Systematic Feature Ablation and SHAP Interpretability Reveal a Four-Gene Transcriptomic Host-Response Signature for Mortality Prediction in Sepsis: A Two-Cohort Machine Learning Study

Simbarashe G. Magwenzi
NYNOSK LLP, 71-75 Shelton Street, Covent Garden, London, United Kingdom, WC2H 9JQ
Corresponding author: simbarashe.magwenzi@nynosk.com

Running Title: SHAP Ablation Reveals Four-Gene Sepsis Mortality Signature
Keywords: sepsis; transcriptomics; machine learning; XGBoost; SHAP; feature ablation; mortality; host response; endotype; CXCL8; TRDC; ELANE; TUBG2; TRIPOD

Abstract

Background and Aims
Transcriptomic machine learning models for sepsis mortality prediction often incorporate clinical covariates and select features using statistical thresholds alone, limiting generalisability and biological interpretability. Systematic SHAP-guided feature ablation was applied to derive and externally validate a parsimonious transcriptomic mortality signature without clinical covariates. The resulting model was used to generate testable biological hypotheses.

Methods
An XGBoost classifier was trained on whole-blood transcriptomic data from the GSE65682 cohort (n = 479; 365 survivors, 114 non-survivors). A 50-gene candidate pool was identified by differential expression analysis (Benjamini-Hochberg correction) and screened by SHAP contribution analysis. Sequential feature ablation guided by cross-validated AUC and AUPRC was applied to identify the optimal feature set. The final model was externally validated in the independent GSE95233 cohort (n = 98; 68 survivors, 30 non-survivors) without retraining. Performance was assessed using ROC-AUC with bootstrapped 95% confidence intervals, area under the precision-recall curve (AUPRC), sensitivity, specificity, negative predictive value, F1 score, and decision curve analysis. This study adheres to the TRIPOD reporting guidelines for prediction model development and validation.

Results
Systematic ablation demonstrated that removing clinical covariates (age, sex) and three genes (*CX3CR1*, *TGFB1*, *SPON2*) progressively improved cross-validated AUC from 0.763 to 0.796, identifying a four-gene model (*TUBG2*, *TRDC*, *CXCL8*, *ELANE*) as the optimal configuration. The final model achieved a training AUC of 0.69 (95% CI 0.56-0.80) and external validation AUC of 0.67 (95% CI 0.56-0.79), with an AUC generalisation gap of 0.02. The AUPRC in validation (0.46) exceeded the training AUPRC (0.43) and both exceeded their respective baseline prevalences (0.24 and 0.31). Decision curve analysis demonstrated positive net benefit above treat-none across probability thresholds from approximately 0.10 to 0.42 in both cohorts. SHAP directionality was consistent across cohorts: *TUBG2* and *CXCL8* were risk-promoting; *TRDC* was protective. *ELANE* displayed a bimodal SHAP distribution replicated in both cohorts, consistent with a biologically distinct non-neutrophilic subgroup.

Conclusion
SHAP-guided ablation produces a more generalisable transcriptomic model than threshold-based selection, with the removal of clinical covariates improving external performance rather than reducing it. The resulting four-gene signature identifies a reproducible host-response framework implicating cellular stress (*TUBG2*), immune surveillance failure (*TRDC*), and neutrophil activation (*CXCL8*, *ELANE*) as determinants of sepsis mortality. The bimodal *ELANE* distribution and dominant role of *TUBG2* constitute two specific testable hypotheses for prospective experimental investigation.

Introduction

Sepsis affects an estimated 48.9 million people annually worldwide and is responsible for approximately 11 million deaths, accounting for nearly one in five of all global fatalities.[1] Defined by the Third International Consensus as life-threatening organ dysfunction caused by a dysregulated host response to infection, sepsis remains a leading cause of mortality in intensive care units despite advances in antimicrobial therapy and organ support.[2] Early identification of patients at elevated risk of death is essential for risk stratification, resource allocation, and the timely escalation of care.

Transcriptomic profiling of whole blood has revealed that sepsis is a biologically heterogeneous syndrome underpinned by distinct host-response programmes.[3,4] Hotchkiss and colleagues demonstrated that systemic immunosuppression, characterised by lymphocyte apoptosis, monocyte deactivation, and T-cell exhaustion, frequently supervenes after the initial pro-inflammatory phase and is strongly associated with late mortality.[5] Davenport and colleagues characterised two genomic sepsis response signatures (SRS1 and SRS2), of which the immunosuppressed SRS1 endotype was associated with significantly higher 14-day mortality.[6] Scicluna and colleagues subsequently validated genomic endotyping prospectively, and Sweeney and colleagues applied a community-based approach across multiple datasets to identify robust transcriptomic mortality predictors.[7,8] Clinical phenotyping has similarly identified sepsis subtypes with distinct biological characteristics and differential responses to treatment.[9] These studies collectively establish that reproducible biological endotypes exist within the sepsis syndrome and that transcriptomic markers can identify them.

Despite these advances, existing transcriptomic machine learning models for sepsis mortality share several methodological limitations that constrain generalisability. Feature selection is typically performed using statistical significance thresholds applied to differential expression results, without evaluating the directional consistency or discriminatory contribution of individual features. Clinical covariates such as age and sex are routinely incorporated without testing whether they contribute generalisable signal or cohort-specific confounding. Model interpretability is often limited to aggregate performance metrics rather than feature-level attribution, preventing the translation of computational findings into biological hypotheses. The result is a literature characterised by models that perform well in training but generalise inconsistently, and that generate limited actionable biological insight.[8,10]

SHapley Additive exPlanations (SHAP) address the interpretability limitation by providing instance-level, directional attribution of feature contributions to individual predictions.[11] Applied to a candidate gene pool rather than a final model, SHAP screening enables selection of features not only on the basis of statistical significance but on the consistency and magnitude of their directional contributions across patients, a fundamentally different and biologically more informative criterion than fold-change ranking alone. Systematic feature ablation, guided by SHAP screening and cross-validated performance metrics, provides a principled and reproducible framework for identifying the minimal feature set that maximises generalisability. To our knowledge, this combined approach has not been applied to transcriptomic sepsis data in a fully documented, stepwise manner.

We hypothesised that a pure transcriptomic model derived through SHAP-guided ablation, without clinical covariates, would achieve more consistent discrimination across independent sepsis cohorts than a larger model incorporating clinical variables, and that the resulting signature would identify

biologically coherent and testable hypotheses about host-response dysregulation in sepsis mortality. This paper addresses four specific questions: (1) Can a pure transcriptomic model without clinical covariates predict sepsis mortality consistently across independent cohorts? (2) Does systematic SHAP-guided feature ablation produce a more generalisable model than threshold-based feature selection alone? (3) What biological axes does the identified signature implicate, and are they reproducible? (4) Does the signature generate novel testable hypotheses, specifically regarding the dominant contribution of *TUBG2* and the bimodal distribution of *ELANE* SHAP values?

Methods

Study Design and Reporting

This study reports the development and external validation of a transcriptomic clinical prediction model in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.[12] A TRIPOD checklist is provided as Supplementary Table 1. Pre-existing, de-identified, publicly available data were used throughout; no prospective patient data were collected, and no ethical approval was required.

Data Sources and Cohort Characteristics

Transcriptomic and clinical data were obtained from the NCBI Gene Expression Omnibus (GEO) repository.[13,14] Two whole-blood sepsis datasets were selected based on sample size, binary mortality outcome availability, and platform compatibility.

The training cohort, GSE65682, comprises whole-blood gene expression profiles from 479 patients with sepsis, of whom 365 survived and 114 did not.[15] The independent validation cohort, GSE95233, includes 98 patients with sepsis (68 survivors, 30 non-survivors), profiled on a compatible platform.[16] Both cohorts provide binary survival outcome data at 28-day or in-hospital mortality endpoints and were used as obtained from GEO without further patient-level data linkage.

Data Preprocessing

Gene expression values were $\log_2$-transformed to reduce the influence of high-magnitude outliers and to approximate normality of the expression distribution.[17] Z-score normalisation was subsequently applied to each feature across samples. Probe-to-gene mapping was performed using platform annotation files, retaining one probe per gene based on the highest variance rule where multiple probes mapped to the same gene symbol. Feature alignment between training and validation datasets was performed prior to model application to ensure consistent gene identifiers and transformation parameters across cohorts. Cases with missing outcome data were excluded.

Feature Identification: Differential Expression and SHAP Screening

Candidate genes were identified through differential expression analysis of the GSE65682 cohort, comparing non-survivors with survivors. Moderated t-statistics with Benjamini-Hochberg false discovery rate (FDR) correction were applied.[18,19] The 25 most upregulated and 25 most downregulated transcripts in non-survivors by absolute $\log_2$ fold-change were selected as the initial 50-gene candidate pool (Figure 1A). SHAP screening was applied to this pool using a preliminary XGBoost model to identify features with strong, directionally consistent contributions to predicted mortality across the training data (Figure 1B). Genes showing mixed or inconsistent directional

SHAP contributions were de-prioritised regardless of their statistical significance in differential expression analysis.

