Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep learning from routine histology improves risk stratification for biochemical recurrence in prostate cancer

Clément Grisi

clement.grisi@radboudumc.nl

Radboud University Nijmegen Medical Centre

Khrystyna Faryna

Radboud University Nijmegen Medical Centre

Nefise Uysal

Radboud University Nijmegen Medical Centre

Vittorio Agosti

Radboud University Nijmegen Medical Centre

Enrico Munari

University of Brescia

Solène-Florence Kammerer-Jacquet

Centre Hospitalier Universitaire de Rennes

Paulo Guilherme de Oliveira Salles

Instituto Mário Penna

Yuri Tolkach

University Hospital Cologne

Reinhard Büttner

University Hospital Cologne

Sofiya Semko

University Hospital Cologne

Maksym Pikul

University Hospital Cologne

Axel Heidenreich

University Hospital Cologne

Jeroen van der Laak

Radboud University Nijmegen Medical Centre

Geert Litjens

Radboud University Nijmegen Medical Centre

Article

Keywords:

Posted Date: March 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9013424/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. J.v.d.L. reports consulting fees from Philips, ContextVision and AbbVie, and grants from Philips, ContextVision and Sectra, outside the submitted work. G.L. reports grants from the Dutch Cancer Society and the NWO, and grants from Philips Digital Pathology Solutions and personal fees from Novartis, outside the submitted work. All other authors declare no relevant competing interests concerning this paper.

Deep learning from routine histology improves risk stratification for biochemical recurrence in prostate cancer

Clément Grisi*[1], Khrystyna Faryna[1], Nefise Uysal[1], Vittorio Agosti[1,2], Enrico Munari[2,3], Solène-Florence Kammerer-Jacquet[4], Paulo Guilherme de Oliveira Salles[5], Yuri Tolkach[6], Reinhard Büttner[6], Sofiya Semko[7], Maksym Pikul[7], Axel Heidenreich[7], Jeroen van der Laak[1], and Geert Litjens[1]

[1]Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
[2]Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
[3]Surgical Pathology Unit, Verona University Hospital Trust, Verona, Italy
[4]Department of Pathology, CHU de Rennes, Rennes, France
[5]Anatomical Pathology Service, Instituto Mário Penna, Belo Horizonte, Brazil
[6]Institute of Pathology, University Hospital Cologne, Cologne, Germany
[7]Clinic of Urology, University Hospital Cologne, Cologne, Germany

Abstract

Accurate prediction of biochemical recurrence (BCR) after radical prostatectomy is critical for guiding adjuvant treatment and surveillance decisions in prostate cancer. However, existing clinicopathological risk models reduce complex morphology to relatively coarse descriptors, leaving substantial prognostic information embedded in routine histopathology underexplored. We present a deep learning-based biomarker that predicts continuous, patient-specific risk of BCR directly from H&E-stained whole-slide prostatectomy specimens. Trained end-to-end on time-to-event outcomes and evaluated across four independent international cohorts, our model demonstrates robust generalization across institutions and patient populations. When integrated with the CAPRA-S clinical risk score, the deep learning risk score consistently improved discrimination for BCR, increasing concordance indices from 0.725–0.772 to 0.749–0.788 across cohorts. To support clinical interpretability, outcome-grounded analyses revealed subtle histomorphological patterns associated with recurrence risk that

*corresponding author: clement.grisi@radboudumc.nl

are not captured by conventional clinicopathological risk scores. This multicohort study demonstrates that deep learning applied to routine prostate histopathology can deliver reproducible and clinically generalizable biomarkers that augment postoperative risk stratification, with potential to support personalized management of prostate cancer in real-world clinical settings.

1 Introduction

Prostate cancer is the second most commonly diagnosed cancer worldwide and the fifth leading cause of cancer-related death, with more than 1.5 million men affected annually [1]. Many patients undergo radical prostatectomy, either as primary treatment or following a period of active surveillance. The success of surgery is routinely monitored through the concentration of prostate-specific antigen (PSA) in the blood, which is expected to drop to undetectable levels (< 0.1 ng/mL) within 4 to 6 weeks after surgery. However, in approximately 30% of the patients, PSA levels subsequently rise, indicating malignant cell regrowth and portending an increased risk of metastasis and disease-specific mortality. Biochemical recurrence (BCR), historically defined as PSA exceeding 0.2 ng/mL on two or more consecutive occasions after reaching nadir, is therefore an early indicator of treatment failure preceding overt clinical progression. Accurate identification of patients at risk for BCR is central to postoperative decision-making, informing the use of adjuvant radiotherapy, advanced imaging, genomic testing, and the intensity of long-term surveillance.

To support these decisions, several statistical risk models combining clinical and pathological variables have been developed [2, 3]. Among these, CAPRA-S is widely adopted and has demonstrated strong prognostic performance across multiple cohorts [4]. Nevertheless, such models rely on manual histopathological assessment and structured reporting, which limits its reproducibility and constrains the granularity at which morphology is represented. Central to most risk models is Gleason grading, which not only suffers from substantial inter-observer variability [5, 6] but also reduces spatially heterogeneous tumor morphology into coarse categorical descriptors. As a result, within-grade architectural patterns and subtle histological features that may carry independent prognostic information are currently overlooked, leaving a substantial fraction of the information embedded in routine histopathology unexplored.

Deep learning has the potential to unlock the true prognostic value of morphological assessment in cancer by capturing fine-grained, spatially distributed histological patterns that are not routinely assessed or quantified in clinical practice. Early approaches based on handcrafted features [7] or simplistic models [8] did not directly optimize on patient outcomes, limiting their ability to discover clinically meaningful prognostic signals. More recent end-to-end models [9–11] have shown improved performance but were largely restricted to tissue microarrays or biopsy cores, which fail to capture the full spatial and morphological heterogeneity present in prostatectomy specimens. In contrast, whole-slide images (WSIs) provide a more comprehensive view of tumor biology, but their gigapixel size poses unique computational challenges. Weakly supervised learning frameworks, particularly multiple instance learning (MIL), have emerged as effective solutions, enabling outcome-driven training on WSIs without pixel-level annotations [12–16]. Incorporating spatial context into these models has further improved performance in grading and prognostic tasks [17–20].

In parallel, advances in self-supervised learning have ushered in the era of pathology foundation models. Pretrained on large, diverse collections of unlabeled histology images, these encoders learn rich, domain-specific representations that transfer effectively to downstream tasks [21, 22]. While these models promise improved robustness and generalization, systematic evaluations in clinically relevant settings remain scarce.

In this study, we present a deep learning approach that predicts risk of BCR directly from whole prostatectomy histology using a continuous time-to-event endpoint. We evaluate three large pathology foundation models alongside a smaller, prostatectomy-specific encoder within a context-aware MIL framework explicitly trained on postoperative recurrence outcomes. Trained on the largest dataset yet assembled for this task, the best models demonstrate robust generalization across four independent international cohorts. Our results show that image-derived risk scores provide prognostic information complementary to CAPRA-S, demonstrating that computational pathology can deliver robust, reproducible biomarkers that augment postoperative risk stratification in prostate cancer.

2 Results

2.1 Data overview and experimental setup

We used five independent cohorts encompassing both historical and contemporary clinical practice, comprising a total of 2485 patients and 7262 whole-slide images and spanning surgeries from 1992 to 2020 across Europe, North America, and South America. This multicohort design covers a wide range of patient populations, follow-up protocols, and slide acquisition systems, enabling a controlled assessment of how *training data heterogeneity* affects model generalization under real-world conditions. The main characteristics of each dataset are summarized in Table 1. Figure 1 illustrates representative whole-slide image thumbnails from each external test cohort, highlighting differences in tissue appearance and color profiles arising from cohort-specific staining protocols and slide acquisition systems. The distribution of International Society of Urological Pathology (ISUP) grades for each cohort is shown in Figure 8, while Figure 9 shows the distribution of biochemical recurrence and censoring times.

Table 1: **Overview of datasets**. Summary of the main characteristics of each cohort used in this study, including collection period, number of patients, number of associated digital pathology slides, and scanner system used for digitization. `TCGA-PRAD` comprises cases from multiple contributing institutions, predominantly from the United States.

Cohort	Location	Years of surgery	# Patients	# Slides	Scanner
`RUMC`	Nijmegen, The Netherlands	1992–2012	608	1723	3DHistech P1000
`PLCO`	United States	1993–2001	723	1971	Leica
`IMP`	Belo Horizonte, Brazil	2016	421	2091	Motic Easy Scan Infinity 60N
`UHC`	Cologne, Germany	2015–2020	330	1028	Hamamatsu NanoZoomer S360
`TCGA-PRAD`	*Multiple locations*	2000–2013	403	449	*Multiple scanners*

Model development. Model training was based on 508 patients from the `RUMC` cohort, publicly released through the `LEOPARD` (LEarning biOchemical Prostate cAncer Recurrence from histopathology sliDes) challenge. To isolate the effect of *training-set heterogeneity* while keeping the downstream architecture fixed, we defined two model development regimes. The first, `RUMC-only`, used the 508 `RUMC` patients. The second, `RUMC+TCGA`, enriched this cohort with 298 `TCGA-PRAD` patients, introducing substantial technical variation (e.g., staining and scanners) and a distinct clinical mix characteristic of a multi-institution resource, for a total of 806 patients. Both

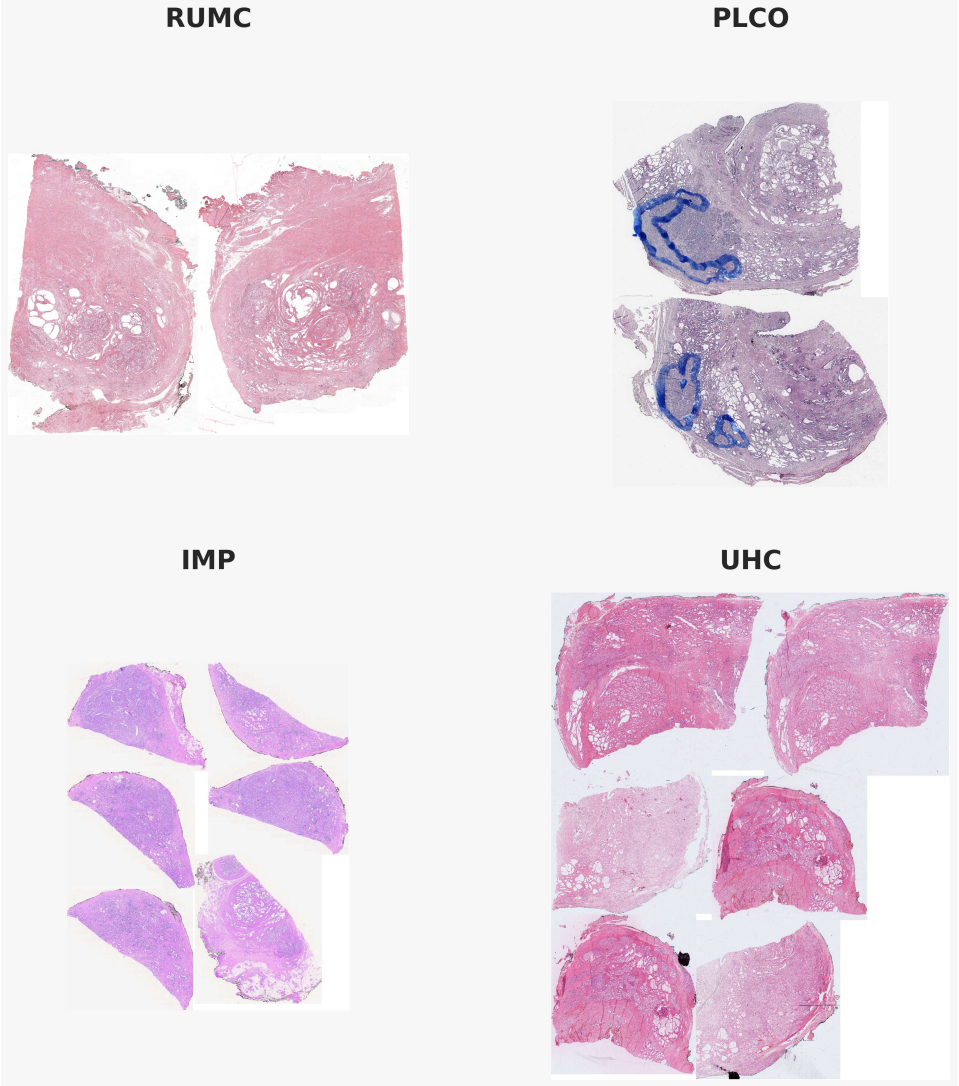


Figure 1: **Representative whole-slide histopathology from external test cohorts.** Low-resolution thumbnails of H&E-stained prostatectomies from the four external test cohorts used in this study. The examples illustrate cohort-specific differences in tissue appearance and color profiles arising from variations in staining protocols, slide preparation, and slide acquisition. All slides were processed using the same preprocessing pipeline prior to model inference.

configurations were split into five cross-validation folds using a stratified procedure to preserve the distribution of recurrence status, Gleason grades, and year of surgery (Section 4.9).

External validation. Model generalization was assessed on four independent cohorts: 100 patients from `RUMC`, 723 from `PLCO`, 421 from `IMP`, and 330 from `UHC`. While the `RUMC` test set is closely related to the training data, the remaining cohorts differ substantially in both technical and clinical characteristics, providing a stringent evaluation of out-of-domain generalization. Importantly, this evaluation spans both historical and contemporary clinical practice. The `RUMC`, `PLCO`, and `IMP` cohorts adopt the historical definition of biochemical recurrence based on two consecutive PSA measurements ≥ 0.2 ng/mL, whereas `UHC` reflects contemporary practice, defining recurrence as a single postoperative PSA measurement ≥ 0.1 ng/mL. This endpoint shift serves as a clinically meaningful test of deployability, assessing whether models trained on historical criteria remain predictive under a contemporary recurrence definition (Section 4.7).

Data preprocessing. For each patient, diagnostic slides were combined into a single patient-level image and tiled into non-overlapping regions at 0.50 microns per pixel (mpp), enabling unified inference across multiple slides while preserving spatially distributed signals present throughout the prostatectomy specimen. This resolution was selected to retain fine-grained morphological detail relevant for risk prediction while remaining compatible with the operating conditions of the encoders evaluated in this study. To avoid masking potentially prognostic cues at the interface between malignant and non-malignant tissue, all available tissue, including benign regions, was retained during tiling, allowing the model to leverage both tumor morphology and contextual non-tumor features potentially associated with recurrence. To further account for scanner variability, a 5% tolerance was applied. When no native spacing fell within this range, regions were extracted at the nearest higher resolution and resampled to approximate 0.50 mpp (Methods, Section 4.9).

2.2 Pretraining diversity drives generalization when models are trained on data from a single source

To assess how pretraining diversity influences generalization when downstream training data originate from a single institution, we evaluated multiple vision encoders within a common weakly supervised learning frame-

work for predicting biochemical recurrence from whole-slide images. All models were implemented using a context-aware multiple instance learning pipeline based on the Hierarchical Image Pyramid Transformer (HIPT) [19], a three-stage architecture that aggregates information at progressively coarser spatial resolutions (Figure 2**C**; Methods, Section 4.3). Within this fixed downstream architecture, we compared four encoders: `Prost40M`, a prostatectomy-specific encoder (Methods, Section 4.1), and three large pathology foundation models – `UNI` [22], `Virchow2` [23], and `H-optimus-0` [24]. Key encoder characteristics are summarized in Table 2.

