Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A prognostic signature based on neutrophil extracellular traps– coagulation crosstalk-related genes in colorectal cancer

Zelin Chen
Guangzhou University of Chinese Medicine
Jinwen Fang
Shantou Hospital of Traditional Chinese Medicine
Xiangyang Yang
Shantou Hospital of Traditional Chinese Medicine
E. Xie
1085918761@qq.com

Shantou Hospital of Traditional Chinese Medicine

Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11458256/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A prognostic signature based on neutrophil extracellular traps–coagulation crosstalk-related genes in colorectal cancer

Zelin Chen1, Jinwen Fang2, Xiangyang Yang2, E. Xie2,*

1 Graduate School, Guangzhou University of Traditional Chinese Medicine, Guangzhou, Guangdong 510006, China

2 Department of Hepatobiliary, Pancreatic, Gastrointestinal, Hernia and Abdominal Wall Surgery, Shantou Hospital of Traditional Chinese Medicine, Shantou, Guangdong 515000, China

*Correspondence: E. Xie, Department of Hepatobiliary, Pancreatic, Gastrointestinal, Hernia and Abdominal Wall Surgery, Shantou Hospital of Traditional Chinese Medicine, No. 3 Shaoshan Road, Longhu District, Shantou, Guangdong 515000, China. E-mail: 1085918761@qq.com

Running title: CHEN et al: NETs-coagulation crosstalk genes in CRC

Tables

Table 1. Baseline characteristics of the TCGA discovery cohort (n = 372).

Characteristic	Value
Age, years, mean ± SD	64.5 ± 13.1
Age, median (IQR)	66.0 (55.0–75.0)
Age range	31–90
Sex, n (%)	Male 205 (55.1); Female 167 (44.9)
Cancer type, n (%)	COAD 279 (75.0); READ 93 (25.0)
Follow-up, months, median (range)	22.1 (0.2–147.9)
Overall survival events, n (%)	85 (22.8)
Tumour stage, n (% of 108 available)	I: 55 (50.9); II: 18 (16.7); III: 8 (7.4); IV: 27 (25.0)
Missing stage data, n (%)	264 (71.0)
Risk group, n (%)	High 186 (50.0); Low 186 (50.0)

Table 2. Immune-cell infiltration differences between risk groups in the extended TCGA PanCanAtlas cohort (n = 566).

Cell type	Mean difference (high−low)	P value	FDR
Macrophage M1	-1.906	1.234e-06	1.974e-05

NK cells	+1.109	0.0006613	0.005291
Tfh	-0.472	0.04934	0.2632
Th17	+0.037	0.139	0.3959
Neutrophils	+0.533	0.1528	0.3959
B cells	-0.619	0.1899	0.3959
Th2	-0.741	0.1985	0.3959
MDSC	-0.443	0.2194	0.3959
Th1	-0.218	0.2262	0.3959
Mast cells	+0.398	0.2474	0.3959
Macrophage M2	-0.080	0.452	0.6575
CD8 T cells	-0.050	0.6661	0.8362
Dendritic cells	-0.191	0.6805	0.8362
Treg	-0.087	0.7317	0.8362
Monocytes	+0.129	0.8467	0.9032
Plasma cells	+0.013	0.9059	0.9059

ssGSEA scores; Mann–Whitney U test with Benjamini–Hochberg adjustment across 16 cell types.

Table 3. Immune-checkpoint gene expression differences in the extended cohort (n = 566).

Gene	**Mean difference (high−low)**	**P value**	**FDR**
IDO1	-99.339	0.006281	0.04451
CD274	-6.851	0.01113	0.04451
LAG3	-22.912	0.03025	0.08068
PDCD1LG2	-3.734	0.1695	0.3173
HAVCR2	-20.373	0.1983	0.3173
PDCD1	-5.471	0.2609	0.3479
CTLA4	+0.371	0.6344	0.7251
TIGIT	+4.839	0.7966	0.7966

Mann–Whitney U test with Benjamini–Hochberg adjustment across eight checkpoint genes.

Table 4. Drug-target genes differentially expressed between risk groups in the extended cohort (n = 566; FDR < 0.05).

Gene	Higher in	P value	FDR
TYMS	Low-risk	2.595e-08	4.153e-07
DHFR	Low-risk	1.671e-07	1.337e-06
EGFR	High-risk	1.895e-05	0.000101
ABCB1	High-risk	6.822e-05	0.0002729
NRAS	Low-risk	0.0001478	0.0004729
PIK3CA	High-risk	0.002706	0.007217
VEGFA	High-risk	0.004487	0.01026
ERBB2	High-risk	0.005993	0.01199
KRAS	Low-risk	0.007655	0.01361

Mann–Whitney U test with Benjamini–Hochberg adjustment across 16 drug-target genes.

Table 5. Univariate and multivariate Cox regression for overall survival (TCGA discovery cohort, n = 372).

Variable	Model	HR	P value
Risk score (per SD)	Univariate	1.570 (1.087–2.266)	0.016
Risk score (per SD)	Multivariate	1.441 (0.971–2.139)	0.070
Age (per year)	Multivariate	1.051	0.0060
Tumour stage (per level)	Multivariate	1.992	1.6×10^{-4}

The multivariate model was adjusted for age and tumour stage (stage available for 108 patients). HR, hazard ratio; SD, standard deviation.

Figures: 12 (+ 1 supplementary) | Tables: 5 | References: 41

Abstract

Cancer-associated thrombosis is a frequent, life-threatening complication of colorectal cancer (CRC). Neutrophil extracellular traps (NETs) and the coagulation cascade engage in extensive crosstalk, yet no prognostic model has been constructed on NETs-coagulation crosstalk-related genes. We analysed transcriptome data of 372 CRC tumours from The Cancer Genome Atlas (TCGA). NETs-coagulation crosstalk genes were defined as NETs-related genes whose expression correlated with a single-sample gene set enrichment analysis (ssGSEA) coagulation activity score. Univariate Cox regression followed by cross-validated feature selection (LASSO-Cox regression, support vector machine recursive feature elimination and random forest) converged on a three-gene signature: risk score = 0.200×ADAMTS13 + 0.407×CRISPLD2 − 0.277×SIGLEC14. High-risk patients had significantly shorter overall survival (log-rank P = 0.00067). In the extended TCGA PanCanAtlas cohort (n = 566), the score stratified survival using a data-derived optimal cut-off (exploratory; P = 0.0004) and added prognostic information to age and stage (3-year continuous net reclassification improvement 0.29; C-index 0.729 to 0.735 in the stage-complete subset). In the discovery cohort, high-risk tumours showed higher neutrophil and myeloid-derived suppressor cell infiltration; a distinct therapeutic-target profile (TYMS, DHFR and TOP1 higher in low-risk tumours; EGFR and ABCB1 higher in high-risk tumours) remained evident after multiple-testing correction in the extended cohort. Across four additional cohorts (n = 702) the association was direction-consistent (I^2 = 0%) and significant in GSE17537 (P = 0.009); microarray validation was limited by absent SIGLEC14 probes on the GPL570 platform. The three-gene model stratified CRC prognosis with modest discrimination, and should be regarded as complementary to TNM staging.

Keywords: colorectal cancer; neutrophil extracellular traps; coagulation; crosstalk; prognostic model; immune infiltration

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed malignancy and the second leading cause of cancer death worldwide [1,2]. Despite advances in surgery, chemotherapy, targeted therapy and immunotherapy, clinical outcomes remain heterogeneous, and robust, easily applicable molecular tools for prognostic stratification are still needed [2].

Patients with cancer have a markedly elevated risk of venous thromboembolism (VTE), and cancer-associated thrombosis (CAT) is a leading cause of death among individuals with solid tumours [3,4]. In CRC, activation of the coagulation cascade not only drives CAT but also fosters tumour growth, angiogenesis and metastatic dissemination, linking haemostatic activation to malignant progression [5,6].

Neutrophil extracellular traps (NETs) are web-like extracellular chromatin structures released by activated neutrophils [7]. In addition to their antimicrobial function, NETs provide a procoagulant scaffold that promotes platelet and coagulation factor activation and contributes to thrombus formation [8,9]. Within the tumour microenvironment, NETs promote circulating tumour-cell adhesion, distant seeding and recurrence [6,10,11], and several NETs-focused gene signatures have been associated with CRC prognosis and immune phenotypes [12–16]. However, prognostic models built explicitly on the crosstalk between NETs and coagulation—two intertwined, tumour-promoting systems—remain largely unexplored in CRC [3,6,17].

Regularised regression, support vector machines and tree-based ensembles are widely used, complementary strategies for prognostic gene selection from high-dimensional transcriptomic data [18–20]; applying them in combination can cross-confirm candidate genes and reduce the instability associated with any single strategy [19].

