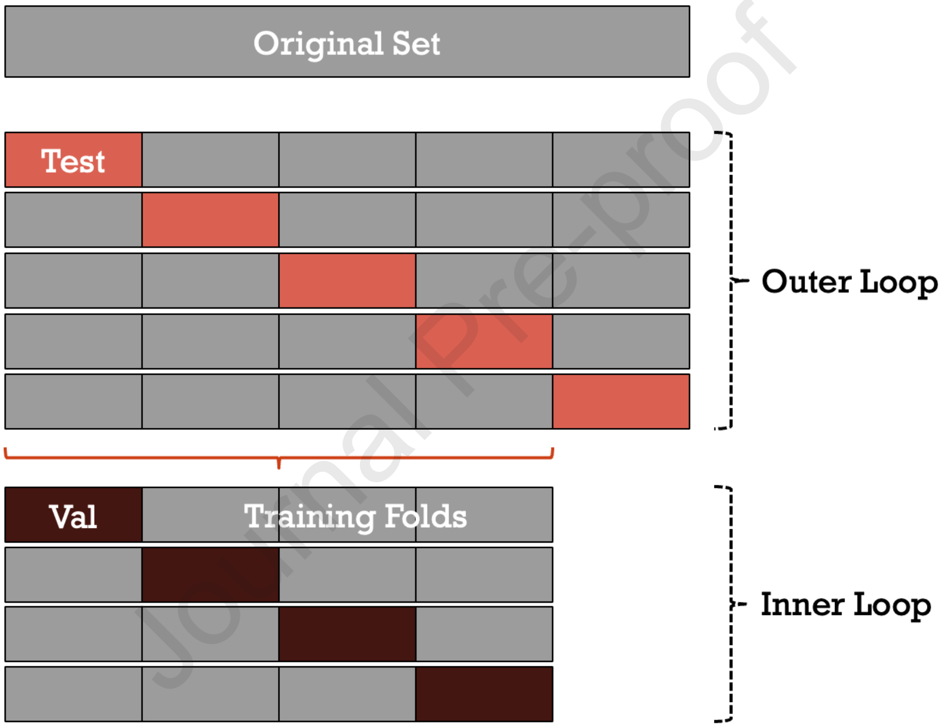