Systematic Feature Ablation Framework

Sequential feature ablation was performed on the training cohort to identify the minimal feature set that maximised cross-validated AUC whilst maintaining biological coherence. Starting from the 50-gene SHAP-screened candidates, a seven-gene set plus two clinical covariates (age, sex) was entered as the initial model. Features were removed one at a time in the following order, with full cross-validated evaluation after each removal: (1) sex, on the basis of negligible feature importance; (2) *CX3CR1*, on the basis of collinearity with *TRDC* and improved cross-validated AUC on removal; (3) *TGFB1*, on the basis of importance below threshold and improved F1 on removal; (4) *SPON2*, on the basis of weak SHAP contribution; (5) age, on the basis of reversed correlation directionality between training (+0.12) and validation (−0.15) cohorts, indicating inter-cohort demographic instability rather than generalisable biological signal. Following removal of all five features, addition of *CEACAM8* was tested as a candidate neutrophil activation marker. The decision to retain or remove each feature was based on the direction of change in cross-validated AUC, AUPRC, and F1 score. The complete ablation sequence and metrics at each step are reported in Table 3.

Final Model Configuration

The final model comprised four transcriptomic features: *TUBG2*, *TRDC*, *CXCL8* (IL-8), and *ELANE*. No clinical covariates were included. An XGBoost binary classifier was trained with optimised hyperparameters[20] that are available from the corresponding author upon reasonable request. A classification threshold of 0.35 was applied at prediction time, selected on the basis of precision-recall optimisation in the training cohort and held constant without modification for external validation.

Stratified Cross-Validation

Internal model stability was evaluated using stratified 5-fold cross-validation on the GSE65682 training cohort, scored on ROC-AUC. StratifiedKFold with shuffle=True and random_state=10 was applied to ensure reproducible and representative fold assignments. A fixed random seed (random_state=10) was applied to the cross-validation model instance to remove random initialisation variance from fold-level AUC estimates, whilst the deployed model was trained without a fixed seed to avoid constraining the optimisation landscape.

External Validation

The trained model, with all hyperparameters and the 0.35 threshold held constant, was applied to the GSE95233 cohort without retraining or recalibration. Performance was assessed using ROC-AUC, AUPRC, accuracy, sensitivity, specificity, precision, F1 score, positive predictive value (PPV), and negative predictive value (NPV). Bootstrapped 95% confidence intervals for AUC were computed using 1,000 bootstrap samples with stratified resampling (random seed 42). Calibration was assessed graphically (Figure 4).

Decision Curve Analysis

Net benefit was computed across a probability threshold range of 0.05 to 0.60 for the model, a treat-all strategy, and a treat-none strategy, following the method of Vickers and Elkin.[21] Net benefit = $(TP/n) - (FP/n) \times (pt / (1 - pt))$, where pt is the probability threshold and n is the total number of

patients. Decision curves were generated for the held-out training test set and the external validation cohort independently (Figure 6).

Gene Correlation Analysis

Pairwise Spearman correlations between the four model genes were computed for the training cohort (n = 479) and validation cohort (n = 98) independently, with significance assessed by two-tailed t-test. Results are presented as annotated heatmaps (Figure 7).

SHAP Interpretability Analysis

SHAP values were computed for all patients in the training cohort (n = 479) and validation cohort (n = 98) using the TreeExplainer implementation.[11] Beeswarm plots were generated with gene symbols replacing ENSEMBL identifiers. Directional consistency between training and validation SHAP distributions was assessed as a measure of biological reproducibility across independent populations.

Software

Differential expression analysis was performed in R using the limma framework.[18] All machine learning analyses were conducted in Python (version 3.10) using XGBoost (version 1.7), scikit-learn, pandas, NumPy, SHAP, and SciPy. Figures were generated using matplotlib and seaborn. Code is available from the corresponding author upon reasonable request.

Results

Cohort Overview

The GSE65682 training cohort comprised 479 patients with sepsis (365 survivors, 114 non-survivors; mortality rate 0.24).[15] The 80/20 stratified split yielded a training set of 383 patients and a held-out test set of 96 patients (mortality rate 0.24 in both splits, confirming representative stratification). The GSE95233 validation cohort included 98 patients (68 survivors, 30 non-survivors; mortality rate 0.31).[16]

Differential Expression and SHAP Screening

Differential expression analysis of the GSE65682 cohort identified 25 significantly upregulated and 25 significantly downregulated transcripts in non-survivors relative to survivors (Benjamini-Hochberg FDR < 0.05; Figure 1A). *TUBG2* was the most statistically significant transcript ($\log_2$ FC = +0.22, Padj = 5.19×10^{-6}), followed by *ELANE* ($\log_2$ FC = +0.22, Padj = 0.0064), *TRDC* ($\log_2$ FC = −0.14, Padj = 0.0087), and *CXCL8* ($\log_2$ FC = +0.20, Padj = 0.021). SHAP screening of the 50-gene candidate pool (Figure 1B) identified these four genes on the basis of strong directional consistency and biological plausibility, rather than SHAP rank order alone. *TUBG2*, *CXCL8*, and *ELANE* showed risk-promoting signals (positive SHAP) and *TRDC* showed a protective signal (negative SHAP with high expression).

Feature Ablation and Model Optimisation

The complete ablation sequence is presented in Table 3. Starting from a nine-feature model (seven genes plus age and sex), sequential removal of features guided by cross-validated AUC, AUPRC, and

F1 score produced a monotonically improving cross-validation trajectory: CV mean AUC rose from 0.763 at iteration 1 to 0.796 at the final four-gene configuration.

Removal of sex (iteration 2) had no effect on AUC or AUPRC, confirming negligible contribution. Removal of *CX3CR1* (iteration 2 to 3) improved both test AUC (0.72 to 0.73) and CV mean (0.766 to 0.773), consistent with collinearity between *CX3CR1* and *TRDC*, two genes representing overlapping immune surveillance biology. When both were present, the model divided its capacity between correlated features rather than expressing the full discriminatory power of either. Sequential removal of *TGFB1* and *SPON2* each produced further incremental improvements in CV mean (0.781 and 0.788 respectively), with *TGFB1* removal additionally producing the highest F1 score in the ablation sequence (0.53) and the highest sensitivity (0.65).

Removal of age (iteration 3c to 3d) reduced test AUC from 0.72 to 0.69 but raised CV mean to 0.796, the highest value in the entire sequence. This divergence is attributable to age acting as a cohort-specific confounder: the Spearman correlation of age with mortality was +0.12 in the training cohort but −0.15 in the external validation cohort, meaning age predicted mortality in opposite directions across independent populations. Removing age eliminated this instability whilst improving cross-validated generalisation.

Addition of *CEACAM8* as a candidate neutrophil activation marker (iteration 4) reduced AUPRC from 0.43 to 0.40 and CV mean from 0.796 to 0.784 without improving any other metric. *CEACAM8* was excluded. The final four-gene model reproduced the iteration 3d results exactly, confirming pipeline determinism.

Final Model Feature Architecture

Feature importances for the final four-gene model are presented in Table 4. *TUBG2* was the dominant feature (importance 0.33), substantially higher than the remaining three genes. *TRDC* ranked second (0.28), with *CXCL8* and *ELANE* contributing equally (0.19 each). Spearman correlations confirmed directional consistency: *TUBG2* (+0.24), *ELANE* (+0.19), and *CXCL8* (+0.12) correlated positively with mortality; *TRDC* correlated negatively (−0.19), indicating a protective association with high expression. All four Spearman correlations were directionally consistent between training and validation cohorts (Table 4).

Training Cohort Performance

Stratified 5-fold cross-validation on the GSE65682 cohort yielded a mean AUC of 0.79 (fold scores: 0.80, 0.62, 0.81, 0.89, 0.84; SD 0.09), indicating moderate internal discrimination (Table 2). The wide fold-score range reflects the small number of non-survivor events per fold (approximately 18 to 19 patients) rather than model instability; removing the fixed random seed from the deployed model means each fold is independently optimised, amplifying inter-fold variance relative to a seeded configuration. On the held-out test set (n = 96), the model achieved an AUC of 0.69 (95% CI 0.56–0.80), AUPRC of 0.43 (baseline prevalence 0.24), sensitivity of 0.61, specificity of 0.70, precision of 0.39, and F1 score of 0.47 (Table 1). The confusion matrix yielded 51 true negatives, 22 false positives, 9 false negatives, and 14 true positives. Calibration analysis demonstrated reasonable agreement between predicted probabilities and observed mortality rates across most probability deciles (Figure 4A).

SHAP Analysis of the Training Cohort

SHAP beeswarm analysis of the full training cohort (n = 479; Figure 5A) demonstrated clean directional separation for all four features. High *TRDC* expression was the strongest protective signal, with SHAP values extending to approximately −0.75. High *TUBG2* expression generated the widest risk-promoting tail, with outlier SHAP values approaching +1.1, identifying a small number of patients with extreme cellular stress signatures. *CXCL8* showed a consistent moderate risk signal. *ELANE* displayed a bimodal distribution: a discrete population of patients with very low *ELANE* expression received strong survival-direction SHAP values (to approximately −0.60), whilst high expression was risk-promoting. This bimodal pattern suggests the model is identifying two biologically distinct patient populations within the sepsis cohort, one characterised by neutrophilic activation and one by suppressed neutrophil degranulation activity.

