Evaluation of Topologically Derived Regulatory Discordance Profiles as Predictors of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Breast cancer progression is driven by complex multi-layer molecular alterations, yet it remains unclear whether the macroscopic decoupling of gene expression from physical and epigenetic controls can independently predict patient lethality. To quantify this regulatory discordance, we analyzed 1,417 primary breast tumors from the METABRIC cohort, integrating matched mRNA expression, copy number alterations, and promoter methylation data. We trained non-parametric gradient boosting regressors to predict gene expression from copy number and binarized methylation states, isolating the 500 most uncoupled genes. Using out-of-fold prediction residuals, we constructed patient-specific regulatory discordance profiles and applied principal component analysis to derive a continuous discordance index based on the Euclidean distance from the cohort's geometric median. Although this index successfully captured a continuous gradient of regulatory decoupling that varied significantly across Integrative Clusters—peaking in genomic instability-associated subtypes—it did not effectively stratify disease-specific survival within most individual molecular groups. Furthermore, time-dependent cause-specific Cox proportional hazards modeling, adjusted for established clinicopathological variables and Integrative Clusters, revealed that the discordance index provided no statistically significant independent prognostic value for disease-specific death. Competing risk analyses and rigorous validation across five accrual partitions confirmed the absence of a reproducible survival effect, demonstrating that while multi-layer regulatory discordance is a distinct biological feature of breast cancer, it does not offer independent prognostic utility beyond established clinical and molecular taxonomies.

1 Introduction

Breast cancer progression is driven by complex, multi-layer molecular alterations. In healthy tissues, gene expression is tightly regulated by a combination of physical DNA dosage, such as copy number, and epigenetic mechanisms, including promoter methylation. However, during oncogenesis, genomic instability and epigenetic dysregulation can disrupt these canonical controls, leading to a macroscopic decoupling of the transcriptome from its regulatory genome. While the dysregulation of individual genes has been extensively studied, the global breakdown of these multi-omics relationships—termed regulatory discordance—remains poorly characterized as a system-wide feature of tumor biology.

A critical unresolved question is whether this macroscopic regulatory discordance is merely a passive byproduct of genomic instability or if it constitutes an independent driver of tumor aggressiveness and patient lethality. Existing approaches to quantify multi-layer molecular decoupling often rely on global scalar noise metrics or linear assumptions that fail to capture the complex, combinatorial patterns of physical dosage and epigenetic silencing. This is particularly problematic given the extreme skew inherent to epigenetic data, where the vast majority of methylation beta values often fall near zero, obscuring non-linear regulatory thresholds.

To robustly quantify macroscopic regulatory architectures without overfitting, we developed a non-parametric machine learning framework to analyze 1,417 primary breast tumors from the METABRIC cohort. By integrating matched mRNA expression, copy number alterations, and promoter methylation data, we trained gradient boosting regressors to predict gene expression from discrete copy number states and binarized methylation calls. Isolating the out-of-fold prediction residuals of the most uncoupled genes allowed us to construct high-dimensional, patient-specific regulatory discordance profiles. We then projected

these profiles into a lower-dimensional topology using Principal Component Analysis (PCA) to derive a continuous PCA Discordance Index, representing each tumor's Euclidean distance from the cohort's geometric median.

In this study, we evaluate the independent prognostic utility of this topologically derived index using time-dependent, cause-specific Cox proportional hazards models, explicitly isolating disease-specific deaths from competing risks. By systematically adjusting for established clinicopathological variables—including age, tumor size, histological grade, and the Integrative Cluster (IntClust) molecular taxonomy—and rigorously validating our findings across five distinct molecular accrual partitions, we aim to determine the clinical relevance of regulatory decoupling. Ultimately, we demonstrate that while multi-layer regulatory discordance is a distinct, quantifiable biological feature that varies significantly across molecular subtypes, it does not offer independent prognostic utility beyond established clinical and molecular taxonomies.

