Robustness of Multimodal Foundation-Model Forecasting for Postoperative Cancer Outcomes

Anonymous Authors[1]

Abstract

Postoperative outcome forecasting is a stringent test of whether frozen medical foundation-model embeddings can support clinically meaningful intelligence: predictions must remain useful under censoring, limited event counts, and comparison with established clinical anchors. We study two-year disease-free survival (DFS) forecasting in an anonymized in-house resected NSCLC/LUAD cohort with paired preoperative computed tomography (CT) and postoperative hematoxylin-and-eosin whole-slide images (WSI). Using frozen patient-level embeddings, we evaluate a complete 2×2 matrix of CT foundation models (Pillar-0, CT-FM) and pathology foundation models (TITAN, Prov-GigaPath), together with WSI-only, CT-only, and score-level fusion models. Simple late averaging is the most stable fusion rule across all model combinations; the strongest TITAN plus Pillar-0 setting reaches a C-index of 0.799 and AUROC of 0.810, improving over its matched WSI-only baseline. However, a stage-only clinical anchor reaches a C-index of 0.837 and AUROC of 0.840 in the same task, and an exploratory TCGA-KIRC stress test similarly favors clinical/Leibovich-like baselines over frozen image embeddings. These results support a clinically anchored view of multimodal foundation embeddings: they are scalable forecasting substrates, but their value should be judged by incremental benefit, calibration, and robustness rather than by standalone image-only performance.

1. Introduction

Forecasting is not merely retrospective classification with delayed labels. It asks whether a model can anticipate future events under uncertainty and remain useful when outcomes are censored, updated, or compared with simpler risk signals already available to clinicians (Gneiting & Raftery, 2007; Makridakis et al., 2020). Postoperative oncology is a natural setting for this framing: recurrence and survival outcomes are delayed, clinically consequential, and shaped by both tumor biology and treatment context.

Postoperative recurrence forecasting is therefore a useful test bed for multimodal foundation models. CT captures macroscopic tumor burden, anatomy, and host context, whereas WSI captures microscopic morphology, stromal composition, and tumor microenvironment. A multimodal representation has a plausible route to incremental value when both modalities are routinely collected around surgery. The hard question is not whether image embeddings contain signal in isolation, but whether that signal remains stable and clinically interpretable when compared with standard prognostic anchors.

Recent pathology and radiology foundation models make this question practically testable. Whole-slide models such as TITAN, UNI, and Prov-GigaPath produce transferable slide- or patient-level representations (Chen et al., 2024; Ding et al., 2025; Xu et al., 2024), and CT foundation models such as Pillar-0 and CT-FM provide transferable volumetric representations (Agrawal et al., 2025; Pai et al., 2025). Prior pathology and multimodal studies have reported cancer outcome signals from histology or CT-WSI fusion (Coudray et al., 2018; Wulczyn et al., 2020; Boeke et al., 2025; Song et al., 2025; Vanguri et al., 2022). Yet it remains unclear whether frozen, general-purpose image embeddings provide robust incremental value in smaller real-world cohorts, and how such value should be interpreted when strong clinical variables are available.

We frame the paper around three questions. First, can frozen CT and WSI foundation embeddings be combined without end-to-end retraining for postoperative DFS forecasting? Second, are the resulting gains stable across a matrix of foundation-model backbones rather than tied to a single favored pairing? Third, how should image-only or image-fusion performance be interpreted against clinical anchors? Our contribution is a compact benchmark of score-level fusion over a complete 2×2 model matrix in an in-house

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2026 Workshop on Forecasting as a New Frontier of Intelligence. Do not distribute.

lung cancer cohort, with explicit clinical anchoring, case-level risk comparison, and a secondary renal-cancer stress test used only to probe robustness.

2. Methods

2.1. Cohorts and Task

The primary cohort contains 620 anonymized patients with resected NSCLC/LUAD and paired preoperative CT and postoperative H&E WSI. The primary benchmark evaluates two-year DFS among cases with sufficient endpoint ascertainment, yielding 466 eligible patients and 56 DFS events. Models are trained and evaluated with balanced cross-validation folds, and all performance values reported below are pooled out-of-fold estimates.

