Methylation-Dependent Transcriptional Residual Variance as a Cause-Specific Predictor of Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Disentangling true epigenetic regulation from the dominant transcriptional effects of macroscopic aneuploidy remains a critical challenge in understanding breast cancer progression. To isolate dosage-independent epigenetic rewiring, we developed an analytical framework utilizing 1,417 multi-omic samples from the METABRIC cohort. We regressed gene expression against discrete copy-number states to extract transcriptional residuals, which were subsequently correlated with promoter methylation to identify subtype-specific epigenetically regulated loci. By aggregating the variance of these residuals, we constructed a patient-level Epigenetic Rewiring Score (ERS). The prognostic utility of the ERS was evaluated using cause-specific Cox proportional hazards models adjusted for baseline clinical covariates and integrative cluster taxonomies, alongside mutational interaction analyses. Despite successfully capturing epigenetic-transcriptomic discordance, the ERS did not demonstrate independent prognostic value for disease-specific mortality (hazard ratio 0.991, $p = 0.843$) or death from competing causes. Furthermore, no significant prognostic synergies were observed between the ERS and key driver mutations, including TP53, PIK3CA, GATA3, and MAP3K1. A random-effects meta-analysis across five distinct molecular accrual cohorts confirmed the stability of this null association, yielding a pooled hazard ratio of 1.002 with zero heterogeneity. These findings indicate that while methylation-dependent transcriptional residual variance is a robustly measurable molecular state, it does not serve as an independent predictor of lethal progression beyond established clinical and molecular classifications in breast cancer.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by a complex interplay of genetic and epigenetic alterations. Among these, macroscopic aneuploidy and widespread copy number alterations (CNAs) are hallmarks of tumor progression, exerting a dominant and pervasive effect on global transcriptional output. While DNA methylation is recognized as a critical mechanism of epigenetic regulation and gene silencing, its specific impact on the transcriptome is frequently masked by the overwhelming dosage effects of genomic instability.

1.1 The challenge of confounding molecular layers

Disentangling true epigenetic rewiring from dosage-driven transcriptional changes remains a significant analytical challenge in molecular oncology. In standard transcriptomic analyses, the expression level of a gene is a composite signal. If a tumor exhibits hypermethylation at a specific promoter alongside a chromosomal deletion encompassing that same locus, it is difficult to partition the relative contribution of the epigenetic modification versus the physical loss of DNA. Consequently, existing prognostic models that incorporate methylation or expression data often conflate these signals, leaving it unclear whether epigenetic alterations independently drive tumor lethality or merely act as passenger events secondary to macroscopic aneuploidy.

To rigorously evaluate the prognostic value of epigenetic regulation, it is necessary to isolate the component of transcriptional variance that is strictly dependent on methylation and independent of DNA dosage.

1.2 Isolating epigenetic rewiring

To address this confounding, we developed a multi-omic analytical framework utilizing 1,417 samples from the METABRIC cohort, integrating matched clinical, transcriptomic, copy number, and methylation data. We hypothesized that by regressing gene expression against discrete copy-number states, we could extract dosage-independent transcriptional residuals. Conceptually, for a given gene i and sample j, the log-transformed expression Y_{ij} can be modeled as:

$$Y_{ij} = \beta_0 + \beta_1 C_{ij} + \beta_2 B_j + \epsilon_{ij} \tag{1}$$

where C_{ij} represents the discrete copy-number state, B_j is the global CNA burden of the tumor (acting as a proxy for macroscopic aneuploidy and tumor purity), and ϵ_{ij} represents the dosage-independent transcriptional residual.

By computing the distance correlation between these residuals (ϵ_{ij}) and promoter methylation beta values, we can systematically identify subtype-specific loci where epigenetic silencing strongly governs the remaining transcriptional variance. Aggregating the variance of these highly dependent genes yields a patient-level Epigenetic Rewiring Score (ERS), a metric designed to capture the magnitude of epigenetic-transcriptomic discordance without the collinear instability of modeling global copy-number burden directly.