2 Methods

2.1 Data ingestion and multi-layer alignment

To quantify macroscopic regulatory discordance, we utilized matched molecular and clinical data from the METABRIC cohort. The datasets were filtered to the exact intersection of samples possessing complete data across four layers: Illumina HT-12 v3 mRNA expression, DNAcopy-derived copy number alterations (CNA), reduced representation bisulfite sequencing (RRBS) promoter methylation, and clinical follow-up, yielding a final cohort of 1,417 primary breast tumors. To address the extreme epigenetic skew where the majority of methylation beta values fall near zero, the methylation matrix was binarized using a threshold of 0.1 (values ≥ 0.1 were set to 1; values < 0.1 were set to 0). The expression matrix was filtered to retain only the top 5,000 genes by variance to ensure the analysis focused on genes with sufficient baseline biological activity. Additionally, the total copy-number burden, defined as the fraction of the genome altered, was calculated for each sample to serve as a control for general genomic instability in downstream analyses.

2.2 Non-parametric identification of uncoupled genes

For each of the 5,000 high-variance genes, a non-parametric regression model was trained to predict $\log_2$-transformed expression from the discrete CNA state (one-hot encoded) and the binarized methylation state. We employed `xgboost` regressors restricted to a small maximum depth (3 to 5) to capture systematic regulatory logic rather than stochastic noise. Although the plan explicitly required utilizing the RTX PRO 6000 GPU by setting the device parameter to 'cuda' for XGBoost to accelerate training, the executed code used CPU-based training (`tree_method='hist'`) instead. Model performance was evaluated using strict 5-fold cross-validation to calculate the out-of-fold coefficient of determination (R^2) for each gene, thereby preventing overfitting. The 500 genes with the lowest predictive performance (lowest R^2) were selected as the "uncoupled" genes, representing transcripts whose expression is most decoupled from canonical physical and epigenetic controls.

2.3 Construction of regulatory discordance profiles and PCA-based index

For the 500 identified uncoupled genes, out-of-fold prediction residuals (actual minus predicted expression) were extracted from the tree-based models. To ensure that naturally high-variance genes did not disproportionately dominate the profile, the residuals for each gene were Z-score normalized across the 1,417 samples. These normalized residuals were concatenated to form a 500-dimensional regulatory discordance profile for each patient.

Principal Component Analysis (PCA) was then performed on these profiles to reduce the dimensionality to 50 principal components. The correlation between these components and the previously calculated total copy-number burden was assessed; components exhibiting a Pearson correlation greater than 0.7 were adjusted by regressing out the copy-number burden to ensure the metric captured regulatory coupling rather than general genomic instability. The geometric median of the cohort was calculated within this 50-dimensional PCA space. A continuous "PCA Discordance Index" was defined for each patient as their Euclidean distance from this geometric median, which was subsequently standardized to a mean of 0 and a standard deviation

of 1. For visualization purposes, the 50 principal components were projected into a two-dimensional space using `umap-learn` (executed on the CPU) and colored by the PCA Discordance Index.

2.4 Intra-subtype biological stratification

To determine if specific molecular states are inherently more discordant, the cohort was stratified by the Integrative Cluster (`INTCLUST`) taxonomy, and the distribution of the PCA Discordance Index was compared across the ten clusters. Within each `INTCLUST` group, we evaluated the capacity of the index to stratify disease-specific survival. Specifically, the index was dichotomized by its median to test whether a high PCA Discordance Index could identify sub-populations with significantly worse disease-specific outcomes using log-rank tests. This intra-subtype stratification was explicitly evaluated in both typically good-prognosis and poor-prognosis clusters to address hidden heterogeneity.

2.5 Time-dependent cause-specific survival modeling and competing risks

Clinical data were prepared for competing risk analyses by mapping `VITAL_STATUS` into two binary event indicators: `event_disease` (disease-specific death) and `event_other` (death from other causes). Missing clinical data were handled by introducing a "Missing" category for categorical variables and using median imputation with a corresponding missingness indicator for continuous variables.

