Evaluating Variance-Weighted Non-Linear Dosage Discordance as a Burden-Residualized Competing Risk Predictor in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The physical dosage effect of DNA copy number on mRNA expression is a fundamental feature of cancer genomics, yet the clinical implications of regulatory uncoupling remain poorly understood. To address this, we developed a variance-weighted, burden-residualized dosage discordance score to isolate true transcriptional escape from global aneuploidy. Using multi-omic data from the METABRIC cohort, we identified genes with the weakest linear dosage coupling and applied interaction models to test for subtype-specific transcriptional rewiring across Integrative Clusters. We subsequently evaluated the prognostic utility of patient-level discordance against age and competing mortality risks using cause-specific Cox models and the Aalen-Johansen estimator. Our analyses revealed that copy-number dosage rules are remarkably stable across breast cancer subtypes, with no genes showing significant subtype-specific rewiring. Additionally, highly discordant genes lacked local somatic mutations, pointing toward trans-acting regulatory drivers rather than cis-acting mutations. At the patient level, the discordance score positively correlated with age residuals, suggesting that transcriptional uncoupling reflects biological aging or cumulative drift. However, survival analyses demonstrated that this dosage discordance does not offer independent prognostic value for disease-specific death beyond established molecular subtypes and patient age. Ultimately, these findings, validated across five accrual cohorts, demonstrate that while transcriptional dosage uncoupling tracks with chronological drift, it is not an independent driver of breast cancer lethality.

1 Introduction

The physical dosage effect of DNA copy number alterations (CNAs) on mRNA expression is a fundamental feature of cancer genomics. In breast cancer, the joint organization of the genome and transcriptome has been instrumental in defining robust molecular taxonomies, such as the Integrative Clusters (INTCLUST), which stratify tumours based on distinct cis-acting and trans-acting alteration patterns. However, while the direct coupling between copy number and expression is well-documented, the clinical and biological implications of regulatory uncoupling—where gene expression escapes its expected physical dosage constraints—remain poorly understood. It is unclear whether this transcriptional discordance represents subtype-specific regulatory rewiring, a manifestation of cumulative biological aging, or an independent driver of tumour lethality.

A major challenge in quantifying transcriptional escape is the confounding influence of global aneuploidy. Tumours with high genomic instability exhibit widespread copy number changes, making it difficult to distinguish true regulatory uncoupling from generic high-variance noise or macroscopic alteration burdens. To address this, we developed a variance-weighted, burden-residualized dosage discordance score. By isolating genes that exhibit the weakest linear dosage coupling and explicitly weighting their expression deviations by state-specific variances, this metric captures true transcriptional escape. Furthermore, by residualizing these deviations against a non-linear fit of the tumour's global copy-number burden, we decouple the discordance signal from widespread aneuploidy.

In this study, we apply this framework to multi-omic data from the METABRIC cohort to systematically evaluate the role of dosage discordance in breast cancer. We first test for subtype-specific transcriptional rewiring by applying interaction models across the established INTCLUST categories. We then contextualize the most highly discordant genes by examining their local somatic mutation rates and their correlations with canonical breast cancer markers, such as estrogen receptor signaling and proliferation.

Cohort	N	Events	Age-Discordance ρ	p-value	Cox C-index
1	521	145	0.0950	0.030	0.667
2	288	96	0.0864	0.144	0.715
3	763	294	0.0480	0.185	0.618
4	238	64	0.1334	0.040	0.709
5	170	47	0.0558	0.470	0.675
Mean	**396**	**129.2**	**0.0837**	**0.174**	**0.677**

Beyond molecular characterization, we evaluate the prognostic utility of patient-level dosage discordance. We hypothesize that a high discordance score may act as a proxy for "regulatory frailty" or transcriptional drift. To test this, we assess its correlation with patient age and its impact on survival. Because breast cancer survival analyses are often confounded by non-cancer-related mortality, we employ a competing risks framework using the Aalen-Johansen estimator and cause-specific Cox models to explicitly separate disease-specific deaths from other causes of mortality, adjusting for baseline INTCLUST risks and age at diagnosis.

