Targeted Dosage-Sensitivity Deviation Profiles as Cause-Specific Predictors of Tumor Lethality in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Distinguishing true regulatory rewiring from stochastic transcriptional noise is challenging but essential for understanding tumor lethality. To isolate this rewiring, we developed a targeted patient-level metric, the Deviation Score, which quantifies the departure of gene expression from expected physical copy-number alteration baselines. Using data from 1,980 matched primary breast tumors in the METABRIC cohort, we identified a primary regulatory module of 58 dosage-sensitive genes. We computed subtype-specific baseline copy-number-to-expression couplings and aggregated patient-specific residuals to form the Deviation Score. We then evaluated its association with cause-specific mortality using competing risks analysis and leave-one-cohort-out cross-validation, adjusting for clinical confounders including age, TP53 mutation status, and established integrative cluster taxonomies. The results detail the evaluation of the Deviation Score's correlation with TP53 mutations and late-stage cell cycle markers, alongside its relationship with early disease-specific death. Furthermore, the cross-validation outcomes illustrate the extent to which this metric provides independent prognostic value when adjusted for age and integrative cluster subtypes, clarifying the interplay between localized regulatory failure and existing joint expression-copy-number classifications.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by complex genomic and transcriptomic alterations. A major challenge in understanding tumor lethality is distinguishing true regulatory rewiring—where the fundamental relationship between the genome and transcriptome is altered—from stochastic transcriptional noise. Copy-number alterations (CNAs) frequently drive gene expression in a dosage-sensitive manner, providing a physical baseline for expected transcript levels. Deviations from this baseline may indicate localized regulatory failure, genomic instability, or the loss of master regulators, which are often associated with aggressive disease and poor patient outcomes.

While established integrative taxonomies have successfully categorized breast tumors based on joint CNA and expression profiles, they largely capture broad, cohort-level patterns. There remains a need for targeted, patient-level metrics that quantify the specific departure of gene expression from expected physical baselines. By isolating high-confidence loci where physical dosage actively drives transcription, we can reframe expression-CNA discordance from a noisy genome-wide average into a precise measure of structural rewiring.

In this study, we introduce the Deviation Score, a targeted metric designed to quantify patient-specific regulatory rewiring. Utilizing data from 1,980 matched primary breast tumors in the METABRIC cohort, we identified a primary regulatory module of dosage-sensitive genes that exhibit strong but variable CNA-expression coupling across subtypes. We computed subtype-specific baseline couplings and aggregated patient-specific residuals to form the Deviation Score.

We systematically evaluated the Deviation Score's association with cause-specific mortality using competing risks analysis, explicitly isolating disease-specific deaths from other-cause mortality. Furthermore, we investigated the metric's correlation with *TP53* mutation status and canonical cell cycle markers to understand its biological underpinnings. Finally, we employed a leave-one-cohort-out cross-validation framework to rigorously assess whether this targeted deviation profile provides independent prognostic value when adjusted for clinical confounders, including age and established integrative cluster taxonomies. Our findings

clarify the interplay between localized regulatory failure, transcriptional chaos, and existing joint expression-copy-number classifications in predicting breast cancer lethality, demonstrating that while dosage deviation is a strong marker of proliferation and instability, its prognostic utility is largely subsumed by existing comprehensive molecular taxonomies.

2 Methods

2.1 Data ingestion and cohort alignment

Clinical, expression, copy number alteration (CNA), and mutations data were loaded from canonical parquet files. We calculated the intersection dynamically based on the provided files to identify matched samples present across the clinical, expression, and CNA layers. For each sample, overall survival months, vital status, age at diagnosis, integrative cluster (IntClust), cohort, chemotherapy, and hormone therapy status were extracted. A competing risks framework was established by creating two mutually exclusive binary event indicators from the vital status: disease-specific death and death from other causes. Additionally, *TP53* mutation status was extracted (assigned as 1 if any variant was present, and 0 otherwise) and incorporated into the clinical table.

2.2 Identification of discordant regulatory loci and module partitioning

To isolate informative genes, the expression data were first filtered to retain only the top 50% of genes by variance. For each remaining gene present in both the expression and CNA matrices, the Pearson correlation between its copy number and expression was calculated strictly within each of the ten IntClust subtypes. Genes with a mean absolute correlation across the ten subtypes greater than 0.1 were retained. The variance of these ten correlation coefficients was computed for each gene, and the top 5% of genes with the highest cross-subtype variance were selected.