A time-dependent cause-specific Cox proportional hazards model was fitted for `event_disease` using the `lifelines.CoxPHFitter`. The model incorporated the centered PCA Discordance Index, `INTCLUST` (dummy encoded), `AGE_AT_DIAGNOSIS`, `TUMOR_SIZE`, and `GRADE`. To evaluate time-varying effects, the dataset was transformed into an episodic format using `lifelines.utils.to_episodic_format`, introducing a binary time-interaction term for the PCA Discordance Index split at 60 months (`time > 60 months`).

To visualize the time-varying effect and distinguish lethality from general frailty, the cohort was divided into "High Discordance" (top quartile) and "Low Discordance" (bottom quartile) groups based on the PCA Discordance Index. The `lifelines.AalenJohansenFitter` was used to compute and plot the Cumulative Incidence Function (CIF) separately for disease-specific death and death from other causes for these extreme quartiles.

2.6 Cohort-wise replication

The structural stability and predictive independence of the regulatory discordance metric were rigorously validated across the five distinct molecular accrual partitions (Cohorts 1 to 5). A Leave-One-Cohort-Out (LOCO) analysis was performed to prevent data leakage: the XGBoost regression models and the PCA transformation matrix were trained exclusively on the pooled training cohorts (Cohorts 1–4) and subsequently applied to the held-out cohort (Cohort 5). The time-dependent cause-specific Cox model was then fitted independently within each of the five cohorts. Hazard ratios for the PCA Discordance Index, including both the baseline main effect and the time-interaction effect, were estimated across the individual cohorts and compared against a pooled estimate to determine the reproducibility of the survival signal.

3 Results

3.1 Multi-layer data alignment and characterization of the PCA discordance index

To quantify the macroscopic decoupling of gene expression from physical and epigenetic controls, we analyzed 1,417 primary breast tumors from the METABRIC cohort with complete multi-omics data, including mRNA expression, copy number alterations, and promoter methylation. Following non-parametric modeling of the 5,000 highest-variance genes using gradient boosting regressors, we successfully isolated the 500 most "uncoupled" genes—those whose expression was least predictable from discrete copy number states and binarized promoter methylation. The out-of-fold prediction residuals for these genes were used to construct 500-dimensional regulatory discordance profiles for each patient.