1.3 Study objectives

The primary objective of this study is to evaluate whether this isolated methylation-dependent transcriptional residual variance serves as a cause-specific predictor of lethality in breast cancer. Because breast cancer patients frequently face competing risks of mortality, we employ cause-specific Cox proportional hazards models to explicitly distinguish disease-specific deaths from competing non-disease deaths.

Furthermore, we investigate whether the prognostic utility of the ERS is context-dependent by testing for synergistic interactions with key driver mutations, including *TP53*, *PIK3CA*, *GATA3*, and *MAP3K1*. Finally, to ensure that our findings represent a reproducible biological phenomenon rather than a cohort-specific artifact, we validate the statistical stability of the ERS across five distinct molecular accrual partitions using a random-effects meta-analysis.

Through this rigorous, multi-layered approach, we aim to define the boundaries of epigenetic prognostication. As detailed in the subsequent sections, while our framework successfully isolates dosage-independent epigenetic rewiring as a robustly measurable molecular state, we demonstrate that it does not provide independent prognostic value beyond established clinical covariates and integrative cluster taxonomies.

2 Methods

2.1 Cohort alignment and data preprocessing

We utilized multi-omic data from the METABRIC cohort, encompassing clinical, transcriptomic, copy number alteration (CNA), DNA methylation, and somatic mutation profiles. Samples were strictly intersected across all molecular layers, yielding a final analytical cohort of 1,417 patients. To facilitate cause-specific survival analyses, the vital status was parsed to create a binary event indicator where disease-specific death was defined as the event of interest, while patients who were living or died of competing causes were right-censored. A per-sample global CNA burden was calculated as the fraction of genes exhibiting non-zero discrete CNA states, serving as a proxy for macroscopic aneuploidy and tumor purity. Missing values in clinical covariates, such as tumor size and histological grade, were addressed by introducing a dedicated "missing" category to retain the full sample size, with complete-case sensitivity analyses planned to ensure robustness.

2.2 Computation of dosage-independent transcriptional residuals

To isolate the component of transcriptional variance independent of DNA dosage, we performed gene-wise ordinary least squares (OLS) regression. For each gene shared across the expression, CNA, and methylation matrices, log2-transformed expression values were regressed against their corresponding discrete five-level CNA states. The CNA states were dummy-encoded with the neutral diploid state (0) set as the reference

level. The global CNA burden was included as an additional covariate to adjust for broad genomic instability. Expanding upon the conceptual model introduced previously, the regression for gene i in sample j was formalized as:

$$Y_{ij} = \beta_0 + \sum_{k \in \{-2,-1,1,2\}} \beta_{1,k} I(C_{ij} = k) + \beta_2 B_j + \epsilon_{ij} \tag{2}$$

where Y_{ij} is the expression level, $I(C_{ij} = k)$ is the indicator function for the discrete CNA state k, B_j is the global CNA burden, and ϵ_{ij} represents the dosage-independent transcriptional residual. These residuals were extracted for all subsequent analyses.

2.3 Subtype-specific identification of methylation-dependent genes

Recognizing the profound molecular heterogeneity of breast cancer, we identified methylation-dependent genes within each precomputed Integrative Cluster (INTCLUST) subtype. For each INTCLUST group, we computed the distance correlation between the promoter methylation beta values and the transcriptional residuals (ϵ_{ij}) for every gene. To account for varying sample sizes across subtypes, empirical p-values were derived using a permutation-based approach, shuffling sample labels 1,000 times per gene. Genes exhibiting a significant distance correlation (false discovery rate [FDR] < 0.05) were selected to define subtype-specific sets of epigenetically regulated loci.

2.4 Calculation of the Epigenetic Rewiring Score (ERS)

For each patient, an Epigenetic Rewiring Score (ERS) was constructed to quantify the magnitude of methylation-driven transcriptional deviation. The ERS was calculated as the mean squared residual (variance) of the transcriptional residuals, restricted strictly to the significant methylation-dependent genes identified for the patient's respective INTCLUST subtype:

$$\text{ERS}_j = \frac{1}{|S_c|} \sum_{i \in S_c} \epsilon_{ij}^2 \tag{3}$$

where S_c is the set of methylation-dependent genes for the INTCLUST subtype c to which patient j belongs. To ensure comparability across the cohort, the raw ERS values were subsequently standardized (Z-scored) within each INTCLUST group. The ERS was intentionally not orthogonalized against the global CNA burden to preserve any biologically relevant interactions with aneuploidy.