Model training used discrete-time survival modeling that enables training with a batch size of one, which is well suited for gigapixel whole-slide inputs. The continuous time scale was partitioned into four non-overlapping intervals defined by quartiles of observed recurrence times among uncensored patients, and models were optimized using a censoring-aware discrete-time log-likelihood (Methods, Section 4.2). To isolate the effect of encoder pretraining under homogeneous downstream data, all models were first trained and evaluated using stratified 5-fold cross-validation on the 508 `RUMC` patients (`RUMC-only` splits). Performance was reported using the mean concordance index across folds, and patient-level ensemble risk scores were obtained by averaging predictions from the five fold-specific models for subsequent statistical testing.

Table 2: **Overview of vision encoders**. Summary of the four vision encoders evaluated in this study, including model architecture, number of parameters, pretraining corpus size and source, and pretraining method.

Encoder	Architecture	Parameters	Pretraining corpus	Pretraining method
`Prost40M`	ViT-Small	22M	40M prostatectomy patches ∼2k WSIs	DINO
`UNI`	ViT-Large	307M	100M patches from >100k WSIs	DINOv2
`Virchow2`	ViT-Huge	632M	1.9B patches from 3.1M WSIs	DINOv2
`H-optimus-0`	ViT-giant	1.1B	273M patches from 500k WSIs	DINOv2

Internal cross-validation performance was uniformly high, with mean c-indices ranging from 0.684 (`H-optimus-0`) to 0.727 (`UNI`), indicating that all encoders achieved strong performance when evaluated on data drawn from the same distribution as the training set (Table 3a). Evaluation on held-out test cohorts, however, revealed marked differences in generalization. `Prost40M` performed reasonably well on in-domain `RUMC` (0.646) but showed substantial performance degradation on external cohorts (0.573 on

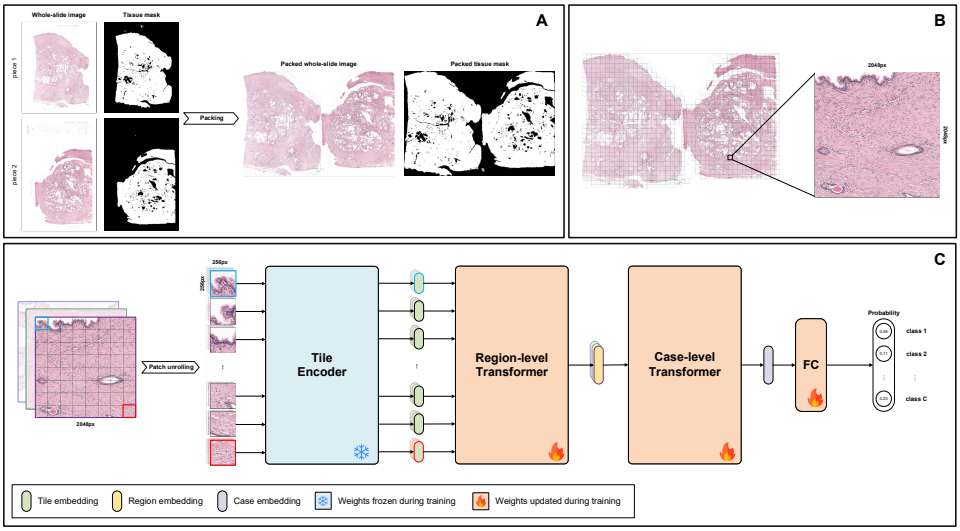


Figure 2: **Integration of multiple histopathology slides into a unified patient-level prediction framework.** **(A)** For patients with multiple slides, segmented tissue regions are cropped using the corresponding tissue masks and stitched into a single larger *packed* slide. This packing process minimizes empty background space while preserving the relative arrangement of tissue, enabling patient-level analysis in a unified coordinate frame. **(B)** The packed slide is tiled into non-overlapping square regions of 2048 pixels at 0.50 mpp. Regions with less than 1% tissue coverage, estimated from the segmentation mask, are excluded. **(C)** Each retained region is then unrolled into non-overlapping 256 pixel tiles, which are processed by a frozen tile encoder to generate tile-level embeddings. For each region, the sequence of tile embeddings is aggregated into a region-level embedding by a first Transformer. The sequence of region-level embeddings for the entire case is then aggregated by a second Transformer into a single patient-level representation, which is projected onto the target classes via a fully connected (FC) layer. The two Transformers and the FC layer are jointly trained for biochemical recurrence risk prediction. Together, these steps enable efficient processing of large whole-slide images, integrating spatially distributed information from multiple slides into a unified patient-level prediction.

PLCO, 0.496 on IMP, 0.626 on UHC). Foundation models exhibited stronger and more consistent generalization to out-of-domain data, with Virchow2 emerging as the most robust encoder across external cohorts (0.701 on PLCO, 0.663 on IMP, and 0.671 on UHC). Pairwise statistical testing confirmed these

trends (Figure 3, Appendix A.1 Table 6). While no significant differences were observed on the internal RUMC test set (all $q > 0.15$), clear performance gaps emerged externally: Virchow2 significantly outperformed all other encoders in both PLCO and IMP, and showed a significant advantage over H-optimus-0 in UHC ($q < 0.01$).

Together, these results indicate that when downstream training data are homogeneous, robustness is strongly influenced by the heterogeneity encountered during pretraining. In this setting, pathology foundation models are preferable to narrow, domain-specific encoders, as their representations are less biased toward a single source domain. Among them, Virchow2 demonstrated the strongest generalization, consistent with its exceptionally large and diverse pretraining corpus, which provides stronger inductive biases than the more limited pretraining of UNI or H-optimus-0.

Table 3: **Cross-validation performance of deep learning models.** Five-fold cross-validation c-index (mean ± std) of models trained on different data splits and evaluated on held-out test cohorts. Best values for each cohort are shown in bold.

(a) Training on RUMC-only splits

Encoder	Tuning	RUMC	PLCO	IMP	UHC
Prost40M	0.712 ± 0.069	0.646 ± 0.062	0.573 ± 0.035	0.496 ± 0.056	0.626 ± 0.043
UNI	**0.727 ± 0.040**	0.676 ± 0.030	0.679 ± 0.009	0.605 ± 0.088	0.629 ± 0.048
Virchow2	0.694 ± 0.066	**0.692 ± 0.040**	**0.701 ± 0.030**	**0.663 ± 0.058**	**0.671 ± 0.040**
H-optimus-0	0.684 ± 0.024	0.674 ± 0.025	0.641 ± 0.045	0.619 ± 0.061	0.625 ± 0.032

(b) Training on RUMC+TCGA splits

Encoder	Tuning	RUMC	PLCO	IMP	UHC
Prost40M	0.723 ± 0.021	**0.717 ± 0.019**	**0.685 ± 0.031**	**0.663 ± 0.057**	0.682 ± 0.023
UNI	0.732 ± 0.025	0.663 ± 0.024	0.647 ± 0.072	0.583 ± 0.086	0.674 ± 0.032
Virchow2	**0.749 ± 0.025**	0.705 ± 0.073	0.655 ± 0.039	0.621 ± 0.070	**0.713 ± 0.024**
H-optimus-0	0.717 ± 0.025	0.647 ± 0.030	0.646 ± 0.037	0.576 ± 0.071	0.663 ± 0.020

2.3 Conditional benefits of training data heterogeneity

Pretraining diversity proved critical for generalization when downstream training data were homogeneous. We next asked whether increasing *training data heterogeneity* yields similar benefits, particularly for encoders with limited pretraining diversity. To this end, we enriched the 508 RUMC cases

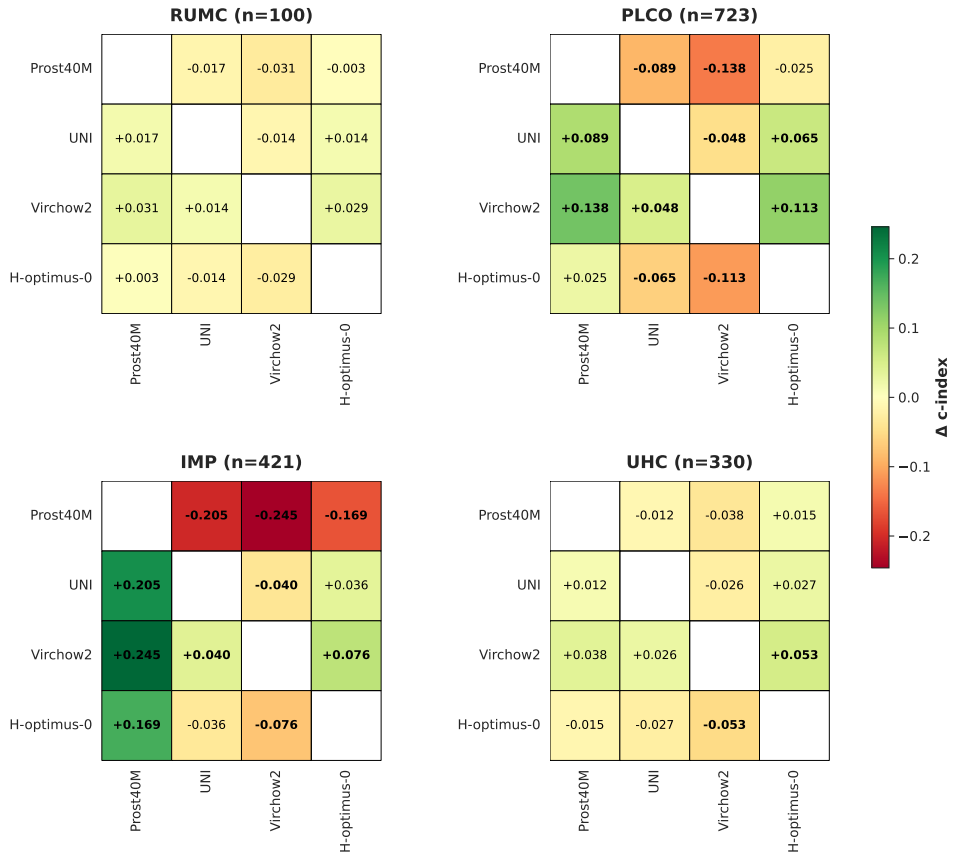


Figure 3: **Pairwise statistical comparisons of encoder performance across test cohorts.** Heatmaps depict differences in concordance index between models trained on the `RUMC-only` splits and evaluated on four independent test cohorts (`RUMC`, `PLCO`, `IMP`, `UHC`). Each cell indicates the difference in performance between the encoder in the row and the encoder in the column. Positive values (green) reflect higher performance for the encoder in the row relative to the encoder in the column, while negative values (red) indicate the opposite. Significant differences are highlighted in bold.

with 298 cases from TCGA-PRAD, introducing substantial technical heterogeneity (staining, scanners) as well as a distinct clinical mix characteristic of a multi-institution cohort. Models were evaluated using stratified 5-fold cross-validation on the enriched RUMC+TCGA splits, and results were confirmed through robust statistical testing (Methods, Section 4.6).

The effect of training data enrichment differed substantially between encoders (Figure 4, Appendix A.2 Table 8). The prostatectomy-specific encoder Prost40M exhibited large and statistically significant performance gains on three of the four external cohorts, suggesting that its narrow pretraining left it vulnerable to exploiting spurious, center-specific correlations when trained on RUMC-only. Adding TCGA-PRAD introduced sufficient technical and clinical variation to steer the model toward more generalizable prognostic features. In contrast, pathology foundation models derived little benefit from enrichment, which we interpret as a robustness effect: their broad and heterogeneous pretraining already yields representations that generalize reasonably well even when downstream training data are comparatively homogeneous. Accordingly, Virchow2 remained strong overall but showed a decline on PLCO ($\Delta = -0.07$, $q < 0.01$), while UNI and H-optimus-0 showed no measurable effect (all $q > 0.1$). Pairwise model comparisons corroborated these findings (Figure 5, Appendix A.1 Table 7), confirming that enrichment significantly improved Prost40M, bringing its performance close to that of Virchow2, which remained the strongest encoder overall.

Interestingly, performance on the contemporary UHC cohort improved for all encoders after enrichment. Both UHC and TCGA-PRAD contain a higher proportion of intermediate-to-high ISUP grades than RUMC, PLCO, and IMP (Figure 8). Because ISUP grade is strongly associated with recurrence risk, training on RUMC-only exposes models to a grade distribution skewed toward lower grades, which can bias learned representations toward that part of the spectrum. Enrichment with TCGA-PRAD increased the representation of higher-grade cases during training, reducing this distribution shift and improving calibration across grades. Consistent with this interpretation, ISUP-stratified analyses on UHC showed that models trained on RUMC-only were skewed toward grade 2, whereas enrichment reduced this bias, improving performance in grades 3 and 5, and yielding a flatter performance profile across grades (Appendix C Figure 10). Approaches that explicitly balance grade representation during training, such as proportion-based sampling or loss weighting, may further mitigate this effect and represent a promising direction for future work.

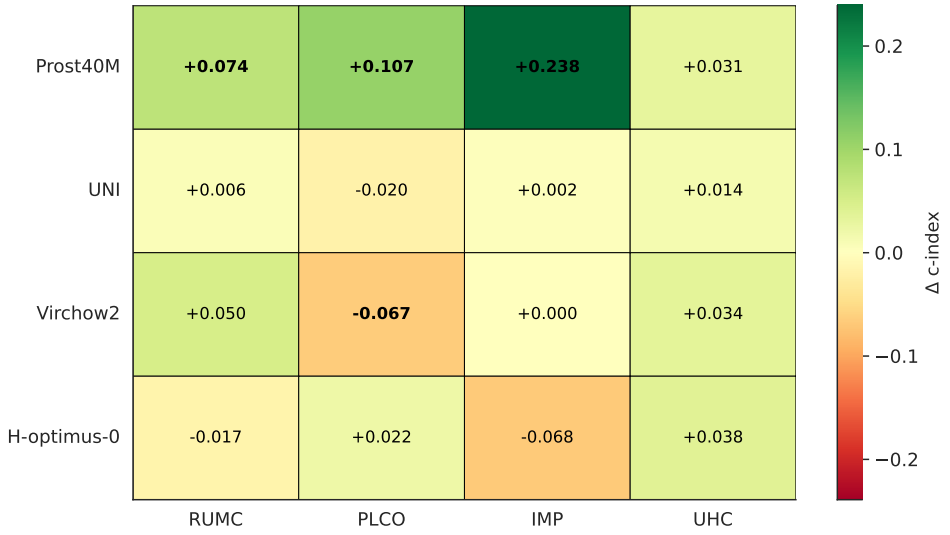


Figure 4: **Effect of training data enrichment on model performance across test cohorts.** The heatmap depicts the change in concordance index (Δ) for each model when trained on the combined `RUMC+TCGA` dataset compared to training on `RUMC-only`. Rows correspond to encoders and columns to test cohorts. Positive values (green) indicate improved performance with training data enrichment, while negative values (red) indicate decreased performance. Significant differences are highlighted in bold.