As a secondary stress test, we reuse the same score-level evaluation logic in TCGA-KIRC with paired CT and WSI, using a recurrence-like survival endpoint. This analysis is deliberately exploratory: it tests whether the interpretation from the lung cohort is directionally consistent in another organ system, not whether renal cancer should become the main empirical claim. For renal cancer, comparison with a Leibovich-like baseline is important because mature postoperative risk systems already summarize strong stage, grade, size, and necrosis signals (Leibovich et al., 2003).

2.2. Frozen Foundation Embeddings

For CT, we evaluate Pillar-0 and CT-FM. Pillar-0 is a radiology foundation model pretrained on multi-organ CT and MRI volumes and evaluated across hundreds of radiologic findings, including long-horizon lung cancer risk prediction (Agrawal et al., 2025). CT-FM is a 3D CT foundation model pretrained on 148,000 CT scans from the Imaging Data Commons by label-agnostic contrastive learning (Pai et al., 2025). For WSI, we evaluate TITAN and Prov-GigaPath. TITAN is a multimodal whole-slide foundation model trained with visual self-supervision and vision-language alignment to produce general-purpose slide representations (Ding et al., 2025). Prov-GigaPath is an open-weight whole-slide foundation model pretrained on 1.3 billion pathology tiles from 171,189 slides across 31 tissue types (Xu et al., 2024). We keep all foundation encoders frozen and train lightweight task heads on their patient-level embeddings.

2.3. Score-Level Fusion

Let z_i^{ct} and z_i^{wsi} denote frozen patient-level CT and WSI embeddings for patient i. Modality-specific Cox heads (Cox, 1972) produce base risk scores

$$s_i^{\mathrm{ct}} = h_{\mathrm{ct}}(z_i^{\mathrm{ct}}), \qquad s_i^{\mathrm{wsi}} = h_{\mathrm{wsi}}(z_i^{\mathrm{wsi}}). \tag{1}$$

We focus on three score-level fusion rules, illustrated in Figure 1. Late average uses no learned combiner:

$$\hat{s}_i = \frac{1}{2}\left(s_i^{\mathrm{ct}} + s_i^{\mathrm{wsi}}\right). \tag{2}$$

Late stacking learns a fold-local linear Cox combiner from base scores, following the principle that stacked models should be trained on out-of-fold predictions rather than in-sample fitted scores (Wolpert, 1992):

$$\hat{s}_i = g_\theta\left([s_i^{\mathrm{wsi}}, s_i^{\mathrm{ct}}]\right). \tag{3}$$

Anchor residual stacking uses WSI as the anchor score and asks the CT score to enter through a residual contrast:

$$\Delta_i = s_i^{\mathrm{ct}} - s_i^{\mathrm{wsi}}, \qquad \hat{s}_i = r_\theta\left([s_i^{\mathrm{wsi}}, \Delta_i]\right). \tag{4}$$

For learned combiners, the training features are generated from inner out-of-fold base scores within each outer training fold, and the fitted combiner is then applied to held-out patients. This prevents the score-level model from seeing in-fold base predictions for the cases it is trained to combine. Two-year event probabilities are obtained by fold-local logistic calibration of the final risk score; ranking metrics use the Cox risk score.

2.4. Evaluation and Clinical Anchoring

We report pooled out-of-fold AUROC for two-year DFS, Harrell's concordance index, Brier score, and PRAUC (Harrell et al., 1996; Graf et al., 1999). Clinical-only Cox baselines are trained on available clinicopathologic variables. The most compact anchor uses pathologic stage alone; broader anchors include TNM, tumor size, grade, STAS, visceral pleural invasion, lymphovascular invasion, and available molecular variables. These anchors are calibration points for whether foundation embeddings add clinically meaningful forecast information beyond variables already available to clinicians (Yang et al., 2017; Steyerberg et al., 2012; Van Calster et al., 2019; Collins et al., 2024; Moons et al., 2025).

