Domain Shift is a Feature, Not a Bug: Distributionally Robust Optimization Outperforms Harmonization in Clinical Foundation Models

Per Magnus Swedenborg[1,]

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

Abstract

Background: The prevailing dogma in multi-site clinical machine learning is harmonizationremoving site-specific batch effects to learn domain-invariant representations. We challenge this paradigm, demonstrating that in oncology, site-specific variance often encodes critical prognostic information (referral patterns correlating with disease severity) that harmonization erases. **Methods:** We trained a multi-omics survival model on TCGA (n=9,415; 33 cancer types) and evaluated on three strictly held-out external cohorts: CPTAC (n=1,031), CGGA (n=970), and SCAN-B (n=3,069). No external data was used during training, hyperparameter selection, or model development. We compared Empirical Risk Minimization (ERM), harmonization (ComBat, CORAL), and **Group Distributionally Robust Optimization (Group DRO)** with a **Pooled Cox** objective. **Results:** We identified a Calibration Paradox: harmonization successfully aligned feature distributions but degraded discrimination. Conversely, Group-Reweighted Pooled Cox DRO achieved a C-index of **0.718** [0.684, 0.750] on CPTAC and **0.708** [0.688, 0.726] on CGGA, outperforming all baselines. Transportability certificates identified a Green tier achieving C=0.744 on 22.2% of the CPTAC cohort. The model correctly failed on SCAN-B (C=0.501), confirming that breast cancer survival requires histopathology data absent from this cohort. **Conclusion:** Domain shift is a predictive feature. Clinical foundation models should abandon aggressive harmonization in favor of robust optimization that enforces global ranking consistency, gated by per-patient transportability certificates.

1. Introduction

The deployment of clinical foundation models is stalled by a reproducibility crisis, where models trained on high-quality curated datasets (e.g., TCGA) fail to generalize to real-world clinical data. The standard computational response is **harmonization**: applying batch-correction methods like ComBat [1] or feature alignment techniques like CORAL [2] to align distributions across sites. This approach rests on the assumption that site-specific variance is technical noise that obscures biological signal.

We test a counter-hypothesis: **site-specific variance is often a prognostic signal.** In oncology, the domain (hospital) is confounded with patient acuity. Tertiary care centers treat more aggressive, late-stage cancers than community clinics due to referral patterns [3]. Aggressive harmonization removes these latent covariates, effectively erasing the signal of disease severity encoded in the batch.

Using the DNAI platforma physics-constrained digital twin integrating multi-omics and histologywe demonstrate that harmonization fails to improve, and often degrades, external validity. Instead,

we propose **Group Distributionally Robust Optimization (Group DRO)** [4] with a **Pooled Cox** objective. Unlike Stratified Cox, which ranks patients only within their hospital, Pooled Cox forces the model to learn a global risk score valid across institutions.

Our contributions are: 1. **The Calibration Paradox:** We demonstrate that ComBat and CORAL improve distribution alignment metrics (Statistical Rejection Rate) but degrade discrimination, confirming that batch effects contain prognostic signal. 2. **Theoretical Framework for Global Ranking:** We show empirically that **Pooled Cox DRO** outperforms **Stratified Cox DRO** by preventing the model from learning environment-specific shortcuts. 3. **Tiered Generalization:** We validate performance across diverse cohorts (CPTAC, CGGA, SCAN-B), delineating the boundaries of omics-only generalization. 4. **Transportability Certificates:** We validate a traffic-light system that ensures safe deployment, identifying a Green tier of highly reliable predictions (22.2% of external cohort).

---

2. Related Work

Batch Effects and Harmonization: In bioinformatics, removing technical variation is standard (ComBat [1], Harmony [5]). However, recent work in causal inference suggests that when the domain correlates with the label (label shift), invariance learning hurts accuracy [6].

Domain Generalization (DG): DG aims to learn robust features from source domains alone. While Invariant Risk Minimization (IRM) [7] seeks invariant causal mechanisms, we argue that in clinical settings, mechanisms are often *not* invariant due to care-context confounding. Group DRO [4] minimizes the loss of the worst-case group, which we adapt to survival analysis.