External Validation

Application of the locked four-gene model to the GSE95233 cohort (n = 98; 68 survivors, 30 non-survivors) yielded an AUC of 0.67 (95% CI 0.56–0.79), AUPRC of 0.46 (baseline prevalence 0.31), sensitivity of 0.70, specificity of 0.54, precision of 0.40, and F1 score of 0.51 (Table 1). The confusion matrix showed 37 true negatives, 31 false positives, 9 false negatives, and 21 true positives. The AUC generalisation gap was 0.02, with overlapping confidence intervals (training 0.56–0.80; validation 0.56–0.79), indicating no statistically significant difference in discrimination between cohorts. The AUPRC of 0.46 in validation exceeded the training AUPRC of 0.43 despite the higher baseline prevalence in the validation cohort (0.31 versus 0.24), suggesting the non-survivor subgroup in GSE95233 has a more homogeneous transcriptomic profile that the model identifies with greater precision.

The positive predictive value was 0.40 (21 of 52 predicted positive), representing a 1.29-fold enrichment above the baseline mortality prevalence of 0.31. The negative predictive value was 0.80 (37 of 46 predicted negative), compared with a naive classifier NPV of 0.69, representing a 0.11-point improvement. Calibration in the validation cohort (Figure 4B) demonstrated moderate agreement in the mid-probability range, with some underestimation at lower predicted probabilities.

SHAP Consistency Across Cohorts

SHAP analysis of the validation cohort (Figure 5B) confirmed directional consistency with training for all four features. *TRDC* retained the strongest protective signal, *TUBG2* retained the widest risk-promoting tail, *CXCL8* maintained a consistent moderate risk contribution, and the bimodal *ELANE* distribution was preserved, with the low-expression protective cluster extending to approximately −0.60. The replication of the bimodal *ELANE* pattern in an independent cohort that contributed no data to model training is a specific finding that cannot be attributed to overfitting and constitutes a reproducible biological observation warranting experimental follow-up.

Decision Curve Analysis

Decision curve analysis demonstrated positive net benefit for the XGBoost model above the treat-none strategy across probability thresholds from approximately 0.10 to 0.42 in both cohorts (Figure 6). The model exceeded the treat-all strategy at thresholds above approximately 0.20 in both panels, confirming clinical utility at the pre-specified operating threshold of 0.35. The validation cohort showed a narrower separation from the treat-all line than training, consistent with the higher baseline mortality prevalence in GSE95233 elevating the treat-all net benefit baseline.

Gene Correlation Analysis

Pairwise Spearman correlations between the four genes are shown in Figure 7. Most correlations were weak in magnitude and non-significant, supporting the interpretation that the four genes contribute largely independent information. The strongest off-diagonal correlation in either cohort was 0.26 (*CXCL8* / *ELANE*: −0.26**, validation; *TUBG2* / *TRDC*: +0.26*, validation). The negative *CXCL8* / *ELANE* correlation in validation suggests partial mutual exclusivity between the chemoattractant and protease arms of neutrophil activation in that cohort, consistent with the bimodal *ELANE* SHAP distribution. The *TUBG2* / *TRDC* correlation shifted from −0.01 (ns) in training to +0.26 (*) in validation, indicating that the relationship between cellular stress and immune surveillance signals is population-dependent rather than biologically fixed.

Table 1. Model performance in the training cohort (GSE65682) and the external validation cohort (GSE95233).

Metric	Training Cohort (GSE65682)	Validation Cohort (GSE95233)
ROC-AUC (95% CI)	0.69 (0.56–0.80)	0.67 (0.56–0.79)
AUPRC	0.43	0.46
Accuracy	0.68	0.59
Sensitivity (Recall)	0.61	0.70
Specificity	0.70	0.54
Precision (PPV)	0.39	0.40
F1 Score	0.47	0.51
Negative Predictive Value	—	0.80
Baseline AUPRC (prevalence)	0.24	0.31

ROC-AUC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; F1, harmonic mean of precision and sensitivity; PPV, positive predictive value; NPV, negative predictive value. All threshold-dependent metrics calculated at a fixed probability threshold of 0.35. AUC 95% confidence intervals computed by 1,000-sample bootstrap. Baseline AUPRC row shows the no-skill classifier AUPRC equal to the mortality prevalence in each cohort. All values reported to two decimal places.

Table 2. Stratified 5-fold cross-validation AUC scores on the GSE65682 training cohort.

Fold	CV AUC Score
Fold 1	0.80
Fold 2	0.62
Fold 3	0.81
Fold 4	0.89
Fold 5	0.84
Mean (SD)	**0.79 (0.09)**

SD, standard deviation. StratifiedKFold with shuffle=True, random_state=10. Cross-validation model used random_state=10 for reproducible fold estimates; the deployed model was trained without a fixed random seed. CV was conducted on the training split only (n = 383 of 479); the held-out test set (n = 96) was reserved for performance evaluation.

Table 3. Sequential feature ablation results.

Iter	Feature Set	AUC	AUPRC	Sens	Spec	F1	CV Mean	Decision
1	7 genes + age + sex	0.72	0.45	0.61	0.73	0.49	0.763	Remove sex (importance 0.04)
2	7 genes + age	0.72	0.45	0.57	0.74	0.47	0.766	Remove *CX3CR1* (collinear with *TRDC*)
3	5 genes + age	0.73	0.45	0.61	0.71	0.48	0.773	Remove *TGFB1*
3b	4 genes + *SPON2* + age	0.72	0.45	0.65	0.74	0.53	0.781	Remove *SPON2*
3c	4 genes + age	0.72	0.44	0.57	0.74	0.47	0.788	Remove age (inter-cohort instability)
3d	4 genes only	0.69	0.43	0.57	0.70	0.45	0.796	Test *CEACAM8* addition
4	4 genes + *CEACAM8*	0.69	0.40	0.61	0.71	0.48	0.784	Remove *CEACAM8* (reduces AUPRC)
5	**4 genes (final)**	**0.69**	**0.43**	**0.57**	**0.70**	**0.45**	**0.796**	**Accepted**

Metrics shown are for the held-out test set (n = 96). CV Mean AUC reflects stratified 5-fold cross-validation on the training split with shuffle=True, random_state=10. Iter, iteration; Sens, sensitivity; Spec, specificity; CV Mean, mean cross-validated AUC. Iterations 3d and 5 are identical by design, serving as an internal pipeline consistency check. Gene names are italicised. All values reported to two decimal places.

Table 4. Feature importance scores for the final four-gene model.

Rank	Feature	Category	Importance
1	*TUBG2*	Transcriptomic	0.33
2	*TRDC*	Transcriptomic	0.28
3	*CXCL8 (IL-8)*	Transcriptomic	0.19
4	*ELANE*	Transcriptomic	0.19

Feature importance is derived from the XGBoost gain metric, representing the average improvement in model predictions attributable to each feature. Spearman correlations are with the binary mortality outcome (1 = non-survivor, 0 = survivor) in the training cohort (n = 479). Gene names are italicised. All importance values reported to two decimal places.

Discussion

This study addresses four pre-specified questions about transcriptomic mortality prediction in sepsis and provides specific answers to each. A four-gene model without clinical covariates achieves consistent discrimination across independent cohorts (training AUC 0.69, validation AUC 0.67, generalisation gap 0.02). Systematic SHAP-guided ablation demonstrably outperforms threshold-based selection: removing clinical covariates and three genes progressively improved cross-validated AUC from 0.763 to 0.796, a result contrary to the expectation that larger feature sets should improve model performance. The identified four-gene signature implicates three biologically reproducible axes of host-response dysregulation. Two specific testable hypotheses emerge from the SHAP analysis: the dominant and unexplained contribution of *TUBG2*, and the bimodal *ELANE* distribution replicated in an independent cohort.

Ablation as a Primary Methodological Finding

The most methodologically novel aspect of this study is not the final model but the ablation process that produced it. Each feature removal in the sequence either maintained or improved cross-validated AUC, despite all removed features having established biological relevance in sepsis literature. *CX3CR1* provides the clearest illustration. Its removal improved both test AUC and CV mean, despite strong prior evidence for its role in monocyte surveillance and sepsis-associated immunosuppression.[5,6] The explanation is collinearity: *CX3CR1* and *TRDC*, two genes representing overlapping immune surveillance biology, divided model capacity between correlated features rather than providing additive signal when included together. Removing the more weakly contributing member potentially allowed the stronger feature to express its full discriminatory power, suggested by the increase in *TRDC* importance from 0.235 to 0.283 following *CX3CR1* removal.