Overall, increasing training set heterogeneity yielded the largest gains for encoders with limited pretraining diversity, whereas models already pretrained on broad, heterogeneous corpora derived little additional benefit. Together with the results of Section 2.2, these findings indicate that robust generalization depends on an interaction between pretraining diversity and training data composition, and that aligning the clinical spectrum between training and target populations is critical for reliable prognostic modeling.

2.4 Deep learning adds prognostic value beyond CAPRA-S

Having established that deep learning models generalize across heterogeneous cohorts, we next asked whether features extracted by the deep learning model from routine prostatectomy histology provide prognostic value beyond established clinical risk models. Specifically, we evaluated whether a deep learning-derived risk score computed directly from digitized H&E slides captures recurrence-associated signals that are complementary to CAPRA-S, a widely used clinicopathological risk score integrating surgical, pathological, and clinical variables. For reference, CAPRA-S alone achieved concordance indices of 0.725 on `RUMC`, 0.761 on `PLCO`, 0.772 on `IMP`, and 0.747 on `UHC` (Methods, Section 4.8). We then tested whether integrating image-derived predictions with CAPRA-S improves prognostic discrimination across independent cohorts.

For each encoder, we derived a single patient-level deep learning risk score (DLRS) by averaging predictions across the five cross-validation folds. To test whether this image-derived score provides prognostic information beyond CAPRA-S, we fit multivariable Cox proportional hazards models including both CAPRA-S and the DLRS as covariates, and evaluated these models separately in each test cohort. We report: (i) the c-index of the DLRS alone, (ii) the c-index of the multivariable Cox model, (iii) the hazard ratio associated with the DLRS after adjustment for CAPRA-S, and (iv) the corresponding p-value from the Wald test. We focus here on models trained on the `RUMC+TCGA` splits, which showed superior discrimination (Table 4). For completeness, results for `RUMC-only` training are provided in Appendix B (Table 9).

Across cohorts, integrating the DLRS with CAPRA-S consistently improved prognostic discrimination relative to either predictor alone. In most settings, the DLRS remained a statistically significant independent predictor after ad-

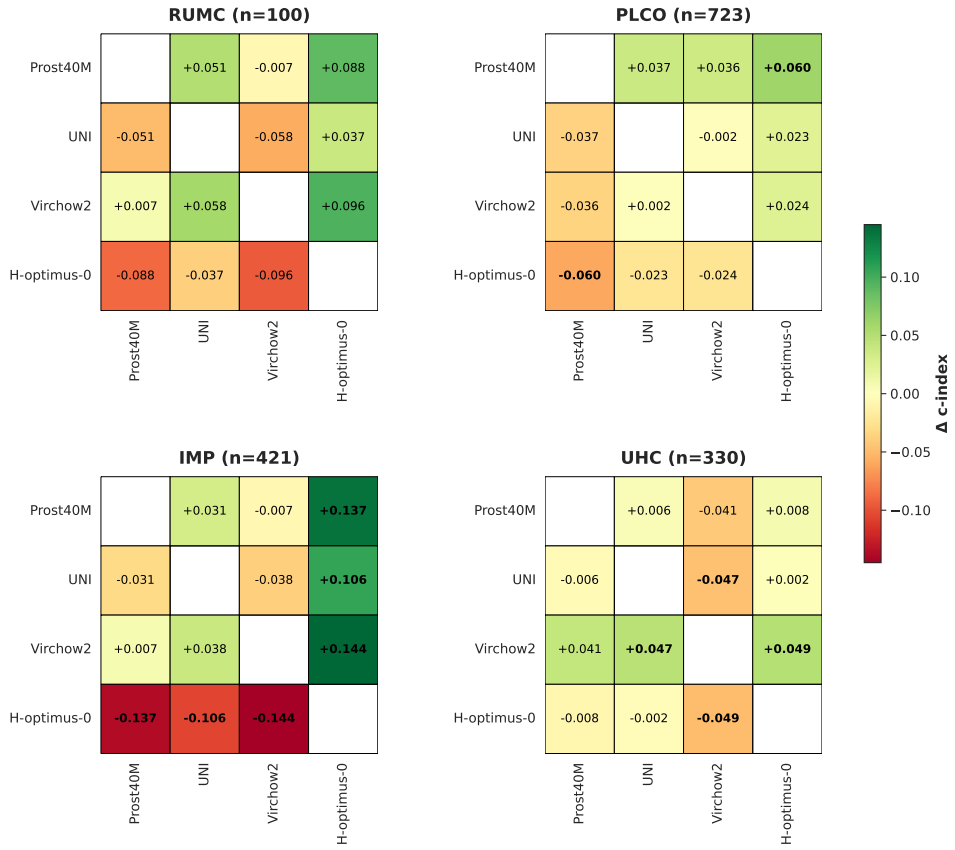


Figure 5: **Pairwise statistical comparisons of encoder performance across test cohorts.** Heatmaps depict differences in concordance index between models trained on the `RUMC+TCGA` splits and evaluated on four independent test cohorts (`RUMC`, `PLCO`, `IMP`, `UHC`). Each cell indicates the difference in performance between the encoder in the row and the encoder in the column. Positive values (green) reflect higher performance for the encoder in the row relative to the encoder in the column, while negative values (red) indicate the opposite. Significant differences are highlighted in bold.

justment for CAPRA-S (hazard ratio > 1, Wald $p < 0.05$), demonstrating that the deep learning model captures recurrence-associated signals not fully accounted for by established clinical risk factors. While average performance varied modestly across encoders, with `Prost40M` achieving the highest mean joint c-index (0.776), followed closely by `Virchow2` (0.769) and `UNI` (0.767), the key finding was consistent across architectures: image-derived risk scores provided complementary prognostic information rather than acting as surrogates for CAPRA-S. Together, these results support the integration of deep learning-based histopathology biomarkers into multivariable risk models to improve postoperative risk stratification in prostate cancer.

2.5 Model interpretability

Having established that our deep learning risk score provides prognostic information complementary to CAPRA-S, a critical question arises: *which histological features actively drive the model's predictions?* Answering this question is essential for understanding why histology-derived risk scores add value beyond established clinicopathological models. Doing so serves two distinct but equally important objectives. First, model interpretability allows clinical validation. By identifying the regions that most strongly influence predictions, we can evaluate whether the model relies on morphological features already known to correlate with prognosis, such as Gleason patterns [25]. Demonstrating alignment with established pathology is crucial for earning clinician trust. Second, interpretability may uncover previously overlooked predictive cues. Because CAPRA-S does not account for within-grade morphology or subtle histopathological patterns, signals extracted directly from histology may explain the complementary prognostic information observed in Section 2.4. Such complementary insights may ultimately support the discovery of new, human-interpretable biomarkers, motivating a systematic investigation into *what* drives the model's predictions.

The term interpretability is widely used in computational pathology, yet it encompasses a range of distinct and sometimes conflicting definitions. In the context of outcome-driven models, we argue that interpretability is most meaningful when it reflects direct contribution to the model's output. Under this definition, an interpretable feature is one whose presence or removal produces a measurable change in the predicted risk of biochemical recurrence, thereby directly linking tissue morphology to the model's prognostic output.

Table 4: **Ensemble and joint modeling performance of models trained on RUMC+TCGA.** Summary of the prognostic performance of deep learning models evaluated on the four held-out test cohorts. For each cohort, the *ensemble* c-index reflects the standalone performance of the deep learning model and is computed by averaging patient-level risk scores from the five RUMC+TCGA cross-validation folds. The *joint* c-index is obtained by fitting a multivariable CoxPH model using both the ensemble risk score and the CAPRA-S score as covariates. For joint models, the hazard ratio (HR) with 95% confidence interval (CI) and the Wald test p-value indicate the independent contribution of the ensemble risk score after adjusting for CAPRA-S. Bold values indicate the highest c-index observed for each cohort.

Cohort	Encoder	Ensemble	Joint	HR (95% CI)	p-value
RUMC	Prost40M	0.760	**0.788**	1.80 (1.27-2.56)	0.001
	UNI	0.709	0.772	1.61 (1.12-2.29)	0.009
	Virchow2	**0.767**	0.774	1.61 (1.14-2.26)	0.006
	H-optimus-0	0.671	0.744	1.44 (1.01-2.06)	0.043
PLCO	Prost40M	**0.711**	**0.779**	1.32 (1.15-1.52)	< 0.001
	UNI	0.674	0.767	1.29 (1.12-1.49)	< 0.001
	Virchow2	0.675	0.765	1.25 (1.10-1.43)	< 0.001
	H-optimus-0	0.651	0.770	1.17 (1.03-1.33)	0.016
IMP	Prost40M	0.700	**0.788**	1.21 (1.05-1.39)	0.008
	UNI	0.669	0.779	1.07 (0.95-1.20)	0.260
	Virchow2	**0.707**	0.779	1.02 (0.90-1.16)	0.754
	H-optimus-0	0.563	0.775	0.89 (0.76-1.04)	0.154
UHC	Prost40M	0.694	0.749	1.31 (1.07-1.62)	0.010
	UNI	0.688	0.753	1.29 (1.08-1.53)	0.004
	Virchow2	**0.735**	**0.760**	1.52 (1.25-1.84)	< 0.001
	H-optimus-0	0.686	0.752	1.23 (1.05-1.45)	0.013

Our risk model is implemented as a Transformer-based multiple instance learning aggregator, which uses attention at multiple spatial scales to aggregate local tile-level representations into a contextualized patient-level prediction. Visualizing attention scores therefore provides a natural starting point for interpretability, as it reveals the regions where the model allocates representational focus in the whole-slide image. However, attention does not establish a causal relationship between these regions and the model's output: it does not indicate whether a given region actively increases the predicted recurrence risk or instead provides evidence for lower risk. To address this limitation, we complement attention-based visualization with an occlusion-based perturbation analysis that directly quantifies the signed contribution of individual tiles to the model's prediction, enabling a direct assessment of which histological patterns increase or decrease recurrence risk.

2.5.1 Attention as a first approximation of model interpretability

Transformer-based architectures derive much of their representational power from the self-attention mechanism, which allows the model to weight different regions of the input according to how informative they are for building internal representations. In the context of whole-slide images, attention maps provide a spatially resolved view of regions the model deems informative during inference. In Hierarchical Vision Transformers (HViTs), attention is computed at multiple spatial resolutions, with different Transformer blocks attending to complementary levels of tissue organization, ranging from local tile-level patterns to global slide-level context. Visualizing the attention map from any single block in isolation therefore provides only a partial view of how information is integrated across scales. To obtain a comprehensive, multi-scale view, we integrate attention scores across the hierarchical architecture by adapting the attention factorization framework of Grisi et al. [20], aggregating attention multiplicatively across Transformer layers (Methods, Section 4.4). The resulting factorized attention maps highlight regions that receive consistent focus across scales.

To illustrate how attention can support a qualitative understanding of regions the model considers informative, we visualize factorized attention maps for two representative patients in Figure 6: one with early biochemical recurrence and a correctly predicted high-risk score (Figure 6**A**), and one with long recurrence-free survival and a correctly predicted low-risk score (Figure 6**B**). For clarity, attention maps are thresholded to display only the most salient regions, with magnified views highlighting prominent hotspots.

These visualizations reveal spatial patterns consistently emphasized by the model and suggest histomorphological features potentially relevant for prognosis. However, attention alone does not indicate whether these regions increase or decrease the predicted recurrence risk. To move beyond localization and assess causal contribution, the next section turns to an outcome-grounded analysis.

2.5.2 Outcome-grounded interpretability via occlusion

To operationalize our contribution-based definition of interpretability, we quantify how individual tissue regions influence predicted recurrence risk using an occlusion-based perturbation analysis (Methods, Section 4.5). For each patient, tiles are removed one at a time from the model's input and the resulting change in predicted risk is measured. This procedure yields a signed contribution score for each tile, directly indicating whether it drives the prediction upward or downward and providing a causal attribution within the model's decision process.

To identify histological patterns that consistently affect model predictions, we focused on the most influential tiles from the held-out `RUMC` test set. For each patient, we selected the ten most risk-increasing and ten most risk-lowering tiles and examined whether shared morphology emerged across patients. To this end, we projected model-refined tile representations using t-SNE [26]. We observed a clear separation between risk-increasing and risk-lowering tiles, with each group forming distinct clusters in the embedding space (Figure 7). Notably, the magnitude of positive contributions was generally stronger than that of negative contributions, consistent with the clinical observation that benign morphology is shared across risk groups, while adverse features are primarily responsible for distinguishing high-risk patients.

This analysis was deliberately restricted to the in-domain `RUMC` test set to ensure that clustering reflects biologically meaningful morphology rather than non-biological confounders such as staining variation or scanner-specific effects, which pathology foundation models are known to encode strongly [27]. By operating on the risk model's internal representations instead of frozen foundation model embeddings, we aimed to further mitigate center-driven batch effects and focus on morphology relevant to recurrence prediction.

An expert pathologist (E.M.) reviewed representative clusters of risk-increasing

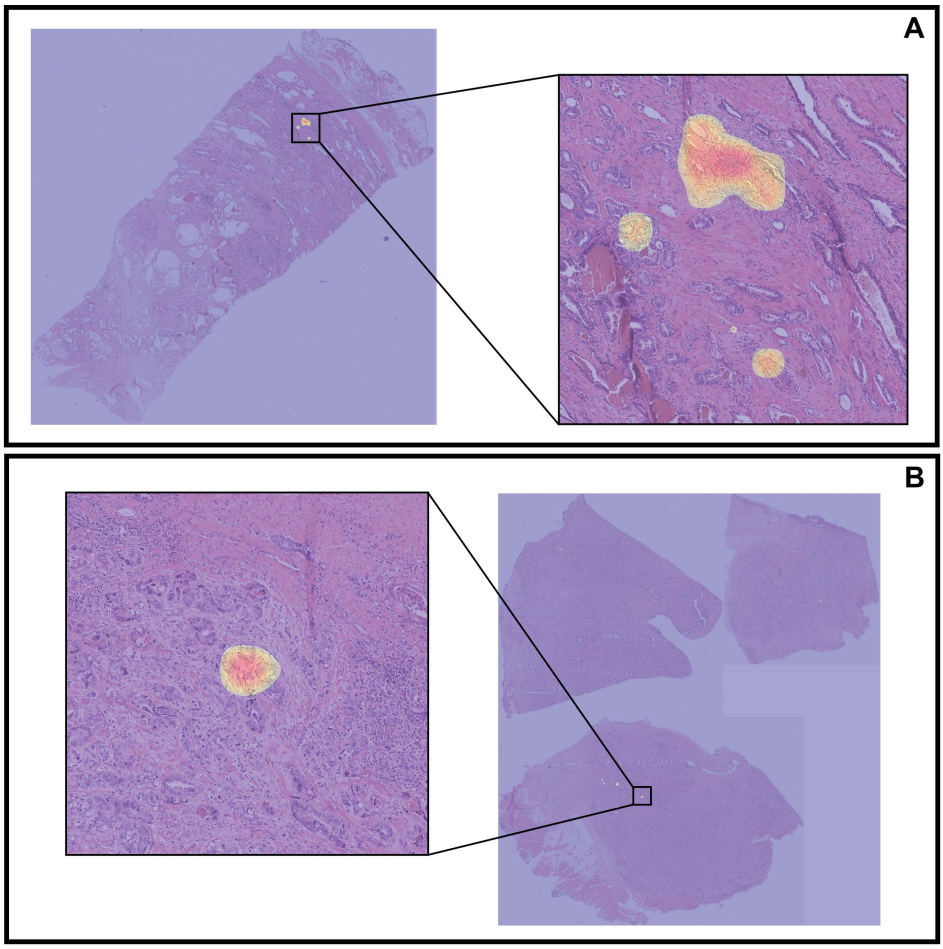


Figure 6: **Factorized attention maps reveal regions the model considers informative in prostatectomy slides.** Factorized attention heatmaps derived from the `Virchow2`-based model and overlaid on H&E-stained slides. Each panel shows the whole-slide image with regions of high model attention, alongside a magnified view of the most salient region. **(A)** True positive case: a patient who experienced biochemical recurrence within a short time and was assigned a high risk score by the model. **(B)** True negative case: a patient who remained recurrence-free with long follow-up and was assigned a low risk score by the model.