Survival Analysis in DG: Most DG work focuses on classification. We extend this to time-to-event data, specifically addressing the distinction between *stratified* partial likelihood (standard in multi-site trials) and *pooled* partial likelihood (necessary for global ranking) [8].

---

3. Methods

3.1 Data, Lockbox Protocol, and Firewall

- **Internal Training (TCGA):** 9,415 patients across 33 cancer types. Groups (g) defined by the intersection of Cancer Type and Tissue Source Site (TSS), yielding 132 environments with $$20 patients each (186 total, merged to $$20 for stable group loss estimation).
- **External Validation 1 (CPTAC): 1,031** patients across 10 cohorts (pan-cancer). Used strictly for final evaluation.
- **External Validation 2 (CGGA): 970** patients (Full Glioma cohort, GBM + LGG).
- **Negative Control (SCAN-B): 3,069** patients (Breast Cancer, RNA-seq only, no WSI). Used to test modality dependence.
- **Total External N:** 5,070 patients across 3 independent cohorts.

Firewall Protocol. No external data was used during any stage of model development: 1. **Training:** Weights optimized on TCGA train split only (7,985 patients). 2. **Validation / early stopping:** TCGA validation split only (1,408 patients). 3. **Hyperparameter selection:** DRO step size (), minimum group size, and learning rate selected on TCGA validation C-index. No external

signal. 4. **Ridge projection:** Fitted on TCGA training split expression latent mapping. External cohorts projected using frozen Ridge weights. 5. **Evaluation:** External cohorts evaluated once after all model decisions were locked. No iterative tuning on external data. 6. **Partition hashes:** SHA-256 hashes of train/val/external sample IDs are recorded for reproducibility (see Code Availability).

3.2 External Inference Pipeline

Since external cohorts lack the exact feature set of the training data, we employ a rigorous projection pipeline: 1. **Feature Intersection:** Identify common features between TCGA and External. 2. **Ridge Projection:** A Ridge regressor (= 10.0) maps external features to the 328-dimensional VAE latent space (z_{full}). **Crucially, this projector is fitted exclusively on the TCGA training split** to prevent data leakage. 3. **Inference:** The projected pseudo-latents are fed to the downstream survival models.

3.3 Baselines: Harmonization and Adaptation

We compare our approach against standard harmonization techniques. Note that while our proposed method is purely inductive (fixed weights), the baselines utilize transductive access to target statistics (unlabeled) for alignment: 1. **Empirical Risk Minimization (ERM):** Standard training minimizing average loss. 2. **ComBat [1]:** Parametric empirical Bayes method to remove batch effects (applied to latent features). 3. **CORAL [2]:** Deep domain adaptation minimizing the distance between feature covariances of source and target. 4. **Affine Calibration:** Simple moment matching (mean/variance) of the target to the source.

3.4 Theoretical Framework: The Geometry of Risk

A critical design choice is the loss function. We contrast **Stratified Cox** vs. **Pooled Cox**.

Stratified Cox (Local Ranking): The loss is the sum of partial likelihoods calculated *within* each group g. The model is never penalized for ranking a high-risk patient in Hospital A lower than a low-risk patient in Hospital B. This allows the model to learn environment-specific calibration (shortcuts).

Pooled Cox (Global Ranking): The risk set $R(t_i)$ includes *all* patients across all groups. This forces the model to learn a globally valid risk score $h(x)$ that is comparable across sites, which is essential for external generalization where site identity is unknown or novel.

Group DRO (Pooled): We minimize the worst-case group loss using the Pooled objective:

$$\min \max_{gG} E_{(x,y)P_g}[L_{pool}(; x, y)]$$

We use exponentiated gradient ascent (step size = 0.01) to update group weights q_g, focusing training on domains where the global ranking is poor.