3. Results

3.1. Cross-Model Feasibility

Table 1 summarizes the complete foundation-model matrix. Across all four CT-WSI pairings, late average improves the matched WSI-only C-index, with gains from 0.0066 to 0.0349. The strongest configuration is TITAN plus Pillar-0, reaching C-index 0.799 and AUROC 0.810. The pattern is informative in both positive and negative directions: WSI-only embeddings provide the stronger unimodal signal, CT-only embeddings are above random but weaker, and fusion is beneficial only when constrained to a low-variance

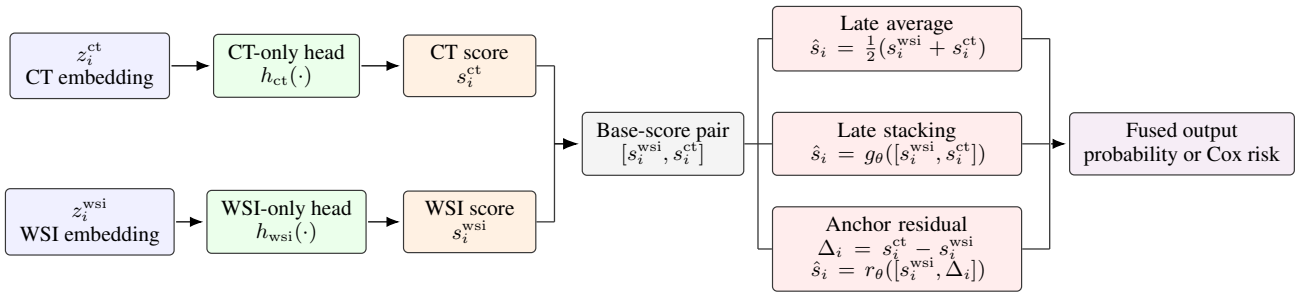

Figure 1. Score-level fusion after unimodal prediction. Frozen CT and WSI patient embeddings are first mapped to modality-specific Cox risk scores. Late average, late stacking, and anchor residual stacking then operate only on these base scores, using out-of-fold training scores for learned combiners.

score-level operation. This suggests that CT and WSI embeddings are weakly complementary, but that the available event count does not support flexible fusion without stronger regularization or additional data.

3.2. Fusion Rules and Clinical Anchors

Table 2 compares fusion variants for the best image-only pairing against clinical anchors. Late average is the strongest and most stable image-only fusion rule. In contrast, learned score stacking and anchor residual stacking underperform, consistent with the risk that even lightweight learned combiners can overfit when only 56 DFS events are available (Peduzzi et al., 1995; Vittinghoff & McCulloch, 2007; Ogundimu et al., 2016; Riley et al., 2019). This is not a failure of multimodal learning in general; it is an empirical constraint on what can be claimed from a modest-event postoperative cohort.

The clinical anchors are deliberately strong. A stage-only clinical anchor reaches C-index 0.837 and AUROC 0.840, exceeding the strongest image-only fusion model. Broader clinicopathologic and molecular feature sets remain competitive but do not exceed stage alone in this endpoint snapshot. This result is central to the paper's message: foundation embeddings are feasible and reusable, but their incremental value must be demonstrated against clinically familiar predictors, not only against unimodal image baselines.

Figure 2 shows this at the patient level: fusion can rescue some WSI-only rankings, clinical anchors can dominate uncertain image scores, and fusion can attenuate plausible false positives. These examples are illustrative audits of risk shifts, not additional evidence of superiority.

The exploratory TCGA-KIRC stress test supports the same cautionary interpretation. In 198 paired cases with 47 recurrence-like events, a Leibovich-like clinical baseline reaches C-index 0.815 and AUROC 0.825, while TITAN WSI-only reaches C-index 0.617 and AUROC 0.640; CT-only and CT-WSI late-average models remain weaker. Thus, the renal cohort is not used here as an additional success claim. Instead, it motivates the clinical-anchoring stance: frozen image embeddings may contain transferable signal, but image-only forecasting can be dominated by mature, organ-specific clinical risk systems.

