MASurv: Risk-Conditional Martingale Adversarial Training for Calibrated Multi-Slide Survival Prediction in Whole-Slide Histopathology Images

Meghdad Sabouri Rad[1], Vincent Huang[2], Mohammad Mehdi Hosseini[1], Farzaneh Nikbakhtsarvestani[1], Amber Bixby[1], Saverio J. Carello[1], Ola El-Zammar[1], Michel R. Nasr[1], and Bardia Rodd[1] (✉)

[1] SUNY Upstate Medical University, Syracuse, NY, USA
✉ Corresponding author: Bard.rodd@gmail.com
[2] University of South Carolina, Columbia, SC, USA

Abstract. Survival prediction from whole slide images (WSIs) is central to computational pathology, yet existing multiple instance learning (MIL) methods optimise exclusively for discrimination, with no explicit calibration objective. Miscalibration is especially harmful in clinical deployment, where systematic prediction errors across patient subgroups lead to inequitable risk estimates. We present **MASurv**, which augments discrete-time survival training with an *Adversarial Conditional Moment Test* (ACMT) on counting-process residual increments, directly encoding the calibration condition as a moment-orthogonality restriction decomposed locally in time within matched risk strata. A second ACMT task regularises within-patient prediction inconsistency for multi-slide patients. On our internal institutional cohort (662 patients, 5-fold cross-validation (CV)), MASurv achieves the best ICI = 0.096 and WG-ICI = 0.188, a 25% subgroup-calibration reduction over the NLL-only baseline (Cohen's $d = -1.54$, very large effect). More critically, on *zero-shot* internal→TCGA-LUAD transfer (509 patients), MASurv is the *only* method retaining both discrimination (C-index = 0.680 ± 0.044) and calibration (D-cal p=0.143, ICI = 0.079 ± 0.002); all baselines collapse (ICI > 0.36, D-cal $p \approx 0$).

1 Introduction

Survival prediction from gigapixel WSIs is increasingly central to precision oncology [2,11]. The dominant paradigm applies MIL to patch embeddings from a pathology foundation model [10,20,19,13], with models selected by Harrell's C-index [9] (a rank-based metric blind to the *absolute* magnitude of predicted probabilities). Calibration [7,1] is indispensable when model outputs inform treatment decisions: systematic overestimation for one sex or ECOG group is a potential source of inequitable risk estimates in clinical deployment, yet has received little attention in WSI survival analysis. Post-hoc recalibration operates

globally, cannot enforce *conditional* calibration under censoring, and requires held-out target data unavailable in cross-institution deployment.

A classical result from counting-process theory provides a handle: under a correctly specified discrete-time hazard, the *martingale residual increments* satisfy

$$\Delta\hat{M}_i(k) = \Delta N_i(k) - Y_i(k)\hat{\lambda}_i(k), \tag{1}$$

where $\Delta N_i(k) \in \{0,1\}$ is the bin-k event indicator, $Y_i(k) \in \{0,1\}$ the at-risk indicator, and $\hat{\lambda}_i(k)$ the predicted per-bin hazard; these increments satisfy $\mathbb{E}[\Delta\hat{M}_i(k) \mid \mathcal{F}_{k-1}] = 0$ under a correctly specified hazard, where $\mathcal{F}_{k-1}$ is the event/censoring history through bin $k-1$ [4]. Systematic non-zero correlation with any nuisance covariate (sex, ECOG, slide count) therefore implies subgroup miscalibration (a falsifiable, covariate-specific signal). We exploit this to formulate calibration as a testable *conditional moment restriction* [8] and enforce it adversarially during training, calling the resulting method the Adversarial Conditional Moment Test (ACMT). Our contributions are:

1. **ACMT loss**: A time-local, risk-conditional adversarial moment test on counting-process residual increments as a principled, grounded calibration objective (Section 3.3).
2. **Multi-slide ACMT (Task B)**: A second ACMT critic enforcing within-patient prediction consistency for multi-slide patients (Section 3.4).
3. **Zero-shot calibration transfer**: To our knowledge, this study represents the first demonstration in WSI survival MIL that moment-orthogonality–based training preserves model calibration under cross-institutional domain shift, whereas existing MIL baselines exhibit significant calibration degradation (Section 4.3).

