UNIVERSITÀ DEGLI STUDI DI SALERNO

Department of Medicine, Surgery and Dentistry

Ph.D. Program in "Translational Medicine of Development and Active Aging"

XXXVIII Cycle

Doctoral Thesis in

IDENTIFICATION OF NOVEL EPITHELIAL-MESENCHYMAL TRANSITION (EMT) AND DNA DAMAGE REPAIR (DDR)-RELATED PROGNOSTIC GENE SIGNATURES AS PREDICTIVE BIOMARKERS IN SIX CANCER COHORTS

Tutor
Dr. Federica Papaccio

Ph.D. Student
Abu Hazafa
Mat: 8861700011

Coordinator
Prof. Palmiero Monteleone

ACADEMIC YEAR 2022/2025

DEDICATION

I dedicated this thesis to my parents, who supported me to complete this work and made me to believe that I dream well, and I can achieve well. Their understanding and belief in me have guided each step of this journey.

I also dedicated this thesis to my beloved wife, whose patience, encouragement, and understanding have been my constant strength.

I also dedicated this thesis to my family and friends, for their faith, kindness, and inspiration that sustained me throughout this endeavour.

Finally, I dedicated this dissertation to God Almighty, who granted me tell my strengths and weaknesses, all sickness and health, and all happiness and sorrow.

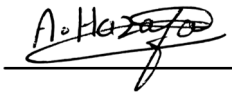
“I was taught that the way of progress was neither swift nor easy.”

— Marie Curie

DECLARATIONS

I hereby declare that this thesis represents my own work which has been done under the supervision of Dr. Federica Papaccio for the degree of Doctor of Philosophy (PhD) at University of Salerno, and that it has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

All data sources, references, and collaborations used in this work have been duly acknowledged, and I affirm that the research has been conducted in accordance with the ethical standards and policies of University of Salerno.

Signature: 

Date: 02-12-2025

ACKNOWLEDGEMENT

It gives me immense pleasure to express my deepest gratitude and sincere thanks to my respected supervisor, **Dr. Federica Papaccio**, for her special attention, consistent encouragement, expertise suggestion and constructive criticism without which, I really would never have been able to complete this thesis. I am extremely grateful for her mentorship and for the supporting me to pursue my PhD under her supervision.

I extend my sincere gratitude to **Prof. Valeria Conti**, in whose laboratory I received extensive training in molecular biology techniques at Ospedali riuniti San Giovanni di Dio e Ruggi d'Aragona, Salerno. I also grateful to **Prof. Antonello Petrella** for accepting me into his laboratory at the Department of Pharmacy, University of Salerno for the duration of my six-month internship. My sincere gratitude goes to all my colleagues and staff members in both laboratories for their cooperation, technical support, and friendship.

Finally, I want to thank all of my family, most of all my parents and my wife for the love, care and faith in me. Their support has been the last but not least power and motivation.

ABSTRACT

Background: Breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), and stomach adenocarcinoma (STAD) account for ~40.97% of all cancer cases and ~46.35% of cancer-related deaths worldwide. Biomarker limitations restrict therapy options, making these cancers difficult to cure. Here, we present a novel dual-signature, multi-cancer comparative analysis integrating epithelial-mesenchymal transition (EMT) and DNA damage repair (DDR) prognostic gene signatures.

Methods: RNA-Seq profiles and corresponding clinical data from The Cancer Genome Atlas (TCGA) project were used to identify differentially expressed genes (DEG). Univariate Cox, Kaplan-Meier (KM), and Least Absolute Shrinkage and Selection Operator (LASSO) regression analyses were applied to identify prognostic EMT- and DDR-related genes and construct final models. Validation was performed using univariate and multivariate Cox, ROC curves, AUC analysis, t-SNE, and Principal Component Analysis (PCA). Kyoto Encyclopaedia of Genes and Genomes (KEGG) and Gene Ontology-Biological Process (GO-BP) were performed to assess pathways involved in the prognostic signatures. CIBERSORT.X and ESTIMATE algorithms were used to analyse immune infiltration, and a nomogram model was developed by integrating clinical factors to identify prognostic signatures. Drug sensitivity correlation analysis was performed between the risk score of cancer cell lines and the IC_{50} values of drugs. The prognostic ability of identified EMT- and DDR-related signatures was validated in Gene Expression Omnibus (GEO) datasets.

Results: Based on Cox, KM, and LASSO analysis, we identified a novel prognostic six-gene related signature in BRCA, six- and two-gene in COAD, nine- and five-gene in LUAD, four-gene in LUSC, ten-, and eight-gene in PAAD, and nine-, and seven-gene related signatures in STAD for EMT and DDR, respectively. ROC and AUC values indicated effective performance across cancers, particularly in PAAD. PCA and t-SNE confirmed strong discrimination at the transcriptomic level. UniCox and MultiCox analyses demonstrated that EMT/DDR-related signatures-based risk scores were independent prognostic variables. Time-dependent ROC analysis showed that the nomogram achieved higher AUCs than individual clinical factors. Immune analysis revealed reduced infiltration of macrophages (M0) and naïve B cells in high-risk group suggesting an immunosuppressive tumour microenvironment in high-risk patients.

Correlation analysis with drug IC_{50} values in cancer cell lines suggested that high-risk scores were associated with predicted sensitivity to (e.g., DNA-damaging agents) in BRCA and COAD, but with predicted resistance to (e.g., certain chemotherapies) in the other cancer types.

Conclusion: The proposed EMT- and DDR-related prognostic gene signatures performed well for overall survival prediction across cancers and could serve as promising biomarkers.

GRAPHICAL ABSTRACT

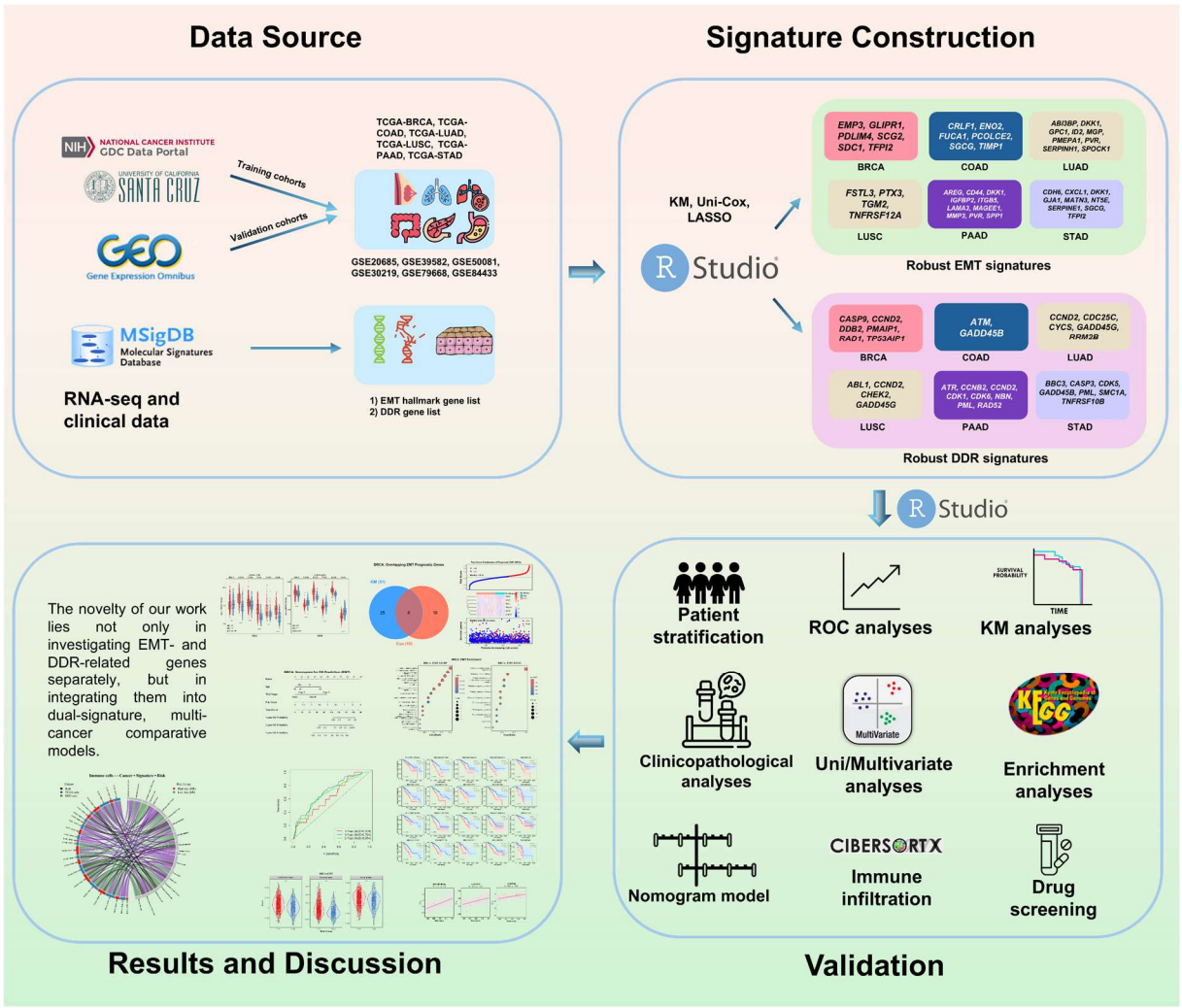


TABLE OF CONTENTS

DECLARATIONS ---------- **iv**

ACKNOWLEDGEMENT ---------- **v**

ABSTRACT ---------- **vi**

GRAPHICAL ABSTRACT ---------- **viii**

LIST OF FIGURES ---------- **xii**

LIST OF TABLES ---------- **xv**

LIST OF ABBREVIATIONS ---------- **xvi**

INTRODUCTION ---------- **1**

1.1. Global Burden of Cancer ---------- 1

1.2. The Clinical Challenge: Limitations of Current Prognostic and Predictive Biomarkers ---------- 2

1.3. Overview of Six Cancer Cohorts ---------- 3

1.3.1. Breast Invasive Carcinoma (BRCA) ---------- 3

1.3.2. Lung Adenocarcinoma (LUAD) ---------- 4

1.3.3. Lung Squamous Cell Carcinoma (LUSC) ---------- 5

1.3.4. Stomach Adenocarcinoma (STAD) ---------- 6

1.3.5. Pancreatic Adenocarcinoma (PAAD) ---------- 6

1.3.6. Colon Adenocarcinoma (COAD) ---------- 7

1.4. Epithelial-Mesenchymal Transition (EMT) as a Driver of Cancer Progression ---------- 8

1.4.1. EMT-Related Gene Signatures as Prognostic Tools ---------- 10

1.5. DNA Damage Response (DDR): From Genome Protection to Cancer Progression ---------- 11

1.5.1. DDR-Related Signature as Prognostic and Predictive Biomarkers ---------- 13

1.6. Molecular Crosstalk Between EMT and DDR ---------- 14

RESEARCH RATIONALE, NOVELTY, AND OBJECTIVES ---------- 17

2.1. Rationale and Novelty of the Study ---------- 17

2.2. Research Hypotheses ---------- 18

2.3. Aim and Objectives ---------- 18

2.3.1. Aim of the Study ---------- 18

2.3.2. Specific Objectives ---------- 19

2.4. Expected Outcome ---------- 19

MATERIALS AND METHODS ---------- 21

3.1. Data Collection ---------- 21

3.2. Differential Gene Expression (DGE) Screening ---------- 23

3.3. Construction and Validation of Novel EMT- and DDR-Related Prognostic Signatures ---- 23

3.4. Independent Prognostic Factor ---------- 25

3.5. Relationship Between Clinical Characteristics and Risk Levels ---------- 25

3.6. Functional Enrichment Analysis ---------- 25

3.7. Immune Cell Infiltration Profiling ---------- 25

3.8. Signature-Based Nomogram Construction and Validation ---------- 26

3.9. Drug Sensitivity Correlation Analysis ---------- 27

3.10. Single-sample GSEA (ssGSEA) ---------- 27

3.11. Statistical Analysis ---------- 27

RESULTS ---------- 29

4.1. Identification and Screening of Prognostic DGE EMT- and DDR-Related Gene Signatures 29

4.2. Validation of EMT and DDR-Related Gene Signatures ---------- 36

4.3. Independent Prognostic Factors of EMT- and DDR-Related Signatures ---------- 42

4.4. Kaplan-Meier Curves for EMT- and DDR-Derived Risk Groups and Clinicopathological Subgroups ---------- 47

4.5. Relationship Between Clinicopathological Characteristics and EMT-/DDR-Derived Risk Groups 56

4.6. Pathways Enrichment Analysis Highlights Humoral Immune Response, Extracellular Matrix Organization, and Extracellular Structure Organization by EMT- and Mitotic Networks by DDR-Related Prognostic Signatures ---------- 61

4.7. Combined Prognostic Analysis of EMT- and DDR-Related Risk Scores and Immune Cell Infiltration---------- 63

4.8. Signature-Based Nomogram Provides Accurate Prognostic Estimates ---------- 79

4.9. Expression and Survival Analysis of Prognostic EMT- and DDR-Related Signature Genes 82

4.10. Low-Risk and High-Risk Groups Possess Distinct ssGSEA Scores Across Cancers ------- 83

4.11. Drug Screening Correlation Analysis Identifies Putative Compounds Showing Anti-Tumoral Activity for EMT- and DDR-Related Prognostic Gene Signatures Cell Lines Across Cancers 85

DISCUSSION ---------- 92

CONCLUSIONS ---------- 100

BIBLIOGRAPHY ---------- 101

ANNEXURES ---------- A

Annexure I – Publication ---------- A

Annexure II – Appendices ---------- 1

Appendix A — Supplementary Files 1 ---------- A-1

Appendix B — Supplementary Files 2 ---------- B-1

Appendix C — Supplementary Files 3 ---------- C-1

Appendix D — Supplementary Files 4 ---------- D-1

LIST OF FIGURES

Figure 1.1. Role of epithelial to epithelial–mesenchymal transition (EMT) in wound heading and cancer progression. This image was created with BioRender (www.biorender.com) (Den Hollander et al., 2024). 9

Figure 1.2. Overview of targeting the DDR pathways and cell-cycle checkpoints. 12

Figure 1.3. The molecular crosstalk between EMT and DDR in cancer progression. 16

Figure 3.1. Schematic representation of the study workflow. 22

Figure 4.1. In the TCGA-BRCA and TCGA-COAD cohorts, a series of analyses were performed to identify EMT- and DDR-related prognostic candidate genes. 30

Figure 4.2. In the TCGA-LUAD and TCGA-LUSC cohorts, a series of analyses were performed to identify EMT- and DDR-related prognostic candidate genes. 32

Figure 4.3. In the TCGA-PAAD and TCGA-STAD cohort, a series of analysis were performed to identify EMT- and DDR-related prognostic candidate genes. 34

Figure 4.4. Results of LASSO regression to construct EMT- and DDR-related prognostic gene signatures models in TCGA-BRCA and TCGA-COAD cohorts. 37

Figure 4.5. Results of LASSO regression to construct EMT- and DDR-related prognostic gene signatures models in TCGA-LUAD and TCGA-LUSC cohorts. 39

Figure 4.6. Results of LASSO regression to construct EMT- and DDR-related prognostic gene signatures models in TCGA-PAAD and TCGA-STAD cohorts. 40

Figure 4.7. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-BRCA cohort. 48

Figure 4.8. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-COAD cohort. 50

Figure 4.9. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-LUAD cohort. 51

Figure 4.10. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-LUSC cohort. 53

Figure 4.11. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-PAAD cohort. 54

Figure 4.12. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-STAD cohort. 55

Figure 4.13. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-BRCA cohort. 65

Figure 4.14. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-COAD cohort. 67

Figure 4.15. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-LUAD cohort. 69

Figure 4.16. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-LUSC cohort. 71

Figure 4.17. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-PAAD cohort. 73

Figure 4.18. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in the TCGA-STAD cohort. 75

Figure 4.19. ssGSEA analysis for EMT and DDR-based risk groups across all six cancers. (a) Relationship between ssGSEA score and risk groups across cancers. (b) Overlapping of samples between EMT and DDR risk groups (low vs high) across cancers. *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$). 84

Figure 4.20. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in BRCA cell lines. 86

Figure 4.21. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in COAD cell lines. 87

Figure 4.22. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in LUAD cell lines. 88

Figure 4.23. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in LUSC cell lines. 89

Figure 4.24. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in PAAD cell lines. 90

Figure 4.25. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in STAD cell lines. 91

LIST OF TABLES

Table 4.1. Univariate and multivariate Cox analysis relationship between clinicopathological factors and risk scores associated with OS across six cancers in test cohort (TCGA) for prognostic EMT and DDR-related signatures. 43

Table 4.2. Univariate and multivariate Cox analysis relationship between clinicopathological factors and risk scores associated with OS across six cancers in the validation cohort (GEO) for prognostic EMT and DDR-related signatures. 45

Table 4.3. Clinicopathological characteristics of low- and high-risk groups across six cancers in training cohort (TCGA)..... 56

Table 4.4. Clinicopathological characteristics of low- and high-risk groups across six cancers in validation cohorts (GEO). 59

Table 4.5. C-index of signature-based nomogram models for all six cancers in training and validation cohorts. 80

LIST OF ABBREVIATIONS

AIC: Akaike Information Criterion

AUC: Area Under the Curve

BRCA: Breast Invasive Carcinoma

CASP9: Caspase-9

CCND2: Cyclin D2

CIBERSORT: Cell-type Identification by Estimating Relative Subsets of RNA Transcripts

C-index: Concordance index

COAD: Colon Adenocarcinoma

DCA: Decision Curve Analysis

DDB2: Damage-Specific DNA Binding Protein 2

DDR: DNA Damage Repair

DEGs: Differentially Expressed Gene Analysis

ESTIMATE: Estimation of Stromal and Immune cells in Malignant Tumour tissues using Expression data

ECM: Extracellular Matrix

EMT: Epithelial-Mesenchymal Transition

FDR: False Discovery Rate

GO-BP: Gene Ontology Biological Process

GEO: Gene Expression Omnibus

KEGG: Kyoto Encyclopaedia of Genes and Genomes

LASSO: Least Absolute Shrinkage and Selection Operator

LUAD: Lung Adenocarcinoma

LUSC: Lung Squamous Cell Carcinoma

MSigDB: Molecular Signature Database

OS: Overall Survival

PAAD: Pancreatic Adenocarcinoma

PCA: Principal Component Analysis

ROC: Receiver Operating Characteristic

RSEM: RNA-Seq by Expectation Maximization

ssGSEA: single-sample Gene Set Enrichment Analysis

STAD: Stomach Adenocarcinoma

TP53AIP1: Tumour Protein p53-Regulated Apoptosis-Inducing Protein 1

t-SNE: t-Distributed Stochastic Neighbour Embedding

UCSC: University of California Santa Cruz

CHAPTER 1

INTRODUCTION

1.1. Global Burden of Cancer

Despite the substantial progress in the past decade, the inherent complexity and heterogeneous nature of cancer, together with tumour evolution, continue to impose the discovery of novel therapies and durable treatments in most patients (Arafeh et al., 2025). The latest report of Cancer Today, France (https://gco.iarc.who.int/today) revealed approximately 19.9 M cancer cases (ASR, 196.9) with total mortality of 9.74 M, globally in 2022, where Europe has the second highest rate of total incidence (22.4%), mortality (20.4%), and 5-year prevalence (25.5%), after Asia. Among them, lung cancer tops the table with approximately 2.48 M (12.42%) of cases and 1.81 M (18.65%) of deaths, while breast cancer is diagnosed in 2.29 M (11.50%) people with a total mortality of 0.66 M (6.84%) globally. Meanwhile, colorectal, stomach, and pancreatic cancer were diagnosed in 1.92 M (9.64%), 0.96 M (4.85%), and 0.51 M (2.56%) people with total mortality of 0.90 M (9.28%), 0.66 M (6.8%), and 0.46 M (4.8%), respectively, worldwide. Lung, breast, colorectal, stomach, and pancreatic cancer possess the cumulative risk of mortality of 2.00, 1.40, 0.84, 0.69, and 0.47, respectively (Ferlay J, 2024). Therefore, cancer remains a significant global health burden, due to its wide-range clinical presentation, complex biology, and intricate pathogenesis, presenting tremendous challenges to both prevention and effective treatment (Xiong et al., 2025).

As the incidence rates of cancer are gradually continuing to rise globally, it is essential to deepen our understanding of the nature of cancer's diverse molecular and pathological fundamentals in order to develop more effective and targeted therapeutic strategies (Siegel et al., 2024). Hence, the identification of biomarkers to guide the selection of the best drug treatment regimens remains a critical step in cancer research.

1.2. The Clinical Challenge: Limitations of Current Prognostic and Predictive Biomarkers

In the practice of modern oncology, biomarkers are defined as measurable indicators of a biological state or condition. In the entire spectrum of cancer care, biomarkers are served as a critical function such as early detection, risk stratification, prognosis, and prediction of response to specific therapeutic interventions (He & Jia, 2021). The integration of validated biomarkers into clinical practice has modernized patient management and has been a foundation of the precision medicine era, including the hormone receptors like oestrogen receptor (ER) and progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) in breast cancer, guide the use of endocrine and anti-HER2 targeted therapies, respectively (He & Jia, 2021). Similarly, the identification mutations in the epidermal growth factor receptor (EGFR) gene and translocations in the anaplastic lymphoma kinase (ALK) gene in non-small cell lung cancer (NSCLC) has led to the development of highly effective tyrosine kinase inhibitors (TKIs) (Patel et al., 2015). Furthermore, in colorectal cancer and other solid tumours, microsatellite instability (MSI) serves as a powerful predictive biomarker for response to immune checkpoint inhibitors (ICIs) (de Assis et al., 2022).

Despite these remarkable achievements, the current biomarker landscape in cancer treatment is constrained by several limitations. Research showed that many patients whose tumours harbour a targetable alteration and initially significantly respond to a biomarker-guided therapy eventually experience disease progression due to the emergence of acquired resistance mechanisms (Wu et al., 2025). For instance, a sub-population of breast cancer patients with HER2-positive did not showed the significant benefits from HER2-targeted agents like trastuzumab, despite the presence of target biomarkers (He & Jia, 2021). However, there are numerous practical and logistic barriers which hinder the clinical translation of biomarkers into everyday practice. Furthermore, the lack of standardized, clinically validated cut-off values for many emerging biomarkers continues to hinder their consistent and reliable application in patient care.

These limitations emphasize a continuous and urgent need for the discovery and validation of novel biomarkers with improved prognostic and predictive abilities (He & Jia, 2021). The

shortcoming of single-analyte biomarkers that do not typically capture the complexity of tumour biology has produced multi-gene expression signatures. Such signatures, derived from transcriptomic data, provide the ability to show how multiple genes and biological pathways are functioning, as well as a stronger, more complete and more reliable representation of tumour behaviour and prognosis than that of one marker (Jones et al., 2020). The transition is critical to the generation of novel, multi-gene prognostic signatures that better describe the complexity of cancer biology.

1.3. Overview of Six Cancer Cohorts

The challenges of tumour heterogeneity and deficiencies of available biomarkers are not just abstract concepts but are put in use in daily clinical management through diagnosis and treatment of cancer patients in medical practice. This thesis focuses on six common and clinically relevant malignancies like breast cancer (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), stomach adenocarcinoma (STAD), pancreatic adenocarcinoma (PAAD), and colon adenocarcinoma (COAD). All cancers have unique bio-clinical and treatment profiles but share numerous important challenges including therapeutic resistance, metastatic progression, and the inadequacy of current patient stratification approaches. Across these six cancer cohorts, we benchmark their biomarker landscape and therapeutic challenges and explore common unmet clinical needs that are likely to benefit from the implementation of an engineered, pan-cancer prognostic model to improve prediction, and personalize treatment decisions.

1.3.1. Breast Invasive Carcinoma (BRCA)

Breast cancer is the most prevalent noncutaneous cancer among women and although there have been significant advances in treatment that have improved outcomes for many patients, the mortality of breast cancer patients has been increasing in recent years. Current therapeutic options are dominated by molecular subtypes based mainly on the expression of ER, PR, and HER2 (He & Jia, 2021). This subtype-based approach has facilitated large-scale therapeutic achievements, such as the use of endocrine therapy for HR+ tumours, and the use of HER2-targeted therapy for HER2-positive tumours. Despite these achievements, substantial challenges remain unclear. Triple-negative breast cancer (TNBC) has remained the most

aggressive and therapeutically challenging subtype that has low therapeutic efficacy as a result of lack of expression of all three standard breast cancer biomarkers (Sukumar et al., 2021). Recent advances including immunotherapy (e.g., pembrolizumab) for PD-L1–positive TNBC and PARP inhibitors for germline BRCA1/2 mutant patients have opened new therapeutic opportunities, but their application is restricted to certain patient subgroups (Sukumar et al., 2021). Intrinsic and acquired drug resistance have emerged as a significant clinical hurdle to the long-term disease control spanning all molecular subtypes and is a major driver of both recurrence and mortality rates of disease.

1.3.2. Lung Adenocarcinoma (LUAD)

Lung adenocarcinoma (LUAD) is the most prevalent subtype of lung cancer and remains a leading cause of cancer-related deaths worldwide (Sung et al., 2021). Over the past two decades, the development of molecular profiling and targeted therapy has transformed treatment options for advanced LUAD. The identification of key oncogenic driver mutations, like mutations in the EGFR gene or rearrangements in ALK and ROS1 led to the development of a new class of drugs called tyrosine kinase inhibitors (TKIs) that induce rapid and durable tumour response which significantly extended the patient survival compared to conventional chemotherapy (Solomon et al., 2014). Despite these important advances, most patients eventually develop acquired resistance to TKIs that typically take on within one to two years of treatment, leads to tumour progression (Oxnard et al., 2011). Although the catalogue of targetable genetic alterations in LUAD continues to expand such as KRAS G12C, MET, RET, and NTRK mutations but a substantial proportion of patients still lack identifiable driver mutations (Skoulidis & Heymach, 2019). The limited treatment options available for these patients are chemotherapy, immunotherapy, or their combinations, which often provide limited and inconsistent efficacy. LUAD clearly represents the biological complexity of modern oncology determined by tumour heterogeneity, the evolution of resistant clones under therapeutic pressure, and dynamic tumour microenvironment interactions that influence response to therapy as well as the emergence of resistance (Herbst et al., 2018). Therefore, there is a critical and ongoing need for integrative, multi-omics novel biomarkers that can predict prognosis more accurately and guide therapeutic decisions beyond the initial driver oncogene, particularly in the context of resistance and for patients eligible for immunotherapy.

1.3.3. Lung Squamous Cell Carcinoma (LUSC)

Lung Squamous Cell Carcinoma (LUSC), representing 25 to 30% of NSCLC cases, remains a major clinical problem owing both to its aggressive nature and restricted therapeutics (Network, 2012). Unlike LUAD, which can benefit from several clinical targeted therapies, LUSC does not have well-defined oncogenic driver mutations that could be targeted and be the keystone to precision treatment (Herbst et al., 2018). Most patients with advanced LUSC are thus treated with platinum-based chemotherapy and immune checkpoint inhibitors (ICIs) that target the PD-1/PD-L1 axis (Reck et al., 2016). While some patients benefitted from immunotherapy, only limited responses have been observed, with the majority of patients either unresponsive or developing acquired resistance. Due to limited effective targeted therapies, and a need to better stratify patients for existing treatments, the identification of robust prognostic and predictive biomarkers remains a major research priority in LUSC (Gandhi et al., 2018).

Molecular studies in recent years have begun to surface biomarkers and may help to improve patient stratification and therapeutic outcomes. Dysregulation of genes central to cell-cycle arrest, DNA repair, and immune signalling were noted using transcriptomic and proteogenomic analyses (Campbell et al., 2016). Among recently reported candidates, the reduced FAM20A expression is associated with worse prognosis and might control tumour immune evasion pathways (Zhang et al., 2024). Mutation-based multigene prognostic predictability signatures with coagulation-related genes also have been reported to be associated with immune-cell infiltration, tumour invasion, and patient survival (Wu et al., 2023). Non-coding RNAs such as MIR205HG and SERPINB5 have also been considered as molecular surrogate markers of disease progression and potential drugs (Chen et al., 2021). However, the translation of biomarkers to clinical practice is still lacking. The majority of proposed markers are also not externally validated, and the very high level of intra-tumoral heterogeneity in LUSC makes consistent application across patients a difficult task (Network, 2012). Hence, efforts continue to integrate genomic, transcriptomic, and immunologic data to create clinically actionable biomarker panels that predict response, track resistance, and inform personalized therapy in LUSC.

1.3.4. Stomach Adenocarcinoma (STAD)

Stomach adenocarcinoma (STAD), the most common histologic type of gastric cancer, continues to be one of the major causes of cancer death globally. It is characterized by high molecular heterogeneity and generally associated with a poor prognosis (Smyth et al., 2020). In recent years, the incorporation of molecular biomarkers has introduced a degree of precision into the management of advanced STAD. Targeted therapies with trastuzumab are used for tumours exhibiting HER2 amplification, while immunotherapies such as pembrolizumab and nivolumab are showed for tumours expressing PD-L1 or demonstrating high microsatellite instability (MSI-H) (Janjigian et al., 2021; Yj, 2010). More recently, the monoclonal antibody like zolbetuximab, has been approved for tumours expressing Claudin-18.2 (CLDN18.2), has emerged as an additional therapeutic option for a subset of biomarker-positive patients (Kubota & Shitara, 2024).