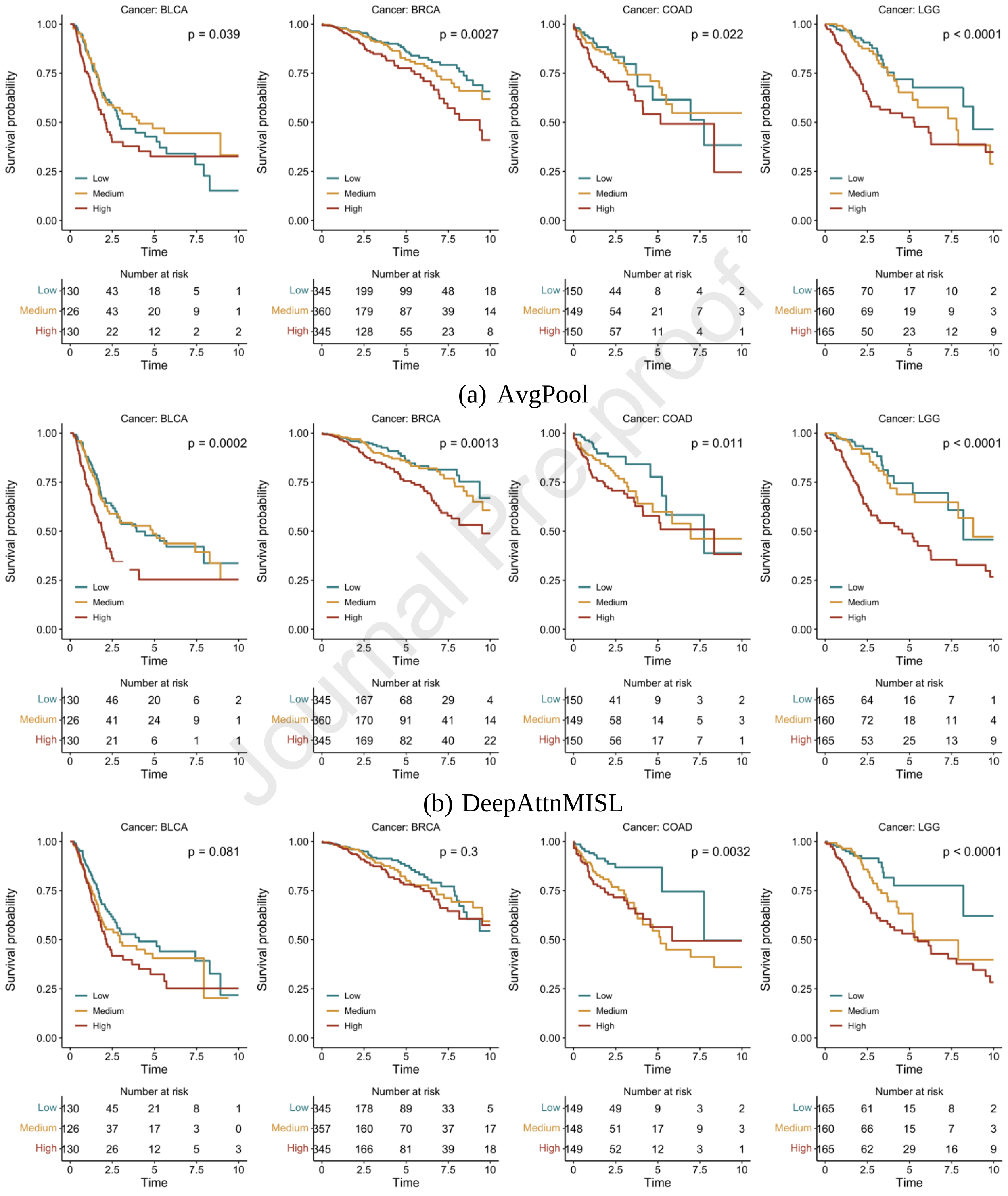
Supplementary Materials