2 Related Work

MIL for WSI survival. ABMIL [10], TransMIL [15], DTFD-MIL [20], and DeepAttnMISL [19] represent the dominant MIL paradigms for WSI survival; all optimise the NLL and report C-index as the primary metric, without any explicit calibration objective. More recent survival-distribution models such as SCMIL [14] refine the architecture (sparse context-aware attention and a mixture-density head) but likewise train with censored NLL and include no conditional-calibration term.

Calibration in survival analysis. D-Calibration [7] tests whether predicted survival probabilities are uniformly distributed at event times via a χ^2 test. ICI [1] measures mean absolute calibration error. X-CAL [16] introduces a differentiable D-calibration proxy as a training regulariser, but is global (not conditional on subgroups) and not designed for the multi-slide MIL setting.

Adversarial training in pathology. Domain-adversarial training (DAT) [5] places a domain classifier on the *feature representation* via gradient reversal, enforcing feature invariance (neither equivalent to calibration in the counting-process

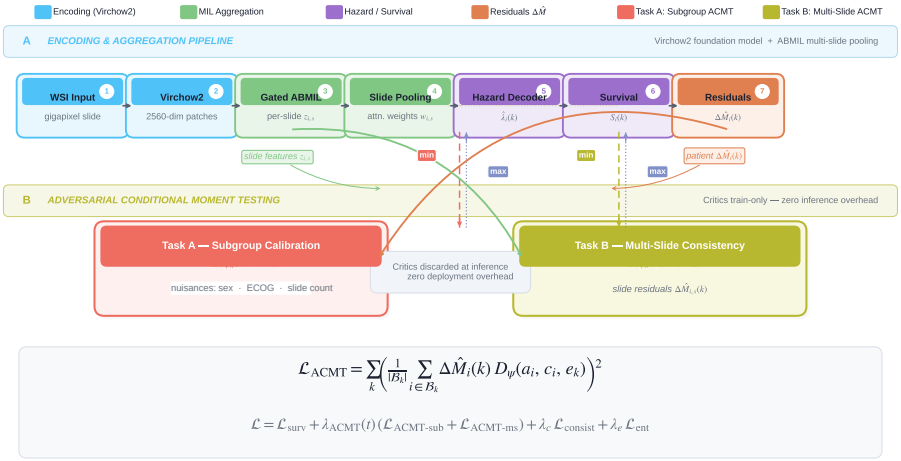

Fig. 1. MASurv pipeline. Virchow2 features are aggregated by gated ABMIL and decoded into discrete-time hazards $\hat{\lambda}_i(k)$. Residuals $\Delta\hat{M}_i(k)$ train ACMT critics for calibration and consistency, *discarded at inference.*

sense nor defined for censored residuals). MASurv instead places its critic on *residual increments* $\Delta\hat{M}_i$, the calibration-relevant quantity. Because conditional calibration requires residual orthogonality to *every* function of the nuisance/risk variables (an infinite moment family a fixed-direction penalty cannot enforce), the critic searches its parametric class for the worst-case residual–nuisance dependence, motivating the adversarial design over fixed moment matching and grounding the objective in survival theory.

3 Method

3.1 Discrete-Time Hazard Model

Time is discretised into $K{=}4$ bins (chosen to balance temporal resolution against minimum bin occupancy; each bin contains $\geq$15% of events in the internal cohort) with boundaries $b_0{<}b_1{<}\cdots{<}b_K$ set to event-time quantiles (0%, 25%, 50%, 75%, 100%) estimated from training event times per fold. The model produces per-bin conditional hazards:

$$\hat{\lambda}_i(k) = \Pr\big(T \in [b_{k-1}, b_k) \mid T \geq b_{k-1},\, x_i\big), \quad k = 1, \ldots, K, \tag{2}$$