Notwithstanding these advances, the clinical utility of targeted agents remains limited to a relatively small number of patients with tumours that exhibit these specific molecular characteristics. The high inter- and intratumoral heterogeneity associated with STAD, however, presents a continuing challenge which has led to variable treatment responses as well as the frequent development of acquired drug resistance (Network, 2014). Most patients continue to be treated with systemic chemotherapy, which highlights a persistent need for discovery of new broadly applicable biomarkers that can help optimize patient stratification and personalized treatment selection.

1.3.5. Pancreatic Adenocarcinoma (PAAD)

Globally, pancreatic ductal adenocarcinoma (PAAD) is considered one of the most lethal and aggressive human malignancies, accounting for more than 90% of all pancreatic cancers. Research has demonstrated that the global incidence of PAAD is projected to rise, making it the second-leading cause of cancer-related death by 2030, which highlights the urgent need for improved detection and management (Sok et al., 2023). Research revealed that its poor prognosis can be associated with several factors with five-year overall survival of approximately 13%. In an early stage, PAAD is usually considered as asymptomatic and in more than 80% of patients it is diagnosed at an advanced or metastatic stage (Kleeff et al.,

2016). However, the considerable heterogeneity and genomic instability of PAAD with a dense fibrotic stroma made it chemo- and radiotherapy resistant.

According to accumulative data, cytotoxic chemotherapy has been considered as one of the most reliable therapies for decades but exhibits minimal survival improvement against PAAD. A serum biomarker such as carbohydrate antigen 19-9 (CA 19-9) is known as commonly used biomarker for diagnosis but it lack sensitivity and specificity for each detection of PAAD (Sok et al., 2023). A significant step towards precision medicines in PAAD has been found that a subset of patients that harbours germline mutations in *BRCA1* or *BRCA2* have tumours susceptibility to poly (ADP-ribose) polymerase (PARP) inhibitors such as Olaparib (Golan et al., 2019). However, the discovery of robust biomarkers for early detection, therapeutic stratification, and prediction of treatment response are remained as crucial unmet clinical need for effective therapies (Golan et al., 2019).

1.3.6. Colon Adenocarcinoma (COAD)

Globally, colon adenocarcinoma (COAD), one of the most common pathological types of colorectal cancer (CRC) is considered as major cause of cancer-related mortality at advanced metastasis stage (Sung et al., 2021). Recent developments in molecular oncology have deepened our insight for COAD, that reveal it to be a disease driven by a complex interaction of genetic, epigenetic and environmental factors (Dienstmann et al., 2017). These insights have led us to better understand the identification of key molecular biomarkers essential for prognosis and treatment particularly in patients with metastatic stages. The clinically identified biomarkers of COAD include mutations in *KRAS*, *NRAS*, and *BRAF*, as well as MSI and mismatch repair (MMR) deficiency (Guinney et al., 2015; Vaughn et al., 2011). Importantly, this type of mutation has strong clinical relevance. Mutations in KRAS or NRAS are critical to clinical decisions; they predict resistance to anti-EGFR therapies, such as cetuximab and panitumumab, and so RAS genotyping is foundational to molecular testing services as well as patient care. In tumours that harbour the *BRAF V600E* mutation, on the contrary, tumours with a poor prognosis might respond to effective combinations composed of BRAF, EGFR and MEK pathway therapies (Kopetz et al., 2019). Furthermore, approximately 15% of COAD patients have high MSI-H/dMMR, which is relevant to the higher mutation rate and high

immunogenicity of the tumour. These tumours are typically responsive to immune checkpoint inhibitors, including pembrolizumab and nivolumab (Le et al., 2015).

Apart from these established biomarkers, we also have a body of emerging research that identifies novel multi-omics and transcriptomic signatures that may allow us to refine patient classification and enhance targeted treatment strategies. The most significant framework for this is provided by the Consensus Molecular Subtypes (CMS) classification, which classifies COAD into four major biological subtypes including CMS1 (MSI-immune) with strong immune activation, CMS2 (canonical) with WNT and MYC pathway activation, CMS3 (metabolic) with metabolic dysregulation, and CMS4 (mesenchymal) with stromal infiltration and poor prognosis (Guinney et al., 2015). The molecular characterization of tumours contributes to a more comprehensive understanding of tumour biology and, therefore, to the development and prognosis of tumours and to informed clinical practice. Concurrently, several new researchers are studying non-invasive biomarkers such as circulating cell-free DNA (cfDNA), microRNAs and exosomal proteins as tools for early diagnostics, follow-up of disease and evaluation of clinical response in real-time (Diehl et al., 2008). However, in colorectal cancer management, these developments signify a transition to more integrated and dynamic models.

1.4. Epithelial-Mesenchymal Transition (EMT) as a Driver of Cancer Progression

Epithelial-mesenchymal transition (EMT) is regarded as a central process of cancer progression by allowing tumour cells to gain motility, invasiveness and resistance against therapy (Yang et al., 2020). At the molecular level, the main feature of EMT is loss of epithelial markers such as E-cadherin (*CDH1*) while induction of mesenchymal markers like N-cadherin (*CDH2*), vimentin (*VIM*), and fibronectin is observed (see Figure 1.1). Such markers were essential for maintaining the stability of epithelial cells and also cell migration and invasion (Tripathi et al., 2023). Nevertheless, EMT transcriptional factors (EMT-TFs), on the basis of E-cadherin repressors can be distinguished as direct or indirect. Direct repressors include zinc-finger proteins from the SNAIL superfamily, such as SNAI1 (SNAIL), SNAI2 (SLUG), SNAI3 (SMUC); zinc finger and E-box binding ZEB proteins ZEB1 (TCF8) and ZEB2 (SIP1); bHLH

factor E47; and Kruppel-like factor KLF8. Indirect repressors consist of TWIST bHLH proteins (TWIST1 and TWIST2), homeobox protein goosecoid (GSC), SIX1, bHLH factor E2.2, and forkhead box protein FOXC2, which inhibit E-cadherin expression through indirect transcriptional mechanisms (Puisieux et al., 2014; Thiery et al., 2009).

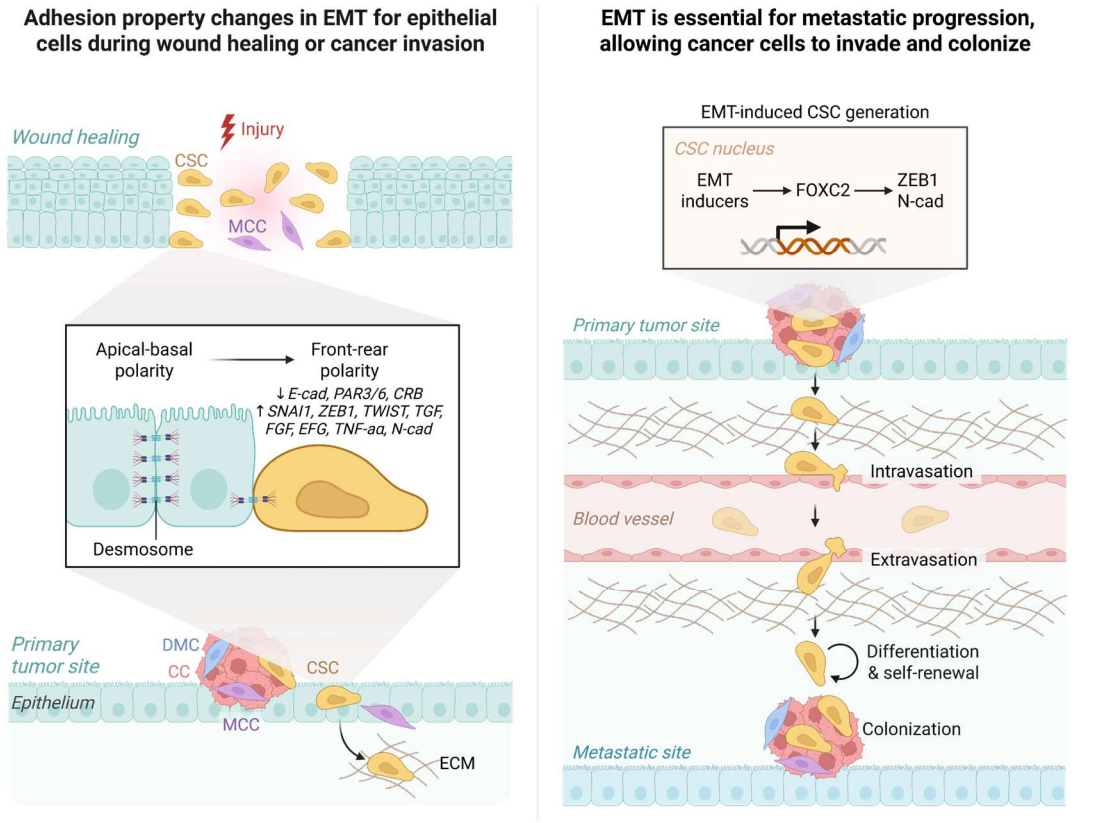


Figure 1.1. Role of epithelial to epithelial–mesenchymal transition (EMT) in wound heading and cancer progression. This image was created with BioRender (www.biorender.com) (Den Hollander et al., 2024).

These EMT-TFs act as transcriptional repressors, which directly bind to the promotor regions of epithelial genes such as *CDH1* (encode E-cadherin) to suppress the expression of epithelial markers and activate the expression of mesenchymal genes. Subsequently, the activation of these EMT-TFs in turn drives by extracellular signalling pathways which are initiated by

tumour microenvironment (TME) (Georgakopoulos-Soares et al., 2020). Among them, Transforming Growth Factor-β (TGF-β) signalling pathway acts as one of the most well-characterized pathways to induce EMT, followed by several other key pathways that played critical role in triggering the EMT programs are Wnt/β-catenin, Notch, and those stimulated by growth factors such as Hepatocyte Growth Factor (HGF) and Epidermal Growth Factor (EGF) (Srinivasan et al., 2025).

EMT is not a process of binary switch but rather proceeds through a spectrum of partial traits. In the partial state, cells exhibit special features of epithelial and mesenchymal transition, displaying markers of both epithelial and mesenchymal cells. Cancer cells can exhibit EMT to different stages and through mesenchymal-epithelial transition (MET), it may also reverse toward epithelial phenotype (Meyer-Schaller et al., 2019). Nonetheless, the mobility of cells associated with EMT stages allows tumour cells to undergo different microenvironmental transformations during local growth, systemic distribution, colonisation of distant tissues, or in response to therapy (Lüönd et al., 2021). Importantly, the role of EMT extends beyond invasion to include the acquisition of stem cell properties and therapy-associated resistance. Recent insights reveal that this resistance is not merely a passive consequence but is actively driven by EMT-associated transcription factors (EMT-TFs), such as ZEB1, which play a direct role in regulating the DNA damage response (DDR) and repair pathways. For instance, ZEB1 has been shown to be phosphorylated by ATM in response to DNA damage, promoting homologous recombination (HR)-dependent repair and contributing to radio resistance (Moyret-Lalle et al., 2022). However, despite the importance of EMT in living organisms in cancer evolution, the accurate identification of cells that undergo EMT still remains a significant challenge (Meyer-Schaller et al., 2019). Although the role of full EMT tumour cells in drug resistance has been established, the relevance of EMT that drives the tumour to spread still remains a subject of ongoing scientific debate (Lüönd et al., 2021).

1.4.1. EMT-Related Gene Signatures as Prognostic Tools

A robust association between EMT activation and aggressive clinical outcomes has inspired extensive research efforts for the development of EMT-related gene signature as prognostic and predictive biomarkers. According to accumulative data, several studies have constructed EMT-related signatures from transcriptomic data of tumour cohorts to stratify patients into high- and

low-risk groups for outcomes in range of cancer types including BRCA, COAD, and LUAD (Cui et al., 2022). The general approach for this type of analysis involves the identification of set of genes whose expression profiles are closely associated with EMT phenotype and effectively linked with patient survival in training cohorts (Han et al., 2025). Subsequently, based on weighted expression of these signature genes a risk score is computed, which is then tested for its prognostic ability in independent validation cohorts.

However, several methodological challenges have emerged that call into question the clinical validity and interpretation of many existing EMT signatures. Most of those signatures have been obtained from bulk tumour tissue, which is known as a heterogeneous mixture of cancer cells and various non-cancerous stromal and immune cells that establish the TME. Many recent comprehensive analyses have revealed that expression scores of EMT, mesenchymal and stemness signatures are largely determined by the stromal cells within the tumour patients especially cancer-associated fibroblasts (CAFs) (Kreis et al., 2023). Naturally, fibroblasts are mesenchymal cells, so a high EMT score in bulk tumour samples may associated with a high stromal content and have no impact on the actual cell-intrinsic EMT programs in the cancer cells. However, this stromal contradictory factor represents a critical challenge in biomarkers development. It poses a significant risk of false conclusions, that appears to measure a cell-specific cancer biological process which may act as surrogate for tumour purity. So, as a result, the clinical potential and biological interpretation of such signatures become uncertain (Kreis et al., 2023). However, for the discovery of prognostic biomarkers to be truly and mechanistically useful, it must accurately represent the biology of cancer cells. To overcome this limitation, there is critical need to develop novel EMT-related gene signatures that can separate cancer-specific signals from those influenced by the tumour stroma.

1.5. DNA Damage Response (DDR): From Genome Protection to Cancer Progression

The DNA damage response (DDR) is a highly dynamic network of proteins that coordinates in such a way as to help manage both DNA damage repair and cell-cycle checkpoint activation, thereby preventing the replication or transmission of damaged genetic material to the daughter cells. DDR is undoubtedly a valuable system for cellular and organism survival that acts as a

safeguard from both intrinsic metabolic processes and environmental sources (Drew et al., 2025). Damage sensors, kinase-based signalling (ATM, ATR, and DNA-PKcs), downstream checkpoint effectors such as CHK1 and CHK2, and numerous repair pathways including base excision repair, nucleotide excision repair, mismatch repair, non-homologous end joining (NHEJ), and homologous recombination (HR) are integrated by DDR (Groelly et al., 2023). The detail about the targeting of DDR is presented in Figure 1.2. DDR acts as an anti-tumour barrier in normal tissues, as it prevents the replications of heavily damaged DNA, induce cell-cycle arrest and apoptosis when lesions are irreparable. However, in established cancers, the frequency of DDR is altered. It was noted that in high-fidelity repairs pathways like HR, the tumour cell often carries defects but are still dependent on alternate repair pathways and checkpoint signalling to overcome the persistent replication stress. However, this selective dependency presents a therapeutic opportunity, which could help the oncologist to disable specific DDR nodes to push the progression of genomically unstable cancer cells past their survivable threshold (Drew et al., 2025; Groelly et al., 2023).

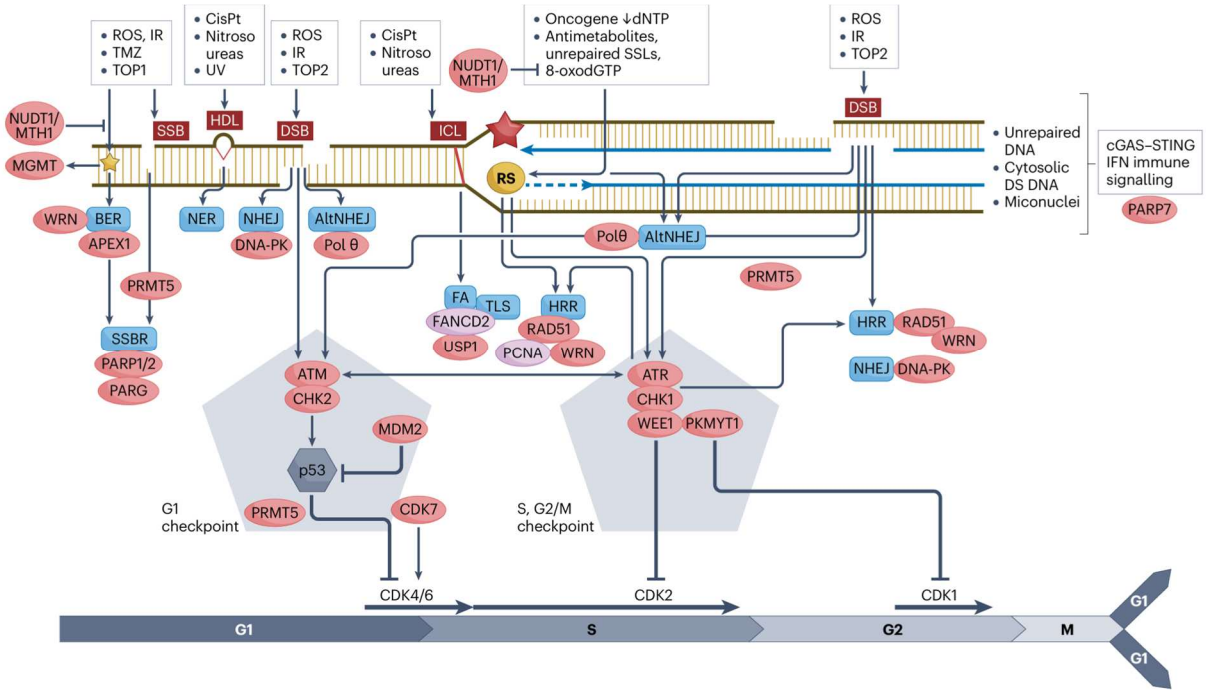


Figure 1.2. Overview of targeting the DDR pathways and cell-cycle checkpoints. Overview of targeting the DDR pathways and cell-cycle checkpoints. Different DNA repair mechanisms are activated according to the type of DNA damaging agents such as base excision repair (BER) for base damage and single-strand breaks, nucleotide excision repair (NER) for bulky adducts,

homologous recombination repair (HRR) and non-homologous end joining (NHEJ) for double-strand breaks (DSBs), and Fanconi anaemia (FA) and translation synthesis (TLS) pathways for interstrand crosslinks (ICLs). This image is taken from (Drew et al., 2025).

However, any defect in the DDR system could lead to genomic instability as a hallmark of cancer; several DDR factors, such as tumour suppressor genes, contributed to cancer progression, including the most notable *BRCA1/BRCA2* and *TP53* genes (Bieging et al., 2014; Hanahan & Weinberg, 2011). In addition to tumour suppressors, the upregulated oncogenes could contribute to compromising the DDR system by pushing cells too early into S phase, which leads to abnormal DNA structure and replication stress. These functions are all required by DDR to restore the DNA (Drew et al., 2025). Hence, DNA repair deficiencies are strongly associated with cancer initiation and progression. However, these same deficiencies represent a critical vulnerability, forming the basis for synthetic lethality approaches in cancer treatment, such as PARP inhibition in BRCA-deficient tumours (Amodio et al., 2024).

1.5.1. DDR-Related Signature as Prognostic and Predictive Biomarkers

Predictive biomarkers are particularly defined as the identification of tumours that may respond to a specific therapy. Predictive biomarkers could be genomic, transcriptomic or functional, and they are likely the hardest of all to measure. Using PARPis as examples which serve as genomic biomarkers (*BRCA1/BRCA2* mutations), these biomarkers likely to be the most consistent, although other genetic changes can restore HRR activity in BRCA-mutated cancers, leading to resistance to PARPi (Noordermeer & van Attikum, 2019). On the other hand, whole-genome sequencing can be used to find alterations in DNA structure caused by HRR-deficient (HRD) or base-substitution signatures, such as COSMIC signature 3. But these signatures indicate genomic ‘scarring’; in other words, they suggest this tumour had HRD at some point in its history but may now be HRR-proficient. Functional assays performed on tumour sample collected in the vicinity of the treatment decision are likely to be the best markers of HRD status for patient stratification, but they are limited to the samples that can be collected at appropriate time 201. *RAD51* foci reflects HRR activity, and the first finding revealed that approximately 60% of ovarian cancers were HRD demonstrated ex vivo *RAD51* focus formation in primary cultures of malignant ascites (Miller et al., 2020). In more recent time, this approach has been

extended to measure *RAD51* foci in formalin-fixed, paraffin-embedded blocks of tumour material (Drew et al., 2025).

DDR pathway status has also become an important clinical predictive biomarker for various cancer treatments. Tumours with DDR deficiencies, which are considered incapable of efficiently repairing the DNA damage caused by platinum-based chemotherapy or radiotherapy, respond more to these treatments (J. Dai et al., 2021). Moreover, the relationship between DDR defects, genomic instability, and high TMB in a tumour can influence immunotherapy on a large scale. The accumulation of mutations in DDR-deficient tumours could drive the production of more neoantigens, which act as a novel protein fragment that can be recognized as foreign by the immune system. This could ultimately enhance immunogenicity that might declared tumours as 'hot,' like they have a greater infiltration of T cells, which leads to greater immune checkpoint inhibition responses.

However, due to the central role of DDR in determining prognosis and therapeutic response, intense research has been conducted to develop DDR-associated gene signatures. The main functionality of these DDR signatures is to predict patient outcomes and guide the selection of appropriate therapies by measuring the expression levels of panels of key DDR genes (Chang et al., 2023). The concept of synthetic lethality and a link between DDR status and immunogenicity are not restricted to a single cancer but are fundamental aspects of cancer biology. This suggests that a well-defined DDR-related gene signature could become a promising pan-cancer biomarker that would be able to identify specific therapeutic targets in various histological contexts and offer strong rationale for a pan-cancer approach.

1.6. Molecular Crosstalk Between EMT and DDR

Recently, a pan-cancer analysis study reported an interesting relationship between EMT and DDR pathways across multiple solid tumours. Ding et al. (Ding et al., 2022) determined the tumour-related DDR score (TR-DDR) based on the expression level of 20 core genes in pan-cancers and reported that the TR-DDR score was significantly associated with genome instability, cell proliferation, and hallmark processes, including EMT and TGF-β signalling. The results revealed that the TR-DDR score association not only enhances stemness and decreases p53/apoptosis signature but also links with EMT-related programs across cancers. A

recent study demonstrated the upregulation of ITGB1 along with DDR family proteins associated with EMT activation and therapeutic resistance in EGFR-mutant non-small cell lung cancer (NSCLC), which suggested the integrin-mediated DDR-EMT crosstalk promotes tumour survival in lung cancer (De Rosa et al., 2024). ITGB1 activation in intracellular signalling is linked to the upregulation of the EMT program and MAPK pathways (Renner et al., 2016). Furthermore, in pancreatic ductal adenocarcinoma (PAPD), ATM deficiency was observed to accelerate EMT via disrupted BMP4/TGF-β signalling, which highlighted its corroborative role as a shared modulator of EMT and DDR across cancers (Russell et al., 2015). Chowdhury et al. (Chowdhury et al., 2024) performed a comprehensive pan-cancer analysis focusing on Kindlin family proteins, reporting that alterations in FERMT1-3 were positively correlated with EMT hallmarks and modulation of DDR pathways across cancers.

A study revealed that during cell cycle, the role of *BRCA1* in HR-mediated DNA repair involves the interplay of CtIP-dependent end processing and help to trigger the mechanism of RAD51-mediated homologous pairing (see Figure 1.3) (Zhao et al., 2017). The components of Fanconi anaemia (FA) pathway, including FANCD2, and mutation of *BRCA1* and *BRCA2*, significantly protected the lapsed forks at DNA replication blocking lesions during DNA replication process (Heitmeir et al., 2022). While in contrast, *BRCA1* actively antagonizes EMT as both direct and indirect. Consequently, by direct promoter binding mechanism, BRCA1 was reported to mediate an inhibitory effect on FOXC1/2 and TWIST1 (known as another EMT-inducing transcription factor downstream of SNAIL and SLUG) (Sengodan et al., 2018). A study reported a fascinating discovery that BRCA1 repairs crosslinks in complexes with SWI/SNF chromatin remodelling factor BRG1, the differentiating maintenance factor NUMB, FANCD2 and NOTCH target gene product HES1 (Wang et al., 2016). However, any failure in the replication lead to DNA damage and upregulate the EMT factors which triggers the luminal to basal and mesenchymal transdifferentiation of mammary epithelial cells (BeLow & Osipo, 2020). The accumulative data revealed that the persistent DNA replication stress could be a denominator of EMT-induction, as the EMT-inducing chemical treatments do generate DNA replication blocking lesions, such as crosslinks (Haynes et al., 2015; Heitmeir et al., 2022).

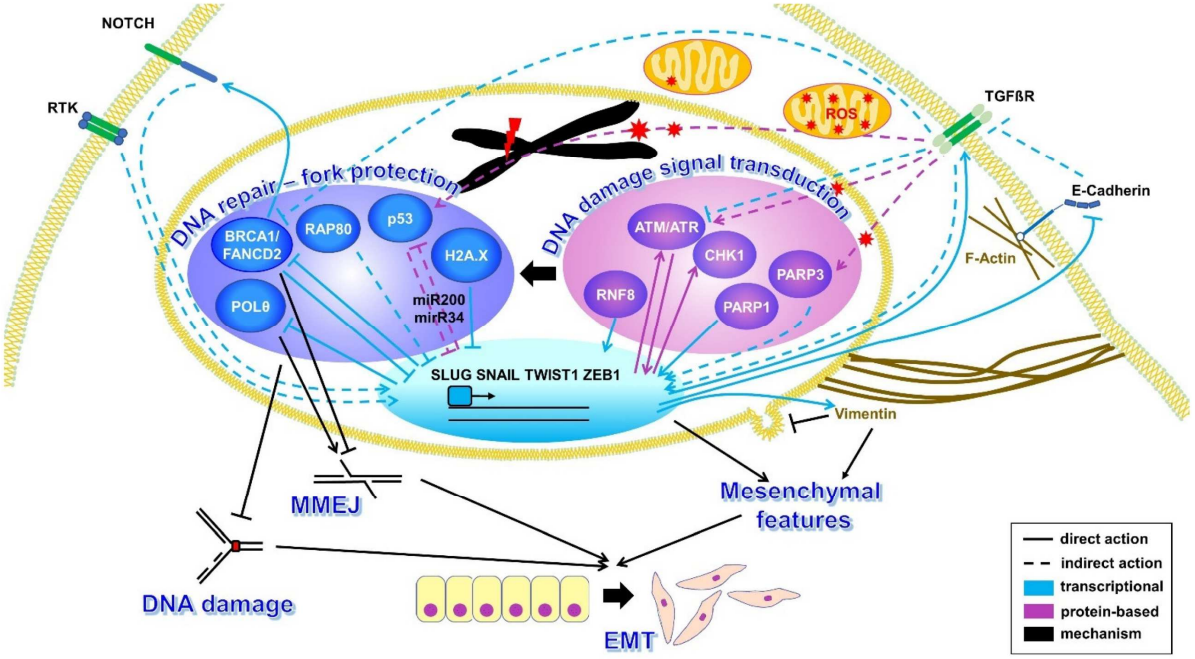


Figure 1.3. The molecular crosstalk between EMT and DDR in cancer progression. This image is taken from (Heitmeir et al., 2022).

Another study reported that in BRCA cells, the DDR of HR pathways was promoted by binding of ZEB1 with USP7 to stabilize CHK1 (Zhang et al., 2014). Similarly, it was observed that during the regulation of DDR pathways, RNF8 triggers the nuclear localization of TWIST (H.-J. Lee et al., 2016). However, by blocking the rearrangement of the early EMT cytoskeleton, cisplatin significantly inhibits BRCA proliferation to benefit patient's prognosis (P. Zhang et al., 2023). Likewise, the metastasis of BRCA cell was significantly restrained by the inhibition of EMT-related pathways with JAM2 to improve prognosis of patients (Peng et al., 2022). Overall, the results showed that DDR and EMT not only related to the prognosis of cancer patients but also interconnected with each other. There is an urgent need to develop a dual EMT/DDR-related gene signatures to fill the research gap and find the novel prognostic biomarkers in solid tumours.

CHAPTER 2

RESEARCH RATIONALE, NOVELTY, AND OBJECTIVES

2.1. Rationale and Novelty of the Study

Globally, cancer has remained as a key barrier to increase life expectancy and one of the leading causes of death. The epidemiological data revealed that the global burden of cancer will increase to approximately 28.4 million cases by 2040 (almost 47% rise compared to 2020) (Sung et al., 2021). Among the molecular mechanisms driving cancer progression, DNA damage response (DDR) and epithelial-mesenchymal transition (EMT) are recognized as key drivers of tumour aggressiveness and resistance to therapies. The functional crosstalk between DDR and EMT demonstrated that these pathways not only modulate DDR gene expression and repair capacity, but that chronic DNA damage can also promote EMT activation and metastatic progression. However, there is a critical need to develop integrated, multi-cancer biomarkers models to fill the research gap.

The development of therapeutic strategies tailored to the genetic and molecular profile of cancer will be essential for advancing cancer therapy. Extensive efforts have been made over the past few decades to illuminate a broad spectrum of genetic and molecular changes that disrupt several regular cellular programs and enable the acquisition and maintenance of malignant phenotypes (Hanahan, 2022). These oncogenic disruptions not only drive tumorigenesis but also help in creating the cancer-specific vulnerabilities or dependencies for cancer, which represent molecular features like genes or proteins that cancer cells critically rely upon for survival and proliferation (Arafeh et al., 2025; Boehm et al., 2021). In order to find potential molecular vulnerabilities in breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), and stomach adenocarcinoma (STAD), we selected two biologically relevant gene sets, EMT and DDR, to systematically identifying novel prognostic genes across all six cancer types. To the best of our knowledge, we conducted a comprehensive study for the first time to identify novel dual prognostic EMT- and DDR-related signatures across six cancers to understand cancer progression and therapy.

The novelty of thesis lies not only in investigating EMT- and DDR-related genes separately, but in integrating them into dual-signature, multi-cancer comparative models. Through this integrative strategy, we aim to enhance understanding of cancer progression mechanisms and improve the identification of robust biomarkers to guide prognosis and therapy selection.

