PS3: A Multimodal Transformer Integrating Pathology Reports with Histology Images and Biological Pathways for Cancer Survival Prediction

Manahil Raza Ayesha Azam Talha Qaiser Nasir Rajpoot
University of Warwick, UK
{manahil.raza, ayesha.azam, talha.qaiser, n.m.rajpoot}@warwick.ac.uk

1. Implementation Details

We employ Pathology Language and Image Pre-Training (PLIP) [8] as our encoder to extract both image and text features from the WSIs and their corresponding pathology reports. As part of our ablation study, we also experiment with QUILT-Net [9] as an alternative feature extractor. Since our approach requires extracting both image and text embeddings, we are inherently constrained to vision-language models for feature extraction.

Both PLIP and QUILT-Net are vision-language models [1] that fine-tune a pretrained contrastive language-image pretraining (CLIP) model [18]. PLIP is trained on OpenPath, a dataset consisting of approximately 200,000 paired pathology image-text pairs, curated from publicly available sources such as medical Twitter [8]. Similarly, QUILT-Net is trained on Quilt-1M, a dataset consisting of 1 million pathology image and text samples, sourced from educational histopathology videos along with other publicly available resources [9].

We train the models to predict disease-specific survival (DSS) [14], employing 5-fold site-stratified cross-validation [7], a widely used approach in the literature. Model performance is evaluated using the concordance index (C-Index) [6], which measures how accurately the model's predicted risks align with actual patient survival outcomes. All models were trained for 50 epochs, utilizing visual and/or text features extracted using the PLIP feature encoder [8]. The training process employed a learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, a cosine learning rate scheduler, and the AdamW optimizer. For MIL-based methods, during training, 4,096 patches were randomly sampled for each WSI. During inference, the entire WSI was processed to generate predictions. MIL-based models were trained using the negative log-likelihood (NLL) loss [24] with a batch size of 1, while prototype-based models were optimized with Cox loss [4] and a batch size of 64. For prototype-based methods, we set the number of histological prototypes to 16, pathway prototypes to 50 and the number of diagnostic prototypes is set to the average length of reports in the training dataset.

2. Multimodal Baselines

Among the Multimodal Baselines, MOTCat [23], MCAT [3], SurvPath [10] and $\text{MMP}_{\text{Trans}}$ [22] utilize transformer-based architectures. With the exception of SurvivMIL [16], all aforementioned models integrate histology images with genomic data for survival prediction. In contrast, SurvivMIL incorporates histology images and pathology reports, making it the only multimodal baseline that integrates text data. Additionally, all pathology-genomics baselines utilize genomic prototypes by grouping genes into either functional categories [3, 13, 23, 25] or biological pathways [5, 10, 19, 22]. However, only the two MMP variants incorporate both histology and pathway prototypes.

3. Clinical Baselines

We conduct both univariate and multivariate Cox regression analyses using clinical variables such as age, sex, and histologic grade. The results in Table.1 highlight our method's performance in comparison to individual clinical variables as well as their combined effect.

4. Attention Visualization

We visualize the histological prototypes created from the WSI and the cross attention between the different modalities [21, 22]. Each WSI is represented by a compact set of 16 histological prototypes. Figure.1.a represents a TCGA-CRC WSI while Figure.1.b displays a heatmap showing the spatial distribution of patches corresponding to each prototype. Figure.1.d illustrates the proportion of patches assigned to each prototype (c), while Figure.1.c highlights representative patches from the most significant prototypes - those with a substantial number of assigned patches. The histological prototypes have been annotated by a pathologist to provide meaningful interpretations. Prototype 0 is associated with normal colon crypts, and 3 captures fibrous connective tissue. Smooth muscle is represented by prototypes 5 and 9, whereas prototype 13 includes both fibrous