In this study, we integrated curated NETs-related genes with the Hallmark coagulation gene set to identify NETs–coagulation crosstalk genes in CRC, applied complementary feature-selection strategies to construct a prognostic risk model, evaluated its performance by survival analyses, time-dependent ROC curves, C-index and a nomogram, and characterised the immune microenvironment and drug-target landscape of the defined risk groups. Within the discovery cohort, the model was additionally examined in the rectal-cancer subset (TCGA-READ) to assess generalisability across anatomical subsites. Independent external validation was performed in the GSE39582 and GSE87211 cohorts, and extended to four further independent cohorts (GSE17536, GSE17537, GSE38832 and the SiLU RNA-sequencing cohort).

Results

Identification of NETs–coagulation crosstalk genes in CRC

The TCGA discovery cohort comprised 372 CRC tumour samples with matched survival data (Table 1). Comparing tumour tissues with 51 adjacent normal samples identified 4,798 DEGs (FDR < 0.05, |log2FC| > 1; Fig. 1a, 1b). Of the 97 curated NETs-related genes, 94 were detectable in the expression matrix; of the 138 coagulation genes, 136 were detectable. An ssGSEA coagulation activity score was computed for each tumour sample, and 58 NETs genes

showed significant correlation with the coagulation score ($|r| > 0.3$, $P < 0.05$), which were defined as NETs–coagulation crosstalk genes (Fig. 2).

Selection of a three-gene prognostic signature

Univariate Cox regression identified three crosstalk genes significantly associated with overall survival at $P < 0.1$: SIGLEC14, ADAMTS13 and CRISPLD2. LASSO-Cox regression ($\lambda = 0.00487$ by ten-fold cross-validation) and SVM-RFE each selected all three genes, while RF selected SIGLEC14 and CRISPLD2; each gene therefore received at least two votes and entered the final signature (Fig. 3a–3d). Multivariate Cox regression yielded the following risk model:

$$\textbf{Risk score} = 0.200 \times \textbf{ADAMTS13} + 0.407 \times \textbf{CRISPLD2} - 0.277 \times \textbf{SIGLEC14}$$

ADAMTS13 (HR 1.221, $P = 0.037$) and CRISPLD2 (HR 1.503, $P = 3.4 \times 10^{-5}$) acted as risk factors, whereas SIGLEC14 (HR 0.758, $P = 6.9 \times 10^{-5}$) was protective.

The risk model stratifies overall survival in the TCGA cohort

Patients were divided into high- and low-risk groups (n = 186 each) by the median risk score. Risk-score distribution and patient survival status are shown in Fig. 4a. The high-risk group had significantly shorter overall survival (log-rank $P = 0.00067$; Fig. 4b).

Discrimination and calibration of the model

Time-dependent AUCs were 0.627, 0.660 and 0.623 at 1, 3 and 5 years, respectively (Fig. 5). In the extended TCGA PanCanAtlas cohort (n = 566 primary tumours), the C-index of the risk score was 0.595 (bootstrap 95% CI 0.538–0.653 over 1000 resamples; optimism-corrected 0.594). The risk score again stratified overall survival by both median dichotomisation (log-rank $P = 0.022$) and the data-derived optimal cut-off identified by log-rank maximisation ($P = 0.0004$; exploratory, because the cut-off was derived from the same data; Fig. 6a). To evaluate incremental value, clinical (age and tumour stage) and combined models were compared in the 553 patients with complete stage information: the C-index increased from 0.729 (clinical model) to 0.735 (combined model), the 3-year continuous (category-free) net reclassification improvement was 0.29, and decision-curve analysis showed greater net benefit for the combined model than for clinical factors alone across a broad range of threshold probabilities at 1, 3 and 5 years (Fig. 6b).

Independent prognostic value

In univariate Cox regression, the risk score was significantly associated with survival (HR 1.570, 95% CI 1.087–2.266, $P = 0.016$), as were age (HR 1.027, $P = 0.081$) and tumour stage (HR 1.974, $P = 5.3 \times 10^{-5}$) (Fig. 7; Table 5). In multivariate analysis adjusting for age and stage, the risk score remained marginally significant (HR 1.441, 95% CI 0.971–2.139, $P = 0.070$), while age (HR 1.051, $P = 0.0060$) and stage (HR 1.992, $P = 1.6 \times 10^{-4}$) were independent predictors. A nomogram combining the risk score, age and stage was constructed (Fig. 8a), and the calibration curve showed reasonable agreement between predicted and observed 3-year survival (Fig. 8b).

The nomogram and its calibration were based on the 108 patients with complete tumour-stage information, so calibration assessment was limited in precision.

Immune-microenvironment differences between risk groups

ssGSEA immune-infiltration analysis showed that the high-risk group had significantly higher scores for neutrophils ($P = 2.6 \times 10^{-5}$), myeloid-derived suppressor cells (MDSCs; $P = 3.9 \times 10^{-4}$), M1 macrophages (P = 0.0044), plasma cells (P = 0.0083), regulatory T cells (P = 0.0088), $CD8^+$ T cells (P = 0.019) and dendritic cells (P = 0.025), whereas Th1 scores were lower in the high-risk group (P = 0.0087) (Fig. 9a, 9b). When the analysis was repeated in the extended PanCanAtlas cohort (n = 566) with Benjamini–Hochberg adjustment across the 16 cell types, M1-macrophage scores were significantly lower and NK-cell scores significantly higher in high-risk tumours (FDR < 0.05), whereas the neutrophil and MDSC differences observed in the discovery cohort were attenuated and no longer significant (Table 2; Supplementary Fig. S1), notably, the direction of the M1-macrophage association was opposite to that in the discovery cohort; these patterns indicate partial cohort-dependence of the immune associations.

Immune-checkpoint and drug-target gene expression

Among immune-checkpoint genes, CD274 (PD-L1; P = 0.0024), LAG3 (P = 0.019) and IDO1 (P = 0.0049) were differentially expressed between risk groups, while PDCD1, CTLA4, HAVCR2, TIGIT and PDCD1LG2 showed no significant difference (Fig. 9c). Several clinically actionable drug-target genes were also differentially expressed: ABCB1 (P = 0.0016), DHFR (P = 0.00036), EGFR (P = 0.00036), NRAS (P = 0.0048), PIK3CA (P = 0.0055), TOP1 (P = 0.042) and TYMS (P = 0.0025); TYMS, DHFR and TOP1 were higher in the low-risk group, whereas EGFR, NRAS, PIK3CA and ABCB1 were higher in the high-risk group (Fig. 10; Table 4). NRAS therefore showed an opposite direction between the discovery and extended cohorts, in which, after false-discovery-rate correction, both NRAS and KRAS were significantly higher in low-risk tumours (Table 4). In the extended cohort with Benjamini–Hochberg correction, CD274 and IDO1 were significantly lower in high-risk tumours (FDR = 0.045; Table 3), again indicating cohort-dependence of the checkpoint associations. For the drug-target genes, the extended-cohort analysis with correction confirmed TYMS, DHFR, NRAS and KRAS as higher in low-risk tumours and EGFR, PIK3CA, VEGFA, ERBB2 and ABCB1 as higher in high-risk tumours (FDR < 0.05; Table 4).

Exploratory external validation

Within the discovery cohort, the rectal-cancer subset (TCGA-READ, n = 93) was analysed separately to assess generalisability across anatomical subsites; because these samples contributed to model development, this analysis is descriptive rather than confirmatory. In this subset, the three-gene model did not significantly stratify survival (log-rank P = 0.619; 1-, 3- and 5-year AUCs 0.605, 0.552 and 0.487, respectively; Fig. 11a). In the independent GSE39582 cohort (n = 573), neither ADAMTS13 (P = 0.838) nor CRISPLD2 (P = 0.713) alone was significantly prognostic, and the exploratory two-gene sub-score (excluding SIGLEC14, which lacks a probe on the GPL570 platform) showed a non-significant trend (P = 0.181; Fig. 11b–11d). When the risk score was analysed as a continuous variable, the direction of association remained consistent in both cohorts, albeit without reaching significance: in TCGA-READ the risk score showed a

hazard ratio of 1.558 per standard-deviation increase (95% CI 0.926–2.624, P = 0.095), and in GSE39582 the two-gene sub-score showed a hazard ratio of 1.144 per standard-deviation increase (95% CI 0.990–1.322, P = 0.068).

In the independent GSE87211 rectal-cancer cohort (n = 196 pre-treatment biopsies), median dichotomisation of the three-gene risk score did not significantly stratify disease-free survival (log-rank P = 0.398; Fig. 11e) or overall survival (P = 0.399). However, as a continuous variable the risk score was associated with shorter disease-free survival (HR 1.647 per SD, 95% CI 0.906–2.997, P = 0.102) and retained a consistent direction of effect after adjustment for age and sex (multivariate HR 1.665, P = 0.098). Among the individual model genes, CRISPLD2 was significantly associated with disease-free survival (HR 1.530 per SD, P = 0.004), consistent with its role as the strongest risk gene in the TCGA model. These external results were therefore interpreted as exploratory.