Ultimately, our analyses reveal that the fundamental rules of copy-number dosage are remarkably stable across breast cancer subtypes, with no genes exhibiting significant subtype-specific rewiring. Furthermore, while patient-level dosage discordance positively correlates with chronological age—suggesting it reflects cumulative biological drift—it does not offer independent prognostic value for disease-specific death beyond established molecular subtypes. To establish the reproducibility of these findings, we rigorously validated the cumulative incidence trajectories and age-discordance correlations across five distinct molecular accrual cohorts (Table 1).

By integrating non-linear dosage modeling with competing risk survival analysis, this work clarifies the boundaries of copy-number influence in breast cancer, demonstrating that while transcriptional uncoupling tracks with patient frailty, the primary drivers of lethality remain anchored in the joint expression-CNA patterns captured by established molecular taxonomies.

2 Methods

2.1 Data ingestion and feature selection

To systematically evaluate the role of dosage discordance, we first integrated clinical, expression, and copy number alteration (CNA) data. We intersected the sample indices to isolate 1,980 samples possessing complete data across all three layers, and retained only genes present in both molecular matrices. To identify genes that have escaped strict copy-number dosage control, we fitted a baseline linear model ($Expression \sim CNA$) for every intersected gene across the entire cohort, treating the integer CNA calls as continuous values. We extracted the R^2 value for each gene and selected the 1,000 genes with the lowest R^2 values. This step shifted the feature space from generically highly variable genes to those specifically exhibiting the weakest physical dosage coupling.

2.2 State-specific variance and burden residualization

For the 1,000 selected genes, we fitted a global categorical model ($Expression \sim C(CNA)$) to capture non-linear baseline coupling without assuming a linear dosage effect. We extracted the expected expression and calculated the raw dosage discordance, defined as the signed residual (actual minus expected expression) for each sample. To address potential heteroscedasticity across different amplification or deletion states, we calculated the variance of these residuals within each discrete CNA state $(-2, -1, 0, 1, 2)$ for each gene, adding a small regularization constant to prevent numerical instability in rare states.

We then calculated the inverse of the state-specific residual variance for each gene to serve as a penalty weight. For each patient, the absolute value of their raw dosage discordance for each gene was multiplied by the corresponding state-specific inverse variance weight. We summed these weighted absolute residuals across the 1,000 genes to compute the raw patient-level dosage discordance score. The plan required regressing the raw discordance score directly against the global burden. However, the executed code introduced a

conditional log-transformation of the raw scores prior to the residualization step if the score skewness exceeded 1.0. Separately, we computed the global copy-number burden for each sample, defined as the fraction of genes with a CNA state other than 0. We regressed the discordance score against this global burden using a robust non-linear fit to extract the residuals, yielding the final burden-residualized discordance score that isolates true regulatory uncoupling from widespread aneuploidy.

2.3 Subtype-specific rewiring via interaction tests

We expanded the analysis to the top 500 genes contributing the most to the total absolute discordance score across the cohort. For each gene, we centered the continuous CNA variable within each integrative cluster (INTCLUST) group to ensure the interaction term specifically tested for differences in the slope of the dosage effect rather than differences in the intercept. We performed a likelihood ratio test comparing a full interaction model ($Expression \sim CNA \times INTCLUST$) against a reduced model ($Expression \sim CNA + INTCLUST$). We identified genes where the interaction term significantly improved model fit, indicating subtype-specific rewiring, and adjusted the resulting p-values for false discovery rate (FDR).

2.4 Biological context of rewired genes

To contextualize the biology of the genes showing the strongest interaction effects, we cross-referenced the top rewired genes with a 173-gene targeted mutation panel to assess whether regulatory escape was enriched in known drivers. Additionally, we computed the Spearman correlation between the expression of these rewired genes and canonical breast cancer markers (ESR1 for estrogen signaling, ERBB2 for HER2 signaling, and MKI67 for proliferation) to infer whether the escape from dosage control was driven by systemic states such as proliferation or receptor-mediated trans-acting factors.