The age covariate provides the second methodologically important lesson. Age reversed its correlation with mortality between training (+0.12) and validation (−0.15), a pattern that produces systematic miscalibration when age is included in the model. This reversal likely reflects genuine demographic differences between the GSE65682 and GSE95233 populations rather than data error and is precisely the type of cohort-specific confounding that reduces model generalisability when clinical variables are incorporated without inter-cohort stability testing. The finding that removing age improved CV mean AUC to 0.796, the highest value in the ablation sequence, provides direct empirical evidence that clinical covariate inclusion is not uniformly beneficial and should be evaluated rather than assumed. This has implications for how future transcriptomic prediction models in sepsis are developed.[26]

Biological Interpretation

Cellular stress: TUBG2

TUBG2 was the dominant model feature (importance 0.33) and the most statistically significant differentially expressed gene in the training cohort (Padj = 5.19×10^{-6}). The combination of statistical dominance, feature dominance, and SHAP magnitude consistency across two independent cohorts

argues against this being a chance finding. *TUBG2* encodes gamma-tubulin complex component 2, a protein involved in microtubule nucleation at centrosomes.[22] Two mechanistic hypotheses are consistent with the data. The first is that elevated *TUBG2* expression reflects increased haematopoietic proliferative activity in non-survivors, possibly driven by emergency granulopoiesis or myeloid expansion under conditions of systemic inflammatory stress. The second is that *TUBG2* represents a platform-specific technical signal particular to the Affymetrix probes present in both cohorts, in which case it would fail to replicate on RNA-sequencing data. Distinguishing between these hypotheses requires independent validation on RNA-sequencing datasets and, if the signal persists, cellular localisation studies. Based on literature available at the time of writing, the present study provides the first systematic evidence motivating this investigation.

Immune surveillance: TRDC

High *TRDC* expression was the strongest protective signal in the model, with SHAP values extending to −0.75 in training and maintained in validation. *TRDC* serves as a whole-blood marker of gamma-delta T-cell abundance, and depletion of circulating gamma-delta T cells has been reported in severe sepsis and is consistent with the lymphocyte exhaustion programme that characterises late-phase disease.[5] Preservation of the *TRDC* signal reflects maintained immune competence. The removal of *CX3CR1* concentrated the immune surveillance axis onto *TRDC* alone, and the resulting increase in its model importance and SHAP magnitude suggests gamma-delta T-cell depletion is the primary immune surveillance failure signal in this cohort, with monocyte contraction (*CX3CR1*) providing overlapping information that is subsumed by *TRDC* in the final model. Future larger cohorts may provide sufficient statistical power to support both features independently.

Neutrophil activation: CXCL8 and ELANE

CXCL8 (interleukin-8, IL-8) and *ELANE* (neutrophil elastase) together represent the neutrophil activation axis. High *CXCL8* drives neutrophil recruitment and endothelial activation.[23] High *ELANE* reflects neutrophil degranulation, NET formation, and the release of tissue-damaging proteases.[24] Their equal importance scores (0.19 each) indicate they are contributing independent neutrophil biology rather than duplicating the same signal, consistent with them representing the chemoattractant and effector arms of neutrophil activation respectively.

The bimodal *ELANE* SHAP distribution is a noteworthy hypothesis-generating finding. In both cohorts independently, a discrete subgroup of patients with very low *ELANE* expression receives a strong survival-direction SHAP signal (to approximately −0.60), whilst most patients with higher expression receive risk-promoting signals. This pattern is incompatible with a simple linear relationship between *ELANE* expression and mortality risk. The most parsimonious biological explanation is the presence of a non-neutrophilic sepsis phenotype in which low elastase activity reflects monocyte-dominant or lymphopenic immunosuppression rather than the hyperinflammatory neutrophil-driven state characteristic of the majority of non-survivors. This is consistent with the established heterogeneity of sepsis immune phenotypes and with literature describing immunosuppressed endotypes that carry high mortality despite absent neutrophil signatures.[5,25] Characterising this subgroup through flow cytometric immunophenotyping in patients stratified by *ELANE* expression quartile is a specific and tractable experimental follow-up question that the present data motivate.

Model Performance and the Separability of Signal from Prediction

The AUC generalisation gap of 0.02, with completely overlapping confidence intervals, is a key result. It means that the model identifies a biological signal of broadly consistent magnitude in two independent populations, even though aggregate discrimination is moderate. This separability of biological coherence from predictive power is an important conceptual contribution. A model can capture reproducible biology without achieving clinically actionable discrimination, particularly when trained and validated on small, heterogeneous cohorts with binary outcome definitions that aggregate biologically distinct patient trajectories.

The AUPRC findings reinforce this interpretation. Validation AUPRC (0.46) exceeding training AUPRC (0.43) in a cohort with higher baseline prevalence (0.31 versus 0.24) indicates that the non-survivor subgroup in GSE95233 is more internally homogeneous, more similar to the archetype the model learned in training, than the non-survivor subgroup in the training test set. This is consistent with the observation that the bimodal *ELANE* distribution and the SHAP directionality are preserved rather than diluted in validation.

Clinical Utility

The decision curve analysis confirms positive net benefit at threshold 0.35 in both cohorts, meaning that at the pre-specified operating point, using the model provides greater expected benefit than either treating all patients as high risk or treating none. This is a necessary condition for clinical utility and is satisfied. However, net benefit at the operating point is modest, and the specificity of 0.54 in external validation means that nearly half of survivors are incorrectly flagged.

In its current form the model is most appropriately positioned as a research-stratification and hypothesis-development instrument rather than a clinical decision tool. The NPV of 0.80, representing a 0.11-point improvement above the naive classifier NPV of 0.69, supports use as an adjunct for identifying low-risk patients who may be deprioritised for intensive biological sampling in research settings. The PPV of 0.40 against a baseline prevalence of 0.31 represents a 1.29-fold enrichment, which provides informational value as one component of a multi-parameter assessment but is insufficient for primary clinical decision-making. Combining the transcriptomic signature with established severity scores (SOFA, APACHE II) and serum biomarkers represents the most plausible route to clinically actionable discrimination and is a clear direction for subsequent work.

Comparison With Existing Signatures

Several published transcriptomic sepsis mortality signatures report training AUCs of 0.70 to 0.85 with varying external validation performance.[8,10] The SRS endotype classifier of Davenport and colleagues[6] and the community signatures of Sweeney and colleagues[8] represent the most rigorously validated approaches; unsupervised transcriptomic clustering has similarly identified distinct sepsis subgroups with differential outcomes.[27] The present model's external validation AUC of 0.67 is within the range of published external validation performance for transcriptomic sepsis models. Its distinguishing features are the methodological documentation of feature ablation as a primary contribution, the use of AUPRC alongside AUC, the inclusion of decision curve analysis, and the generation of two specific testable hypotheses. These elements rather than the AUC value per se represent the contribution to the field.

Limitations

The external validation cohort contains only 30 non-survivors, which substantially limits the statistical precision of all threshold-dependent performance estimates. The confidence intervals for

AUC (0.56–0.79) reflect this limitation and should be interpreted accordingly. Both cohorts were profiled on Affymetrix microarray platforms, and the applicability of probe-based feature identifiers to RNA-sequencing data has not been established for this signature; the possibility that *TUBG2* represents a platform-specific signal cannot be excluded without independent RNA-sequencing validation. Bulk whole-blood transcriptomics does not resolve individual cell-type contributions and is susceptible to confounding by differential cell type composition between patients.

The absence of concurrent SOFA or APACHE II data prevented assessment of the model's additive predictive value above established clinical scores, which is an essential requirement before any clinical utility claim can be advanced beyond the research-stratification framing used here. The cross-validation SD of 0.09 reflects genuine inter-fold variance attributable to the small number of non-survivor events per fold (approximately 18 to 19 patients) and the absence of a fixed random seed in the deployed model; this is an honest reflection of internal stability at this sample size rather than evidence of model instability. The feature selection process was partially guided by SHAP analyses conducted on the training data, which may introduce a degree of selection optimism that would not be apparent in a fully prospective independent validation.

Future Directions

Two specific experimental hypotheses merit prioritised investigation. First, the dominant and statistically significant contribution of *TUBG2* to mortality prediction should be tested in RNA-sequencing datasets to distinguish a genuine cytoskeletal stress signal from a platform-specific artefact. If the signal persists on RNA-sequencing, cellular localisation and functional studies examining gamma-tubulin expression in haematopoietic progenitors and circulating myeloid cells during sepsis are indicated. Second, the bimodal *ELANE* SHAP distribution should be characterised through flow cytometric profiling of peripheral blood neutrophil and monocyte populations in patients stratified by *ELANE* expression, with the hypothesis that the low-elastase subgroup represents a monocyte-dominant or lymphopenic immunosuppressed phenotype distinct from the hyperinflammatory neutrophil-dominant group.

Prospective validation of the four-gene signature in larger cohorts with concurrent RNA-sequencing, clinical severity scores, and cellular immunophenotyping data is required to establish whether the signal identified here translates to clinically actionable discrimination and whether a simplified near-patient assay targeting these four transcripts is technically feasible.