We additionally evaluated the model in four further independent cohorts (Fig. 12). Because SIGLEC14 is not covered by the Affymetrix HG-U133_Plus_2 platform, the microarray cohorts (GSE17536, n = 177; GSE17537, n = 55; GSE38832, n = 122) were analysed using the two-gene ADAMTS13–CRISPLD2 sub-score, whereas the full three-gene score was applied in the RNA-sequencing SiLU cohort (n = 348). The sub-score significantly stratified overall survival in GSE17537 (HR 1.91 per SD, log-rank P = 0.009; Fig. 12a) and showed a consistent direction of effect in GSE17536 (HR 1.33, P = 0.56; Fig. 12b), GSE38832 (HR 1.12, P = 0.90; Fig. 12c) and the SiLU cohort (sub-score HR 1.10 per SD, Cox P = 0.54; the full three-gene score gave a similar result; Fig. 12d). Fixed-effect meta-analysis of the ADAMTS13–CRISPLD2 sub-score across the four cohorts (n = 702; the sub-score was used for all four cohorts to ensure comparability) yielded a pooled HR of 1.22 (95% CI 0.94–1.57, P = 0.14) with no heterogeneity (I^2 = 0%; Fig. 12e), and every cohort showed a hazard ratio above unity.

Discussion

In this study we constructed, to our knowledge, the first prognostic model in CRC based specifically on the crosstalk between NETs and the coagulation cascade. By linking a curated NETs gene set to a data-derived coagulation activity score and applying complementary feature-selection strategies, we derived a compact three-gene signature—ADAMTS13, CRISPLD2 and SIGLEC14—that stratified overall survival in the TCGA cohort with modest discrimination (5-year AUC 0.623; C-index 0.595 with negligible bootstrap optimism in the extended cohort) and captured a biologically coherent, neutrophil-centred immune phenotype.

Each model gene is plausibly embedded in the NETs–coagulation axis. ADAMTS13 is the von Willebrand factor (VWF)-cleaving metalloproteinase that limits platelet-rich microthrombosis; a reduced ADAMTS13 activity relative to VWF has been linked to cancer-associated thrombosis and worse survival across malignancies [21,22]. Its positive coefficient in our model suggests that elevated tumour-tissue ADAMTS13 expression marks a thrombo-inflammatory microenvironment, although transcript abundance need not parallel protease activity and this direction requires functional confirmation. Notably, CRISPLD2 has been identified as a NETs-related prognostic gene in two independent recent CRC studies [12,15], supporting the robustness of its selection in our model. SIGLEC14 forms a paired activating/inhibitory receptor system with Siglec-5 on neutrophils and modulates neutrophil activation and NET formation [23–25]; its protective coefficient may reflect altered myeloid composition rather than a simple causal effect, and its precise role in CRC warrants dedicated study.

The immune landscape of high-risk tumours provides mechanistic support for the model. High-risk samples were enriched in neutrophils and MDSCs—cells central to NET formation and tumour-associated immunosuppression [26–28] and recently shown to drive NETosis in the pre-metastatic niche of colorectal liver metastasis [17]—alongside increased M1 macrophages, regulatory T cells and $CD8^+$ T cells, indicating an inflamed yet suppressed microenvironment. Differential expression of CD274, LAG3 and IDO1 suggests that high-risk patients may constitute a subgroup with distinct immunotherapy-relevant biology [12,13,26]; whether this translates into differential response to immune-checkpoint blockade requires direct testing. Notably, the immune and checkpoint associations were only partially reproducible in the extended PanCanAtlas cohort with multiple-testing correction (lower M1 macrophages and higher NK cells, together with lower CD274 and IDO1, in high-risk tumours), underscoring that the immunological interpretation of the risk groups is cohort-dependent and hypothesis-generating.

The drug-target analysis adds a translational dimension. TYMS and DHFR, targets of fluoropyrimidine-based chemotherapy, and TOP1, the target of topoisomerase-I inhibitors, were relatively higher in low-risk tumours, whereas EGFR, PIK3CA and the efflux transporter ABCB1 were elevated in high-risk tumours in both cohorts. NRAS showed a cohort-dependent direction—higher in high-risk tumours in the discovery cohort but significantly higher in low-risk tumours after false-discovery-rate correction in the extended cohort—and KRAS was significantly associated with risk group only in the extended cohort (Table 4). These patterns are consistent with molecular subtype-stratified drug-response analyses in recent CRC NETs models [13,14,16] and are hypothesis-generating; if validated, they may inform stratified treatment strategies in CRC.

Several limitations must be emphasised. First, although internal performance was acceptable, the model was not significantly validated by dichotomised analyses in GSE39582 or in the TCGA-READ rectal subset of the discovery cohort; in the independent GSE87211 rectal-cancer cohort the three-gene risk score showed only a consistent direction of effect in continuous and multivariate analyses (multivariate DFS HR 1.665, P = 0.098), although CRISPLD2 was individually prognostic (P = 0.004), and the GSE87211 population (exclusively locally advanced rectal cancers destined for neoadjuvant radiochemotherapy) differed from the mixed-stage discovery cohort. A key technical reason is that SIGLEC14 lacks a probe on the Affymetrix HG-U133_Plus_2 platform, forcing sub-score analyses that discard one-third of the model; platform batch effects and biological heterogeneity between colon and rectal cohorts likely also contributed. Notably, when analysed as continuous variables, both the READ risk score (HR 1.558 per SD, P = 0.095) and the GSE39582 sub-score (HR 1.144 per SD, P = 0.068) retained the expected direction of effect, suggesting that the negative dichotomised results partly reflect information loss from median-based grouping rather than complete model failure; taken together, the TCGA-READ rectal subset, GSE39582 and GSE87211 all showed hazard ratios above unity for the risk score in continuous analyses, supporting the robustness of the model direction if not its dichotomised significance. The four additional cohorts examined here extended this pattern, with hazard ratios above unity in every cohort, no between-cohort heterogeneity ($I^2 = 0\%$) and a pooled hazard ratio of 1.22 (95% CI 0.94–1.57), although only GSE17537 reached significance, consistent with limited per-cohort power. Second, tumour-stage data were available for only 108 of 372 patients, limiting the multivariate adjustment and the calibration assessment. Third, the small three-gene signature yielded only marginal independence (P = 0.070), a finding corroborated in the larger extended TCGA cohort (multivariate P = 0.071); the score is therefore best regarded as complementary to TNM staging rather than an independent prognostic determinant. Immune-infiltration, checkpoint and drug-target comparisons were exploratory and partially cohort-dependent; they are intended to generate hypotheses rather than to support definitive biological conclusions. Fourth, all analyses were retrospective and in silico, and the biological functions of the model genes were not experimentally validated. In addition, the median follow-up in the discovery cohort was 22.1 months, which limits the precision of the 5-year survival estimates. Prospective, multi-cohort validation and functional experiments are essential before any clinical application.

In conclusion, we identified a NETs–coagulation crosstalk gene signature (ADAMTS13, CRISPLD2, SIGLEC14) that stratifies prognosis and characterises the immune microenvironment of CRC. Despite partial external validation, with a consistent direction of effect across seven independent cohorts and significant survival stratification in GSE17537, the model offers a mechanistically grounded, low-cost candidate tool and a rationale for further study of NETs–coagulation interplay in CRC.

Methods

Data acquisition and preprocessing

Level-3 normalised RNA-sequencing expression data (Illumina HiSeqV2) and corresponding clinical information for colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ)

were downloaded from the UCSC Xena data portal (https://xenabrowser.net) [29]. Tumour samples with complete overall-survival information were retained, yielding a discovery cohort of 372 CRC samples. The discovery cohort comprised both colon (COAD, n = 279) and rectal (READ, n = 93) tumours; the READ samples were additionally analysed as a rectal-cancer subset to assess generalisability across anatomical subsites (a descriptive analysis, as these samples contributed to model development). The Gene Expression Omnibus cohort GSE39582, profiled on the Affymetrix HG-U133_Plus_2 (GPL570) platform, was additionally used [30]; probe-level data were mapped to gene symbols using the GPL570 annotation and log2-transformed. The Gene Expression Omnibus cohort GSE87211, comprising pre-treatment tumour biopsies from 196 patients with locally advanced rectal cancer, was used as an additional independent validation cohort; probe-level data were mapped to gene symbols using the GPL13497 platform annotation. The extended TCGA PanCanAtlas cohort (coadread_tcga_pan_can_atlas_2018; primary tumours with survival data, n = 566) was retrieved through the cBioPortal API [31,32] and used for the robustness, immune-infiltration and clinical-utility analyses.