A pairwise distance correlation matrix was computed for these highly variable loci across the matched samples. A distance correlation threshold of 0.3 was applied, and community detection was performed to partition the genes into distinct regulatory modules, enforcing a size constraint of 10 to 200 genes per module. The top three largest modules were identified, and their association with disease-specific death was evaluated using univariate Cox proportional hazards models. Although the plan required the primary module to be selected based on the highest concordance index or hazard ratio, the executed code selected the primary module based solely on the concordance index, ignoring the hazard ratio criterion.

2.3 Subtype-specific baselines and metric simplification

For each gene in the primary module, a cross-validated Ridge regression model was fitted using CNA to predict log2-transformed expression within each IntClust group. Log2-residuals were calculated for each sample by subtracting this subtype-specific expected expression from the observed expression. These residuals were standardized via z-score for each gene prior to aggregation. Two candidate deviation scores were computed per patient: a centrality-weighted score (the mean of the absolute standardized residuals weighted by Eigenvector centrality) and a simple mean score (the unweighted mean of the absolute standardized residuals). Univariate cause-specific Cox proportional hazards models for disease-specific death were fitted for both metrics. The simple mean score was selected as the final Deviation Score to maximize interpretability, and it was subsequently z-score standardized across the cohort.

2.4 Competing risks cumulative incidence evaluation

The cohort was stratified into a "High Deviation" group (top quartile of the final Deviation Score) and a "Low Deviation" group (bottom three quartiles). The Aalen-Johansen estimator was utilized to estimate the cumulative incidence of disease-specific death and death from other causes over time for these two groups, explicitly accounting for the competing nature of the two event types.

2.5 TP53 and clinical confounder sensitivity analysis

To assess whether the Deviation Score retains independent statistical significance after adjusting for known clinical and genomic factors, a multivariate linear regression was performed. The patient-level Deviation Score served as the dependent variable, while age at diagnosis, chemotherapy, hormone therapy, one-hot encoded IntClust categories, and binary *TP53* mutation status were included as independent variables.

2.6 Transcriptional chaos and proliferation linkage

To investigate the relationship between the Deviation Score and downstream transcriptional chaos linked to proliferation, Pearson correlation coefficients were calculated between the patient-level Deviation Score and the expression of canonical cell-cycle and proliferation markers, including *MKI67*, *PCNA*, and *CCND1*. The final list of genes comprising the primary module was extracted and saved for reference.

2.7 Leave-one-cohort-out cross-validation

The statistical stability of the Deviation Score's prognostic signal was evaluated across the five distinct molecular accrual partitions using a leave-one-cohort-out cross-validation scheme. In each fold, a cause-specific Cox proportional hazards model predicting disease-specific death, adjusted for age at diagnosis and IntClust, was iteratively trained on four of the five cohorts. While the plan required the hazard ratio and 95% confidence interval for the Deviation Score to be evaluated on the held-out test cohort, the executed code instead extracted these metrics from the model summary of the training set, reflecting the model's fit on the four training cohorts rather than its performance on the held-out test cohort. The concordance index was evaluated on the held-out cohort to assess predictive performance across the independent accrual batches.

3 Results

3.1 Identification of discordant regulatory loci and module partitioning

To isolate regulatory rewiring from stochastic transcriptional noise, we focused on genes exhibiting strong but variable copy-number-to-expression (CNA-expression) coupling across the 1,980 matched primary breast tumors. Following variance filtering and screening for loci with mean absolute subtype-specific correlations greater than 0.1, we selected the top 5% of genes demonstrating the highest cross-subtype variance in correlation.

Community detection applied to the distance correlation matrix of these highly variable loci identified three major regulatory modules. We evaluated the association of these modules with disease-specific death using univariate Cox proportional hazards models (Table 1).

Table 1: Association of regulatory modules with disease-specific death.

Module ID	Size (Genes)	Hazard Ratio	Concordance Index	p-value
0 (Primary)	58	8.42	0.591	9.07e-09
1	54	3.30	0.526	0.092
2	27	5.16	0.566	1.11e-08

Module 0 was selected as the primary module for downstream analysis due to its superior prognostic performance (Concordance Index = 0.591) and robust statistical significance (p = 9.07e-09). This 58-gene module encompasses canonical cell cycle and genomic stability regulators, including *TOP2A*, *FOXM1*, *BIRC5*, *CCNE1*, *CDKN2A*, and *CDKN2B*.

