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.

An outcome-blind H&E architecture score is associated with prostate cancer-specific mortality

Anh Lê
Cedars-Sinai Medical Center
Priyanka Vasanthakumari
Cedars-Sinai Medical Center
Itzel Valencia
Cedars-Sinai Medical Center
Maryam Ranjpour Aghmiouni
Cedars-Sinai Medical Center
Marta Osrodek
Cedars-Sinai Medical Center
Mohamed Omar
mohamed.omar@csmc.edu
Cedars-Sinai Medical Center

Article

Keywords:

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10990195/v1

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

Additional Declarations: No competing interests reported.

Abstract

Grade Group (GG) is the dominant prognostic variable in localized prostate cancer (PCa), but it does not encode the proportion of high-grade tumor or adverse subtypes such as cribriform growth. We developed an outcome-blind hematoxylin and eosin (H&E) architecture score from a Gleason-pattern classifier trained on SICAPv2 annotations (155 slides, 95 patients) with frozen Prov-GigaPath embeddings to generate patient-level architecture scores in two external radical prostatectomy cohorts. In PLCO (n = 1,086 patients, 66 prostate cancer deaths), the score was associated with prostate cancer-specific mortality (hazard ratio [HR] 2.57, 95% confidence interval [CI] 1.73-3.84) but not with death from other causes (HR 0.98, 95% CI 0.89-1.07), and it remained significant after clinical adjustment (HR 1.88, 95% CI 1.28-2.77). The association replicated for disease-free survival in TCGA-PRAD (n = 394 patients, 72 events; HR 2.83, 95% CI 1.64-4.88). The signal was consistent across different pathology foundation-model encoders and within the Gleason 7 stratum, and decomposed into high-grade and cribriform architecture. Transcriptomically, the score was specifically correlated with proliferation and genomic-instability programs, and its survival association in TCGA-PRAD was robust to tissue-source site. Among men with biopsy Gleason ≤6, the top score tertile had a higher 20-year prostate cancer mortality than the low-score group (4.20% versus 0.96%), flagging a candidate biopsy-occult higher-risk subgroup.

INTRODUCTION

Gleason Grade Group (GG) is the dominant histologic prognostic variable in localized prostate cancer (PCa), but it is a categorical reduction of a morphology that is continuous, spatially heterogeneous, and architecturally specific. Tumors with the same GG can differ substantially in the amount and architectural subtype of high-grade cancer. Continuous measures of pattern 4/5 refine outcome stratification beyond categorical grade[1,2]. In a recent radical prostatectomy (RP) cohort restricted to GG2-4 disease, the absolute volume of pattern 4 was more strongly associated with advanced pathological stage and biochemical recurrence than GG[3]. These findings distinguish the presence of a high-grade component, which contributes to GG assignment, from its quantity, which varies within a grade category.

Architectural subtype provides an additional distinction. Poorly formed, fused, glomeruloid, and cribriform glands are all classified as pattern 4, yet their prognostic associations differ. In a meta-analysis of cribriform histology, the pattern was associated with biochemical recurrence, metastasis, and PCa-specific mortality (PCSM)[4], and within Gleason score 7 tumors, the stratum in which grade least determines the treatment decision, cribriform growth was independently associated with postoperative metastasis and disease-specific death[5]. Cribriform and intraductal carcinoma also mark a distinct genome. In The Cancer Genome Atlas (TCGA) and Canadian Prostate Cancer Genome Network (CPC-GENE) cohorts, tumors with cribriform or intraductal growth and Gleason score 7 or higher had greater copy-number alteration burden, with associated alterations including PTEN and TP53 loss and MYC gain[6]. The clinical relevance of these features is reflected in the 2026 European Association of Urology

guidelines, which recommend excluding patients with cribriform or intraductal histology on biopsy from active surveillance[7]. Nevertheless, pathology reports record cribriform growth as present or absent and do not quantify its burden or its spatial extent within the specimen.

Deep learning applied to hematoxylin and eosin (H&E) sections recovers prognostic information beyond the assigned grade. Gleason grading models have been associated with PCSM, and prespecified multimodal or outcome-derived systems have been validated on RP tissue[8–10]. Gleason-pattern intermixing have also predicted recurrence beyond 10 years independently of clinical variables[11], and automated cribriform detectors now reach pathologist-level agreement across international cohorts[12,13]. In parallel, outcome-supervised models trained across randomized radiotherapy trials have prognosticated distant metastasis and PCSM and predicted benefit from treatment duration[14–17], and one such model has been validated in an RP cohort[18]. Other image-only models predict recurrence or derive prognostic information from estimated Gleason-pattern proportions[19]. These studies establish the prognostic value of H&E, but they do not provide an outcome-independent histomorphologic model that is specifically associated with PCa death rather than other-cause mortality, and whose predictions can be decomposed into defined adverse histoarchitectural patterns.

To address this gap, we developed an outcome-blind histomorphologic model from a Gleason-pattern classifier trained on SICAPv2 biopsy annotations over frozen Prov-GigaPath tile embeddings[20] and fixed aggregation into patient-level architecture scores (Fig. 1). We then applied the model to RP tissue from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, with approximately two decades of mortality follow-up[21,22] and to TCGA prostate adenocarcinoma cohort (TCGA-PRAD), with disease-free survival (DFS) and molecular data[23]. We tested whether the score was associated specifically with PCSM, whether the association persisted after clinicopathological adjustment, which architectural and molecular features accounted for the signal, and whether the findings were robust to tissue encoder, site, and tissue sampling.

RESULTS

Gleason-pattern classifiers distinguish grade and cribriform growth

We developed a histomorphologic model based on a multilabel tile-level Gleason-pattern classifier that uses frozen Prov-GigaPath embeddings and is applied, without retraining, to two external RP cohorts with independent endpoints (**Fig. 1, Table 1**). We trained the classifier on the SICAPv2 needle-biopsy dataset, which comprises 155 whole-slide images (WSIs) from 95 patients spanning benign tissue through GG5, 119 of them containing cancer, to estimate independent probabilities for non-cancer, Gleason patterns 3, 4, and 5, and cribriform pattern 4 (**Fig. 2a-e**). A separate attention-based slide-level classifier was trained on the same embeddings to assess patient-level grade discrimination.

Table 1. Clinicopathologic characteristics of the study cohorts.

Characteristic	PLCO (n = 1086)	TCGA-PRAD (n = 394)
Age at diagnosis, years, median (IQR)	61.0 (58.0-65.0)	61.0 (56.0-66.0)
Diagnostic PSA, ng/mL, median (IQR)	5.6 (4.4-8.0)	not available
Gleason Grade Group		
GG1	545 (50.2)	40 (10.2)
GG2 and GG3 (Gleason 7)	404 (37.2)	208 (52.8)
GG4 (Gleason 8)	76 (7.0)	50 (12.7)
GG5 (Gleason 9-10)	54 (5.0)	96 (24.4)
Missing	7 (0.6)	0 (0.0)
Tumor stage		
Stage II / T2	790 (72.7)	154 (39.1)
Stage III / T3	230 (21.2)	227 (57.6)
Stage IV / T4	37 (3.4)	8 (2.0)
Missing	29 (2.7)	5 (1.3)
Endpoint	PCSM	DFS
Evaluable patients	1086 (100)	394 (100)
Events	66 (6.1)	72 (18.3)
Competing non-PCa deaths	480 (44.2)	not applicable
Median follow-up	22.6 years	35.0 months

The separate slide-level classifier distinguished GG5 from all other classes with an area under the receiver operating characteristic curve (AUC) of 0.927, with lower discrimination across GG2-GG4 (**Supplementary Fig. S1a**). Its macro one-versus-rest AUC was 0.840 (95% CI: 0.796-0.881) across all 155 slides and 0.763 (95% CI: 0.713-0.816) across the 119 cancer-containing slides. Across all slides, it distinguished GG ≥3 from benign tissue and GG1-GG2 with an AUC of 0.923 (95% CI: 0.876-0.963; **Supplementary Fig. S1b**). Fold-specific six-class macro AUCs ranged from 0.76 to 0.92, compared with a pooled out-of-fold macro AUC of 0.840 (**Supplementary Fig. S1c**). Patient-grouped out-of-fold evaluation of 12,081 annotated tiles yielded one-versus-rest AUCs of 0.957 for non-cancer, 0.894 for pattern 3, 0.877 for pattern 4, 0.929 for pattern 5, and 0.911 for cribriform pattern 4 (**Fig. 2g**). Because the cribriform comparison included benign and low-grade negative tiles, we also trained a separate cribriform-versus-non-cribriform pattern 4 classifier. Across 63 patients, this classifier achieved a cross-validated AUC of 0.859 (standard deviation [SD] 0.057; **Fig. 2h**). This standalone classifier was used to assess recognizability of the pattern and did not contribute to the architecture score.

The model-derived architecture score is specifically associated with prostate cancer mortality in PLCO

After SICAPv2 training, we applied the frozen Gleason-pattern classifier and fixed aggregation rule to generate patient-level architecture scores, without using survival outcomes. Each tile's adverse score was the maximum predicted probability across the pattern 4, pattern 5, and cribriform-like channels. Tumor tiles from all slides for a patient were pooled, and the mean of the highest-scoring 10% defined the patient-level architecture score (**Fig. 1**). The score showed partial concurrent validity with pathology grade: Spearman ρ was 0.38 (95% CI: 0.31-0.45) with worst biopsy Gleason in PLCO and 0.53 (95% CI: 0.46-0.59) with RP Gleason in TCGA-PRAD (**Fig. 2h**). It was therefore grade-associated, but not grade-equivalent.