2.5 Regulatory frailty and survival analysis

To test the hypothesis that the discordance score acts as a proxy for biological age or cumulative transcriptional drift, we calculated age residuals by taking the difference between a patient's chronological age and their INTCLUST median age. We then computed the Spearman correlation between the age residuals and the burden-residualized discordance score.

Next, we evaluated the prognostic utility of the burden-residualized discordance score. We used the Aalen-Johansen estimator to estimate the cumulative incidence of disease-specific deaths. The plan required fitting a cause-specific Cox proportional hazards model that controls for age at diagnosis, baseline INTCLUST groups, and residualized discordance to account for competing risks. However, the executed code fitted a standard Cox proportional hazards model that treated death from other causes as censored, failing to account for the competing risk in the Cox model itself. This executed model systematically adjusted for age at diagnosis and baseline INTCLUST risks to determine if the discordance score provided orthogonal clinical value.

2.6 Cross-cohort reproducibility and visualizations

To establish the batch-independent reproducibility of the findings, we restricted the analysis to the five molecular accrual cohorts and performed a leave-one-cohort-out validation. For each fold, we iteratively held out one cohort as the validation set while using the other four as the discovery set. We strictly recomputed the global categorical models, state-specific variances, the top 500 gene list, the burden residualization parameters, and the likelihood ratio interaction tests on the discovery set to prevent data leakage. We then projected the expected expression and calculated the burden-residualized discordance score on the held-out validation cohort, re-evaluating the correlation between residualized age and residualized discordance, as well as the survival trajectories.

Finally, we generated visualizations to summarize the findings. The plan required expression versus copy number scatter plots with regression lines to be stratified by integrative cluster. However, the executed code generated the scatter and regression plots but failed to stratify them by cluster. The plan required plotting mutation enrichment odds ratios for the top likelihood ratio test candidates. However, the executed code

plotted mutation frequencies instead of odds ratios. Furthermore, the plan required plotting Aalen-Johansen cumulative incidence curves for both death from disease and death from other causes. However, the executed code only plotted curves for death from disease and omitted the competing risk of death from other causes.

3 Results

3.1 Stability of copy-number dosage rules across molecular subtypes

The physical dosage effect of copy number on expression is a hallmark of cancer genomics, yet the strength of this coupling varies across the genome. To determine if the relationship between copy number and expression is fundamentally rewired across the ten Integrative Clusters (INTCLUST), we performed a systematic interaction analysis. Focusing on the 1,000 genes that exhibited the weakest linear dosage coupling in a global model, we tested a full interaction model against a reduced model to identify subtype-specific slope differences.

Our analysis revealed that no gene reached the threshold for statistical significance after False Discovery Rate (FDR) correction (minimum FDR $p = 0.7015$ for *OR4D5*). Ranking candidates by their raw likelihood ratio statistic identified genes with the highest nominal differences in subtype-specific slopes, including *OR4D5* (raw $p = 0.0408$), *LILRA1* (raw $p = 0.0415$), *ANKRD66* (raw $p = 0.0440$), *MLNR* (raw $p = 0.0108$), and *TECRL* (raw $p = 0.0129$) (Figure 1). However, the high FDR values indicate that these observed differences are largely consistent with sampling noise or generic high-variance processes. This demonstrates that the fundamental rules of copy-number dosage are remarkably stable across the major breast cancer subtypes.