Model	BLCA	LUAD	KIRC	STAD	CRC	HNSC	Avg (↑)
Age	0.562 ± 0.064	0.485 ± 0.093	0.558 ± 0.075	0.542 ± 0.096	0.452 ± 0.153	0.490 ± 0.030	0.523
Sex	0.484 ± 0.053	0.533 ± 0.050	0.521 ± 0.051	0.554 ± 0.055	0.556 ± 0.065	0.488 ± 0.046	0.515
Grade	0.512 ± 0.011	n/a	0.731 ± 0.052	0.560 ± 0.039	n/a	0.544 ± 0.059	n/a
All	0.557 ± 0.062	0.494 ± 0.093	0.723 ± 0.044	0.583 ± 0.051	0.496 ± 0.099	0.516 ± 0.090	0.561
PS3	0.684 ± 0.026	0.662 ± 0.102	0.774 ± 0.067	0.638 ± 0.045	0.826 ± 0.101	0.627 ± 0.066	0.702

Table 1. Survival Prediction Using Clinical Variables: The variables include age, sex, and histologic grade, collectively referred to as "All."

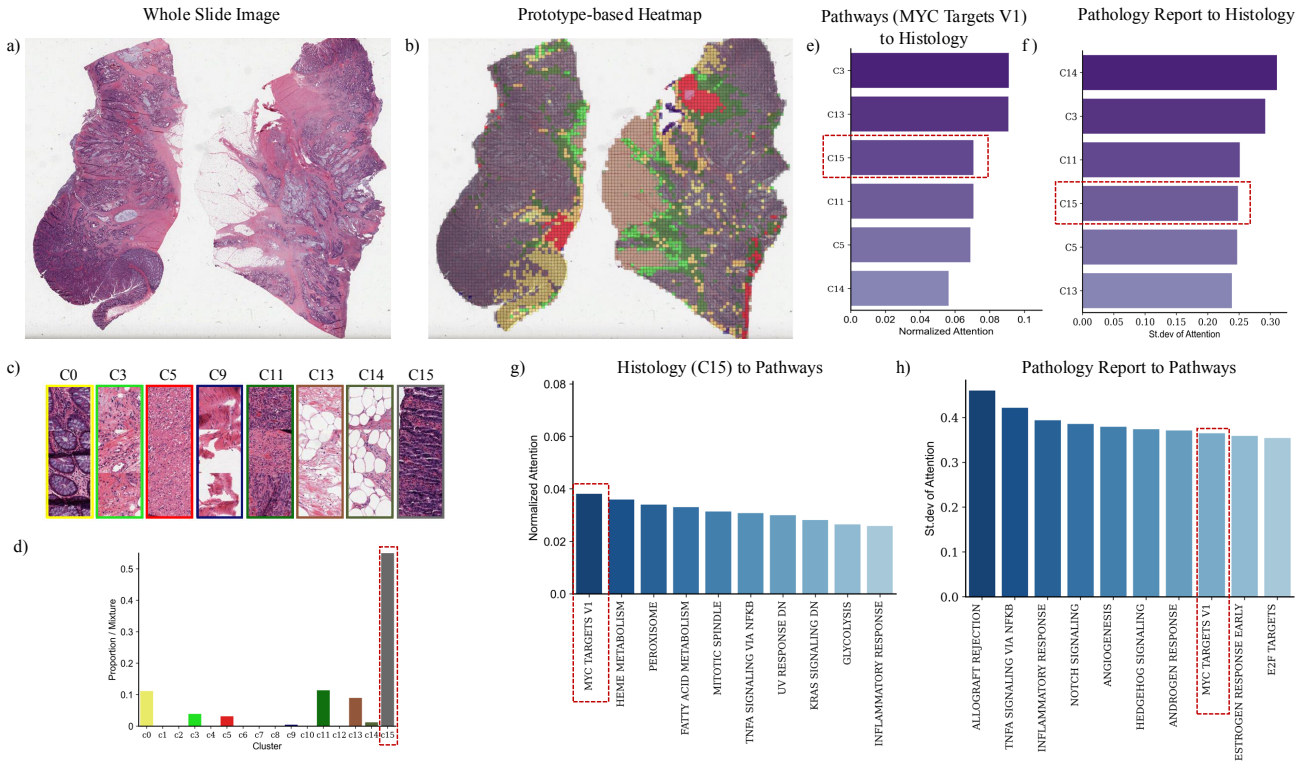

Figure 1. (a) A WSI for a CRC patient. (b) Prototype-based heatmap showing the closest morphological prototype for each patch in the WSI. (c) Top three representative patches for the most significant prototypes. (d) Proportion of each prototype in the WSI. (e) Top six histological prototypes highly attended by the pathway MYC Targets V1. (f) Top six histological prototypes highly attended by the pathology report. (g) Top ten pathways highly attended by C15 (tumor prototype). (h) Top ten pathways highly attended by the pathology report.

and adipose tissue. Prototype 14 corresponds specifically to adipose tissue. Lastly, Prototype 15 represents tumor regions, while 11 corresponds to tumor stroma.