PLCO was the main external mortality cohort and included 1,086 men, 66 PCa deaths, and 480 deaths from other causes (**Table 1**). In an age-adjusted cause-specific Cox model, each 1 SD increase in the score was associated with a PCSM HR of 2.57 (95% CI: 1.73-3.84; $p < 0.001$), but not with non-PCa-death (HR = 0.98, 95% CI: 0.89-1.07; **Fig. 3a**). A stacked Lunn-McNeil model distinguished the two effects (HR ratio = 2.64; 95% CI: 1.75-3.97; interaction $P < 0.001$). The pooled non-PCa-death null did not mask an association with any specific competing cause. Decomposing the 480 non-PCa deaths into their recorded causes, the score was not significantly associated with cardiovascular death (HR = 0.95, 95% CI: 0.82-1.11; 170 events) and death from a cancer other than PCa (HR = 0.98, 95% CI: 0.82-1.17; 130 events). The PCSM effect exceeded the effect for each cause by an HR ratio of 2.70 (95% CI: 1.76-4.14) and 2.62 (95% CI: 1.70-4.06), respectively, both interaction $P <0.001$.

A Fine-Gray model gave a concordant subdistribution HR of 2.67 (95% CI: 1.60-4.44; **Fig. 3a**). The proportional-hazards assumption was not fully satisfied in the age-adjusted PCSM model (global Grambsch-Therneau P = 0.048). The association remained in the same direction before and after the median event time of 16.6 years, but was stronger earlier (HR = 4.08; 95% CI: 2.06-8.07) than later (HR = 1.86; 95% CI: 1.14-3.04). The overall HR of 2.57 is therefore a time-averaged effect.

PCSM cumulative incidence separated across score tertiles. At 15 years, incidence was 5.1% in the high tertile and 0.6% in the low tertile; 41 and 11 PCSM events, respectively, accrued over complete follow-up among 362 patients per tertile (Gray test $\chi^2 = 26.7$, $P = 2.0 \times 10^{-6}$; **Fig. 3b**). Across the full PLCO cohort, 15-year PCSM incidence increased from 0.57% to 1.94% and 5.05% across low, middle, and high score tertiles, whereas non-PCa mortality remained at 13.14%, 12.83%, and 13.48%, respectively (**Fig. 3c**). Scores were higher in men who died of PCa than in the other outcome groups (Kruskal-Wallis $P = 4.4 \times 10^{-9}$; **Supplementary Fig. S2a**). The direction was stable across alternative tile-aggregation rules (**Supplementary Fig. S2b**) and in 1,000 patient-level bootstrap samples, and it was not explained by pen ink, tissue quantity, or scanner and stain effects. On whole-slide overlays, the score localized to poorly formed, fused, and cribriform tumor architecture and remained low over well-formed glands (**Fig. 3d**).

Clinical adjustment attenuates the association without establishing incremental discrimination

After adjustment for age, log diagnostic prostate-specific antigen (PSA), biopsy Gleason grade, and clinical stage, the score remained associated with PCSM (cause-specific HR = 1.88; 95% CI: 1.28-2.77; P = 0.0014; n = 1,027; 65 events; **Fig. 3a**). Adjustment for RP Gleason grade and pathologic stage, which uses pathology from the same specimen on which the score was measured, further attenuated the association (cause-specific HR = 1.63; 95% CI: 1.10-2.41; P = 0.014; n = 1,015; 64 events; **Fig. 3a**); the corresponding Fine-Gray HR was 1.74 (95% CI: 1.07-2.82).

In a standardized drop-one comparison within the pretreatment clinical model, biopsy Gleason grade had the largest independent contribution (likelihood-ratio $\chi^2 = 24.0$), followed by the architecture score (12.7), log PSA (3.8), age (1.6), and clinical stage (approximately 0; **Fig. 4a**). The score was only weakly correlated with biopsy Gleason ($\rho = 0.27$; n = 910), with broad overlap within each grade. Median scores increased from Gleason 6 to 8 (0.889, 0.926, and 0.965) and then plateaued at Gleason 9-10 (**Fig. 4b**). Adding the score changed the out-of-fold Harrell C-index from 0.763 to 0.790 ($\Delta C = 0.027$; 95% CI: −0.017-0.066; **Fig. 4c**). Incremental discrimination over the clinical model is therefore unresolved at the current event count. Time-dependent discrimination, calibration, and decision-curve analyses are shown in **Supplementary Fig. S3**.

High-grade and cribriform-like components account for the mortality association

We examined architecture sub-scores derived from the Gleason-pattern classifier's high-grade and cribriform-like outputs. In a joint age-adjusted model, the combined pattern 4/5 component (HR per SD = 1.76; 95% CI: 1.22-2.66) and the cribriform-like component (HR = 1.52; 95% CI: 1.22-1.91) were independently associated with PCSM. Tumor fraction (HR = 1.15; 95% CI: 0.89-1.48) and tile-level heterogeneity (HR = 0.98; 95% CI: 0.75-1.30) were not (**Fig. 4d**). The pattern 4/5 component recovered approximately 90% of the full score log-HR. The cribriform-like sub-score derives from the deployed multilabel Gleason pattern classifier's cribriform output, not the separate detector described above.

Risk was nonlinear across the score distribution. Relative to the median, the spline HR was 2.37 at the 75th percentile and 4.33 at the 90th percentile, while remaining nearly flat through the lower half of the range (likelihood-ratio test for nonlinearity, $\chi^2 = 10.1$, P = 0.0015; **Supplementary Fig. S2c,d**). The PCSM association was therefore concentrated in the upper score range rather than distributed uniformly across the scale.

The score shows a concordant association with disease-free survival in TCGA-PRAD

We applied the fixed histomorphologic model to TCGA-PRAD, an independent RP cohort with DFS rather than PCSM as the endpoint. Each 1-SD score increase was associated with DFS in age-adjusted Cox regression (HR = 2.83; 95% CI: 1.64-4.88; n = 394; 72 events; **Fig. 5a**). Because TCGA-PRAD does not report biopsy Gleason grade, a pretreatment clinical model comparable with PLCO could not be constructed. In the closest available postsurgical model, adjusted for RP grade and pathologic stage, the HR attenuated to 1.43 (95% CI: 0.83-2.47; n = 389; 72 events). We therefore interpret TCGA-PRAD as a

concordant prognostic association with a different endpoint and incremental value over pathology in this cohort is not established.

Ten-year recurrence incidence was 17% in the low tertile (11 events), 40% in the middle tertile (17 events), and 83% in the high tertile (44 events; log-rank $\chi^2 = 39.3$, 2 degrees of freedom, $P = 3.0 \times 10^{-9}$; **Fig. 5b**). Corresponding 5-year incidences were 13.4%, 19.5%, and 47.3%. Scores were higher among the 72 patients who recurred or progressed than among the 322 who remained disease-free (Mann-Whitney $P = 3.4 \times 10^{-7}$; **Fig. 5c**). The deployed probability aggregate approached its ceiling in this cohort, but a prespecified log-odds aggregate spread the distribution (interquartile range [IQR] = 2.80 for the logit aggregate versus 0.09 for the top-10% probability aggregate; **Supplementary Fig. S4a**) and retained a continuous DFS association (HR per SD = 1.89; 95% CI: 1.51-2.37; **Supplementary Fig. S4b,c**). This analysis indicates scale compression rather than loss of prognostic signal.

The direction was stable across alternative aggregation rules: HRs per SD were 4.19 (95% CI: 1.92-9.13) for the top 5%, 2.83 (95% CI: 1.64-4.88) for the top 10%, 2.44 (95% CI: 1.59-3.75) for the top 20%, 1.97 (95% CI: 1.53-2.54) for the fraction of tiles above 0.5, and 1.97 (95% CI: 1.53-2.53) for the maximum-region rule (**Supplementary Fig. S4d**). Whole-slide overlays again localized high scores to adverse glandular architecture (**Fig. 5d**).

The architecture score is associated with proliferation and genomic instability programs

In TCGA-PRAD, the score correlated with a transcriptomic proliferation index and genomic instability quantified as the fraction of genome altered (FGA). After adjustment for tumor purity, Gleason grade, and measured stromal fraction, partial Spearman correlations were 0.245 (95% CI: 0.117-0.369) for proliferation and 0.285 (95% CI: 0.150-0.403) for FGA (n = 230; Fig. 6a).

Gene-set enrichment analysis (GSEA) of all expressed genes recovered the same biology: E2F Targets (normalized enrichment score [NES] of 3.44), G2/M Checkpoint (3.10), MYC Targets V1 (2.79), MYC Targets V2 (2.45), DNA Repair (2.16), and Mitotic Spindle (2.09) were the leading positively enriched MSigDB hallmark sets, all at false discovery rate (FDR) $q < 0.01.25$. Channel-resolved analysis within tumors containing any pattern 4 (n = 192) showed related but distinct programs (**Fig. 6b**). Pattern 4 and cribriform-like content were enriched for cell-cycle, MYC, DNA-repair, and mTORC1 programs. Pattern 5 content instead showed enrichment for immune, inflammatory, angiogenic, and epithelial-mesenchymal transition (EMT) programs (NES = 3.12) and depletion of androgen response. The cribriform-like channel was associated with 1,634 protein-coding genes at $q < 0.05$ and with 11 enriched hallmark sets, led by E2F Targets (NES = 2.23), G2/M Checkpoint (1.99), and MYC Targets V1 (1.94); EMT was depleted (NES = −3.78; **Fig. 6c**). The E2F and PTEN-loss modules were also positively associated (partial $\rho = 0.30$, $q = 9.1 \times 10^{-5}$; and $\rho = 0.17$, $q = 0.032$). Restricting to tumors containing Gleason pattern 4 reduced but did not eliminate grade confounding.

Prespecified control modules supported biological specificity. After adjustment for purity, grade, and stromal fraction (n = 229), proliferation/cell-cycle and E2F/replication modules were associated with the

score (partial ρ = 0.267, q = 2.8 × 10^{-4}; and ρ = 0.197, q = 0.0088), whereas ischemia/immediate early response, ribosomal/housekeeping expression, and immune infiltration were null (all q = 0.86; **Fig. 6d**). The raw correlation with HoVer-Net-derived stromal fraction was weakly negative (ρ = −0.22; 95% CI, −0.33 to −0.10; n = 265; **Fig. 6e**), arguing against stromal abundance as the main source of the score.