2.2. Research Hypotheses

1. Distinct yet interrelated EMT- and DDR-related gene signatures across six cancers−BRCA, COAD, LUAD, LUSC, PAAD and STAD−can be used to develop prognostic models which can stratify patients according to overall survival (OS) and cancer progression.
2. A dual-signature prognostic model integrating EMT and DDR in multi-cancer framework will demonstrate superior prognostic performance compared to single signature model. Furthermore, the computed risk scores will correlate with immune infiltration patterns and drug sensitivity that could provide a biological tool for patient stratification and treatment optimization in precision oncology.

2.3. Aim and Objectives

2.3.1. Aim of the Study

The aim of the study is to identify, construct and validate novel dual EMT- and DDR-related prognostic gene signatures across six major cancers through an integrative multi-omics and clinical data analysis approach. The present study moves beyond correlation to build dual prognostic gene signatures for both EMT and DDR across six solid cancers and explained a comparative, multi-cancer prognostic framework that integrates EMT and DDR signatures into clinically validated models which revealed their combined potential as biomarkers for patient stratification and treatment optimization. This approach will address a critical unmet clinical need in precision oncology by moving beyond single-pathway or single-cancer analyses to develop multi-cancer, mechanism-based prognostic biomarkers with potential clinical applicability.

2.3.2. Specific Objectives

- To identify differentially expressed EMT- and DDR-related genes across six cancer cohorts using RNA-sequencing and corresponding clinical data from The Cancer Genome Atlas (TCGA).
- To perform univariate Cox regression, Kaplan–Meier survival analysis, and least absolute shrinkage and selection operator (LASSO) regression to identify genes which significantly associated with overall survival and construct robust EMT- and DDR-related prognostic signatures for each cancer type.
- To validate the predictive accuracy and independence of the risk score of identified gene signatures using multivariate Cox regression, receiver operating characteristic (ROC) curves, area under the curve (AUC), and external datasets from the Gene Expression Omnibus (GEO).
- To perform functional enrichment analyses, including Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway analyses to elucidate the biological mechanisms underlying the identified prognostic signatures.
- To evaluate the association between EMT/DDR-based risk scores and tumour immune microenvironment characteristics using ESTIMATE and CIBERSORT.X algorithms, thereby assessing the immunological relevance of the prognostic models.
- To construct an integrated prognostic nomogram that combines molecular risk scores with clinical parameters to improve individualized survival prediction and facilitate clinical translation.
- To explore the potential relationship between EMT/DDR signature-derived risk scores and drug sensitivity profiles in cancer cell lines to identify therapeutic vulnerabilities and inform precision treatment strategies.

2.4. Expected Outcome

This study is expected to yield novel dual EMT–DDR prognostic gene signatures that can:

- Precisely stratify patients into risk subgroups with distinct survival outcomes.
- Represent tumour-intrinsic biology rather than stromal confounding.
- Correlate with immune landscape and therapeutic response.

- Establish a framework that can be applied to develop biomarker-based prognostic models across multiple cancer types.

Overall, these findings may contribute to the development of systematically updated clinically biomarkers that improve patient stratification and guide targeted therapy selection in precision oncology.

CHAPTER 3

MATERIALS AND METHODS

3.1. Data Collection

Gene expression RNAseq (polyA+ IlluminaHiSeq) data and corresponding survival data for TCGA Breast invasive carcinoma (BRCA, n = 1218), TCGA Colon adenocarcinoma (COAD, n = 329), TCGA Lung adenocarcinoma (LUAD, n = 576), TCGA Lung squamous cell carcinoma (LUSC, n = 553), TCGA Pancreatic adenocarcinoma (PAAD, n = 183), and TCGA Stomach adenocarcinoma (STAD, n = 417), were retrieved from the University of California Santa Cruz (UCSC) XENA database (https://tcga.xenahubs.net). The gene expression RNAseq data for BRCA, COAD, LUAD, LUSC, and PAAD were quantified using the RSEM (RNA-Seq by Expectation Maximization) method, then $\log_2$-transformed as $\log_2$ (norm_count + 1). At the same time, STAD was quantified using the RPKM (Reads Per Kilobase per Million mapped reads) method, then $\log_2$-transformed as $\log_2$ (RPKM + 1). For the validation purposes, we retrieved the microarray and corresponding clinical data of GSE20685 (conducted by platform of GPL570, Affymetrix Human Genome U133 Plus 2.0 Array), GSE39582 (platform of GPL570, Affymetrix Human Genome U133 Plus 2.0 Array), GSE50081 (platform of GPL570, Affymetrix Human Genome U133 Plus 2.0 Array), GSE30219 (platform of GPL570, Affymetrix Human Genome U133 Plus 2.0 Array), GSE79668 (conducted by platform of GPL11154, Illumina HiSeq 2000 (*Homo sapiens*), and GSE84433 (platform of GPL6947, Illumina HumanHT-12 V3.0 expression beadchip) from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo). Similarly, data of GDSC2 half-maximal inhibitory concentration (IC_{50}) of drugs for all six cancer cell lines and data of mRNA expression level for these cell lines were downloaded from the CCLE database (https://sites.broadinstitute.org/ccle)

The EMT and DDR hallmark gene sets were obtained from the Molecular Signature Database (MSigDB, https://www.gsea-msigdb.org/gsea/index.jsp) using the standard and systemic names HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION (M5930) and WP_DNA_DAMAGE_RESPONSE (M39339), respectively. The source species, *Homo sapiens*, was used for both gene sets, which contained 200 and 69 genes, respectively,

associated with specific biological pathways and processes (see Supplementary File 1, Table S1). RStudio software (version 2024.12.1, Build 563) was employed for data manipulation, analysis, and visualization. Different RStudio libraries were used for these purposes (see Supplementary File 1, Table S2). The study overview is presented in Figure 3.1.

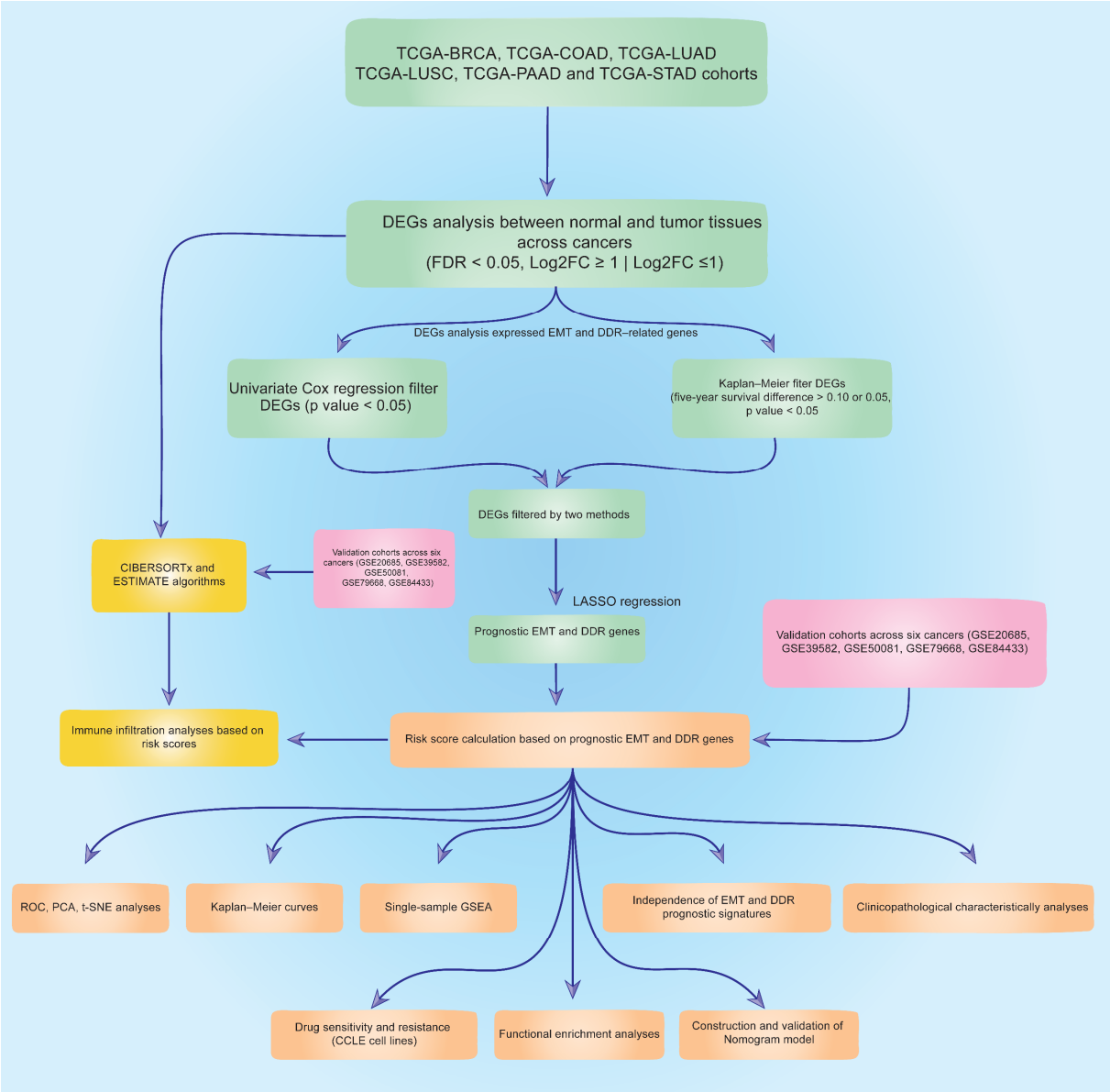


Figure 3.1. Schematic representation of the study workflow.

3.2. Differential Gene Expression (DGE) Screening

Differential Gene Expression Analysis (DGEA) was performed using the "limma" R package on the normalized gene expression matrices of all six cancers. Based on phenotype data, the expression metrics for all cancers were divided into primary tumours and adjacent normal tissues. The selection criteria for DGE were adjusted as $|\log_2\text{-fold change (FC)}| \geq 1$ and FDR-adjusted p value < 0.05. The results were visualized using a volcano plot via ggplot2 in R, which showed significant DGE by plotting $\log_2$ FC against adjusted p-values and highlighting upregulated and downregulated genes (Chi et al., 2022).

3.3. Construction and Validation of Novel EMT- and DDR-Related Prognostic Signatures

Two pre-established independent gene sets, EMT and DDR, served as candidate signature sets. Univariate Cox regression and Kaplan-Meier survival analysis were performed on tumour samples using overall survival (OS) as the dependent variable, along with survival status, against all TCGA cohorts to determine the prognostic value of EMT and DDR gene candidates. However, the significant threshold criteria to screen prognostic EMT and DDR-related DGE were applied as p-value < 0.05 for Univariate Cox analysis and five-year overall survival difference $\geq |0.10|$ for Kaplan-Meier for LUAD, LUSC, STAD, PAAD, and COAD. In contrast, for BRCA, the five-year overall survival difference was set as $\geq |0.05|$. Expression values were treated as continuous covariates, Wald P values, hazard ratio (HR), and 95% confidence intervals (CI) were extracted. Venn diagrams were plotted to visualize the overlapped genes between Cox and OS in all six cohorts for EMT and DDR-related genes using the "VennDiagram" R package. At the same time, forest plots summarize univariate HR. Only genes that passed these criteria were retained and used for downstream analysis. Correlation network analysis was performed for overlapping prognostic genes for each signature. The Pearson correlation was calculated, and edges with $|r| < 0.05$ were pruned; the results were visualized using "igraph" R package.

The LASSO Cox regression analysis was performed using the "glmnet" R package to screen EMT and DDR prognostic genes further and construct a Cox module. Lasso regression was employed to minimize the risk of overfitting and address collinearity through tenfold cross-

validation. A relationship between regression coefficients and lambda (punishment coefficient) was established using "glmnet". However, based on minimal cross-validation error (λ min), the optimal penalty parameter was determined, and genes with non-zero coefficients at this optimal λ were identified as final EMT and DDR prognostic signatures. The prognostic EMT and DDR-related risk score of each patient was computed as the linear predictor using the following formula:

$$Risk\ score = \sum_{i=1}^{n} \beta i * Expi$$

Where *Exp* represents the gene expression (which was identified by Lasso regression) and β represents the risk coefficients of a gene. This algorithm is significant in removing the over-fitting of the model, helping to remove highly co-expressed genes and ultimately designing a simplified model (W. Dai et al., 2021).

The patients were divided into high and low-risk groups based on optimum cut-off values that were calculated using the "surv_cutpoint" function in R. Kaplan-Meier survival curves comparing risk groups (high-risk and low-risk groups) were plotted. Univariate Cox models on the risk factor yielded HR, 95% CI, and log-rank P values using the "survminer" R package to determine OS. The composite figures, including risk score distribution with median thresholds, heatmaps (normal vs tumour), and survival status plots, were generated using the "pheatmap" R package.

The prognostic performance of novel EMT and DDR signature models was evaluated by plotting receiver operating characteristic (ROC) curves, time-dependent ROC curves, and area under the curve (AUC) values using the "timeROC" R package at 1-, 3-, and 5-year intervals to quantify the discrimination of the signatures. Moreover, t-distributed stochastic neighbour embedding (t-SNE) and principal component analysis (PCA) were performed using "Rtsne" and "ggplot2" R packages to examine the distribution of patients in high-risk and low-risk groups (Shi et al., 2022).

3.4. Independent Prognostic Factor

The clinicopathological characteristics, factors, and risk scores for EMT and DDR prognostic signatures were investigated using Univariate as well as Multivariate Cox regression analysis to determine whether risk scores were an independent prognostic factor for OS or not using the "survival" R package. Different factors, including pathological T stage, N stage, M stage, age (<60 vs ≥60), gender, TNM stage, and risk score, were used in Univariate and Multivariate Cox analysis.

3.5. Relationship Between Clinical Characteristics and Risk Levels

The patients in all six cancers were divided into subgroups based on clinical variables age (<60 versus (vs) ≥ 60), gender (Male vs Female), T stage (T1-T2 vs T3-T4), N stage (N0 vs N1-N3), M stage (M0 vs M1), and TNM stage (Stage I-II vs Stage III-IV) to determine the predictive ability of prognostic EMT and DDR-related gene signature. A correlation analysis was performed between risk levels and clinical characteristics using the "ggpubr" R package. Binary heatmaps representing the distribution of clinical features were plotted.

3.6. Functional Enrichment Analysis

Kyoto Encyclopaedia of Genes and Genomes (KEGG) and Gene Ontology-Biological Process (GO-BP) pathway enrichment analysis were performed using R packages "clusterProfiler" and "org.Hs.eg.db." The R package "limma" was used to obtain DGE in the high-risk and low-risk groups that were subjected to KEGG and GO-BP. The absolute threshold was set as $|Log_2FC| \geq 1$ and P value < 0.05. Overrepresentation analysis used Benjamini-Hochberg (BH) adjustment ($q \leq 0.05$), and dot plots were plotted with "enrichplot."

3.7. Immune Cell Infiltration Profiling

The Cell-type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) algorithm was used to calculate the infiltration ratio of 22 immune cells in all six cancer cohorts. CIBERSORT is a bioinformatics algorithm that is used to calculate cell composition from gene expression matrices of complex tissues. The scores of 22 types of human immune cells can be calculated using a combination of LM22 (leukocyte signature matrix) and the CIBERSORT algorithm. The web version of CIBERSORTx

(https://cibersortx.stanford.edu/runcibersortx.php) was used to calculate the immune score in each sample of all six cancer cohorts. CIBERSORTx deconvolution outputs were joined to clinical risk groups for each signature set in R. The stacked bar plots and raincloud plots were generated using the R package "ggplot2" to compare high- and low-risk groups for each signature. Wilcoxon rank-sum tests were executed, and *p*-values were mentioned in raincloud plots. Based on the top six most significant Wilcoxon *P* values, the Kaplan-Meier curves were generated using the "survminer" R package with Cox proportional HR stratified by low/high cell fraction and risk groups to assess the combined prognostic impact. Pairwise log-rank tests were computed with the R package "survival" to identify significant inter-group differences.

However, for the quantification of non-tumour cellular components within each bulk transcriptome sample, we applied the ESTIMATE (Estimate score, immune score, and stromal score) algorithm via the "estimate" R package from R-Forge. Pearson's correlation analysis was used to fully ensure the reproducibility and quantification of the tumour microenvironment between ESTIMATE metrics and risk scores.

3.8. Signature-Based Nomogram Construction and Validation

Based on the results of the Multivariate Cox regression analysis, a nomogram for OS prediction was constructed using the "cph" function of the R package "rms". Survival probability functions at 1-, 3-, and 5-year were predicted in the nomogram model. The model performance was quantified using the Concordance index (C-index) and the Akaike Information Criterion (AIC). The calibration curves were plotted between predicted OS and observed OS in tertiles by grouping predicted probabilities and observed Kaplan-Meier estimates with 95% CI against median predicted values. Nomogram linear predictors, including age, TNM stage, and risk score, were evaluated by time-dependent ROC curves using the R package "timeROC" at 1-, 3-, and 5-year OS. Moreover, Decision curve analysis (DCA) was performed using the R package "rmda" to assess the clinical utility of the nomogram by computing net benefits across threshold probabilities (0-0.8) for the nomogram and comparator models, including age, TNM stage, risk score, and age + TNM stage. Net benefits curves were plotted with ideal and treat-none/treat-all references.

3.9. Drug Sensitivity Correlation Analysis

The relevant data for performing drug sensitivity correlation analysis on prognostic EMT and DDR-related signatures were acquired from the Cancer Cell Line Encyclopaedia (CCLE) database. Drug sensitivity correlation analysis was performed to identify the candidate compounds whose *in vitro* efficacy is associated with prognostic EMT and DDR-related signatures. First, we computed the EMT and DDR risk scores for each cell line of the respective cancer using the same method mentioned above. Then Pearson's correlation coefficient (ρ) was calculated between the IC_{50} of drugs and the risk score of respective cell lines to compute correlation. We selected the significant drugs based on *p*-values and FDR <0.05.

3.10. Single-sample GSEA (ssGSEA)

The EMT and DDR scores were calculated using the single-sample Gene Set Enrichment Analysis (ssGSEA) method, a non-parametric, rank-based statistical approach designed to evaluate the enrichment of a predefined gene signature within individual samples of gene expression data (Y. Zhang et al., 2023). EMT and DDR scores were calculated using the R package "GSVA (v 2.0.4)". The key parameters were defined by the 'ssgseaParam' function, including the input data as a processed gene expression matrix, the EMT and DDR gene signature as gene sets, an alpha value of 0.25 to adjust the weighting of ranked genes during score calculation, and normalization to make scores comparable across samples. The computed EMT and DDR-based scores were merged with independently derived risk groups (low and high) for each sample, and box plots were generated using the "ggplot2" R package. The two-sided Mann-Whitney U test was used to assess the statistical significance between two groups.

3.11. Statistical Analysis

R Programming Language software (RStudio 2024.12.1+563) was used to perform all statistical analyses. DGE were distinguished between primary tumours and normal tissues by the Wilcoxon test. The Wald test was used to extract p-values in Cox proportional-hazards models. Log-rank text was used in Kaplan-Meier curves to confirm whether those survival curves diverge beyond random fluctuation. Moreover, the Chi-square test was used to detect different proportions. The Mann-Whitney U test was used in the ssGSEA boxplots. Different R packages, including "ggplot2", "grid", "patchwork", "survminer", "VennDiagram", "pheatmap",

"RColorBrewer", "igraph", and "gghalves + ggdist", were used to generate the figures for results. Results were considered statistically significant below $p < 0.05$.

CHAPTER 4

RESULTS

4.1. Identification and Screening of Prognostic DGE EMT- and DDR-Related Gene Signatures

To systematically investigate the role of EMT and DDR genes and identify prognostic EMT and DDR-related genes in six cancers, such as BRCA, LUAD, LUSC, PAAD, STAD, and COAD, we initially performed the gene expression profiles using TCGA cohort in 1,218 BRCA, 329 COAD, 576 LUAD, 553 LUSC, 183 PAAD, and 417 STAD patients. The results reported that a total of 1,416 genes were significantly upregulated, and 2,815 genes were downregulated in TCGA-BRCA tissues compared to adjacent normal breast tissues (see Figure 4.1a), while 1,568 genes were upregulated and 3,238 were downregulated in TCGA-COAD (see Figure 4.1j). It was also reported that 1812 genes were significantly upregulated and 2611 downregulated in TCGA-LUAD (see Figure 4.2a), 2694 genes were upregulated and 3669 downregulated in TCGA-LUSC (see Figure 4.2j), 109 genes were significantly upregulated, and 80 genes were downregulated in TCGA-PAAD (see Figure 4.3a), and 1107 genes were upregulated and 461 were downregulated in TCGA-STAD tissues compared to normal tissues (see Figure 4.3j), $|\log_2 FC| \geq 1$, p value < 0.05).

We subsequently performed Univariate Cox regression and Kaplan-Meier curves analysis to further explore the relationship between EMT and DDR genes and identified DGE to find prognostic biomarkers associated with patient outcomes. The genes that passed these two criteria and overlapped between Cox and KM analyses were selected for further LASSO regression analysis to find robust EMT and DDR-related prognostic signatures. The results showed that only 31 EMT-related DGE passed the criteria for KM and 16 genes for Cox, while for DDR 23 genes passed the criteria for KM and 9 for Cox in BRCA, and only six genes were intersected between KM and Cox in EMT and 6 for DDR in BRCA (see Figure 4.1b,c). Results reported that based on Univariate Cox analysis and KM, 19 genes out of 123 were interconnected with OS for EMT and 2 out of 22 genes for DDR in TCGA-COAD (see Figure 4.1k,l), while for LUAD, 12 out of 90 genes for EMT and 6 out of 33 genes for DDR were

overlapped (see Figure 4.2b,c). Similarly, for LUSC, 11 genes overlapped between KM and Cox analysis for EMT and 4 for DDR (see Figure 4.2k,l). Results revealed that 78 and 16 genes satisfied the criteria for EMT and DDR in PAAD (see Figure 4.3b,c) and 95 and 8 genes for EMT and DDR in STAD, respectively (see Figure 4.3k,l). These overlapped genes represented the candidate genes for subsequent analysis, such as the LASSO regression model for EMT and DDR.

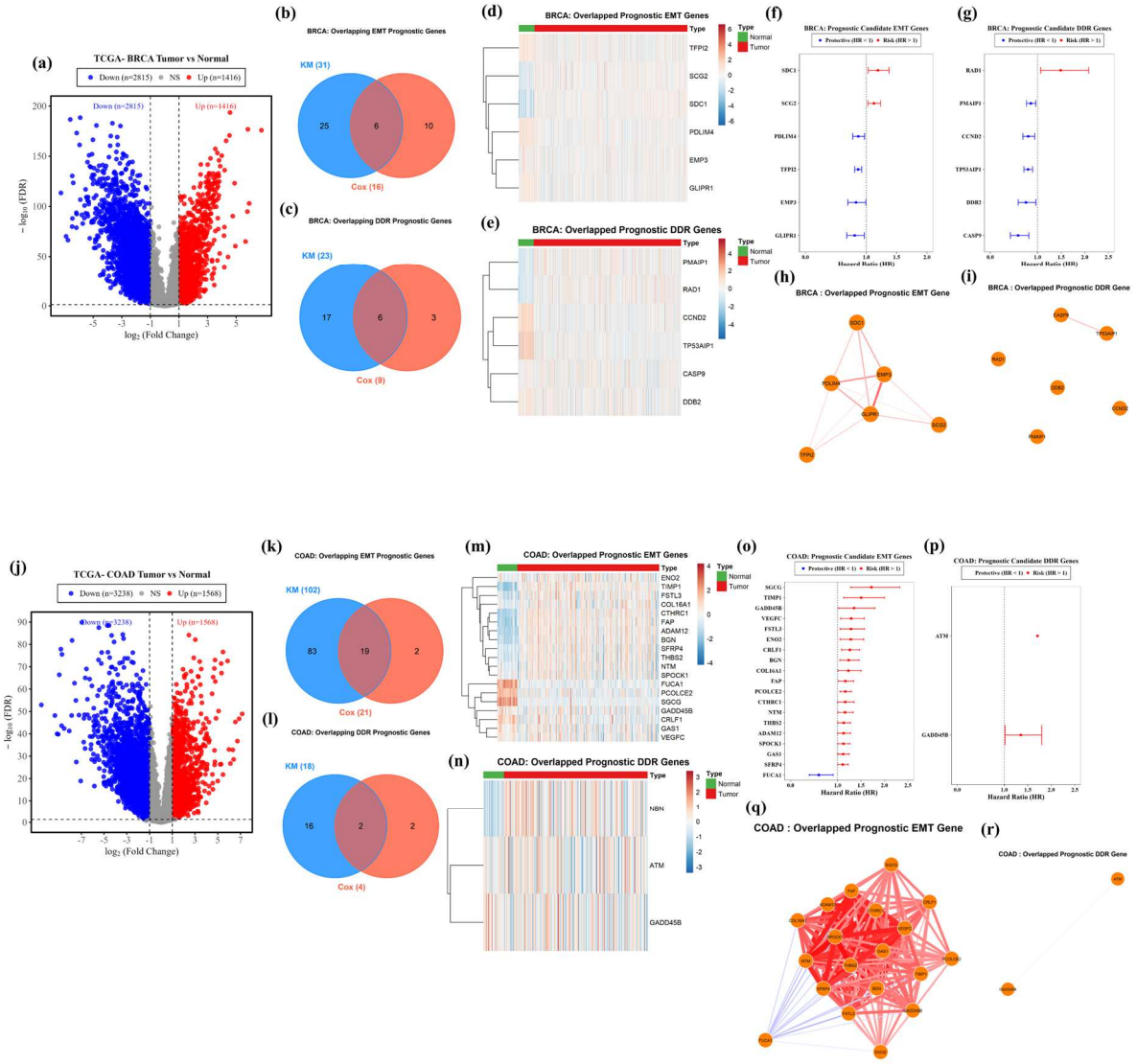


Figure 4.1. In the TCGA-BRCA and TCGA-COAD cohorts, a series of analyses were performed to identify EMT- and DDR-related prognostic candidate genes. The volcano plot showed differentially expressed genes (DEGs) between tumour and normal breast tissues, emphasizing significantly upregulated and downregulated genes (a for TCGA-BRCA; j for TCGA-COAD). Venn diagrams showed the overlap of EMT- and DDR-related genes identified

by both Kaplan–Meier and Cox regression analyses (b-c for TCGA-BRCA; k-l for TCGA-COAD). Heatmaps in panels (d/m) and (e/n) visualize the expression patterns of overlapping EMT- and DDR-related genes, respectively (d-e for TCGA-BRCA; m-n for TCGA-COAD). Univariate Cox regression analyses identify EMT- and DDR-related genes significantly associated with overall survival (OS) (f-g for TCGA-BRCA; o-p for TCGA-COAD). The prognostic candidate EMT- and DDR-related genes are further demonstrated through correlation network analyses (h-i for TCGA-BRCA; q-r for TCGA-COAD), showing their interrelated expression patterns.

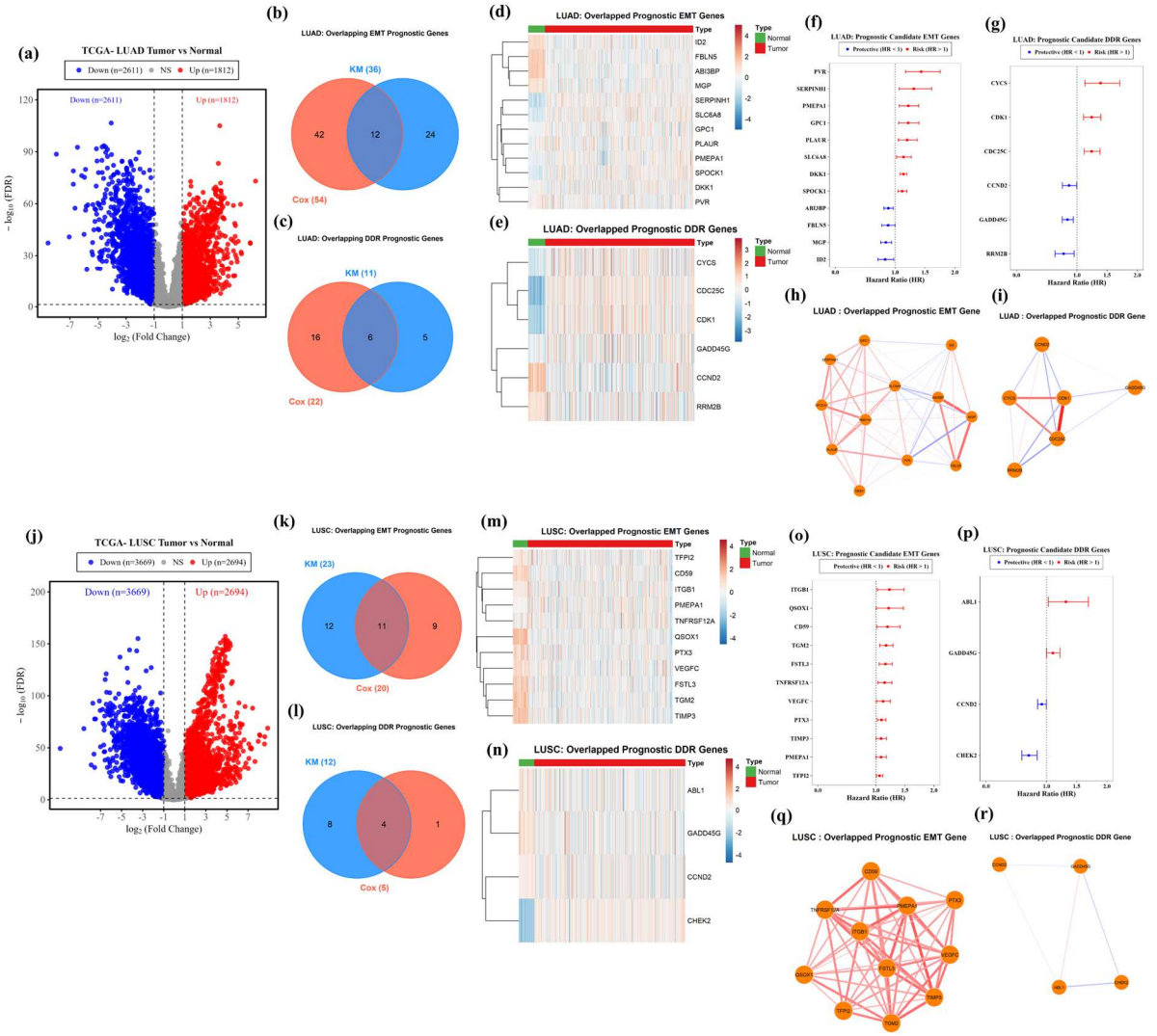


Figure 4.2. In the TCGA-LUAD and TCGA-LUSC cohorts, a series of analyses were performed to identify EMT- and DDR-related prognostic candidate genes. The volcano plot showed differentially expressed genes (DEGs) between tumour and normal breast tissues, emphasizing significantly upregulated and downregulated genes (a for TCGA-LUAD; j for TCGA-LUSC). Venn diagrams showed the overlap of EMT- and DDR-related genes identified by both Kaplan–Meier and Cox regression analyses (b-c for TCGA-LUAD; k-l for TCGA-LUSC). Heatmaps in panels (d/m) and (e/n) visualize the expression patterns of overlapping EMT- and DDR-related genes, respectively (d-e for TCGA-LUAD; m-n for TCGA-LUSC). Univariate Cox regression analyses identify EMT- and DDR-related genes significantly associated with overall survival (OS) (f-g for TCGA-LUAD; o-p for TCGA-LUSC). The prognostic candidate EMT- and DDR-

related genes are further demonstrated through correlation network analyses (h-i for TCGA-LUAD; q-r for TCGA-LUSC), showing their interrelated expression patterns.