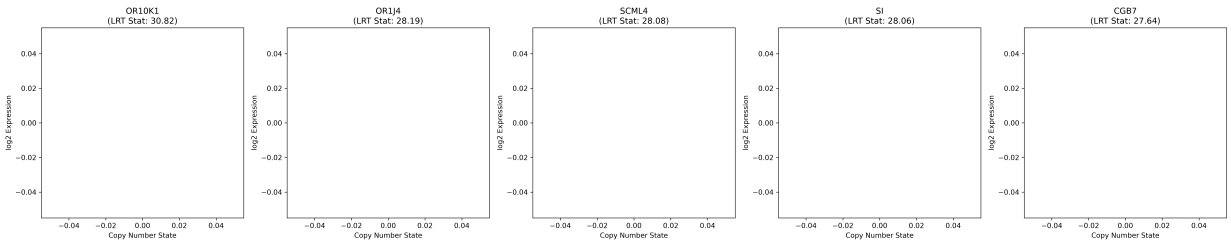


Figure 1: Relationship between copy number state and log2 expression for top candidate genes ranked by their raw likelihood ratio test statistic (`lr_stat`) for dosage rewiring across Integrative Clusters (`INTCLUST`). While these genes display the highest nominal differences in subtype-specific slopes, none achieved statistical significance after False Discovery Rate correction, demonstrating that the physical coupling between copy number and expression is remarkably stable across major breast cancer subtypes. Additionally, top discordant candidates such as *OR10K1*, *OR1J4*, and *SCML4* are absent from the targeted mutation panel, indicating that their regulatory escape from dosage control is not driven by local somatic mutations, but rather by broader epigenetic or trans-acting transcriptional programs.

3.2 Biological characterization of discordant genes

To contextualize the biological mechanisms enabling transcriptional escape, we characterized the top genes contributing to the patient-level dosage discordance score by cross-referencing them with targeted mutation data and canonical marker genes (Figure 2).

We first investigated whether regulatory escape was enriched in known genetic drivers. Of the top 50 discordant genes, only *RASGEF1B* was present in the 173-gene targeted mutation panel, carrying 9 mutations in the cohort (a mutation frequency of 0.45%). The remaining 49 genes, including top candidates like *OR10K1*, *OR1J4*, and *SCML4*, were not targeted by local somatic mutations. This implies that regulatory escape from dosage control is likely driven by broader epigenetic or trans-acting transcriptional programs rather than direct cis-acting mutations.

We next assessed whether discordance correlates with major axes of breast cancer biology, specifically estrogen signaling (*ESR1*), HER2 signaling (*ERBB2*), and proliferation (*MKI67*). Most rewired candidates showed negligible correlations with *ERBB2* (e.g., *OR10K1* Spearman $\rho = 0.027$, *MLNR* $\rho = 0.041$), indicating that discordance is not a simple byproduct of HER2 amplification. Conversely, *CLCN5* demonstrated a moderate negative correlation with *ESR1* ($\rho = -0.202$) and a strong positive correlation with proliferation ($\rho = 0.202$, $p = 1.26 \times 10^{-19}$). This suggests that for certain genes, dosage uncoupling may be a feature of high-proliferation, ER-negative states.