with discrete-time survival $\hat{S}_i(k) = \prod_{j=1}^{k}\big(1{-}\hat{\lambda}_i(j)\big)$. The negative log-likelihood (NLL) loss is:

$$\mathcal{L}_{\text{surv}} = -\frac{1}{n}\sum_i \Big[\delta_i \log \hat{\lambda}_i(k_i) + \sum_{j<k_i} \log\big(1{-}\hat{\lambda}_i(j)\big)\Big], \tag{3}$$

where k_i is the event-time bin index and $\delta_i \in \{0, 1\}$ the event indicator ($\delta_i = 0$ for censored observations).

3.2 Multi-Slide Feature Aggregation

Patient i has M_i slides; Virchow2 [17] encodes each slide into patch embeddings of dimension 2560, extracting up to 4096 patches per slide. A gated ABMIL module [10] produces per-slide attention weights and patient embedding:

$$w_{i,s} = \frac{\exp\big(g(z_{i,s})\big)}{\sum_{s'} \exp\big(g(z_{i,s'})\big)}, \qquad z_i = \sum\nolimits_s w_{i,s}\, z_{i,s}. \tag{4}$$

Patient hazards $\hat{\lambda}_i(k)$ are decoded from the pooled embedding z_i; slide-level hazards $\hat{\lambda}_{i,s}(k)$ are additionally decoded from each individual $z_{i,s}$ for Task B.

3.3 Adversarial Conditional Moment Test — Task A

Counting-process residual increments. Let $\Delta N_i(k) \in \{0, 1\}$ be the bin-k event indicator and $Y_i(k) \in \{0, 1\}$ the at-risk indicator. The residual increment $\Delta\hat{M}_i(k)$ (Eq. 1) satisfies $\mathbb{E}[\Delta\hat{M}_i(k) \mid \mathcal{F}_{k-1}] = 0$ under a correctly specified hazard [4]. Systematic non-zero correlation with a nuisance covariate a_i implies subgroup miscalibration. We use three nuisances (sex, ECOG, and slide count) spanning the demographic, clinical, and multi-slide axes; age and Charlson Comorbidity Index are excluded as they overlap with legitimate prognostic signal, so orthogonalising against them would risk discarding true risk rather than miscalibration. They are one-hot concatenated into a_i and fed to a single shared critic D_{sub,ψ_A}.

Risk-conditional matching. The conditioning vector $c_i = [\tilde{H}_i(t_1^*),\ \tilde{H}_i(t_2^*),\ \tilde{H}_i(t_3^*),\ f_i]$ encodes the cumulative-hazard profile at the 25th/50th/75th event-time percentiles (fixed across folds) and a coarse follow-up bin $f_i \in \{1, 2, 3, 4\}$. $\tilde{H}_i = -\log \hat{S}_i$ is computed with stop-gradient (no gradient through c_i). Let $\mathcal{B}_k$ be the matched patient set for bin k, e_k a learnable bin embedding; bins with $|\mathcal{B}_k| < 8$ are skipped.

ACMT loss (Task A).

$$\mathcal{L}_{\text{ACMT-sub}} = \sum_{k=1}^{K} \left(\frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \Delta\hat{M}_i(k)\ D_{\mathrm{sub},\psi_A}(a_i, c_i, e_k) \right)^2, \tag{5}$$

a squared empirical moment restriction in the spirit of GMM [8]. The survival model *minimises* (5); the critic *maximises* it.

3.4 Multi-Slide ACMT — Task B

Slide-level residuals $\Delta\hat{M}_{i,s}(k) = \Delta N_i(k) - Y_i(k)\hat{\lambda}_{i,s}(k)$ feed a separate critic D_{ms,ψ_B} ($\psi_B \neq \psi_A$), where $\mathcal{B}'_k$ denotes the set of (patient, slide) pairs (i, s) with patient i

in risk stratum k:

$$\mathcal{L}_{\text{ACMT-ms}} = \sum_k \left(\frac{1}{|\mathcal{B}'_k|} \sum_{(i,s)\in\mathcal{B}'_k} \Delta\hat{M}_{i,s}(k)\; D_{\text{ms},\psi_B}(a_{i,s}, c_i, e_k) \right)^2, \tag{6}$$

regularising within-patient consistency independently of Task A.