Furthermore, heatmap clustering and correlation network analysis reported the distinct expression patterns and interaction networks among the identified prognostic EMT and DDR genes. In BRCA, Cox proportional hazard ratio (HR) revealed that two genes (*SDC1* and *SCG2*), which overlapped, have HR > 1, while all other four genes have protective HR for EMT; for DDR, only *RAD1* has HR > 1 (see Figure 4.1f,g). Similarly, in COAD, only the *FUCA1* gene was associated with a low risk of death for EMT, while both ATM and GADD45B were associated with a high risk of death for DDR (see Figure 4.1o,p). It was also noted that *ABI3BP, FBLN5, MGP*, and *ID2* were reported as protective genes for EMT in LUAD, while *CCND2*, *GADD45G*, and *RRM2B* were reported as protective genes for DDR in LUAD (see Figure 4.2f,g). Interestingly, no gene with HR < 1 was observed for EMT in LUSC, whereas two genes, such as *CCND2* and *CHEK2,* have HR < 1 (see Figure 4.2o,p). In the case of PAAD, *CCND2*, *GADD45G, SESN1, RAD 52,* and *RAD17* have HR < 1, and other genes have above 1 for DDR (see Figure 4.3f,g), whereas only a single gene, *GADD45B,* has a higher risk factor for DDR in STAD (see Figure 4.3o,p).

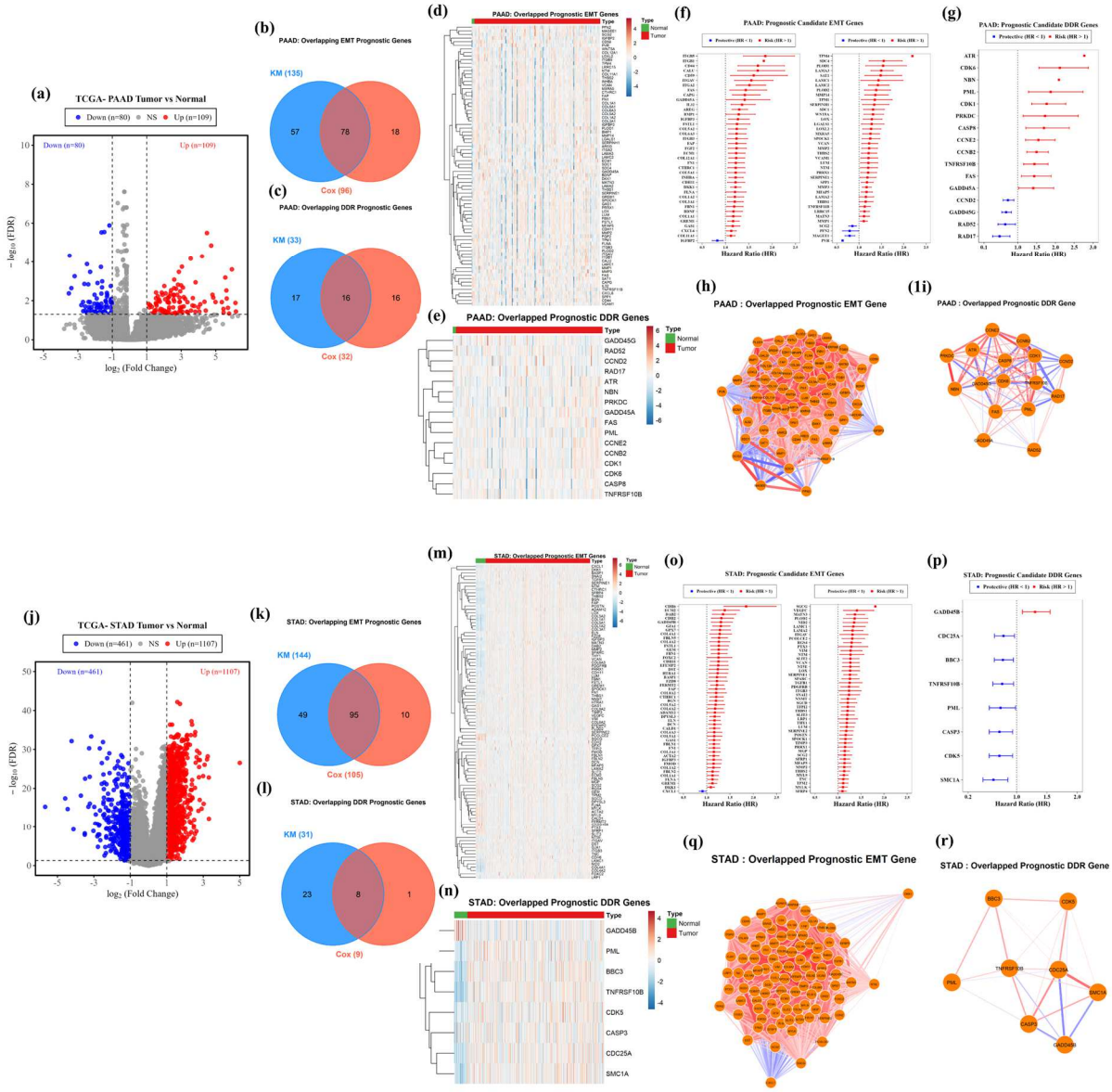


Figure 4.3. In the TCGA-PAAD and TCGA-STAD cohort, a series of analysis were performed to identify EMT- and DDR-related prognostic candidate genes. The volcano plot showed differentially expressed genes (DEGs) between tumour and normal breast tissues, highlighting significantly upregulated and downregulated genes (a for TCGA-PAAD; j for TCGA-STAD). Venn diagrams showed the overlap of EMT- and DDR-related genes identified by both Kaplan–Meier and Cox regression analyses (b-c for TCGA-PAAD; k-l for TCGA-STAD). Heatmaps in panels (d/m) and (e/n) visualize the expression patterns of overlapping EMT- and DDR-related genes, respectively (d-e for TCGA-PAAD; m-n for TCGA-STAD). Univariate Cox regression analyses identify EMT- and DDR-related genes significantly associated with overall survival

(OS) (f-g for TCGA-PAAD; o-p for TCGA-STAD). The prognostic candidate EMT- and DDR-related genes are further demonstrated through correlation network analyses (h-i for TCGA-PAAD; q-r for TCGA-STAD), showing their interrelated expression patterns. We employed the least absolute shrinkage and selection operator (LASSO) Cox regression analysis to build EMT and DDR-related prognostic signature models from these genes, thereby reducing multicollinearity. The coefficient profiles of candidate genes revealed a progressive shrinkage of coefficients as the penalty increased for all cancers (see Figures 4.4–4.6a,b,o,p,q,r). The optimal Lambda (λ) value was determined using 10-fold cross-validation to minimize the partial likelihood deviance and select a robust gene signature with the strongest prognostic relevance. Based on the optimal cut-off threshold of λ, we identified the EMT and DDR-related prognostic gene signatures. However, the prognostic EMT risk scores were identified as follows: risk score = (-0.0977)*(*EMP3* expression) + (-0.1181)* (*GLIPR1* expression) + (-0.1100)*(*PDLIM4* expression) + (0.1227)*(*SCG2* expression) + (-0.2275)*(*SDC1* expression) + (-0.1036)*(*TFPI2* expression) for BRCA, risk score = (0.01958)*(*CRLF1* expression) + (0.0827)*(*ENO2* expression) + (-0.2856)*(*FUCA1* expression) + (0.0360)*(*PCOLCE2* expression) + (0.2715)*(*SGCG* expression) + (0.1146)*(*TIMP1* expression) for COAD, risk score = (-0.0294)*(*ABI3BP* expression) + (0.0986)*(*DKK1* expression) + (0.0714)*(*GPC1* expression) + (-0.0325)*(*ID2* expression) + (-0.0980)*(*MGP* expression) + (0.0251)*(*PMEPA1* expression) + (0.1359)*(*PVR* expression) + (0.1315)*(*SERPINH1* expression) + (0.0125)*(*SPOCK1* expression) for LUAD, risk score = (0.0399)*(*FSTL3* expression) + (0.0336)*(*PTX3* expression) + (0.0799)*(*TGM2* expression) + (0.0197)*(*TNFRSF12A* expression) for LUSC, risk score = (0.0465)*(*AREG* expression) + (0.0133)*(*CD44* expression) + (0.0233)*(*DKK1* expression) + (-0.1040)*(*IGFBP2* expression) + (0.0523)*(*ITGB5* expression) + (0.1930)*(*LAMA3* expression) + (-0.1080)*(*MAGEE1* expression) + (0.0745)*(*MMP3* expression) + (-0.0837)*(*PVR* expression) + (0.0568)*(*SPP1* expression) for PAAD, and risk score = (0.1797)*(*CDH6* expression) +(-0.0419)*(*CXCL1* expression) + (0.0365)*(*DKK1* expression) + (0.0709)*(*GJA1* expression) + (0.1246)*(*MATN3* expression) + (0.1222)*(*NT5E* expression) + (0.0860)*(*SERPINE1* expression) + (0.0362)*(*SGCG* expression) + (0.0369)*(*TFPI2* expression) for STAD.

Whereas the prognostic risk score for DDR-related gene signature was noted as follows: risk score = (-0.3225)*(*CASP9* expression) + (-0.1229)*(*CCND2* expression) + (-0.1499)*(*DDB2*

expression) + (-0.1507)*(*PMAIP1* expression) + (0.0921)*(*RAD1* expression) + (-0.1532)*(*TP53AIP1* expression) for BRCA, risk score = (0.4927)*(*ATM* expression) + (0.2743)*(*GADD45B* expression) for COAD, risk score = (-0.0269)*(*CCND2* expression) + (0.1154)*(*CDC25C* expression) + (0.1733)*(*CYCS* expression) + (-0.1356)*(*GADD45G* expression) + (-0.1241)*(*RRM2B* expression) for LUAD, risk score = (0.0548)*(*ABL1* expression) + (-0.0429)*(*CCND2* expression) + (-0.2465)*(*CHEK2* expression) + (0.0038)*(*GADD45G* expression) for LUSC, risk score = (0.2186)*(*ATR* expression) + (0.0738)*(*CCNB2* expression) + (-0.0639)*(*CCND2* expression) + (0.2086)*(*CDK1* expression) + (0.3949)*(*CDK6* expression) + (0.2328)*(*NBN* expression) + (0.2488)*(*PML* expression) + (-0.1497)*(*RAD52* expression) for PAAD, and risk score = (-0.0907)*(*BBC3* expression) + (-0.0433)*(*CASP3* expression) + (-0.1396)*(*CDK5* expression) + (0.1650)*(*GADD45B* expression) + (-0.2377)*(*PML* expression) + (-0.2243)*(*SMC1A* expression) + (-0.0078)*(*TNFRSF10B* expression) for STAD. The prognostic EMT and DDR-related genes were risk factors for OS. The results revealed that, based on absolute values of coefficients, *SCG2* has the most influence on OS prediction for EMT and *RAD1* for DDR in BRCA, while *SGCG* was found to be the most influential gene for EMT and *ATM* for DDR in COAD. Similarly, for LUAD and LUSC, the *PVR* and *TGM2* genes have a significant impact on OS prediction for EMT, while *CYCS* and *ABL1* affect OS prediction for DDR, respectively. The *LAMA3* and *MATN3* genes were found to have the most significant influence on OS prediction for EMT, while *CDK6* and *GADD45B* were found to have the most significant influence on DDR in PAAD and STAD, respectively.

4.2. Validation of EMT and DDR-Related Gene Signatures

Optimal cut-off values for all six cancers with selected EMT and DDR-related prognostic gene signatures were calculated using the "surv_cutpoint" function of the R package "survminer." Based on optimal cut-off values, patients in all six cancers were stratified into high- and low-risk groups for both gene signatures. Risk score distribution and survival status plots revealed that patients in high-risk groups for all six cancers have markedly higher mortality rates compared to those in the low-risk group for both EMT and DDR-related gene signatures (see Figures 4.4–4.6e,f,s,t). We performed PCA and t-SNE algorithms for dimensionality reduction, and results demonstrated that patients in low- and high-risk groups are dispersed in various directions for both gene signatures in all cancers, although there is some overlap between low-

and high-risk groups in BRCA (see Figures 4.4–4.6g,j,u,x). The results of survival status plots are consistent with Kaplan-Meier curves, which show a significant separation in overall survival between high- and low-risk groups for EMT and DDR-related gene signatures ($p < 0.001$) (see Figures 4.4–4.6k,l,y,z).

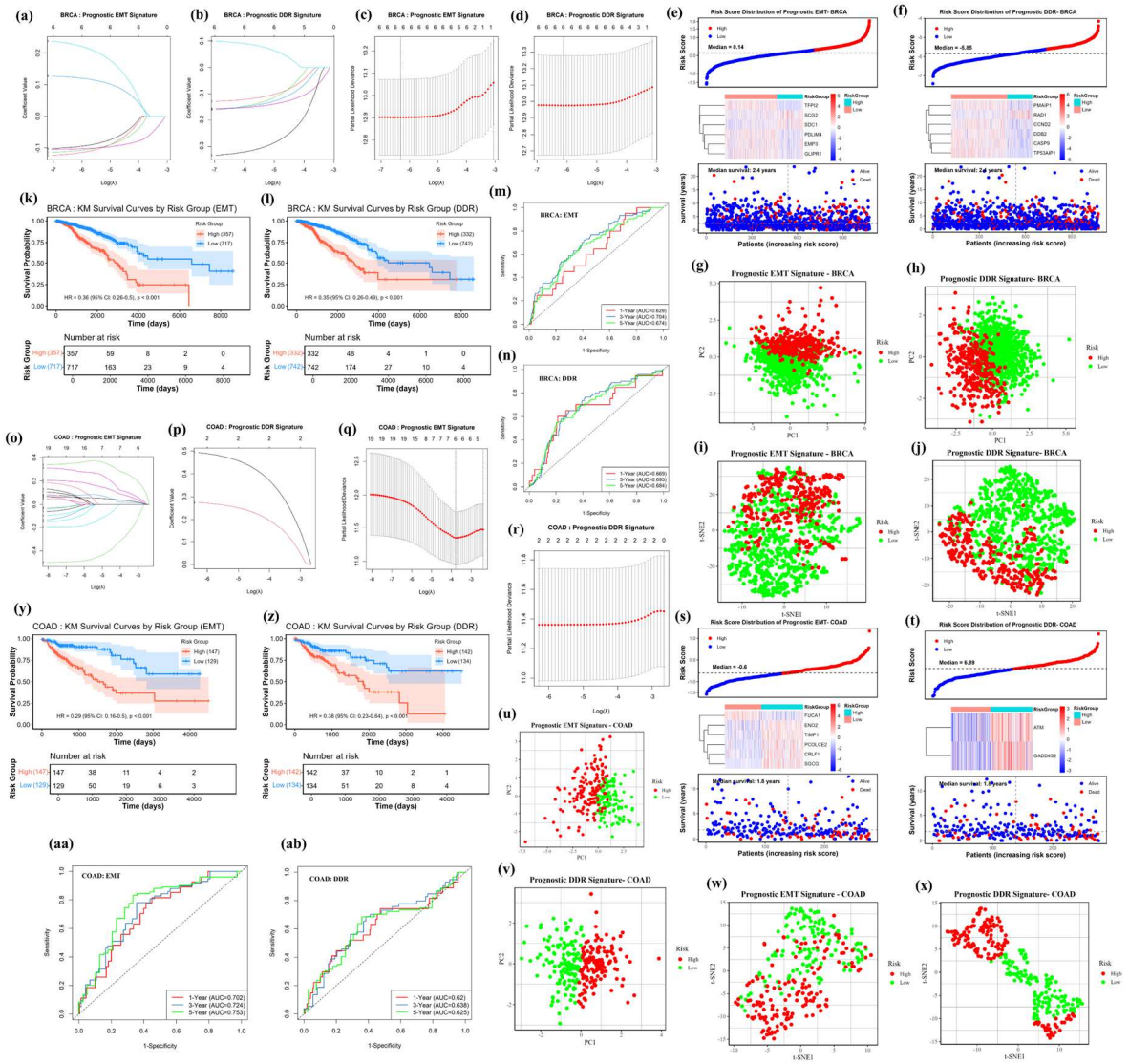


Figure 4.4. Results of LASSO regression to construct EMT- and DDR-related prognostic gene signatures models in TCGA-BRCA and TCGA-COAD cohorts. Coefficient profiles (a-b for TCGA-BRCA; o-p for TCGA-COAD) illustrate the gene selection process, while cross-validation plots (c-d for TCGA-BRCA; q-r for TCGA-COAD) identify optimal lambda values for minimizing prediction error. Risk score analyses (e-f for TCGA-BRCA; s-t for TCGA-COAD) include score distributions, survival status, and heatmaps of gene expression, showing

clear separation between high- and low-risk groups. Dimensionality reduction analyses using PCA (g-h for TCGA-BRCA; u-v for TCGA-COAD) and t-SNE (i-j for TCGA-BRCA; w-x for TCGA-COAD) further validate the models by effectively distinguishing risk groups. Kaplan-Meier survival curves (k-l for TCGA-BRCA; y-z for TCGA-COAD) indicate that high-risk patients had significantly poorer overall survival based on both EMT and DDR signatures. Time-dependent ROC curves (m-n for TCGA-BRCA; aa-ab for TCGA-COAD) illustrate the predictive performance of models across 1-, 3-, and 5-year survival periods.

Similarly, we performed the identical analysis in the validation cohorts across all six cancers to confirm the prognostic power of our robust EMT and DDR gene signatures. The results of validation cohorts (GSE20685, GSE39582, GSE50081, GSE79668, and GSE84433) were consistent with the training cohorts (TCGA), and Kaplan-Meier curves reported a significant separation in high- and low-risk groups in validation cohorts (p <0.001) of BRCA, COAD, LUAD, PAAD, and STAD (see Supplementary File 3, Figures S1–6d,i). While, in LUSC (GSE30219), due to the low number of samples, the results are a little bit varied, the OS results for prognostic EMT gene signature were significant (p = 0.030) while for DDR prognostic gene signature, it was just over the significant threshold (p = 0.051) (see Supplementary File 3, Figure S4d,i). PCA and t-SNE algorithms results of validation cohorts (GSE20685, GSE39582, GSE50081, GSE30219, GSE79668, and GSE84433) across cancers present similar results to training cohorts and revealed the patients are dispersed in various directions for both EMT and DDR prognostic signatures (see Supplementary File 3, Figures S1–6b,c,g,h). The results indicated that the selected EMT and DDR-related prognostic signatures provide strong discrimination power at the transcriptome level, supporting the biological relevance of risk stratification.

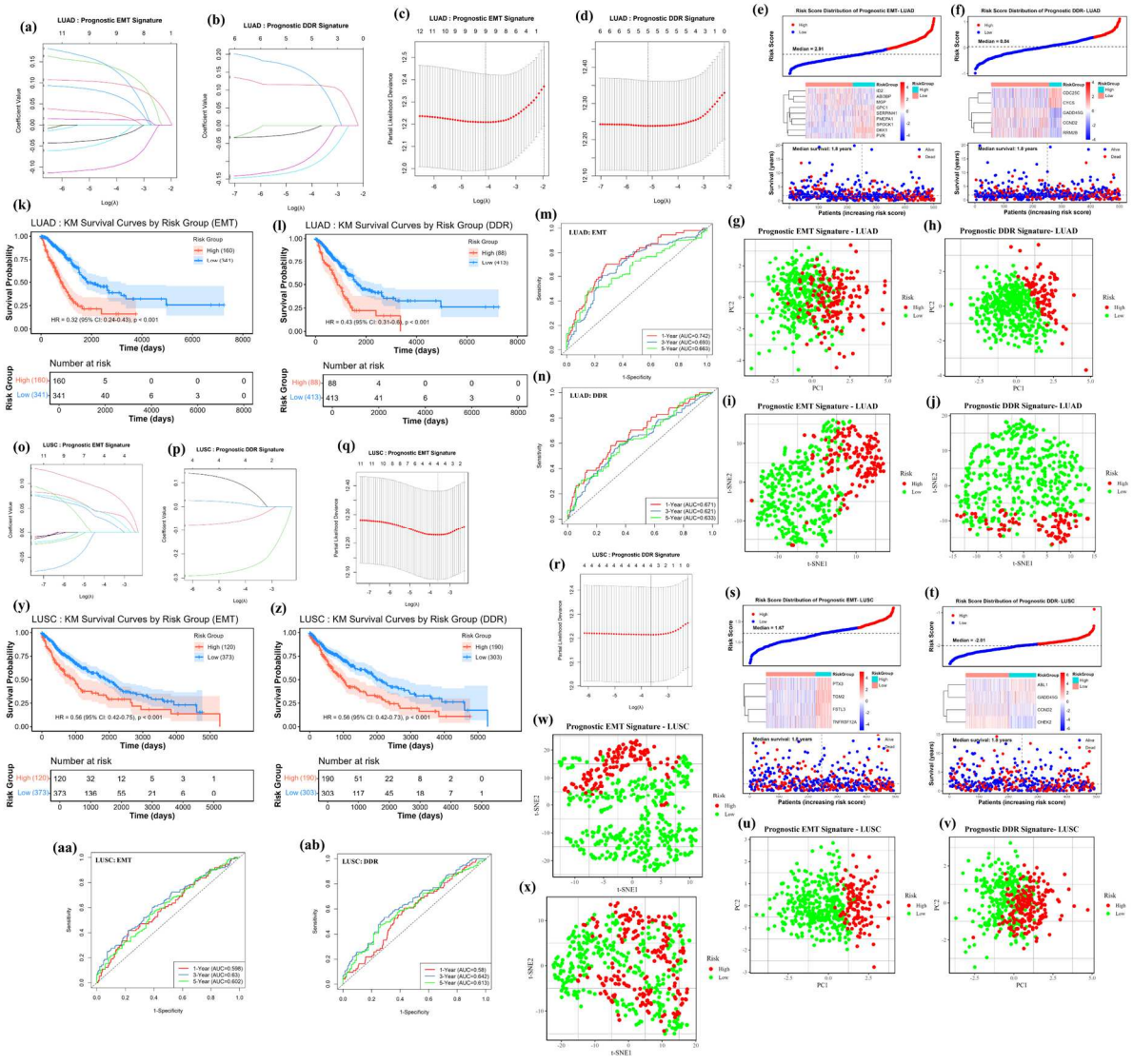


Figure 4.5. Results of LASSO regression to construct EMT- and DDR-related prognostic gene signatures models in TCGA-LUAD and TCGA-LUSC cohorts. Coefficient profiles (a-b for TCGA-LUAD; o-p for TCGA-LUSC) illustrate the gene selection process, while cross-validation plots (c-d for TCGA-LUAD; q-r for TCGA-LUSC) identify optimal lambda values for minimizing prediction error. Risk score analyses (e-f for TCGA-LUAD; s-t for TCGA-LUSC) include score distributions, survival status, and heatmaps of gene expression, showing clear separation between high- and low-risk groups. Dimensionality reduction analyses using PCA (g-h for TCGA-LUAD; u-v for TCGA-LUSC) and t-SNE (i-j for TCGA-LUAD; w-x for TCGA-LUSC) further validates the models by effectively distinguishing risk groups. Kaplan-Meier survival curves (k-l for TCGA-LUAD; y-z for TCGA-LUSC) indicate that high-risk patients had significantly poorer overall survival based on both EMT and DDR signatures.

Time-dependent ROC curves (m-n for TCGA-LUAD; aa-ab for TCGA-LUSC) illustrate the predictive performance of models across 1-, 3-, and 5-year survival periods.

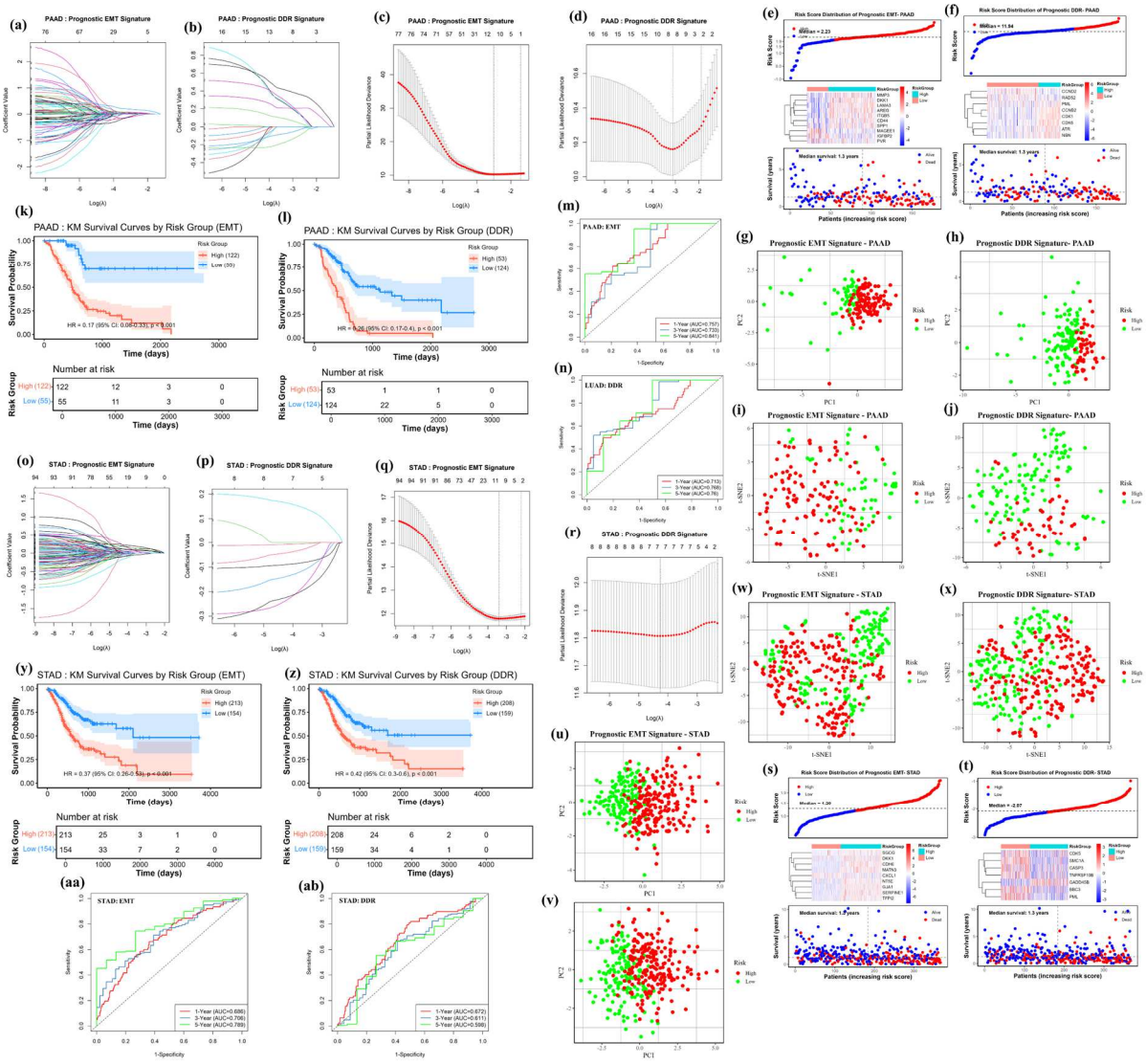


Figure 4.6. Results of LASSO regression to construct EMT- and DDR-related prognostic gene signatures models in TCGA-PAAD and TCGA-STAD cohorts. Coefficient profiles (a-b for TCGA-PAAD; o-p for TCGA-STAD) illustrate the gene selection process, while cross-validation plots (c-d for TCGA-PAAD; q-r for TCGA-STAD) identify optimal lambda values for minimizing prediction error. Risk score analyses (e-f for TCGA-PAAD; s-t for TCGA-STAD) include score distributions, survival status, and heatmaps of gene expression, showing clear separation between high- and low-risk groups. Dimensionality reduction analyses using PCA (g-h for TCGA-PAAD; u-v for TCGA-STAD) and t-SNE (i-j for TCGA-PAAD; w-x for TCGA-STAD) further validate the models by effectively distinguishing risk groups. Kaplan-

Meier survival curves (k-l for TCGA-PAAD; y-z for TCGA-STAD) indicate that high-risk patients had significantly poorer overall survival based on both EMT and DDR signatures. Time-dependent ROC curves (m-n for TCGA-PAAD; aa-ab for TCGA-STAD) illustrate the predictive performance of models across 1-, 3-, and 5-year survival periods.