2.5 Cause-specific survival and mutational interaction analysis

The prognostic utility of the ERS was evaluated using cause-specific Cox proportional hazards models, explicitly estimating the hazard of disease-specific mortality in the presence of competing risks. The models were adjusted for baseline clinical and molecular confounders, including age at diagnosis, chemotherapy, hormone therapy, tumor size, histological grade, tumor stage, global CNA burden, and INTCLUST category. The proportional hazards assumption was tested for all covariates.

To investigate potential synergies between epigenetic rewiring and specific genetic backgrounds, we cross-referenced the cohort with somatic mutation data for key breast cancer drivers (*TP53*, *PIK3CA*, *GATA3*, and *MAP3K1*). For each driver gene, an interaction model was fitted incorporating the main effects of the ERS, the binary mutation status, and their interaction term (ERS $\times$ mutation status), adjusting for the same clinical covariates. Interaction p-values were corrected for multiple testing using the Benjamini-Hochberg procedure.

2.6 Meta-analysis across accrual cohorts

To validate the statistical stability and reproducibility of the ERS effect, we leveraged the multi-center design of the METABRIC study. The cohort was partitioned into its five distinct molecular accrual batches (Cohorts 1 through 5). The adjusted cause-specific Cox models were fitted independently within each partition. The resulting hazard ratios and standard errors were then synthesized using a random-effects meta-analysis to

calculate a pooled hazard ratio. Heterogeneity across the accrual cohorts was assessed using Cochran's Q statistic and the I^2 metric, ensuring that any observed effects—or lack thereof—were not driven by batch-specific artifacts but represented a consistent biological phenomenon.

3 Results

3.1 Isolation of dosage-independent epigenetic rewiring

To operationalize our framework for isolating epigenetic regulation from macroscopic aneuploidy, we analyzed the core intersection of 1,417 METABRIC samples with complete expression, copy number alteration (CNA), and methylation profiles. Following the ordinary least squares (OLS) regression of log2-scaled expression against its five-level discrete copy-number state and the global per-sample CNA burden, we successfully extracted dosage-independent transcriptional residuals. Subsequent subtype-specific filtering via distance correlation and a 1,000-fold permutation test (FDR < 0.05) identified distinct sets of loci where these residuals were significantly coupled to promoter methylation within each Integrative Cluster (INTCLUST).

By calculating the mean squared residual variance across these subtype-specific, methylation-dependent gene sets, we computed the Epigenetic Rewiring Score (ERS) for each patient. To ensure comparability across the diverse molecular landscape of the cohort, the raw ERS values were standardized (Z-scored) within each INTCLUST group. The resulting distribution provides a normalized, patient-level metric of transcriptomic deviation driven by epigenetic modulation rather than DNA dosage ().