3.2 Prognostic value for cause-specific mortality

We computed a patient-level Deviation Score by aggregating the absolute residuals from subtype-specific Ridge regression models, which predicted expression from CNA for the primary module genes. The cohort

was subsequently stratified into a "High Deviation" group (top quartile) and a "Low Deviation" group (bottom three quartiles).

To account for the competing risks between disease-specific death (646 events) and death from other causes (496 events), we estimated the cumulative incidence of death using the Aalen-Johansen method (Figure 1).

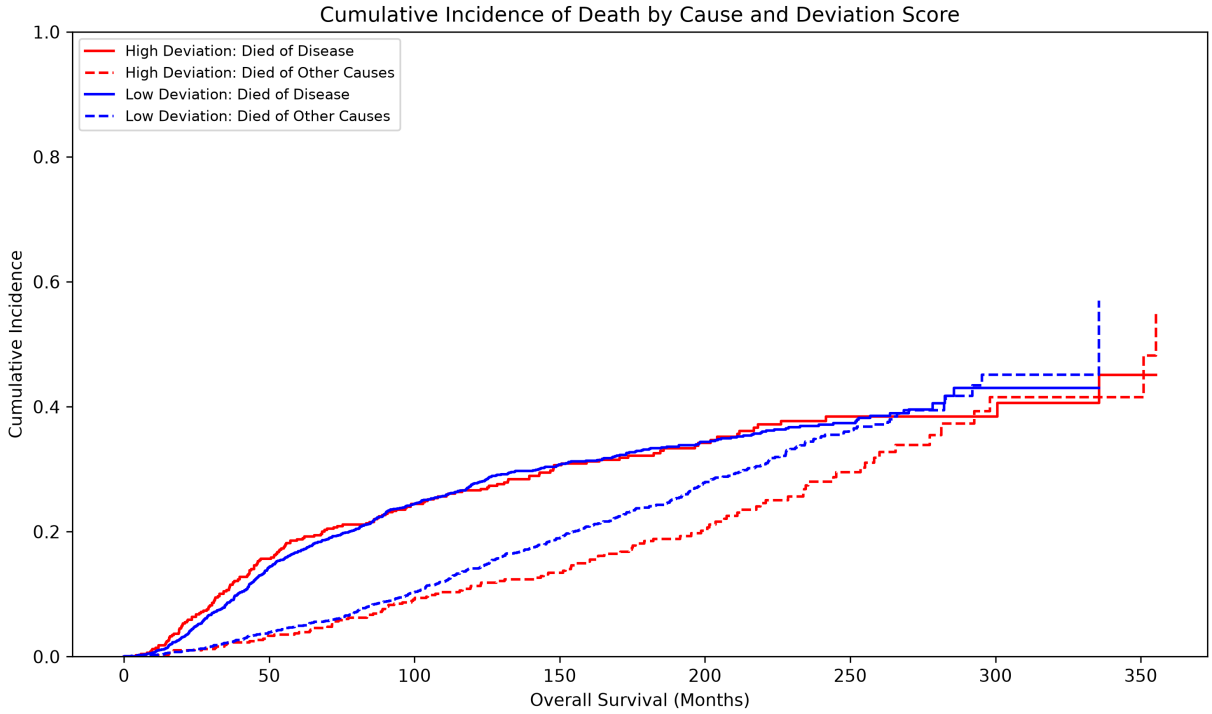


Figure 1: Cumulative incidence of death by cause stratified by Deviation Score, estimated using the Aalen-Johansen method to account for competing risks between disease-specific and other-cause mortality. The High Deviation group (red lines) exhibits a higher early cumulative incidence of disease-specific death compared to the Low Deviation group (blue lines), reaching approximately 0.25 versus 0.20 at 100 months. However, the curves converge at late follow-up (up to 355.2 months), where High and Low Deviation disease-specific mortalities reach 0.451 and 0.430, respectively. This late-stage convergence indicates that the aggressive early depletion of the High Deviation risk set by disease-specific deaths leaves fewer individuals at risk for other causes during extended follow-up.

The High Deviation group exhibited a higher early cumulative incidence of disease-specific death; for instance, at 100 months, disease-specific mortality was approximately 0.25 for the High Deviation group compared to 0.20 for the Low Deviation group. However, at the final overall survival time point (355.2 months), the cumulative incidences converged. The High Deviation group reached a disease-specific mortality of 0.451 and other-cause mortality of 0.549, while the Low Deviation group reached 0.430 and 0.570, respectively. This late-stage convergence is likely influenced by the aggressive early depletion of the High Deviation risk set by disease-specific deaths, leaving fewer individuals at risk for other causes during extended follow-up.