We predicted the accuracy of the model across 1-, 3-, and 5-year survival periods using time-dependent ROC curves for training (TCGA) and validation cohorts (GEO). The results demonstrated that the area under the curves (AUC) for 1-, 3-, and 5-year OS were 0.629, 0.704, and 0.674 for EMT-related prognostic signature and 0.660, 0.695, and 0.684 for DDR-related prognostic signature in the TCGA-BRCA cohort, respectively (see Figure 4.4m,n). While in the validation cohort BRCA-GSE20685, AUC values for 1-, 3-, and 5-year OS were noted as 0.861, 0.646, and 0.661 for EMT-related prognostic signature and 0.799, 0.585, and 0.597 for DDR-related prognostic signature, respectively (see Supplementary File 3, Figure S1e,j). In TCGA-COAD, AUC values for EMT-related prognostic signature were 0.702, 0.724, and 0.753, while for DDR, the AUC values were 0.62, 0.638, and 0.625, respectively (see Figure 4.4aa,ab). Similarly, in validation cohort COAD-GSE39582, AUC values for EMT and DDR-related signatures were observed as (0.555, 0.542, and 0.550) and (0.562, 0.550, and 0.566) for 1-, 3- and 5-year OS, respectively (see Supplementary File 3, Figure S2e,j). It was noted that AUC values in LUAD were 0.742, 0.693, and 0.663 for EMT-related signature and 0.671, 0.621, and 0.633 for DDR-related signature (see Figure 4.5m,n), while in LUSC, AUC values were 0.598, 0.63, and 0.602 for EMT-related signature and 0.58, 0.642, and 0.613 for DDR-related signature, respectively (see Figure 4.5aa,ab). Almost similar results were noted in the validation cohorts of LUAD (GSE50081) and LUSC (GSE30219), where AUC values were observed as (0.696, 0.638, and 0.638) and (0.701, 0.665, and 0.568) for EMT-related prognostic signature, and (0.72, 0.636, and 0.637) and (0.521, 0.664, and 0.62) for DDR-related prognostic signature, respectively (see Supplementary File 3, Figures S3–4e,j).

Furthermore, EMT-related signatures show the 0.757, 0.733, and 0.841 AUC values in PAAD and 0.686, 0.706, and 0.789 in the STAD cohort for 1-, 3-, and 5-year OS, whereas DDR-signature presented 0.713, 0.768, and 0.76 AUC values in PAAD, 0.672, 0.611, and 0.598 in STAD, respectively (see Figure 4.6m,n/aa,ab). Similar to the training cohort, the validation cohort presented the optimal AUC results in PAAD-GSE79668 for EMT (0.669, 0.857, and 0.725) and DDR-related (0.651, 0.724, and 0,939) prognostic signatures (see Supplementary

File 3, Figure S5e,j). While in STAD-GSE84433, AUC values were noted as 0.579, 0.611, and 0.628 for EMT-related prognostic signature and 0.544, 0.595, and 0.605 for DDR-prognostic gene signature (see Supplementary File 3, Figure S6e,j). It is reported that the EMT and DDR-related gene signatures model performed best in PAAD, with higher AUC values. In comparison, LUSC showed lower AUC values compared to the others, both in the training cohort (TCGA) and in the validation cohorts (GEO). Overall, based on the results, it is reported that our EMT and DDR-related gene signatures performed well for OS prediction in both the training and validation cohorts.

4.3. Independent Prognostic Factors of EMT- and DDR-Related Signatures

We performed the univariate and multivariate Cox analysis to identify the independent prognostic factors of EMT- and DDR-related gene signatures associated with OS in training (TCGA) and validation cohorts (GEO). Results reported that, based on univariate Cox analysis, age, TNM stage, T stage, N stage, M stage and risk score were identified as independent prognostic factors for OS with EMT- and DDR-related signatures, while age, TNM stage and risk score were identified as independent prognostic factors for OS by univariate and multivariate Cox analysis in TCGA-BRCA cohort (see Table 4.1). In TCGA-COAD cohort, univariate and multivariate Cox analysis revealed that M stage and risk score were independent prognostic factors for OS in EMT-related signature. In contrast, in the DDR-related signature, only the M stage was identified as an independent prognostic factor for OS in both Cox analyses. By the univariate analysis in TCGA-LUAD for EMT-related signature, TNM stage, T stage, N stage, and risk score were significantly associated with OS but in multivariate analysis N stage (HR = 1.69, 95% CI = 1.17-2.44, $p < 0.005$) and risk score (HR = 2.98, 95% CI = 2.14-4.16, $p < 0.001$) remained as an independent prognostic factors for OS, while in DDR-related signature T stage (HR = 1.63, 95% CI = 1.05-2.54, $p < 0.030$), N stage (HR = 1.74, 95% CI = 1.21-2.51, $p < 0.003$) and risk score (HR = 2.59, 95% CI = 1.68-4.00, $p < 0.001$) remained as independent prognostic factors for OS. But, among them, the risk score showed the highest significance ($p < 0.001$) for OS by univariate and multivariate analysis (see Table 4.1).

Table 4.1. Univariate and multivariate Cox analysis relationship between clinicopathological factors and risk scores associated withs OS across six cancers in test cohort (TCGA) for prognostic EMT and DDR-related signatures.

Variable	Univariate analysis		Multivariate analysis		Univariate analysis		Multivariate analysis	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value
	EMT				DDR			
TCGA-BRCA								
Age (< 60 vs ≥60)	1.94 (1.41-2.69)	< **0.001**	1.77 (1.25-2.51)	**0.001**	1.94 (1.41-2.69)	< **0.001**	2.07 (1.48-2.92)	< **0.001**
Gender (Male vs Female)	0.83 (0.12-5.93)	0.851	0.73 (0.10-5.27)	0.755	0.83 (0.12-5.93)	0.851	0.54 (0.08-3.92)	0.545
TNM Stage (Stage I-II vs Stage III-IV)	2.54 (1.83-3.52)	< **0.001**	2.53 (1.54-4.14)	< **0.001**	2.54 (1.83-3.52)	< **0.001**	2.13 (1.30-3.47)	**0.002**
T Stage (T1-T2 vs T3-T4)	1.72 (1.19-2.50)	**0.004**	0.87 (0.54-1.40)	0.567	1.72 (1.19-2.50)	**0.004**	1.00 (0.62-1.60)	0.985
N Stage (N0 vs N1-N3)	2.27 (1.60-3.24)	< **0.001**	1.38 (0.88-2.14)	0.158	2.27 (1.60-3.24)	< **0.001**	1.67 (1.08-2.58)	**0.021**
M Stage (M0 vs M1)	1.86 (1.25-2.79)	**0.002**	1.21 (0.77-1.90)	0.402	1.86 (1.25-2.79)	**0.002**	1.25 (0.80-1.96)	0.321
Risk Score	2.88 (2.10-3.95)	< **0.001**	3.10 (2.17-4.43)	< **0.001**	2.91 (2.03-4.17)	< **0.001**	2.90 (2.01-4.18)	< **0.001**
TCGA-COAD								
Age (< 60 vs ≥60)	1.34 (0.78-2.30)	0.289	1.84 (0.91-3.70)	0.089	1.34 (0.78-2.30)	0.289	1.92 (0.95-3.85)	0.068
Gender (Male vs Female)	1.45 (0.88-2.38)	0.142	1.23 (0.68-2.22)	0.503	1.45 (0.88-2.38)	0.142	1.25 (0.70-2.26)	0.452
TNM Stage (Stage I-II vs Stage III-IV)	3.00 (1.73-5.22)	< **0.001**	1.87 (0.41-8.50)	0.415	3.00 (1.73-5.22)	< **0.001**	1.25 (0.23-6.92)	0.797
T Stage (T1-T2 vs T3-T4)	2.42 (0.97-6.04)	0.059	1.19 (0.39-3.66)	0.760	2.42 (0.97-6.04)	0.059	1.23 (0.40-3.81)	0.718
N Stage (N0 vs N1-N3)	2.34 (1.42-3.85)	**0.001**	1.14 (0.25-5.19)	0.863	2.34 (1.42-3.85)	**0.001**	1.74 (0.32-9.52)	0.521
M Stage (M0 vs M1)	2.88 (1.70-4.87)	< **0.001**	2.39 (1.27-4.49)	**0.007**	2.88 (1.70-4.87)	< **0.001**	2.15 (1.15-4.02)	**0.016**
Risk Score	3.88 (2.19-6.88)	< **0.001**	2.96 (1.51-5.77)	**0.001**	2.80 (1.46-5.35)	**0.002**	1.92 (0.84-4.40)	0.122
TCGA-LUAD								
Age (< 60 vs ≥60)	1.04 (0.75-1.45)	0.800	1.25 (0.88-1.77)	0.211	1.04 (0.75-1.45)	0.800	1.22 (0.86-1.72)	0.264
Gender (Male vs Female)	1.05 (0.79-1.41)	0.721	1.04 (0.77-1.42)	0.783	1.05 (0.79-1.41)	0.721	1.03 (0.76-1.40)	0.847
TNM Stage (Stage I-II vs Stage III-IV)	2.63 (1.92-3.59)	< **0.001**	1.42 (0.92-2.19)	0.110	2.63 (1.92-3.59)	< **0.001**	1.37 (0.89-2.11)	0.158
T Stage (T1-T2 vs T3-T4)	2.32 (1.58-3.39)	< **0.001**	1.52 (0.97-2.38)	0.069	2.32 (1.58-3.39)	< **0.001**	1.63 (1.05-2.54)	**0.030**
N Stage (N0 vs N1-N3)	2.49 (1.86-3.34)	< **0.001**	1.69 (1.17-2.44)	**0.005**	2.49 (1.86-3.34)	< **0.001**	1.74 (1.21-2.51)	**0.003**
M Stage (M0 vs M1)	1.02 (0.73-1.41)	0.928	0.97 (0.69-1.36)	0.848	1.02 (0.73-1.41)	0.928	1.04 (0.74-1.47)	0.817
Risk Score	3.23 (2.37-4.41)	< **0.001**	2.98 (2.14-4.16)	< **0.001**	2.92 (1.95-4.36)	< **0.001**	2.59 (1.68-4.00)	< **0.001**
TCGA-LUSC								
Age (< 60 vs ≥60)	1.28 (0.87-1.90)	0.215	1.16 (0.76-1.77)	0.476	1.28 (0.87-1.90)	0.215	1.15 (0.76-1.75)	0.512
Gender (Male vs Female)	1.19 (0.87-1.64)	0.28	1.34 (0.96-1.88)	0.086	1.19 (0.87-1.64)	0.280	1.32 (0.94-1.85)	0.105

TNM Stage (Stage I-II vs Stage III-IV)	1.50 (1.08-2.09)	**0.015**	1.06 (0.66-1.71)	0.810	1.50 (1.08-2.09)	**0.015**	1.08 (0.67-1.74)	0.747
T Stage (T1-T2 vs T3-T4)	1.66 (1.20-2.29)	**0.002**	1.38 (0.89-2.14)	0.154	1.66 (1.20-2.29)	**0.002**	1.40 (0.90-2.17)	0.132
N Stage (N0 vs N1-N3)	1.15 (0.87-1.51)	0.337	1.19 (0.85-1.66)	0.303	1.15 (0.87-1.51)	0.337	1.16 (0.83-1.62)	0.392
M Stage (M0 vs M1)	1.64 (1.13-2.40)	**0.010**	1.51 (1.01-2.26)	**0.046**	1.64 (1.13-2.40)	**0.010**	1.40 (0.93-2.09)	0.106
Risk Score	3.99 (2.02-7.91)	< **0.001**	4.88 (2.37-10.03)	< **0.001**	4.12 (2.16-7.86)	< **0.001**	4.39 (2.22-8.70)	< **0.001**
TCGA-PAAD								
Age (< 60 vs ≥60)	1.40 (0.89-2.21)	0.143	1.45 (0.91-2.32)	0.120	1.40 (0.89-2.21)	0.143	1.34 (0.84-2.14)	0.223
Gender (Male vs Female)	0.82 (0.55-1.24)	0.350	0.92 (0.60-1.42)	0.721	0.82 (0.55-1.24)	0.350	1.01 (0.66-1.55)	0.966
TNM Stage (Stage I-II vs Stage III-IV)	0.99 (0.24-4.03)	0.986	0.95 (0.22-4.07)	0.942	0.99 (0.24-4.03)	0.986	1.14 (0.26-4.89)	0.863
T Stage (T1-T2 vs T3-T4)	2.05 (1.09-3.87)	**0.026**	1.02 (0.53-1.97)	0.957	2.05 (1.09-3.87)	**0.026**	1.06 (0.55-2.08)	0.856
N Stage (N0 vs N1-N3)	2.04 (1.22-3.41)	**0.007**	1.66 (0.95-2.90)	0.076	2.04 (1.22-3.41)	**0.007**	1.57 (0.89-2.78)	0.123
M Stage (M0 vs M1)	0.82 (0.54-1.24)	0.343	1.11 (0.72-1.73)	0.631	0.82 (0.54-1.24)	0.343	1.24 (0.80-1.93)	0.333
Risk Score	5.15 (2.95-9.00)	< **0.001**	4.73 (2.62-8.54)	< **0.001**	4.62 (2.77-7.71)	< **0.001**	4.30 (2.45-7.54)	< **0.001**
TCGA-STAD								
Age (< 60 vs ≥60)	1.47 (1.02-2.12)	**0.040**	1.89 (1.27-2.81)	**0.002**	1.47 (1.02-2.12)	**0.040**	2.01 (1.34-3.02)	**0.001**
Gender (Male vs Female)	1.34 (0.94-1.90)	0.105	1.37 (0.93-2.01)	0.112	1.34 (0.94-1.90)	0.105	1.26 (0.85-1.85)	0.248
TNM Stage (Stage I-II vs Stage III-IV)	1.83 (1.26-2.65)	**0.001**	1.34 (0.77-2.30)	0.298	1.83 (1.26-2.65)	**0.001**	1.36 (0.80-2.31)	0.258
T Stage (T1-T2 vs T3-T4)	1.68 (1.11-2.55)	**0.014**	1.00 (0.59-1.68)	0.991	1.68 (1.11-2.55)	**0.014**	1.22 (0.73-2.05)	0.453
N Stage (N0 vs N1-N3)	1.84 (1.24-2.74)	**0.002**	1.43 (0.78-2.61)	0.25	1.84 (1.24-2.74)	**0.002**	1.38 (0.76-2.49)	0.285
M Stage (M0 vs M1)	1.57 (0.93-2.64)	0.088	1.57 (0.90-2.76)	0.114	1.57 (0.93-2.64)	0.088	1.57 (0.90-2.77)	0.115
Risk Score	4.15 (2.72-6.32)	< **0.001**	4.86 (3.09-7.63)	< **0.001**	3.29 (2.01-5.37)	< **0.001**	3.89 (2.21-6.82)	< **0.001**

The validation cohort revealed that in BRCA-GSE20685, TNM stage, T stage, N stage, M stage and risk score were observed as independent prognostic factors ($p < 0.001$) for OS in univariate Cox analysis, while with multivariate Cox analysis only T stage (HR = 2.37, 95% CI = 1.22-4.61, $p = 0.011$), N stage (HR = 3.08, 95% CI = 1.73-5.48, $p < 0.001$) and risk score (HR = 3.07, 95% CI = 1.59-5.90, $p = 0.001$) remained as prognostic risk factor for prognostic EMT and (T stage: HR = 2.62, 95% CI = 1.35-5.06, $p = 0.004$), (N stage: HR = 3.23, 95% CI = 1.82-5.72, $p < 0.001$) and (risk score: HR = 2.24, 95% CI = 1.15-4.38, $p = 0.001$) for DDR-related gene signatures (see Table 4.2). In validation cohort of COAD-GSE39582, only M stage had remained as independent prognostic factor for OS in EMT (M stage: HR = 4.18, 95% CI = 2.93-5.96, $p < 0.001$) and DDR-related (M stage: HR = 4.07, 95% CI = 2.86-5.80, $p < 0.001$) gene signature based on multivariate Cox analysis (see Table 4.2).

Table 4.2. Univariate and multivariate Cox analysis relationship between clinicopathological factors and risk scores associated with OS across six cancers in the validation cohort (GEO) for prognostic EMT and DDR-related signatures.

Variable	Univariate analysis		Multivariate analysis		Univariate analysis		Multivariate analysis	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value
	EMT				DDR			
BRCA-GSE20685								
Age (< 60 vs ≥60)	1.12 (0.61-2.08)	0.709	0.98 (0.52-1.86)	0.948	1.12 (0.61-2.08)	0.709	1.08 (0.57-2.07)	0.805
TNM Stage (Stage I-II vs Stage III-IV)	4.94 (2.67-9.13)	**< 0.001**	1.22 (0.40-3.69)	0.730	4.94 (2.67-9.13)	**< 0.001**	1.99 (0.65-6.05)	0.225
T Stage (T1-T2 vs T3-T4)	4.29 (2.64-6.96)	**< 0.001**	2.37 (1.22-4.61)	**0.011**	4.29 (2.64-6.96)	**< 0.001**	2.62 (1.35-5.06)	**0.004**
N Stage (N0 vs N1-N3)	3.79 (2.16-6.62)	**< 0.001**	3.08 (1.73-5.48)	**< 0.001**	3.79 (2.16-6.62)	**< 0.001**	3.23 (1.82-5.72)	**< 0.001**
M Stage (M0 vs M1)	5.20 (2.39-11.33)	**< 0.001**	1.23 (0.37-4.09)	0.739	5.20 (2.39-11.33)	**< 0.001**	0.74 (0.22-2.48)	0.621
Risk Score	3.90 (2.06-7.39)	**< 0.001**	3.07 (1.59-5.90)	**0.001**	2.13 (1.13-4.03)	**0.019**	2.24 (1.15-4.38)	**0.018**
COAD-GSE39582								
Age (< 60 vs ≥60)	1.41 (1.00-1.99)	0.052	1.54 (1.08-2.18)	0.016	1.41 (1.00-1.99)	0.052	1.55 (1.10-2.20)	0.013
Gender (Male vs Female)	1.31 (0.98-1.74)	0.067	1.28 (0.96-1.72)	0.094	1.31 (0.98-1.74)	0.067	1.27 (0.95-1.70)	0.109
TNM Stage (Stage I-II vs Stage III-IV)	1.80 (1.35-2.39)	**< 0.001**	1.39 (0.81-2.38)	0.237	1.80 (1.35-2.39)	**< 0.001**	1.34 (0.79-2.27)	0.272
T Stage (T1-T2 vs T3-T4)	2.37 (1.21-4.62)	**0.012**	1.85 (0.94-3.64)	0.075	2.37 (1.21-4.62)	**0.012**	1.79 (0.91-3.53)	0.091
N Stage (N0 vs N1-N3)	1.76 (1.32-2.34)	**< 0.001**	0.97 (0.57-1.66)	0.917	1.76 (1.32-2.34)	**< 0.001**	1.00 (0.59-1.69)	0.991
M Stage (M0 vs M1)	4.85 (3.50-6.72)	**< 0.001**	4.18 (2.93-5.96)	**< 0.001**	4.85 (3.50-6.72)	**< 0.001**	4.07 (2.86-5.80)	**< 0.001**
Risk Score	1.49 (0.92-2.41)	0.101	1.06 (0.66-1.69)	0.806	2.14 (1.22-3.75)	**0.008**	1.58 (0.92-2.71)	0.094
LUAD-GSE50081								
Age (< 60 vs ≥60)	1.04 (0.75-1.45)	0.800	1.25 (0.88-1.77)	0.211	1.04 (0.75-1.45)	0.800	1.22 (0.86-1.72)	0.264
Gender (Male vs Female)	1.05 (0.79-1.41)	0.721	1.04 (0.77-1.42)	0.783	1.05 (0.79-1.41)	0.721	1.03 (0.76-1.40)	0.847
TNM Stage (Stage I-II vs Stage III-IV)	2.63 (1.92-3.59)	**< 0.001**	1.42 (0.92-2.19)	0.110	2.63 (1.92-3.59)	**< 0.001**	1.37 (0.89-2.11)	0.158
T Stage (T1-T2 vs T3-T4)	2.32 (1.58-3.39)	**< 0.001**	1.52 (0.97-2.38)	0.069	2.32 (1.58-3.39)	**< 0.001**	1.63 (1.05-2.54)	**0.030**
N Stage (N0 vs N1-N3)	2.49 (1.86-3.34)	**< 0.001**	1.69 (1.17-2.44)	**0.005**	2.49 (1.86-3.34)	**< 0.001**	1.74 (1.21-2.51)	**0.003**
M Stage (M0 vs M1)	1.02 (0.73-1.41)	0.928	0.97 (0.69-1.36)	0.848	1.02 (0.73-1.41)	0.928	1.04 (0.74-1.47)	0.817
Risk Score	3.23 (2.37-4.41)	**< 0.001**	2.98 (2.14-4.16)	**< 0.001**	2.92 (1.95-4.36)	**< 0.001**	2.59 (1.68-4.00)	**< 0.001**
LUSC-GSE30219								
Age (< 60 vs ≥60)	2.17 (1.06-4.42)	**0.033**	2.67 (1.26-5.64)	**0.010**	2.17 (1.06-4.42)	**0.033**	2.58 (1.19-5.61)	**0.016**
Gender (Male vs Female)	0.44 (0.15-1.26)	0.125	0.59 (0.20-1.74)	0.337	0.44 (0.15-1.26)	0.125	0.56 (0.19-1.66)	0.297
TNM Stage (Stage I-II vs Stage III-IV)	6.57 (1.46-29.61)	**0.014**	2.30 (0.34-15.51)	0.394	6.57 (1.46-29.61)	**0.014**	3.79 (0.58-24.65)	0.162
T Stage (T1-T2 vs T3-T4)	2.97 (1.23-7.14)	**0.015**	1.91 (0.61-6.03)	0.268	2.97 (1.23-7.14)	**0.015**	1.41 (0.48-4.19)	0.531

N Stage (N0 vs N1-N3)	2.33 (1.06-5.11)	**0.035**	2.70 (1.00-7.29)	**0.050**	2.33 (1.06-5.11)	**0.035**	2.45 (0.90-6.61)	0.078
Risk Score	6.95 (0.66-72.94)	0.106	9.18 (0.84-99.88)	0.069	4.47 (0.76-26.25)	0.098	2.50 (0.43-14.67)	0.311
PAAD-GSE79668								
Age (< 60 vs ≥60)	1.18 (0.62-2.23)	0.613	2.07 (0.98-4.40)	0.058	1.18 (0.62-2.23)	0.613	1.22 (0.60-2.48)	0.587
Gender (Male vs Female)	1.24 (0.66-2.30)	0.501	1.05 (0.53-2.10)	0.885	1.24 (0.66-2.30)	0.501	1.09 (0.55-2.16)	0.804
TNM Stage (Stage I-II vs Stage III-IV)	1.21 (0.54-2.73)	0.649	1.53 (0.46-5.03)	0.487	1.21 (0.54-2.73)	0.649	0.85 (0.27-2.61)	0.771
T Stage (T1-T2 vs T3-T4)	1.97 (0.98-3.99)	0.058	2.96 (1.17-7.48)	**0.022**	1.97 (0.98-3.99)	0.058	1.55 (0.65-3.68)	0.326
N Stage (N0 vs N1-N3)	1.44 (0.72-2.87)	0.297	0.64 (0.26-1.59)	0.339	1.44 (0.72-2.87)	0.297	1.38 (0.60-3.17)	0.450
M Stage (M0 vs M1)	1.45 (0.44-4.75)	0.536	0.15 (0.02-1.27)	0.083	1.45 (0.44-4.75)	0.536	2.03 (0.40-10.36)	0.395
Risk Score	1.05 (1.02-1.07)	**0.001**	1.09 (1.05-1.14)	**< 0.001**	1.11 (1.03-1.19)	**0.005**	1.11 (1.02-1.20)	**0.017**
STAD-GSE84433								
Age (< 60 vs ≥60)	1.84 (1.34-2.52)	**< 0.001**	1.90 (1.38-2.60)	**< 0.001**	1.84 (1.34-2.52)	**< 0.001**	1.89 (1.38-2.59)	**< 0.001**
Gender (Male vs Female)	1.27 (0.91-1.76)	0.159	1.21 (0.87-1.68)	0.266	1.27 (0.91-1.76)	0.159	1.17 (0.84-1.63)	0.351
T Stage (T1-T2 vs T3-T4)	3.53 (1.81-6.91)	**< 0.001**	3.01 (1.53-5.91)	**0.001**	3.53 (1.81-6.91)	**< 0.001**	2.99 (1.52-5.89)	**0.001**
N Stage (N0 vs N1-N3)	2.29 (1.46-3.58)	**< 0.001**	1.97 (1.26-3.09)	**0.003**	2.29 (1.46-3.58)	**< 0.001**	2.02 (1.29-3.17)	**0.002**
Risk Score	2.58 (1.62-4.09)	**< 0.001**	2.16 (1.37-3.40)	**0.001**	3.74 (1.79-7.82)	**< 0.001**	2.75 (1.33-5.69)	**0.006**

In the case of TCGA-LUSC, univariate and multivariate analysis revealed that M stage (HR = 1.64, 95% CI = 1.13-2.40, p = 0.010 and HR = 1.51, 95% CI = 1.01-2.26, $p < 0.046$, respectively) and risk score (HR = 3.99, 95% CI = 2.02-7.91, p < 0.001 and HR = 4.88, 95% CI = 2.37-10.03, $p < 0.001$, respectively) were identified as independent prognostic factors in EMT-related signature, whereas in DDR-related signature only risk score (HR = 4.12, 95% CI = 2.16-7.86, p < 0.001 and HR = 4.39, 95% CI = 2.22-8.70, $p < 0.001$, respectively) remained as independent prognostic factor for OS. Similarly, in TCGA-PAAD cohort, univariate and multivariate analyses identified only risk scores as independent prognostic factors for OS in EMT (HR = 5.15, 95% CI = 2.95-9.00, p < 0.001 and HR = 4.73, 95% CI = 2.62-8.54, $p < 0.001$, respectively) and DDR (HR = 4.62, 95% CI = 2.77-7.71, p < 0.001 and HR = 4.30, 95% CI = 2.45-7.54, $p < 0.001$, respectively) related signatures. Similar results were identified in validation cohort of PAAD-GSE79668, where only risk score was found as independent prognostic factor for OS in EMT (HR = 1.05, 95% CI = 1.02-1.07, p = 0.001 and HR = 1.09, 95% CI = 1.05-1.14, $p < 0.001$, respectively) and DDR (HR = 1.11, 95% CI = 1.03-1.19, $p < 0.005$ and HR = 1.11, 95% CI = 1.02-1.20, $p = 0.017$, respectively) related signature (see Table 4.2). Furthermore, in univariate and multivariate analysis, clinical factor age along with risk

score remained as independent prognostic factors for OS in EMT (age: HR = 1.89, 95% CI = 1.27-2.81, $p = 0.002$; risk score: HR = 4.86, 95% CI = 3.09-7.63, $p < 0.001$) and DDR (age: HR = 2.01, 95% CI = 1.34-3.02, $p = 0.001$; risk score: HR = 3.89, 95% CI = 2.21-6.82, $p < 0.001$) related gene signatures in STAD cohort, whereas in the validation cohort of STAD-GSE84433, for EMT-related prognostic signature, the clinical variables such as age (univariate: $p < 0.001$; multivariate: $p < 0.001$), T stage (univariate: $p < 0.001$; multivariate: $p = 0.001$), N stage (univariate: $p < 0.001$; multivariate: $p = 0.003$) and risk score (univariate: $p < 0.001$; multivariate: $p = 0.001$) was noted as independent prognostic factors, similarly for DDR-related gene signature, the same variables were observed as independent prognostic factors but with slight different p-values only for N stage (univariate: $p < 0.001$; multivariate: $p = 0.002$) and risk score (univariate: $p < 0.001$; multivariate: $p = 0.006$) (see Table 4.2). The finding suggested that risk scores derived from EMT- and DDR-related prognostic signatures remained as independent prognostic factors for OS across all six cancers (except COAD for DDR signature) in the training cohort (TCGA) and were reported as a robust OS predictor of patient outcomes. By contrast, in the validation cohorts, the risk score was identified as an independent prognostic factor for OS only in BRCA, LUAD, PAAD, and STAD. Overall, it is concluded that identified EMT- and DDR-related prognostic gene signatures could be effective biomarkers in these six cancers.