3.5 Multi-Objective Model

The training objective combines all loss terms:

$$\mathcal{L} = \mathcal{L}_{\text{surv}} + \lambda_{\text{ACMT}}(t)\big(\mathcal{L}_{\text{ACMT-sub}} + \mathcal{L}_{\text{ACMT-ms}}\big) + \lambda_c\,\mathcal{L}_{\text{consist}} + \lambda_e\,\mathcal{L}_{\text{ent}}, \tag{7}$$

where $\mathcal{L}_{\text{consist}} = \frac{1}{n}\sum_i \frac{1}{M_i}\sum_{s,k}\big(\hat{\lambda}_{i,s}(k) - \text{sg}[\hat{\lambda}_i(k)]\big)^2$ is a stop-gradient patient–slide hazard consistency term (sg$[\cdot]$ stops the gradient, pulling each slide hazard towards the detached patient hazard), and $\mathcal{L}_{\text{ent}} = -\frac{1}{n}\sum_i H(w_i)$, with $H(w_i) = -\sum_s w_{i,s}\log w_{i,s}$ the entropy of patient i's slide-attention distribution; minimising it with $\lambda_e > 0$ maximises attention entropy and discourages collapse onto a single slide. The full objective is solved as a min–max problem: the critics D_{sub,ψ_A} and D_{ms,ψ_B} *maximise* the ACMT terms $\mathcal{L}_{\text{ACMT-sub}} + \mathcal{L}_{\text{ACMT-ms}}$ while the survival model *minimises* the total loss, via alternating updates. Four stabilisation measures prevent adversarial collapse: (i) a 20-epoch NLL-only warmup; (ii) linear ramp of λ_{ACMT} from 0 to 0.1 over 20 epochs; (iii) Task B activated 10 epochs after Task A; (iv) spectral normalisation [12] on both critics. Key hyperparameters: λ_{ACMT}=0.1, λ_c=0.5, Adam 5×10^{-5}, batch 32, 70 epochs. Critics add only 7,618 parameters (<1%).

4 Experiments

4.1 Data, Protocol, and Metrics

Internal cohort. 662 lung cancer patients from an internal institutional cohort; covariates: age, sex, ECOG, Charlson Comorbidity Index; Virchow2 features [17], 4096 patches/slide; 5-fold CV with strict patient-level splits (all slides from one patient stay in the same fold).

TCGA-LUAD (external, zero-shot). 509 adenocarcinoma patients, 1,567 slides; public TCGA repository. **Zero-shot protocol**: all five internally trained checkpoints (one per fold) are independently applied to all 509 TCGA patients. TCGA results are mean ± std across the five internal fold checkpoints; all baselines receive the identical evaluation.

Metrics. C-index (C-statistics) [9] (↑) for discrimination; IBS [6] (↓) for integrated Brier score; ICI [1] (↓) for mean absolute calibration error; WG-ICI (↓): $\max_{g\in\mathcal{G}}$ ICI(g) where $\mathcal{G}$ = {male, female, ECOG≤1, ECOG≥2, single-slide, multi-slide} (6 groups); D-cal p [7] (↑), with p>0.05 indicating non-rejection.

Table 1. internal 5-fold CV (mean ± std across folds; patient-level splits). **Bold** = best per column. WG-ICI = worst-group ICI (sex, ECOG, slide-count subgroups). †High D-cal variance (±0.332) reflects fold-level χ^2 instability under small effective event counts. MASurv targets *reduction* in calibration error, not a statistical D-cal pass; the latter is demonstrated on the larger TCGA cohort (p=0.143).