We then tested whether the TCGA-PRAD survival association is a between-site batch effect of the foundation model encoder. A classifier decoded the tissue-source site from the frozen embeddings far above chance, at a cross-validated balanced accuracy of 0.92 against 0.083 expected for 12 sites, indicating that the encoder captures site information. Nevertheless, the age-adjusted DFS association persisted after stratification by tissue-source site (HR = 2.13; 95% CI: 1.21-3.74; P = 0.009), and all 12 leave-one-site-out estimates remained above 1 (range: 2.30-4.14; **Fig. 6f**). PLCO did not contain a comparable multisite axis.

These associations were measured in bulk RNA, so we tested whether they hold in situ. We related the architecture score to proliferation and androgen receptor (AR) programs across 12 spatial transcriptomic PRAD slides from five patients and a 3D reconstructed prostatectomy specimen of seven serial sections, using an autocorrelation-preserving spatial null with patient-level aggregation across the six patients (**Supplementary Fig. S5**). Across slides, stromal colocalization was inversely related to the score-proliferation (Pearson r = −0.91) and score-AR (r = −0.98) associations. The two internally replicated multi-slice patients showed positive epithelial associations, but patient-level evidence was insufficient: the small-cluster test for proliferation gave P = 0.093, the AR association was not significant, and the cribriform-like channel did not robustly colocalize with proliferation across four gated patients (mean patient-level correlation, 0.08; P = 0.32). The in-situ evidence is therefore consistent with the bulk associations where the score localizes to epithelium, and the small number of patients precludes a cross-patient in-situ claim in either direction.

Robustness depends on pathology-specific representation rather than one foundation encoder

The prognostic stratification of the architecture score did not depend on the foundation model tissue encoder used to compute it. Because each encoder defines a different embedding space, we refit the Gleason pattern classifier on SICAPv2 for each alternative encoder, with no survival supervision and no exposure to either evaluation cohort. In PLCO, the top score tertile versus the lower two was associated with PCSM for Prov-GigaPath (HR = 3.36; 95% CI: 2.04-5.53), Virchow2 (4.04; 95% CI: 2.43-6.72), and UNI2-h (4.89; 95% CI: 2.89-8.26). Corresponding DFS HRs in TCGA-PRAD were 4.02 (95% CI: 2.49-6.48), 3.46 (95% CI: 2.14-5.59), and 2.89 (95% CI: 1.81-4.61; **Supplementary Fig. S6a**).

Concurrent correlations with grade were nearly identical across encoders: ρ = 0.38, 0.38, and 0.39 in PLCO and 0.53, 0.53, and 0.54 in TCGA-PRAD (**Supplementary Fig. S6b**). Within the Gleason 7 stratum in PLCO, the score remained associated with PCSM under both the biopsy definition (HR per SD = 2.07; 95% CI: 1.19-3.60; n = 282; 33 events) and the RP definition (HR = 2.48; 95% CI: 1.29-4.76; n = 402; 27 events; **Supplementary Fig. S6c**). By contrast, replacing the pathology representation with an ImageNet-pretrained ResNet50 weakened the association. The age-adjusted PCSM HR decreased from 2.57 (95%

CI: 1.73-3.84) with Prov-GigaPath to 1.38 (0.99-1.94), and the fully adjusted HR decreased from 1.88 (95% CI: 1.28-2.77) to 1.17 (95% CI: 0.85-1.61; **Supplementary Fig. S6d**). Thus, the association was robust across pathology foundation encoders but was not recovered by a generic natural image encoder.

The score identifies a candidate higher-risk subgroup within low biopsy grade

We compared the fixed histomorphologic model with an attention-based multiple-instance Cox model trained directly on the survival outcome of one cohort and evaluated in the other, using the same Prov-GigaPath tile embeddings (**Supplementary Fig. S6e**). When trained on TCGA-PRAD and evaluated in PLCO, the supervised model had a deployment C-index of 0.648 versus 0.725 for the histomorphologic model. In the reverse direction, the values were 0.620 and 0.700. The supervised model achieved in-sample C-indices of 0.892 and 0.964 in its training cohorts; these values are optimistic and not directly comparable with held-out deployment. These results show that the histomorphologic model achieved higher cross-cohort C-indices than the untuned outcome-supervised comparator in both evaluated transfer directions.

Among 706 men with biopsy Gleason ≤6, a prespecified within-stratum split placed 11 of 17 PCa deaths in the high score tertile. Twenty-year PCSM cumulative incidence was 4.20% (95% CI: 1.85-7.18) in the high tertile, 0.45% in the middle tertile, and 0.96% in the low tertile. The high-versus-low difference in restricted mean time lost to PCa death was 0.18 years (95% CI: 0.03-0.35), and remained 0.19 years (95% CI: 0.02-0.43) when tertiles were recomputed within biopsy Gleason 6 alone. This finding is statistically unstable since the cumulative-incidence intervals overlapped, and the continuous cause-specific HR was 1.96 (95% CI: 0.99-3.87; **Supplementary Fig. S6f**). Expanding the stratum to biopsy Gleason ≤7 (n = 988; 50 PCSM events) yielded a high-tertile 20-year incidence of 8.03% and an HR per SD of 2.44 (1.57-3.81).

DISCUSSION

Our histomorphologic model, incorporating a biopsy-trained Gleason-pattern classifier, generated architecture scores from RP histology associated with PCSM in PLCO and with DFS in TCGA-PRAD. In PLCO, the association persisted after clinicopathological adjustment and was absent for non-PCa death. The score decomposed into high-grade and cribriform-like components, and its molecular correlates were dominated by proliferation and genomic instability. These results support adverse-architecture burden as a measurable histological phenotype, while leaving its incremental clinical value unresolved.

Neither the Gleason-pattern classifier nor patient-level aggregation was fitted to survival events, so the architecture score's association with PCa death in PLCO was evaluated separately from training of the histomorphologic model. Previous work has shown that outcome-supervised, multimodal, or outcome-derived pathology models can stratify PCSM, recurrence, and treatment benefit[9,10,14–18]. The present model addresses a different question, which is whether scores derived from Gleason architecture are associated with later clinical outcome. In a cross-cohort comparison, the histomorphologic model

transferred to each cohort better than an outcome-supervised model built from the same image representation, which discriminated far less outside the cohort in which it was trained. This is evidence for the outcome-blind design rather than a benchmark result, since a model fitted to outcomes in one cohort reflects that cohort's composition whereas a score fixed on architecture does not.

The outcome-blind design leaves open whether the association reflects the tumor architecture the score measures or a general burden of illness that accompanies it. PLCO allowed disease specificity to be tested within the same patients. The score was associated with PCa death but not with pooled non-PCa death, cardiovascular death, or death from another cancer, which argues against general frailty as the principal explanation. The negative-control analysis is nevertheless conditional on the competing endpoint sharing the relevant unmeasured causes of bias without being caused by tumor architecture[24]. Under that assumption, the tight null for non-PCa death confined shared frailty or comorbidity confounding to a multiplicative factor of about 0.93 to 1.12 on the HR scale, leaving the fully adjusted PCSM HR essentially unchanged after calibration. The E-value[25] for the fully adjusted HR of 1.88 was 3.17, with 1.87 for its lower confidence bound. Recorded primary treatment and additional demographic and clinical covariates did not change the association significantly, but unrecorded salvage radiotherapy or androgen-deprivation therapy (ADT) could still confound postoperative outcomes.

The PCSM association reflects specific adverse architectures and not the amount of tumor on the slide. The multilabel Gleason-pattern classifier estimates separate probabilities for adverse growth patterns, allowing high-grade and cribriform-like sub-scores to be evaluated alongside the patient-level architecture score. In the joint age-adjusted model, pattern 4/5 burden and the cribriform-like channel were independently associated with PCSM, while tumor fraction and tile-level heterogeneity were not. Cribriform growth is already recognized as adverse, with Russo et al.[4] reporting that cribriform growth is associated with biochemical recurrence and cancer-specific mortality, and Cornford et al.[7] recommend excluding cribriform or intraductal histology from active surveillance at baseline. Our results extend that clinical referent from a binary pathological feature to a continuous model-derived burden.

The molecular analyses were consistent with established biology of adverse prostate architecture. In TCGA-PRAD, the score and its pattern 4/cribriform-like components correlated with E2F, G2/M, MYC, DNA-repair, and PTEN-loss programs, while pattern 5 content was associated with inflammatory, angiogenic, EMT, and androgen-low programs. These are cross-sectional associations in bulk tissue and do not establish causal mechanisms. The bulk associations were consistent in situ, where the score localized to tumor epithelium, but a spatial multi-slide analysis was underpowered across the few available patients.

The architecture score correlated weakly and negatively with stromal fraction and decreased as stromal abundance rose, indicating that stromal content is not its main determinant. Additionally, while the encoder captures site information, the DFS association still held under stratification by tissue-source site and across refits that excluded each site in turn, indicating that tissue-source site does not account for it. We found that the score is grade-associated but is not a re-derived grade. Its within-grade spread was

broad, its median plateaued across the highest grades, and it remained associated with PCSM within Gleason 7. These findings are compatible with information carried by continuous pattern 4 burden and architectural subtype[1,2,5].Mediation analysis[26] attributed 14% of the cause-specific association to a continuous machine-derived grade and 17% to biopsy Gleason grade. The expected increase in the mediated proportion with a finer grade measure did not occur, and a nonzero direct effect on the hazard does not by itself establish incremental discrimination. Nevertheless, GG remained the dominant independent predictor in the adjusted model, consistent with the durability of Gleason grading.