Gene sets and definition of NETs–coagulation crosstalk genes

A curated NETs-related gene set of 97 genes was assembled from the Gene Ontology term for NET formation (GO:0140645) and classical NETosis-associated genes reported in the literature [7,23]. The coagulation gene set comprised the 138 genes of the MSigDB Hallmark coagulation gene set [33]. Differentially expressed genes (DEGs) between CRC tumour and adjacent normal tissues were identified with thresholds of $|\log2 \text{ fold change}| > 1$ and false discovery rate (FDR) < 0.05. A per-sample coagulation activity score was computed by ssGSEA [34]. NETs–coagulation crosstalk genes were defined as NETs genes whose expression showed significant Pearson correlation with the coagulation ssGSEA score ($|r| > 0.3$, $P < 0.05$).

Feature selection and risk model construction

Crosstalk genes with $P < 0.1$ in univariate Cox regression for overall survival were retained as candidates. Three algorithms were applied: (1) LASSO-Cox regression with ten-fold cross-validation to select the penalty parameter [18]; (2) SVM-RFE with a linear support vector machine [19]; and (3) RF with 500 trees, retaining genes with importance above the mean [20]. Genes selected by at least two of the three algorithms constituted the final signature, and a risk score was calculated from multivariate Cox regression coefficients [35]. Patients were dichotomised into high- and low-risk groups by the median risk score.

Survival analysis and model evaluation

Kaplan–Meier curves were compared using the log-rank test [36]. Time-dependent ROC curves were used to estimate the area under the curve (AUC) at 1, 3 and 5 years [37]. Harrell's concordance index (C-index) was computed, and optimism was quantified by bootstrap resampling (1000 iterations) [38]. Univariate and multivariate Cox regression adjusting for age and tumour stage assessed the independence of the risk score [35]. A nomogram integrating the risk score with clinical variables was constructed, and a calibration curve evaluated agreement between predicted and observed 3-year survival. Analyses were performed in Python 3.13 using lifelines [39] and scikit-learn [40]. As a sensitivity analysis, classification was additionally evaluated using the optimal risk-score cut-off identified by maximising the log-rank statistic across

quantiles. Clinical utility was assessed by decision-curve analysis (net benefit across threshold probabilities at 1, 3 and 5 years) and by the continuous (category-free) net reclassification improvement (NRI) at 3 years, in which cases were defined as death within 3 years and controls as patients surviving beyond 3 years, with patients censored before 3 years excluded. Two-sided $P < 0.05$ was considered statistically significant.

Immune infiltration, checkpoints and drug targets

Relative abundances of 16 immune cell types were estimated by ssGSEA using established marker gene sets 32 and compared between risk groups (Mann–Whitney U test). Expression of eight immune-checkpoint genes (PDCD1, CD274, PDCD1LG2, CTLA4, LAG3, HAVCR2, TIGIT and IDO1) and sixteen clinically relevant drug-target genes was likewise compared between groups. P values were nominal unless stated otherwise; in the extended-cohort analyses, Benjamini–Hochberg false discovery rates (FDR) were additionally computed within each analysis family (immune cell types, checkpoint genes and drug-target genes).

External validation

The final model was first examined in the rectal-cancer subset of the discovery cohort (TCGA-READ); because this subset contributed to model development, that analysis is descriptive, and independent external validation relied on the cohorts below. For GSE39582, because SIGLEC14 has no probe on the Affymetrix HG-U133_Plus_2 platform, single-gene survival analyses were performed for ADAMTS13 and CRISPLD2 and a two-gene exploratory sub-score was evaluated; these analyses were interpreted as exploratory. The full three-gene risk score was then validated in the independent GSE87211 cohort using median-dichotomised Kaplan-Meier analysis and continuous Cox regression of disease-free and overall survival, with multivariate adjustment for age and sex. Validation was extended to four further independent cohorts: the colon-cancer cohorts GSE17536 (n = 177) and GSE17537 (n = 55) [41], the disease-specific-survival cohort GSE38832 (n = 122), and an RNA-sequencing cohort of Chinese patients with CRC (SiLU; cBioPortal study coad_silu_2022; n = 348). Because SIGLEC14 is not covered by the Affymetrix HG-U133_Plus_2 platform, microarray cohorts were analysed using the two-gene ADAMTS13–CRISPLD2 sub-score, whereas the full three-gene score was applied in the SiLU RNA-sequencing cohort; cohort-level hazard ratios were combined by inverse-variance fixed-effect meta-analysis.

Ethics approval and consent to participate

This study used publicly available, de-identified data from TCGA and GEO. Ethical approval was not required in accordance with institutional and national guidelines.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed in this study are publicly available from the UCSC Xena data portal (https://xenabrowser.net), The Cancer Genome Atlas (https://www.cancer.gov/tcga) and the Gene Expression Omnibus (GSE39582, GSE87211, GSE17536, GSE17537 and GSE38832) and the

cBioPortal for Cancer Genomics (SiLU cohort, coad_silu_2022). Analysis code is provided as a supplementary code archive (Code S1) and is available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this work.

Authors' contributions

Zelin Chen and E. Xie conceived and designed the study. Zelin Chen performed the bioinformatic analyses and drafted the manuscript. Jinwen Fang and Xiangyang Yang contributed to data interpretation and critically revised the manuscript. E. Xie supervised the study and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the staff of the Department of Hepatobiliary, Pancreatic, Gastrointestinal, Hernia and Abdominal Wall Surgery, Shantou Hospital of Traditional Chinese Medicine for their support and encouragement. We also acknowledge the Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases for making the data used in this study publicly available.

Use of artificial intelligence tools

During the preparation of this work, the authors used ChatGPT (OpenAI, San Francisco, CA, USA; https://chatgpt.com) to improve the readability and language of the manuscript, and reviewed and edited the content as needed, taking full responsibility for the content of the publication.

References

1. Bray, F., Laversanne, M., Sung, H. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 74, 229–263 (2024).

2. Siegel, R. L., Wagle, N. S., Cercek, A. et al. Colorectal cancer statistics, 2023. CA: A Cancer Journal for Clinicians 73, 233–254 (2023).

3. Hisada, Y. & Mackman, N. Mechanisms of cancer-associated thrombosis. Research and Practice in Thrombosis and Haemostasis 7, 100123 (2023).

4. Mulder, F. I., Horváth-Puhó, E., van Es, N. et al. Venous thromboembolism in cancer patients: A population-based cohort study. Blood 137, 1959–1969 (2021).

5. Khorana, A. A., Kuderer, N. M., McCrae, K. et al. Cancer-associated thrombosis and mortality in patients with cancer stratified by Khorana score risk levels. Cancer Medicine 9, 8062–8073 (2020).

6. Wolach, O. & Martinod, K. Casting a NET on cancer: The multiple roles for neutrophil extracellular traps in cancer. Current Opinion in Hematology 29, 53–62 (2022).

7. Brinkmann, V., Reichard, U., Goosmann, C. et al. Neutrophil extracellular traps kill bacteria. Science 303, 1532–1535 (2004).

8. Fuchs, T. A., Brill, A., Duerschmied, D. et al. Extracellular DNA traps promote thrombosis. Proceedings of the National Academy of Sciences 107, 15880–15885 (2010).

9. Zhou, Y., Xu, Z. & Liu, Z. Impact of neutrophil extracellular traps on thrombosis formation: New findings and future perspective. Frontiers in Cellular and Infection Microbiology 12, 910908 (2022).

10. Cools-Lartigue, J., Spicer, J., McDonald, B. et al. Neutrophil extracellular traps sequester circulating tumor cells and promote metastasis. Journal of Clinical Investigation 123, 3446–3458 (2013).

11. Zhong, W., Wang, Q., Shen, X. et al. Neutrophil extracellular trap is surrogate biomarker for prognosis and response to neoadjuvant therapy in locally advanced rectal cancer. Journal of Inflammation Research 16, 6443–6455 (2023).

12. Dong, J., Zhao, J., Wu, Z. et al. The predictive value of neutrophil extracellular trap-related risk score in prognosis and immune microenvironment of colorectal cancer patients. Molecular Biotechnology 67, 1509–1525 (2025).

13. Qian, J., Duan, J. & Cao, D. Identification of a novel four-gene prognostic model related to neutrophil extracellular traps for colorectal cancer. Turkish Journal of Gastroenterology 35, 849–858 (2024).

14. Wang, C., Chen, J., Song, Z. et al. Neutrophil extracellular traps as biomarkers for predicting prognosis and chemotherapy response in colorectal cancer. Scientific Reports 16, 3074 (2026).

15. Jin, Y., Liao, L., Chen, Q. et al. Multi-omics analysis reveals that neutrophil extracellular traps related gene TIMP1 promotes CRC progression and influences ferroptosis. Cancer Cell International 25, 31 (2025).