Method	C-idx ↑	IBS ↓	ICI ↓	WG-ICI ↓	D-cal p ↑
ABMIL [10]	$0.649_{\pm.029}$	$0.153_{\pm.008}$	$0.099_{\pm.016}$	$0.222_{\pm.035}$	$0.002_{\pm.003}$
TransMIL [15]	$0.642_{\pm.027}$	$0.165_{\pm.009}$	$0.108_{\pm.031}$	$0.245_{\pm.042}$	$\mathbf{0.166}^{\dagger}_{\pm.332}$
DTFD-MIL [20]	$\mathbf{0.652}_{\pm.030}$	$0.155_{\pm.013}$	$0.116_{\pm.039}$	$0.222_{\pm.059}$	$0.069_{\pm.138}$
DeepAttnMISL [19]	$\mathbf{0.652}_{\pm.029}$	$0.153_{\pm.016}$	$0.103_{\pm.028}$	$0.239_{\pm.041}$	$0.029_{\pm.057}$
Baseline (NLL only)	$0.651_{\pm.028}$	$0.153_{\pm.011}$	$0.117_{\pm.037}$	$0.249_{\pm.044}$	$0.010_{\pm.020}$
MASurv (ours)	$\mathbf{0.710}_{\pm.034}$	$\mathbf{0.149}_{\pm.009}$	$\mathbf{0.096}_{\pm.015}$	$\mathbf{0.188}_{\pm.024}$	$0.018_{\pm.027}$

Baselines. ABMIL [10], TransMIL [15], DTFD-MIL [20], DeepAttnMISL [19], and **Baseline** (identical architecture, λ_{ACMT}=0). All share Virchow2 features, identical patient-level splits, training schedules (Adam, lr = 5×10^{-5}, 70 epochs, batch 32, same seeds per fold), and early-stopping criteria (min val-NLL, patience 15). TransMIL is excluded from TCGA: OOM (>24 GB) despite gradient checkpointing and patch/slide reductions.

4.2 Internal Cohort Validation

Results in Table 1 and Fig. 2. C-index increase does not reach significance (Wilcoxon p=0.44, n=5); MASurv achieves the lowest ICI (0.096, −18%), WG-ICI (0.188, −25%; Cohen's d=−1.54, very large effect; t-test p=0.071), and IBS (0.149), winning WG-ICI 4/5 folds (single exception F2, +0.018, smaller than any win margin). Largest subgroup gains (Fig. 2): single-slide (0.246→0.157, −36%; Task A alone, confirming Task A independently effective), male (−17%), ECOG-low (−18%, fold std 0.006 vs. 0.037).

4.3 Zero-Shot Cross-Institution Transfer to TCGA-LUAD

Preprocessing and leakage sanity checks. To rule out artefacts that could inflate the MASurv advantage: (i) strict TCGA patient-level aggregation and de-duplication (all slides from each patient pooled; no duplicate patients in the evaluation set); (ii) identical Virchow2 patch extraction across all methods (4,096 patches/slide at 0.5 mpp, same tissue segmentation mask); (iii) hazard bins frozen from internal training quantiles, no TCGA data informs bin boundaries; (iv) identical event/censoring definitions (δ_i=1 all-cause death, δ_i=0 otherwise) and slide inclusion criteria (≥200 foreground patches); All externally evaluable baselines were rejected by D-calibration on TCGA ($p\approx0$) with ICI > 0.36 (Fig. 3). **MASurv is the only method non-rejected by D-calibration on external data** (p=0.143), with ICI = 0.079, 4.6× lower than the best baseline (DTFD-MIL,

Fig. 2. Top: Per-fold internal cohort results (MASurv vs. Baseline, F1–F5). MASurv wins WG-ICI 4/5, ICI 3/5, IBS 4/5. **Bottom:** Subgroup ICI across six internal cohort groups; MASurv reduces ICI in all six (largest: single-slide −36%).