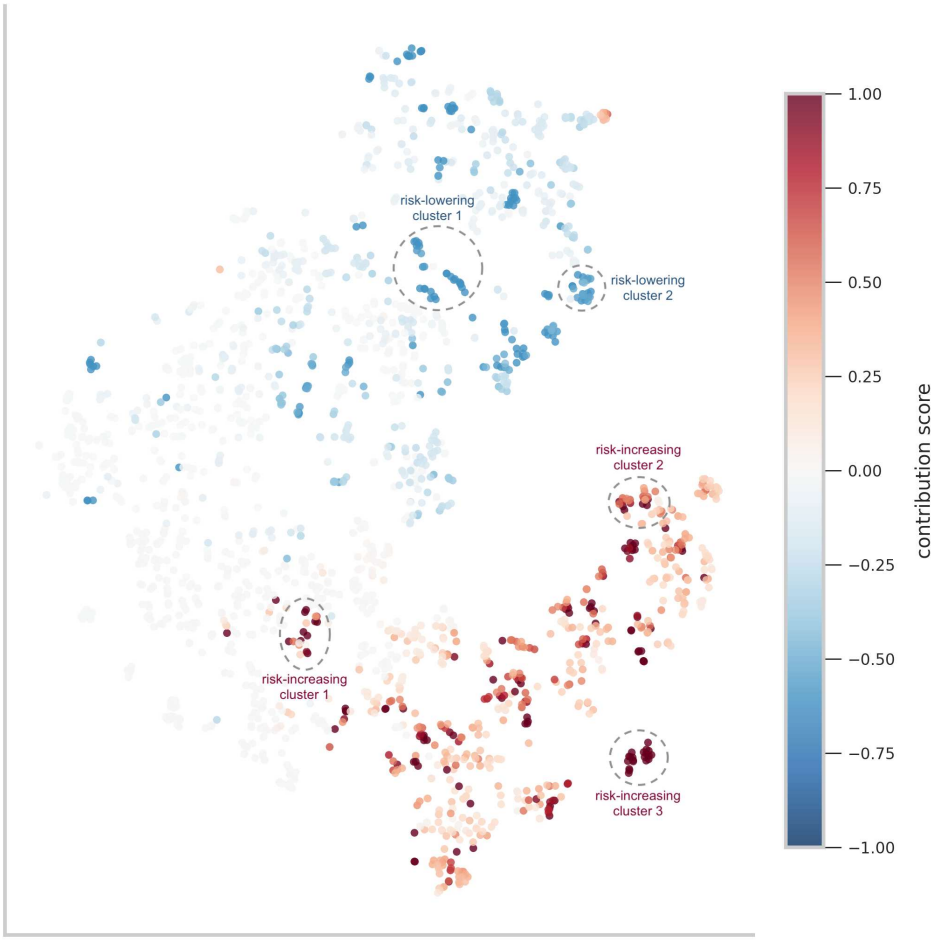


Figure 7: **t-SNE plot of model-refined tile representations reveals recurrence-associated morphology.** Across the `RUMC` test cohort, we selected tiles with the strongest positive or negative contributions to the recurrence risk estimated by the `Virchow2`-based model. Their model-refined embeddings are projected into two dimensions using t-SNE. Color encodes each tile's signed contribution to the final prediction, with red indicating tiles associated with higher recurrence risk and blue indicating tiles associated with lower risk, highlighting gradients of prognostic relevance within clusters. Grey outlines denote the representative clusters subsequently reviewed by an expert pathologist.

and risk-lowering tiles to characterize the underlying morphology (Appendix D). Risk-increasing clusters frequently captured adverse histologies, including large cribriform glands (Figure 11), fused Gleason pattern 4 glands, and single-cell pattern 5 carcinoma (Figure 12), whereas risk-lowering clusters were enriched for well-formed glands or benign tissue (Figure 14). This alignment with established prognostic signals supports the biological and clinical validity of the model.

Beyond tumor-centric features, one risk-increasing cluster contained little or no carcinoma and was dominated by fibromuscular stroma or visually heterogeneous non-tumor tissue (Figure 13). Its association with increased recurrence risk suggests that the model captures prognostic cues beyond tumor morphology – signals that current risk models do not account for. Similarly, the presence of risk-lowering clusters containing specific Gleason pattern 4 morphologies (Figure 15) highlights within-grade heterogeneity that is not distinguished by CAPRA-S. Together, these findings provide biological grounding for the independent risk signal identified in Section 2.4. Comprehensive investigation of these signals is an important direction for future work but lies beyond the scope of this study.

This outcome-grounded analysis provides a general blueprint for causal interpretation of pathology prognostic models. Together, our findings reveal that the model integrates both well-established adverse tumor morphology and additional prognostic features not routinely assessed in clinical practice. This duality underscores its alignment with expert pathology while highlighting deep learning's potential to uncover novel biomarkers and broaden our understanding of the mechanisms underlying biochemical recurrence.

3 Discussion

In this study, we developed and validated a deep learning-based biomarker for predicting biochemical recurrence directly from routine H&E-stained prostatectomy specimens, trained end-to-end on time-to-event outcomes and evaluated across multiple independent cohorts. Beyond demonstrating strong prognostic performance, our results provide broader insights into how computational pathology models can be trained, evaluated, and interpreted to support clinically meaningful risk stratification.

First, our results indicate that robust generalization in histology-based prog-

nostic modeling depends on an interaction between pretraining diversity and downstream training data composition, rather than on either factor in isolation. When downstream training data were drawn from a single institution, models pretrained on large, heterogeneous corpora generalized substantially better than narrowly pretrained, organ-specific encoders. Conversely, enriching downstream training with multi-institutional cases primarily benefited encoders with limited pretraining diversity, while providing little additional gain for foundation models. Beyond variability in data acquisition protocols, our results show that differences in the underlying patient mix also shape generalization. Performance on the `UHC` cohort improved for all encoders after adding `TCGA-PRAD` to the training set, consistent with these cohorts sharing a similar clinical spectrum distinct from `RUMC`, `PLCO`, and `IMP`. Together, these results point to a general principle: pretraining diversity and downstream data composition play complementary roles in shaping model generalization, with their relative importance depending on where heterogeneity is introduced in the pipeline. While broad pretraining can mitigate limited downstream data diversity, aligning the clinical spectrum between training and deployment settings remains critical for all models. This insight is particularly relevant as pathology foundation models are increasingly applied across institutions with heterogeneous patient populations, scanners, and clinical practices.

Second, our study provides a mechanistic explanation for why deep learning-derived histology scores add prognostic value beyond established clinical risk models such as CAPRA-S. Through outcome-grounded interpretability analyses, we show that the model captures both well-established adverse tumor morphologies and additional recurrence-associated features, including within-grade architectural heterogeneity and non-tumor or stromal patterns. These features may encode biologically relevant interactions that are not explicitly captured in current clinicopathological risk models relying on coarse categorical grades, providing a biological rationale for the independent prognostic signal observed when combining the deep learning risk score with CAPRA-S. More broadly, our findings illustrate how deep learning applied to histopathology can reveal prognostic signals that may not be readily accessible through routine human assessment. By learning from large numbers of cases and operating across multiple spatial scales, the model can integrate subtle and spatially distributed cues that are difficult to conceptualize or quantify systematically in clinical practice.

Third, our results have practical implications for deploying pathology AI

systems in real-world clinical settings. We demonstrate consistent performance across cohorts spanning diverse institutions, scanners, patient populations, and definitions of biochemical recurrence. Generic postoperative risk models are inherently limited: patients with identical CAPRA-S scores can experience markedly different disease trajectories. Moving toward precision oncology therefore requires incorporating additional patient-specific information into risk stratification. Routine histopathology encodes rich, spatially resolved information about tumor architecture, grade heterogeneity, and tumor-stroma context that is systematically available for nearly all patients but remains largely unquantified in current risk models. Deep learning offers a principled framework to extract and integrate these signals into patient-level predictions at scale. While future progress will likely involve multimodal models integrating imaging, genomics, and clinical data, building robust and well-validated unimodal histology models is a critical and necessary step toward that goal.

This work also has limitations that point to important directions for future research. Although pathology foundation models improved robustness, their performance was not uniformly stable across all settings, indicating that large-scale pretraining alone does not eliminate susceptibility to cohort-specific biases. More systematic strategies for balancing clinically relevant variables during training may further improve generalization. In addition, while occlusion-based interpretability provides a more faithful assessment of tile-level contributions than attention alone, rigorous validation of model-derived patterns – such as blinded, multi-expert review across cohorts – will be essential to translate these insights into reproducible pathological concepts. Finally, prospective studies will be required to assess how histology-based risk scores influence clinical decision-making and patient outcomes when integrated into real-world workflows.

In summary, our findings highlight a broader role for histology-based AI in precision oncology. Deep learning models trained directly on routine prostatectomy slides can uncover prognostic signals that complement established clinical risk scores and capture aspects of tumor biology not currently used in postoperative care. By enabling systematic, fine-grained characterization of tumor morphology, computational pathology offers a path toward more individualized risk stratification and treatment planning in prostate cancer.

4 Methods

4.1 Introducing a small-scale, prostatectomy-specific encoder

While pathology foundation models are pretrained on extremely large and diverse datasets, their generic features, though effective across many organs and diseases, may lack sensitivity to the subtle morphological patterns of a specific organ. To evaluate the potential benefits of organ-specific pretraining, we developed `Prost40M`, a ViT-Small pretrained with DINO [28].

In DINO, a student network g_{θ_s} is trained to match the output of a teacher network g_{θ_t} under different augmented crops of the same image, encouraging consistency across local and global contexts. Given a $(256, 256)$ pixel image $\mathbf{x}$, a set V of different views is generated: 8 local crops of $(96, 96)$ pixels and 2 global crops of $(224, 224)$ pixels. All views are passed through the student, while only the global views are passed through the teacher. Given a pair of input views, both networks output probability distributions over K dimensions, denoted p_s and p_t, respectively. These distributions are matched by minimizing the cross-entropy loss with respect to the student parameters θ_s:

$$\min_{\theta_s} \sum_{\mathbf{v} \in \left\{\mathbf{v}_1^g, \mathbf{v}_2^g\right\}} \sum_{\substack{\mathbf{v}' \in V \\ \mathbf{v}' \neq \mathbf{v}}} H\left(p_t\left(\mathbf{v}\right), p_s\left(\mathbf{v}'\right)\right)$$

where $H(a, b) = -a \log b$.

`Prost40M` was trained on approximately 40 million tiles extracted from 1888 H&E-stained prostatectomy slides – 449 sourced from the 403 patients in the `TCGA-PRAD` cohort and 1439 sourced from the 508 patients in the `RUMC` development set. Using the `slide2vec` library, we extracted non-overlapping $(256, 256)$ pixel tiles at 0.50 mpp, discarding any with less than 25% tissue coverage. We pretrained `Prost40M` for a single epoch, as additional pretraining epochs yield only marginal gains due to the high visual redundancy between histology tiles.

4.2 Discrete modeling of biochemical recurrence

Biochemical recurrence prediction is an ordinal regression task that models time-to-event distributions, where the event (recurrence) is not always observed and may be right-censored. Patients with observed BCR were treated as events, and patients without observed BCR were right-censored at their

last contact.

Let T be a continuous random variable denoting the time to biochemical recurrence. The recurrence-free survival function $f_{\text{bcr}}(T > t \mid X)$ gives the probability that a patient with covariates X remains recurrence-free beyond time t, and the hazard function $f_{\text{hazard}}(T = t \mid T \geq t, X)$ denotes the probability of BCR at time t, defined as:

$$f_{\text{hazard}}(T = t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Traditionally, a Cox Proportional Hazards (CoxPH) model is used to estimate the hazard function, and in deep learning this is implemented by optimizing the Cox partial log-likelihood. However, this loss depends on pairwise comparisons between patients within a batch, making it unsuitable for whole-slide image training where memory constraints and variable number of tiles make training with large batch sizes impractical. To address this, we adopt the discrete-time survival modeling from Chen et al. [29], which allows training with a batch size of one. Specifically, we partition the continuous time scale into four non-overlapping bins: $[t_0, t_1)$, $[t_1, t_2)$, $[t_2, t_3)$, and $[t_3, t_4)$, defined by the quartiles of observed recurrence times among uncensored patients. Each patient is then assigned a discrete label $Y_i \in \{0, 1, 2, 3\}$ indicating the interval in which recurrence was observed, or in which censoring occurred if no event was recorded. Given a patient-level feature vector h_i, the model predicts the discrete hazard for each interval. The hazard function for bin b is defined as:

$$f_{\text{hazard}}(b \mid h_i) = P(T = b \mid T \geq b, h_i)$$

which relates to the recurrence-free survival function:

$$f_{\text{bcr}}(b \mid h_i) = \prod_{u=1}^{b} (1 - f_{\text{hazard}}(u \mid h_i))$$

We train the model using the log-likelihood function for a discrete survival model [30], accounting for censorship:

$$\begin{aligned}\mathcal{L} = & -c_i \cdot \log\left(f_{\text{bcr}}(Y_i \mid h_i)\right) \\ & -(1-c_i) \cdot \log\left(f_{\text{bcr}}(Y_i - 1 \mid h_i)\right) \\ & -(1-c_i) \cdot \log\left(f_{\text{hazard}}(Y_i \mid h_i)\right)\end{aligned}$$

where $c_i = 1$ for censored patients and 0 otherwise. Following prior work [29], we up-weight the contribution of uncensored patients during training, leading to the final objective:

$$\mathcal{L}_{\text{bcr}} = (1-\alpha) \cdot \mathcal{L} + \alpha \cdot \mathcal{L}_{\text{uncensored}} \tag{1}$$

where $\mathcal{L}_{\text{uncensored}}$ corresponds to the loss term for uncensored patients:

$$\begin{aligned}\mathcal{L}_{\text{uncensored}} = & -(1-c_i) \cdot \log\left(f_{\text{bcr}}(Y_i - 1 \mid h_i)\right) \\ & -(1-c_i) \cdot \log\left(f_{\text{hazard}}(Y_i \mid h_i)\right)\end{aligned}$$

This formulation enables stable training on large-scale WSIs using a batch size of one, without relying on inter-sample risk comparisons.