4. Discussion

The central message is clinically and methodologically conservative. Frozen CT and WSI foundation embeddings can be converted into postoperative forecasting scores, and a complete 2×2 matrix shows that simple late averaging provides repeatable multimodal gain over matched WSI-only baselines. However, pathologic stage remains the best compact anchor in the lung cohort, and a Leibovich-like baseline dominates image-only embeddings in the renal stress test. The right interpretation is therefore not that images beat clinical variables, but that frozen image embeddings provide scalable candidate forecasts whose incremental value must be tested against clinical anchors, calibration, event fraction, and censoring structure (Van Calster et al., 2019; Vickers & Elkin, 2006).

Why does late average outperform learned score combiners here? The likely reason is bias-variance tradeoff under event scarcity. CT and WSI base scores appear weakly complementary but noisy; a simple average has almost no tuning capacity and can behave as a variance-reducing ensemble (Breiman, 1996). In contrast, stacking must estimate combination weights from limited events, and those weights can become unstable if base predictions are noisy, imperfectly calibrated, or correlated differently across folds. This negative stacking result is therefore an important finding: frozen embeddings reduce the representation-learning burden, but they do not remove the need for event-aware model development.

Several methodological extensions remain future work. Penalized Cox models, survival ensembles, random survival forests, gradient-boosted survival models, and neural sur-

Table 1. Primary in-house DFS benchmark across the complete frozen foundation-model matrix. Values are pooled out-of-fold estimates on 466 eligible patients with 56 two-year DFS events. ΔC is the late-average C-index gain over the matched WSI-only model.

WSI encoder	CT encoder	WSI-only C	WSI-only AUROC	CT-only C	CT-only AUROC	Late avg C	Late avg AUROC	ΔC
TITAN	Pillar-0	0.782	0.804	0.610	0.619	**0.799**	**0.810**	+0.017
TITAN	CT-FM	0.782	0.804	0.562	0.562	0.789	0.801	+0.007
Prov-GigaPath	Pillar-0	0.742	0.762	0.610	0.619	0.777	0.789	+0.035
Prov-GigaPath	CT-FM	0.742	0.762	0.562	0.562	0.758	0.768	+0.016