We model cross-modal attention across histology, pathways, and text, capturing their interrelationships. We analyze histology-to-pathway and pathway-to-histology attention to link histological prototypes with relevant biological pathways. Additionally, we model text-to-pathway and text-to-histology interactions to understand how pathology reports emphasize biological pathways and align with morphological features in WSIs.

To analyze pathology reports, we compute the standard deviation of cross-attention scores across all text segments within a single report to identify key pathways and clusters (Figures.1.h,e). Instead of focusing on individual text segments, we consider the entire report to capture the overall diagnostic context. Standard deviation is used instead of averaging attention scores, as it better highlights pathways that receive selective but strong attention from certain segments while being ignored by others, preventing dilution of meaningful signals.

For Prototype 15 ($C = 15$), which represents tumor regions and is the most dominant prototype in the WSI, we identify MYC Targets V1, TNFA Signaling via NF-κB,

and Inflammatory Response as key pathways consistently emphasized by both histology-based and pathology report-based attention (Figures 1.g,h). These pathways have been shown to be important for prognosis [11, 12, 15]. Among these, we visualize the highly attended histological prototypes corresponding to MYC targets V1 and the pathology report, noting that C15 emerges as a highly attended prototype in both (Figures1.e,f) This finding underscores strong bidirectional cross attention between the three modalities.

4.1. Word Clouds

We stratified patients for TCGA-CRC into low- and high-risk groups based on the median cutoff of their predicted risk scores. Using cross-attention mechanisms, we identified the most highly attended text segment within each pathology report, determined by the average attention from all histological prototypes. To explore risk-associated textual patterns, we use the top-ranked text segment for each patient and generated two word clouds—one representing the high-risk group and another for the low-risk group as shown in Fig.2. The provided word clouds categorize two-word phrases instead of single words. The low-risk word cloud (blue) includes terms like "margins negative" and "lymph nodes negative," which indicate that cancer has not spread and are associated with a better prognosis [17]. Additionally, phrases such as "moderately differentiated" and "well differentiated" align well with low-risk pathology, as tumors with these characteristics tend to be less aggressive compared to poorly differentiated ones. Conversely, the high-risk word cloud (red) contains terms that indicate advanced disease and poor prognosis, such as "lymph nodes positive," "poorly differentiated," "serosal involvement," and "radial margin" [20]. These terms reflect features linked to higher recurrence risk, deeper tissue invasion, and metastatic potential, making them indicators of more aggressive colorectal cancer.

5. Kaplan-Meier Analysis

Figure.3 presents Kaplan-Meier survival curves for the predicted high-risk and low-risk groups. Patients with risk scores above the cohort median are classified as high-risk (red), while those below the median are considered low-risk (blue). We compare our proposed model against key baselines, including the best overall multimodal model (MMP_{OT}), the top transformer-based multimodal baseline ($\text{MMP}_{\text{Trans}}$), and the sole histology-text baseline (SurvivMIL). We use the log-rank test [2] to assess whether the difference between high- and low-risk groups is statistically significant, considering a p-value threshold of 0.05.

