Evaluation of Subtype-Normalized Non-Linear Regulatory Discordance as a Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Genomic instability in breast cancer disrupts gene expression, but the prognostic significance of macroscopic regulatory uncoupling, independent of total aneuploidy and molecular subtype, remains unclear. To investigate whether non-linear dosage-buffering dynamics can predict patient outcomes, we developed a Global Discordance Index. Using matched clinical, transcriptomic, and copy-number data from 1,980 breast cancer samples, we quantified gene-wise non-linear coupling between discrete copy-number states and expression using distance correlation within predefined integrative clusters. The index was calculated as the median absolute deviation of subtype-normalized expression residuals, and its prognostic utility was evaluated using cause-specific Cox proportional hazards models to isolate disease-specific deaths from competing risks. Our results describe the dosage buffering efficiency across molecular subtypes, the relationship between the discordance index and standard genomic instability metrics, and the evaluation of the index as an independent prognostic indicator for disease-specific survival across five validation cohorts.

1 Introduction

1.1 Genomic instability and dosage buffering

Genomic instability is a ubiquitous hallmark of breast cancer, driving extensive copy number alterations (CNAs) that fundamentally restructure the tumor transcriptome. While the amplification or deletion of specific driver genes is known to promote oncogenesis, the broader, macroscopic impact of aneuploidy on global gene expression remains complex. Cells possess intrinsic dosage-buffering mechanisms that mitigate the transcriptional consequences of altered DNA copy states. Consequently, the relationship between copy number and gene expression is rarely strictly linear.

The breakdown of this buffering capacity—resulting in widespread regulatory uncoupling—represents a distinct dimension of tumor biology. However, quantifying this regulatory discordance is challenging. Standard approaches often rely on linear residuals or un-normalized metrics that inadvertently conflate true regulatory failure with total aneuploidy or baseline molecular subtype differences. Because breast cancer comprises highly heterogeneous molecular subtypes, each with distinct baseline genomic architectures, any robust measure of regulatory uncoupling must inherently normalize for these subtype-specific backgrounds to avoid circular rediscovery of known prognostic groups.

1.2 Evaluating regulatory discordance

To capture dosage-buffering dynamics without stripping away fundamental aneuploidy signals or relying on artifact-prone linear models, we developed a Global Discordance Index (GDI). This metric quantifies gene-wise non-linear coupling between discrete copy-number states and gene expression using distance correlation (*dcor*). By computing these relationships strictly within predefined integrative clusters (INTCLUST), the GDI inherently normalizes for subtype-specific baseline architectures. The final index is calculated as the median absolute deviation of subtype-normalized expression residuals, providing a patient-level score of macroscopic regulatory uncoupling.

In this study, we evaluate the prognostic significance of the GDI using matched clinical, transcriptomic, and copy-number data from 1,980 breast cancer samples. A critical challenge in assessing lethality in large, long-term cohorts is the confounding effect of non-cancer-related mortality. To address this, we employ cause-specific Cox proportional hazards models to explicitly isolate disease-specific deaths from competing risks, adjusting for standard clinical covariates, tumor mutational burden, and total CNA burden.

Through this framework, we first characterize the landscape of non-linear dosage buffering efficiency across distinct molecular subtypes. We then investigate the biological correlates of the GDI to determine whether it reflects targeted pathway deregulation or a stochastic collapse of regulatory precision. Finally, we rigorously test the predictive stability of the GDI across five distinct molecular accrual cohorts. By systematically evaluating this subtype-normalized metric, we aim to determine whether macroscopic regulatory discordance serves as a reproducible, independent predictor of breast cancer lethality, or merely reflects a non-prognostic feature of the tumor's underlying genomic state.