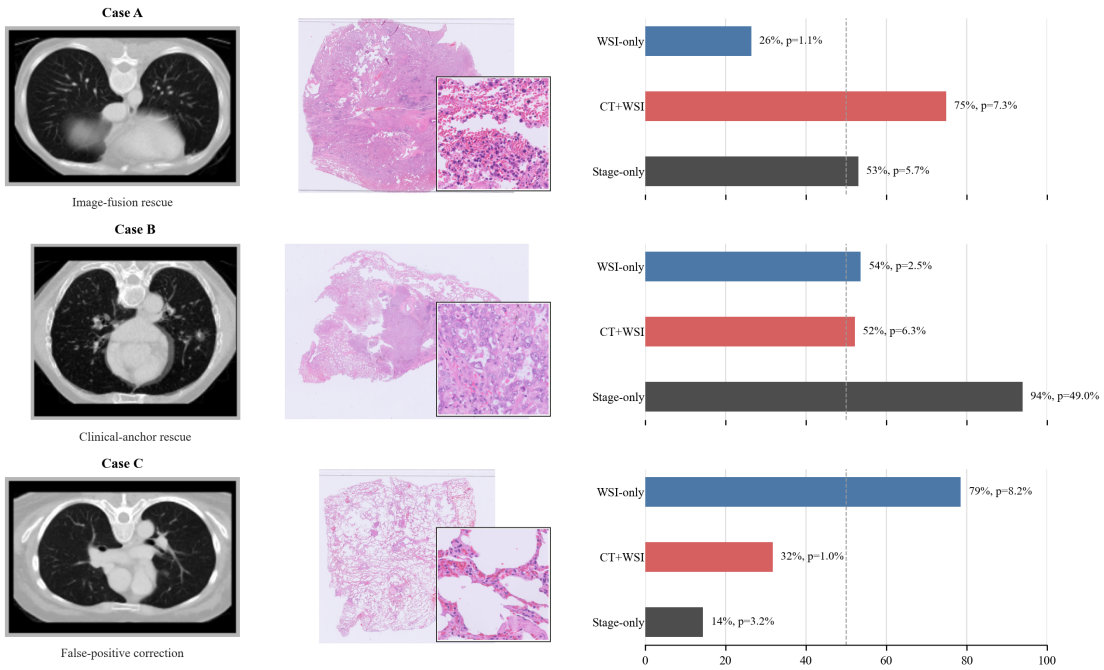

Figure 2. Representative CT, WSI, and out-of-fold risk comparisons for three held-out patients. Each row pairs a representative CT slice, WSI thumbnail, and risk bars from the same case: Case A illustrates image-fusion rescue, Case B clinical-anchor rescue, and Case C false-positive correction. Bars compare WSI-only, CT+WSI late-average fusion, and stage-only predictions; labels report percentile rank and calibrated two-year event probability.

Table 2. Clinical anchors and fusion variants in the primary in-house benchmark. Image rows use TITAN plus Pillar-0.

Model	C-index	AUROC	Brier	PRAUC
Stage-only clinical anchor	**0.837**	**0.840**	0.083	0.406
Extended molecular anchor	0.808	0.826	0.087	0.419
Clinicopathologic core anchor	0.796	0.815	0.084	**0.440**
WSI-only	0.782	0.804	0.103	0.365
CT-only	0.610	0.619	0.115	0.185
Late average	0.799	0.810	0.100	0.346
Late stacking	0.535	0.747	0.097	0.291
Anchor residual	0.577	0.742	0.097	0.294

vival objectives may become useful when the multimodal cohort is larger or externally validated (Hothorn et al., 2006; Ishwaran et al., 2008; Katzman et al., 2018). Representation-level fusion and clinically anchored residual modeling are also promising, but the present benchmark supports a low-capacity score-level strategy and transparent clinical anchoring.

Limitations follow from this design: the primary cohort is single-institution; the two-year DFS event count is modest; CT and WSI are not synchronous measurements; TCGA-KIRC is exploratory rather than external validation; and all encoders are frozen. These constraints sharpen the workshop message: reusable multimodal embeddings require anchoring, calibration, event-maturity sensitivity, and external validation.

5. Conclusion

Frozen multimodal foundation embeddings support feasible CT-WSI postoperative forecasting across multiple model pairings, with late averaging providing the most reliable image-only gain. Their most defensible role is as clinically anchored forecasting components whose incremental value is tested across cohorts, endpoint definitions, and foundation-model backbones.

References

Agrawal, K. K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L. V., Fintelmann, F., Darrell, T., Bai, Y., Chung, M., and Yala, A. Pillar-0: A new frontier for radiology foundation models, 2025. arXiv:2511.17803.

Boeke, D., Blommestijn, C., Wray, R. N., Chupetlovska, K., Gao, S., Gao, Z., Beets-Tan, R. G. H., Crispin-Ortuzar, M., Jones, J. O., Silva, W., and Machado, I. P. Integrating pathology and CT imaging for personalized recurrence risk prediction in renal cancer, 2025. arXiv:2508.21581.

Breiman, L. Bagging predictors. *Machine Learning*, 24: 123–140, 1996. doi: 10.1007/BF00058655.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30: 850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24:1559–1567, 2018. doi: 10.1038/s41591-018-0177-5.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972.

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., Robertson, H., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Chen, C. S., Gerber, G., Le, L. P., and Mahmood, F. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 2025. doi: 10.1038/s41591-025-03982-3.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18):2529–2545, 1999. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2529::AID-SIM274⟩3.0.CO;2-5.