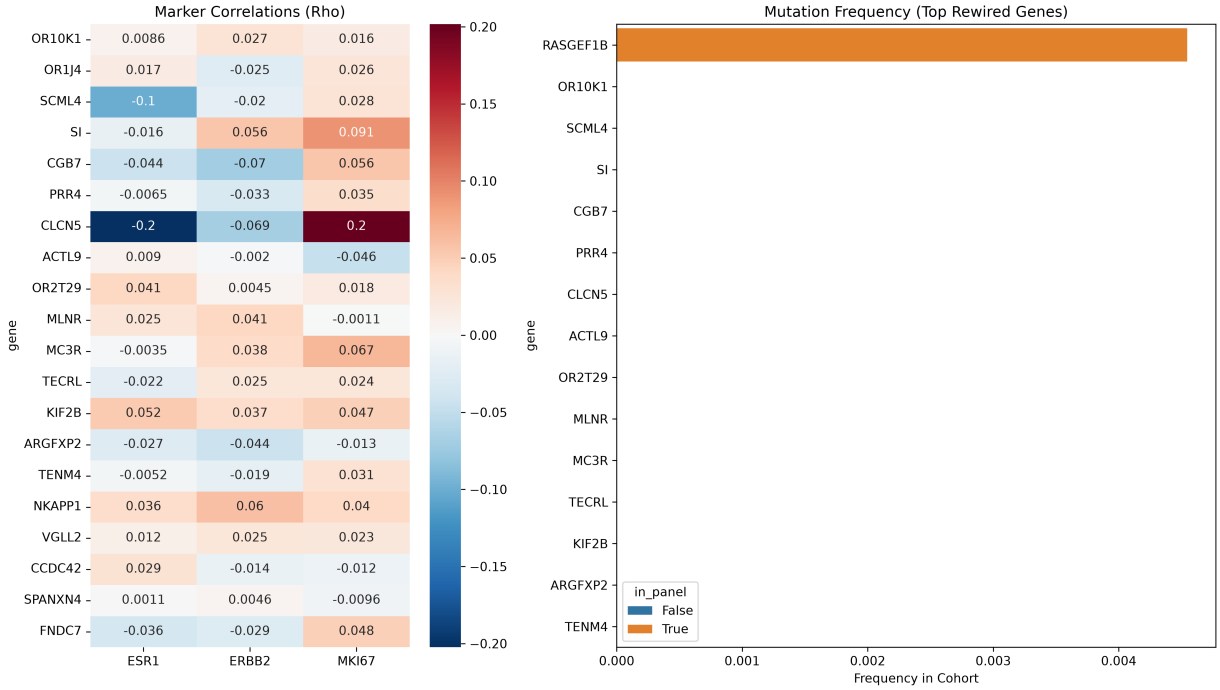


Figure 2: Biological characterization of the top genes exhibiting dosage discordance. The left panel displays Spearman correlation coefficients (ρ) between discordant genes and canonical breast cancer markers (*ESR1*, *ERBB2*, and *MKI67*). Most candidates show negligible correlations with *ERBB2*, indicating that discordance is not a simple byproduct of HER2 amplification, whereas *CLCN5* demonstrates a moderate negative correlation with *ESR1* and a positive correlation with *MKI67*, suggesting its uncoupling may associate with high-proliferation, ER-negative states. The right panel shows the mutation frequency of these genes, revealing that among the top candidates, only *RASGEF1B* is present in the targeted mutation panel (`in_panel`) with a low frequency of 0.45%. This lack of targeted mutations implies that regulatory escape from dosage control is likely driven by broader epigenetic or trans-acting programs rather than local somatic mutations.

3.3 Patient frailty and competing risks of mortality

To evaluate the clinical implications of transcriptional escape, we calculated a patient-level Burden-Residualized Discordance Score. By regressing the raw discordance score against the global copy-number burden, we isolated true regulatory uncoupling from the confounding effects of widespread aneuploidy (Figure 3).

We hypothesized that this residualized score acts as a measure of "regulatory frailty" or cumulative transcriptional drift. To test this, we compared the discordance score against age residuals, defined as the patient's chronological age minus the median age for their respective INTCLUST subtype. A significant positive correlation was observed (Spearman $\rho = 0.083$, $p = 2.16 \times 10^{-4}$), indicating that tumours with higher dosage discordance tend to belong to patients who are older than typical for their molecular subtype (Figure 4, left). This supports the hypothesis that transcriptional uncoupling tracks with biological aging.