4.3 Weakly supervised prediction of BCR with Hierarchical Vision Transformers

Whole-slide images exhibit a natural hierarchical structure, with information distributed across multiple scales. At the finest level, cell-centric regions reveal nuclear morphology and subcellular detail. At intermediate scales, patterns of cell-to-cell interaction become apparent. At progressively larger scales, broader tissue architecture emerges, capturing macro-level organization and interactions between cellular clusters. Drawing inspiration from this layered structure, Chen et al. [19] introduced the Hierarchical Image Pyramid Transformer (HIPT), a three-stage model that aggregates information at progressively coarser spatial resolutions. This multi-scale modeling is essential for capturing the complexity of intra-tumoral heterogeneity observed across entire slides.

To predict biochemical recurrence from whole-slide images, we adopt a

weakly supervised multiple instance learning framework largely based on the HIPT. Its multiscale approach is particularly suited to biochemical recurrence prediction: fine-grained nuclear morphology may reveal subtle indicators of tumor aggressiveness, whereas larger-scale tissue architecture reflects growth patterns and tumor-stroma interactions. By jointly modeling these complementary scales, HIPT learns rich, context-aware slide representations and has demonstrated state-of-the-art performance in tasks such as cancer subtyping and prostate cancer grading [19, 20].

Because of GPU memory constraints, the model cannot be trained end-to-end. We therefore decouple feature extraction from feature aggregation. In the feature extraction stage, each slide is divided into non-overlapping 2048 pixel regions at 0.50 mpp, which are further unrolled into $(256, 256)$ tiles. Each tile is passed through a pretrained encoder, producing region-level features of shape $(64, d)$, where d is the encoder's embedding dimension. Stacking all region embeddings from a case yields a patient-level feature h_i of shape $(M_{i,2048}, 64, d)$, where $M_{i,2048}$ denotes the number of 2048 pixel regions extracted for patient i. In the feature aggregation stage, a region-level Transformer compresses each $(64, d)$ region feature into a 192-dimensional embedding. The resulting sequence of shape $(M_{i,2048}, 192)$ is processed by a slide-level Transformer, which pools across regions to form a single 192-dimensional whole-slide embedding. A fully connected layer finally projects this embedding to discrete recurrence labels. All aggregator components are jointly trained using the discrete-time survival loss defined in Equation 1 (Section 4.2). We evaluated three recent pathology foundation models as tile encoders, chosen for their strong performance in the HEST-Benchmark [31]:

- `UNI` [22] is a ViT-L/16 model with 307 million parameters trained on over 100 thousand WSIs using DINOv2.
- `Virchow2` [23] is a ViT-H/14 model with 632 million parameters trained on 3.1 million WSIs using a modified DINOv2 framework tailored for histopathology.
- `H-optimus-0` [24] is a ViT-g/14 model with 1.1 billion parameters trained on a proprietary collection of more than 500 thousand WSIs using DINOv2.

In all experiments, aggregator components are optimized for up to 100 epochs using Adam [32] with a base learning rate of 2×10^{-4} (halved every 20 epochs) and weight decay of 10^{-5}. We apply early stopping after 20

consecutive epochs without improvement on the tuning loss, with a minimum of 50 epochs completed. We use a batch size of 1 with 32 gradient accumulation steps.

4.4 Factorized attention maps

To factorize attention scores across the entire hierarchical architecture, we build on the work of Grisi et al. [20], replacing summation with multiplication across transformers. Let N be the total number of Transformers involved in the pretraining or the training processes. Let n be the number of frozen Transformers. For the pixel with (x, y) coordinates in the slide, we denote by $a^j_{(x,y)}$ the attention score of the last self-attention layer of the j-th Transformer T_j. We compute the factorized attention score for that pixel, $a_{(x,y)}$, as:

$$a_{(x,y)} = \prod_{j=0}^{N-1} a^j_{(x,y)} \left[\gamma(1 - \mathbb{1}_{\mathrm{F}}(\mathrm{T}_j)) + (1-\gamma)\mathbb{1}_{\mathrm{F}}(\mathrm{T}_j)\right] \tag{2}$$

with $\mathbb{1}_{\mathrm{F}}(\mathrm{T}_j)$ the indicator function defined on the set of Transformers $\{\mathrm{T}_j\}_{j=0}^{N-1}$ and equal to 1 if T_j is frozen, 0 otherwise. In this study, we set $\gamma = 1$ in Equation 2 to nullify the contribution of the frozen encoder and focus on the attention from the region-level and slide-level Transformers, which are explicitly trained on the prediction task.

Attention scores were normalized at the slide level, mapped to RGB colors using a diverging colormap, and overlaid on top of their respective spatial locations in the whole-slide image. To facilitate interpretation, scores were smoothed with a Gaussian kernel and values below 0.5 were filtered out, leaving only salient regions visible. Hotspots of high attention, shown in red, indicate positive evidence with strong relative contribution to the model prediction compared to other regions. Heatmaps were overlaid on the original H&E slide with a transparency value of 0.5 to simultaneously visualize the underlying tissue morphology.

4.5 Tile-level contribution quantification via occlusion

We assessed the influence of individual tiles on the model's predictions using an occlusion-based perturbation approach. For each patient i, let S_i denote the sequence of feature vectors produced by a pretrained encoder for all non-overlapping $(256, 256)$ pixel tiles sampled at 0.50 mpp in that case:

$$\mathrm{S}_i = (x_1^i, x_2^i, \ldots, x_{n_i}^i)$$

where n_i is the number of tiles for case i. The trained recurrence-risk model is written as:

$$\phi : (x_1^i, x_2^i, \ldots, x_{n_i}^i) \mapsto r_i \in \mathbb{R}$$

We first computed the patient-level risk score:

$$r_i^{\text{legacy}} = \phi\big(\mathrm{S}_i\big)$$

To assess the influence of tile t on the model's prediction, we constructed an occluded version of the case by removing that tile from the sequence, resulting in a sequence of length $n_i - 1$:

$$\mathrm{S}_i^t = (x_1^i, \ldots, x_{t-1}^i, x_{t+1}^i, \ldots, x_{n_i}^i)$$

Thus S_i^t contains one fewer element than S_i. The corresponding occluded risk score is:

$$r_i^t = \phi\big(\mathrm{S}_i^t\big)$$

The contribution score of tile t in case i was defined as:

$$c_i^t = r_i^{\text{legacy}} - r_i^t$$

By construction, $c_i^t > 0$ indicates a *risk-increasing* tile (occlusion decreases predicted risk), whereas $c_i^t < 0$ indicates a *risk-lowering* tile (occlusion increases predicted risk). For each case i, we retained the $K = 10$ tiles with the most positive contributions and the $K = 10$ tiles with the most negative contributions. To mitigate outliers, contribution scores were clipped at the 5th (q_5) and 95th (q_{95}) percentiles computed across all tiles from all 100 cases in the `RUMC` test set:

$$\hat{c}_i^t = \min\big(\max(c_i^t, q_5),\, q_{95}\big)$$

Clipped values were normalized by the maximum absolute contribution score, yielding normalized scores in $[-1, 1]$:

$$\bar{c}_i^t = \frac{\hat{c}_i^t}{\max_{j,k} |\hat{c}_j^k|}$$

For each selected tile, we extracted the model-refined representation, corresponding to the output of the final region-level transformer layer prior to attention pooling. These embeddings were projected into two dimensions using t-SNE, with points colored by their normalized contribution $\bar{c}_i^t$. Within five visually coherent clusters, we randomly sampled up to 20 representative tiles for qualitative assessment by an expert pathologist (Appendix D).

4.6 Statistics

We assessed statistical significance of performance differences using non-parametric resampling methods. All analyses were conducted at the patient level using ensemble-derived risk scores. Model performance was quantified using Harrell's concordance index (c-index), with 95% confidence intervals estimated by event-stratified bootstrapping with 4000 resamples. We report p-values and adjusted q-values in Appendix A.

Pairwise model comparisons. To compare encoders within each test cohort, we performed paired bootstrap resampling of the difference in c-index between two models, using two-sided tests. Resulting p-values were adjusted for multiple comparisons with the Benjamini-Hochberg procedure.

Training set enrichment effect. To evaluate the effect of enriching the training set with `TCGA-PRAD` cases, we applied the same procedure to compare ensemble risk scores from models trained on `RUMC-only` versus `RUMC+TCGA`, stratified by encoder and test cohort.

4.7 Datasets

We conducted large-scale evaluation across five multi-institutional radical prostatectomy cohorts. All datasets included digitized whole-slide images at approximately 0.25 microns per pixel and clinical follow-up for biochemical recurrence, providing time-to-event or time-to-censoring information.

Additional dataset details, including inclusion and exclusion criteria and scanner specifications, are provided in Appendix E.

4.7.1 RUMC

The Radboud University Medical Center (`RUMC`) cohort includes 608 patients who underwent radical prostatectomy between 1992 and 2012 in Nijmegen, The Netherlands. For each patient, histopathology slides corresponding to the highest-grade tumor – identified from the original pathology report – were retrieved from the hospital archive and digitized using a 3DHistech P1000 scanner. This cohort was part of the `LEOPARD` challenge (LEarning biOchemical Prostate cAncer Recurrence from histopathology sliDes). We used the 508 publicly released cases for model development and 100 held-out test cases, totaling 1723 slides. Clinico-pathological variables included preoperative PSA, extracapsular extension, surgical margin status, seminal vesicle invasion, lymph node involvement, and prostatectomy Gleason grade. Gleason grading for this cohort was based on the original pathology report and was not re-evaluated.

4.7.2 TCGA-PRAD

The Cancer Genome Atlas Prostate Adenocarcinoma (`TCGA-PRAD`) cohort consists of 403 patients with at least one associated prostatectomy diagnostic slide. Although TCGA provides annotations for BCR, we found these labels to be inconsistent across different clinical files. To construct a reliable ground truth for our study, we curated BCR status and time-to-event information by parsing the case-specific `.xml` files available through the TCGA portal. We identified 298 patients with BCR information (62 events), while the remaining 105 patients were unlabeled and used only for self-supervised pretraining (Section 4.1), for a total of 449 slides.

4.7.3 PLCO

The Prostate, Lung, Colorectal, and Ovarian (`PLCO`) Cancer Screening Trial enrolled approximately $155,000$ participants between 1993 and 2001, with prostate cancer follow-up available through December 31, 2009. For this study, we selected prostate cancer patients whose primary treatment was radical prostatectomy. An experienced pathologist (J.B.) manually reviewed all available histopathology slides to identify those containing cancer, allowing us to exclude benign slides from further analysis. In the end, we included 723 patients, for a total of 1971 slides. BCR status was determined from

standardized PSA follow-up records, and all cases were regraded according to the latest ISUP guidelines [33] by pathology residents under supervision of expert uropathologists. These cases were also part of the `LEOPARD` challenge.

4.7.4 IMP

The Instituto Mário Penna (`IMP`) cohort comprised 421 patients who underwent radical prostatectomy in 2016 in Belo Horizonte, Brazil. Patients were all followed up until 2022. Histopathology slides for each case were digitized using a Motic Easy Scan Infinity 60N scanner, resulting in a total of 2091 digital prostatectomy slides. Patient-level clinico-pathological data included preoperative PSA levels, prostatectomy Gleason grades, extracapsular extension, surgical margin status, seminal vesicle invasion, and lymph node involvement.

4.7.5 UHC

The University Hospital Cologne (`UHC`) cohort included 330 patients who underwent radical prostatectomy between 2015 and 2020 in Cologne, Germany. Patients were followed through 2023 in accordance with European Association of Urology (EAU) guidelines. Histopathology slides for each case were digitized using a Hamamatsu NanoZoomer S360 scanner, resulting in a total of 1028 digital prostatectomy slides. Patient-level clinico-pathological data included preoperative PSA, prostatectomy Gleason grades, and pathological staging.

4.8 CAPRA-S

CAPRA-S combines six clinicopathological variables into a 0–12 point scale (Table 5) to stratify patients into three groups at low ($0--2$), intermediate ($3--5$), and high (≥ 6) risk of recurrence [4].

Several of these clinicopathological variables were missing or partially available in some cohorts. When possible, we inferred surrogate values using information derived from pathological staging, based on the American Joint Committee on Cancer 8th edition TNM system for prostate cancer, which is routinely used in clinical and pathological assessment. Specifically:

- Extracapsular extension: assumed for pT-stage $\geq$ T3a
- Seminal vesicle invasion: assumed for pT-stage $\geq$ T3b
- Surgical margin positivity: assumed for pT-stage $\geq$ T3

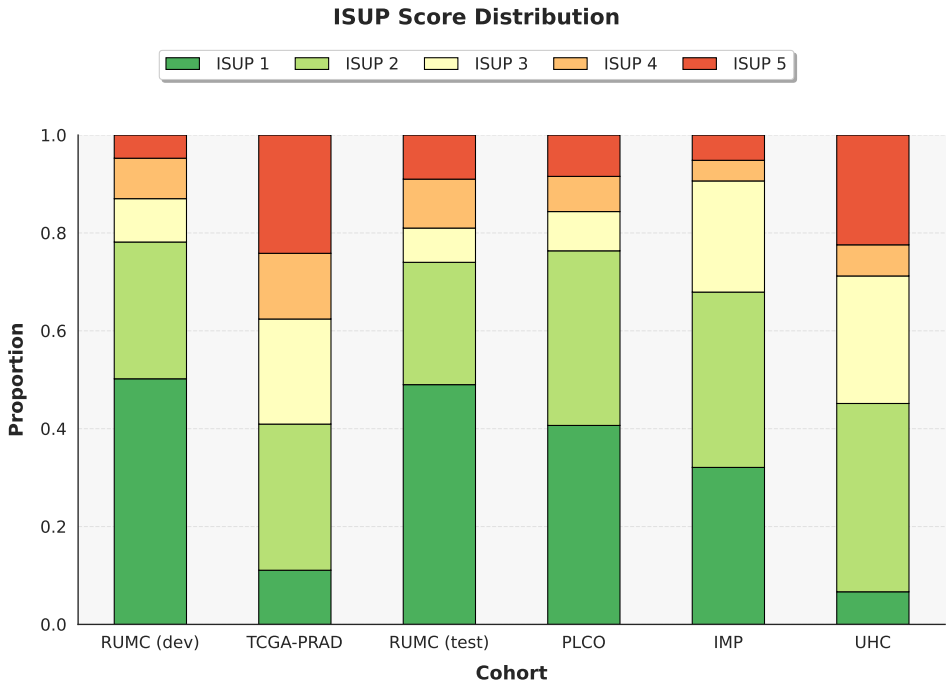


Figure 8: **ISUP grade distribution across cohorts.** Stacked bar plots show the proportion of patients in each cohort used in this study, grouped by ISUP grade. Each bar represents a cohort, with colors indicating ISUP grades 1 (green) to 5 (red). Differences in grade distributions reflect variation in recruitment periods, inclusion criteria, and treatment guidelines.

