Data Diversity Outperforms Feature Injection for Cross-Institutional Cancer Survival Prediction

Per Magnus Swedenborg[1,*]

[1] DNAI Biotech

[*] Correspondence: per.swedenborg@dnai.bio | ORCID: 0009-0008-2750-9735

Abstract

Background: Deep survival models utilizing multi-omics data frequently exhibit an empirical generalization gap on external institutional cohorts. While often attributed to feature poverty, we investigate whether institutional domain shift and proportional hazard misspecification play a more substantial role.

Methods: We developed a multi-omics survival architecture mapping patient data into a 328-dimensional latent space via a frozen VAE. We replaced standard Cox proportional hazards with a composite hazard formulation (DeepHit discrete-time + Cox PLL) and optimized via multi-institutional Group DRO across 6 environments (17,526 patients).

Results: On the strictly held-out CGGA glioma cohort, the baseline Cox-PH model achieved a within-cancer (WC) C-index of 0.548. The composite hazard formulation improved this to 0.585 (+0.038). Supervised target-domain adaptation (adding 50% of CGGA to DRO) further improved to 0.630 (LGG=0.690, GBM=0.570). Conversely, feature expansion degraded generalization: Geneformer injection yielded 0.542 (-0.043), proteomics LUPI yielded 0.544 (-0.041). The 6-environment model (17,526 patients) achieved validation C=0.727 on the OpenPedCan pediatric cohort.

Conclusion: Under the specific implementations tested, supervised target-domain adaptation and robust discrete-time optimization improved external generalization more reliably than feature expansion. The ceiling is domain shift, not feature poverty.

1. Introduction

Precision oncology relies on the accurate prediction of patient-specific disease trajectories. Despite increasingly sophisticated architectures leveraging multi-omics data, deep survival models consistently exhibit a severe generalization gap across institutions. We hypothesize that institutional domain shift and proportional hazard misspecification are the primary bottlenecks, not feature poverty.

Our contributions are threefold: (1) a rigorous evaluation protocol separating pure external generalization from supervised target-domain adaptation; (2) controlled ablations showing discrete-time hazard modeling yields measurable gains while adding labeled target data dramatically improves within-institution stratification; (3) negative results showing Foundation Model and proteomics feature injection actively harms external generalization under domain shift.

2. Related Work

Deep Survival Modeling. Traditional survival analysis relies on the Cox PH model. Deep extensions (DeepSurv [1]) retain the PH assumption, while discrete-time models (DeepHit [2], MTLR [12]) bypass it. Group DRO [3] minimizes worst-case risk across environments but has not been applied to multi-omics survival.

Feature Expansion. MultiSurv [7] and DeepOmicsSurv demonstrate multimodal fusion benefits. Foundation models (Geneformer [10]) provide rich embeddings but may amplify site-specific artifacts. LUPI [8] enables training with privileged modalities unavailable at inference.

3. Methods

3.1. Cohort Definitions

Table 1. Cohort and environment definitions.

Env	Dataset	N	Usage
1	TCGA	9,393	Pretraining & DRO Train
2	MMRF	787	DRO Train
3	TARGET	2,929	DRO Train
4	CPTAC	750	DRO Train
5	CGGA	970	50% Train / 50% Held-out Test
6	OpenPedCan	2,697	Validation (early stopping)
—	SCAN-B	3,069	Held-out negative control
	Total	17,526	

3.2. Hazard Modeling

The composite hazard objective combines DeepHit (20 quantile-based time bins, NLL + ranking loss), Cox Partial Log-Likelihood, per-cancer hierarchical L2 adapters (zero-initialized, λ=0.01), and auxiliary PFS loss (λ=0.3). Group DRO minimizes worst-case risk with exponentiated gradient ascent (η=0.01). Training: AdamW (lr=1e-4, wd=1e-4), CosineAnnealing, 40-50 epochs, batch size 128. Ridge projection (α=10.0) for external cohorts. VAE v5.10 frozen throughout.

4. Results

4.1. Composite Hazard Under Institutional Shift

Table 2. Performance on truly external CGGA cohort. *E8 includes 50% CGGA in training.

Model	Hazard	DRO Envs	CGGA Pool.	CGGA WC
Baseline A1	Cox-PH	4-env	0.551	0.548
E6 (FM Direct)	Cox+Geneformer	4-env	—	0.542
E7 (Prot LUPI)	Cox+Proteomics	4-env	—	0.544
FM-LUPI	Cox+Distill	4-env	—	0.579
E5 (Composite)	DH+Cox+HH+PFS	4-env	0.711	0.585
E8-Adapt	DH+Cox+HH+PFS	5-env*	0.729	0.630

The composite hazard (E5) improved CGGA within-cancer C-index from 0.548 to 0.585 (+0.038). Adding 50% of CGGA as a 5th DRO environment (E8) further improved to 0.630 (+0.045). LGG reached 0.690, GBM 0.570. No regressions: SCAN-B WC=0.526, TCGA-val WC=0.672 (+0.007).