3.5 Transportability Certificate & ISS

We compute a status S $\{\text{Green}, \text{Yellow}, \text{Red}\}$ for every patient x based on the Information Sufficiency Score (ISS). The ISS is a bounded metric $[0, 1]$ composed of normalized uncertainty and density terms:

$$\text{ISS} = 0.35(1\ U) + 0.35(1\ D) + 0.20C_{data} + 0.10N_{density}$$

Where: * U: MC Dropout variance, min-max normalized to [0,1] based on training distribution. * D: Mahalanobis distance in latent space, normalized via Chi-squared CDF. * C_{data}: Data completeness ratio (0 to 1). * $N_{density}$: Local manifold density (0 to 1).

Thresholds: Green if $D_M\ _{OOD}$ and $ISS\ 0.7$; Red if $D_M >\ _{OOD}$ (where $_{OOD} = 10.31$).

3.6 Statistical Analysis

- **Performance Metric:** Harrells C-index (concordance index) for censored data.
- **Uncertainty:** 95% Confidence Intervals (CI) derived from 1,000-sample patient-level bootstrapping, stratified by cancer type.
- **Hypothesis Testing:** Log-rank test for survival curve separation.

---

4. Experiments & Results

4.1 The Calibration Paradox: Alignment Improves, Discrimination Worsens

We tested whether harmonization improves external validity. Table 1 shows the Calibration Paradox: methods that aggressively align distributions (ComBat, CORAL) reduce the Statistical Rejection Rate significantly but degrade the C-index.

Table 1: The Calibration Paradox (CPTAC Pooled Validation)

Method	Statistical Rejection Rate* ()	Mean Shift ()	C-index (95% CI)
Baseline (ERM)	93.3%	0.84	0.569 [0.534, 0.602]
ComBat (Harmonization)	12.4%	0.31	0.551 [0.522, 0.580]
CORAL (Adaptation)	**0.5%**	**0.15**	0.531 [0.505, 0.560]

**Note: Statistical Rejection Rate is defined as the percentage of external samples distinguishable from the training distribution via a Kolmogorov-Smirnov test ($p < 0.05$) on latent features. This differs from the clinical safety tiers in Table 4.*

Mechanism: In oncology, the batch (hospital) is often a proxy for patient acuity (referral bias). By forcing the feature statistics of a high-acuity tertiary center to match those of a general hospital via ComBat or CORAL, we strip away the latent covariate of disease severity, destroying prognostic signal.

4.2 Mechanism: Why Pooled Cox DRO Wins

To validate our theoretical framework, we compared Stratified vs. Pooled objectives.

Table 2: Objective Function Ablation (CPTAC Pooled)

Objective	Optimization	C-index (95% CI)	Interpretation
Stratified Cox	ERM	0.588 [0.560, 0.615]	Local ranking only
Stratified Cox	Group DRO	0.623 [0.595, 0.650]	Robust local ranking
Pooled Cox	**Group DRO**	**0.718 [0.684, 0.750]**	**Robust global ranking**

Result: Stratified DRO (0.623) optimizes within-site rank, failing to learn a global baseline. Pooled DRO (0.718) forces the model to rank patients correctly *across* disparate environments. This confirms that for a foundation model, the goal is global ordering, not just local discrimination.

4.3 External Validation & Negative Controls

We evaluated the best model (Pooled Group DRO) across diverse external cohorts.

Table 3: External Validation by Cohort

Cohort	N	Baseline (ERM) C	Group DRO C (95% CI)		Status
CPTAC (Pan-Cancer)	**1,031**	0.569	**0.718 [0.684, 0.750]**	+14.9pp	Success
CGGA (Glioma)	**970**	0.661	**0.708 [0.688, 0.726]**	+4.7pp	Success
SCAN-B (BRCA)	**3,069**	0.504	0.501 [0.465, 0.533]	-0.3pp	Fail (Expected)

Negative Result: The model fails on SCAN-B (C=0.501, near random). This is expected and informative: SCAN-B contains only RNA-seq (no WSI). In breast cancer, survival is heavily driven by grade and spatial histology, which are missing here. Internally, BRCA omics-only C-index is 0.365 [0.256, 0.503] below random confirming that BRCA is fundamentally WSI-dependent in our architecture. This validates that the model is not hallucinating performance and that the CPTAC/CGGA gains are genuine.