16. Zhang, D., Zhao, Y., Wang, S. et al. A prognostic model of angiogenesis and neutrophil extracellular traps related genes manipulating tumor microenvironment in colon cancer. Journal of Cancer 14, 2109–2127 (2023).

17. Cao, X., Lan, Q., Xu, H. et al. Granulocyte-like myeloid-derived suppressor cells: The culprits of neutrophil extracellular traps formation in the pre-metastatic niche. International Immunopharmacology 143, 113500 (2024).

18. Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58, 267–288 (1996).

19. Guyon, I., Weston, J., Barnhill, S. et al. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002).

20. Breiman, L. Random forests. Machine Learning 45, 5–32 (2001).

21. Colonne, C. K., Favaloro, E. J. & Pasalic, L. The intriguing connections between von Willebrand factor, ADAMTS13 and cancer. Healthcare 10, 557 (2022).

22. Obermeier, H. L., Riedl, J., Ay, C. et al. The role of ADAMTS-13 and von Willebrand factor in cancer patients: Results from the Vienna Cancer and Thrombosis Study. Research and Practice in Thrombosis and Haemostasis 3, 503–514 (2019).

23. Wang, H., Kim, S. J., Lei, Y. et al. Neutrophil extracellular traps in homeostasis and disease. Signal Transduction and Targeted Therapy 9, 235 (2024).

24. Ali, S. R., Fong, J. J., Carlin, A. F. et al. Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. Journal of Experimental Medicine 211, 1231–1242 (2014).

25. Macauley, M. S., Crocker, P. R. & Paulson, J. C. Siglec-mediated regulation of immune cell function in disease. Nature Reviews Immunology 14, 653–666 (2014).

26. Zhao, Y., Du, J. & Shen, X. Targeting myeloid-derived suppressor cells in tumor immunotherapy: Current, future and beyond. Frontiers in Immunology 14, 1157537 (2023).

27. Barry, S. T., Gabrilovich, D. I., Sansom, O. J. et al. Therapeutic targeting of tumour myeloid cells. Nature Reviews Cancer 23, 216–237 (2023).

28. Wu, Y., Yi, M., Niu, M. et al. Myeloid-derived suppressor cells: An emerging target for anticancer immunotherapy. Molecular Cancer 21, 184 (2022).

29. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012).

30. Marisa, L., de Reyniès, A., Duval, A. et al. Gene expression classification of colorectal cancer into molecular subtypes. PLoS Medicine 10, e1001453 (2013).

31. Cerami, E., Gao, J., Dogrusoz, U. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery 2, 401–404 (2012).

32. Hoadley, K. A., Yau, C., Hinoue, T. et al. Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 173, 291–304 (2018).

33. Liberzon, A., Birger, C., Thorvaldsdóttir, H. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Systems 1, 417–425 (2015).

34. Barbie, D. A., Tamayo, P., Boehm, J. S. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–112 (2009).

35. Cox, D. R. Regression models and life-tables. Journal of the Royal Statistical Society: Series B 34, 187–220 (1972).

36. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481 (1958).

37. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56, 337–344 (2000).

38. Harrell, F. E. Jr, Califf, R. M., Pryor, D. B. et al. Evaluating the yield of medical tests. JAMA 247, 2543–2547 (1982).

39. Davidson-Pilon, C. lifelines: Survival analysis in Python. Journal of Open Source Software 4, 1317 (2019).

40. Pedregosa, F., Varoquaux, G., Gramfort, A. et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).

41. Smith, J. J., Deane, N. G., Wu, F. et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 139, 959–968 (2010).

Figure legends

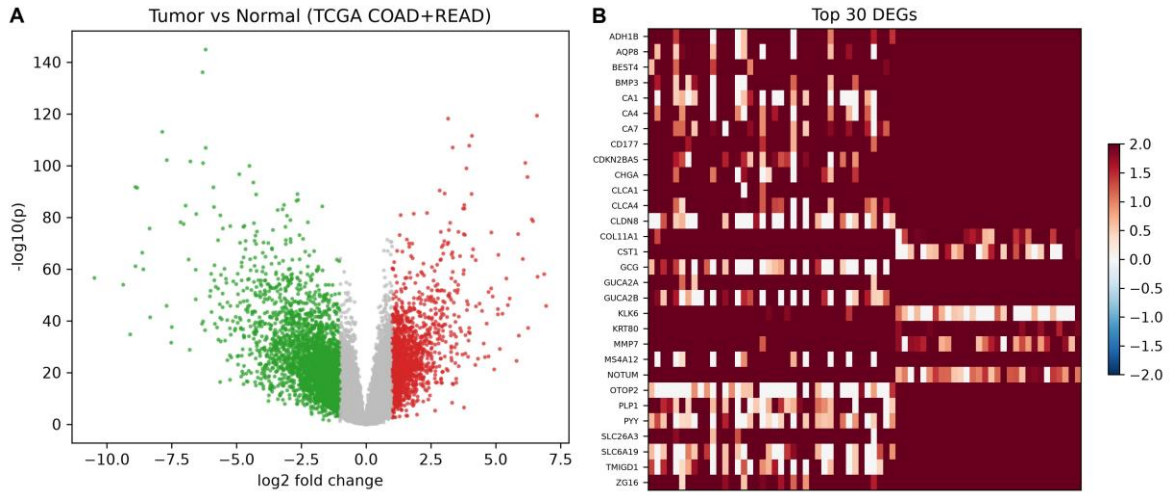


Fig. 1a. Volcano plot of differentially expressed genes between CRC tumour and adjacent normal tissues (4,798 DEGs; FDR < 0.05, |log2FC| > 1).

Fig. 1b. Heatmap of the top differentially expressed genes between tumour and normal samples.

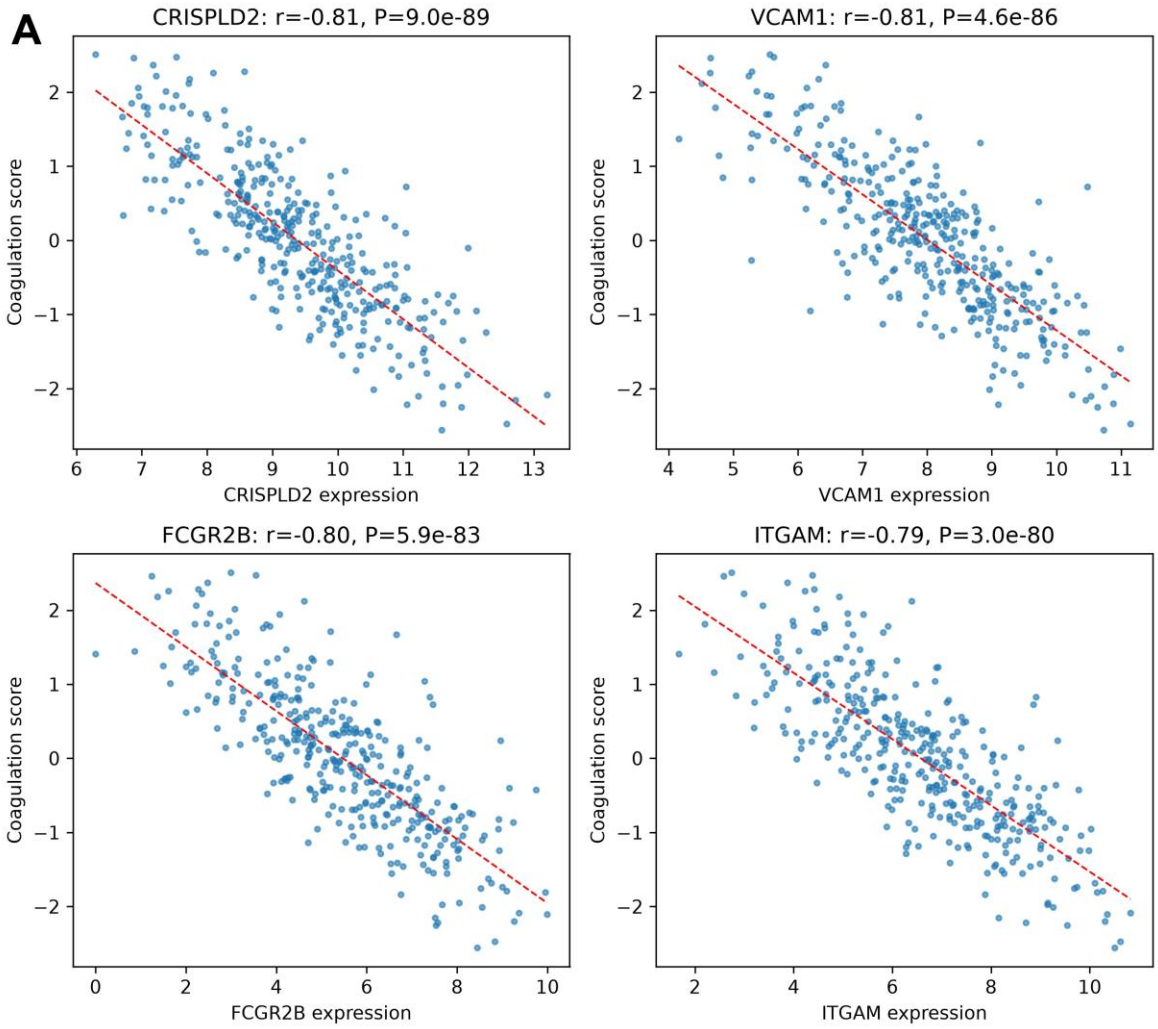


Fig. 2. Identification of NETs–coagulation crosstalk genes. Pearson correlation between the expression of NETs-related genes and the ssGSEA coagulation activity score; 58 genes met the criteria $|r| > 0.3$ and $P < 0.05$.