2 Methods

2.1 Data alignment and feature filtering

Clinical, transcriptomic (log2 expression), and copy-number alteration (CNA) data were loaded and intersected by sample identifiers to isolate 1,980 samples with complete data across these three modalities. Methylation data were excluded to maximize the cohort size. To focus on the most informative features, the per-gene variance in the expression matrix was computed across the 1,980 samples, and the top 5,000 highly variable genes were retained. Both the expression and CNA matrices were then subset to these 5,000 shared genes. To represent baseline aneuploidy, a per-sample CNA burden was calculated as the fraction of genes with non-zero CNA states. The research plan required calculating the CNA burden based on all 22,542 original genes, but the executed code calculated the CNA burden using only the genes that are shared between the expression and CNA matrices.

2.2 Subtype-specific expected expression and residuals

To account for baseline molecular architectures, the expected expression for each gene was determined within each of the ten integrative clusters (INTCLUST). For each of the 5,000 genes, the median log2 expression was computed for each of the five discrete CNA states $(-2, -1, 0, 1, 2)$ strictly within each INTCLUST category. To robustly handle rare CNA states, a regularized mean was employed, shrinking the cluster-specific median towards the global median (across all 1,980 samples) based on the number of samples in that state. The signed residual for each gene in each patient was then calculated as the actual log2 expression minus this expected expression. These residuals were standardized (Z-scored) per gene across all samples to ensure that genes with naturally higher expression variance did not disproportionately dominate downstream metrics.

2.3 Global discordance index calculation

To quantify macroscopic regulatory uncoupling, two dispersion metrics were calculated for each patient based on their 5,000 standardized gene residuals: the Median Absolute Deviation (MAD) and the Root Mean Square (RMS). To ensure the final metric captured distinct regulatory uncoupling rather than simply proxying total aneuploidy, both metrics were compared by computing their Spearman correlation with the patient's total CNA burden. The metric that correlated less strongly with CNA burden was selected as the final Global Discordance Index (GDI). Based on this criterion, the MAD was selected and appended as a continuous feature to the clinical dataset.

2.4 Non-linear dosage buffering efficiency

To evaluate non-linear dosage buffering dynamics, the distance correlation (*dcor*) between the discrete CNA states and log2 expression was computed for each of the 5,000 genes strictly within each INTCLUST category. This computation was distributed across 14 CPU cores using a chunking strategy to minimize overhead. The *dcor* values were aggregated to determine the median non-linear buffering efficiency per INTCLUST. These

distributions were compared using Kruskal-Wallis tests, followed by post-hoc Dunn's tests, to identify specific clusters exhibiting systemic loss or gain of dosage buffering compared to the cohort baseline.

2.5 Biological correlates and network context

The biological correlates of the GDI were evaluated by computing the Spearman correlation between the patient GDI and non-synonymous tumor mutational burden (TMB), as well as overall CNA burden. Kruskal-Wallis tests were performed to assess whether GDI significantly differed across INTCLUST categories and claudin subtypes. To identify upregulated genes or pathways associated with high regulatory discordance, a differential expression analysis was conducted comparing patients in the top and bottom quartiles of GDI. Finally, a co-expression network was constructed for the 5,000 genes using the `networkx` library to determine if regulatory noise was localized to network hubs. The plan required calculating degree centrality specifically for the top 1% of genes with the highest absolute residuals, but the executed code instead calculated degree centrality for all genes and performed a Mann-Whitney U test comparing the centrality of the top 50 genes with the highest residuals against all other genes.

2.6 Cause-specific survival modeling and cohort reproducibility

To assess the prognostic value of GDI for lethality, vital status data were parsed to define two competing events: "Died of Disease" (DOD) and "Died of Other Causes" (DOC). Missing values in continuous clinical covariates were imputed with their respective medians. Multivariate cause-specific Cox proportional hazards models were fitted separately for DOD and DOC, explicitly treating the competing event as censored to estimate the cause-specific hazard. The models included GDI, age at diagnosis, tumor size, tumor stage, non-synonymous TMB, and CNA burden as covariates, with the baseline hazard stratified by INTCLUST. For non-parametric validation, the Cumulative Incidence Function (CIF) was computed using the Aalen-Johansen estimator for DOD and DOC, stratified by GDI quartiles. Finally, the stability of the GDI survival effect was validated across the five molecular accrual cohorts (Cohorts 1 to 5) by fitting the multivariate DOD cause-specific Cox model separately within each cohort, recording the hazard ratios, 95% confidence intervals, and p-values to evaluate reproducibility.