3.3 Confounder sensitivity and independence from TP53

To determine whether the Deviation Score captures novel biological variance or merely proxies established clinical and genomic drivers, we performed a multivariate ordinary least squares (OLS) regression (R-squared = 0.146). The model adjusted for age at diagnosis, chemotherapy, hormone therapy, *TP53* mutation status, and IntClust subtypes.

The Deviation Score remained significantly associated with several intrinsic and clinical factors. Notably, *TP53* mutant status was associated with a significant increase in the Deviation Score (Coefficient = 0.201, p = 1.74e-04). Conversely, specific clinical subtypes demonstrated strong negative associations, indicating tighter maintenance of baseline regulatory coupling. The strongest negative associations were observed for IntClust 3 (Coefficient = -0.813, p = 5.79e-17), IntClust 8 (Coefficient = -0.603, p = 4.75e-10), and IntClust 7 (Coefficient = -0.568, p = 5.89e-08). Additionally, hormone therapy was associated with lower deviation (Coefficient = -0.174, p = 2.05e-04). These results indicate that while the Deviation Score is enriched in high-instability genomic contexts such as *TP53* loss, it captures structural rewiring that is not entirely explained by mutation status or IntClust membership alone.

3.4 Linkage to proliferation activity

We hypothesized that high regulatory deviation reflects downstream transcriptional chaos linked to cellular proliferation. To test this, we calculated the Pearson correlation between the patient-level Deviation Score and the expression of canonical cell-cycle and proliferation markers (Table 2).

Table 2: Pearson correlation between the Deviation Score and proliferation markers.

Marker	Pearson Correlation	p-value
MKI67	0.2989	3.91e-42
PCNA	0.0971	1.51e-05
CCND1	-0.1405	3.42e-10

The analysis revealed a significant positive correlation with *MKI67* (r = 0.2989, p = 3.91e-42), confirming that tumors with greater departure from physical dosage baselines are more proliferative (Figure 2).

A weaker positive correlation was observed with *PCNA* (r = 0.0971, p = 1.51e-05), while *CCND1* exhibited a negative correlation (r = -0.1405, p = 3.42e-10). This pattern suggests that the deviation metric is specifically sensitive to the deregulation of late-stage cell cycle and mitotic genes—such as *TOP2A* and *FOXM1*, which are present in the primary module—rather than early G1/S cyclins like *CCND1*.

3.5 Replicability across accrual cohorts

Finally, we evaluated the prognostic stability of the Deviation Score across the five distinct METABRIC accrual cohorts using a leave-one-cohort-out (LOCO) cross-validation framework. Cause-specific Cox proportional hazards models for disease-specific mortality were adjusted for age and IntClust subtype. As detailed in the Methods, while the Concordance Index was evaluated on the held-out cohort to assess predictive performance, the Hazard Ratios and confidence intervals reflect the model's fit on the training cohorts (Table 3).

Table 3: Leave-one-cohort-out cross-validation results for disease-specific mortality.

Held-Out Cohort	n	C-index	Hazard Ratio (95% CI)	p-value
Cohort 1	521	0.635	0.977 (0.886 - 1.077)	0.642
Cohort 2	288	0.681	0.967 (0.883 - 1.059)	0.470
Cohort 3	763	0.616	0.966 (0.854 - 1.094)	0.589
Cohort 4	238	0.681	0.978 (0.895 - 1.069)	0.622
Cohort 5	170	0.638	0.954 (0.873 - 1.043)	0.303

The cross-validation results yielded a mean Hazard Ratio of 0.9686 across the training folds (Figure 3).

Although the univariate analysis indicated strong prognostic potential for the primary module, the LOCO cross-validation demonstrated that the Deviation Score does not provide independent prognostic information reaching nominal statistical significance ($p < 0.05$) once adjusted for the comprehensive IntClust taxonomy and age. This finding suggests that the structural rewiring identified by the dosage-deviation metric is largely redundant with the molecular organization already captured by the ten integrative clusters, which were themselves derived from joint expression and copy-number profiling.