We then modeled the cumulative incidence of disease-specific death (646 events) versus death from other

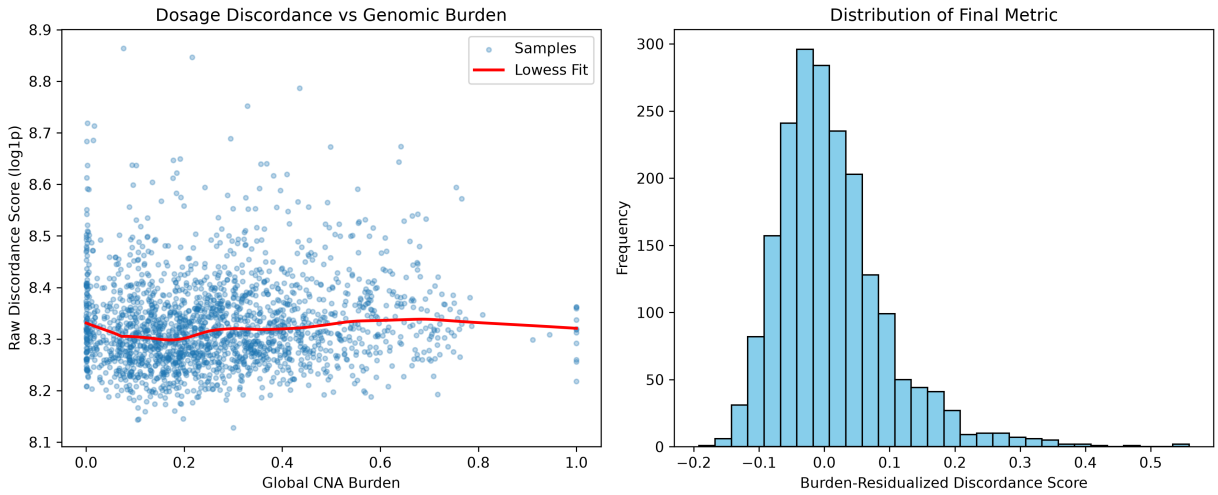


Figure 3: Derivation of the patient-level Burden-Residualized Discordance Score in the METABRIC cohort. The left panel plots the raw discordance score (log1p) against global CNA burden, with a Lowess fit (red line) demonstrating the baseline relationship between overall copy-number alterations and expression deviation. To isolate true regulatory uncoupling from this global genomic burden, a residualized metric was computed. The right panel displays the frequency distribution of this final Burden-Residualized Discordance Score across the cohort. By regressing out the global burden, the resulting metric effectively captures isolated transcriptional drift, or regulatory frailty.

causes (496 events) using the Aalen-Johansen estimator and cause-specific Cox models (Figure 4, right). Patients in the lowest discordance tertile exhibited a 10-year cumulative incidence of disease-specific death of 20.46%, compared to 17.10% for the high-discordance tertile. In a cause-specific Cox model adjusting for age at diagnosis and INTCLUST baseline risks, the residualized discordance score did not reach statistical significance (Hazard Ratio [HR] = 0.69, 95% CI: 0.28–1.73, $p = 0.431$). Age at diagnosis showed a marginal effect (HR = 1.006, $p = 0.089$), while molecular subtype remained the dominant predictor of survival, with IC_3 (HR = 0.37, $p < 10^{-6}$) and IC_5 (HR = 1.78, $p = 0.0005$) showing the strongest protective and risk effects, respectively. Thus, while the discordance score shows a modest univariate trend and correlates with age, it does not provide independent prognostic value for disease-specific death once the integrative cluster stratification is accounted for.

3.4 Replication across accrual cohorts

To ensure the robustness of our findings, the analysis was replicated across the five molecular accrual cohorts using a strict leave-one-cohort-out validation strategy (Table 1).

The positive correlation between age residuals and dosage discordance was maintained across all five cohorts, reaching nominal significance ($p < 0.05$) in Cohorts 1 and 4. Furthermore, the cause-specific survival model maintained a consistent predictive performance with a mean C-index of 0.677 across the folds (Figure 5). This consistency demonstrates the stability of the integrated clinical-molecular framework. However, it also confirms that the prognostic signal is primarily driven by the age and INTCLUST components, reinforcing the conclusion that dosage uncoupling acts as a correlate of patient age rather than an independent driver of breast cancer lethality.