Supplementary Figure S1. Illustration of the Nested Cross-Validation Procedure. We randomly split the original dataset into 5 folds with equal sample sizes. The splitting is stratified by patient's survival status. In the first outer loop, we use the first split as the testing dataset, and the remaining 4 splits for model training and validation. Specifically, we run a 4-fold cross-validation using these 4 splits to determine the dropout rate and early stopping epoch. Then we evaluate the selected model using the test split. We repeat the outer loop 5 times, each time using a different testing split. The entire procedure will fit 20 (i.e., 5×4) models for each hyperparameter configuration.



Journal Pre-proof

Supplementary Figure S2. Kaplan-Meier curves for the baseline methods. Patients were stratified into three risk groups based on tertiles of testing c-index.



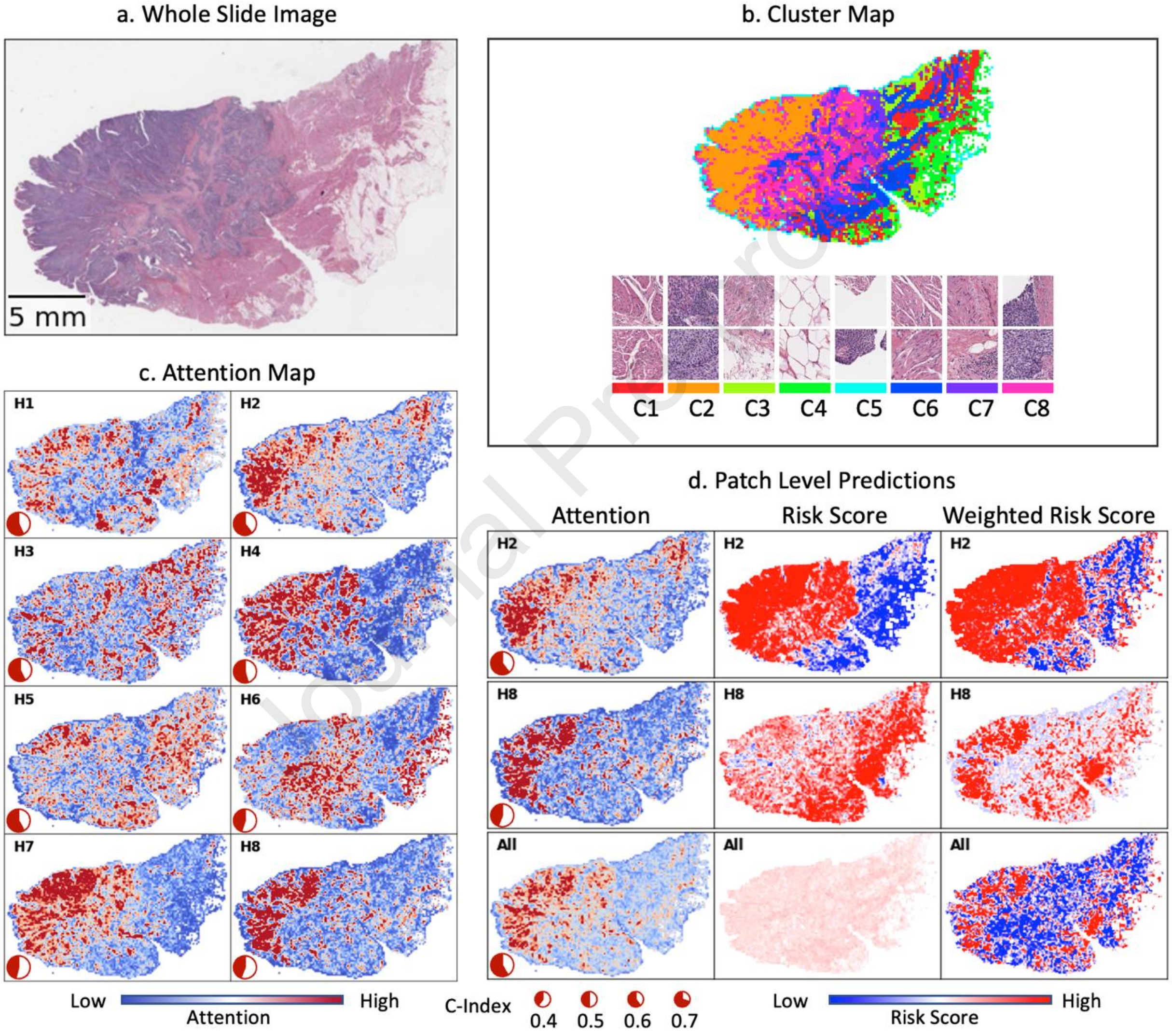
(a) AvgPool

(b) DeepAttnMISL

(c) PatchGCN

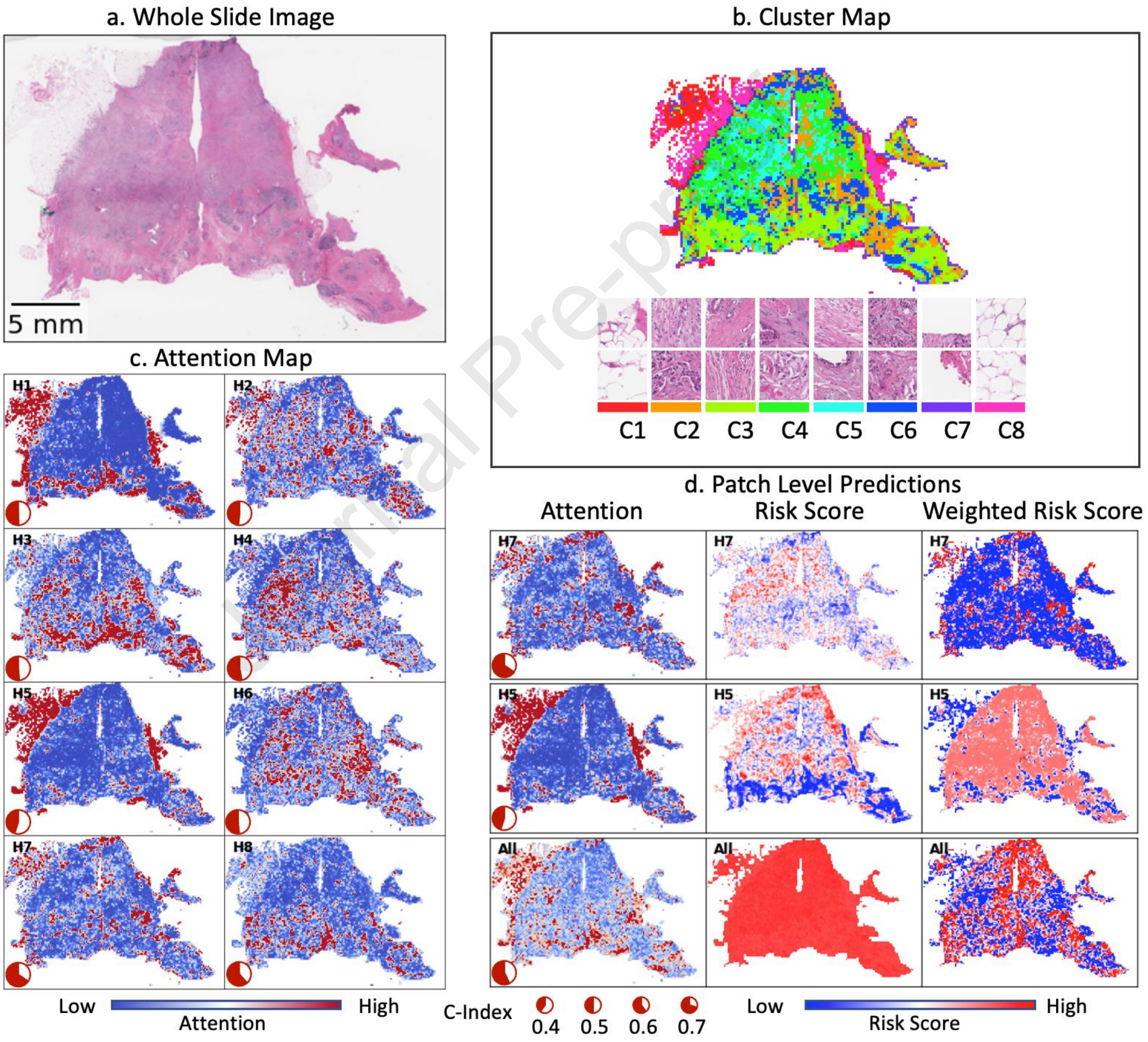
Journal Pre-proof

Supplementary Figure S3. Visualization of head-wise attention map and patch clusters for one sample WSI from BLCA. (a) Whole Slide Image. (b) Patch clusters on the WSI level and example patches from each cluster. (c) Head-wise attention map. Red color: rescaled-attention weights > 2; Blue color: rescaled-attention weights = 0. Pie plot in the lower left corner shows the head-wise c-index. (d) Patch level prediction for the selected heads. Rows: best performing head, worst performing head, and all heads combined. Columns: attention map, unscaled risk score for each patch, and weighted risk scores (i.e., attention weight × risk score). “High” and “low” risk scores refer to the maximum and minimum head-wise patient-level risk scores.