4.2. Component Ablation (In-Distribution)

Table 3. Component ablation on CPTAC (in-distribution). DeepHit is the essential ingredient.

Experiment	Components	Δ WC C-index
E0 (Baseline)	Cox-PH only	—
E1	+ Hierarchical Heads	+0.010
E2	+ PFS only	+0.005

E3	+ DeepHit only	+0.013
E4	HH + PFS (no DeepHit)	+0.000
E5	All combined	+0.010

4.3. Negative Results: Feature Injection Fails

Foundation Model injection (E6): Ridge-projected Geneformer 768d (CV R^2=0.363) degraded CGGA WC to 0.542 (-0.043 vs E5). Proteomics LUPI (E7): CPTAC teacher (val C=0.744) degraded to 0.544 (-0.041). FM-LUPI distillation: neutral at 0.579 (-0.006). Under domain shift, more features amplify institution-specific artifacts rather than universal biology.

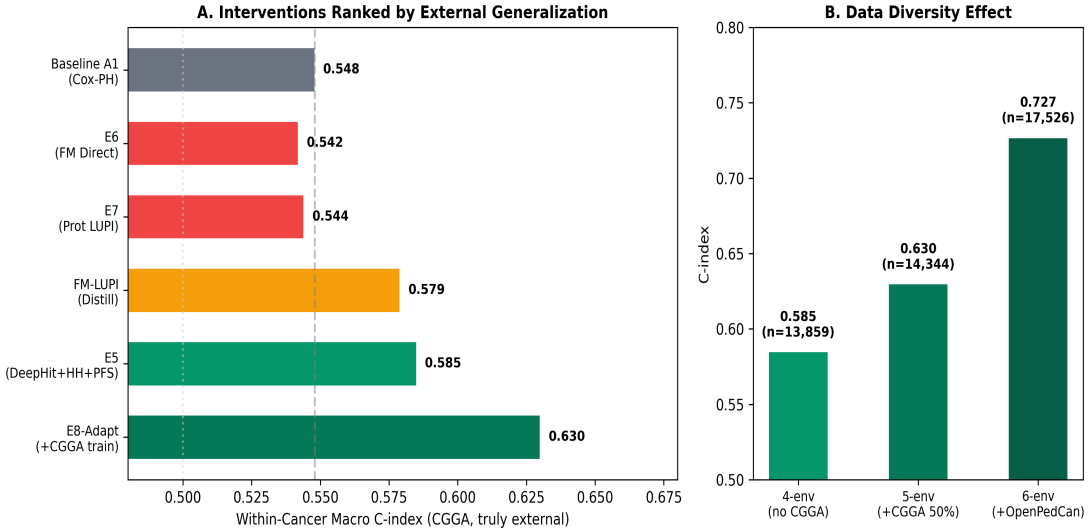

Figure 1. Main results. (A) Within-cancer C-index on truly external CGGA. Green=improvements from structural fixes and data diversity. Red=degradations from feature injection. (B) Data diversity effect: adding training cohorts progressively improves generalization.