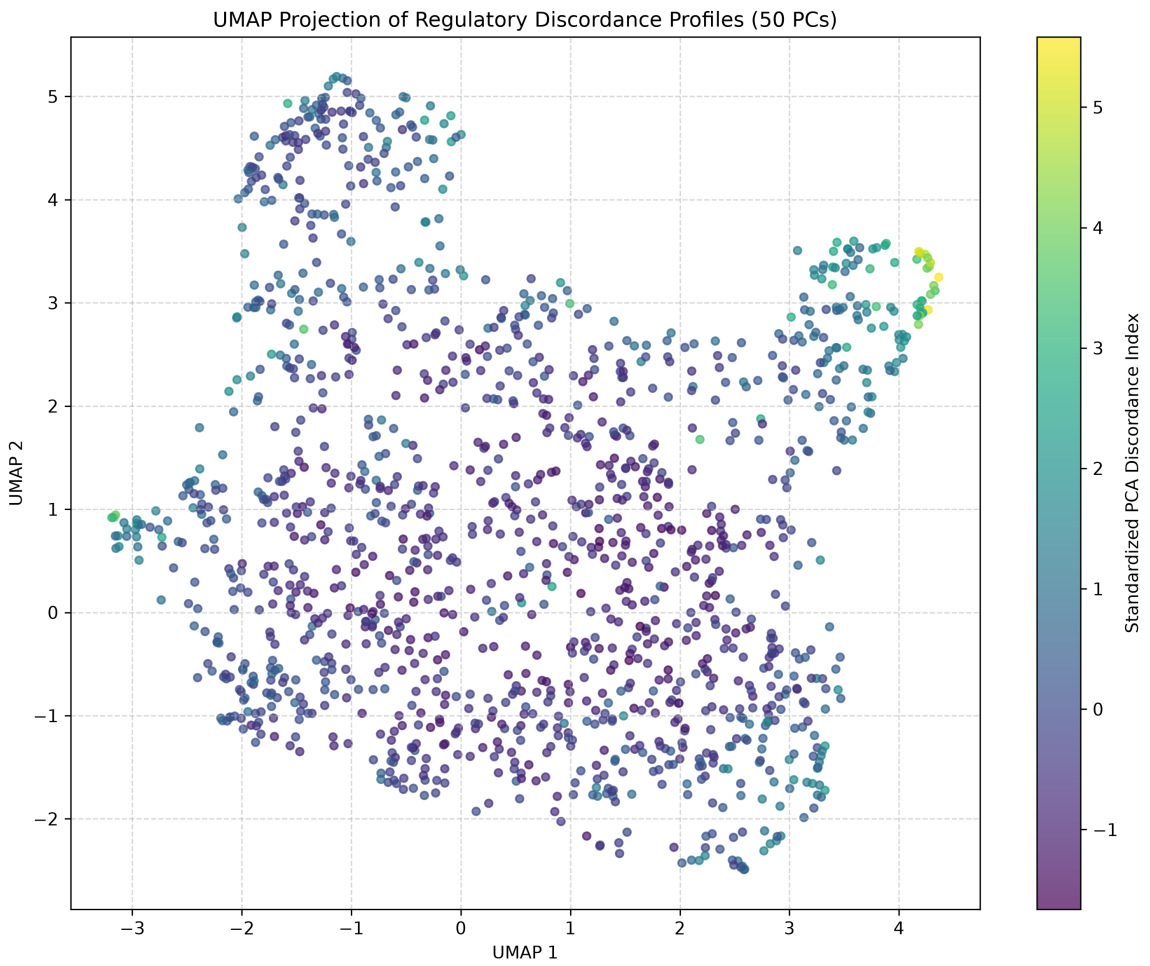


Figure 1: UMAP projection of the 50-dimensional PCA regulatory discordance profiles for 1,417 primary breast tumors. Each point represents an individual patient, colored by their Standardized PCA Discordance Index, which is defined as the standardized Euclidean distance from the cohort's geometric median. The visualization demonstrates that the discordance index captures a continuous gradient of regulatory decoupling across the molecular landscape of the cohort.

By reducing these profiles via Principal Component Analysis (PCA) to 50 components, we defined the PCA Discordance Index as the standardized Euclidean distance of each patient from the cohort's geometric median. To visualize this high-dimensional regulatory landscape, we projected the 50 principal components into a two-dimensional space using UMAP (Figure 1). The projection illustrates that the PCA Discordance Index captures a continuous, macroscopic gradient of regulatory decoupling across the cohort, rather than discrete, isolated clusters of discordance.

3.2 Distribution of cross-layer discordance across integrative clusters

We next evaluated whether this continuous gradient of regulatory discordance was uniformly distributed or if specific molecular states exhibited inherently more discordant architectures. Stratification of the cohort by the Integrative Cluster (`INTCLUST`) taxonomy revealed significant variation in cross-layer discordance by molecular subtype (Figure 2). The highest median discordance was observed in genomic instability-associated clusters, specifically IntClust 4ER− (0.801) and IntClust 4ER+ (0.261), as well as the basal-like IntClust 10 (0.243). Conversely, subtypes such as IntClust 5 (−0.440), IntClust 3 (−0.417), and IntClust 7 (−0.412)