- Lymph node involvement: based on pN1 status

These assumptions applied to `RUMC`, `UHC`, and `IMP`. For `PLCO`, pathological T staging was based on earlier AJCC editions: the 4th edition in 22% of cases and the 5th edition in the remaining 78%. In this cohort, the only deviation from the above assumptions concerned seminal vesicle invasion, which was assumed for pT-stage $\geq$ T3c under the AJCC 4th edition, rather than $\geq$ T3b. No change was required for cases staged according to the AJCC 5th edition. Missing values were handled via median imputation within each cohort.

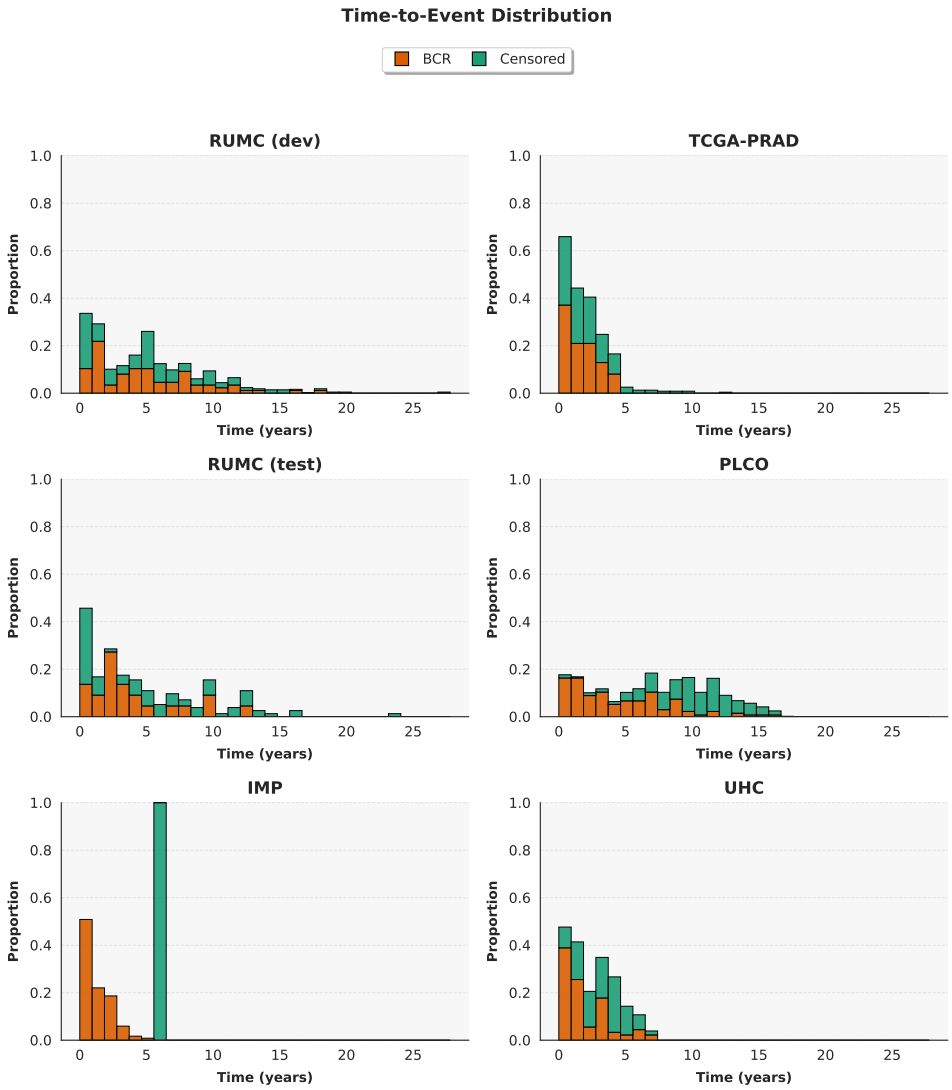


Figure 9: **Time-to-event distributions across cohorts.** Stacked histograms show times from radical prostatectomy to biochemical recurrence or last follow-up in each cohort used in this study. The horizontal axis represents time in years, and the vertical axis shows the proportion of patients within each cohort, subdivided into BCR (orange) and censored (green) cases. Consistent bin edges are used to facilitate comparison across cohorts. Cohort-specific differences in follow-up times reflect variations in recruitment periods, surveillance protocols, and clinical practices.

Table 5: Scoring system used for calculating the CAPRA-S score following radical prostatectomy. Each variable contributes a fixed number of points: preoperative prostate-specific antigen (PSA) level in ng/mL, prostatectomy Gleason score (pGS), status of the surgical margin (SM), seminal vesicle invasion (SVI), extracapsular extension (ECE), and lymph node involvement (LNI). Points are summed to yield the CAPRA-S score.

Variable	Value	Points	Variable	Value	Points
PSA	0–6	0	**pGS**	2–6	0
	6.01–10	1		3+4	1
	10.01–20	2		4+3	2
	>20	3		8–10	3
SM	Negative	0	**ECE**	No	0
	Positive	2		Yes	1
SVI	No	0	**LNI**	No	0
	Yes	2		Yes	1

4.9 Data preprocessing and experimental setup

Preprocessing. To prepare the data for model training, we first apply a deep learning-based tissue segmentation model to isolate tissue from the background in each whole-slide image. For patients where multiple slides are available, the segmented tissue regions are stitched into a unified, larger slide to enable patient-level analysis in a unified coordinate frame (Figure 2**A**) using a tissue-packing algorithm [34]. Slides from the `PLCO` dataset contain pen markings, which we detect and remove from the tissue masks using an artifact segmentation model [35]. We then use the Python library `slide2vec` to extract non-overlapping 2048 pixel regions at a resolution of 0.50 mpp. To account for slight variations in resolution introduced by different scanners, `slide2vec` uses a spacing tolerance parameter τ, which defines how much the native spacing can deviate from the target spacing during tiling. We set τ to 5% to accept native spacings between 0.475 and 0.525 mpp. If no native spacing falls within this range, a larger region is extracted at a higher resolution and resized to approximate 0.50 mpp. We filter out low-content regions by retaining only those with at least 1% tissue coverage, as estimated by the segmentation mask (Figure 2**B**).

Experimental setup. We partition the development data using five-fold cross-validation, stratifying jointly by recurrence status, prostatectomy ISUP grade, and whether the radical prostatectomy occurred before or after 2005, reflecting a change in clinical guidelines that year. We use the open-source Python library `scikit-multilearn`, which supports stratification across multiple discrete labels. To evaluate the impact of training data diversity on model performance, we maintain two distinct versions of the development data. The first includes 508 patients from the `RUMC` cohort, while the second combines these 508 `RUMC` patients with the 298 patients from the `TCGA-PRAD` cohort. Each version is split independently into five cross-validation folds, referred to as `RUMC-only` and `RUMC+TCGA` splits, respectively.

5 Human Ethics Approval

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the Netherlands Code of Conduct for Research Integrity.

The Medical Ethics Committee of Radboud University Medical Center approved the use of the `RUMC` data in this study on 5/1/2016 (Approval No. 20162275). Contact person: Geert Litjens, Department of Pathology, Radboud Institute for Health Science, Radboud University Medical Center, Nijmegen, The Netherlands (`geert.litjens@radboudumc.nl`).

The `UHC` cohort was created in the course of two studies, each approved by the institutional review board (Approval Nos. 20-1583 and 22-1233). Contact person: Iurii Tolkach, Institute of Pathology, University Hospital Cologne, Cologne, Germany (`iurii.tolkach@uk-koeln.de`).

The use of retrospective patient data from `IMP`, Belo Horizonte, Brazil, was approved by the institutional review board. Contact person: Paulo Guilherme de Oliveira Salles, Anatomical Pathology Service, Instituto Mario Penna, Belo Horizonte, Brazil (`paulo.salles@mariopenna.org.br`).

6 Consent to Participate

For the `RUMC` cohort, the Medical Ethics Committee of Radboud University Medical Center waived the requirement for informed consent. For the `UHC` cohort, the data were collected as part of two studies approved by the

institutional review board of the University Hospital Cologne. All consent procedures were carried out in accordance with the requirements of the approving committee. The institutional review board approval was waived for all analyses of retrospective, de-identified IMP patient data. Patients were informed at registration and were able to opt out at any time.

7 Data Availability

The development cohort from RUMC is publicly available under a CC-BY-NC-SA license via AWS at https://registry.opendata.aws/leopard. The curated dataset of 298 patients from TCGA-PRAD is available under a CC-BY license at https://huggingface.co/datasets/waticlems/tcga-prad-bcr. Data from PLCO is available upon submission of a project proposal through the Cancer Data Access System (CDAS).

8 Code Availability

The code used in this study is publicly available at https://github.com/clemsgrs/bcr-risk-prediction. We have also open-sourced our Prost40M model, which can be accessed at https://huggingface.co/waticlems/Prost40M.

Acknowledgements

The creation of the cohort of University Hospital Cologne was financed by the Wilhelm Sander Foundation, Munich, Germany (Y. Tolkach: grant 2022.040.1) and by Federal Ministry of Education and Research of Germany (BMBF; Project FED-PATH, Y. Tolkach and R. Büttner).

V. Agosti has been supported by the Beretta Foundation (Italy). E. Munari has been partially supported by the European Union - NextGenerationEU, through the Italian Ministry of University and Research under PNRR - M4C2-I1.3 Project PE_00000019 "HEAL ITALIA", CUP: B33C22001030006.

Author Contributions

C.G. conceptualization, data curation, model development, result analyses; K.F. data curation; N.U. data curation; V.A. data curation; E.M. result

analyses; S.F.K.J. data curation; P.G.d.O.S. IMP data curation; Y.T. UHC pathology data preparation; R.B. UHC pathology data preparation; S.S. UHC clinical data preparation; M.P. UHC clinical data preparation; A.H. UHC clinical data preparation. J.v.d.L. conceptualization, supervision; G.L. conceptualization, supervision. All authors reviewed the manuscript.

Competing Interests

J.v.d.L. reports consulting fees from Philips, ContextVision and AbbVie, and grants from Philips, ContextVision and Sectra, outside the submitted work. G.L. reports grants from the Dutch Cancer Society and the NWO, and grants from Philips Digital Pathology Solutions and personal fees from Novartis, outside the submitted work. All other authors declare no relevant competing interests concerning this paper.

References

[1] Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024). 2

[2] Cooperberg, M. R. *et al.* The university of california, san francisco cancer of the prostate risk assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy. *J. Urol.* **173**, 1938–1942 (2005). 2

[3] Sandeman, K. *et al.* Prostate mri added to capra, mskcc and partin cancer nomograms significantly enhances the prediction of adverse findings and biochemical recurrence after radical prostatectomy. *PLOS ONE* **15**, 1–14 (2020). 2

[4] Cooperberg, M. R., Hilton, J. F. & Carroll, P. R. The CAPRA-S score: A straightforward tool for improved prediction of outcomes after radical prostatectomy. *Cancer* **117**, 5039–5046 (2011). 2, 32

[5] Ozkan, T. A. *et al.* Interobserver variability in gleason histological grading of prostate cancer. *Scand. J. Urol.* **50**, 420–424 (2016). 2

[6] van Leenders, G. J. L. H. *et al.* The 2019 international society of urological pathology (ISUP) consensus conference on grading of prostatic carcinoma. *Am. J. Surg. Pathol.* **44**, e87–e99 (2020). 2

[7] Yamamoto, Y. *et al.* Automated acquisition of explainable knowledge from unannotated histopathology images. *Nat. Commun.* **10**, 5642 (2019). 3

[8] Leo, P. *et al.* Computer extracted gland features from H&E predicts prostate cancer recurrence comparably to a genomic companion diagnostic test: a large multi-site study. *NPJ Precis. Oncol.* **5**, 35 (2021). 3

[9] Pinckaers, H. *et al.* Predicting biochemical recurrence of prostate cancer with artificial intelligence. *Commun. Med. (Lond.)* **2**, 64 (2022). 3

[10] Dietrich, E. *et al.* Towards explainable end-to-end prostate cancer relapse prediction from he images combining self-attention multiple instance learning with a recurrent neural network. In Roy, S. *et al.* (eds.) *Proceedings of Machine Learning for Health*, vol. 158 of *Proceedings of Machine Learning Research*, 38–53 (PMLR, 2021). 3

[11] Farrokh, M., Kumar, N., Gann, P. H. & Greiner, R. Learning to predict prostate cancer recurrence from tissue images. *J. Pathol. Inform.* **15**, 100344 (2024). 3

[12] Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019). 3

[13] Hou, L. *et al.* Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016). 3

[14] Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In Dy, J. & Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, 2127–2136 (PMLR, 2018). 3

[15] Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019). 3

[16] Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021). 3

[17] Lerousseau, M., Vakalopoulou, M., Deutsch, E. & Paragios, N. Sparsec-onvmil: Sparse convolutional context-aware multiple instance learning for whole slide image classification. In Atzori, M. *et al.* (eds.) *Proceedings of the MICCAI Workshop on Computational Pathology*, vol. 156 of *Proceedings of Machine Learning Research*, 129–139 (PMLR, 2021). 3

[18] Shao, Z. *et al.* Transmil: Transformer based correlated multiple instance learning for whole slide image classification (2021). 2106.00908. 3

[19] Chen, R. J. *et al.* Scaling vision transformers to gigapixel images via hierarchical self-supervised learning (2022). 2206.02647. 3, 7, 26, 27

[20] Grisi, C. *et al.* Hierarchical vision transformers for prostate biopsy grading: Towards bridging the generalization gap. *Medical Image Analysis* **105**, 103663 (2025). 3, 17, 27, 28

[21] Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* **81**, 102559 (2022). 3

[22] Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024). 3, 7, 27

[23] Zimmermann, E. *et al.* Virchow2: Scaling self-supervised mixed magnification models in pathology (2024). 2408.00738. 7, 27

[24] Saillard, C. *et al.* H-optimus-0 (2024). 7, 27

[25] Nguyen, J. K. *et al.* Proposal for an optimised definition of adverse pathology (unfavourable histology) that predicts metastatic risk in prostatic adenocarcinoma independent of grade group and pathological stage. *Histopathology* **85**, 598–613 (2024). 15

[26] van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008). URL http://jmlr.org/papers/v9/vandermaaten08a.html. 18

[27] de Jong, E. D., Marcus, E. & Teuwen, J. Current pathology foundation models are unrobust to medical center differences (2025). 2501.18055. 18

[28] Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2021). 24

[29] Chen, R. J. *et al.* Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4025 (2021). 25, 26

[30] Zadeh, S. G. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3126–3137 (2021). 25

[31] Jaume, G. *et al.* Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems* (2024). 27

[32] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2017). 1412.6980. 27

[33] Epstein, J. I. & Hirsch, M. S. A comparison of genitourinary pathology society (GUPS) and international society of urological pathology (ISUP) prostate cancer grading guidelines. *Am. J. Surg. Pathol.* **45**, 1005–1007 (2021). 32

[34] Aswolinskiy, W. *et al.* Attention-based whole-slide image compression achieves pathologist-level pre-screening of multi-organ routine histopathology biopsies. *medRxiv* (2024). 35

[35] Lucassen, R. T., Blokx, W. A. M. & Veta, M. Tissue cross-section and pen marking segmentation in whole slide images. In Tomaszewski, J. E. & Ward, A. D. (eds.) *Medical Imaging 2024: Digital and Computational Pathology*, vol. 12933, 129330B. International Society for Optics and Photonics (SPIE, 2024). 35

A Statistical Testing

A.1 Pairwise Model Comparisons

Table 6: **Pairwise comparison of encoders trained on `RUMC-only` data.** The table reports the difference in concordance index (Δ) between pairs of encoders when evaluated on each test cohort. For each comparison, Δ values are shown together with 95% bootstrap confidence intervals (CI) and false discovery rate (FDR) adjusted q values. Positive Δ values indicate higher performance of the encoder listed first in the comparison, whereas negative values indicate higher performance of the encoder listed second. Values that remain significant after correction ($q < 0.05$) are shown in bold.