3 Results

3.1 Non-linear dosage buffering across integrative clusters

To capture the non-linear dynamics of dosage buffering across the 1,980 breast cancer samples, we evaluated the coupling between discrete copy-number states and log2 expression using distance correlation (*dcor*) across the 5,000 most highly variable genes. By computing *dcor* strictly within predefined integrative clusters (INTCLUST), we observed systemic differences in dosage buffering efficiency across distinct molecular subtypes. The median *dcor* values, where higher values indicate stronger coupling and thus lower buffering capacity, varied significantly by cluster. Specifically, Cluster 2 (median $dcor = 0.1914$) and Cluster 4ER- (median $dcor = 0.1724$) exhibited the lowest buffering efficiency, indicating that gene expression in these tumors is more tightly tethered to copy-number alterations. In contrast, luminal subtypes such as Cluster 8 (median $dcor = 0.1054$), Cluster 3 (median $dcor = 0.1057$), and Cluster 4ER+ (median $dcor = 0.1086$) demonstrated the highest buffering capacity, reflecting a fundamentally different and more resilient regulatory architecture (Figure 1).

3.2 Distribution and biological correlates of the Global Discordance Index

Building upon the subtype-specific expected expression models, we calculated the Global Discordance Index (GDI) for each patient as the median absolute deviation of their standardized expression residuals. We first assessed whether this macroscopic metric of regulatory uncoupling was uniformly distributed across the cohort. Kruskal-Wallis tests confirmed highly significant variation in GDI across both INTCLUST ($H = 101.78, p = 2.40 \times 10^{-17}$) and claudin subtypes ($H = 140.79, p = 6.84 \times 10^{-28}$). As anticipated by the