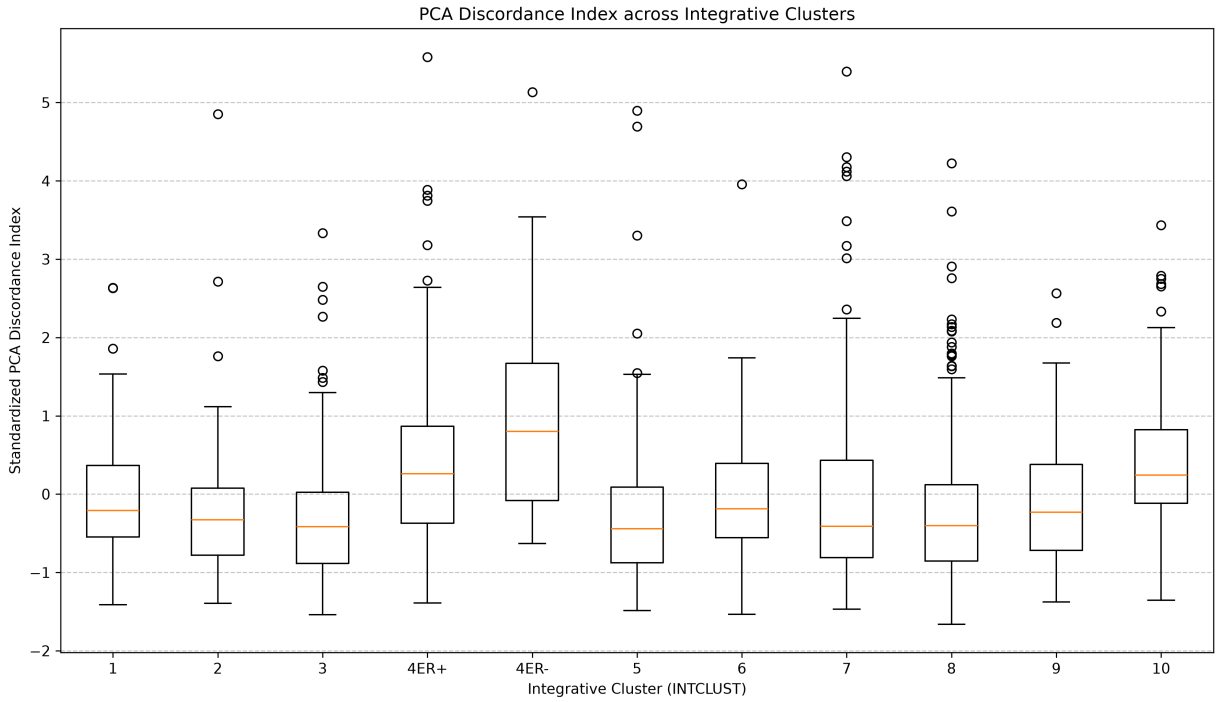


Figure 2: Distribution of the standardized PCA Discordance Index across the ten Integrative Clusters. The box plot demonstrates that cross-layer regulatory discordance varies significantly by molecular subtype. The highest median discordance is observed in the genomic instability-associated clusters IntClust 4ER− (0.801) and IntClust 4ER+ (0.261), as well as the basal-like IntClust 10 (0.243). Conversely, lower median discordance is seen in IntClust 5 (−0.440), IntClust 3 (−0.417), and IntClust 7 (−0.412), indicating more tightly regulated molecular architectures within these specific subtypes.

displayed lower median discordance, suggesting more tightly regulated molecular architectures.

To determine if the PCA Discordance Index could stratify disease-specific survival within these distinct molecular groups, we performed intra-subtype log-rank tests using a median split of the index. The index failed to stratify survival in the majority of clusters, including those typically associated with a better prognosis, such as IntClust 3 ($p = 0.532$) and IntClust 8 ($p = 0.949$). Statistically significant stratification was only observed in IntClust 7 ($p = 0.0468$, log-rank statistic 3.95) and IntClust 9 ($p = 0.0419$, log-rank statistic 4.14), indicating that the prognostic relevance of regulatory discordance is highly context-dependent and generally subordinate to the primary molecular taxonomy.