4.4. Kaplan-Meier Curves for EMT- and DDR-Derived Risk Groups and Clinicopathological Subgroups

Based on the risk groups of EMT and DDR-related gene signatures, the patients were divided into subgroups both in training (TCGA) and validation (GEO) cohorts according to age (< 60 and ≥ 60), TNM stage (I-II and III-IV), N stage (N0 and N1-3), T stage (T1-T2 and T3-T4), and M stage (M0 and M1). The results indicated that patients in the high-risk group possess worse OS as compared to low-risk group for EMT-related gene signature in training TCGA-BRCA cohort in <60 ($p < 0.001$), ≥60 ($p < 0.001$), TMN stage I-II ($p < 0.001$), TNM stage III-IV ($p < 0.001$), N0 ($p < 0.001$), N1-N3 ($p < 0.001$), T1-T2 ($p < 0.001$), T3-T4 ($p < 0.001$), M0 ($p < 0.001$) and M1 ($p = 0.001$) subgroups. Similarly, for DDR-related gene signature the results revealed the worse OS in high-risk group such as < 60 ($p = 0.002$), ≥ 60 ($p < 0.001$), TNM stage I-II ($p < 0.001$), TNM stage III-IV ($p < 0.001$), N0 ($p < 0.001$), N1-N3 ($p < 0.001$), T1-T2 ($p <$

0.001), T3-T4 ($p = 0.001$), M0 ($p < 0.001$) and M1 ($p = 0.002$) for TCGA-BRCA cohort (see Figure 4.7a,b).

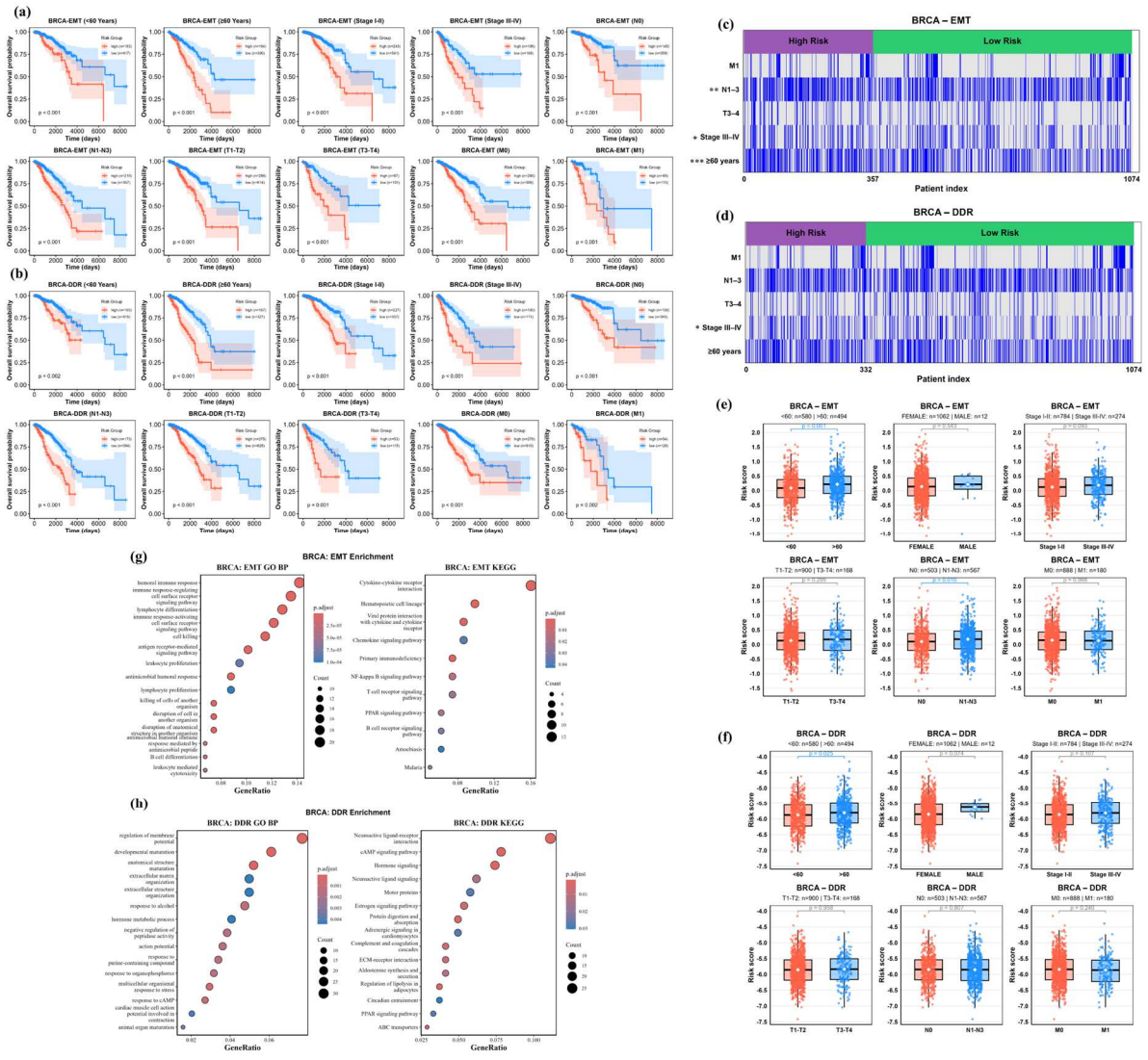


Figure 4.7. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-BRCA cohort. (a, b) Kaplan-Meier survival analysis of subgroups stratified by age, TNM stage, T stage, N stage, and M stage, illustrating the prognostic value of EMT- and DDR-related signatures. Heatmap (c, d) plots of clinical parameters and prognostic EMT- and DDR-derived risk scores. The relationships between risk scores and clinicopathological characteristics (e, f) explained significant correlations for both EMT- and DDR-related signature models. Functional enrichment analysis (KEGG and GO-BP) for low-

and high-risk groups in prognostic EMT and DDR signature models (g, h). *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$)

In the validation cohort for BRCA, high-risk group presented the worse OS in < 60 ($p < 0.001$), ≥ 60 ($p < 0.001$), TNM stage I-II ($p < 0.001$), N0 ($p < 0.001$), N1-N3 ($p < 0.001$), T1-T2 ($p < 0.001$), T3-T4 ($p = 0.007$), and M0 ($p < 0.001$) for EMT and < 60 ($p = 0.006$), ≥ 60 ($p = 0.006$), TNM stage I-II ($p < 0.001$), N0 ($p = 0.010$), N1-N3 ($p = 0.001$), T1-T2 ($p < 0.001$), and M0 ($p < 0.001$) for DDR-related gene signature (see Supplementary File 3, Figure S7a,b). Like the BRCA cohort, other cancer cohorts, including COAD, showed similar results for all clinical variables, except for TNM stage I-II and T1-T2 stage, which were not statistically significant for the EMT-related signature. In contrast, for DDR-related signature the clinical subgroups including < 60 ($p = 0.001$), ≥ 60 ($p = 0.013$), N1-N3 ($p = 0.009$), T1-T2 ($p < 0.031$), T3-T4 ($p = 0.002$), and M0 ($p = 0.012$) showed worse OS in high risk groups compared to low-risk group (see Figure 4.8a,b). By contrast, in the validation cohort of COAD-GSE39582, the high-risk group showed the worse OS only in < 60 ($p = 0.041$) and TNM stage III-IV ($p = 0.036$) subgroups for EMT-related gene signature; similarly, in DDR signature, ≥ 60 ($p < 0.001$), TNM stage I-II ($p = 0.006$), TNM stage III-IV ($p = 0.005$), N0 ($p < 0.001$), N1-N3 ($p = 0.017$), T3-T4 ($p < 0.001$), and M0 ($p < 0.001$) subgroups presented the worse OS (see Supplementary File 3, Figure S7g,h).

Whereas for TCGA-LUAD cohort, all clinical sub-characteristics were strongly significant for OS in EMT- and DDR-related gene signatures (see Figure 4.9a,b). The results of validation cohort (LUAD-GSE50081) were consistent with the training cohort, where all clinical sub-characteristics were strongly significant for OS for high-risk group ($p < 0.001$) in both gene signatures (see Supplementary File 3, Figure S8a,b). Furthermore, Kaplan-Meier curves for TCGA-LUSC cohort revealed that in all subgroups, patients in the high-risk group showed worse OS, except for TNM stage I-II, T3-T4, and M1 stage for the EMT-related gene signature; similar trends were also observed for the DDR-related signature (see Figure 4.10a,b). In the validation cohort of LUSC-GSE30219, patients in the high-risk group presented worse OS in TMN stage I-II ($p = 0.017$), N0 ($p < 0.011$), T1-T2 ($p = 0.024$), and M0 ($p = 0.027$) for EMT, and TMN stage I-II ($p = 0.045$) and M0 ($p = 0.039$) for DDR-related prognostic signature (see Supplementary File 3, Figure S8g,h).

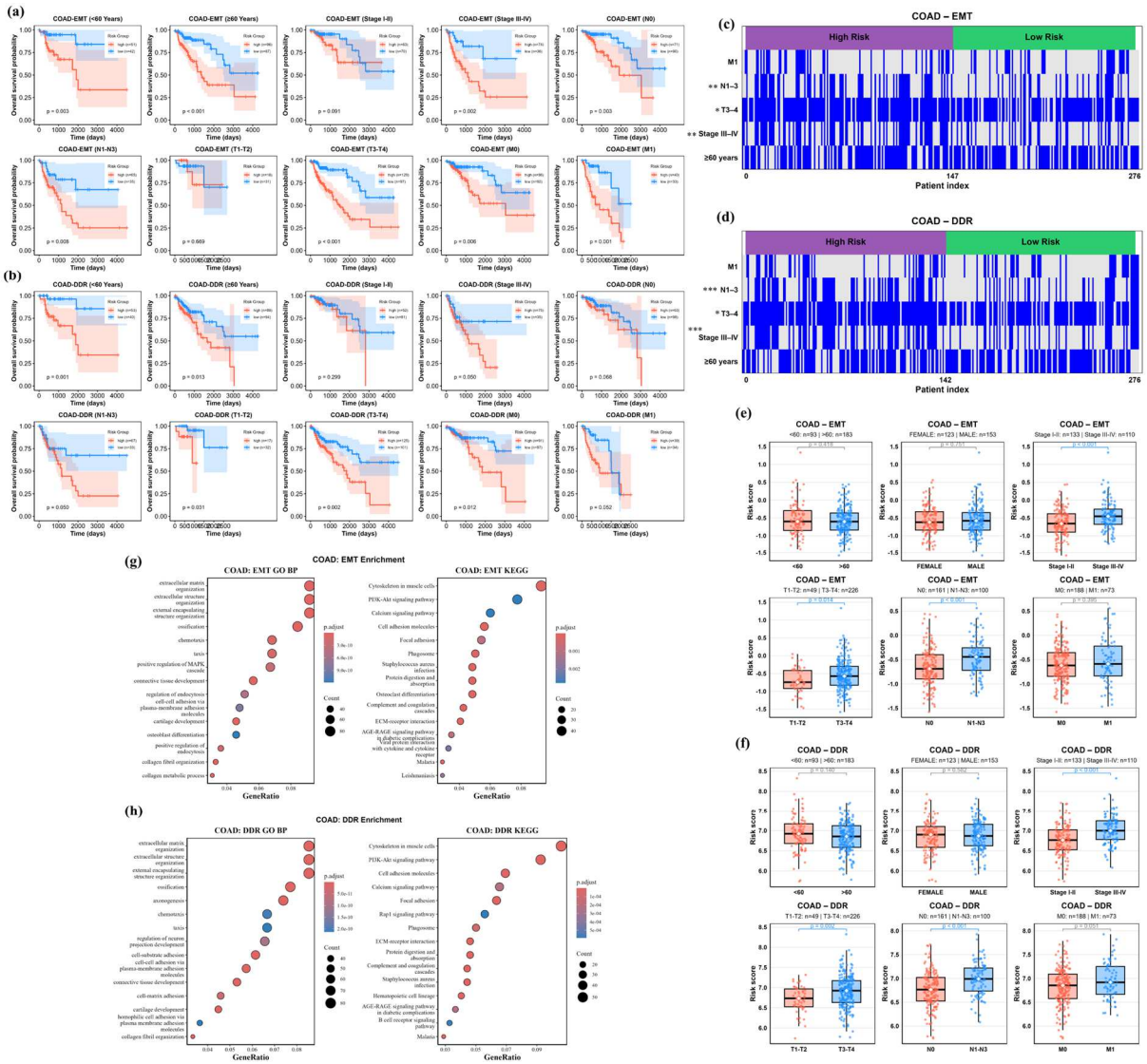


Figure 4.8. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-COAD cohort. (a, b) Kaplan-Meier survival analysis of subgroups stratified by age, TNM stage, T stage, N stage, and M stage, illustrating the prognostic value of EMT- and DDR-related signatures. Heatmap (c, d) plots of clinical parameters and prognostic EMT- and DDR-derived risk scores. The relationships between risk scores and clinicopathological characteristics (e, f) explained significant correlations for both EMT- and DDR-related signature models. Functional enrichment analysis (KEGG and GO-BP) for low- and high-risk groups in prognostic EMT and DDR signature models (g, h). *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$)

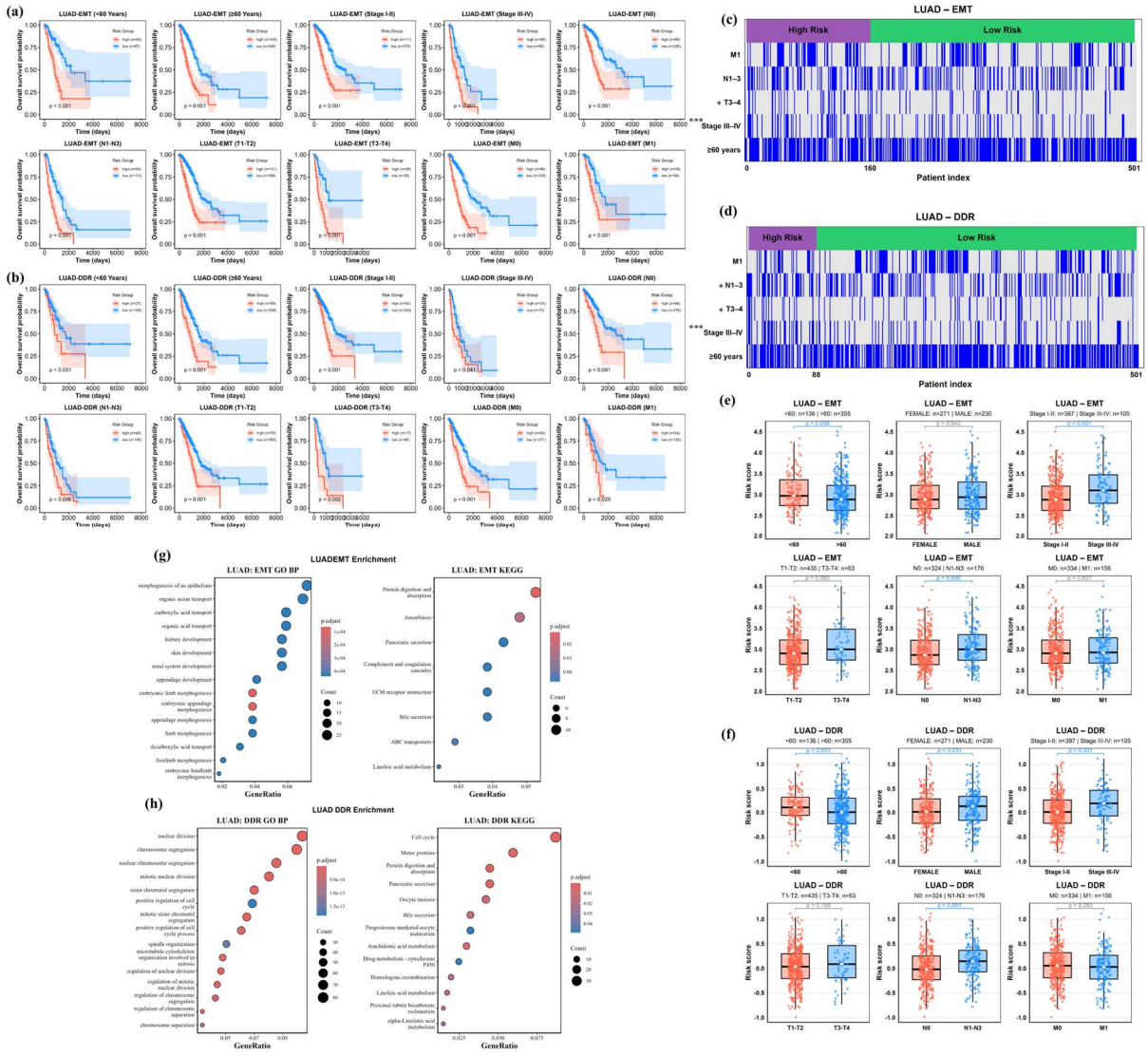


Figure 4.9. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-LUAD cohort. (a, b) Kaplan-Meier survival analysis of subgroups stratified based on age, TNM stage, T stage, N stage, and M stage, illustrating the prognostic values of EMT- and DDR-related signatures. Heatmap (c, d) plots of clinical parameters and prognostic EMT- and DDR-derived risk scores. The relationships between risk scores and clinicopathological characteristics (e, f) explained significant correlations for both EMT- and DDR-related signature models. Functional enrichment analysis (KEGG and GO-BP) for low- and high-risk groups in prognostic EMT and DDR signature models (g, h). *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$)

For TCGA-PAAD cohort, patients in high-risk groups showed significant worse OS for EMT-related signature in subgroups of < 60 ($p = 0.002$), ≥ 60 ($p < 0.001$), TMN stage I-II ($p < 0.001$), N0 ($p = 0.009$), N1-N3 ($p < 0.001$), T1-T2 ($p < 0.001$), T3-T4 ($p < 0.001$), M0 ($p = 0.004$) and M1 ($p < 0.001$); similar trend was noted for DDR-related gene signature. In TCGA-PAAD, since TNM stage III-IV subgroup had very few patients, it was excluded from the results (see Figure 4.11a,b). In TCGA-STAD, patients in high-risk groups in subgroups TNM stage I-II ($p = 0.043$) and M1 ($p = 0.044$) had poorer OS compared to low-risk groups based on the EMT-related prognostic signature; by contrast, in other subgroups, such as < 60 ($p < 0.001$), ≥ 60 ($p < 0.001$), TNM stage III-IV ($p < 0.001$), N1-N3 ($p < 0.001$), T3-T4 ($p < 0.001$) and M0 ($p < 0.001$) we observed significant worse OS in high-risk groups compared to low-risk groups (see Figure 4.12a). For DDR-related prognostic signature, poorer OS in high-risk groups was observed in TNM stage I-II ($p < 0.015$), while a worse OS was noted in < 60 ($p = 0.003$), ≥ 60 ($p < 0.001$), TNM stage III-IV ($p < 0.001$), N1-N3 ($p < 0.001$), T3-T4 ($p < 0.001$), and M0 ($p < 0.001$) in high-risk groups in STAD cohort (see Figure 4.12b). Likewise, in training cohort (TCGA), the results were consistent in the validation cohort of PAAD-GSE79668 and STAD-GSE84433, where patients in high-risk groups showed significantly worse OS in most of the clinical sub-groups (see Supplementary File 3, Figure S9a,b,g,h).

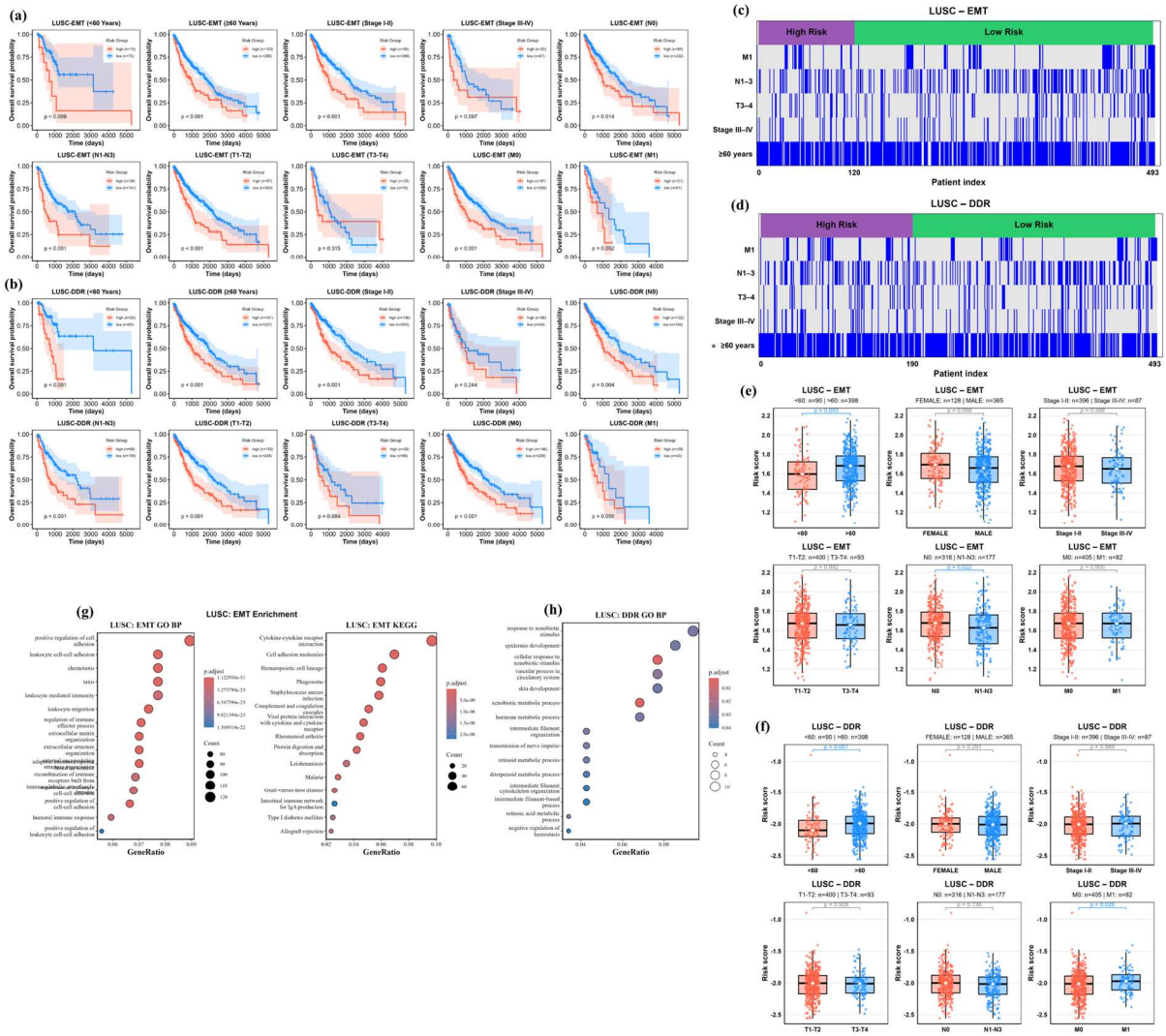


Figure 4.10. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-LUSC cohort. (a, b) Kaplan-Meier survival analysis of subgroups stratified based on age, TNM stage, T stage, N stage, and M stage, illustrating the prognostic values of EMT- and DDR-related signatures. Heatmap (c, d) plots of clinical parameters and prognostic EMT- and DDR-derived risk scores. The relationships between risk scores and clinicopathological characteristics (e, f) explained significant correlations for both EMT- and DDR-related signature models. Functional enrichment analysis (KEGG and GO-BP) for low- and high-risk groups in prognostic EMT and DDR signature models (g, h). *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$)

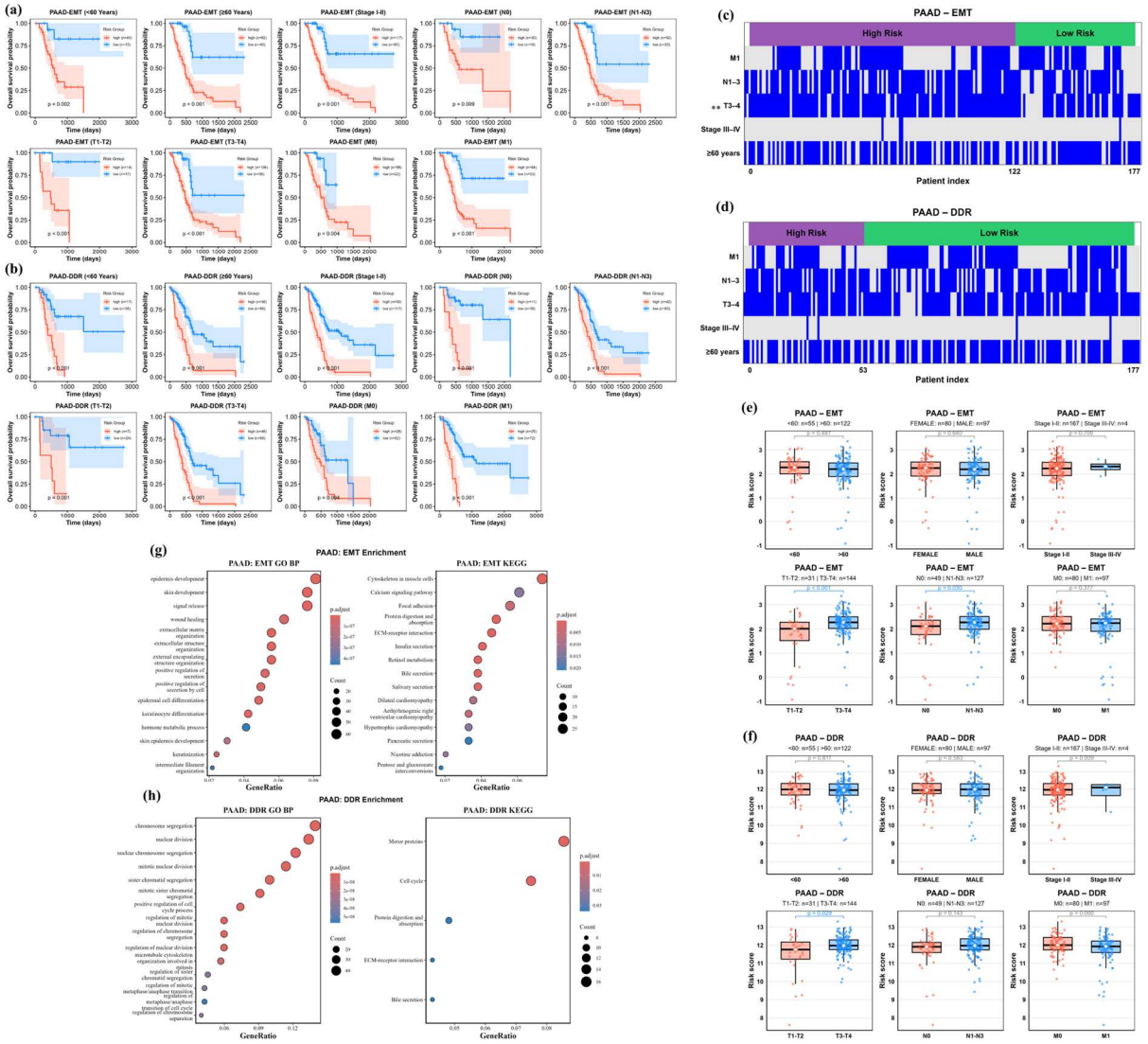


Figure 4.11. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-PAAD cohort. (a, b) Kaplan-Meier survival analysis of subgroups stratified based on age, TNM stage, T stage, N stage, and M stage, illustrating the prognostic values of EMT- and DDR-related signatures. Heatmap (c, d) plots of clinical parameters and prognostic EMT- and DDR-derived risk scores. The relationships between risk scores and clinicopathological characteristics (e, f) explained significant correlations for both EMT- and DDR-related signature models. Functional enrichment analysis (KEGG and GO-BP) for low- and high-risk groups in prognostic EMT and DDR signature models (g, h). *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$)

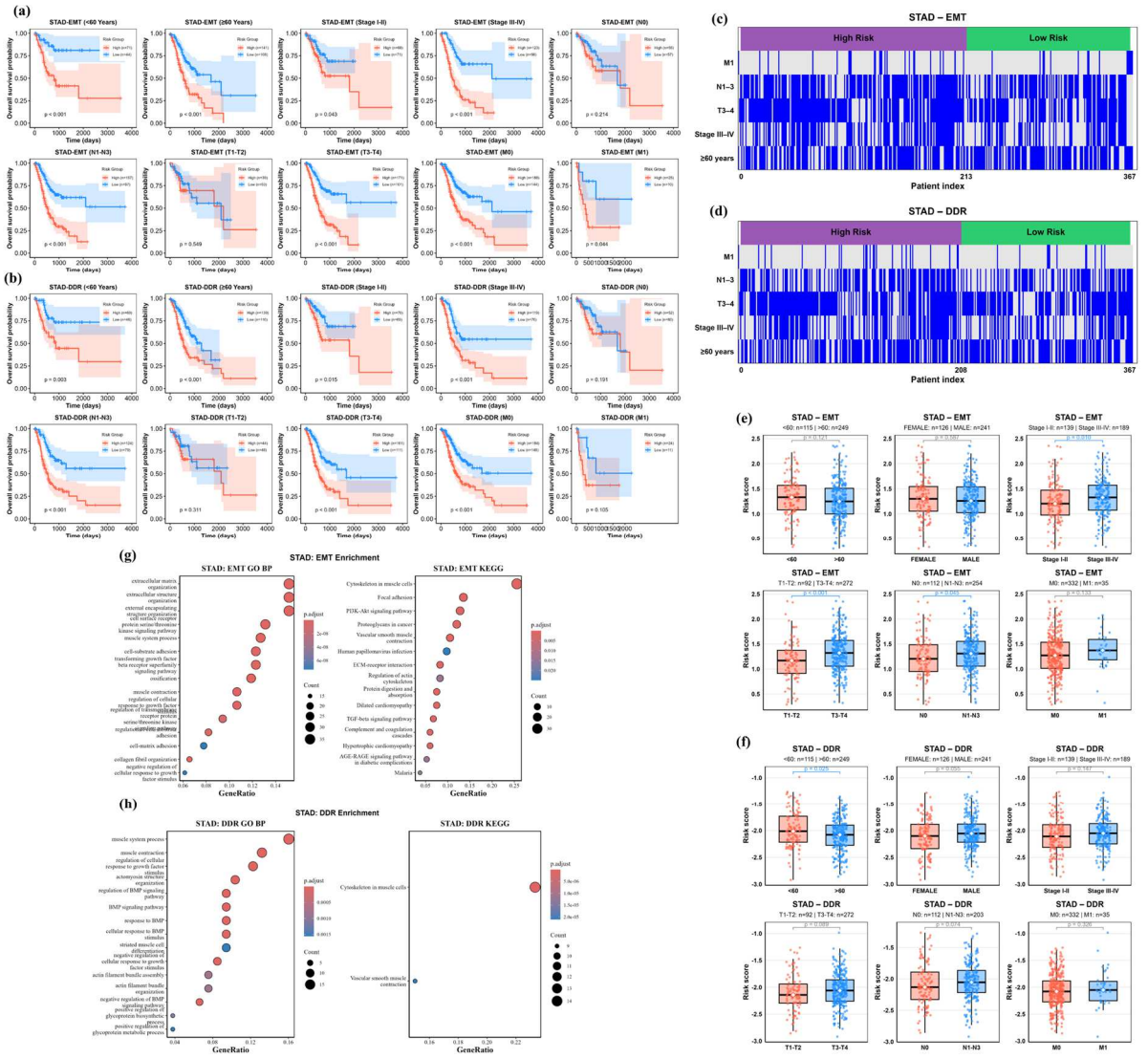


Figure 4.12. Relationship between EMT- and DDR-based risk groups and clinical characteristics factors in the TCGA-STAD cohort. (a, b) Kaplan-Meier survival analysis of subgroups stratified based on age, TNM stage, T stage, N stage, and M stage, illustrating the prognostic values of EMT- and DDR-related signatures. Heatmap (c, d) plots of clinical parameters and prognostic EMT- and DDR-derived risk scores. The relationships between risk scores and clinicopathological characteristics (e, f) explained significant correlations for both EMT- and DDR-related signature models. Functional enrichment analysis (KEGG and GO-BP) for low- and high-risk groups in prognostic EMT and DDR signature models (g, h). *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$)

4.5. Relationship Between Clinicopathological Characteristics and EMT-/DDR-Derived Risk Groups

To further explore the correlation between EMT- and DDR-related risk scores and clinical parameters, we compared clinical characteristics across high- and low-risk groups in all six cancer TCGA and GEO validation cohorts via heatmap and box plots. Heatmap analysis revealed that the high-risk group possesses more cases in ≥ 60 ($p < 0.001$), TNM stage III-IV ($p = 0.024$), and N1-N3 ($p = 0.005$) for EMT-related signature and only TNM stage III-IV ($p = 0.024$) for DDR-related signature in the BRCA TCGA cohort (see Table 4.3; Figure 4.7c,d). In BRCA-GSE20685 validation cohort, ≥ 60 ($p = 0.019$), TNM stage III-IV ($p = 0.014$), and T3-T4 ($p = 0.001$) showed the higher percentage of cases for EMT-related signature in high-risk group, while no clinical factor presented a higher significant percentage of cases for high-risk group for DDR-related signature (see Table 4.4; Supplementary File 3, Figure S7c,d). Similarly, in TCGA-COAD cohort, there were more cases in TNM stage III-IV, T3-T4, and N1-N3 in the high-risk group compared to low-risk group for EMT- (TNM stage III-IV, $p = 0.003$; T3-T4, $p = 0.015$; N1-N3, $p = 0.002$) and DDR- (TNM stage III-IV, $p < 0.001$; T3-T4, $p = 0.014$; N1-N3, $p < 0.001$) related signatures (see Table 4.3; Figure 4.8c,d). In the validation cohort of COAD-GSE39582, the high-risk group shows no significant difference in cases compared to the low-risk group for both EMT- and DDR-related signatures (see Table 4.4; Supplementary File 3, Figure S7i,j).