Several robustness analyses strengthen the reported architecture-associated prognostic findings. Top-tertile stratification remained after the Gleason classifier was retrained over three different pathology foundation encoders, whereas a generic ImageNet representation did not recover a statistically supported association. The TCGA DFS association also persisted after stratification by tissue-source site and in every leave-one-site-out analysis. These results show that the adverse-architecture definition is not specific to one pathology foundation model.

The study is limited mainly by endpoint count and tissue sampling. PLCO contained 66 PCa deaths, which was insufficient to resolve incremental discrimination over grade and made the biopsy Gleason ≤6 analysis especially fragile. Each PLCO patient was represented by serial sections from one RP block rather than by spatially separated tumor regions. Between-core reliability was modest, although a single-core analysis remained associated with PCSM in 99% of resamples. Similarly, TCGA whole-section downsampling reduced both agreement and the DFS association. Whole-section and multiblock validation is therefore needed to test whether a core-derived burden represents the specimen, which the TMA substrate cannot establish. In TCGA-PRAD, the constraint was the scale instead, since the score approached its ceiling there and under the alternative encoders, leaving tertile contrasts as the comparison, and those contrasts are not comparable across cohorts because PCSM and DFS are not interchangeable. Both cohorts were scored by a classifier trained on SICAPv2, which is small and single-source and discriminates the intermediate grades less strongly than extreme grades. External validation on multi-institutional biopsy and prostatectomy whole sections, with standardized pathologist annotation of cribriform and intraductal growth and complete salvage-treatment data, is required. Finally, salvage radiotherapy and ADT after RP were not reported in either cohort, which prevents testing the confounding effect of post-surgical treatment on the score's association with PCSM.

In summary, the outcome-blind histomorphologic model generated architecture scores specifically associated with PCa mortality, with associations also observed for high-grade and cribriform-like sub-scores. Especially important is our finding that the score identified a subgroup with higher long-term risk of PCa death among men whom biopsy grade would place in a low-risk category. This finding motivates future work to determine whether the score improves risk assessment in larger, independent biopsy cohorts with adequate number of PCa mortality events.

METHODS

Ethics and data governance

This retrospective study used de-identified histopathology and clinical data from public or controlled-access sources. SICAPv2 was released under a Creative Commons Attribution 4.0 license. Access to the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial RP images and linked clinical records was obtained through the National Cancer Institute (NCI) Cancer Data Access System (CDAS) under approved project PLCOI-1803, and all PLCO records were handled as de-identified controlled-access data under the terms of that data use agreement. TCGA-PRAD slides and clinical annotations were obtained from the Genomic Data Commons and cBioPortal. All datasets were de-identified at source, no protected health information was accessed, and no new consent was required under the governing data-use terms.

Study cohorts and endpoints

For model training and internal validation, we used SICAPv2, which comprises 155 H&E WSIs from prostate needle biopsies of 95 patients; 119 slides contain cancer. The release includes 18,783 512-pixel patches with slide-level grade labels, 12,081 patches with tile-level Gleason-architecture labels, and a patient-disjoint cross-validation partition. A separate set of 4,494 patches from 63 patients (763 cribriform-positive; 17.0%) was used only for the standalone cribriform classifier.

PLCO was used as the main external evaluation cohort for PCSM. The image-linked source contained 2,999 RP TMA slides from 1,095 men. Each patient contributed 1-6 cores derived from serial sections of a single tissue block. Thus, between-core analyses quantify within-block repeatability rather than sampling across the tumor. Of 2,999 slides, 2,997 were scanned on one Aperio instrument (NCI-S096017) at 40× with an x-axis resolution of 0.2532 μm per pixel (MPP) and a native 10× pyramid level. One slide lacked that level and one was unreadable.

The evaluable PLCO cohort included 1,086 men, 66 PCa deaths, and 480 competing deaths, with median follow-up of 22.6 years. Follow-up extended from enrollment to death or censoring. PCSM was analyzed with death from any other cause as the competing event. TCGA-PRAD was an independent external cohort of RP whole sections and included 394 patients with 72 DFS events. DFS was the reported composite of recurrence or progression. TCGA-PRAD reports RP grade but not biopsy grade; only postsurgical PSA was available, so PSA was not included in TCGA models. For every survival model, the patient-level score was standardized to zero mean and unit SD within each cohort after excluding patients with an undefined score, so all HRs are per 1 SD.

Image preprocessing and quality control

PLCO and TCGA-PRAD whole-slide images were read with OpenSlide at the pyramid level nearest 10× and partitioned into non-overlapping 512-pixel tiles. At the PLCO resolution of 0.2532 MPP, each tile represented a 518.6-μm field. TCGA fields ranged from approximately 464 to 518 μm because source magnification varied across sites. SICAPv2 provides 512-pixel patches at the same nominal 10×

geometry. Tiles were retained when at least 25% of pixels were classified as tissue by Otsu thresholding of saturation and were Macenko-normalized to a fixed target[27].

Each normalized tile was resized to 224 pixels, center-cropped, and encoded with the fixed Prov-GigaPath tile encoder[20], producing a 1,536-dimensional embedding. The larger source field was used consistently in training and evaluation to preserve gland-level context. Artifact exclusion combined two masks. The first was the GrandQC tissue-and-artifact segmentation model38, run with its multi-class model in an occlusion gate so that a tile was dropped when its center pixel fell on any GrandQC artifact class, namely tissue fold, dark spot, pen mark, bubble or slide edge, out-of-focus region, or background, and was retained only when its center pixel was segmented as clean tissue. The second was an embedding-norm filter that removed the lowest 25% of tiles per slide. Low-content fragments that passed the initial tissue threshold had low embedding norms and were otherwise assigned spuriously high adverse probabilities (artifact-norm Pearson $r \approx -0.66$). The 25% threshold was selected at the knee of a grade-concordance sweep (Spearman $\rho = 0.30$, 0.34, and 0.38 after dropping 0%, 10%, and 25%, with little further change at 30%). The GrandQC and norm masks were intersected, and the percentile was calculated over all slide tiles.

Tile-level multilabel Gleason classifier

The histomorphologic model used a multi-label Gleason pattern classifier with fixed 1,536-dimensional embeddings. It comprised layer normalization, a 256-unit linear layer with rectified-linear activation, dropout of 0.3, and a five-output linear layer with independent sigmoid activations for non-cancer, Gleason pattern 3, pattern 4, pattern 5, and cribriform pattern 4. Independent sigmoids allowed a tile to be positive for both pattern 4 and its cribriform subtype. The classifier processed each tile independently and did not contain any attention or learned slide-level pooling.

This classifier was trained on SICAPv2 tile labels with binary cross-entropy, class-specific positive weights capped at 10 to offset class imbalance, AdamW optimization (learning rate 1e-4; weight decay 1e-4), a batch size of 256, and up to 100 epochs. Early stopping used mean validation AUC with patience of 12 epochs. Discrimination was assessed by patient-grouped five-fold cross-validation using GroupKFold, with a held-out patient-grouped validation split for early stopping, producing patient-grouped out-of-fold predictions across 12,081 architecture-labeled tiles from 95 patients. Per-class AUC intervals were estimated by 2,000 patient-level bootstrap samples. After training, the classifier weights were frozen before application to PLCO or TCGA-PRAD.

Slide-level grade classifier and separate cribriform detector

A separate slide-level GG classifier was trained on the same SICAPv2 embeddings using attention-based multiple-instance learning (MIL)[28]. Tile embeddings were projected to 512 units with rectified-linear activation and dropout of 0.25, pooled through a 256-dimensional gated-attention layer, and classified across benign tissue and GG1-GG5. It was evaluated by nested patient-grouped five-fold cross-validation with patient-level bootstrap intervals. This model did not contribute to the architecture score.

A separate binary cribriform detector used the same layer-normalized 256-unit architecture as the tile classifier and was trained on the 4,494-patch cribriform set using class-weighted binary cross-entropy. Repeated patient-grouped stratified five-fold cross-validation yielded an AUC of 0.859 (SD = 0.057). This detector was not used for scoring. The term cribriform-like refers to the cribriform channel of the multi-label Gleason classifier, which has not been tested in an external cribriform-annotated cohort.

Patient-level architecture score

For each tile, the adverse score was max[P(pattern 4), P(pattern 5), P(cribriform pattern 4)]. A tile was classified as tumor-containing when P(non-cancer) ≤0.50, so that benign tissue, which varies in quantity across microarray cores, could neither dilute focal adverse architecture nor inflate the denominator. Tumor-tile scores were pooled across all slides for a patient, and the case score was the mean of the highest 10%, using the ceiling of 0.10 × the tumor-tile count and at least one tile. Patients with no tumor tiles had undefined scores and were excluded. The 10% fraction and 0.50 tumor threshold were specified before survival analysis. Alternative summaries included the top 1%, 5%, and 20%, the fraction above 0.50, and a maximum-region rule. Pattern 4/5, cribriform-like, tumor-fraction, and tile-heterogeneity components were computed from the same fixed tile outputs.

External evaluation and encoder robustness

PLCO and TCGA-PRAD were analyzed after the Prov-GigaPath encoder, Gleason pattern classifier weights, and patient-level aggregation had been fixed. The same image geometry, stain target, artifact masks, and aggregation were used in both cohorts. All reported survival models used one patient-level score per cohort.

To test whether the prognostic stratification depends on the specific foundation encoder, the entire feature pipeline was rebuilt using Virchow2[29] and UNI2-h[30] and compared with Prov-GigaPath[20]. Virchow2 embeddings concatenated the class token and mean patch-token representation (2,560 dimensions)[29]; UNI2-h produced a 1,536-dimensional representation[30]. Because the embedding spaces differ, the tile-level Gleason classifier was trained separately on SICAPv2 for each encoder, without survival labels or evaluation-cohort data. The cross-encoder comparison used the top-tertile contrast because Virchow2 and UNI2-h compressed the per-SD probability scale. A generic ImageNet-pretrained ResNet50 was evaluated with the same training, quality-control, aggregation, and survival procedures.