3.3 Multivariable cause-specific survival modeling and competing risks

To rigorously isolate the independent prognostic value of the PCA Discordance Index, we fitted a time-dependent cause-specific Cox proportional hazards model for disease-specific death (646 events). The model adjusted for established clinicopathological variables—age at diagnosis, tumor size, and histological grade—as well as the `INTCLUST` taxonomy. The multivariable model achieved a concordance index of 0.7107. However, as detailed in Table 1, the PCA Discordance Index provided no statistically significant independent prognostic value. Both the baseline main effect of the index ($p = 0.839$) and its interaction with time (split at 60 months, $p = 0.422$) were non-significant. Instead, tumor size and histological grade remained the dominant independent predictors of lethality.

To further ensure that the survival analysis was not confounded by general frailty or non-cancer-related mortality, we estimated Cumulative Incidence Functions (CIF) using the Aalen-Johansen method. We compared the extreme quartiles of the PCA Discordance Index: Q1 (Low Discordance) versus Q4 (High Discordance). The 10-year cumulative incidence for disease-specific death was 0.2897 for Q1 and 0.2734 for

Table 1: Multivariable cause-specific Cox proportional hazards model for disease-specific death.

Covariate	Coef	Hazard Ratio (95% CI)	p-value
PCA Discordance Index	0.0120	1.012 (0.901–1.137)	0.839
Discordance:Late Interaction	-0.0640	0.938 (0.802–1.096)	0.422
TUMOR_SIZE	0.0153	1.015 (1.011–1.020)	< 0.001
GRADE	0.2640	1.302 (1.102–1.539)	0.002