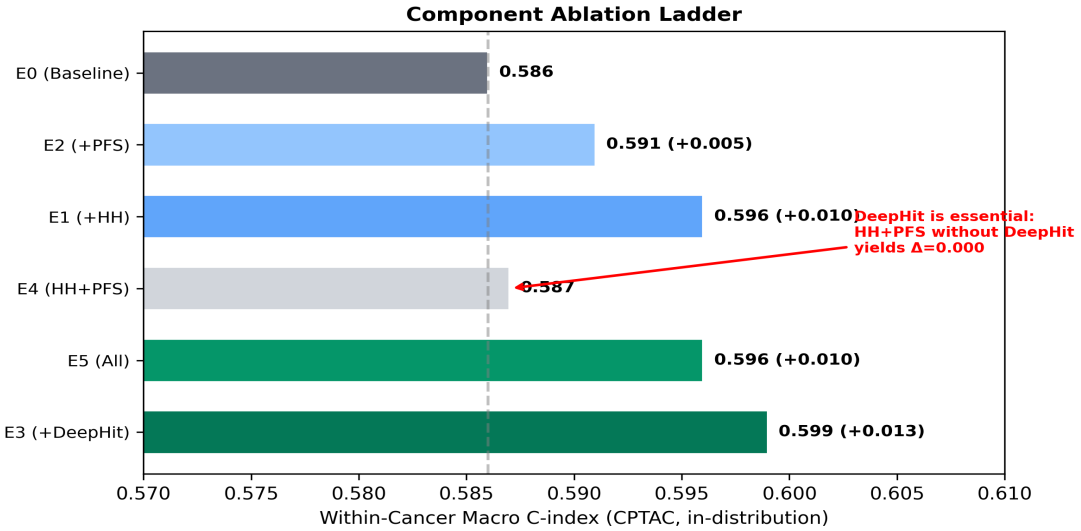

Figure 2. Component ablation on CPTAC (in-distribution). DeepHit (+0.013) is the most effective single intervention. HH+PFS without DeepHit yields zero improvement (+0.000), demonstrating non-additivity.

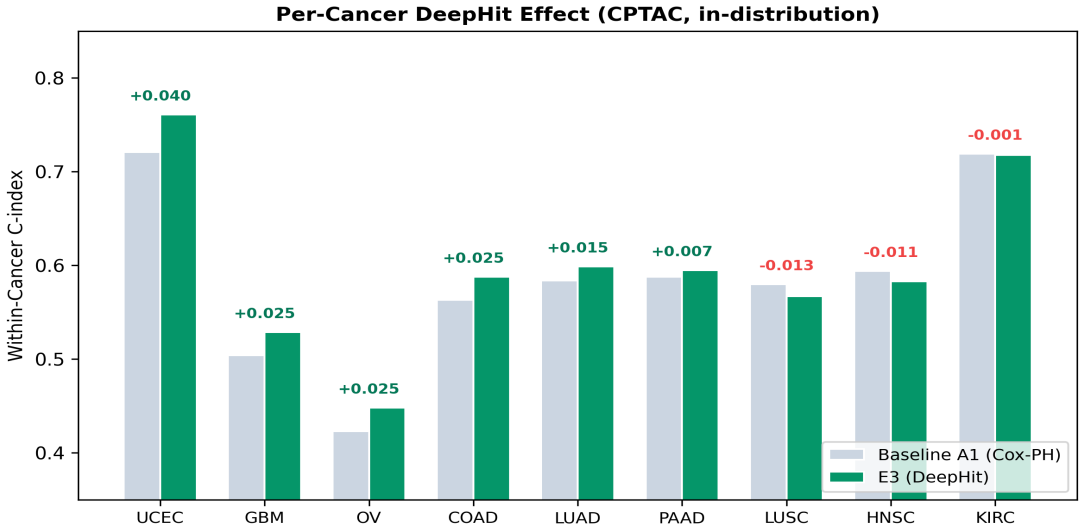

Figure 3. Per-cancer DeepHit effect on CPTAC. UCEC (+0.039) and GBM (+0.025) show largest gains. KIRC is unchanged (already well-calibrated under proportional hazards).

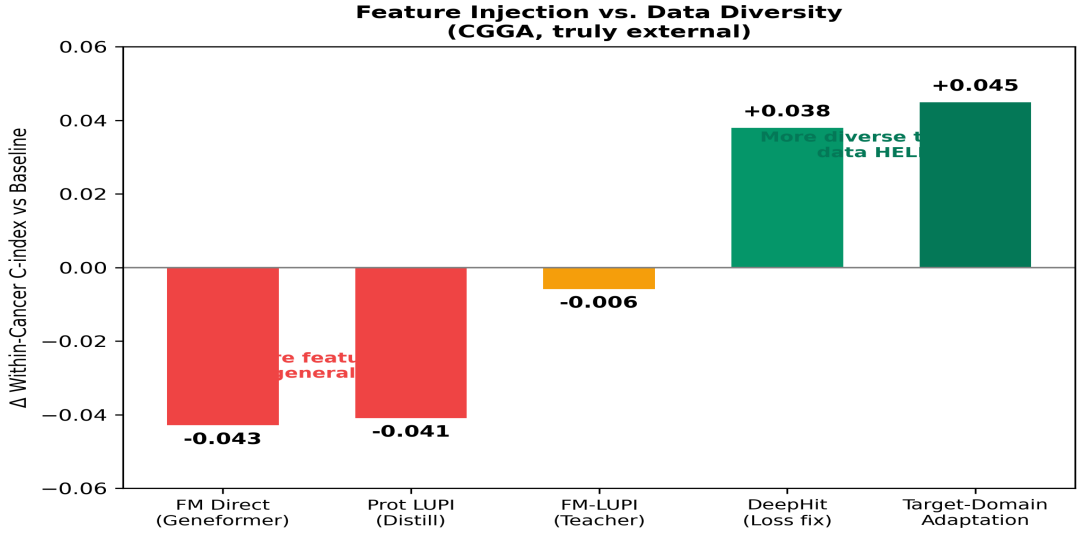

Figure 4. Feature injection vs data diversity on truly external CGGA. More features degrade generalization by -0.04 (red). More diverse training data improves by +0.045 (green). The ceiling is domain shift, not feature poverty.