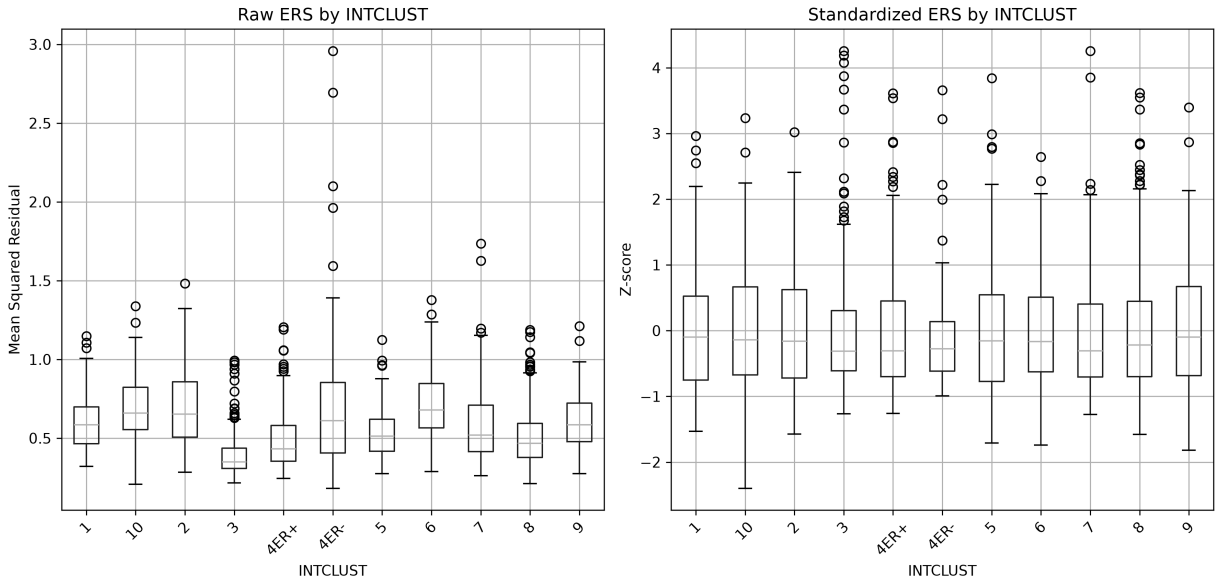


Figure 1: Distribution of the Epigenetic Rewiring Score (ERS) across `INTCLUST` subgroups. The left panel displays the raw ERS, calculated as the mean squared residual variance across subtype-specific, methylation-dependent gene sets for each patient. The right panel illustrates the standardized ERS, which has been Z-scored within each `INTCLUST` category to facilitate comparison across the cohort. This standardization provides a normalized measure of how much a tumor's transcriptome has deviated from its expected dosage-based output due to epigenetic modulation.

3.2 Cause-specific survival outcomes

With the ERS established as a measure of epigenetic-transcriptomic discordance, we evaluated its prognostic utility using cause-specific Cox proportional hazards models. This approach explicitly distinguished death from disease ($n = 467$) from death due to competing causes ($n = 363$). The models were rigorously adjusted

for baseline clinical covariates, including age at diagnosis, chemotherapy, hormone therapy, global CNA burden, tumor size, histological grade, tumor stage, and INTCLUST category.

The analysis revealed that the ERS does not serve as a significant independent predictor for either cause of mortality. For disease-specific death, the hazard ratio (HR) was 0.991 (95% CI: 0.904–1.086, $p = 0.843$). Similarly, no significant association was observed for death from other causes (HR = 0.952, 95% CI: 0.862–1.051, $p = 0.332$). In contrast, standard clinical factors remained robust predictors of lethal progression, with tumor stage (HR = 1.58, $p < 0.001$) and histological grade (HR = 1.24, $p = 0.025$) demonstrating strong prognostic value (). The proportional hazards assumption was tested for all variables; while the ERS satisfied the assumption for disease-specific death ($p = 0.157$), violations were noted for hormone therapy ($p = 0.031$) and INTCLUST 10 ($p < 0.001$). Ultimately, the lack of significance for the ERS indicates that it provides no additional prognostic information beyond established clinical and molecular classifications.