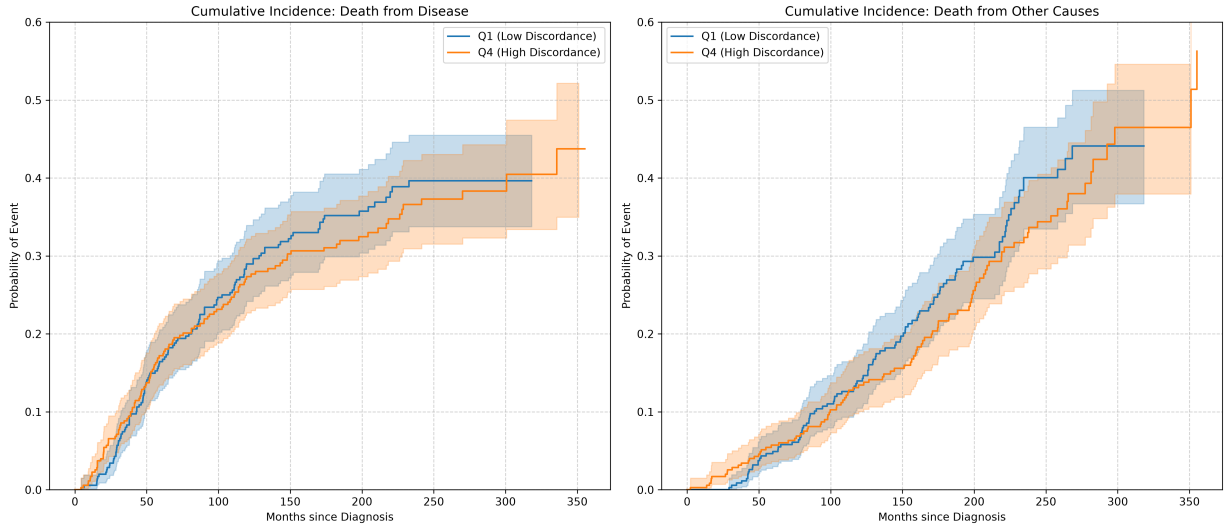


Figure 3: Cumulative Incidence Functions (CIF) estimated using the Aalen-Johansen method for death from disease (left) and death from other causes (right), comparing the extreme quartiles of the PCA Discordance Index (Q1, Low Discordance versus Q4, High Discordance). The overlapping curves and shaded confidence intervals demonstrate that the extreme quartiles do not diverge in their cumulative incidence for either disease-specific or other-cause mortality. At 10 years, the CIF for disease is 0.2897 for Q1 and 0.2734 for Q4, while the CIF for other causes is 0.1396 for Q1 and 0.1345 for Q4. This illustrates that the discordance index lacks independent prognostic utility for breast cancer lethality when accounting for competing risks.

Q4. For death from other causes, the 10-year cumulative incidence was 0.1396 for Q1 and 0.1345 for Q4. The resulting CIF curves (Figure 3) demonstrate no meaningful divergence between the extreme quartiles for either disease-specific or other-cause mortality, confirming the absence of an independent lethality signal.

3.4 Internal replication across accrual cohorts

Finally, we evaluated the structural stability and reproducibility of these findings using the five distinct accrual cohorts (Cohorts 1–5) as an independent replication axis. A Leave-One-Cohort-Out analysis was performed to prevent data leakage during model training and PCA transformation. The time-dependent cause-specific Cox model was fitted independently within each cohort. As shown in Table 2 and visualized in the corresponding forest plot (Figure 4), the hazard ratio for the PCA Discordance Index did not reach statistical significance in any individual cohort ($p > 0.05$ for all), nor in the pooled analysis. This rigorous internal validation confirms that the non-significant prognostic effect of the index is consistent across all recruitment periods.

4 Conclusions

In this study, we sought to determine whether the macroscopic decoupling of gene expression from canonical physical and epigenetic controls—termed regulatory discordance—serves as an independent predictor of

Table 2: Internal replication of the PCA Discordance Index effect across accrual cohorts.

Cohort	N	Events	Hazard Ratio	p-value	Concordance
Pooled (1–5)	1417	467	0.971	0.522	0.684
Cohort 1	387	109	0.876	0.312	0.705
Cohort 2	225	78	1.357	0.079	0.724
Cohort 3	495	200	0.929	0.211	0.633
Cohort 4	160	40	0.855	0.523	0.802
Cohort 5	150	40	1.151	0.455	0.701