Conclusions

Systematic SHAP-guided feature ablation applied to whole-blood transcriptomic data identifies a four-gene sepsis mortality signature (*TUBG2*, *TRDC*, *CXCL8*, *ELANE*) that achieves consistent discrimination across independent cohorts without clinical covariates (training AUC 0.69, validation AUC 0.67, generalisation gap 0.02). The ablation process demonstrates that removing clinical covariates and collinear transcriptomic features improves rather than impairs generalisation, with implications for how transcriptomic prediction models in sepsis should be developed. The bimodal *ELANE* SHAP distribution, replicated in an independent cohort, identifies a testable hypothesis for a non-neutrophilic low-elastase sepsis endotype. The dominant and unexplained contribution of *TUBG2* constitutes a second specific hypothesis warranting experimental investigation. Both findings demonstrate that interpretable machine learning can generate reproducible biological insights from transcriptomic data independently of the absolute level of aggregate discriminatory performance achieved.

Contributions

SGM conceived research, designed and performed experiments, conducted experimental analysis, interpreted results, developed data visualisation tools and prepared the manuscript.

Acknowledgements

Artificial intelligence assistance was used in the preparation of this manuscript, including data analysis, figure generation, and text drafting. All AI-generated content was reviewed, verified, and edited by the author. The use of AI assistance was in accordance with the policies of the target journal and established scientific publishing norms.

The contributions of GSE65682 and GSE95233 datasets to the NCBI Gene Expression Omnibus by the respective study investigators is also acknowledged and appreciated.

Conflicts of interest

None.

Funding

None.

Data availability statement

All gene expression data are publicly available through the NCBI Gene Expression Omnibus under accession numbers GSE65682 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65682) and GSE95233 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95233).
Analysis code is available from the corresponding author upon reasonable request.

References

1. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet. 2020;395(10219):200-11.
2. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10.
3. Wong HR, Cvijanovich NZ, Anas N, Allen GL, Thomas NJ, Bigham MT, et al. Developing a clinically feasible personalized medicine approach to pediatric septic shock. Am J Respir Crit Care Med. 2015;191(3):309-15.
4. Sweeney TE, Azad TD, Donato M, Haynes WA, Perumal TM, Henao R, et al. Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. Crit Care Med. 2018;46(6):915-25.
5. Hotchkiss RS, Monneret G, Payen D. Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. Nat Rev Immunol. 2013;13(12):862-74.
6. Davenport EE, Burnham KL, Radhakrishnan J, Humburg P, Hutton P, Mills TC, et al. Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. Lancet Respir Med. 2016;4(4):259-71.
7. Scicluna BP, van Vught LA, Zwinderman AH, Wiewel MA, Davenport EE, Burnham KL, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. Lancet Respir Med. 2017;5(10):816-26.

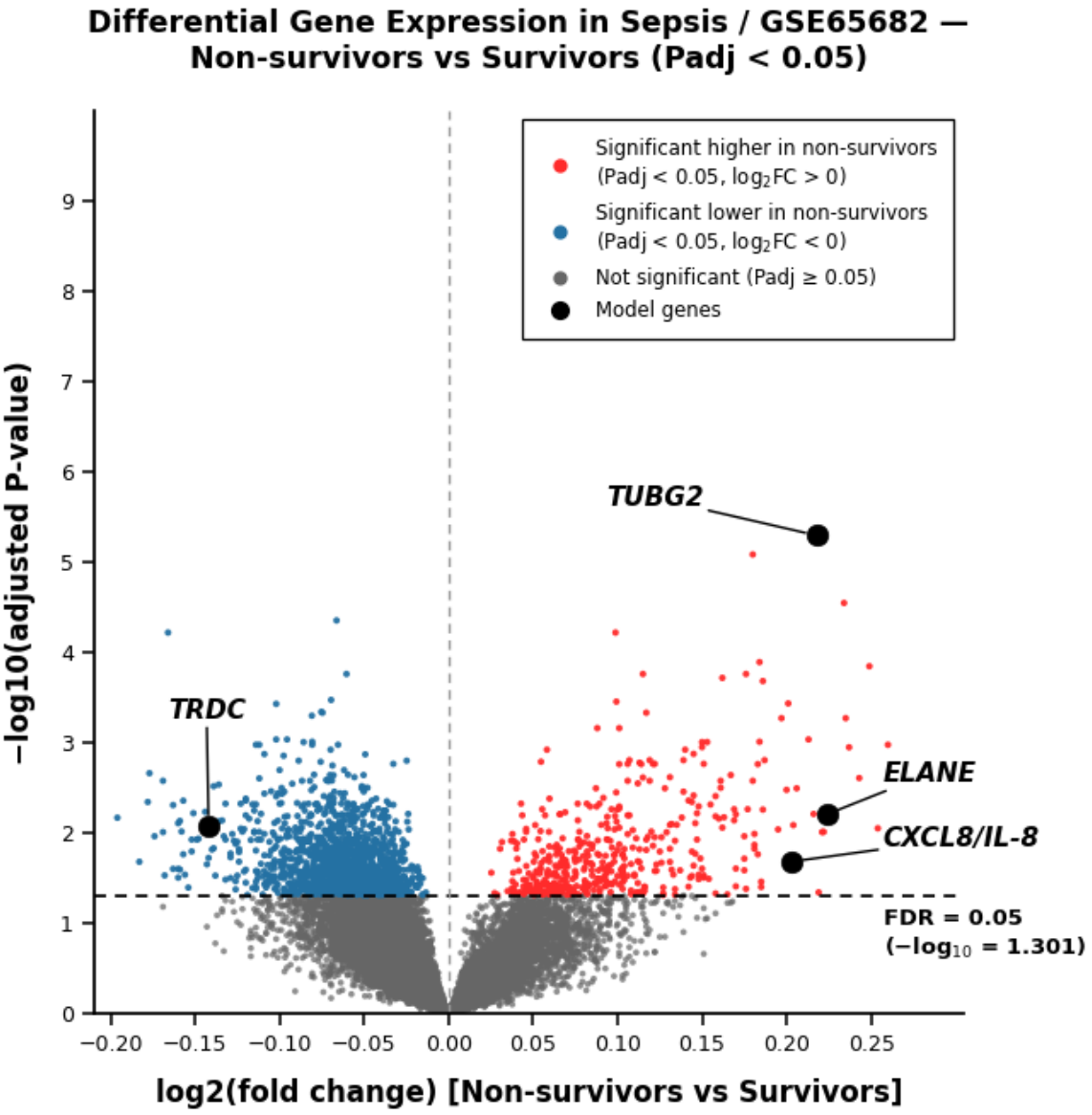
8. Sweeney TE, Perumal TM, Henao R, Nichols M, Howrylak JA, Choi AM, et al. A community approach to mortality prediction in sepsis via gene expression analysis. Nat Commun. 2018;9(1):694.
9. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321(20):2003-17.
10. Burnham KL, Davenport EE, Radhakrishnan J, Humburg P, Gordon AC, Hutton P, et al. Shared and distinct aspects of the sepsis transcriptomic response to fulminating bacterial and viral pathogens. Am J Respir Crit Care Med. 2017;196(10):1260-72.
11. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30:4765-74.
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.
13. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207-10.
14. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets: update. Nucleic Acids Res. 2013;41(Database issue):D991-5.
15. GSE65682. Gene Expression Omnibus [Internet]. National Center for Biotechnology Information; 2015 [cited 2024]. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65682
16. GSE95233. Gene Expression Omnibus [Internet]. National Center for Biotechnology Information; 2017 [cited 2024]. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95233
17. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57(1):289-300.
20. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785-94.
21. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74.
22. Murphy SM, Urbani L, Stearns T. The mammalian gamma-tubulin complex contains homologues of the yeast spindle pole body components spc97p and spc98p. J Cell Biol. 1998;141(3):663-74.
23. Bozza FA, Salluh JI, Japiassu AM, Soares M, Assis EF, Gomes RN, et al. Cytokine profiles as markers of disease severity in sepsis: a multiplex analysis. Crit Care. 2007;11(2):R49.
24. Korkmaz B, Moreau T, Gauthier F. Neutrophil elastase, proteinase 3 and cathepsin G: physicochemical properties, activity and physiopathological functions. Biochimie. 2008;90(2):227-42.
25. Venet F, Monneret G. Advances in the understanding and treatment of sepsis-induced immunosuppression. Nat Rev Nephrol. 2018;14(2):121-37.
26. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.

27. Vantourout P, Hayday A. Six-of-the-best: unique contributions of gammadelta T cells to immunology. Nat Rev Immunol. 2013;13(2):88-100.