Table 4.3. Clinicopathological characteristics of low- and high-risk groups across six cancers in training cohort (TCGA).

Characteristics level	Whole cohort	Low risk	High risk	*P* value	Whole cohort	Low risk	High risk	*P* value
	EMT-related signature				DDR-related signature			
TCGA-BRCA								
Age								
<60	580 (54.0%)	417 (58.2%)	163 (45.7%)	**< 0.001**	580 (54.0%)	415 (55.9%)	165 (49.7%)	0.068
≥60	494 (46.0%)	300 (41.8%)	194 (54.3%)		494 (46.0%)	327 (44.1%)	167 (50.3%)	
TNM stage								
Stage I-II	784 (74.1%)	541 (76.3%)	243 (69.6%)	**0.024**	784 (74.1%)	557 (76.2%)	227 (69.4%)	**0.024**
Stage III-IV	274 (25.9%)	168 (23.7%)	106 (30.4%)		274 (25.9%)	174 (23.8%)	100 (30.6%)	
T stage								
T1-T2	900 (84.3%)	614 (85.9%)	286 (81.0%)	0.050	900 (84.3%)	625 (84.5%)	275 (83.8%)	0.869
T3-T4	168 (15.7%)	101 (14.1%)	67 (19.0%)		168 (15.7%)	115 (15.5%)	53 (16.2%)	
N stage								
N0	503 (47.0%)	358 (50.1%)	145 (40.8%)	**0.005**	503 (47.0%)	345 (46.7%)	158 (47.7%)	0.801
N1-N3	567 (53.0%)	357 (49.9%)	210 (59.2%)		567 (53.0%)	394 (53.3%)	173 (52.3%)	
M stage								
M0	888 (83.1%)	598 (83.9%)	290 (81.7%)	0.418	888 (83.1%)	610 (82.9%)	278 (83.7%)	0.797
M1	180 (16.9%)	115 (16.1%)	65 (18.3%)		180 (16.9%)	126 (17.1%)	54 (16.3%)	

TCGA-COAD								
Age								
<60	93 (33.7%)	42 (32.6%)	51 (34.7%)	0.805	93 (33.7%)	40 (29.9%)	53 (37.3%)	0.236
≥60	183 (66.3%)	87 (67.4%)	96 (65.3%)		183 (66.3%)	94 (70.1%)	89 (62.7%)	
TNM stage								
Stage I-II	133 (54.7%)	70 (66.0%)	63 (46.0%)	**0.003**	133 (54.7%)	81 (69.8%)	52 (40.9%)	**< 0.001**
Stage III-IV	110 (45.3%)	36 (34.0%)	74 (54.0%)		110 (45.3%)	35 (30.2%)	75 (59.1%)	
T stage								
T1-T2	49 (17.8%)	31 (24.2%)	18 (12.2%)	**0.015**	49 (17.8%)	32 (24.1%)	17 (12.0%)	**0.014**
T3-T4	226 (82.2%)	97 (75.8%)	129 (87.8%)		226 (82.2%)	101 (75.9%)	125 (88.0%)	
N stage								
N0	161 (61.7%)	90 (72.0%)	71 (52.2%)	**0.002**	161 (61.7%)	98 (74.8%)	63 (48.5%)	**< 0.001**
N1-N3	100 (38.3%)	35 (28.0%)	65 (47.8%)		100 (38.3%)	33 (25.2%)	67 (51.5%)	
M stage								
M0	188 (72.0%)	92 (73.6%)	96 (70.6%)	0.687	188 (72.0%)	97 (74.0%)	91 (70.0%)	0.555
M1	73 (28.0%)	33 (26.4%)	40 (29.4%)		73 (28.0%)	34 (26.0%)	39 (30.0%)	
TCGA-LUAD								
Age								
<60	136 (27.7%)	87 (26.1%)	49 (31.0%)	0.307	136 (27.7%)	109 (26.9%)	27 (31.4%)	0.477
≥60	355 (72.3%)	246 (73.9%)	109 (69.0%)		355 (72.3%)	296 (73.1%)	59 (68.6%)	
TNM stage								
Stage I-II	387 (78.7%)	276 (83.1%)	111 (69.4%)	**< 0.001**	387 (78.7%)	333 (82.2%)	54 (62.1%)	**< 0.001**
Stage III-IV	105 (21.3%)	56 (16.9%)	49 (30.6%)		105 (21.3%)	72 (17.8%)	33 (37.9%)	
T stage								
T1-T2	435 (87.3%)	304 (89.7%)	131 (82.4%)	**0.033**	435 (87.3%)	365 (88.8%)	70 (80.5%)	0.051
T3-T4	63 (12.7%)	35 (10.3%)	28 (17.6%)		63 (12.7%)	46 (11.2%)	17 (19.5%)	
N stage								
N0	324 (64.8%)	228 (67.1%)	96 (60.0%)	0.150	324 (64.8%)	276 (67.0%)	48 (54.5%)	**0.036**
N1-N3	176 (35.2%)	112 (32.9%)	64 (40.0%)		176 (35.2%)	136 (33.0%)	40 (45.5%)	
M stage								
M0	334 (68.2%)	235 (70.6%)	99 (63.1%)	0.118	334 (68.2%)	271 (67.2%)	63 (72.4%)	0.417
M1	156 (31.8%)	98 (29.4%)	58 (36.9%)		156 (31.8%)	132 (32.8%)	24 (27.6%)	
TCGA-LUSC								
Age								
<60	90 (18.4%)	75 (20.3%)	15 (12.7%)	0.088	90 (18.4%)	65 (21.5%)	25 (13.4%)	**0.034**
≥60	398 (81.6%)	295 (79.7%)	103 (87.3%)		398 (81.6%)	237 (78.5%)	161 (86.6%)	
TNM stage								
Stage I-II	396 (82.0%)	298 (81.6%)	98 (83.1%)	0.835	396 (82.0%)	250 (83.6%)	146 (79.3%)	0.288
Stage III-IV	87 (18.0%)	67 (18.4%)	20 (16.9%)		87 (18.0%)	49 (16.4%)	38 (20.7%)	
T stage								
T1-T2	400 (81.1%)	303 (81.2%)	97 (80.8%)	1.000	400 (81.1%)	245 (80.9%)	155 (81.6%)	0.936
T3-T4	93 (18.9%)	70 (18.8%)	23 (19.2%)		93 (18.9%)	58 (19.1%)	35 (18.4%)	
N stage								
N0	316 (64.1%)	232 (62.2%)	84 (70.0%)	0.150	316 (64.1%)	194 (64.0%)	122 (64.2%)	1.000
N1-N3	177 (35.9%)	141 (37.8%)	36 (30.0%)		177 (35.9%)	109 (36.0%)	68 (35.8%)	
M stage								
M0	405 (83.2%)	308 (83.5%)	97 (82.2%)	0.858	405 (83.2%)	259 (85.8%)	146 (78.9%)	0.067
M1	82 (16.8%)	61 (16.5%)	21 (17.8%)		82 (16.8%)	43 (14.2%)	39 (21.1%)	
TCGA-PAAD								
Age								
<60	55 (31.1%)	15 (27.3%)	40 (32.8%)	0.577	55 (31.1%)	38 (30.6%)	17 (32.1%)	0.991
≥60	122 (68.9%)	40 (72.7%)	82 (67.2%)		122 (68.9%)	86 (69.4%)	36 (67.9%)	
TNM stage								
Stage I-II	167 (97.7%)	50 (98.0%)	117 (97.5%)	1.000	167 (97.7%)	117 (98.3%)	50 (96.2%)	0.755
Stage III-IV	4 (2.3%)	1 (2.0%)	3 (2.5%)		4 (2.3%)	2 (1.7%)	2 (3.8%)	
T stage								
T1-T2	31 (17.7%)	17 (32.1%)	14 (11.5%)	**0.002**	31 (17.7%)	24 (19.7%)	7 (13.2%)	0.416
T3-T4	144 (82.3%)	36 (67.9%)	108 (88.5%)		144 (82.3%)	98 (80.3%)	46 (86.8%)	
N stage								
N0	49 (27.8%)	19 (35.2%)	30 (24.6%)	0.206	49 (27.8%)	38 (30.9%)	11 (20.8%)	0.233
N1-N3	127 (72.2%)	35 (64.8%)	92 (75.4%)		127 (72.2%)	85 (69.1%)	42 (79.2%)	
M stage								
M0	80 (45.2%)	22 (40.0%)	58 (47.5%)	0.441	80 (45.2%)	52 (41.9%)	28 (52.8%)	0.242

M1	97 (54.8%)	33 (60.0%)	64 (52.5%)		97 (54.8%)	72 (58.1%)	25 (47.2%)	
TCGA-STAD								
Age								
<60	115 (31.6%)	44 (28.9%)	71 (33.5%)	0.421	115 (31.6%)	46 (29.5%)	69 (33.2%)	0.526
≥60	249 (68.4%)	108 (71.1%)	141 (66.5%)		249 (68.4%)	110 (70.5%)	139 (66.8%)	
TNM stage								
Stage I-II	139 (42.4%)	71 (51.8%)	68 (35.6%)	**0.005**	139 (42.4%)	69 (49.6%)	70 (37.0%)	**0.030**
Stage III-IV	189 (57.6%)	66 (48.2%)	123 (64.4%)		189 (57.6%)	70 (50.4%)	119 (63.0%)	
T stage								
T1-T2	92 (25.3%)	53 (34.4%)	39 (18.6%)	**< 0.001**	92 (25.3%)	48 (30.2%)	44 (21.5%)	0.075
T3-T4	272 (74.7%)	101 (65.6%)	171 (81.4%)		272 (74.7%)	111 (69.8%)	161 (78.5%)	
N stage								
N0	112 (30.6%)	57 (37.0%)	55 (25.9%)	**0.031**	112 (30.6%)	60 (38.0%)	52 (25.0%)	**0.011**
N1-N3	254 (69.4%)	97 (63.0%)	157 (74.1%)		254 (69.4%)	98 (62.0%)	156 (75.0%)	
M stage								
M0	332 (90.5%)	144 (93.5%)	188 (88.3%)	0.132	332 (90.5%)	148 (93.1%)	184 (88.5%)	0.189
M1	35 (9.5%)	10 (6.5%)	25 (11.7%)		35 (9.5%)	11 (6.9%)	24 (11.5%)	

In TCGA-LUAD cohort, high-risk group has more cases in TNM stage III-IV and T3-T4 for EMT-related signature (TNM stage III-IV, $p < 0.001$; T3-T4, $p = 0.033$) and in TNM stage III-IV and N1-N3 for DDR-related signature (TNM stage III-IV, $p < 0.001$; N1-N3, $p = 0.036$) (see Table 4.3; Figure 4.9c,d). The results revealed that no clinical parameter had more cases in high-risk groups than in low-risk groups in TCGA-LUSC for the EMT-related signature, while only > 60 ($p < 0.034$) clinical factor had more cases in high-risk groups for the DDR-related signature (see Table 4.3; Figure 4.10c,d). In the validation cohort of LUAD-GSE50081, only N1-N3 presented a significant ($p = 0.033$) difference in high-risk group for DDR-related signature (see Table 4.4; Supplementary File 3, Figure S8c,d). Similar, results were observed in LUSC-GSE30219, where the number of cases were almost equally distributed in low- and high-risk groups for both EMT- and DDR-related signatures (see Table 4.4; Supplementary File 3, Figure S7i,j). Similarly, results were also noted for PAAD, where only clinical level T3-T4 has more cases in a high-risk group for EMT-related signature in the training cohort (TCGA) (see Table 4.3; Figure 4.11c,d) and no clinical factor presented the significant percentage of cases for high-risk group in validation cohort (see Table 4.4; Supplementary File 3, Figure S9c,d). In STAD, more cases were noted in high-risk group in TNM stage III-IV ($p = 0.005$), T3-T4 ($p < 0.001$), and N1-N3 ($p < 0.031$) stages for EMT-related signature in the training cohort (TCGA) (see Figure 4.12c). In contrast, in the validated cohort, T3-T4 ($p = 0.003$) and N1-N3 ($p = 0.008$) clinical factors showed higher cases in the high-risk group compared to the low-risk group (see Table 4.4; Supplementary File 3, Figure S9i). However, for DDR-related signature TNM stage III-IV ($p = 0.030$) and N1-N3 ($p = 0.01$) exhibit more cases in high-risk group compared to low-risk group in the training cohort (TCGA) (see Table 4.3; Figure 4.12c,d), and N1-N3 ($p = 0.01$) in the validation cohort (GEO) (see Table 4.4; Supplementary

File 3, Figure S9j). The results reported that the patients in the high-risk group were significantly more likely to have advanced tumour stage (stage III-IV), distant metastasis (M1), tumour invasion (T3-T4) and lymph node (N1-N3) especially in BRCA, COAD, LUAD, and STAD cancers for both signatures in the training cohort, while in the validation cohort, due to low number of samples, in most of the cancers (COAD, LUAD, LUSC, and PAAD) there were no significant difference in percentage of cases between high- and low-risk groups.

Table 4.4. Clinicopathological characteristics of low- and high-risk groups across six cancers in validation cohorts (GEO).

Characteristics level	Whole cohort	Low risk	High risk	*P* value	Whole cohort	Low risk	High risk	*P* value
	EMT-related signature				DDR-related signature			
BRCA-GSE20685								
Age								
<60	278 (85.0%)	255 (86.7%)	23 (69.7%)	**0.019**	278 (85.0%)	248 (85.5%)	30 (81.1%)	0.640
≥60	49 (15.0%)	39 (13.3%)	10 (30.3%)		49 (15.0%)	42 (14.5%)	7 (18.9%)	
TNM stage								
Stage I-II	311 (95.1%)	283 (96.3%)	28 (84.8%)	**0.014**	311 (95.1%)	276 (95.2%)	35 (94.6%)	1.000
Stage III-IV	16 (4.9%)	11 (3.7%)	5 (15.2%)		16 (4.9%)	14 (4.8%)	2 (5.4%)	
T stage								
T1-T2	289 (88.4%)	266 (90.5%)	23 (69.7%)	**0.001**	289 (88.4%)	256 (88.3%)	33 (89.2%)	1.000
T3-T4	38 (11.6%)	28 (9.5%)	10 (30.3%)		38 (11.6%)	34 (11.7%)	4 (10.8%)	
N stage								
N0	137 (41.9%)	125 (42.5%)	12 (36.4%)	0.622	137 (41.9%)	121 (41.7%)	16 (43.2%)	1.000
N1-N3	190 (58.1%)	169 (57.5%)	21 (63.6%)		190 (58.1%)	169 (58.3%)	21 (56.8%)	
M stage								
M0	319 (97.6%)	287 (97.6%)	32 (97.0%)	1.000	319 (97.6%)	283 (97.6%)	36 (97.3%)	1.000
M1	8 (2.4%)	7 (2.4%)	1 (3.0%)		8 (2.4%)	7 (2.4%)	1 (2.7%)	
COAD-GSE39582								
Age								
<60	153 (26.2%)	108 (28.2%)	45 (22.4%)	0.156	153 (26.2%)	128 (26.2%)	25 (26.3%)	1.000
≥60	431 (73.8%)	275 (71.8%)	156 (77.6%)		431 (73.8%)	361 (73.8%)	70 (73.7%)	
TNM stage								
Stage I-II	309 (52.8%)	213 (55.5%)	96 (47.8%)	0.092	309 (52.8%)	266 (54.4%)	43 (44.8%)	0.107
Stage III-IV	276 (47.2%)	171 (44.5%)	105 (52.2%)		276 (47.2%)	223 (45.6%)	53 (55.2%)	
T stage								
T1-T2	61 (10.4%)	45 (11.7%)	16 (8.0%)	0.204	61 (10.4%)	56 (11.5%)	5 (5.2%)	0.099
T3-T4	524 (89.6%)	339 (88.3%)	185 (92.0%)		524 (89.6%)	433 (88.5%)	91 (94.8%)	
N stage								
N0	314 (53.7%)	214 (55.7%)	100 (49.8%)	0.197	314 (53.7%)	270 (55.2%)	44 (45.8%)	0.116
N1-N3	271 (46.3%)	170 (44.3%)	101 (50.2%)		271 (46.3%)	219 (44.8%)	52 (54.2%)	
M stage								
M0	499 (85.3%)	336 (87.5%)	163 (81.1%)	0.051	499 (85.3%)	422 (86.3%)	77 (80.2%)	0.167
M1	86 (14.7%)	48 (12.5%)	38 (18.9%)		86 (14.7%)	67 (13.7%)	19 (19.8%)	
LUAD-GSE50081								
Age								
<60	19 (14.6%)	18 (15.5%)	1 (7.1%)	0.662	19 (14.6%)	5 (11.6%)	14 (16.1%)	0.679
≥60	111 (85.4%)	98 (84.5%)	13 (92.9%)		111 (85.4%)	38 (88.4%)	73 (83.9%)	
TNM stage								
Stage I-II	130 (100.0%)	116 (100.0%)	14 (100.0%)	–	130 (100.0%)	43 (100.0%)	87 (100.0%)	–
T stage								
T1-T2	128 (98.5%)	114 (98.3%)	14 (100.0%)	1.000	128 (98.5%)	43 (100.0%)	85 (97.7%)	0.807
T3-T4	2 (1.5%)	2 (1.7%)	0 (0.0%)		2 (1.5%)	0 (0.0%)	2 (2.3%)	
N stage								
N0	95 (73.1%)	84 (72.4%)	11 (78.6%)	0.864	95 (73.1%)	37 (86.0%)	58 (66.7%)	**0.033**
N1-N3	35 (26.9%)	32 (27.6%)	3 (21.4%)		35 (26.9%)	6 (14.0%)	29 (33.3%)	

M stage								
M0	130 (100.0%)	116 (100.0%)	14 (100.0%)	–	130 (100.0%)	43 (100.0%)	87 (100.0%)	–
LUSC-GSE30219								
Age								
<60	20 (32.8%)	17 (37.8%)	3 (18.8%)	0.279	20 (32.8%)	4 (57.1%)	16 (29.6%)	0.302
≥60	41 (67.2%)	28 (62.2%)	13 (81.2%)		41 (67.2%)	3 (42.9%)	38 (70.4%)	
TNM stage								
Stage I-II	59 (96.7%)	43 (95.6%)	16 (100.0%)	0.968	59 (96.7%)	7 (100.0%)	52 (96.3%)	1.000
Stage III-IV	2 (3.3%)	2 (4.4%)	0 (0.0%)		2 (3.3%)	0 (0.0%)	2 (3.7%)	
T stage								
T1-T2	55 (90.2%)	40 (88.9%)	15 (93.8%)	0.943	55 (90.2%)	7 (100.0%)	48 (88.9%)	0.799
T3-T4	6 (9.8%)	5 (11.1%)	1 (6.2%)		6 (9.8%)	0 (0.0%)	6 (11.1%)	
N stage								
N0	52 (85.2%)	37 (82.2%)	15 (93.8%)	0.480	52 (85.2%)	7 (100.0%)	45 (83.3%)	0.546
N1-N3	9 (14.8%)	8 (17.8%)	1 (6.2%)		9 (14.8%)	0 (0.0%)	9 (16.7%)	
M stage								
M0	61 (100.0%)	45 (100.0%)	16 (100.0%)	–	61 (100.0%)	7 (100.0%)	54 (100.0%)	–
PAAD-GSE79668								
Age								
<60	16 (31.4%)	6 (21.4%)	10 (43.5%)	0.166	16 (31.4%)	3 (60.0%)	13 (28.3%)	0.345
≥60	35 (68.6%)	22 (78.6%)	13 (56.5%)		35 (68.6%)	2 (40.0%)	33 (71.7%)	
TNM stage								
Stage I-II	44 (86.3%)	23 (82.1%)	21 (91.3%)	0.591	44 (86.3%)	5 (100.0%)	39 (84.8%)	0.799
Stage III-IV	7 (13.7%)	5 (17.9%)	2 (8.7%)		7 (13.7%)	0 (0.0%)	7 (15.2%)	
T stage								
T1-T2	15 (29.4%)	8 (28.6%)	7 (30.4%)	1.000	15 (29.4%)	1 (20.0%)	14 (30.4%)	1.000
T3-T4	36 (70.6%)	20 (71.4%)	16 (69.6%)		36 (70.6%)	4 (80.0%)	32 (69.6%)	
N stage								
N0	14 (27.5%)	8 (28.6%)	6 (26.1%)	1.000	14 (27.5%)	1 (20.0%)	13 (28.3%)	1.000
N1-N3	37 (72.5%)	20 (71.4%)	17 (73.9%)		37 (72.5%)	4 (80.0%)	33 (71.7%)	
M stage								
M0	48 (94.1%)	27 (96.4%)	21 (91.3%)	0.860	48 (94.1%)	5 (100.0%)	43 (93.5%)	1.000
M1	3 (5.9%)	1 (3.6%)	2 (8.7%)		3 (5.9%)	0 (0.0%)	3 (6.5%)	
STAD-GSE84433								
Age								
<60	156 (43.7%)	53 (43.8%)	103 (43.6%)	1.000	156 (43.7%)	70 (41.9%)	86 (45.3%)	0.597
≥60	201 (56.3%)	68 (56.2%)	133 (56.4%)		201 (56.3%)	97 (58.1%)	104 (54.7%)	
T stage								
T1-T2	46 (12.9%)	25 (20.7%)	21 (8.9%)	**0.003**	46 (12.9%)	28 (16.8%)	18 (9.5%)	0.058
T3-T4	311 (87.1%)	96 (79.3%)	215 (91.1%)		311 (87.1%)	139 (83.2%)	172 (90.5%)	
N stage								
N0	71 (19.9%)	34 (28.1%)	37 (15.7%)	**0.008**	71 (19.9%)	45 (26.9%)	26 (13.7%)	**0.003**
N1-N3	286 (80.1%)	87 (71.9%)	199 (84.3%)		286 (80.1%)	122 (73.1%)	164 (86.3%)	

Box plot results reported that, in TCGA-BRCA cohort, the patients in clinical parameters of > 60 ($p < 0.001$) and N1-N3 ($p = 0.010$) have a higher risk score for EMT-related signature, while for DDR-related signature only patients in > 60 ($p = 0.025$) have a higher risk score (see Figure 4.7e,f). In TCGA-COAD cohort, a higher risk score was noted in stage III-IV ($p < 0.001$), T3-T4 ($p = 0.014$), and N1-N3 ($p < 0.001$) for EMT-related signature, while stage III-IV ($p < 0.001$), T3-T4 ($p = 0.002$), and N1-N3 ($p < 0.001$) had a higher risk score for DDR-related signature (see Figure 4.8e,f). It was reported that, for EMT-related signature in TCGA-LUAD cohort, the higher risk score was observed in patients with clinical variables of > 60 ($p = 0.038$), stage III-IV ($p < 0.001$), and N1-N3 ($p = 0.006$), while > 60 ($p = 0.003$), male ($p = 0.031$), stage

III-IV ($p < 0.001$), and N1-N3 ($p < 0.001$) had higher risk scores for DDR-related signature (see Figure 4.9e,f). In TCGA-LUSC cohort, > 60 ($p = 0.003$) and N1-N3 ($p = 0.022$) parameters showed a higher risk score for EMT-related signature, and >60 ($p = 0.007$) and M1 ($p = 0.028$) for DDR-related signatures (see Figure 4.10e,f). Similarly, in TCGA-PAAD cohort, T3-T4 ($p < 0.001$) and N1-N3 ($p = 0.030$) parameters have higher risk scores for EMT-related signature, and T3-T4 ($p = 0.029$) for DDR-related signature (see Figure 4.11e,f). Whereas in TCGA-STAD cohort, the higher risk score was observed in clinical parameters of stage III-IV ($p = 0.010$), T3-T4 ($p < 0.001$), and N1-N3 ($p = 0.045$) for EMT-related signature, <60 ($p = 0.025$) for DDR-related signature (see Figure 4.12e,f).

Similar results were noted in validation cohorts, where higher significant risk-scores for EMT-related signature were observed in > 60 ($p = 0.003$) and TNM stage III-IV ($p = 0.012$) in BRCA-GSE20685 (see Supplementary File 3, Figure S7e); TNM stage III-IV ($p = 0.032$) in COAD-GSE39582 (see Supplementary File 3, Figure S7k); T-stage T3-T4 ($p = 0.003$) and N-stage N1-N3 ($p = 0.005$) in STAD-GSE84433 (see Supplementary File 3, Figure S9k). In contrast, for DDR-related signature, in T-stage T3-T4 ($p = 0.033$) presented the highest significant risk score in COAD-GSE39582 (see Supplementary File 3, Figure S7l); N-stage N1-N3 ($p = 0.005$) in LUAD-GSE50081 (see Supplementary File 3, Figure S8f); male ($p = 0.011$), T-stage T3-T4 ($p = 0.009$), and N-stage N1-N3 ($p = 0.036$) in STAD-GSE84433 (see Supplementary File 3, Figure S9l). It is concluded that the higher the tumour stage, the higher the risk score for all six cancers in training and validation cohorts.