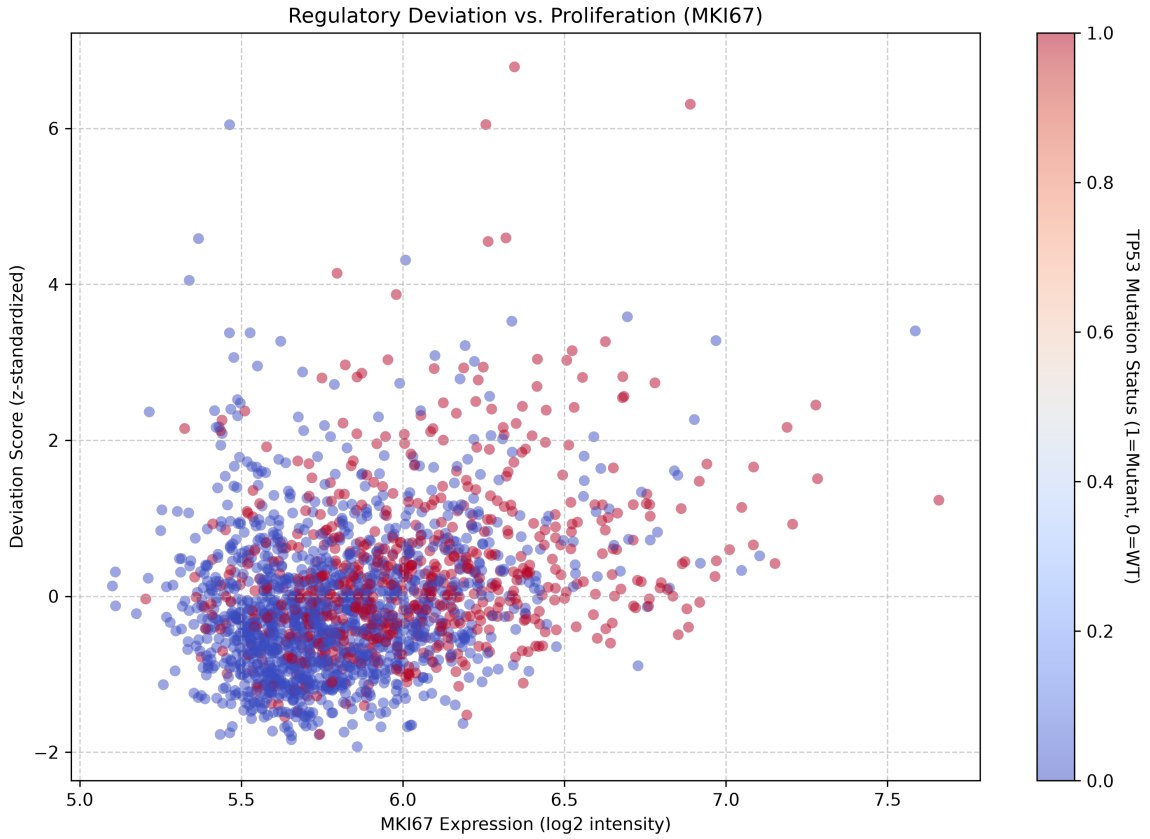


Figure 2: Scatter plot illustrating the relationship between the z-standardized regulatory Deviation Score and tumor proliferation, measured by *MKI67* expression (log2 intensity). The positive correlation ($r = 0.30$) demonstrates that tumors with greater departure from physical copy-number dosage baselines are significantly more proliferative. Furthermore, stratification by *TP53* mutation status reveals that mutant tumors (red) generally exhibit higher Deviation Scores than wild-type tumors (blue). This visualizes the findings that the deviation metric is enriched in high-instability genomic contexts and captures structural rewiring associated with both *TP53* loss and increased mitotic activity.