4.5 Transportability Certificate: Green Tier Performance

Within the CPTAC cohort, the ISS-based certificate identifies a high-confidence subset:

Table 5: Green Tier Discrimination

Cohort	Green N (%)	Green C (95% CI)	Green vs Red OS	p-value
CPTAC	229 (22.2%)	**0.744 [0.692, 0.793]**	+310 days	8.6e-7
CGGA	596 (61.4%)	**0.675 [0.643, 0.706]**	+992 days	7.2e-40
SCAN-B	0 (0%)	N/A	N/A	N/A

The Green tier achieves clinical-grade discrimination (C>0.65) on both CPTAC and CGGA. SCAN-B correctly receives zero Green certificates (all YELLOW/RED), confirming that the certificate framework does not issue false confidence. Across all external cohorts, 20.1% of patients (1,421/7,071 including internal validation) achieve Green status with C>0.6.

4.6 The Two-Checkpoint Strategy

An unexpected finding is that no single checkpoint optimizes both internal and external performance:

Table 6: Checkpoint Strategy

Checkpoint	Internal Global C	CPTAC External C	CGGA External C	Best For
Phase 0 (CATE)	**0.704**	0.569	**0.718**	Internal + single-disease external
Group DRO	0.727	**0.718**	0.708	Multi-site external

We recommend a **dual-checkpoint deployment**: the Phase 0 checkpoint for internal use and single-disease cohorts (where the same cancer type dominates), and the DRO checkpoint for multi-site external generalization. The certificate framework operates identically on both checkpoints.

4.4 Transportability Certificate Coverage

We applied the ISS-based certificate to the CPTAC cohort to identify safe-to-predict patients. Unlike the statistical rejection rate in Table 1 (which measures strict distribution identity), these tiers measure *informational sufficiency* for clinical decision support.

Table 4: Certificate Stratification (CPTAC Cohort)

Status	Definition	Count (N)	Percentage
Green	D_M , ISS 0.7	**229**	**22.2%**
Yellow	D_M , $ISS < 0.7$	456	44.2%
Red	$D_M >$	346	33.6%

Note: The Green tier represents the subset of patients where the models internal representation is both statistically in-distribution and informationally sufficient.

---

5. Discussion

Our findings challenge the harmonization is always good heuristic that dominates multi-site clinical AI. The **Calibration Paradox** (Table 1) reveals the mechanism: when we force the latent statistics of disparate hospitals to align, we remove the context that implies disease severity. In oncology,

hospital identity is a proxy for referral patterns, care intensity, and patient acuity all prognostic signals that harmonization erases.

The superiority of **Pooled Cox DRO** over Stratified Cox (Table 2) is the theoretical linchpin. Stratified models learn to rank well *within* a hospital but fail to establish a global baseline they can exploit environment-specific shortcuts (e.g., learning that Hospital A patients are generally sicker without learning *why*). Pooled DRO forces the model to confront cross-site comparisons, resulting in globally valid risk scores. The +9.5pp gap between Pooled DRO (0.718) and Stratified DRO (0.623) on CPTAC demonstrates that this distinction is not merely theoretical.

The Two-Checkpoint Strategy (Table 6) is an unexpected practical finding. No single training objective dominates across all deployment contexts. The Phase 0 checkpoint excels internally and on single-disease external cohorts (CGGA C=0.718 for glioma), while DRO excels on multi-site pan-cancer evaluation (CPTAC C=0.718). This suggests that the optimal robustness-discrimination trade-off depends on the deployment context, and that clinical AI systems may need checkpoint selection as part of their deployment specification.

The failure on **SCAN-B** (Table 3, C=0.501, n=3,069) is a critical validation of limits. It confirms that for breast cancer, RNA alone is insufficient for prognosis without spatial histology (WSI), consistent with prior METABRIC analyses showing BRCA omics-only C=0.365 internally. The certificate framework correctly assigns zero Green certificates to SCAN-B, demonstrating that the transportability system does not issue false confidence.

Generalizability beyond this platform. While we demonstrate DRO using a specific multi-omics architecture, the core finding that Pooled Cox DRO outperforms harmonization for external generalization applies to any survival model trained on multi-site data. The Group DRO algorithm requires only group labels (TSS in our case) and a differentiable loss function. We encourage adoption in other clinical foundation models regardless of input modality.