References

[1] Mohsin Bilal, Manahil Raza, Youssef Altherwy, Anas Alsuhaibani, Abdulrahman Abduljabbar, Fahdah Almarshad, Paul Golding, Nasir Rajpoot, et al. Foundation models in computational pathology: A review of challenges, opportunities, and impact. *arXiv preprint arXiv:2502.08333*, 2025. 1

[2] J Martin Bland and Douglas G Altman. The logrank test. *Bmj*, 328(7447):1073, 2004. 3

[3] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4025, 2021. 1

[4] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34 (2):187–202, 1972. 1

[5] Haitham A Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021. 1

[6] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996. 1

[7] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021. 1

[8] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 1

[9] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024. 1

[10] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11579–11590, 2024. 1

[11] Kyu Sang Lee, Yoonjin Kwak, Kyung Han Nam, Duck-Woo Kim, Sung-Bum Kang, Gheeyoung Choe, Woo Ho Kim, and Hye Seung Lee. c-myc copy-number gain is an independent prognostic factor in patients with colorectal cancer. *PLoS One*, 10(10):e0139727, 2015. 3

[12] Yichao Liang, Xin Wu, Qi Su, Yujie Liu, and Hong Xiao. Identification and validation of a novel inflammatory response-related gene signature for the prognosis of colon cancer. *Journal of inflammation research*, pages 3809–3821, 2021. 3

[13] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molec-

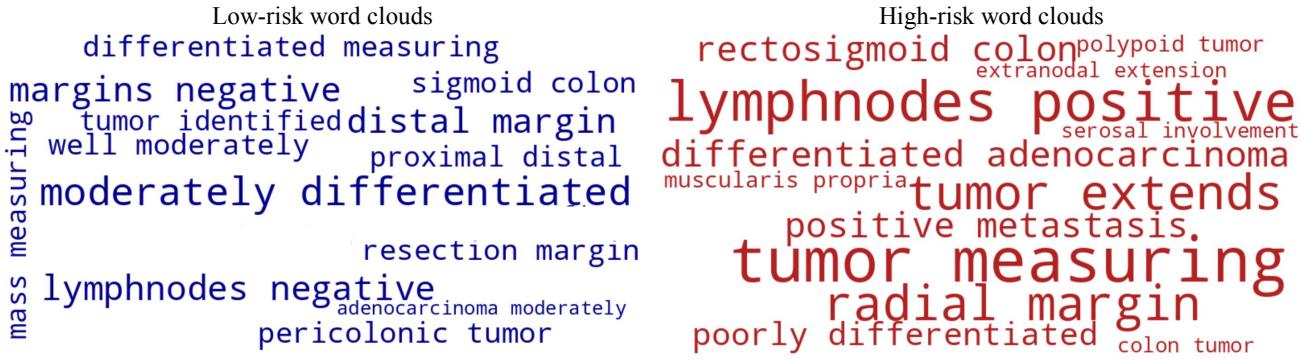

Figure 2. Two-phrase wordclouds for high-risk group (red) and low-risk (blue) group for TCGA-CRC depicting words from top text segments based on histology prototypes.

ular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015. 1

[14] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018. 1

[15] Matthew Martin, Mengyao Sun, Aishat Motolani, and Tao Lu. The pivotal player: components of nf-κb pathway as promising biomarkers in colorectal cancer. *International Journal of Molecular Sciences*, 22(14):7429, 2021. 3

[16] Reed Naidoo, Olga Fourkioti, Matt De Vries, and Chris Bakal. Survivmil: A multimodal, multiple instance learning pipeline for survival outcome of neuroblastoma patients. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2024. 1

[17] Shuji Ogino, Katsuhiko Nosho, Natsumi Irahara, Kaori Shima, Yoshifumi Baba, Gregory J Kirkner, Mari Mino-Kenudson, Edward L Giovannucci, Jeffrey A Meyerhardt, and Charles S Fuchs. Negative lymph node count is associated with survival of colorectal cancer patients, independent of tumoral molecular alterations and lymphocytic reaction. *Official journal of the American College of Gastroenterology— ACG*, 105(2):420–433, 2010. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[19] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019. 1

[20] Sameer Shivji, James R Conner, Valeria Barresi, and Richard Kirsch. Poorly differentiated clusters in colorectal cancer: a current review and implications for future practice. *Histopathology*, 77(3):351–368, 2020. 3

[21] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11566–11578, 2024. 1

[22] Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag Jayant Vaidya, Alexander Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. In *Forty-first International Conference on Machine Learning*, 2024. 1