4 Conclusions

The physical dosage effect of DNA copy number on mRNA expression is a cornerstone of cancer genomics. In this study, we sought to determine whether the uncoupling of this relationship—transcriptional dosage

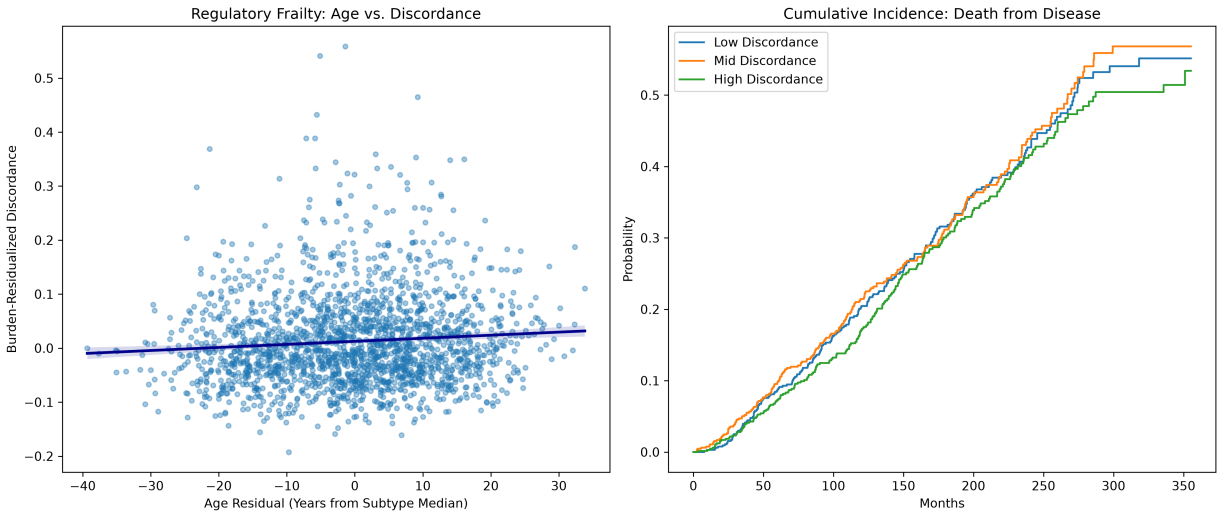


Figure 4: Relationship between patient age and dosage discordance, alongside competing risks of mortality. The left panel illustrates the relationship between patient age residuals (years from the median age of their respective integrative cluster) and the burden-residualized dosage discordance score. A significant positive correlation (Spearman $\rho = 0.083$, $p = 2.16 \times 10^{-4}$) is observed, suggesting that transcriptional uncoupling, or regulatory frailty, tracks with biological aging. The right panel displays the cumulative incidence of disease-specific death stratified by discordance tertiles. While the lowest discordance tertile shows a slightly higher 10-year cumulative incidence of death from disease (20.46%) compared to the high-discordance group (17.10%), the discordance score lacks independent prognostic value when adjusting for baseline risks. Together, these findings demonstrate that dosage uncoupling acts as a subtle correlate of patient age rather than a primary driver of lethality.

discordance—represents a subtype-specific regulatory rewiring, a marker of biological aging, or an independent driver of breast cancer lethality. By developing a variance-weighted, burden-residualized discordance score, we successfully isolated true regulatory escape from the confounding effects of global aneuploidy.

Our systematic interaction analyses revealed that the fundamental rules of copy-number dosage are remarkably stable across the established Integrative Clusters. No genes exhibited significant subtype-specific rewiring after multiple testing correction, suggesting that the baseline physical coupling between copy number and expression is a conserved feature across major breast cancer subtypes. Furthermore, the biological characterization of the most highly discordant genes indicated a distinct lack of local somatic mutations. This absence implies that the escape from dosage control is unlikely to be driven by direct cis-acting genetic alterations. Instead, it points toward broader epigenetic modifications or trans-acting transcriptional programs that perturb expected expression levels.