ICI = 0.366). This demonstrates that calibration-regularised representations generalise *broadly*. Fig. 4 shows MASurv's curves tracking the diagonal across all four bins, while ABMIL deviates by up to 0.41 absolute units. Notably, the internal C-index ranking does not predict TCGA performance (Spearman ρ=−0.70): despite leading on both internal calibration and C-index, MASurv gains further ground on TCGA, while DeepAttnMISL (joint top on the internal cohort) falls to last place. This rank reversal is consistent with moment-orthogonality training preventing discrimination-centric overfitting that causes severe calibration degradation in all baselines.

4.4 Multi-Slide Stress Test and Ablation

Among 312 multi-slide patients in the internal cohort, MASurv achieves ICI = 0.101 vs. Baseline 0.117, with within-patient risk variance reduced by **26.7%** (4.43×10^{-4} vs. 5.99×10^{-4}). Fig. 5 shows a 12-variant ablation on internal 5-fold CV. Variants include: NLL-only (λ_{ACMT}=0); BCE-adversary [5]; adversary on features z_i not residuals; no risk conditioning; global (not time-local) moments; no Task B; no ramp; and further ablations. Removing the adversary raises WG-ICI by 0.033

Table 2. Zero-shot internal cohort → TCGA-LUAD (509 patients). C-index, ICI, IBS, WG-ICI: mean ± std *across five internal cohort fold checkpoints* (std = training-fold variability, not patient resampling). D-cal: χ^2 test computed on TCGA patient predictions (single point estimate, no ±). $p \approx 0$ = rejection at all conventional levels. TransMIL excluded (OOM; see Baselines, Sec. 4).

Method	C-idx ↑	IBS ↓	ICI ↓	WG-ICI ↓	D-cal p ↑
ABMIL [10]	$0.592_{\pm.006}$	$0.098_{\pm.004}$	$0.392_{\pm.011}$	$0.405_{\pm.012}$	≈0
DTFD-MIL [20]	$0.594_{\pm.006}$	$0.104_{\pm.008}$	$0.366_{\pm.027}$	$0.376_{\pm.029}$	≈0
DeepAttnMISL [19]	$0.588_{\pm.008}$	$0.102_{\pm.014}$	$0.382_{\pm.064}$	$0.395_{\pm.064}$	≈0
MASurv (ours)	$\mathbf{0.680}_{\pm.044}$	$\mathbf{0.087}_{\pm.003}$	$\mathbf{0.079}_{\pm.002}$	$\mathbf{0.175}_{\pm.014}$	**0.143**

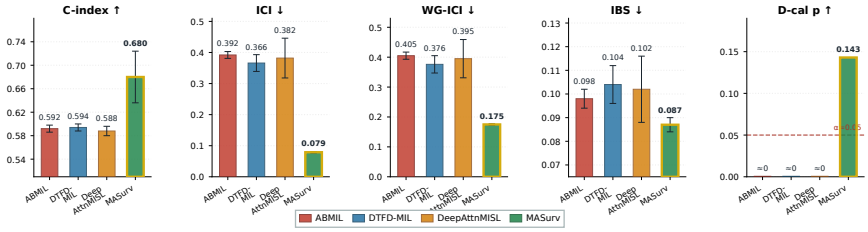

Fig. 3. Zero-shot internal cohort → TCGA-LUAD across five metrics. MASurv (green) leads all; dashed line α=0.05; all baselines D-cal $p \approx 0$.

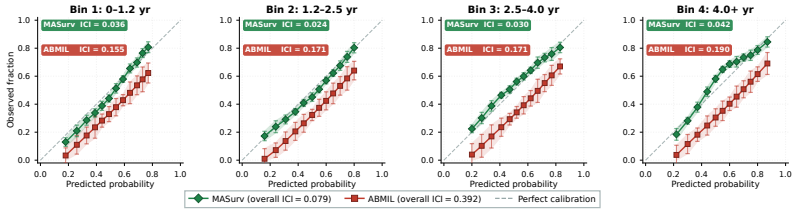

Fig. 4. TCGA-LUAD calibration curves by time bin. MASurv (green, ICI = 0.079) tracks the diagonal well; ABMIL (red, ICI = 0.392) systematically overestimates survival.