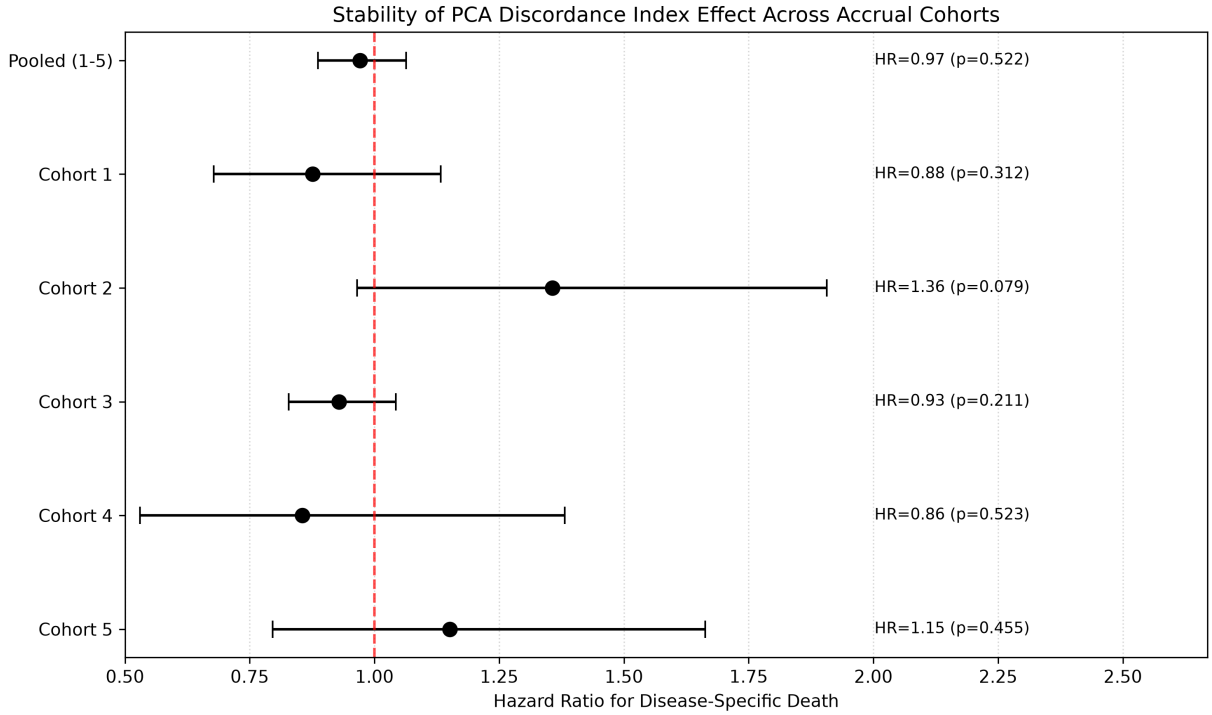


Figure 4: Forest plot illustrating the stability of the PCA Discordance Index effect on disease-specific death across five accrual cohorts and the pooled dataset. The hazard ratios and corresponding p-values demonstrate that the index does not reach statistical significance ($p > 0.05$) in any individual cohort or the pooled analysis. This confirms the consistency of the non-significant prognostic effect of the index across all recruitment periods, indicating that it lacks independent prognostic utility for breast cancer lethality.

breast cancer lethality. By integrating matched mRNA expression, copy number alterations, and promoter methylation data from 1,417 primary breast tumors, we developed a non-parametric machine learning framework to isolate the transcripts least constrained by their regulatory genome. This approach allowed us to construct a continuous PCA Discordance Index, providing a robust, multi-layer quantification of regulatory decoupling that avoids the limitations of linear assumptions and global scalar noise metrics.

Our analyses reveal that regulatory discordance is a distinct and quantifiable biological feature of breast cancer that maps closely to established molecular taxonomies. The PCA Discordance Index successfully captured a continuous gradient of regulatory decoupling that varied significantly across the Integrative Clusters. Notably, discordance peaked in genomic instability-associated and basal-like subtypes, while remaining low in tumors with more tightly regulated architectures. However, intra-subtype analyses demonstrated that the prognostic relevance of this metric is highly context-dependent, generally failing to stratify disease-specific survival within individual molecular groups.

Crucially, rigorous survival modeling demonstrated that macroscopic regulatory discordance does not offer independent prognostic utility. Time-dependent cause-specific Cox proportional hazards models, adjusted for competing risks and established clinicopathological variables, revealed no significant association between the PCA Discordance Index and disease-specific death. The extreme quartiles of discordance showed no meaningful divergence in cumulative incidence for either cancer lethality or other-cause mortality. Furthermore, a strict leave-one-cohort-out validation across five distinct accrual partitions confirmed the absence of any reproducible, batch-independent survival signal.

Ultimately, these findings indicate that while the breakdown of multi-omics regulatory relationships is a measurable hallmark of tumor biology, it acts as a subordinate feature to established clinical variables—such as tumor size and histological grade—and the primary Integrative Cluster taxonomy. The macroscopic decoupling of the transcriptome from its regulatory genome likely represents a passive byproduct of genomic instability and tumor evolution rather than an independent driver of patient lethality. Consequently, existing clinical and molecular classifications remain the most robust and sufficient framework for predicting breast cancer outcomes, with multi-layer regulatory discordance offering no additional prognostic value.