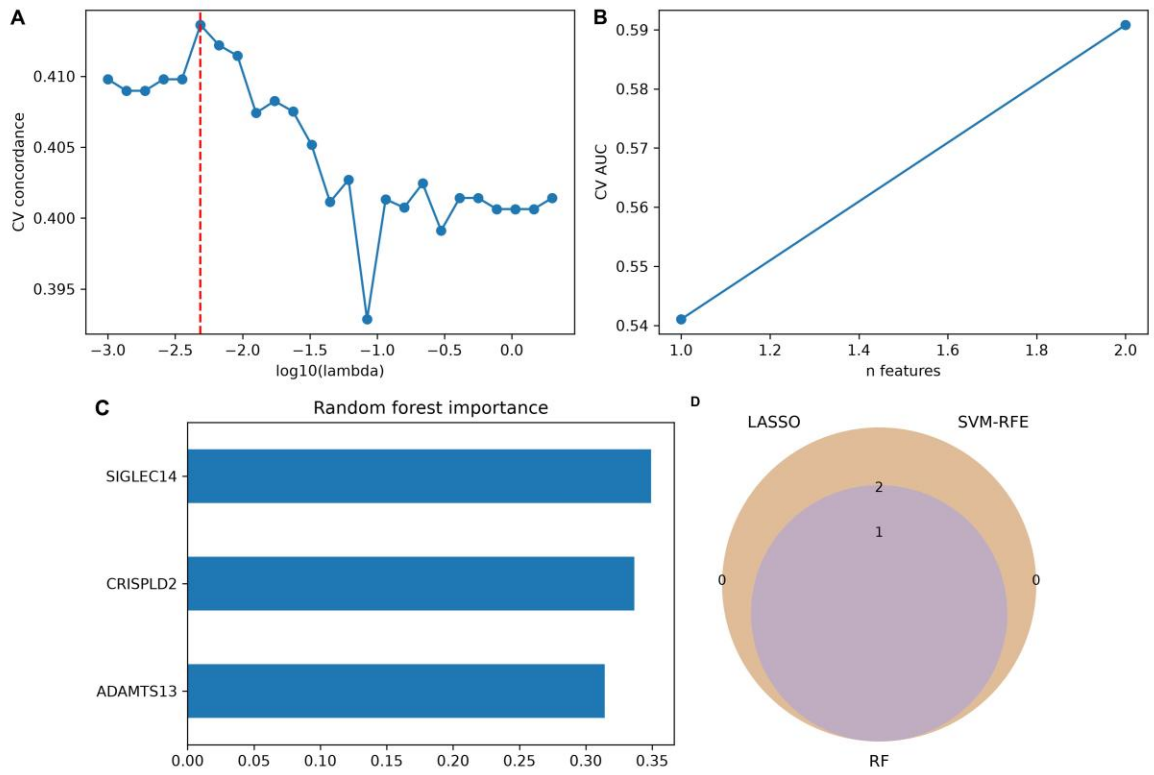


Fig. 3a. LASSO-Cox regression: partial-likelihood deviance across the penalty parameter; λ = 0.00487 was selected by ten-fold cross-validation.

Fig. 3b. SVM-RFE: cross-validation accuracy across the number of retained features.

Fig. 3c. Random forest gene importance (500 trees); genes above the mean importance are retained.

Fig. 3d. Venn diagram of genes selected by LASSO, SVM-RFE and RF. ADAMTS13, CRISPLD2 and SIGLEC14 each received ≥2 votes.

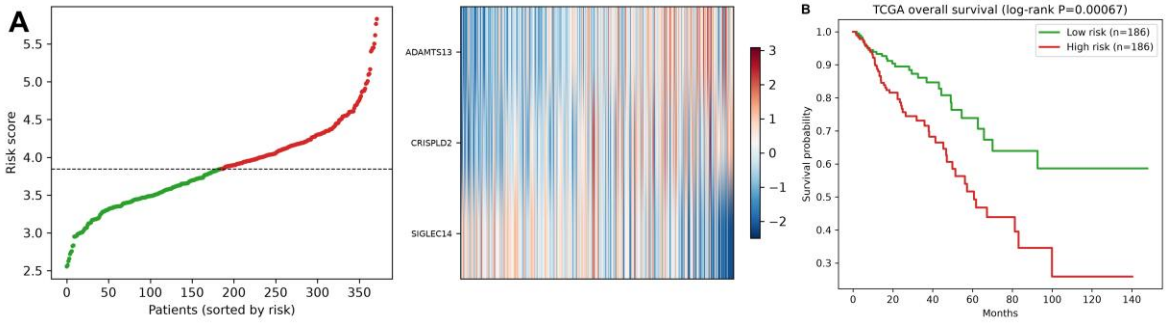


Fig. 4a. Distribution of the risk score and survival status in the TCGA discovery cohort (n = 372).

Fig. 4b. Kaplan–Meier curves of overall survival for high- and low-risk groups (log-rank P = 0.00067).

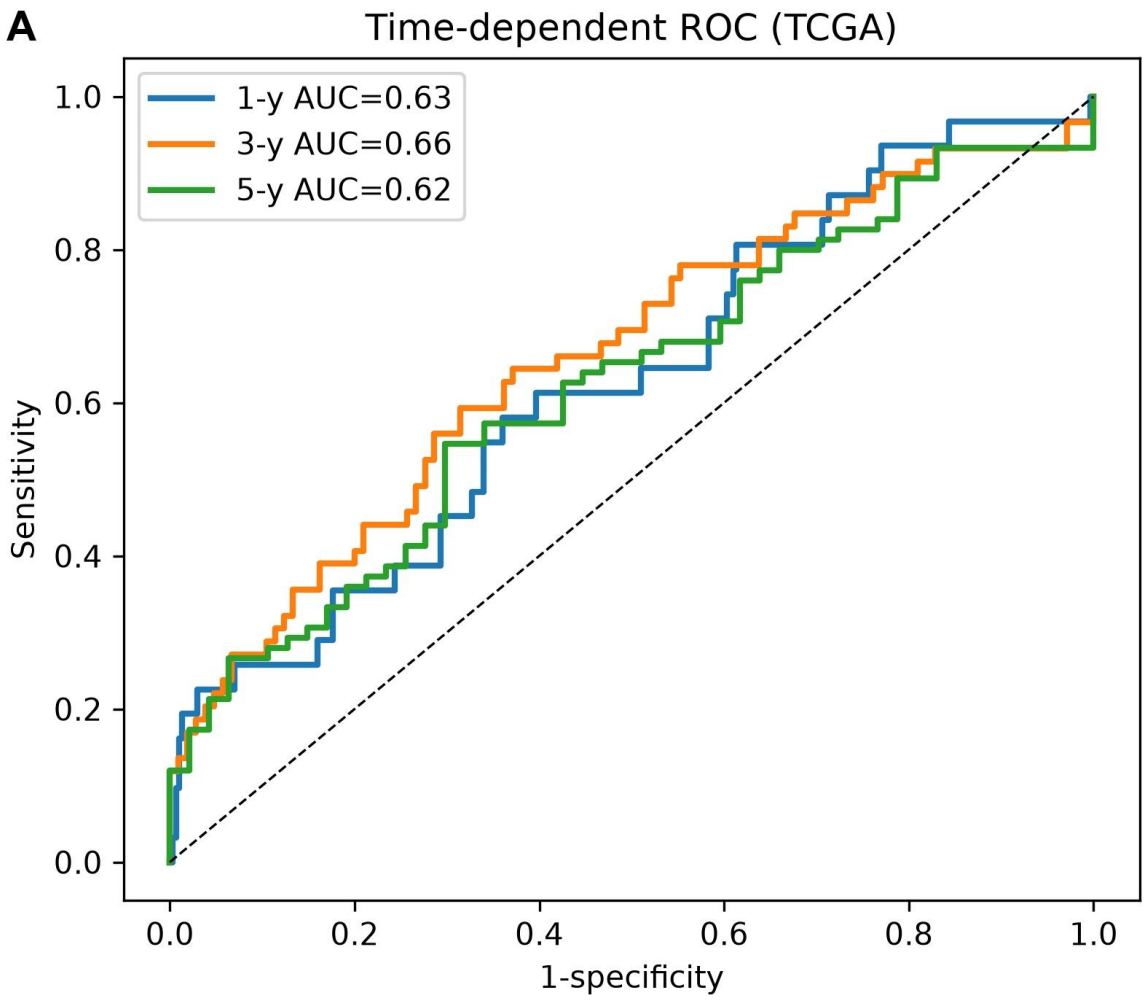


Fig. 5. Time-dependent ROC curves for 1-, 3- and 5-year overall survival prediction (AUC 0.627, 0.660 and 0.623).