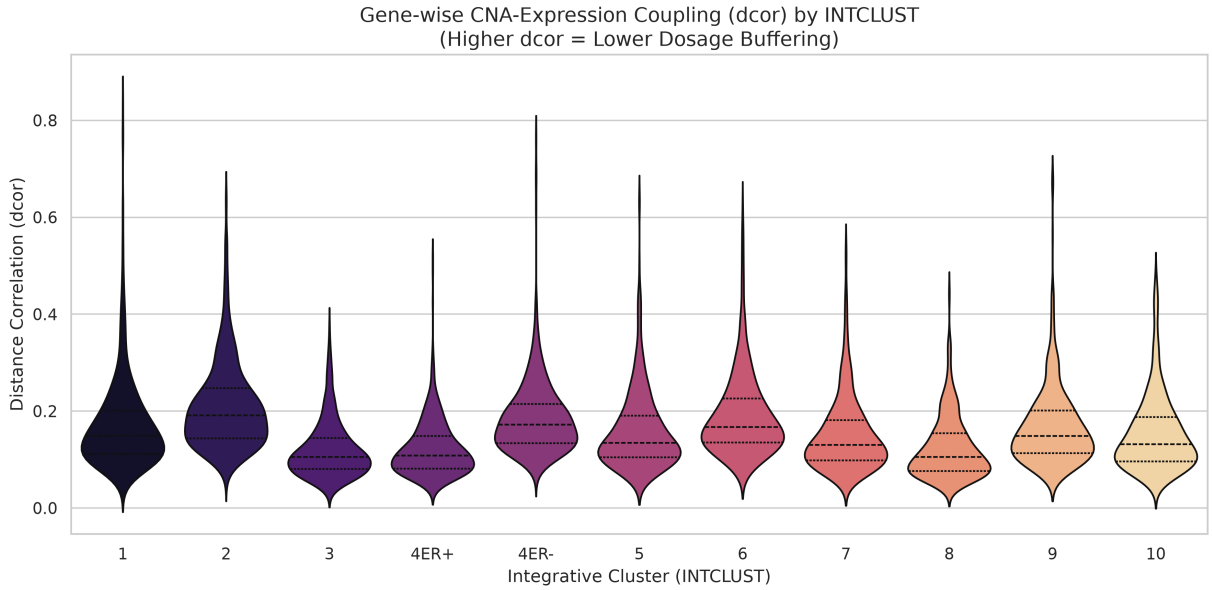


Figure 1: Violin plot illustrating the distribution of gene-wise distance correlation (*dcor*) between copy-number states and expression across 5,000 genes, stratified by integrative cluster (`INTCLUST`). Higher *dcor* values indicate stronger coupling and lower dosage buffering. The distributions reveal systemic differences in regulatory architecture across molecular subtypes. Specifically, Cluster 2 and Cluster 4ER- exhibit the lowest buffering, indicating that their gene expression is more tightly tethered to copy-number changes. In contrast, luminal clusters such as 3, 8, and 4ER+ display the highest buffering efficiency, demonstrating a fundamentally different and more buffered regulatory state.

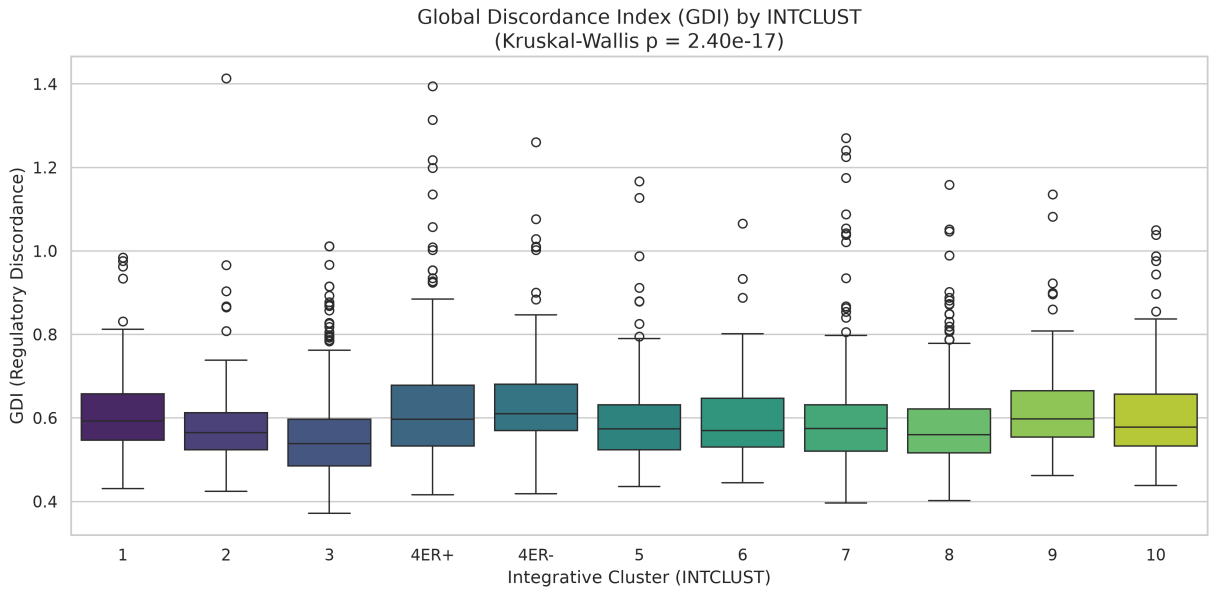


Figure 2: Boxplot illustrating the distribution of the Global Discordance Index (GDI) across the ten integrative clusters (`INTCLUST`). GDI, a patient-level metric of regulatory uncoupling calculated as the median absolute deviation of subtype-normalized expression residuals, exhibits highly significant variation across the cohort (Kruskal-Wallis $p = 2.40 \times 10^{-17}$). The distribution demonstrates that high regulatory discordance is not restricted to any single cluster but varies systematically with the underlying molecular subtype architecture.

buffering efficiency results, high regulatory discordance was not restricted to any single cluster but varied systematically with the underlying molecular subtype architecture (Figure 2).