Outcome-supervised cross-cohort comparator

To test whether the fixed histomorphologic model transfers across cohorts as well as a model trained directly on outcome, we built an outcome-supervised comparator from the same image representation. The outcome-supervised model was an image-only gated-attention Cox model that used the same fixed Prov-GigaPath tile embeddings as the histomorphologic model. A 512-unit tile projection and 256-dimensional attention layer were followed by a scalar Cox log-risk output. Training minimized the Breslow negative partial log-likelihood, with AdamW (learning rate and weight decay of 1e-4), dropout of

0.25, a fixed 50 epochs, and a 400,000-tile batch budget. The model was fit once on all patients in one cohort and evaluated in the other, in both directions, namely trained on TCGA-PRAD DFS and deployed to PLCO PCSM, and trained on PLCO PCSM and deployed to TCGA-PRAD DFS. On each held-out cohort we computed the Harrell concordance index (C-index) of the supervised risk and, for comparison, of the patient-level architecture score.

Molecular and spatial correlates

TCGA-PRAD RNA-seq data were analyzed on the $\log_2$ scale. Genomic instability was measured by the cBioPortal FGA variable. Partial Spearman correlations related the architecture score to proliferation and FGA while sequentially adjusting for tumor purity, RP grade, and stromal fraction. Genes were ranked by Spearman correlation with the score or with each adverse pattern channel. Preranked GSEA used the 50 MSigDB hallmark gene sets and BH q values[31]. Stromal fraction was derived from precomputed HoVer-Net nuclear segmentations[32].

To assess site confounding, we used a multinomial logistic regression to predict TCGA tissue-source site from mean tumor-tile embeddings under stratified five-fold cross-validation. The DFS association was then refit with tissue-source-site strata and after exclusion of each site in turn. Spatial analyses used 12 PCa slides representing 6 patients, including serial sections from a 3D prostatectomy reconstruction. Spot-level score associations with proliferation, AR activity, and stroma were evaluated against an autocorrelation-preserving spatial null. Serial sections were collapsed to the biological unit before cross-patient tests. The three adverse-pattern channels were evaluated separately for colocalization with proliferation.

Statistical analysis

The primary PLCO analysis used an age-adjusted cause-specific Cox proportional-hazards model for PCSM, with death from other causes treated as a competing event and censored, estimated by unpenalized maximum likelihood. Disease specificity was assessed by applying the same model to non-PCa death and by contrasting event-specific coefficients in a stacked Lunn-McNeil model[33]. A Fine-Gray subdistribution model provided an absolute-risk sensitivity analysis[34]. PCSM cumulative incidence was estimated by Aalen-Johansen and compared across tertiles with Gray's test. All tests were two-sided with $\alpha = 0.05$ unless an FDR threshold was specified. The pretreatment PLCO model included age, log diagnostic PSA, biopsy Gleason grade, and clinical stage. A postsurgical sensitivity model substituted RP Gleason grade and pathologic stage. Incremental discrimination was the difference in out-of-fold Harrell C-index between the clinical model and the same model plus the score, estimated by patient-grouped five-fold cross-validation with patient-level bootstrap CIs. Restricted cubic splines tested nonlinearity. The four component features were entered jointly with age; profile-likelihood intervals were used for component HRs. TCGA DFS was analyzed by Cox regression and 1 minus Kaplan-Meier, with a multigroup log-rank test for tertiles.

The proportional-hazards assumption was assessed for every reported Cox model with scaled Schoenfeld residuals and the Grambsch-Therneau test across Kaplan-Meier, rank, identity, and log time transforms[35]. The age-adjusted PLCO model showed a global violation; early and late HRs were therefore reported after a prespecified split at the median PCSM event time. The TCGA DFS model and PLCO postsurgical model did not show evidence of violation for the score. Wald intervals were used unless otherwise specified. The main PCSM HR was also evaluated in 1,000 patient bootstrap samples, and classifier AUCs in 2,000 patient-grouped bootstrap samples.

The age-adjusted cause-specific PCSM model was the single confirmatory analysis. Component, aggregation, spline, subgroup, mediation, and sensitivity analyses were supportive and reported with nominal P values. BH correction was applied to gene-level and gene-set analyses. Recorded competing causes were analyzed separately only when event counts supported the fitted model; cardiovascular death and death from a non-PCa met this criterion. Non-PCa death was also used as a negative-control outcome, and E-values were calculated for the fully adjusted association[24,25]. The score HR was refit after adding recorded primary treatment and, separately, family history, body-mass index, and self-reported race.

Causal mediation through grade used a rare-outcome survival approximation with a continuous machine-derived grade and biopsy Gleason as mediators; patient-clustered bootstrap intervals were reported.32 Within biopsy Gleason ≤6, and secondarily ≤7, score tertiles were formed within stratum and PCSM cumulative incidence was estimated with patient-level bootstrap intervals. Restricted mean time lost to PCa death was the area under the cause-specific cumulative-incidence curve[36]. Out-of-fold decision curves compared the clinical model with and without the score at 15 and 20 years, with paired bootstrap intervals and a shrinkage sensitivity analysis[37]. Between-core reliability was measured with ICC(1,1) and cluster bootstrap intervals. Core sampling sensitivity used one random core per patient and leave-one-core-out analyses. TCGA whole-section sensitivity sampled fixed tile fractions and compared each sampled score with the full section. A prespecified log-odds aggregate was the mean logit of tumor-tile adverse probabilities; its DFS association and spline shape were analyzed on the standardized scale.

Declarations

DATA AVAILABILITY

The SICAPv2 prostate needle-biopsy dataset analyzed in this study is a public deposit released by Silva-Rodríguez and colleagues under a Creative Commons Attribution 4.0 (CC BY 4.0) license and is available from Mendeley Data (https://doi.org/10.17632/9xxm58dvs3.1). The Cancer Genome Atlas prostate adenocarcinoma (TCGA-PRAD) diagnostic slides and clinical annotations are open-access, de-identified data available through the National Cancer Institute (NCI) Genomic Data Commons (https://portal.gdc.cancer.gov) and cBioPortal (https://www.cbioportal.org). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial radical prostatectomy images and linked clinical

records are controlled-access data obtained from the NCI Cancer Data Access System (CDAS; https://cdas.cancer.gov/plco/) under approved project PLCOI-1803 and a data-use agreement, and they cannot be redistributed by the authors; qualified investigators may apply for access through CDAS.

CODE AVAILABILITY

The analysis code supporting the findings of this study is available on GitHub at https://github.com/TheOmarLab/he-architecture-score-prostate. The frozen Gleason-pattern classifier weights and fixed aggregation rule for calculating architecture scores will be deposited in Zenodo with a permanent DOI upon acceptance.

ACKNOWLEDGEMENTS

This work was supported by the Department of Defense grant HT9425-24-1-0381 (to M.O). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

AUTHOR CONTRIBUTIONS

M.O. conceived and supervised the project. A.L. and M.O. developed the methodology and software. A.L. performed the analyses and visualization. M.O. and A.L. wrote the original draft. P.V., M.R.A., and M. Osrodek. edited the draft. All authors have read and approved the submitted manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

References

1. Sauter, G. *et al.* Clinical Utility of Quantitative Gleason Grading in Prostate Biopsies and Prostatectomy Specimens. *Eur Urol* **69**, 592–598 (2016).
2. Seyrek, N. *et al.* Alternative prostate cancer grading systems incorporating percent pattern 4/5 (IQ-Gleason) and cribriform architecture (cGrade) improve prediction of outcome after radical prostatectomy. *Virchows Arch* **480**, 1149–1157 (2022).
3. The Absolute Volume of Gleason Pattern 4 on Radical Prostatectomy Is More Strongly Associated With Advanced Stage and Biochemical Recurrence Than Gleason Grade Groups | Journal of Urology. *The Journal of Urology* https://www.auajournals.org/doi/10.1097/JU.0000000000004484 (2025).
4. Russo, G. I. *et al.* Oncological outcomes of cribriform histology pattern in prostate cancer patients: a systematic review and meta-analysis. *Prostate Cancer Prostatic Dis* **26**, 646–654 (2023).
5. Kweldam, C. F. *et al.* Cribriform growth is highly predictive for postoperative metastasis and disease-specific death in Gleason score 7 prostate cancer. *Mod Pathol* **28**, 457–464 (2015).

6. Böttcher, R. *et al.* Cribriform and intraductal prostate cancer are associated with increased genomic instability and distinct genomic alterations. *BMC Cancer* **18**, 8 (2018).
7. Cornford, P. *et al.* EAU-EANM-ESTRO-ESUR-ISUP-SIOG Guidelines on Prostate Cancer-2024 Update. Part I: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol* **86**, 148–163 (2024).
8. Wulczyn, E. *et al.* Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading. *Commun Med (Lond)* **1**, 10 (2021).
9. Li, E. V. *et al.* An Artificial Intelligence-Digital Pathology Algorithm Predicts Survival After Radical Prostatectomy From the Prostate, Lung, Colorectal, and Ovarian Cancer Trial. *J Urol* **213**, 600–608 (2025).
10. Eminaga, O. *et al.* Artificial intelligence unravels interpretable malignancy grades of prostate cancer on histology images. *npj Imaging* **2**, 6 (2024).
11. Fernandez-Mateos, J. *et al.* Tumor evolution metrics predict recurrence beyond 10 years in locally advanced prostate cancer. *Nat Cancer* **5**, 1334–1351 (2024).
12. Ambrosini, P. *et al.* Automated detection of cribriform growth patterns in prostate histology images. *Sci Rep* **10**, 14904 (2020).
13. Szolnoky, K. *et al.* Finding Holes: Pathologist-Level Performance Using AI for Cribriform Morphology Detection in Prostate Cancer. *European Urology Open Science* **87**, 31–39 (2026).
14. Esteva, A. *et al.* Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *npj Digit. Med.* **5**, 71 (2022).
15. Ross, A. E. *et al.* External Validation of a Digital Pathology-based Multimodal Artificial Intelligence Architecture in the NRG/RTOG 9902 Phase 3 Trial. *Eur Urol Oncol* **7**, 1024–1033 (2024).
16. Tward, J. D. *et al.* Prostate Cancer Risk Stratification in NRG Oncology Phase III Randomized Trials Using Multimodal Deep Learning With Digital Histopathology. *JCO Precis Oncol* e2400145 (2024) doi:10.1200/PO.24.00145.
17. Armstrong, A. J. *et al.* Development and Validation of an Artificial Intelligence Digital Pathology Biomarker to Predict Benefit of Long-Term Hormonal Therapy and Radiotherapy in Men With High-Risk Prostate Cancer Across Multiple Phase III Trials. *J Clin Oncol* **43**, 3494–3504 (2025).
18. Bjartell, A. *et al.* Validation of a Digital Pathology–Based Multimodal Artificial Intelligence Biomarker in a Prospective, Real-World Prostate Cancer Cohort Treated with Prostatectomy. *Clin Cancer Res* **31**, 1546–1553 (2025).
19. Patkar, S. *et al.* A selective CutMix approach improves generalizability of deep learning-based grading and risk assessment of prostate cancer. *Journal of Pathology Informatics* **15**, 100381 (2024).
20. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
21. Prorok, P. C. *et al.* Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials* **21**, 273S-309S (2000).