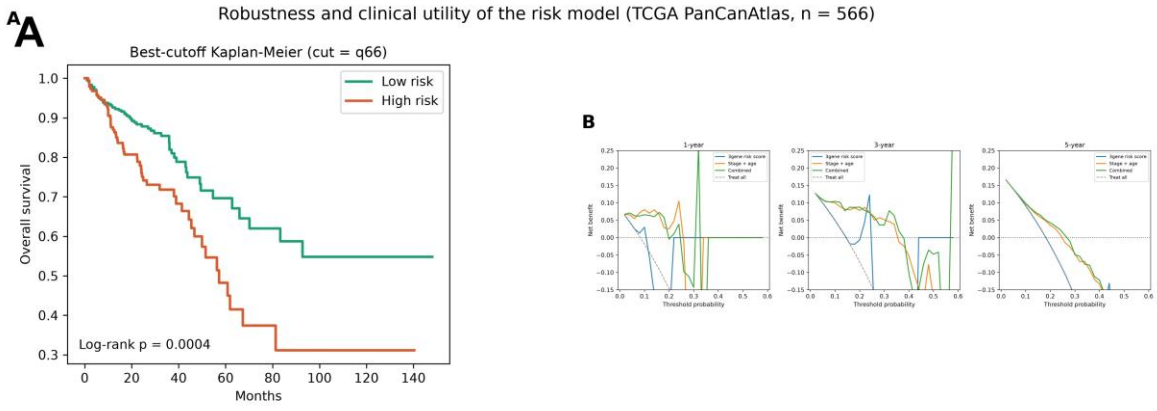


Fig. 6. Robustness and clinical utility of the risk model in the extended TCGA PanCanAtlas cohort (n = 566). (A) Kaplan–Meier curves using the optimal risk-score cut-off (log-rank P =

0.0004). (B) Decision-curve analysis at 1, 3 and 5 years comparing the three-gene risk score, clinical factors (age and tumour stage) and the combined model.

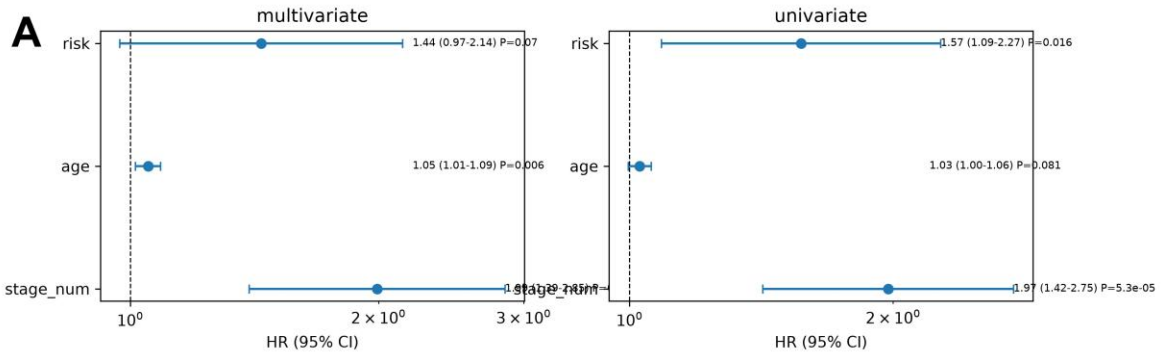


Fig. 7. Forest plots of univariate and multivariate Cox regression for overall survival.

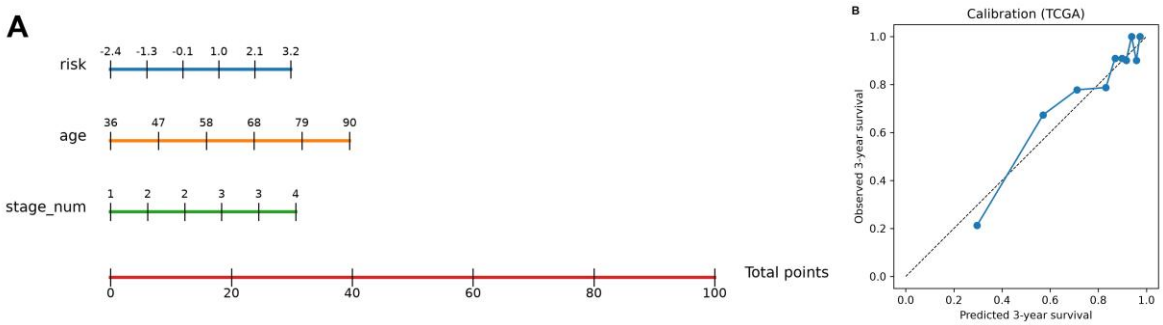


Fig. 8a. Nomogram integrating risk score, age and tumour stage for predicting 3-year overall survival.

Fig. 8b. Calibration curve of the nomogram for 3-year survival.

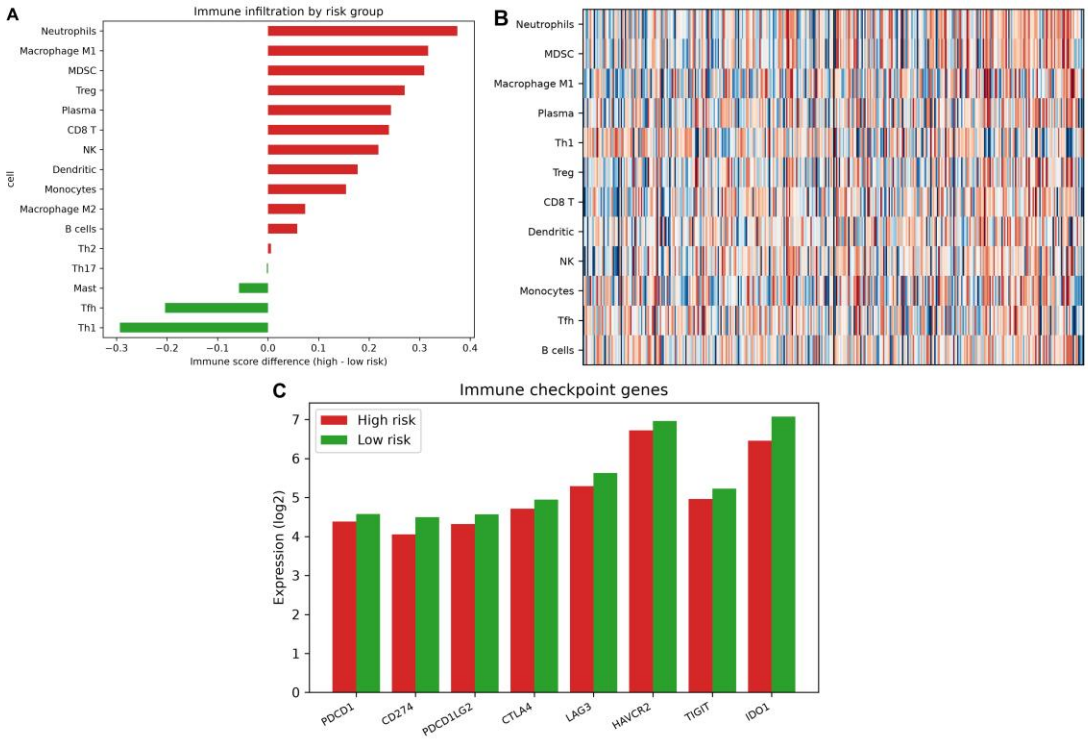


Fig. 9a. Differences in ssGSEA immune-cell scores between high- and low-risk groups in the TCGA discovery cohort (Mann–Whitney U test; nominal P values).

Fig. 9b. Heatmap of immune-cell infiltration across individual samples by risk group.

Fig. 9c. Expression of immune-checkpoint genes between high- and low-risk groups.