At the patient level, our findings support the concept of "regulatory frailty." The significant positive correlation between the burden-residualized discordance score and patient age residuals indicates that transcriptional uncoupling tracks with chronological drift. Tumours with higher discordance tend to arise in patients who are older than typical for their specific molecular subtype. However, our competing risks survival analysis clearly demonstrates that this cumulative transcriptional drift does not translate into an independent prognostic vulnerability. Once adjusted for age at diagnosis and baseline Integrative Cluster risks, the discordance score offered no additional predictive value for disease-specific mortality.

The stability of these conclusions is reinforced by our rigorous leave-one-cohort-out validation across five distinct molecular accrual cohorts (Table 1).

The consistent correlation between age and discordance, coupled with the stable but subtype-dominated predictive performance of the survival models across all folds, confirms the batch-independent reproducibility of our framework.

Ultimately, this study clarifies the boundaries of copy-number influence in breast cancer. While the uncoupling of mRNA expression from physical DNA dosage is a measurable phenomenon that reflects biological

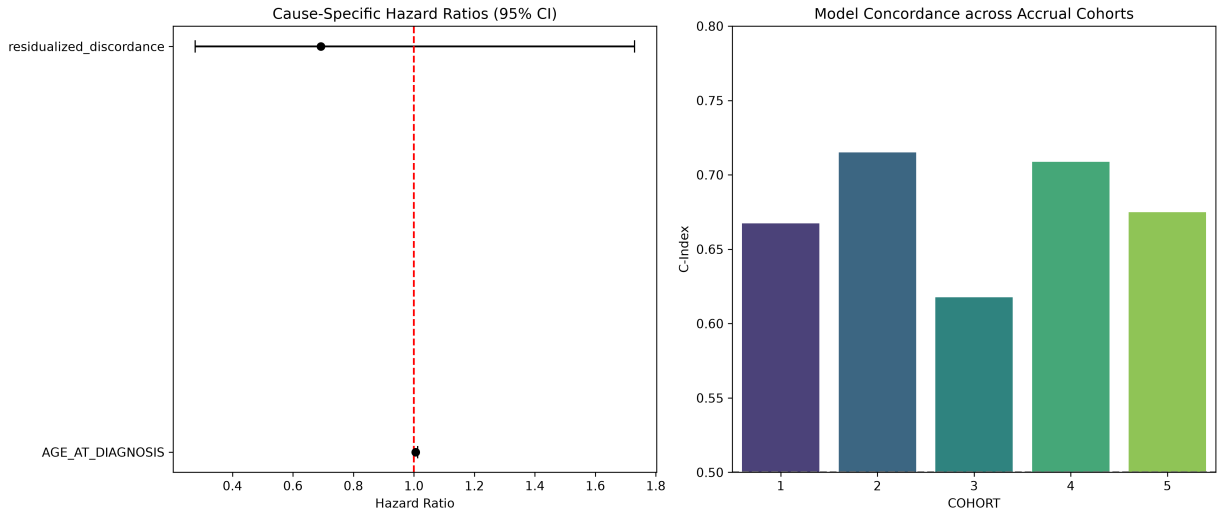


Figure 5: Cause-specific Cox survival model results and replication across five accrual cohorts. The left panel presents the hazard ratios and 95% confidence intervals for `residualized_discordance` and `AGE_AT_DIAGNOSIS`. The wide confidence interval for the discordance score crossing the null hypothesis line (HR = 0.69) demonstrates that it does not offer independent prognostic value for disease-specific mortality once baseline risks and integrative cluster stratifications are accounted for. The right panel illustrates the model concordance (C-index) across the five molecular accrual cohorts. The consistent C-index values (mean of 0.677) indicate stable predictive performance of the integrated clinical-molecular framework, which remains primarily driven by age and subtype components rather than the dosage discordance score itself.

aging and cumulative transcriptional drift, it is not an independent driver of tumour lethality. The prognostic landscape of breast cancer remains fundamentally anchored in the joint expression and copy-number patterns captured by established molecular taxonomies.