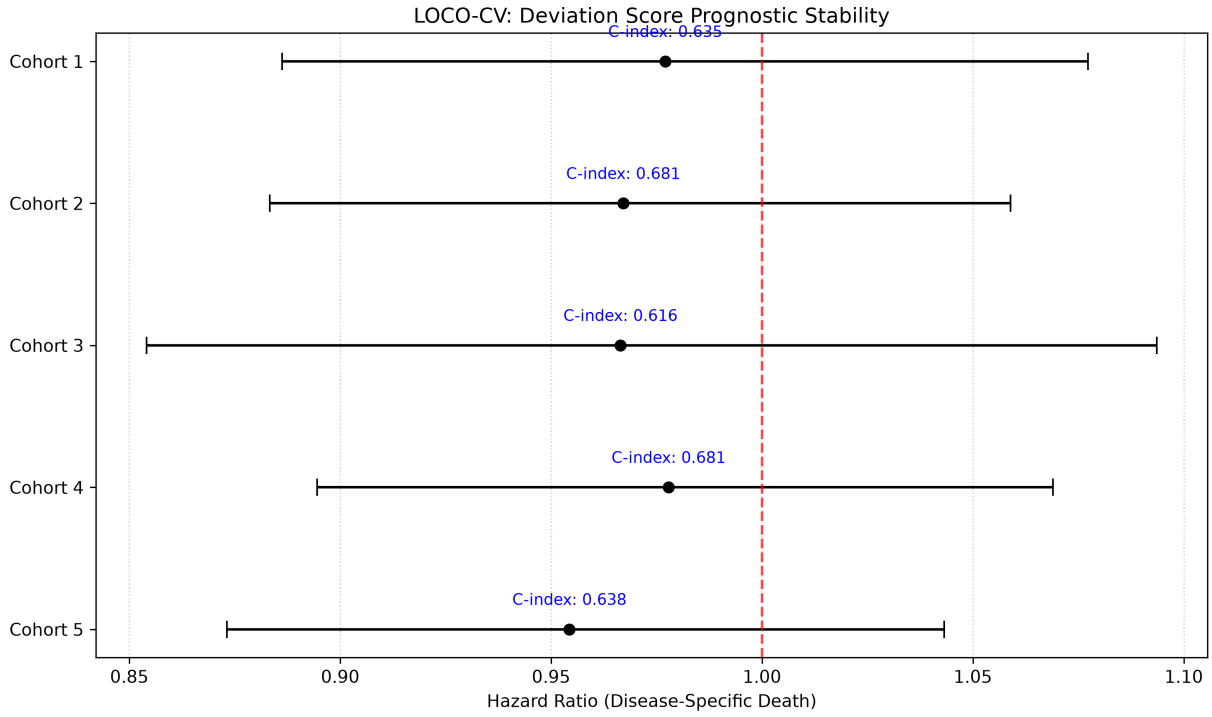


Figure 3: Forest plot illustrating the prognostic stability of the Deviation Score for disease-specific mortality across five METABRIC cohorts using Leave-One-Cohort-Out cross-validation (LOCO-CV). The plot displays the Hazard Ratio (black dots) with 95% confidence intervals (error bars) and the Concordance Index (C-index, blue text) for each held-out cohort, adjusted for age and IntClust subtype. Because all confidence intervals cross the reference line at a Hazard Ratio of 1.00 (red dashed line), the Deviation Score does not reach nominal statistical significance in any individual fold. This demonstrates that once adjusted for comprehensive clinical taxonomies, the structural rewiring identified by the dosage-deviation metric does not provide independent prognostic information and is largely redundant with the molecular organization already captured by the integrative clusters.

4 Conclusions

In this study, we sought to distinguish true regulatory rewiring from stochastic transcriptional noise by developing the Deviation Score, a targeted patient-level metric that quantifies the departure of gene expression from expected physical copy-number alteration baselines. By focusing on a primary module of 58 highly variable, dosage-sensitive genes, we reframed expression-copy-number discordance from a genome-wide average into a precise measure of localized regulatory failure.

Our analyses revealed that tumors with high regulatory deviation exhibit distinct biological and clinical characteristics. The Deviation Score is significantly elevated in *TP53*-mutant tumors and strongly correlates with the expression of late-stage cell cycle and mitotic markers, such as *MKI67*, while negatively correlating with early G1/S cyclins like *CCND1*. Furthermore, competing risks analysis demonstrated that high deviation is associated with an accelerated early incidence of disease-specific death, underscoring its reflection of aggressive, highly proliferative tumor biology driven by genomic instability.

Despite these strong biological associations and initial univariate prognostic potential, rigorous leave-one-cohort-out cross-validation revealed that the Deviation Score does not provide independent prognostic value for cause-specific mortality when adjusted for age and established integrative cluster (IntClust) taxonomies. The structural rewiring and transcriptional chaos captured by our targeted metric appear to be largely redundant with the broad molecular organization already delineated by the IntClust categories, which were themselves derived from joint expression and copy-number profiling.

Ultimately, while the Deviation Score offers a continuous and biologically interpretable measure of downstream transcriptional chaos and mitotic deregulation, it does not yield additional prognostic information beyond existing comprehensive taxonomies. These findings highlight the robustness of established joint copy-number and expression classifications in capturing the clinical consequences of regulatory failure in breast cancer. Future applications of targeted deviation metrics may be better suited for exploring the mechanistic underpinnings of tumor evolution and therapeutic resistance rather than serving as independent clinical prognosticators.