22. Andriole, G. L. *et al.* Mortality Results from a Randomized Prostate-Cancer Screening Trial. *New England Journal of Medicine* **360**, 1310–1319 (2009).
23. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).
24. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388 (2010).
25. VanderWeele, T. J. & Ding, P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med* **167**, 268–274 (2017).
26. VanderWeele, T. J. Causal mediation analysis with survival data. *Epidemiology* **22**, 582–585 (2011).
27. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107–1110 (2009). doi:10.1109/ISBI.2009.5193250.
28. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based Deep Multiple Instance Learning. Preprint at https://doi.org/10.48550/arXiv.1802.04712 (2018).
29. Zimmermann, E. *et al.* Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology. Preprint at https://doi.org/10.48550/arXiv.2408.00738 (2024).
30. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat Med* **30**, 850–862 (2024).
31. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
32. Graham, S. *et al.* Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* **58**, 101563 (2019).
33. Lunn, M. & McNeil, D. Applying Cox Regression to Competing Risks. *Biometrics* **51**, 524–532 (1995).
34. Fine, J. P. & Gray, R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* **94**, 496–509 (1999).
35. GRAMBSCH, P. M. & THERNEAU, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994).
36. Andersen, P. K. Decomposition of number of life years lost according to causes of death. *Statistics in Medicine* **32**, 5278–5285 (2013).
37. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* **26**, 565–574 (2006).

Figures

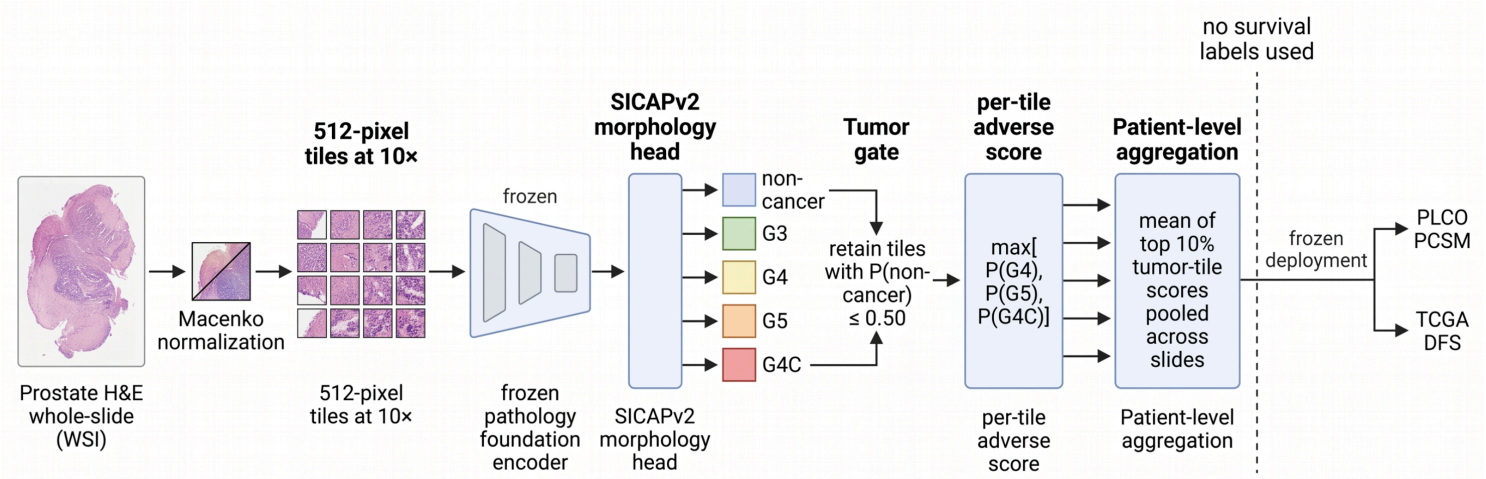

Figure 1

Development and external evaluation of the outcome-blind H&E histomorphologic model.

Hematoxylin and eosin (H&E)-stained prostate whole-slide images (WSIs) were divided into 512 × 512-pixel tiles at 10× magnification. Tiles were Macenko stain-normalized and encoded using Prov-GigaPath. A multilabel Gleason classifier trained on pathologist-annotated SICAPv2 prostate biopsy tiles produced independent probabilities for non-cancer, Gleason patterns 3, 4, and 5 (G3-G5), and cribriform pattern 4 (G4C). Tiles with P(non-cancer) ≤ 0.50 were retained and each received an adverse score. Tumor-tile scores were pooled across all slides from each patient, and the mean of the highest-scoring 10% defined the patient-level architecture score. The classifier weights and aggregation rule were applied unchanged to two radical prostatectomy cohorts: PLCO (1,086 patients) and TCGA-PRAD (394 patients). The respective endpoints were prostate cancer-specific mortality (PCSM), with other-cause death treated as a competing event, and disease-free survival (DFS).

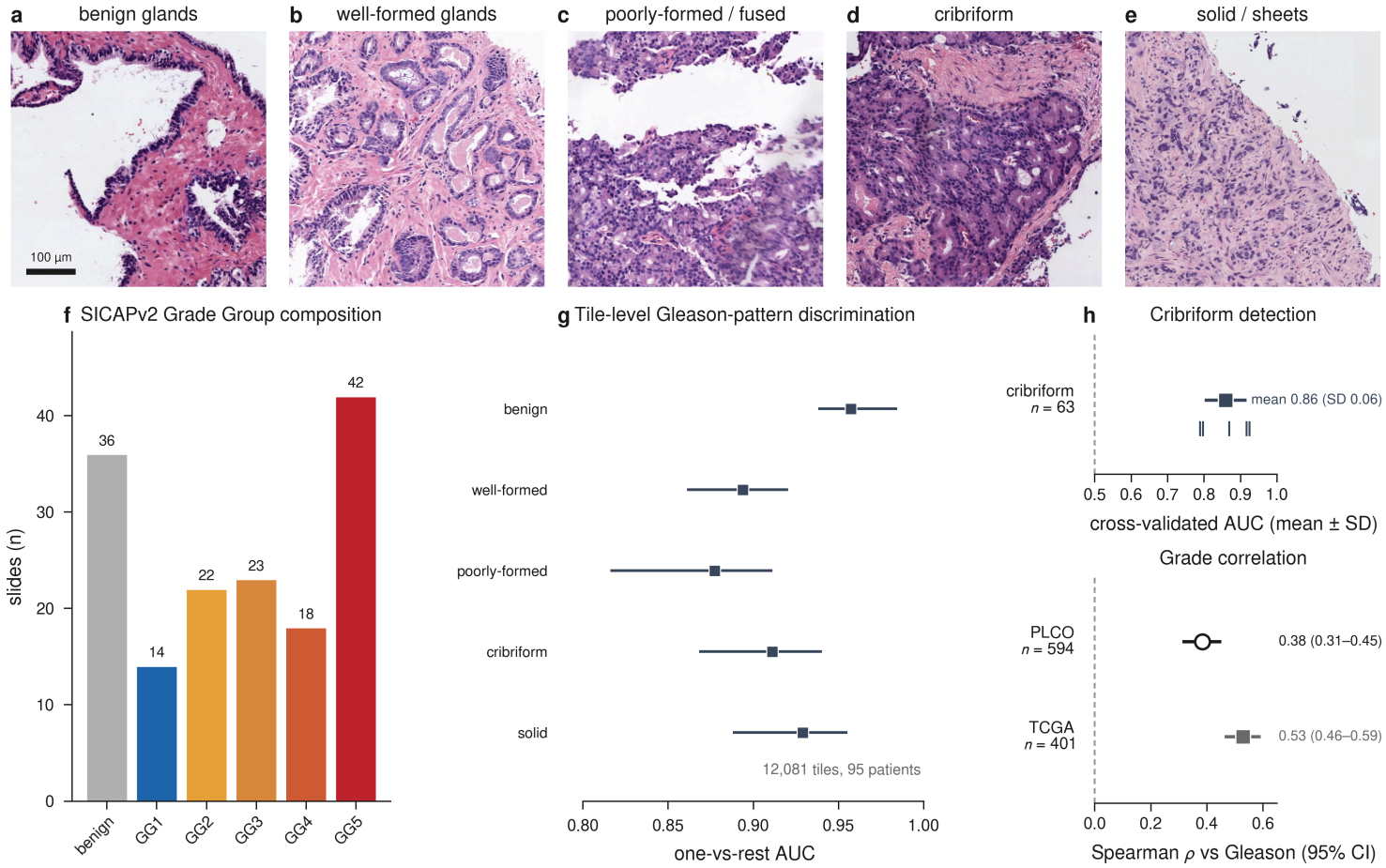

Figure 2

Development and evaluation of the Gleason-pattern classifiers on SICAPv2.