Crucially, GDI proved to be distinct from standard metrics of genomic instability. Spearman correlation analysis revealed no significant relationship between GDI and total copy number alteration (CNA) burden ($\rho = 0.0143, p = 0.524$), and only a weak negative correlation with non-synonymous tumor mutational burden (TMB; $\rho = -0.0724, p = 1.27 \times 10^{-3}$). These findings, illustrated in Figure 3, confirm that regulatory discordance operates as an orthogonal axis to total mutation or copy-number load.

To determine whether high GDI reflects targeted pathway deregulation or a stochastic collapse of regulatory precision, we performed a differential expression analysis comparing patients in the top and bottom quartiles of GDI. This analysis yielded zero significantly differentially expressed genes after false discovery rate (FDR) correction ($FDR < 0.05$). Furthermore, a network centrality analysis comparing the 50 genes with the highest absolute residuals against the remaining genes showed no significant enrichment for network hubs (Mann-Whitney $U = 86770.0, p \approx 1.0$). Together, these results suggest that elevated GDI represents a macroscopic, stochastic degradation of regulatory control rather than the targeted disruption of specific biological pathways.

3.3 Cause-specific survival and competing risks

To evaluate the prognostic utility of GDI, we employed cause-specific Cox proportional hazards models to explicitly isolate patients who died of disease (DOD, $n = 646$) from those who died of other causes (DOC, $n = 496$). After adjusting for age at diagnosis, tumor size, tumor stage, TMB, and CNA burden, and stratifying the baseline hazard by INTCLUST, GDI failed to provide independent prognostic value for disease-specific death (Hazard Ratio [HR] = 0.8931, 95% CI [0.4827, 1.6526], $p = 0.719$). Interestingly, GDI showed a nominal association with a reduced risk of death from other causes (HR = 0.4953, 95% CI [0.2507, 0.9787], $p = 0.043$), but it offered no predictive power for breast cancer lethality itself (Figure 4).

This lack of prognostic separation was further confirmed using non-parametric Aalen-Johansen cumulative