Figures

Figure 1

A)



B)

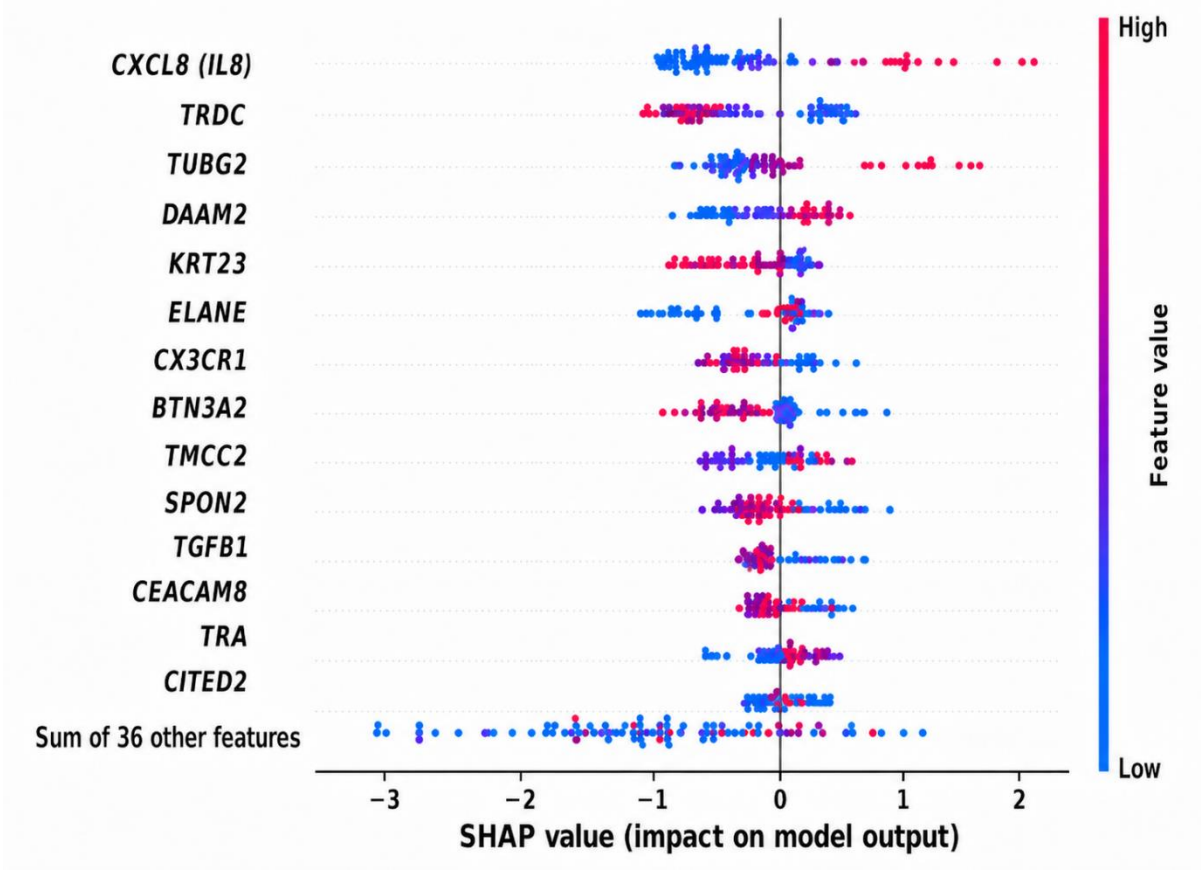


Figure 1. Feature identification and selection. Panel A: volcano plot of differential gene expression in the GSE65682 training cohort (non-survivors versus survivors, n = 479). Each point represents one transcript. The x-axis denotes the $\log_2$ fold-change in non-survivors relative to survivors; positive values indicate higher expression in non-survivors. The y-axis denotes the negative $\log_{10}$ of the Benjamini-Hochberg FDR-adjusted p-value. The horizontal dashed line indicates FDR = 0.05. Red: higher in non-survivors (Padj < 0.05). Blue: lower in non-survivors (Padj < 0.05). Grey: not significant. Black circles indicate the four final model genes. Panel B: SHAP beeswarm screening plot applied to the 50-gene candidate pool. Each point represents one patient. Genes are ranked by mean absolute SHAP value. Features were selected based on strong directional consistency and biological plausibility. The four final model genes (*TUBG2*, *TRDC*, *CXCL8*, *ELANE*) are visible among the top-ranked features.

Figure 2

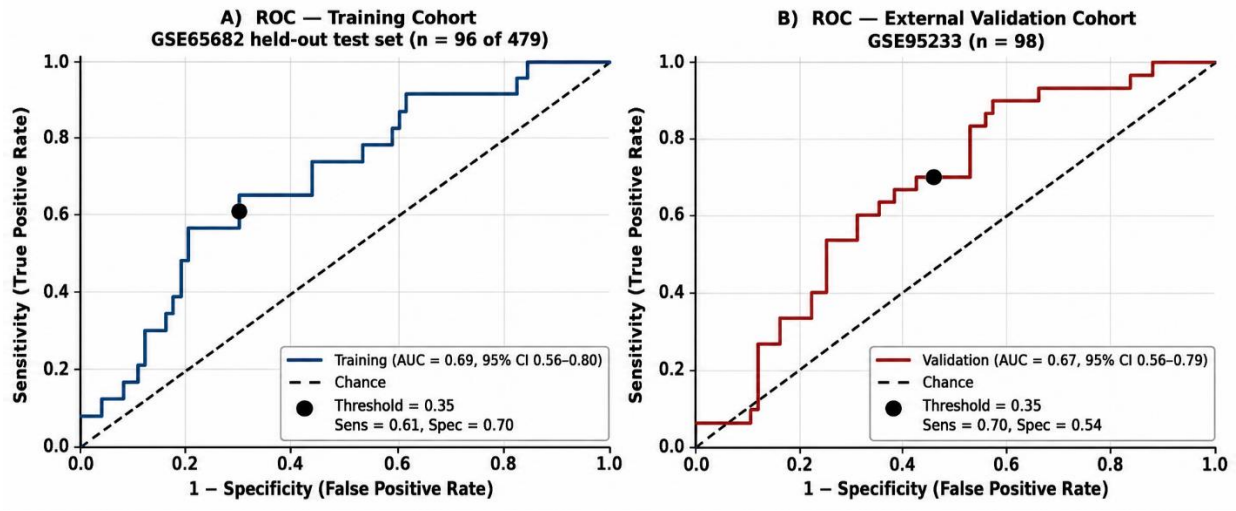


Figure 2. Receiver operating characteristic (ROC) curves. Panel A: training cohort held-out test set (GSE65682, n = 96 of 479). Panel B: external validation cohort (GSE95233, n = 98). The diagonal dashed line represents chance discrimination. The black circle indicates the operating point at the pre-specified probability threshold of 0.35. AUC values are reported with bootstrapped 95% confidence intervals (1,000 samples). The step-like curve appearance reflects the small number of samples in each evaluation set. The overlapping confidence intervals across panels indicate no statistically significant difference in discrimination between cohorts.

Figure 3

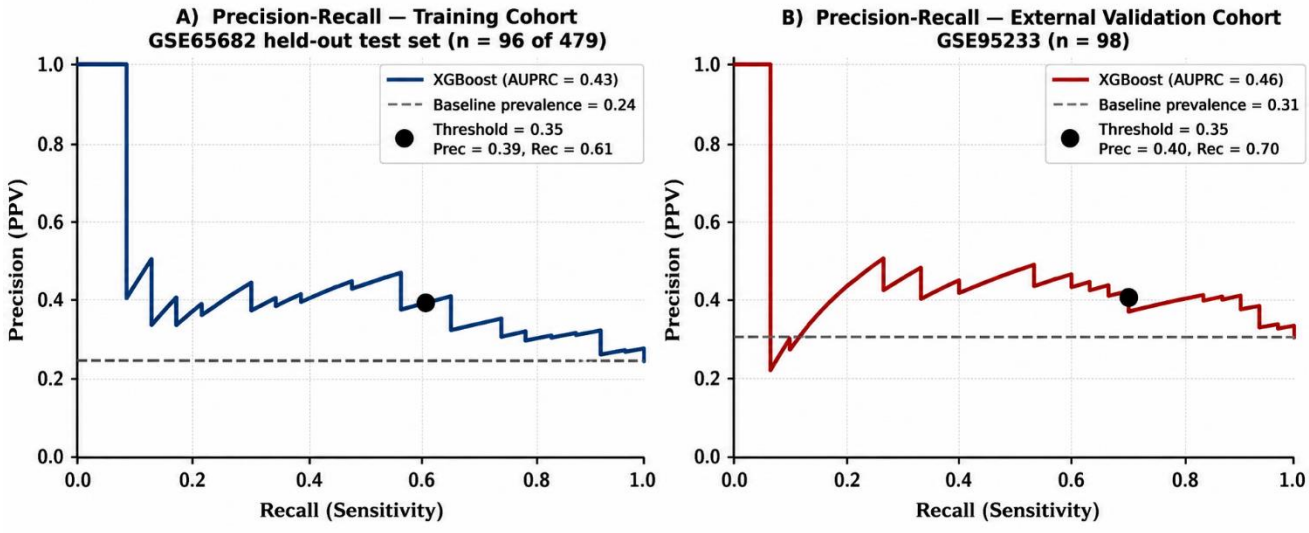


Figure 3. Precision-recall curves. Panel A: training cohort held-out test set (GSE65682, n = 96 of 479). Panel B: external validation cohort (GSE95233, n = 98). The dashed horizontal line indicates

the no-skill classifier precision equal to the cohort mortality prevalence (Panel A: 0.24; Panel B: 0.31). The black circle indicates the operating point at threshold 0.35. The AUPRC in validation (0.46) exceeds both the training AUPRC (0.43) and the validation baseline prevalence (0.31), indicating improved precision in non-survivor identification in the external cohort. The precision spike at low recall in Panel B reflects high-confidence predictions at stringent probability thresholds, a common feature of precision-recall curves in small imbalanced datasets.