(a-e) Representative SICAPv2 tiles: (a) benign glands, (b) well-formed glands (Gleason pattern 3), (c) poorly-formed or fused glands (Gleason pattern 4), (d) cribriform growth (Gleason pattern 4 cribriform), and (e) solid sheets (Gleason pattern 5).

(f) Distribution of 155 SICAPv2 slides across benign tissue and GG1-GG5.

(g) One-versus-rest AUCs with patient-bootstrap 95% CIs for the deployed tile-level Gleason-pattern classifier, evaluated using patient-grouped out-of-fold predictions across 12,081 annotated tiles from 95 patients.

(h) Upper: cross-validated AUCs and mean ± SD for the standalone cribriform-versus-non-cribriform pattern 4 detector, which did not contribute to the architecture score. Lower: Spearman correlations with 95% CIs between the architecture score and biopsy Gleason in PLCO or radical prostatectomy Gleason in TCGA-PRAD.

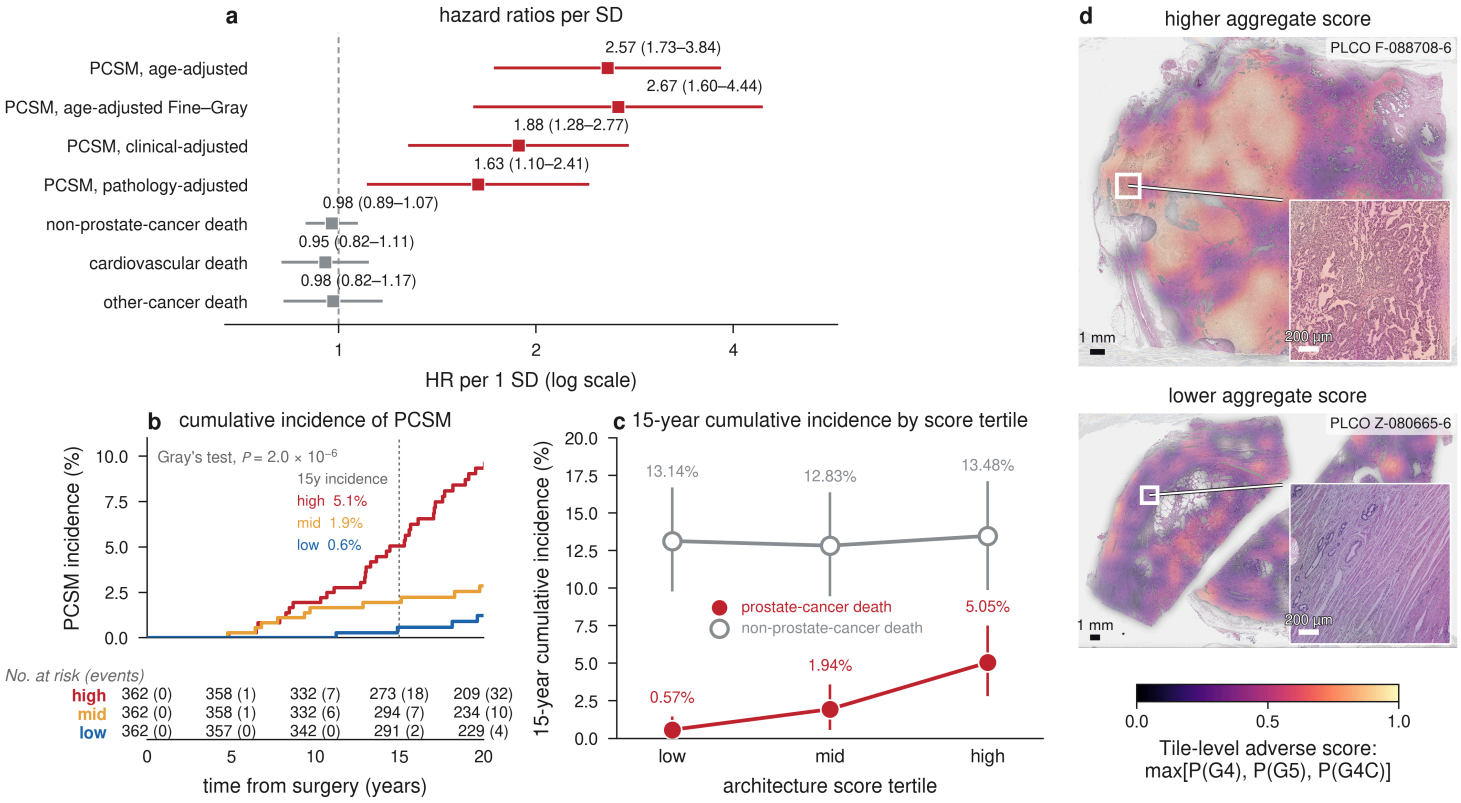

Figure 3

Association of the architecture score with disease-specific mortality in PLCO.

(a) Forest plot of the hazard ratio (HR) per 1 standard deviation (SD) increase in architecture score for prostate cancer-specific mortality (PCSM) in age-adjusted cause-specific and Fine-Gray models, in the adjusted cause-specific model, and for non-prostate-cancer death.

(b) Aalen-Johansen cumulative-incidence functions for PCSM by score tertile; P value from Gray's test.

(c) Score distribution by observed outcome.

(d) Score overlays on two PLCO slides. The color bar shows max[P(pattern 4), P(pattern 5), P(cribriform pattern 4)] from 0 to 1. Scale bars: 1 mm on whole cores and 200 µm in insets.

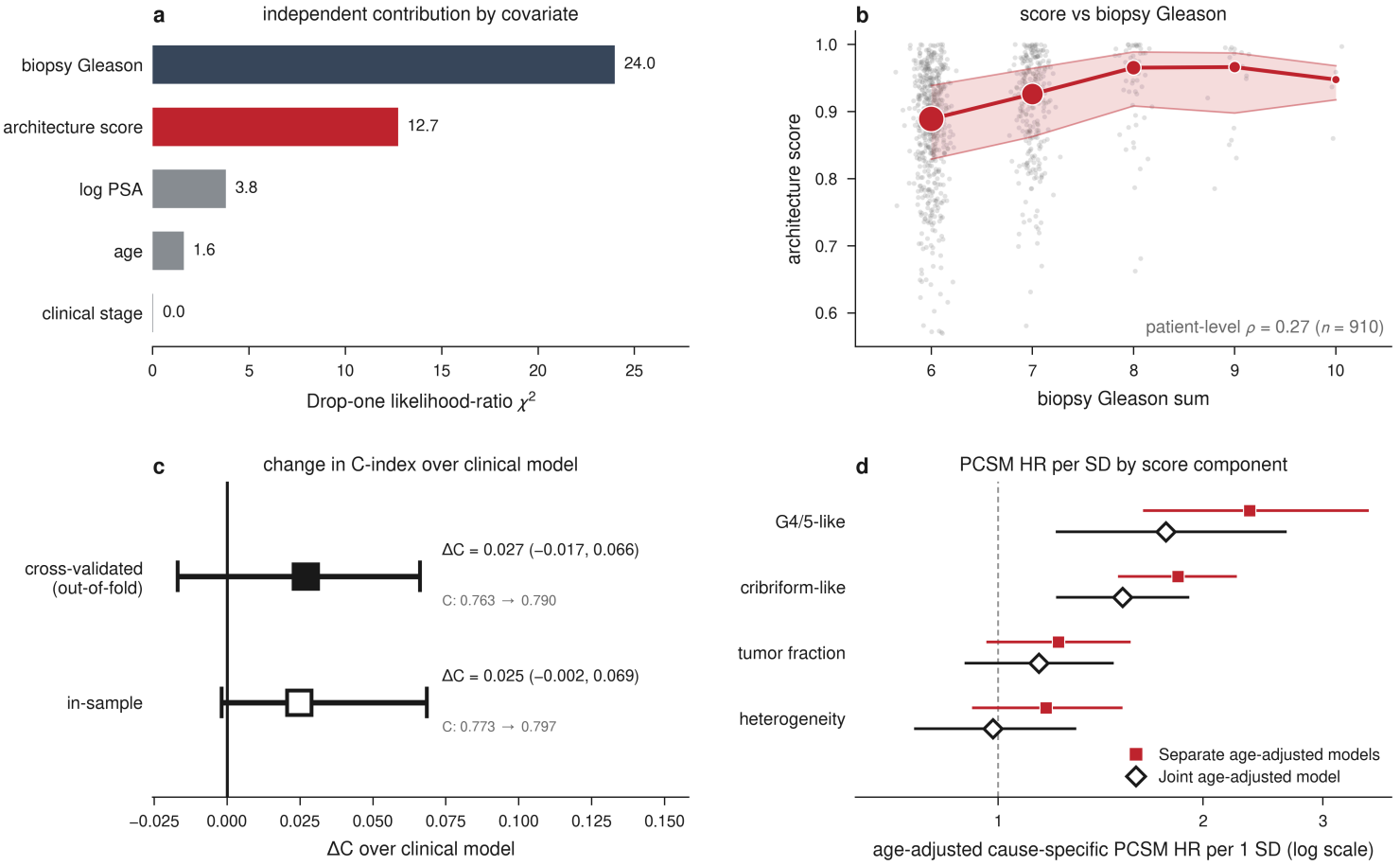

Figure 4

Grade context and incremental discrimination in PLCO.

(a) Standardized drop-one likelihood-ratio χ² for covariates in the pretreatment clinical model plus architecture score.

(b) Per-sample architecture score (jittered scatter) against biopsy Gleason sum, summarized per grade by the median score with interquartile-range bands.

(c) Delta-C (score-augmented minus clinical baseline) with the 95% bootstrap confidence interval.