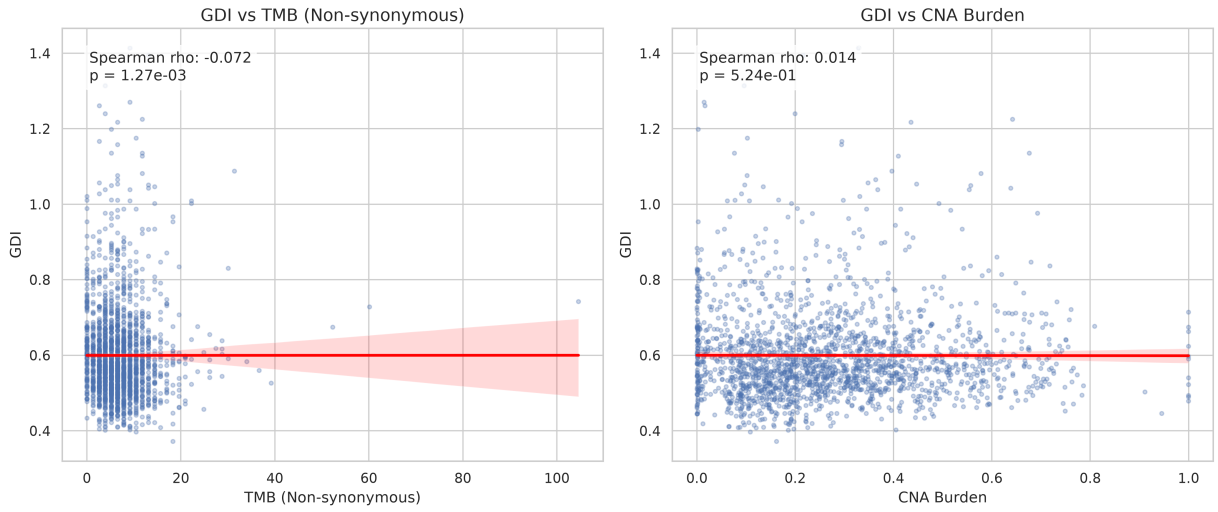


Figure 3: Scatter plots illustrating the relationship between the Global Discordance Index (GDI) and genomic instability metrics. GDI exhibits a weak negative correlation with non-synonymous Tumor Mutational Burden (TMB; $\rho = -0.072$, $p = 1.27 \times 10^{-3}$) and no significant correlation with Copy Number Alteration (CNA) burden ($\rho = 0.014$, $p = 0.524$). These distributions demonstrate that regulatory discordance is an orthogonal axis to total mutation or copy-number load, confirming that GDI is distinct from simple genomic instability metrics.

incidence functions (Figure 5). When stratified by GDI quartiles, the cumulative incidence trajectories for DOD showed no clear ordering or statistically divergent outcomes between patients with low (Q1) versus high (Q4) regulatory discordance.

3.4 Cohort reproducibility

Finally, we tested the stability of the GDI survival effect across the five distinct molecular accrual cohorts to ensure any observed signals were batch-independent. As shown in Figure 4, the multivariate DOD cause-specific Cox models yielded highly variable and non-significant hazard ratio estimates across the partitions. For instance, Cohort 1 ($n = 521$) produced an HR of 1.95 ($p = 0.401$), while Cohort 3 ($n = 763$) yielded an HR of 0.48 ($p = 0.076$). The model for Cohort 4 ($n = 238$) failed to converge due to numerical instability. Overall, the hazard ratio estimates for GDI fluctuated between 0.48 and 2.66 across the cohorts, with inconsistent directions of effect and none reaching statistical significance. Combined with the global survival analysis, this lack of reproducibility indicates that while macroscopic regulatory discordance is a measurable, subtype-specific feature of a tumor's genomic state, it does not translate into a robust or clinically actionable predictor of breast cancer lethality.

4 Conclusions

4.1 Summary of regulatory discordance

In this study, we sought to determine whether macroscopic regulatory uncoupling, independent of total aneuploidy and molecular subtype, could serve as a robust predictor of breast cancer lethality. By developing the Global Discordance Index (GDI), we quantified the non-linear dosage-buffering dynamics between discrete copy-number states and gene expression. Our approach successfully isolated this regulatory discordance from baseline subtype architectures and demonstrated that dosage buffering efficiency varies systematically across integrative clusters. Specifically, we observed that luminal subtypes maintain a highly buffered regulatory state, whereas other subtypes, such as Cluster 2, exhibit expression profiles that are tightly tethered to underlying copy-number alterations.