Figure 4

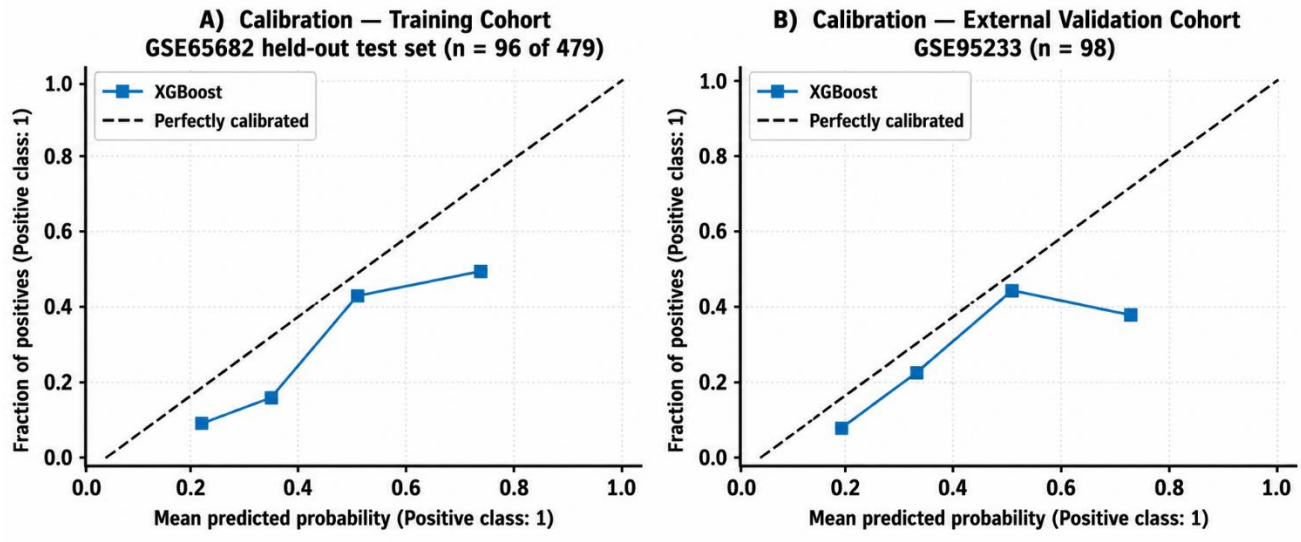


Figure 4. Calibration curves. Panel A: training cohort held-out test set (GSE65682, n = 96 of 479). Panel B: external validation cohort (GSE95233, n = 98). The dashed diagonal represents perfect calibration. Both panels show the relationship between mean predicted probability and observed mortality fraction across five probability bins.

Figure 5

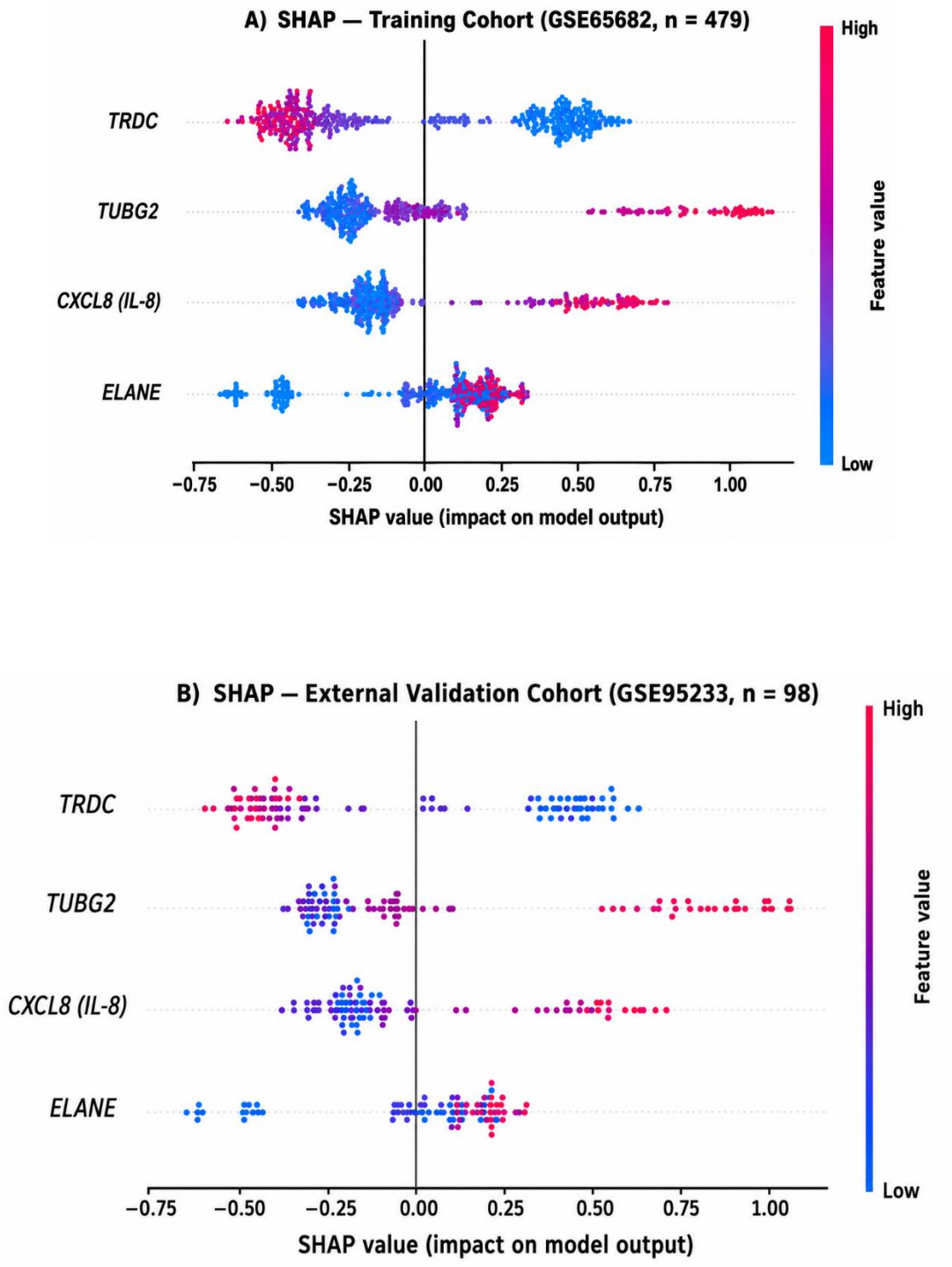


Figure 5. SHAP beeswarm plots for the final four-gene model. Panel A: full training cohort (GSE65682, n = 479). Panel B: external validation cohort (GSE95233, n = 98). Each point represents one patient. The x-axis denotes the SHAP value (contribution to predicted log-odds of mortality);

colour encodes feature expression level (red: high; blue: low). Features are ranked by mean absolute SHAP value. The bimodal distribution of *ELANE*, visible in both panels, reflects a discrete subgroup of patients with very low *ELANE* expression that receives a strong survival-direction prediction. The greater density of points in Panel A reflects the larger training cohort (n = 479 versus n = 98).

Figure 6

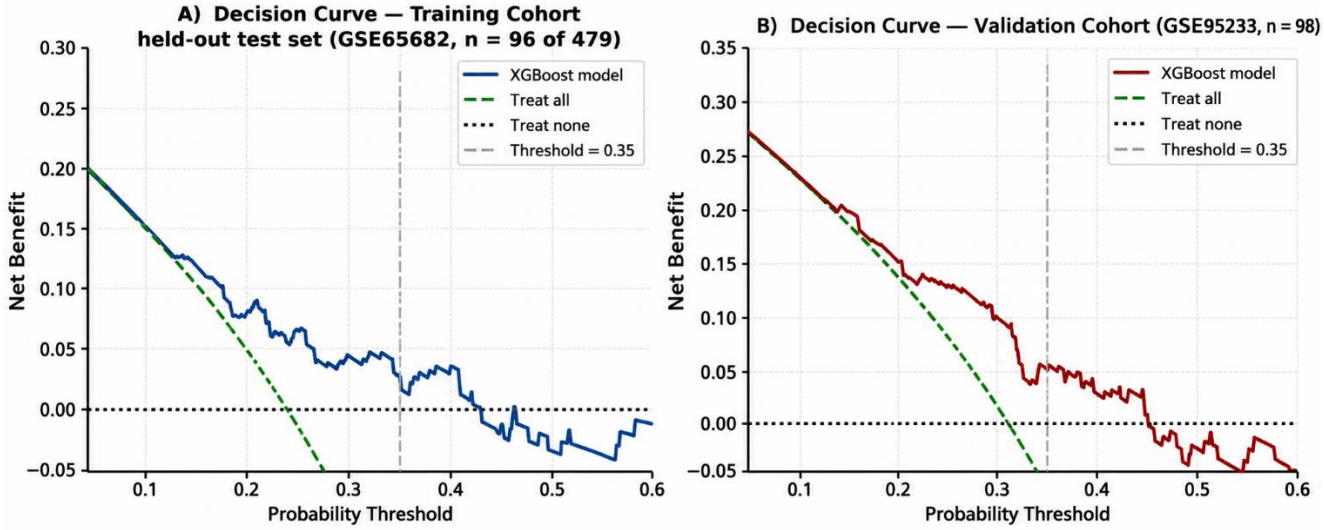


Figure 6. Decision curve analysis. Panel A: training cohort held-out test set (GSE65682, n = 96 of 479). Panel B: external validation cohort (GSE95233, n = 98). Net benefit is shown for the XGBoost model, a treat-all strategy, and a treat-none strategy across probability thresholds from 0.05 to 0.60. The vertical dashed line indicates the pre-specified operating threshold of 0.35. The model provides positive net benefit above the treat-none strategy across thresholds from approximately 0.10 to 0.42 in both cohorts.

