, L. Ruff, M. Alber, J. Teuwen, F. Klauschen, K.-R. Müller, Towards robust foundation models for digital pathology (2025).

[25] J. K. Chan, The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology, International Journal of Surgical Pathology 22 (2014) 12–32. URL: https://scholar.google.com/scholar_url?url=https://journals.sagepub.com/doi/pdf/10.1177/1066896913517939&hl=de&sa=T&oi=ucasa&ct=usl&ei=FkOkaNe2O5XVieoPtoHAGA&scisig=AAZF9b_-l8-reiiqiD1lzGzd-KQV. doi:10.1177/1066896913517939/ASSET/76A6DB59-E393-4796-829E-398C992CD091/ASSETS/IMAGES/LARGE/10.1177_1066896913517939-FIG38.JPG.

[26] G. Gandaglia, R. Leni, F. Bray, N. Fleshner, S. J. Freedland, A. Kibel, P. Stattin, H. V. Poppel, C. L. Vecchia, Epidemiology and prevention of prostate cancer, European Urology Oncology 4 (2021) 877–892. doi:10.1016/J.EUO.2021.09.006.

[27] 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.

[28] D. F. Gleason, G. T. Mellinger, Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging,

The Journal of urology 111 (1974) 58–64. doi:10.1016/s0022-5347(17)59889-4.

[29] R. N. Flach, P. P. M. Willemse, B. B. Suelmann, I. A. Deckers, T. N. Jonges, C. van Dooijeweert, P. J. van Diest, R. P. Meijer, Significant inter-and intralaboratory variation in gleason grading of prostate cancer: A nationwide study of 35,258 patients in the netherlands, Cancers 13 (2021) 5378. URL: https://www.mdpi.com/2072-6694/13/21/5378/htmhttps://www.mdpi.com/2072-6694/13/21/5378. doi:10.3390/CANCERS13215378/S1.

[30] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, V. Lempitsky, Domain-adversarial training of neural networks, Journal of machine learning research 17 (2016) 1–35. doi:10.48550/arXiv.1505.07818.

[31] 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.

[32] 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.

[33] 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.

[34] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, J. Chen, A benchmark of batch-effect correction methods for single-cell rna sequencing data, Genome biology 21 (2020) 1–32. doi:10.1186/s13059-019-1850-9.

[35] W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, Biostatistics 8 (2007) 118–127. doi:10.1093/biostatistics/kxj037.

[36] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Augmix: A simple data processing method to improve robustness and uncertainty, 8th International Conference on Learning Representations, ICLR 2020 (2019). URL: https://arxiv.org/abs/1912.02781v2.

[37] 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.

[38] M. Aubreville, C. A. Bertram, T. A. Donovan, C. Marzahl, A. Maier, R. Klopfleisch, A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research, Scientific data 7 (2020) 417. doi:10.1038/s41597-020-00756-z.