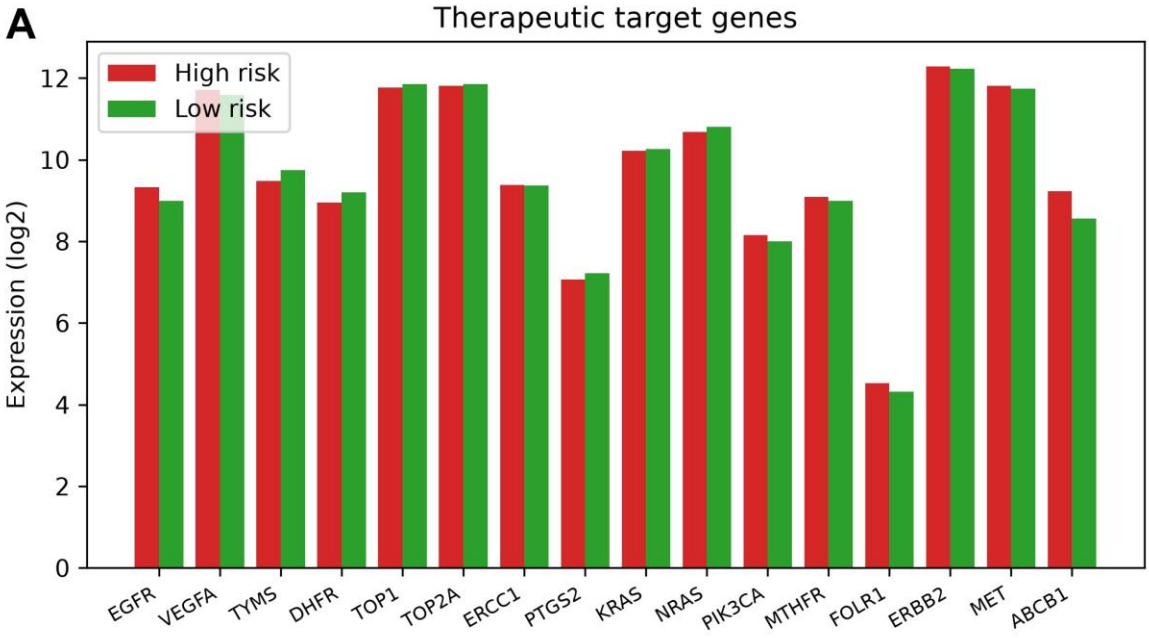


Fig. 10. Expression of clinically relevant drug-target genes between high- and low-risk groups.

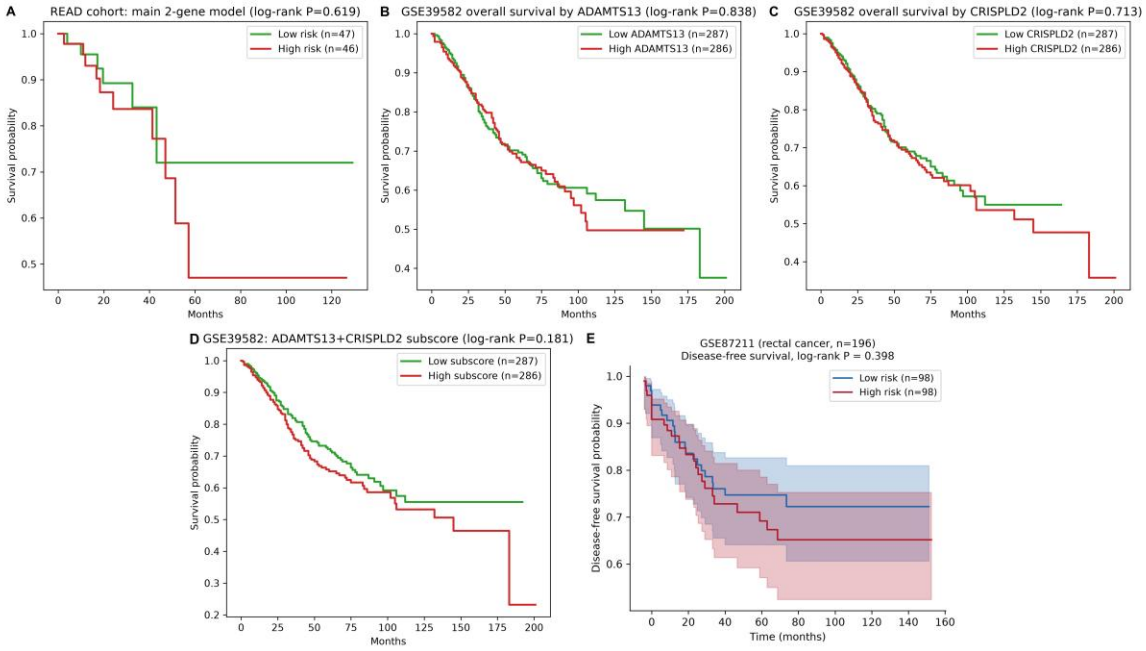


Fig. 11a. Kaplan–Meier curves of the three-gene model in the rectal-cancer subset (TCGA-READ) of the discovery cohort (n = 93; descriptive analysis, as these samples contributed to model development; log-rank P = 0.619).

Fig. 11b. Kaplan–Meier curves for ADAMTS13 expression in GSE39582 (log-rank P = 0.838).

Fig. 11c. Kaplan–Meier curves for CRISPLD2 expression in GSE39582 (log-rank P = 0.713).

Fig. 11d. Kaplan–Meier curves for the exploratory two-gene sub-score in GSE39582 (log-rank P = 0.181).

Fig. 11e. Kaplan–Meier curves of the three-gene risk score in the GSE87211 rectal-cancer validation cohort (n = 196; disease-free survival by median risk-score split, log-rank P = 0.398).

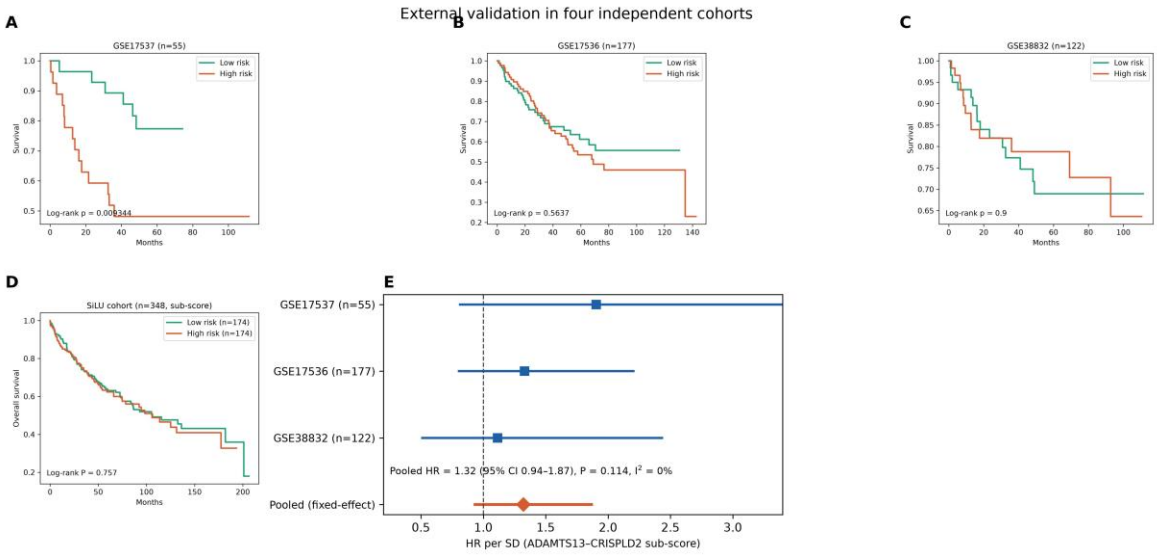


Fig. 12. External validation in four independent cohorts. (A) Kaplan–Meier curves of the ADAMTS13–CRISPLD2 sub-score in GSE17537 (log-rank P = 0.009). (B) GSE17536 (P = 0.56). (C) GSE38832 (P = 0.90). (D) SiLU RNA-sequencing cohort, analysed with the ADAMTS13–CRISPLD2 sub-score (median-split log-rank P = 0.76; Cox HR 1.10 per SD, P = 0.54; the full three-gene score gave a similar non-significant result). (E) Fixed-effect meta-analysis of the sub-score across the four cohorts (n = 702; the sub-score was used in all four cohorts for comparability); all cohorts showed hazard ratios above unity with no heterogeneity (I^2 = 0%).

Supplementary Fig. S1. Immune-infiltration analysis in the extended TCGA PanCanAtlas cohort (n = 566). (a) Differences in ssGSEA immune-cell scores between risk groups; (b) heatmap of immune-cell scores for individual samples ordered by risk score. P values were adjusted across the 16 cell types by the Benjamini–Hochberg method; FDR < 0.05 for M1 macrophages (lower in high-risk) and NK cells (higher in high-risk).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CodeS1.zip
- SciRepS1.docx