[23] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21241–21251, 2023. 1

[24] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020. 1

[25] Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21485–21494, 2023. 1

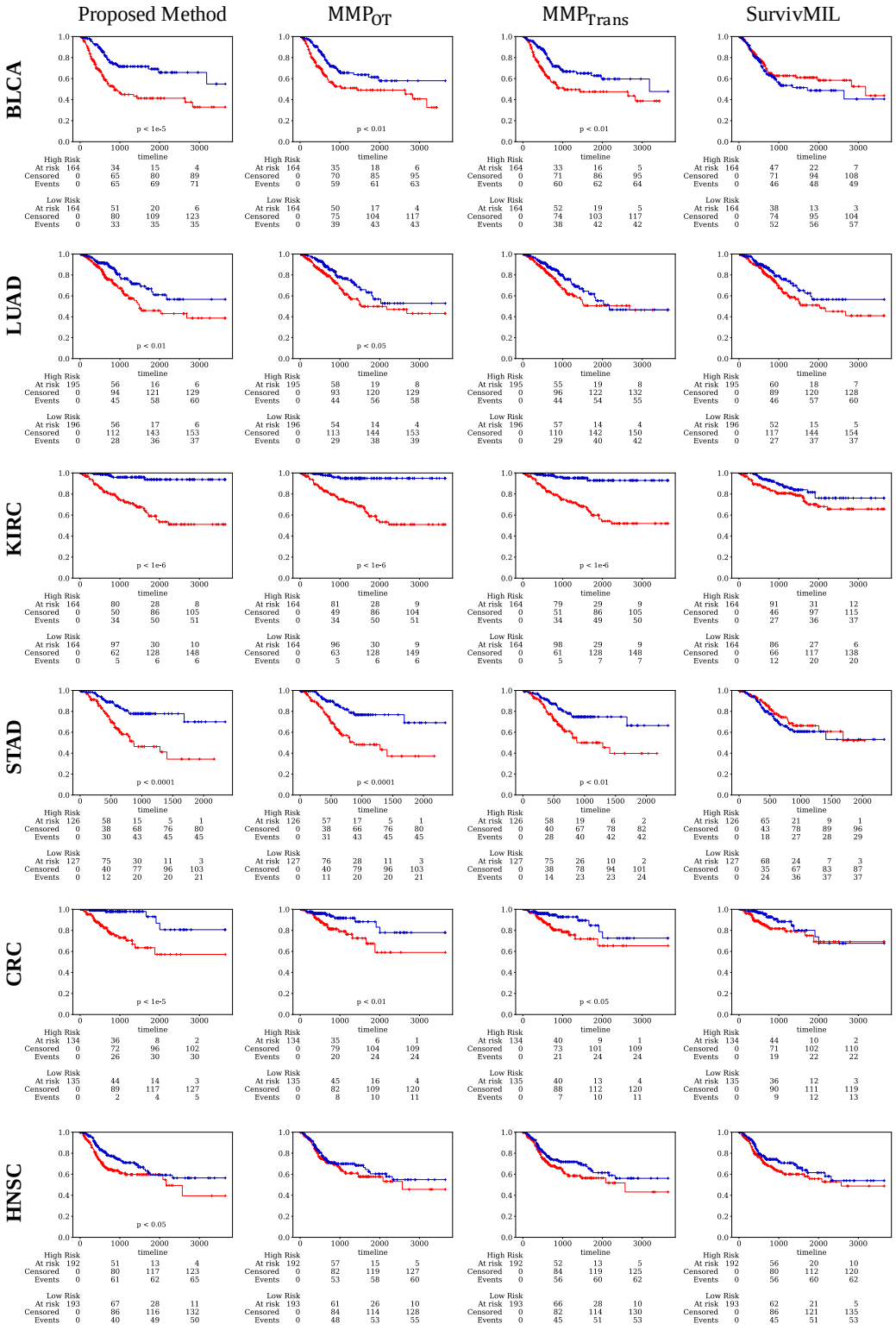

Figure 3. Kaplan-Meier curves comparing the proposed method with multimodal baselines. High-risk (red) and low-risk (blue) groups were stratified using the median predicted risk. Statistical significance was assessed using the log-rank test.