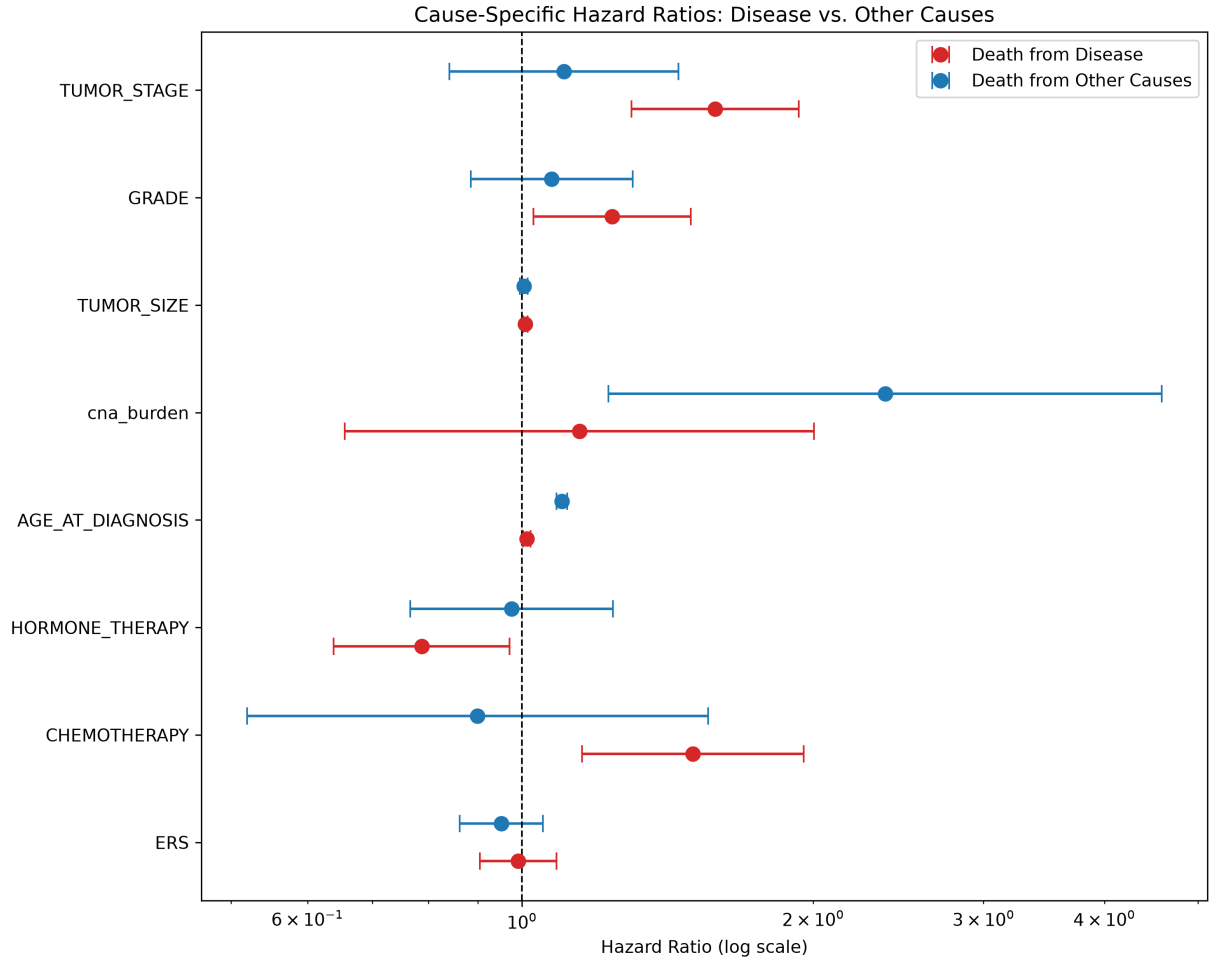


Figure 2: Forest plot displaying cause-specific hazard ratios for death from disease (red) and death from other causes (blue) evaluated via Cox proportional hazards models. The Epigenetic Rewiring Score (ERS) does not emerge as a significant independent predictor for either mortality cause, with its confidence intervals crossing the null threshold. Conversely, standard clinical covariates, notably `TUMOR_STAGE` and `GRADE`, remain strong predictors of disease-specific mortality. The models adjust for additional factors including `AGE_AT_DIAGNOSIS`, `CHEMOTHERAPY`, `HORMONE_THERAPY`, `cna_burden`, and `TUMOR_SIZE`. These results demonstrate that while the ERS successfully quantifies epigenetic-transcriptomic discordance, it provides no independent prognostic value beyond established clinical variables.

3.3 Mutational interaction analysis

Given that the prognostic impact of epigenetic alterations can be context-dependent, we investigated whether the effect of the ERS was conditional on the mutational status of key breast cancer drivers, specifically *TP53*, *PIK3CA*, *GATA3*, and *MAP3K1*. Interaction models incorporating the main effects of the ERS, the binary mutation status, and their interaction term were fitted for disease-specific survival.