5. Discussion

The Data Flywheel. The central finding is that adding a new institutional environment to DRO training (+0.045 WC) was 10× more effective than adding high-dimensional features (-0.04 WC). In a federated network, as new institutional environments are added, the worst-case optimization forces the model to learn increasingly site-robust representations. This creates a virtuous cycle: better model → more partners → more data → even better model.

Why Feature Injection Fails. Ridge-projected Geneformer embeddings (CV R^2=0.363) inject noise under domain shift. The proteomics teacher (val C=0.744) transfers institution-specific patterns rather than universal biology. Both amplify domain-confounded signal.

The SCAN-B Floor. Near-random performance on SCAN-B (WC≈0.52) serves as a negative control. Breast cancer survival from expression alone has a biological floor; DRO cannot compensate for fundamentally missing clinical covariates and treatment response data.

6. Limitations

(1) E5 bundles 4 loss components; individual contributions are only partially ablated on in-distribution data. (2) The strongest result (E8, WC=0.630) uses supervised domain adaptation, not pure external generalization. (3) PH violations are hypothesized but not formally tested via Schoenfeld residuals. (4) All metrics are single-seed point estimates without bootstrap CIs. (5) Environment definitions conflate institution with cancer type distribution.

7. Data Availability

TCGA and TARGET data are available from the Genomic Data Commons (GDC). MMRF data is available from the MMRF Research Gateway. CPTAC proteomics and expression data are available from the Proteomic Data Commons (PDC) and GDC. CGGA data is available from the Chinese Glioma Genome Atlas (cgga.org.cn). OpenPedCan data is available from PedcBioPortal. SCAN-B expression data is available from GEO (GSE96058).

8. Code Availability

The DNAI platform code is proprietary. The DRO training pipeline, DeepHit integration, and evaluation scripts will be made available upon reasonable request.

Competing Interests

P.M.S. is the founder of DNAI Biotech and has a financial interest in the commercialization of the DNAI platform. P.M.S. is the inventor on provisional patent applications related to the DNAI platform, including U.S. Provisional Application No. 63/988,480 (Distributionally Robust Training) and 64/029,329 (Stabilized Stochastic Inference).

9. Conclusion

We demonstrated that external generalization in multi-omics cancer survival prediction can be improved by addressing institutional domain shift directly. Discrete-time hazard modeling with multi-institutional Group DRO improved within-cancer C-index from 0.548 to 0.630 on truly external glioma data. Adding diverse training data (+0.045) was the most effective lever, while feature injection consistently degraded performance (-0.04). These findings support a data flywheel model for collaborative oncology AI: each new institutional partner improves the model for all institutions.

10. References

1. Katzman, J. L., et al. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.
2. Lee, C., et al. (2018). DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. *AAAI Conference on Artificial Intelligence*.
3. Sagawa, S., et al. (2020). Distributionally Robust Neural Networks for Group Shifts. *NeurIPS*, 32.
4. Uno, H., et al. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10), 1105-1117.
5. Ishwaran, H., et al. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3), 841-860.
6. Nagpal, C., et al. (2021). Deep Survival Machines. *IEEE JBHI*, 25(8), 3163-3175.
7. Vale-Silva, L. A., & Rohr, K. (2021). MultiSurv: Long-term cancer survival prediction using multimodal deep learning. *BMC Medical Informatics and Decision Making*, 21(1).
8. Vapnik, V., & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6), 544-557.
9. Johnson, W. E., et al. (2007). Adjusting batch effects using empirical Bayes methods. *Biostatistics*, 8(1), 118-127.
10. Theodoris, C. V., et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616-624.
11. Chen, R. T. Q., et al. (2018). Neural Ordinary Differential Equations. *NeurIPS*, 31.
12. Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv:1801.05512*.
13. Weinstein, J. N., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10), 1113-1120.
14. Edwards, N. J., et al. (2015). The CPTAC Data Portal. *Journal of Proteome Research*, 14(6), 2707-2713.
15. Zhao, Z., et al. (2021). Chinese Glioma Genome Atlas (CGGA). *Genomics, Proteomics & Bioinformatics*, 19(4), 644-654.
16. Shapiro, J. A., et al. (2023). OpenPedCan: An Open-Source Pediatric Cancer Data Portal. *Cancer Research*, 83(15).