---

6. Limitations

1. **Ridge Projection Ceiling:** External validation relies on Ridge-projected pseudo-latents (Rš=0.638 on TCGA cross-validation), which introduces an information bottleneck. Direct VAE encoding of external RNA-seq is possible but produces higher-entropy latents due to missing DNA/CNV/Meth modalities (24% of the 328d latent is unobserved). The Ridge approach trades fidelity for stability.
2. **Heuristic ISS Weights:** The ISS component weights (0.35/0.35/0.20/0.10) are empirically derived from validation performance. While we demonstrate robustness across three external cohorts, the optimal weights may differ for new modalities or deployment contexts. Per-cancer PolicyProfiles partially address this by adjusting abstention thresholds.
3. **Retrospective Validation:** Despite a strict lockbox protocol with SHA-256 partition hashes, all evaluations are retrospective. Prospective validation on data collected after model freeze is required to confirm clinical utility.
4. **Single Training Dataset:** The model was trained exclusively on TCGA. While DRO improves robustness to distribution shift, it cannot address shifts in the underlying biology (e.g., novel cancer subtypes not represented in TCGA).
5. **Modality Dependence:** The SCAN-B failure demonstrates that omics-only generalization has cancer-type-specific limits. For WSI-dependent cancers (BRCA, CESC), external valida-

tion requires matched histopathology.

6. **ComBat/CORAL Implementation:** Harmonization baselines were applied in latent space (post-Ridge projection), not in gene expression space. Performance of gene-level ComBat may differ, though prior work suggests similar limitations when batch correlates with outcome [6].

---

7. Code and Data Availability

Code: All training and evaluation scripts, including the Group DRO implementation and ISS calculation, are available from the corresponding author upon reasonable request. **Data:** * **TCGA:** Available via the GDC Data Portal (dbGaP phs000178). * **CPTAC:** Available via the CPTAC Data Portal. * **CGGA:** Available via the Chinese Glioma Genome Atlas (http://www.cgga.org.cn). * **SCAN-B:** Available via NCBI GEO (GSE96058). **Models:** Pre-trained VAE v5.10 and Hypernet v3.2 weights are available upon request for non-commercial research.

Ethics Statement

This study exclusively utilized de-identified, publicly available retrospective datasets (TCGA, CPTAC, CGGA, SCAN-B). All datasets were accessed in accordance with their respective data use agreements. No patient contact, intervention, or collection of new human biological material was performed. As all data were previously collected, de-identified, and publicly released under institutional review, additional IRB approval was not required per the Common Rule (45 CFR 46.104(d)(4)).

Author Contributions (CRediT)

P.M.S.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing Original Draft, Writing Review & Editing, Visualization.

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 for Cross-Institutional Generalization).

---

8. Conclusion

We present evidence across 5,070 external patients that domain shift is a predictive feature in clinical foundation models. By abandoning harmonization in favor of **Group-Reweighted Pooled Cox DRO**, we achieved C-indices of **0.718** [0.684, 0.750] on CPTAC and **0.708** [0.688, 0.726] on CGGA, with a transportability-certified Green tier reaching **0.744** on 22.2% of the CPTAC cohort. The model correctly abstains on SCAN-B (C=0.501), demonstrating calibrated self-awareness. We recommend: (1) a dual-checkpoint deployment strategy matching robustness profile to deployment context, and (2) per-patient transportability certificates gating clinical output by information sufficiency, ensuring that AI-assisted oncology is both robust and transparent about its limits.

9. References

1. Johnson, W. E., et al. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*.
2. Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation Alignment for Deep Domain Adaptation. *ECCV*.
3. Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*.
4. Sagawa, S., et al. (2019). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *ICLR*.
5. Korsunsky, I., et al. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*.
6. Zhao, H., et al. (2019). On Learning Invariant Representations for Domain Adaptation. *ICML*.
7. Arjovsky, M., et al. (2019). Invariant Risk Minimization. *arXiv*.
8. Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.