Harrell, F. E., Lee, K. L., and Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(SICI)1097-0258(19960229)15:4⟨361::AID-SIM168⟩3.0.CO;2-4.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and van der Laan, M. J. Survival ensembles. *Biostatistics*, 7 (3):355–373, 2006. doi: 10.1093/biostatistics/kxj011.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:24, 2018. doi: 10.1186/s12874-018-0482-1.

Leibovich, B. C., Blute, M. L., Cheville, J. C., Lohse, C. M., Frank, I., Kwon, E. D., Weaver, A. L., Parker, A. S., and Zincke, H. Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma: a stratification tool for prospective clinical trials. *Cancer*, 97(7):1663–1671, 2003. doi: 10.1002/cncr.11234.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74, 2020. doi: 10.1016/j.ijforecast.2019.04.014.

Moons, K. G. M., Damen, J. A. A. G., Kaul, T., et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, 388: e082505, 2025. doi: 10.1136/bmj-2024-082505.

Ogundimu, E. O., Altman, D. G., and Collins, G. S. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, 76:175–182, 2016. doi: 10.1016/j.jclinepi.2016.02.031.

Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B. H., Fedorov, A., Mak, R. H., and Aerts, H. J. W. L. Vision foundation models for computed tomography, 2025. arXiv:2501.09001.

Peduzzi, P., Concato, J., Feinstein, A. R., and Holford, T. R. Importance of events per independent variable

in proportional hazards regression analysis. II. accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, 48(12):1503–1510, 1995. doi: 10.1016/0895-4356(95)00048-8.

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., and Collins, G. S. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*, 38(7):1276–1296, 2019. doi: 10.1002/sim.7992.

Song, B., Leroy, A., Yang, K., Dam, T., Wang, X., Maurya, H., Pathak, T., Lee, J., Stock, S., Li, X. T., Fu, P., Lu, C., Toro, P., Chute, D. J., Koyfman, S., Saba, N. F., Patel, M. R., and Madabhushi, A. Deep learning informed multimodal fusion of radiology and pathology to predict outcomes in hpv-associated oropharyngeal squamous cell carcinoma. *EBioMedicine*, 114:105663, 2025. doi: 10.1016/j.ebiom.2025.105663.

Steyerberg, E. W., Pencina, M. J., Lingsma, H. F., Kattan, M. W., Vickers, A. J., and Van Calster, B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *European Journal of Clinical Investigation*, 42(2):216–228, 2012. doi: 10.1111/j.1365-2362.2011.02562.x.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17:230, 2019. doi: 10.1186/s12916-019-1466-7.

Vanguri, R. S., Luo, J., Aukerman, A. T., Egger, J. V., Fong, C. J., Horvat, N., Pagano, A. M., Araujo-Filho, J. A. B., Geneslaw, L., Rizvi, H., et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nature Cancer*, 3(10):1151–1164, 2022. doi: 10.1038/s43018-022-00416-8.

Vickers, A. J. and Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361.

Vittinghoff, E. and McCulloch, C. E. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6):710–718, 2007. doi: 10.1093/aje/kwk052.

Wolpert, D. H. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Wulczyn, E., Steiner, D. F., Xu, Z., Sadhwani, A., Wang, H., Flament-Auvigne, I., Mermel, C. H., Chen, P.-H. C., Liu, Y., and Stumpe, M. C. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLOS ONE*, 15(6):e0233678, 2020. doi: 10.1371/journal.pone.0233678.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., Gonzalez, 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., and Poon, H. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Yang, L., Wang, S., Zhou, Y., Lai, S., Xiao, G., Gazdar, A., and Xie, Y. Evaluation of the 7th and 8th editions of the AJCC/UICC TNM staging systems for lung cancer in a large North American cohort. *Oncotarget*, 8(40):66784–66795, 2017. doi: 10.18632/oncotarget.18158.