Cohort	Model Comparison	Δ	95% CI	q
RUMC	Prost40M - UNI	-0.017	[-0.107, 0.080]	0.852
	Prost40M - Virchow2	-0.031	[-0.128, 0.075]	0.852
	Prost40M - H-optimus-0	-0.003	[-0.103, 0.079]	0.977
	UNI - Virchow2	-0.014	[-0.070, 0.042]	0.852
	UNI - H-optimus-0	+0.014	[-0.069, 0.086]	0.852
	Virchow2 - H-optimus-0	+0.029	[-0.051, 0.102]	0.852
PLCO	Prost40M - UNI	**-0.089**	[-0.138, -0.041]	< 0.001
	Prost40M - Virchow2	**-0.138**	[-0.181, -0.095]	< 0.001
	Prost40M - H-optimus-0	-0.025	[-0.064, 0.014]	0.208
	UNI - Virchow2	**-0.048**	[-0.074, -0.024]	< 0.001
	UNI - H-optimus-0	**+0.065**	[0.031, 0.101]	0.002
	Virchow2 - H-optimus-0	**+0.113**	[0.083, 0.146]	< 0.001
IMP	Prost40M - UNI	**-0.205**	[-0.266, -0.145]	< 0.001
	Prost40M - Virchow2	**-0.245**	[-0.298, -0.194]	< 0.001
	Prost40M - H-optimus-0	**-0.169**	[-0.215, -0.124]	< 0.001
	UNI - Virchow2	**-0.040**	[-0.073, -0.008]	0.028
	UNI - H-optimus-0	+0.036	[-0.008, 0.083]	0.115
	Virchow2 - H-optimus-0	**+0.076**	[0.044, 0.111]	< 0.001
UHC	Prost40M - UNI	-0.012	[-0.054, 0.031]	0.608
	Prost40M - Virchow2	-0.038	[-0.079, 0.003]	0.148
	Prost40M - H-optimus-0	+0.015	[-0.032, 0.063]	0.608
	UNI - Virchow2	-0.026	[-0.053, 0.000]	0.148
	UNI - H-optimus-0	+0.027	[-0.007, 0.060]	0.180
	Virchow2 - H-optimus-0	**+0.053**	[0.031, 0.075]	< 0.001

Table 7: **Pairwise comparison of encoders trained on RUMC+TCGA data.** The table reports the difference in concordance index (Δ) between pairs of encoders when evaluated on each test cohort. For each comparison, Δ values are shown together with 95% bootstrap confidence intervals (CI) and false discovery rate (FDR) adjusted q values. Positive Δ values indicate higher performance of the encoder listed first in the comparison, whereas negative values indicate higher performance of the encoder listed second. Values that remain significant after correction ($q < 0.05$) are shown in bold.

Cohort	Model Comparison	Δ	95% CI	q
RUMC	Prost40M - UNI	+0.051	[-0.021, 0.141]	0.283
	Prost40M - Virchow2	-0.007	[-0.124, 0.115]	0.882
	Prost40M - H-optimus-0	+0.088	[0.005, 0.183]	0.151
	UNI - Virchow2	-0.058	[-0.153, 0.024]	0.283
	UNI - H-optimus-0	+0.037	[-0.056, 0.137]	0.548
	Virchow2 - H-optimus-0	+0.096	[-0.000, 0.201]	0.151
PLCO	Prost40M - UNI	+0.037	[0.000, 0.077]	0.114
	Prost40M - Virchow2	+0.036	[-0.001, 0.076]	0.114
	Prost40M - H-optimus-0	**+0.060**	[0.022, 0.102]	0.024
	UNI - Virchow2	-0.002	[-0.035, 0.033]	0.943
	UNI - H-optimus-0	+0.023	[-0.009, 0.057]	0.198
	Virchow2 - H-optimus-0	+0.024	[-0.009, 0.059]	0.198
IMP	Prost40M - UNI	+0.031	[-0.011, 0.075]	0.216
	Prost40M - Virchow2	-0.007	[-0.060, 0.048]	0.793
	Prost40M - H-optimus-0	**+0.137**	[0.079, 0.197]	< 0.001
	UNI - Virchow2	-0.038	[-0.097, 0.023]	0.271
	UNI - H-optimus-0	**+0.106**	[0.038, 0.174]	0.003
	Virchow2 - H-optimus-0	**+0.144**	[0.092, 0.196]	< 0.001
UHC	Prost40M - UNI	+0.006	[-0.028, 0.042]	0.901
	Prost40M - Virchow2	-0.041	[-0.078, -0.003]	0.070
	Prost40M - H-optimus-0	+0.008	[-0.031, 0.046]	0.901
	UNI - Virchow2	**-0.047**	[-0.079, -0.018]	0.009
	UNI - H-optimus-0	+0.002	[-0.030, 0.034]	0.903
	Virchow2 - H-optimus-0	**+0.049**	[0.015, 0.084]	0.011

A.2 Effect of Training Dataset Enrichment

Table 8: **Effect of training data enrichment on model performance.** The table shows the change in concordance index (Δ) between models trained on the combined `RUMC+TCGA` splits and those trained on `RUMC-only` splits. For each test cohort and encoder, Δ values are reported alongside 95% bootstrap CI, the exact two-sided p values, and FDR-adjusted q values. Positive Δ values indicate higher performance after training data enrichment, whereas negative values indicate lower performance. Values that remain significant after FDR correction ($q < 0.05$) are highlighted in bold.

Cohort	Encoder	Δ	95% CI	p	q
RUMC	Prost40M	**+0.074**	[0.023, 0.136]	0.002	0.008
	UNI	+0.006	[-0.060, 0.067]	0.866	0.990
	Virchow2	+0.050	[-0.019, 0.139]	0.190	0.380
	H-optimus-0	-0.017	[-0.112, 0.075]	0.708	0.871
PLCO	Prost40M	**+0.107**	[0.052, 0.161]	< 0.001	< 0.001
	UNI	-0.020	[-0.064, 0.021]	0.353	0.565
	Virchow2	**-0.067**	[-0.103, -0.033]	< 0.001	< 0.001
	H-optimus-0	+0.022	[-0.031, 0.072]	0.414	0.602
IMP	Prost40M	**+0.238**	[0.159, 0.318]	< 0.001	< 0.001
	UNI	+0.002	[-0.050, 0.053]	0.933	0.993
	Virchow2	0.000	[-0.032, 0.031]	0.993	0.993
	H-optimus-0	-0.068	[-0.130, -0.006]	0.033	0.104
UHC	Prost40M	+0.031	[-0.024, 0.083]	0.260	0.462
	UNI	+0.014	[-0.040, 0.064]	0.593	0.790
	Virchow2	+0.034	[-0.000, 0.068]	0.052	0.139
	H-optimus-0	+0.038	[-0.012, 0.084]	0.133	0.303

B Ensemble and joint modeling performance of models trained on RUMC-only splits

Table 9: **Ensemble and joint modeling performance of models trained on RUMC-only.** Summary of the prognostic performance of deep learning models evaluated on the four held-out test cohorts. For each cohort, the *ensemble* c-index reflects the standalone performance of the deep learning model and is computed by averaging patient-level risk scores from the five RUMC-only cross-validation folds. The *joint* c-index is obtained by fitting a multivariable CoxPH model using both the ensemble risk score and the CAPRA-S score as covariates. For joint models, the hazard ratio (HR) with 95% confidence interval (CI) and the Wald test p-value indicate the independent contribution of the ensemble risk score after adjusting for CAPRA-S. Bold values indicate the highest c-index observed for each cohort.

Test set	Encoder	Ensemble	Joint	HR (95% CI)	p-value
RUMC	Prost40M	0.686	0.754	1.50 (1.05-2.15)	0.028
	UNI	0.703	0.739	1.31 (0.93-1.85)	0.122
	Virchow2	**0.717**	**0.759**	1.50 (1.04-2.17)	0.030
	H-optimus-0	0.688	0.733	1.17 (0.86-1.60)	0.327
PLCO	Prost40M	0.604	0.766	1.07 (0.98-1.17)	0.121
	UNI	0.694	0.769	1.26 (1.11-1.43)	< 0.001
	Virchow2	**0.742**	**0.778**	1.27 (1.13-1.43)	< 0.001
	H-optimus-0	0.629	0.763	1.18 (1.04-1.32)	0.007
IMP	Prost40M	0.462	0.771	0.88 (0.75-1.03)	0.123
	UNI	0.667	0.775	1.05 (0.92-1.20)	0.497
	Virchow2	**0.707**	**0.779**	1.17 (1.01-1.36)	0.037
	H-optimus-0	0.631	0.774	1.10 (0.93-1.29)	0.271
UHC	Prost40M	0.663	**0.750**	1.04 (0.93-1.16)	0.528
	UNI	0.675	0.748	1.20 (1.00-1.44)	0.052
	Virchow2	**0.701**	0.747	1.26 (1.04-1.54)	0.018
	H-optimus-0	0.648	0.746	1.08 (0.90-1.29)	0.412

C ISUP-stratified performance on UHC

The UHC cohort consists of 330 cases, distributed as follows: 22 ISUP 1, 127 ISUP 2, 86 ISUP 3, 21 ISUP 4, and 74 ISUP 5.

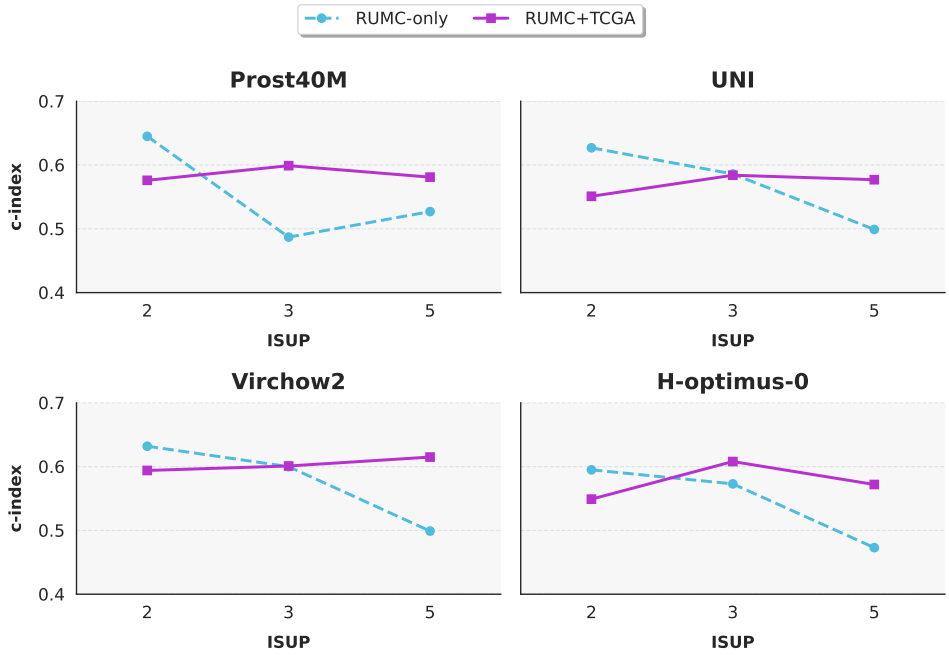


Figure 10: **ISUP-stratified predictive performance on the UHC cohort before and after training set enrichment.** Ensemble c-index of models trained on RUMC-only versus RUMC+TCGA, stratified by ISUP grade within the UHC cohort. Results exclude grade 1 (all censored, preventing c-index calculation) and grade 4 (small sample size). Models trained on RUMC-only show performance skewed toward ISUP 2, reflecting the grade distribution in the training set. Adding TCGA-PRAD cases reduces this bias, improving performance in higher-grade cases (ISUP 3 and 5) and yielding a flatter profile across grades, consistent with better calibration of recurrence risk across grades.

D Contribution cluster analysis

For three risk-increasing and two risk-lowering clusters, we present grids of representative tiles. Each grid displays individual tiles highlighted within their surrounding whole-slide context, facilitating interpretation of the histomorphological patterns that the model associates with increased or reduced recurrence risk.

D.1 Risk-increasing clusters

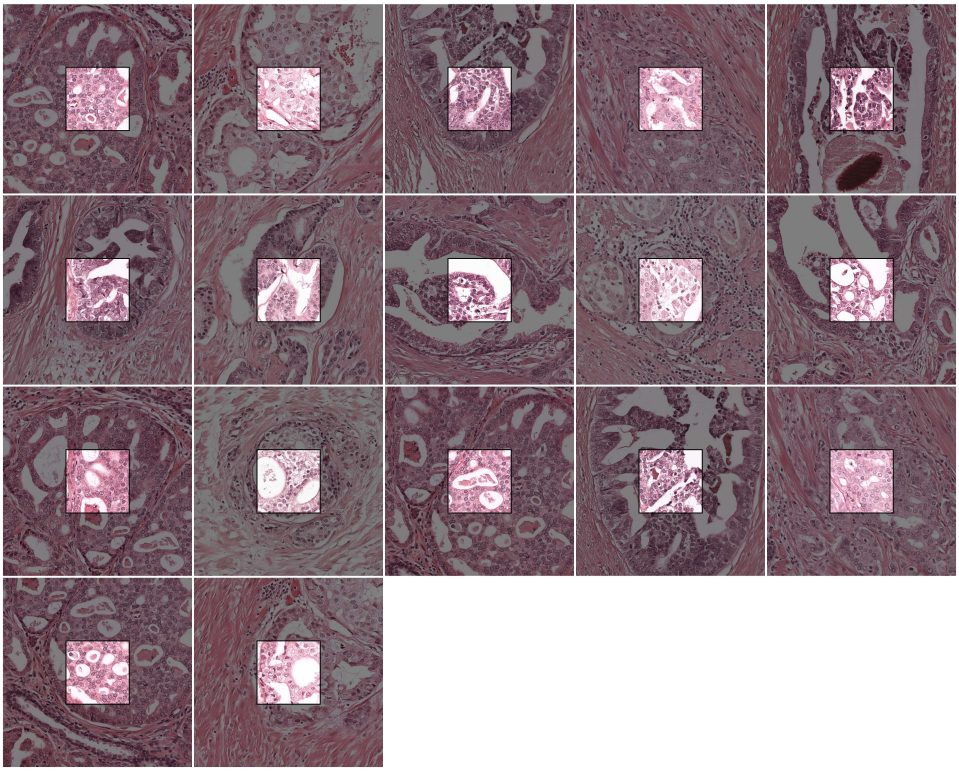


Figure 11: **Representative tiles from risk-increasing cluster 1.** This cluster is dominated by adverse histology, including extensive cribriform structures and fused glands forming anastomosing cords, aligning with the model's interpretation of this cluster as increasing risk of recurrence.