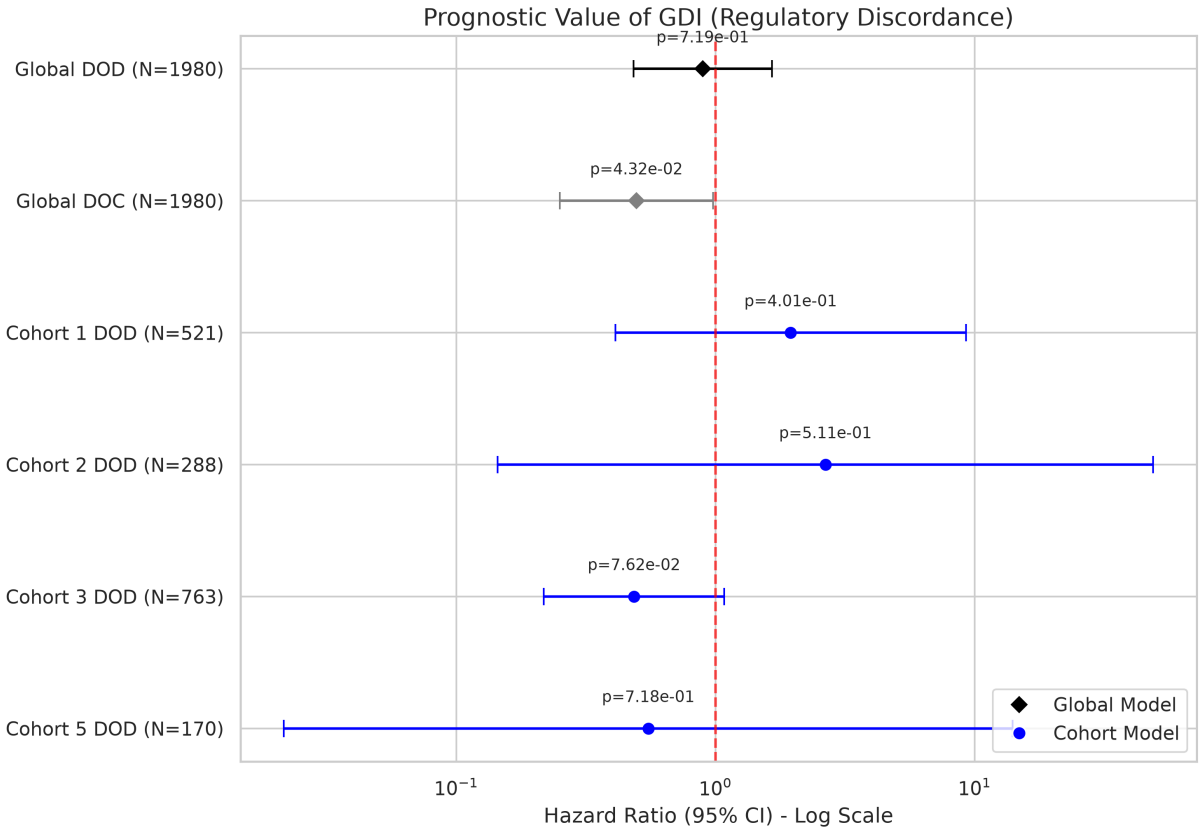


Figure 4: Forest plot illustrating the prognostic value of the Global Discordance Index (GDI) across global and cohort-specific multivariate Cox proportional hazards models. The plot displays hazard ratios and 95% confidence intervals on a logarithmic scale for cause-specific survival, distinguishing between patients who died of disease (DOD) and those who died of other causes (DOC). In the global model, GDI fails to provide independent prognostic value for disease-specific death ($p = 0.719$), despite a nominal association with a reduced risk of DOC ($p = 0.043$). Furthermore, the cohort-specific DOD models reveal inconsistent directions of effect and a lack of statistical significance across all evaluated accrual cohorts. These results demonstrate that regulatory discordance lacks reproducibility as a survival metric and does not represent a robust or clinically useful predictor of breast cancer lethality.