No significant synergistic effects were identified between the ERS and any of the evaluated driver mutations after multiple testing considerations (). We observed a marginal trend for the interaction between the ERS and *MAP3K1* (HR = 1.308, $p = 0.092$), suggesting a potential context where high epigenetic rewiring might be more detrimental in a *MAP3K1*-mutant background. However, this trend did not reach the threshold for statistical significance. Interactions with *PIK3CA* (HR = 1.144, $p = 0.158$), *TP53* (HR = 0.891, $p = 0.222$), and *GATA3* (HR = 0.959, $p = 0.791$) were similarly null, demonstrating that the prognostic neutrality of the ERS is consistent across these major genetic backgrounds.

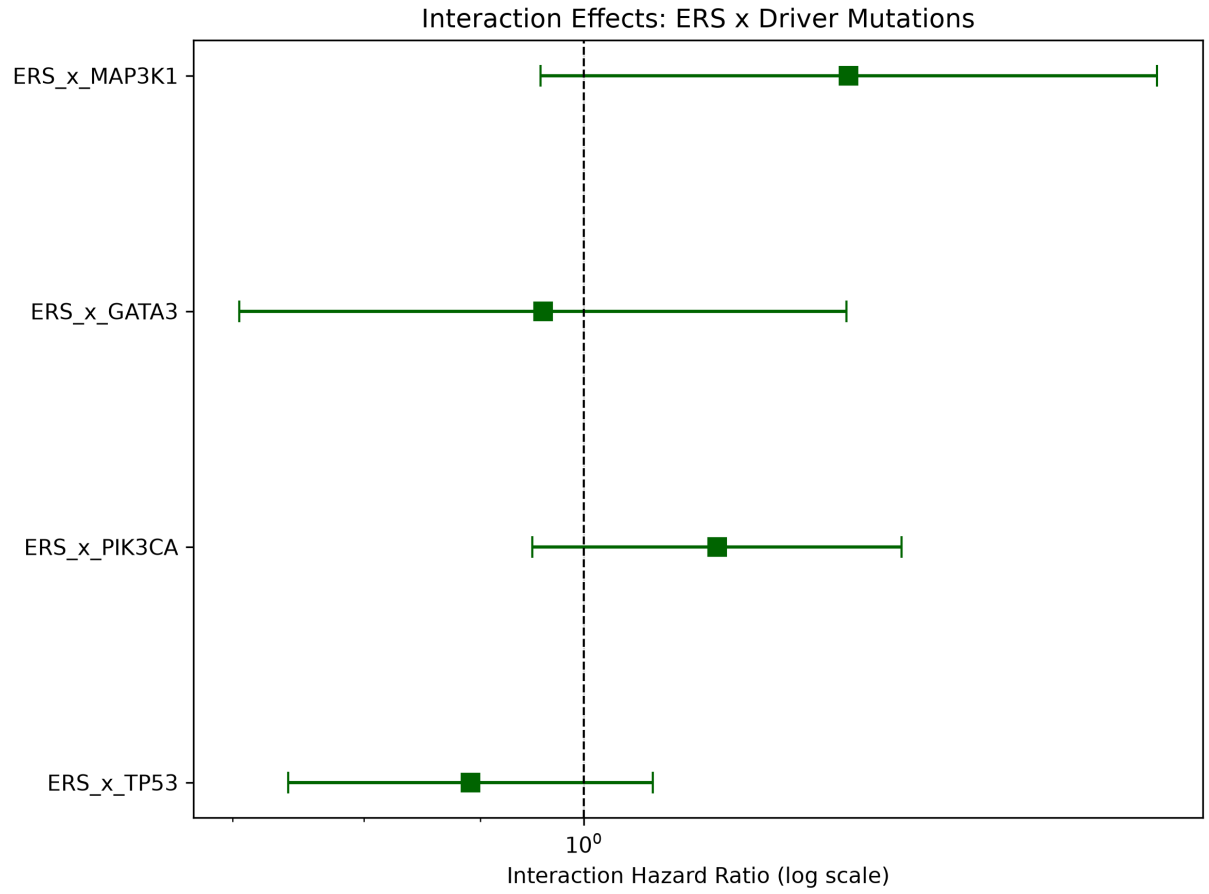


Figure 3: Forest plot displaying the interaction effects between the Epigenetic Rewiring Score (ERS) and key breast cancer driver mutations (*MAP3K1*, *GATA3*, *PIK3CA*, and *TP53*) on disease-specific survival. The interaction hazard ratios and their 95% confidence intervals are plotted on a log scale, with the vertical dashed line representing a hazard ratio of 1 (no interaction effect). The analysis reveals no significant synergies between the ERS and the mutational status of these drivers. While a marginal trend is observed for the ERS × *MAP3K1* interaction, suggesting that a high ERS may be more detrimental in a *MAP3K1*-mutant background, this does not reach the threshold for statistical significance. Ultimately, the plot demonstrates that the prognostic effect of the ERS is not conditional on the presence of these key driver mutations.

3.4 Reproducibility across accrual cohorts

To ensure that the observed lack of association was not an artifact of the aggregated dataset, we performed an internal validation by evaluating the ERS effect independently across the five distinct molecular accrual partitions of the METABRIC cohort. Adjusted cause-specific Cox models were fitted within each partition, and the results were synthesized using a random-effects meta-analysis.

The effect of the ERS on disease-specific survival was remarkably consistent and null across all five cohorts (Cohort 1: Coef = -0.067, $p = 0.537$; Cohort 2: Coef = 0.080, $p = 0.532$; Cohort 3: Coef = -0.008, $p = 0.906$; Cohort 4: Coef = -0.093, $p = 0.580$; Cohort 5: Coef = 0.231, $p = 0.212$;). The pooled hazard ratio was 1.002 (95% CI: 0.912–1.101, $p = 0.966$). Crucially, the heterogeneity across the batches was measured at $I^2 = 0.00\%$ (Cochran's $Q = 2.65$, df = 4, $p = 0.97$). This complete lack of variance between the accrual cohorts confirms that the absence of an independent prognostic association between methylation-dependent transcriptional residual variance and lethal progression is a highly stable and reproducible property of this dataset.