(p=0.078); *Feature adv.* (critic on z_i) produces the *worst* WG-ICI, validating residual targeting over feature confusion; *ACMT→BCE* raises ICI; no variant changes C-index (all p>0.17).

5 Conclusion

Limitations and Conclusion. Bins with fewer than 8 at-risk patients are skipped, reducing adversarial supervision in heavily censored cohorts; the study covers a single encoder (Virchow2) and cancer type, with cross-encoder validation (UNI [3], GigaPath [18]) and multi-cancer extension as primary future directions; X-CAL [16] comparison is deferred. Unlike post-hoc recalibration, ACMT produces

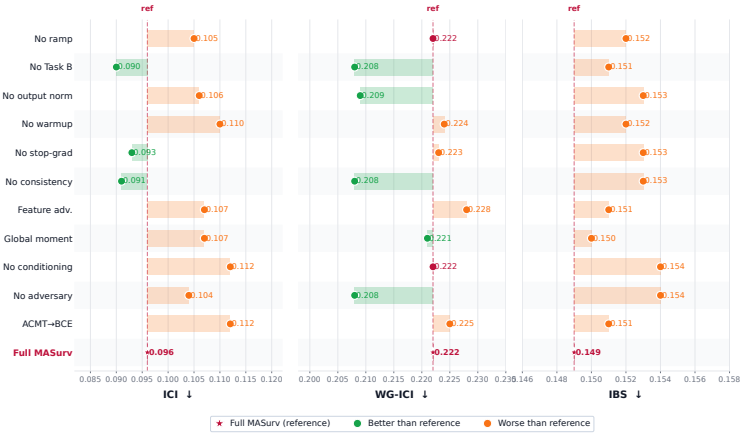

Fig. 5. 12-variant ablation (internal 5-fold CV): ICI, WG-ICI, IBS. **Red** = MASurv; orange/green = worse/better. *Feature adv.* has worst WG-ICI; *No adversary* highest IBS.

built-in conditional calibration requiring no target data, transferring across institutions, an essential property for equitable clinical deployment.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Austin, P.C., Harrell Jr., F.E., van Klaveren, D.: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat. Med. **39**(21), 2714–2742 (2020)
2. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR. pp. 16144–16155 (2022)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nat. Med. **30**, 850–862 (2024)
4. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. Wiley (1991)
5. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(59), 1–35 (2016)
6. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Stat. Med. **18**(17–18), 2529–2545 (1999)

7. Haider, H., Hoehn, B., Davis, S., Greiner, R.: Effective ways to build and evaluate individual survival distributions. J. Mach. Learn. Res. **21**(85), 1–63 (2020)
8. Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica **50**(4), 1029–1054 (1982)
9. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
10. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: ICML. PMLR, vol. 80, pp. 2127–2136 (2018)
11. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**, 555–570 (2021)
12. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
13. Sabouri Rad, M., Huang, J., Hosseini, M.M., Choudhary, R., Siezen, H., Akabari, R., Jamaspishvili, T., El-Zammar, O., Patel, P.G., Carello, S.J., Nasr, M.R., Rodd, B.: Deep learning for digital pathology: a critical overview of methodological framework. J. Pathol. Inform. **19**, 100514 (2025)
14. Yang, Z., Liu, H., Wang, X.: SCMIL: sparse context-aware multiple instance learning for predicting cancer survival probability distribution in whole slide images. In: MICCAI 2024. LNCS, pp. 448–458. Springer, Cham (2024)
15. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In: NeurIPS. vol. 34, pp. 2136–2147 (2021)
16. Goldstein, M., Han, X., Puli, A., Perotte, A.J., Ranganath, R.: X-CAL: explicit calibration for survival analysis. In: NeurIPS. vol. 33, pp. 18296–18307 (2020)
17. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: scaling self-supervised mixed magnification models in pathology. arXiv:2408.00738 (2024)
18. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**, 181–188 (2024)
19. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med. Image Anal. **65**, 101789 (2020)
20. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: CVPR. pp. 18802–18812 (2022)