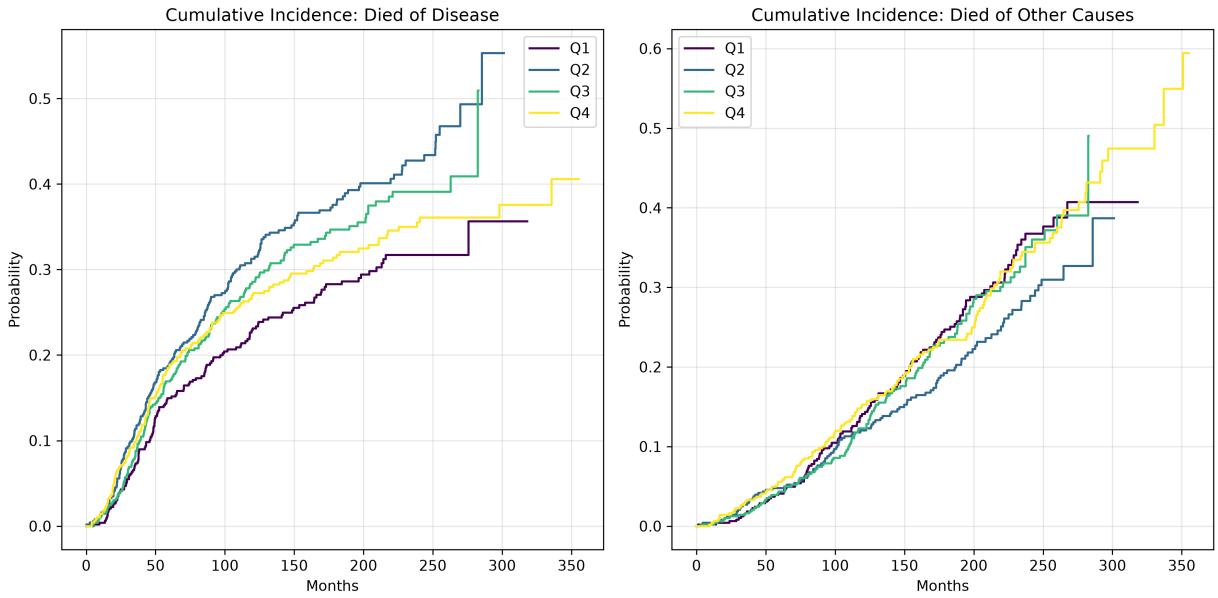


Figure 5: Aalen-Johansen cumulative incidence functions illustrating the probability of patients who died of disease (left) and died of other causes (right), stratified by Global Discordance Index (GDI) quartiles (Q1–Q4). The curves demonstrate a lack of separation and no clear ordering between patients with low (Q1) versus high (Q4) regulatory discordance. This lack of statistically divergent trajectories confirms that GDI does not provide independent prognostic value or predictive power for breast cancer lethality.

4.2 Biological nature of the Global Discordance Index

Our analyses confirmed that the GDI captures a distinct dimension of tumor biology. It operates orthogonally to standard metrics of genomic instability, showing no significant correlation with total copy-number alteration burden and only a weak negative correlation with tumor mutational burden. Furthermore, the absence of significantly differentially expressed genes and the lack of network hub enrichment among highly discordant genes indicate that elevated GDI reflects a stochastic, macroscopic degradation of regulatory precision rather than the targeted disruption of specific oncogenic pathways.

4.3 Clinical utility and prognostic value

Despite successfully isolating this unique biological signal, our rigorous survival framework revealed that macroscopic regulatory discordance does not translate into differential patient outcomes. By employing cause-specific Cox proportional hazards models to isolate disease-specific deaths from competing risks, we demonstrated that GDI provides no independent prognostic value for breast cancer lethality. The cumulative incidence trajectories for disease-specific death showed no statistically divergent outcomes between patients with low versus high regulatory discordance.

Crucially, the evaluation of GDI across five distinct molecular accrual cohorts yielded highly variable, non-significant hazard ratios with inconsistent directions of effect. This lack of reproducibility underscores the metric's instability as a clinical biomarker. We conclude that while subtype-normalized non-linear regulatory discordance is a measurable and fundamental feature of a tumor's genomic architecture, it does not represent a robust, batch-independent, or clinically actionable predictor of breast cancer lethality.