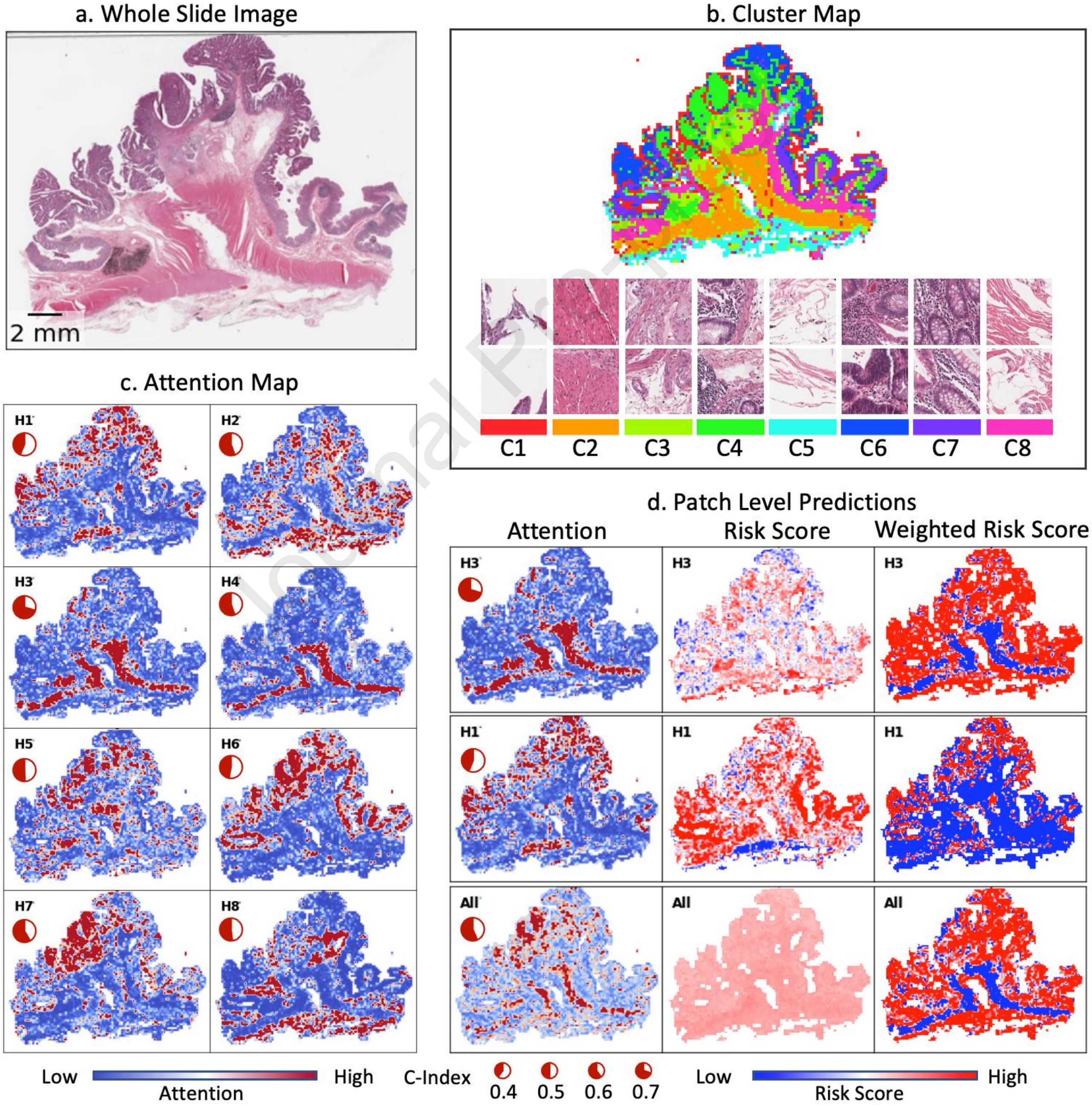
Journal Pre-proof

Supplementary Figure S4. Visualization of head-wise attention map and patch clusters for one sample WSI from BRCA. (a) Whole Slide Image. (b) Patch clusters on the WSI level and example patches from each cluster. (c) Head-wise attention map. Red color: rescaled-attention weights > 2; Blue color: rescaled-attention weights = 0. Pie plot in the lower left corner shows the head-wise c-index. (d) Patch level prediction for the selected heads. Rows: best performing head, worst performing head, and all heads combined. Columns: attention map, unscaled risk score for each patch, and weighted risk scores (i.e., attention weight × risk score). "High" and "low" risk scores refer to the maximum and minimum head-wise patient-level risk scores.



Journal Pre-proof

Supplementary Figure S5. Visualization of head-wise attention map and patch clusters for one sample WSI from COAD. (a) Whole Slide Image. (b) Patch clusters on the WSI level and example patches from each cluster. (c) Head-wise attention map. Red color: rescaled-attention weights > 2; Blue color: rescaled-attention weights = 0. Pie plot in the lower left corner shows the head-wise c-index. (d) Patch level prediction for the selected heads. Rows: best performing head, worst performing head, and all heads combined. Columns: attention map, unscaled risk score for each patch, and weighted risk scores (i.e., attention weight × risk score). “High” and “low” risk scores refer to the maximum and minimum head-wise patient-level risk scores.



Journal Pre-proof

Supplementary Table S1. The effect of dropout rates on c-index, evaluated from 4-fold cross-validation using data from the first outer fold. **Boldface**: best for each column.

Dropout	BLCA	BRCA	COAD	LGG
0.00	**0.604**	0.618	0.633	0.757
0.20	0.598	0.622	0.654	0.758
0.50	0.597	0.624	**0.668**	0.757
0.80	0.595	**0.643**	0.657	**0.761**
0.95	0.599	0.636	0.631	0.736

Journal Pre-proof