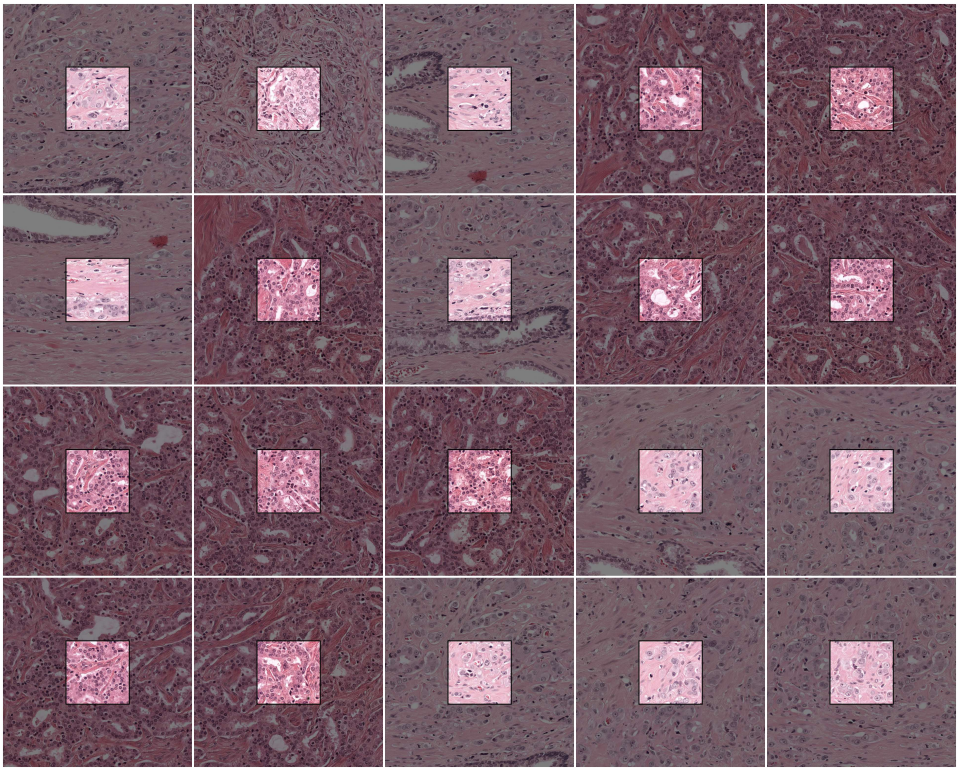


Figure 12: **Representative tiles from risk-increasing cluster 2.** Tiles predominantly contain high-grade carcinoma, including single infiltrating cells characteristic of Gleason pattern 5 and fused pattern 4 glands forming anastomosing cords. These morphologies are strongly prognostically adverse and align with large positive contributions to risk of recurrence.

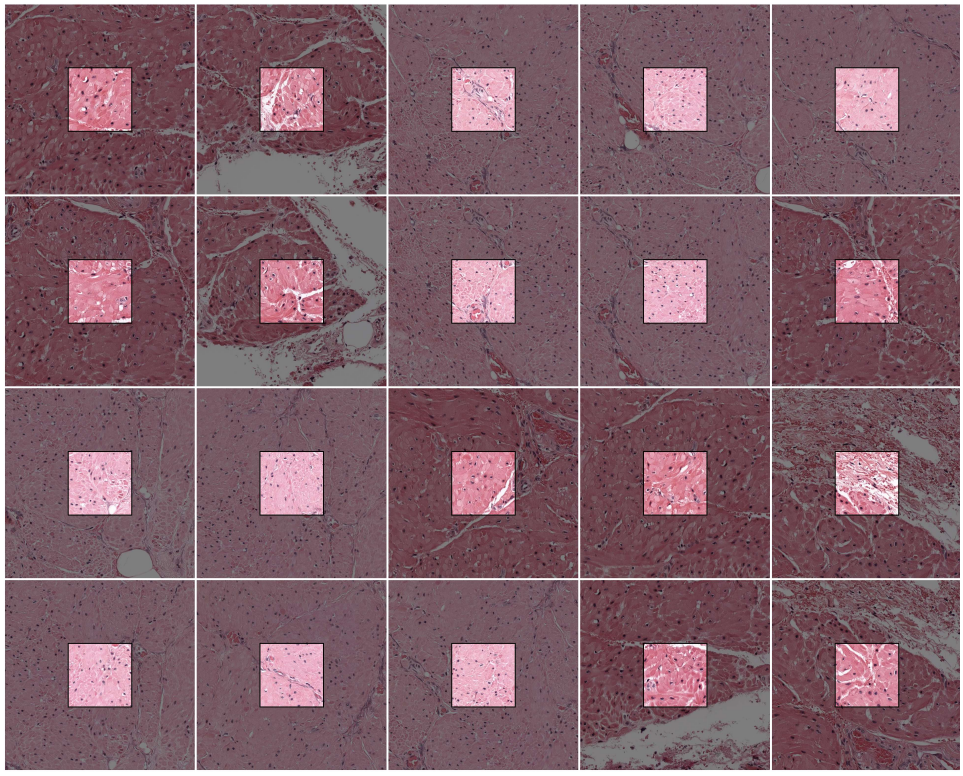


Figure 13: **Representative tiles from risk-increasing cluster 3.** Tiles primarily depict fibromuscular stroma rather than carcinoma, suggesting that the model may partially attribute recurrence risk based on patient- or slide-specific non-tumor features.

D.2 Risk-lowering clusters

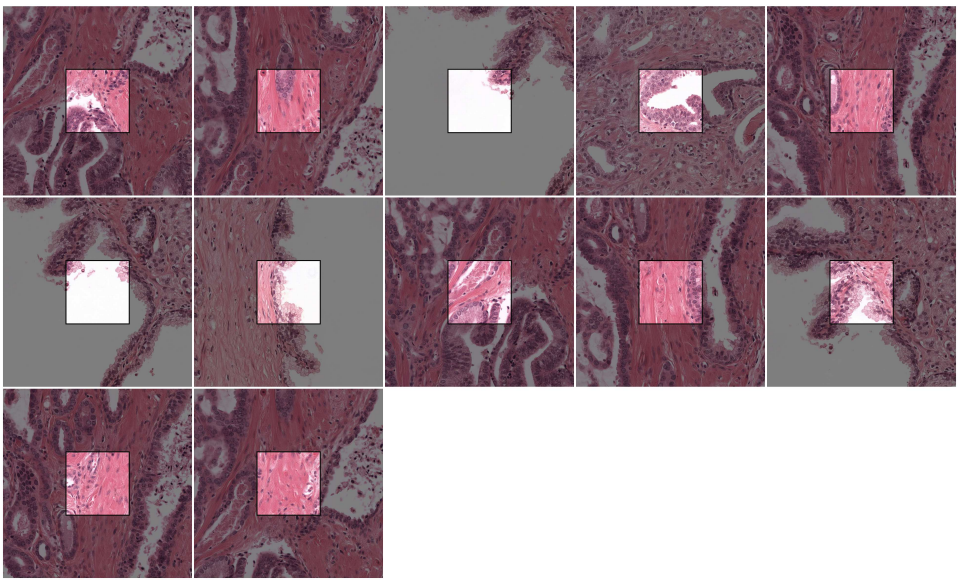


Figure 14: **Representative tiles from risk-lowering cluster 1.** Most tiles display predominantly benign tissue with only occasional proximity to carcinoma.

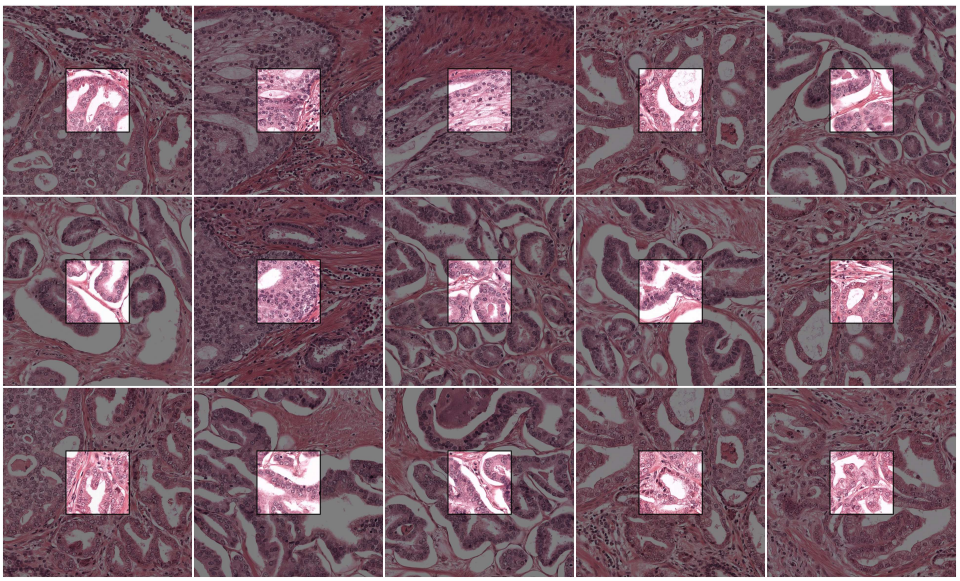


Figure 15: **Representative tiles from risk-lowering cluster 2.** While some tiles show cribriform morphology which should confer higher risk, the overall Gleason pattern 4 complexity seen in other tiles is more consistent with a favorable histological appearance.

E Dataset details

We provide detailed inclusion and exclusion criteria, scanner specifications, tumor subtype considerations, BCR label definitions, and follow-up data extraction procedures for all datasets used in this study.

E.1 RUMC

The Radboud University Medical Center (RUMC, Nijmegen, The Netherlands) cohort initially consisted of 903 patients who underwent radical prostatectomy between 1992 and 2012. Patients were monitored for biochemical recurrence through PSA follow-up, typically within one year at the hospital and thereafter by general practitioners. The following cases were excluded: no consent (2), no follow-up information (22), PSA levels failing to reach nadir after prostatectomy (83), history of neoadjuvant treatment (12), absence of sufficiently digitizable slides (120), unclear tumor location (6), and predominant ductal carcinoma subtype (1). This resulted in a final cohort of 657 patients. Of these, 49 were used as an intermediate evaluation cohort during the `LEOPARD` challenge and are excluded from this study. We used the remaining 608 patients: 508 publicly released development cases and 100 held-out test cases.

E.2 TCGA-PRAD

TCGA includes multiple recurrence-related fields that were found to be inconsistent across data releases. To construct reliable BCR labels, we curated recurrence status and event times using case-specific clinical XML files. A recurrence event was assigned if there was either (1) a positive BCR indicator with a corresponding date of recurrence, (2) evidence of biochemical disease progression recorded as a new tumor event, or (3) a recurrence date provided without explicit indicator status. Patients without any sign of recurrence were labeled as censored. Patients with ambiguous or conflicting information were excluded. Through this process, we identified 298 patients with BCR labels, including 62 recurrence events and 236 censored cases. These patients were linked to 332 WSIs. Although the remaining 105 patients had incomplete or unreliable recurrence information, their slides were retained for self-supervised pretraining. In total, this resulted in 403 patients and 449 slides from `TCGA-PRAD` used in this study.

E.3 PLCO

The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial enrolled approximately 155,000 participants between 1993 and 2001, with follow-up through December 31, 2009. We identified patients with BCR using the `rp_bcr_0p2consec` variable, for which we could get time to BCR as the time from surgery to two consecutive positive PSA measurements (`rp_bcr_0p2consec_mo`). For patients without BCR, time to censoring or last follow-up was measured from the date of surgery (calculated as `primary_trtp_days - candxdaysp`) to the date of study exit (`pcp_bcr_exit_days`). Cases with reported distant metastasis present at the time of surgery (1) and negative time to follow-up (2) were excluded. In the end, we obtained 773 patients with at least one digital pathology slide. All slides had been scanned using a Leica scanner at a resolution of 0.25 mpp. Like the `RUMC` cohort, these cases were included in the `LEOPARD` challenge, where 50 were used for intermediate evaluation and 723 were held out for final testing. In this study, we only used the 723 hold-out cases for testing, excluding the 50 cases used during the challenge evaluation phase.

E.4 IMP

The Instituto Mário Penna (IMP) cohort initially consisted of 437 patients who underwent radical prostatectomy in 2016. Patients were all followed up until 2022: every three months during the first year, every four months in the second year, every six months during the third and fourth years, and annually from the fifth year onward. No patients had documented distant metastasis at the time of surgery or received neoadjuvant hormonal therapy. In all cases, the predominant cancer component was not ductal adenocarcinoma. We excluded cases without any digitized slide (10), and cases where PSA levels failed to reach a nadir after radical prostatectomy (6), resulting in a final cohort of 421 patients. Histopathology slides for each case were digitized using a Motic Easy Scan Infinity 60N scanner at a resolution of 0.26 mpp, resulting in a total of 2091 digital prostatectomy slides.

E.5 UHC

The University Hospital Cologne (UHC) cohort initially consisted of 337 patients who underwent radical prostatectomy between 2015 and 2020 in Cologne, Germany. All patients had no history of hormonal therapy or

evidence of therapy-induced regression, as per the cohort's predefined inclusion criteria. We further excluded cases where the majority of the tumor was classified as ductal adenocarcinoma (3), cases with negative or missing time to recurrence (3), and cases where PSA levels failed to reach a nadir after radical prostatectomy (1), resulting in a final cohort of 330 patients. Histopathology slides for each case were digitized using a Hamamatsu NanoZoomer S360 scanner at a resolution of 0.23 mpp, resulting in a total of 1028 digital prostatectomy slides.

Importantly, the definition of BCR differed in the `UHC` cohort compared with the other datasets. Whereas `RUMC`, `PLCO`, and `IMP` adopt the historical standard of two consecutive PSA measurements ≥ 0.2 ng/mL, `UHC` defines BCR as a single postoperative PSA ≥ 0.1 ng/mL. This shift reflects the evolution of clinical practice and improvements in PSA assay sensitivity, with the second confirmatory test rarely altering classification. Another key difference is the distribution of tumor grades: the `UHC` cohort contains relatively few low-grade (ISUP 1) prostate cancers. This again reflects changes in treatment guidelines: active surveillance is now commonly recommended for low-grade disease, resulting in fewer such patients undergoing surgery. The grade distribution in `UHC` is more similar to that of the `TCGA-PRAD` cohort, and contrasts with the `RUMC`, `PLCO`, and `IMP` cohorts, which include a higher proportion of low-grade cases (Figure 8). While introducing heterogeneity, these differences make the `UHC` cohort particularly valuable for testing model generalization to contemporary clinical practice.