Figure 7

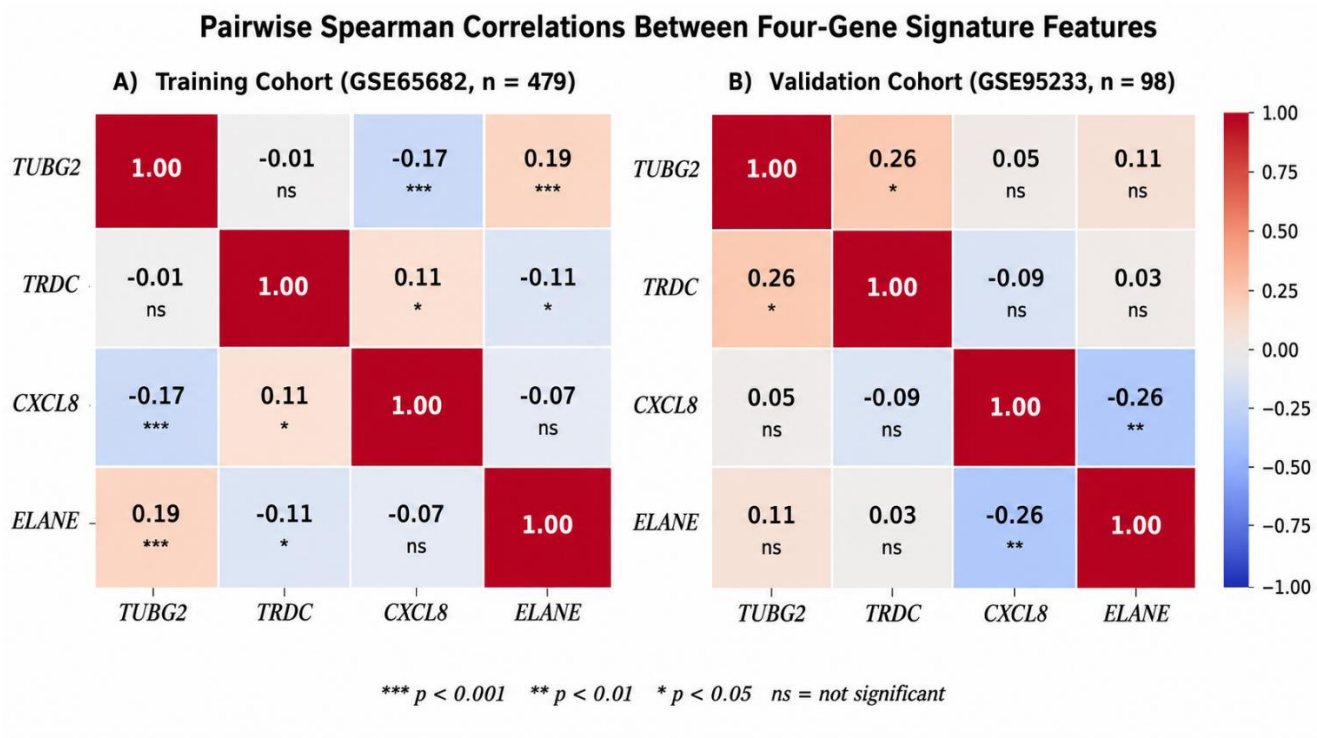


Figure 7. Pairwise Spearman correlations between the four final model genes. Panel A: training cohort (GSE65682, n = 479). Panel B: external validation cohort (GSE95233, n = 98). Correlation coefficients are shown in each cell with significance annotations below (*** p < 0.001; ** p < 0.01; * p < 0.05; ns = not significant). The colour scale denotes correlation direction and magnitude (red: positive; blue: negative). Most correlations are weak and non-significant, supporting the interpretation that the four genes contribute largely independent information. The negative *CXCL8* / *ELANE* correlation in validation (−0.26**) is consistent with the bimodal *ELANE* SHAP distribution observed in Figure 5B.

Figure 8

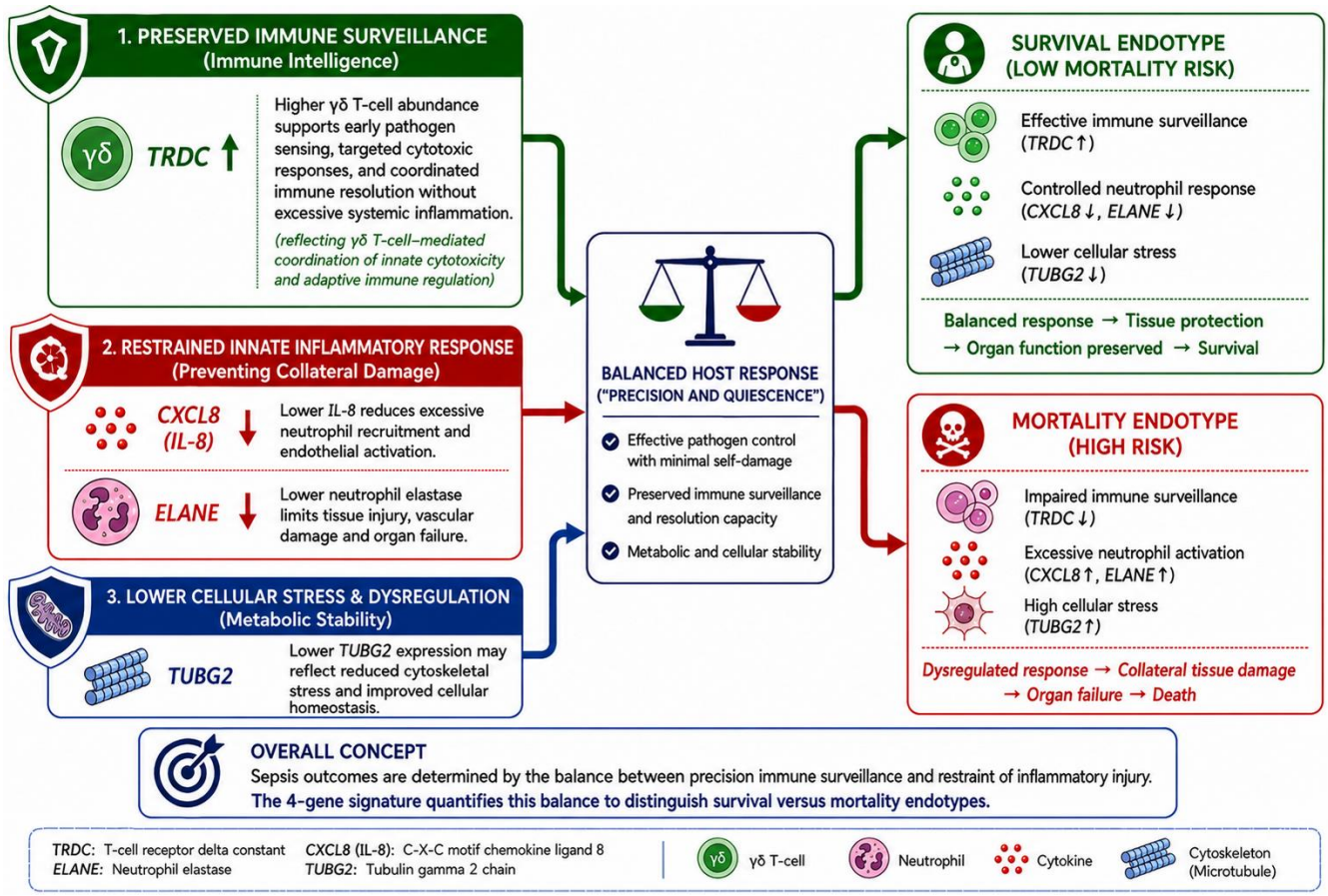


Figure 8. Proposed mechanistic framework of host-response imbalance in sepsis mortality. Three biological axes are identified by the four-gene signature: (i) immune surveillance, maintained by gamma-delta T-cell activity (*TRDC*), associated with survival; (ii) neutrophil-mediated inflammatory activation through chemoattractant signalling (*CXCL8*) and protease release (*ELANE*), associated with mortality; and (iii) cellular cytoskeletal stress (*TUBG2*), the dominant predictor, the mechanism of which warrants experimental characterisation. The balance scale represents the balanced host response associated with survival. Disruption of immune surveillance combined with excessive neutrophil activation and elevated cellular stress drives the dysregulated endotype associated with death.