(d) Harrell C-index for the pre-treatment clinical baseline versus the baseline plus architecture score, shown for both the cross-validated out-of-fold model and the in-sample model.

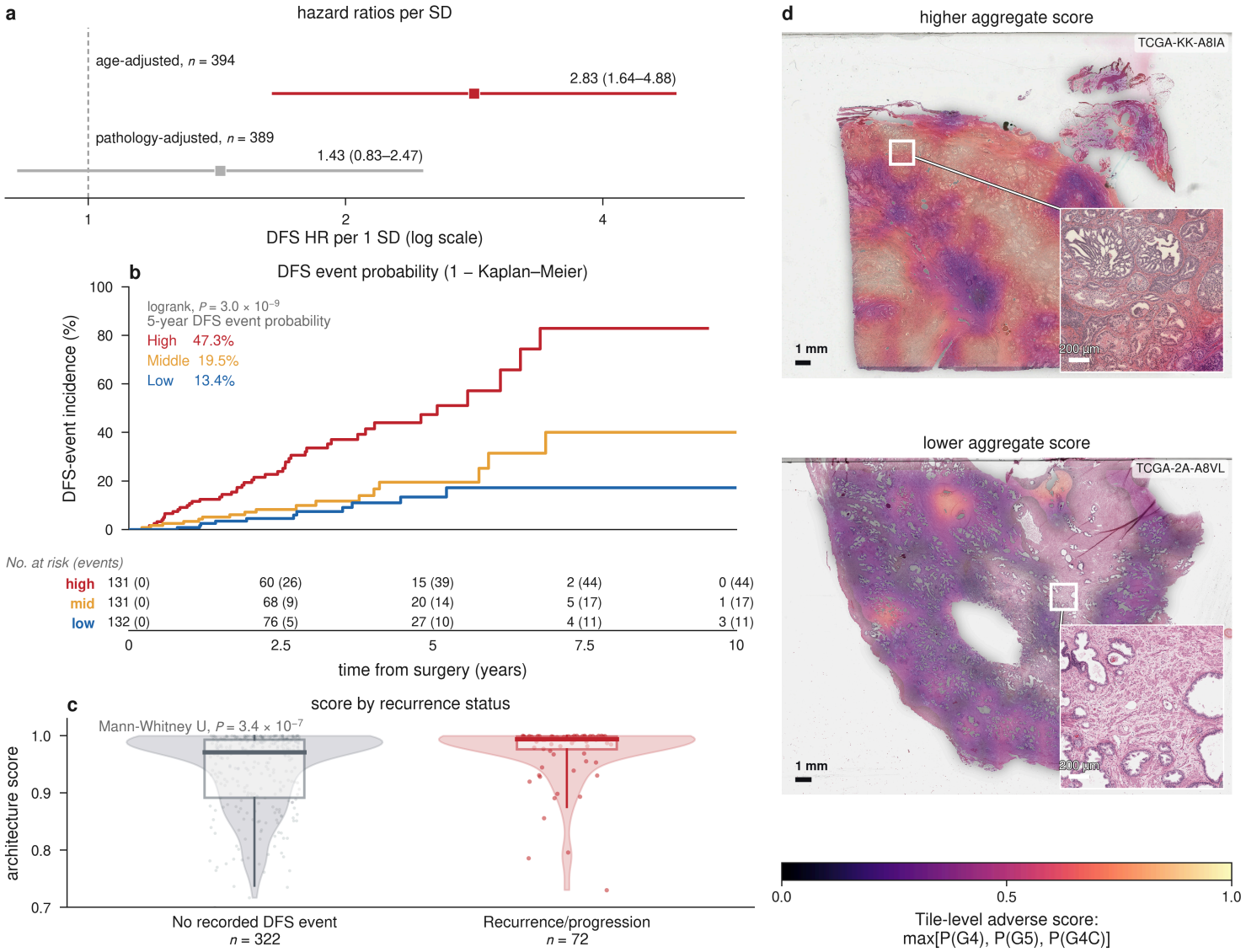

Figure 5

Association with disease-free survival in TCGA-PRAD.

(a) Forest plot of the disease-free survival (DFS) HR per 1-SD increase in score adjusted for age and pathological variables.

(b) One minus Kaplan-Meier recurrence incidence by score tertile, with numbers at risk and events; P value from the multigroup log-rank test.

(c) Score distribution by DFS status.

(d) Age-adjusted DFS HR per SD under alternative aggregation rules.

(e) Whole-slide score overlays in two TCGA-PRAD radical prostatectomy cases spanning the aggregate-score range, shown as the tile-level adverse score (max P(G4/G5/G4C)) overlaid on H&E.

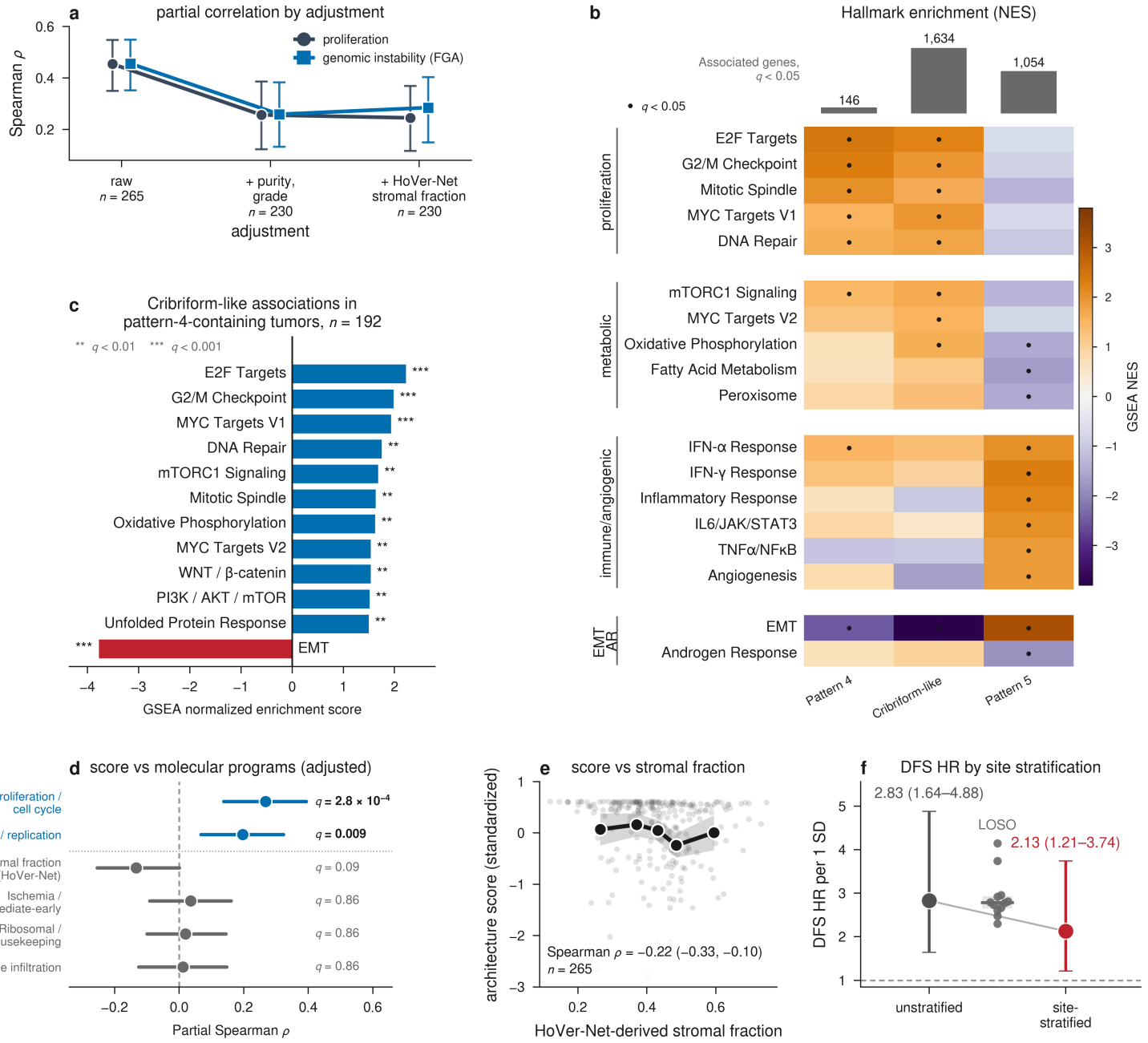

Figure 6

Molecular correlates and confound controls in TCGA-PRAD.

(a) Partial Spearman correlations between the architecture score and both proliferation and genomic instability quantified as fraction of genome altered (FGA), across the raw estimate, then partialling out tumor purity and Gleason grade, and additionally the HoVer-Net stromal fraction.

(b) Normalized enrichment score (NES) from pre-ranked MSigDB Hallmark gene set enrichment analysis (GSEA) for pattern 4, cribriform-like, and pattern 5 content within tumors containing pattern 4; dots denote q < 0.05.

(c) Significant Hallmark sets associated with cribriform-like content in the same stratum.

(d) Forest plot of the partial Spearman correlation between the architecture score and six pre-specified gene modules, each adjusted for tumor purity, Gleason grade, and HoVer-Net stromal fraction (the stromal-fraction module is adjusted for purity and grade only).

(e) Per-sample scatter of the standardized architecture score against the HoVer-Net-quantified stromal fraction (n = 265), with a binned-mean trend and 95% CI over the points.

(f) TCGA DFS HR for the architecture score, age-adjusted, unstratified versus stratified by source site.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS1.pdf
- SupplementaryFigureS2.pdf
- SupplementaryFigureS3.pdf
- SupplementaryFigureS4.pdf
- SupplementaryFigureS5.pdf
- SupplementaryFigureS6.pdf
- SupplementaryFigureS7.pdf
- SUPPLEMENTARYINFORMATION.docx