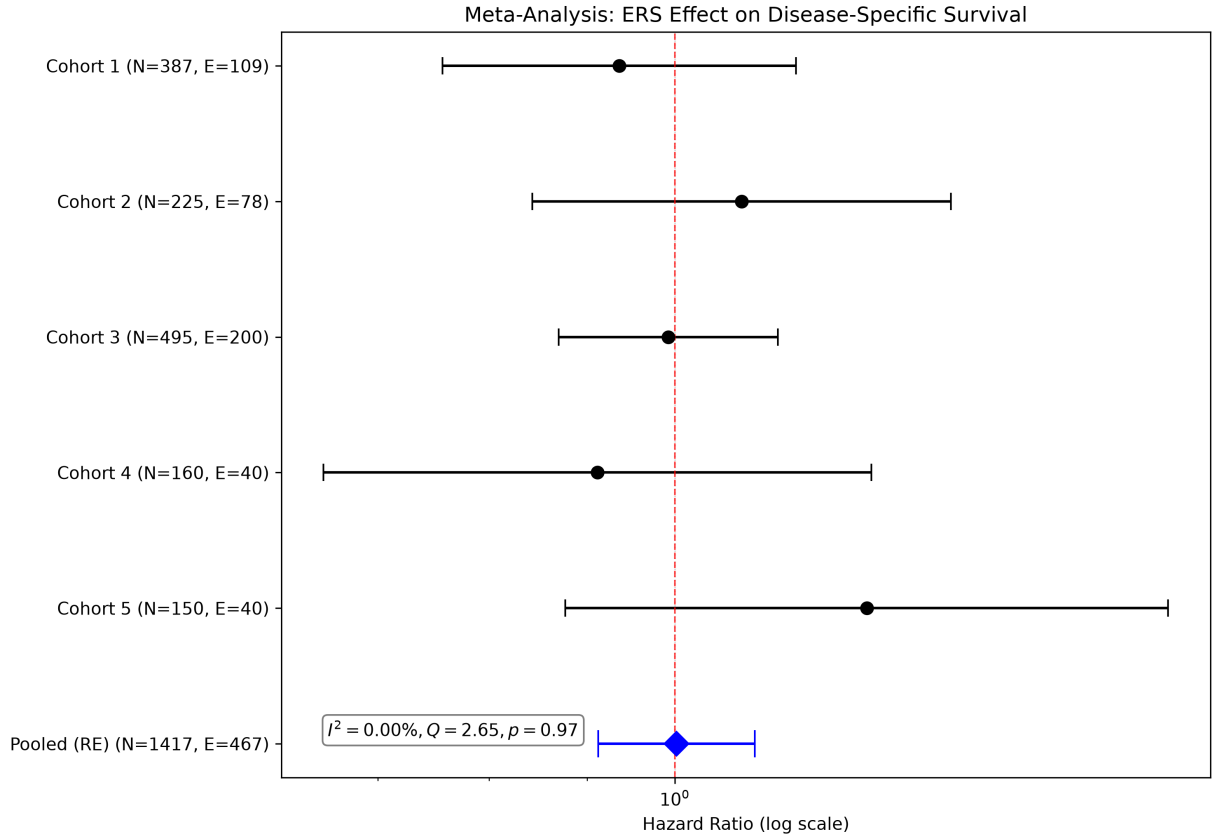


Figure 4: Forest plot detailing a random-effects meta-analysis of the Epigenetic Rewiring Score (ERS) effect on disease-specific survival across five independent molecular accrual cohorts. Black markers indicate the hazard ratio and 95% confidence intervals for each individual cohort, while the blue diamond represents the pooled estimate. The vertical dashed red line denotes a hazard ratio of 1 (no effect). As demonstrated by the pooled estimate and the individual cohort intervals crossing the null line, the ERS does not provide significant independent prognostic value for lethal progression. The accompanying heterogeneity metrics ($I^2 = 0.00\%$, $Q = 2.65$, $p = 0.97$) reveal a complete lack of variance between the batches, confirming that the absence of association between ERS and disease-specific mortality is a highly stable property across the METABRIC cohort.

4 Conclusions

4.1 Summary of findings

In this study, we addressed the critical analytical challenge of confounding molecular layers in breast cancer by developing a high-resolution framework to isolate dosage-independent epigenetic rewiring. By regressing gene expression against discrete copy-number states and global genomic instability, we successfully extracted transcriptional residuals that reflect true epigenetic modulation. The aggregation of these signals into the Epigenetic Rewiring Score (ERS) provided a biologically grounded, patient-level metric of methylation-dependent transcriptomic discordance.

Despite successfully capturing this isolated epigenetic signal, our comprehensive cause-specific survival analyses revealed that the ERS does not provide independent prognostic value for lethal progression. The score failed to predict disease-specific mortality or death from competing causes beyond the prognostic baseline established by standard clinical covariates, such as tumor stage and histological grade, and the Integrative Cluster (INTCLUST) taxonomies. Furthermore, this prognostic neutrality was consistent across major genetic backgrounds, as we observed no significant synergistic interactions between the ERS and key driver mutations, including *TP53*, *PIK3CA*, *GATA3*, and *MAP3K1*.

4.2 Biological and clinical implications

The defining strength of these findings lies in their remarkable statistical stability. A random-effects meta-analysis across five independent molecular accrual cohorts confirmed the null association with zero heterogeneity. This absolute lack of variance across batches underscores that the absence of an independent prognostic effect is not a statistical artifact, but a highly reproducible biological reality within the METABRIC cohort.

Ultimately, these results indicate that while methylation-dependent transcriptional residual variance is a robustly measurable molecular state, it does not act as a primary, independent driver of breast cancer lethality. Epigenetic-transcriptomic discordance, once rigorously decoupled from the dominant transcriptional effects of macroscopic aneuploidy, appears to offer no additional clinical utility for survival stratification. This study highlights the critical importance of controlling for broad genomic instability when evaluating the prognostic potential of epigenetic biomarkers, demonstrating that true epigenetic rewiring is largely superseded by established clinical and molecular classifications in predicting patient outcomes.