4.6. Pathways Enrichment Analysis Highlights Humoral Immune Response, Extracellular Matrix Organization, and Extracellular Structure Organization by EMT- and Mitotic Networks by DDR-Related Prognostic Signatures

To explore the biological significance of the EMT- and DDR-related prognostic signature in all six cancers (BRCA, COAD, LUAD, LUSC, PAAD, and STAD), we performed functional enrichment analysis using GO and KEGG databases comparing high- and low-risk groups. GO-BP results revealed that EMT-related signature was highly enriched in immune related processes, including humoral immune response, immune response-regulating cell surface receptor signalling pathway, and lymphocyte differentiation while in KEGG analysis signalling

related pathways such as cytokine-cytokine receptor interaction, hematopoietic cell lineage, chemokine signalling pathway and NF-κB signalling pathway were most enriched in BRCA (see Figure 4.7g). In COAD, the extracellular matrix organization, extracellular structure organization and external encapsulation structure organization were the top three enriched pathways based on GO-BP, and cytoskeleton in muscle cells, P13K-AKT signalling pathway, calcium signalling pathway, and focal adhesion were most enriched according to KEGG (see Figure 4.8g). In LUAD, GO-BP results revealed significant enrichment in morphogenesis of epithelium, organic anion transport, and skin development, while KEGG results noted significant enrichment of protein digestion and absorption, amoebiasis, ECM-receptor interaction and ABC transporters (see Figure 4.9g). While in the case of LUSC, positive regulation of cell adhesion, leukocyte cell-cell adhesion, extracellular matrix organization and extracellular structure organization were most significant enriched pathways by GO-BP and cytokine-cytokine receptor interaction, cell adhesion molecules, and hematopoietic cell lineage were the top three most enriched pathways by KEGG analysis (see Figure 4.10g). Furthermore, epidermis development, skin development, signal release, and extracellular matrix organization pathways are most significant by GO analysis and cytoskeleton in muscle cells, calcium signalling pathway, focal adhesion and ECM-receptor interaction were most enriched in PAAD by KEGG analysis (see Figure 4.11g). Whereas in STAD, extracellular matrix organization, extracellular structure organization, and external encapsulating structure organization are significant top three enrichment pathways by GO-BP analysis along with cytoskeleton in muscle cells, focal adhesion, and P13K-AKT signalling pathway by KEGG analysis (see Figure 4.12g).

For DDR-related prognostic gene signature, GO-BP results reported that regulation of membrane potential, development maturation, and anatomical structure maturation were most significantly enrichment pathways in BRCA cohort (see Figure 4.7h), extracellular matrix organization, extracellular structure organization, and external encapsulating structure organization in COAD cohort (see Figure 4.8h), nuclear division, chromosome segregation, and nuclear chromosome segregation in LUAD cohort (see Figure 4.9h), response to xenobiotic stimulus, epidermis development, and cellular response to xenobiotic stimulus in LUSC cohort (see Figure 4.10h), chromosome segregation, nuclear division, and nuclear chromosome segregation in PAAD cohort (see Figure 4.11h), and muscle system process, muscle

contraction, and regulation of cellular response to growth factor stimulus in STAD cohort (see Figure 4.12h), respectively. Similarly, KEGG analysis revealed that the most significant enrichment of neuroactive and-receptor interaction and cAMP signalling pathway in BRCA (see Figure 4.7h), cytoskeleton in muscle cells and P13K-AKT signalling pathway in COAD (see Figure 4.8h), cell cycle and motor proteins in LUAD (see Figure 4.9h) and PAAD (see Figure 4.11h), and cytoskeleton in muscle cells and vascular smooth muscle contraction in STAD (see Figure 4.12h), respectively.

Our results show that the EMT-related prognostic gene signature is not only significantly enriched in classical epithelial-mesenchymal transition pathways, such as extracellular matrix organization, extracellular structure organization, and cytoskeleton remodelling, but also in immune-related signalling pathways across multiple cancer types. In BRCA and LUSC, EMT-related prognostic signatures strongly aligned with immune-modulatory processes, while in COAD, PAAD, and STAD, they showed significant enrichment in structure reorganization pathways. In contrast, the DDR-related prognostic signature was closely related to cell cycle regulation and genomic stability in LUAD and PAAD. In COAD and STAD, it presented significant enrichment in the structural remodelling process, and in BRCA, it was linked to cellular signalling.

4.7. Combined Prognostic Analysis of EMT- and DDR-Related Risk Scores and Immune Cell Infiltration

We applied the CIBERSORT algorithm to further investigate the relationship between patient risk groups and the immune infiltration ratios of 22 immune cells across all six cancers in both training (TCGA) and validation (GEO) cohorts. The stacked bar plot revealed the heterogeneous distribution of 22 immune cell types in both high- and low-risk groups. However, certain immune cells, such as macrophages, T cells, and B cells, showed clear differences between high- and low-risk samples in all six cancers in both the training (see Figures 4.13-4.18a,b) and validation cohorts (see Supplementary File 3, Figures S10–15a,d). Furthermore, we compared the distribution of each immune cell type between low- and high-risk groups via raincloud graphs. TCGA-BRCA showed enrichment of macrophage M2 ($p < 0.001$) in high-risk group for EMT and of macrophage M0 ($p < 0.001$) for DDR-related signature, while B cells naïve ($p = 0.016$), T cells CD8 ($p < 0.001$), and T cells follicular helper

($p = 0.006$) were significantly enriched in low-risk groups for EMT and B cells naïve ($p < 0.001$), T cells CD8 ($p < 0.001$), and mast cells resting ($p = 0.006$) for DDR-related signature (see Figure 4.13c,d). In the validation cohort of BRCA-GSE20685, mast cell resting ($p = 0.005$) and macrophages M0 ($p < 0.042$) were significantly enriched in high-risk group for EMT-related signature (see Supplementary File 3, Figure S10b), but no significant enrichment was observed for DDR-related signature (see Supplementary File 3, Figure S10e). Notably, macrophage M0 and macrophage M2 ($p < 0.001$) showed high infiltration in the high-risk group in TCGA-COAD cohort for EMT- (macrophage M0, $p = 0.001$; macrophage M2, $p < 0.001$) and DDR- (macrophage M0, $p = 0.002$; macrophage M2, $p < 0.001$) related signatures (see Figure 4.14c,d). Similar, results were noted for validation cohort (COAD-GSE39582), where macrophage M0, macrophage M1, macrophage M2 and neutrophils showed higher infiltration in the high-risk group for EMT- (macrophage M0, $p = 0.001$; macrophage M1, $p = 0.001$; macrophage M2, $p < 0.001$; neutrophils, $p = 0.001$) and DDR-related (macrophage M0, $p = 0.033$; macrophage M2, $p < 0.023$; neutrophils, $p = 0.001$) signatures (see Supplementary File 3, Figure S11b,e).

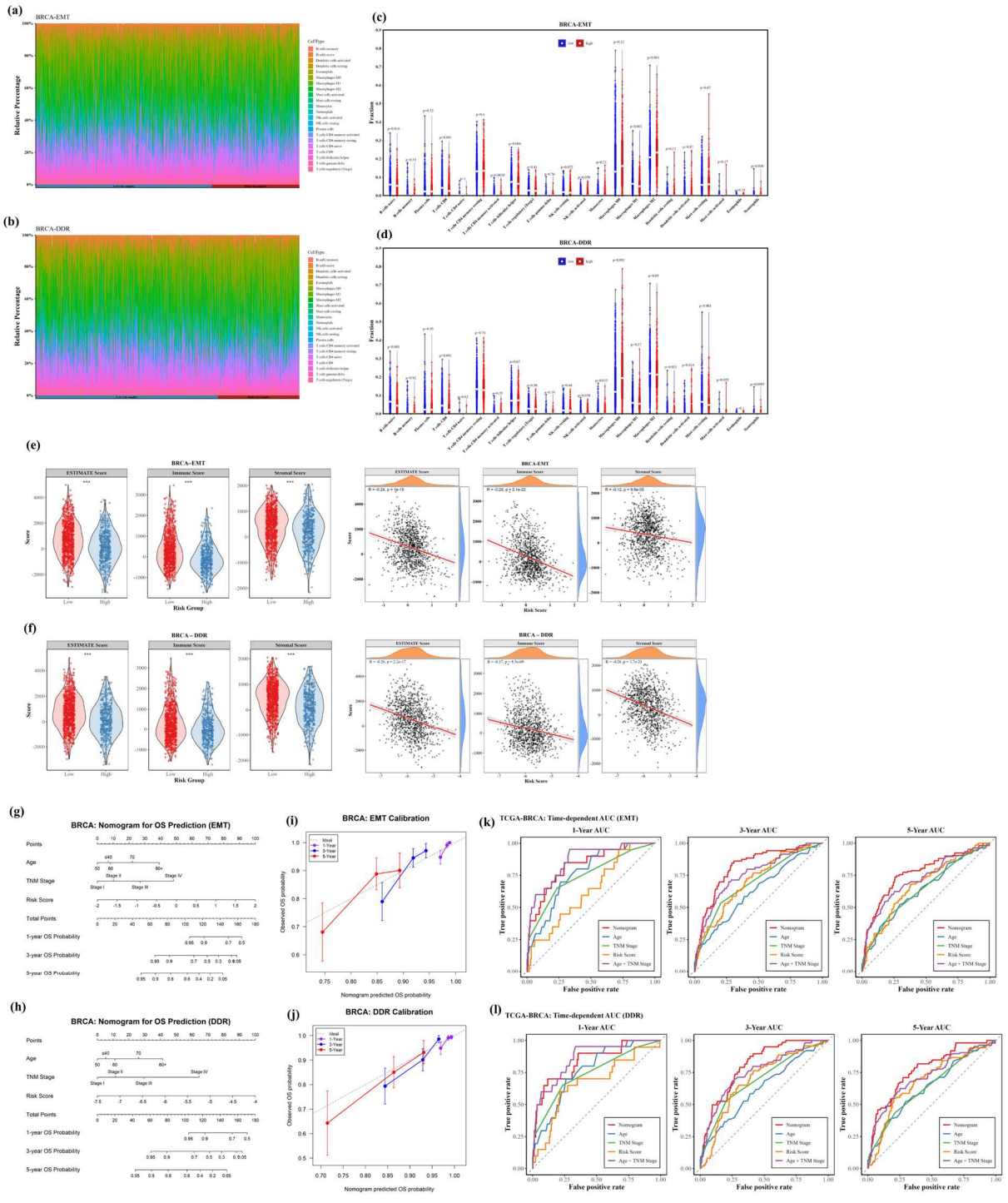


Figure 4.13. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-BRCA cohort. Bar chart plots of 22 immune cells infiltration across high- and low-risk groups of EMT- (a) and DDR-related (b) signatures. Raincloud plots between

EMT- (c) and DDR-based (d) risk groups and 22 kinds of immune cells. (e) ESTIMATE stromal and immune scores between low-risk and high-risk in TCGA-BRCA patients for EMT and DDR-related signatures. (f) Dual-margin scatterplots correlation between risk score and ESTIMATE-derived scores for EMT and DDR-related prognostic signatures in TCGA-BRCA. Construction and validation of nomogram models provide survival predictions based on EMT- (g) and DDR-related (h) signatures. Calibration curves further demonstrated agreement between predicted and observed survival outcomes at all three OS for EMT- (i) and DDR-related (j) signature. Time-dependent ROC curves validate the predictive ability of the nomogram alone, age, TNM stage, risk score, and combined age + TNM stage at 1-, 3-, and 5-year OS for EMT- (k) and DDR-related (l) signature.

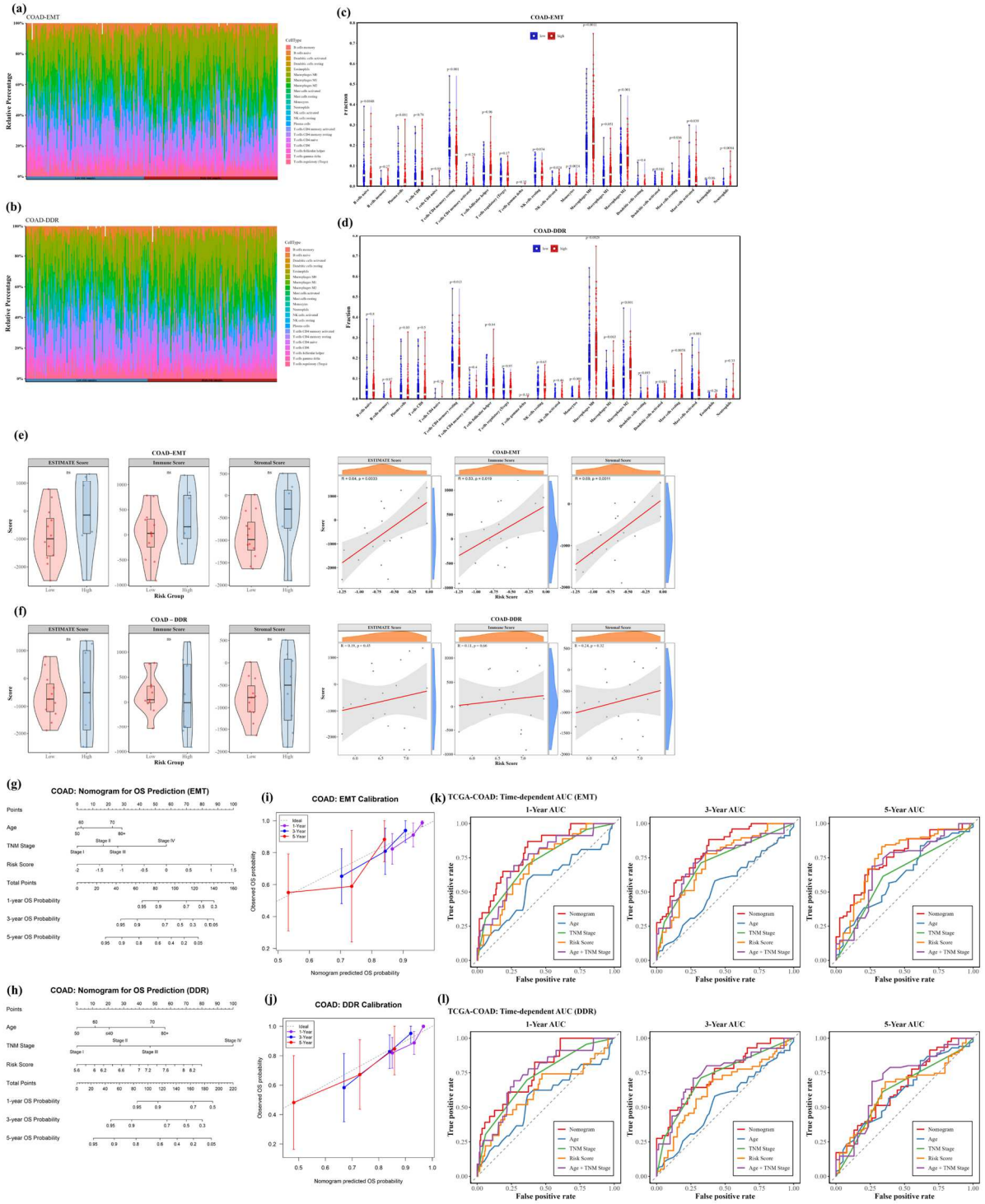


Figure 4.14. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-COAD cohort. Bar chart plots of 22 immune cells infiltration across high- and low-risk groups of EMT- (a) and DDR-related (b) signatures. Raincloud plots between

EMT- (c) and DDR-based (d) risk groups and 22 kinds of immune cells. (e) ESTIMATE, stromal and immune scores between low-risk and high-risk in TCGA-COAD patients for EMT and DDR-related signatures. (f) Dual-margin scatterplots correlation between risk score and ESTIMATE-derived scores for EMT and DDR-related prognostic signatures in TCGA-COAD. Construction and validation of nomogram models provide survival predictions based on EMT- (g) and DDR-related (h) signatures. Calibration curves further demonstrated agreement between predicted and observed survival outcomes at all three OS for EMT- (i) and DDR-related (j) signature. Time-dependent ROC curves validate the predictive ability of the nomogram alone, age, TNM stage, risk score, and combined age + TNM stage at 1-, 3-, and 5-year OS for EMT- (k) and DDR-related (l) signature.

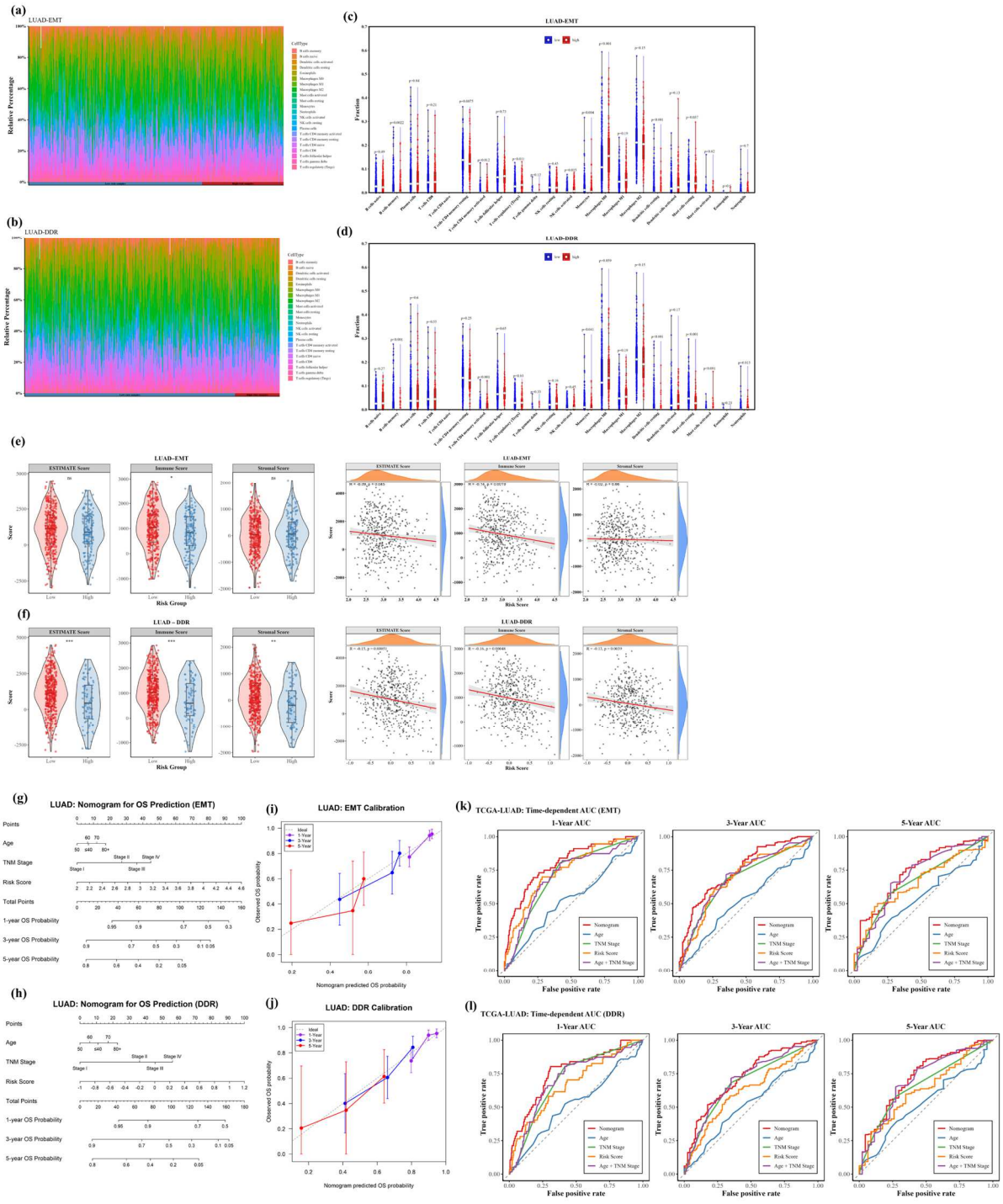


Figure 4.15. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-LUAD cohort. Bar chart plots of 22 immune cells infiltration across high- and low-risk groups of EMT- (a) and DDR-related (b) signatures. Raincloud plots between EMT- (c) and DDR-based (d) risk groups and 22 kinds of immune cells. (e) ESTIMATE,

stromal and immune scores between low-risk and high-risk in TCGA-LUAD patients for EMT and DDR-related signatures. (f) Dual-margin scatterplots correlation between risk score and ESTIMATE-derived scores for EMT and DDR-related prognostic signatures in TCGA-LUAD. Construction and validation of nomogram models provide survival predictions based on EMT- (g) and DDR-related (h) signature. Calibration curves further demonstrated agreement between predicted and observed survival outcomes at all three OS for EMT- (i) and DDR-related (j) signature. Time-dependent ROC curves validate the predictive ability of alone nomogram, age, TNM stage, risk score and combined age + TNM stage at 1-, 3-, and 5-year OS for EMT- (k) and DDR-related (l) signature.

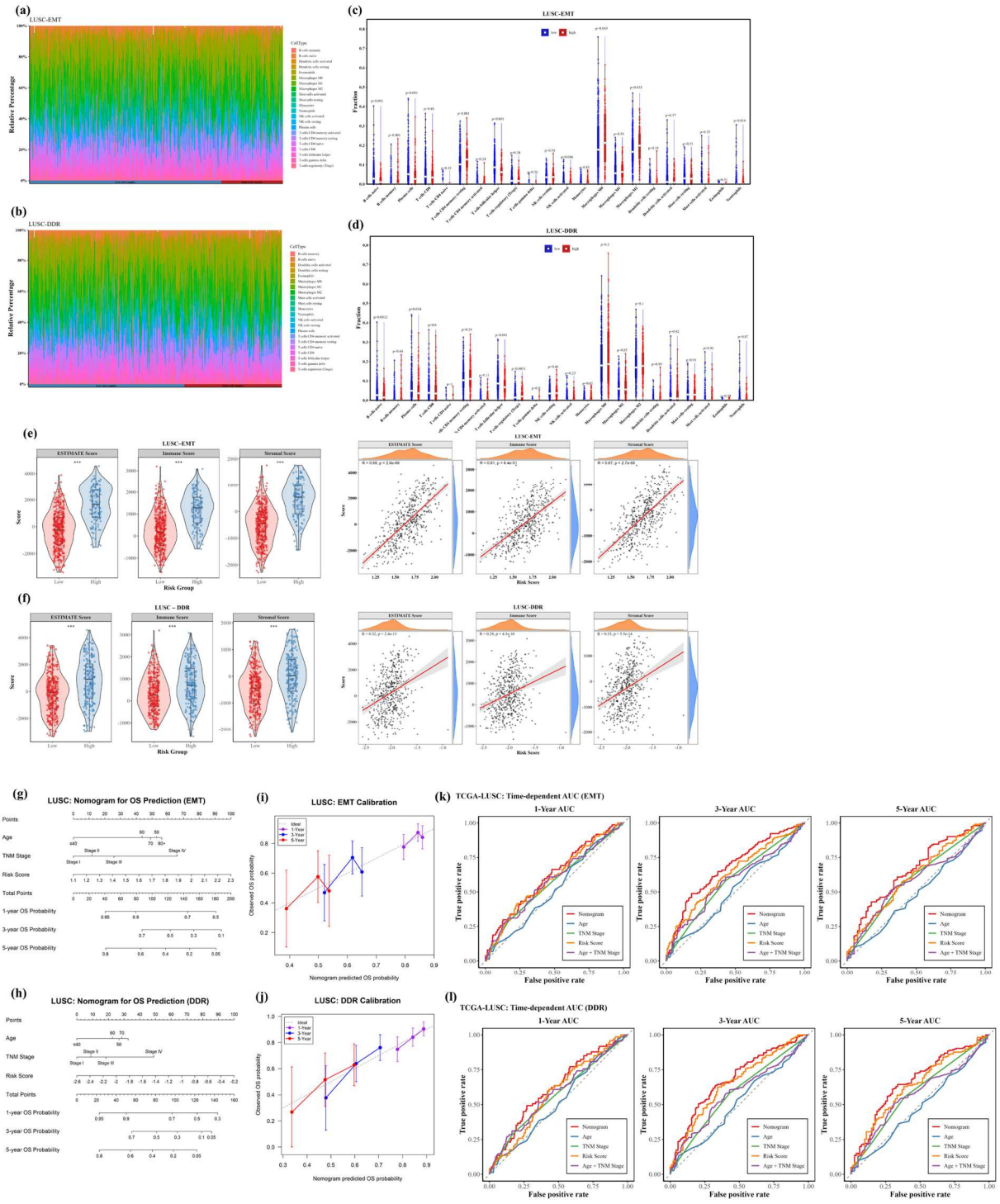


Figure 4.16. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-LUSC cohort. Bar chart plots of 22 immune cells infiltration across high- and low-risk groups of EMT- (a) and DDR-related (b) signatures. Raincloud plots between

EMT- (c) and DDR-based (d) risk groups and 22 kinds of immune cells. (e) ESTIMATE, stromal and immune scores between low-risk and high-risk in TCGA-LUSC patients for EMT and DDR-related signatures. (f) Dual-margin scatterplots correlation between risk score and ESTIMATE-derived scores for EMT and DDR-related prognostic signatures in TCGA-LUSC. Construction and validation of nomogram models provide survival predictions based on EMT- (g) and DDR-related (h) signatures. Calibration curves further demonstrated agreement between predicted and observed survival outcomes at all three OS for EMT- (i) and DDR-related (j) signature. Time-dependent ROC curves validate the predictive ability of the nomogram alone, age, TNM stage, risk score, and combined age + TNM stage at 1-, 3-, and 5-year OS for EMT- (k) and DDR-related (l) signatures.

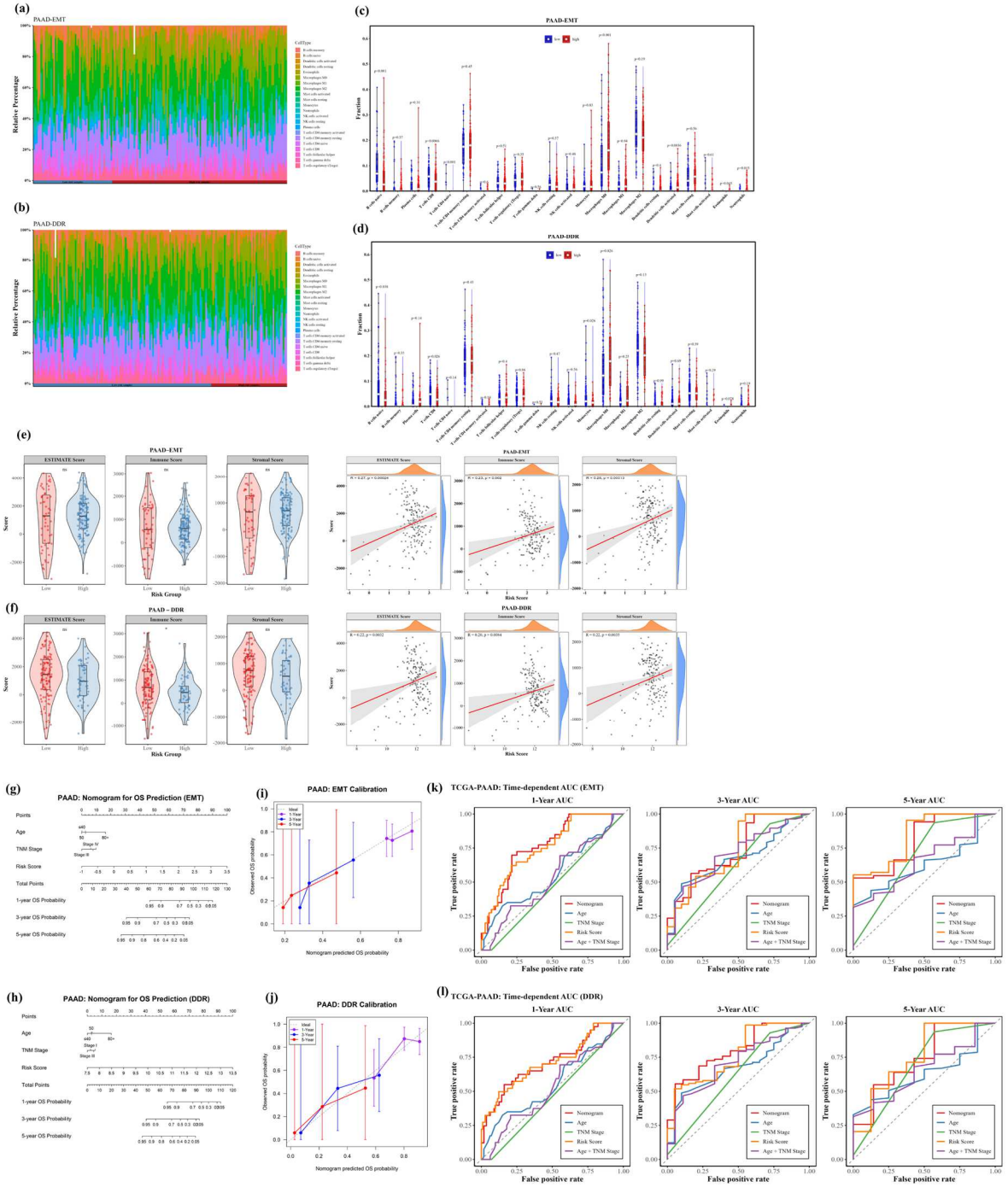


Figure 4.17. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in TCGA-PAAD cohort. Bar chart plots of 22 immune cells infiltration across high- and low-risk groups of EMT- (a) and DDR-related (b) signatures. Raincloud plots between EMT- (c) and DDR-based (d) risk groups and 22 kinds of immune cells. (e) ESTIMATE,

stromal and immune scores between low-risk and high-risk in TCGA-PAAD patients for EMT and DDR-related signatures. (f) Dual-margin scatterplots correlation between risk score and ESTIMATE-derived scores for EMT and DDR-related prognostic signatures in TCGA-PAAD. Construction and validation of nomogram models provide survival predictions based on EMT- (g) and DDR-related (h) signatures. Calibration curves further demonstrated agreement between predicted and observed survival outcomes at all three OS for EMT- (i) and DDR-related (j) signature. Time-dependent ROC curves validate the predictive ability of the nomogram alone, age, TNM stage, risk score, and combined age + TNM stage at 1-, 3-, and 5-year OS for EMT- (k) and DDR-related (l) signature.

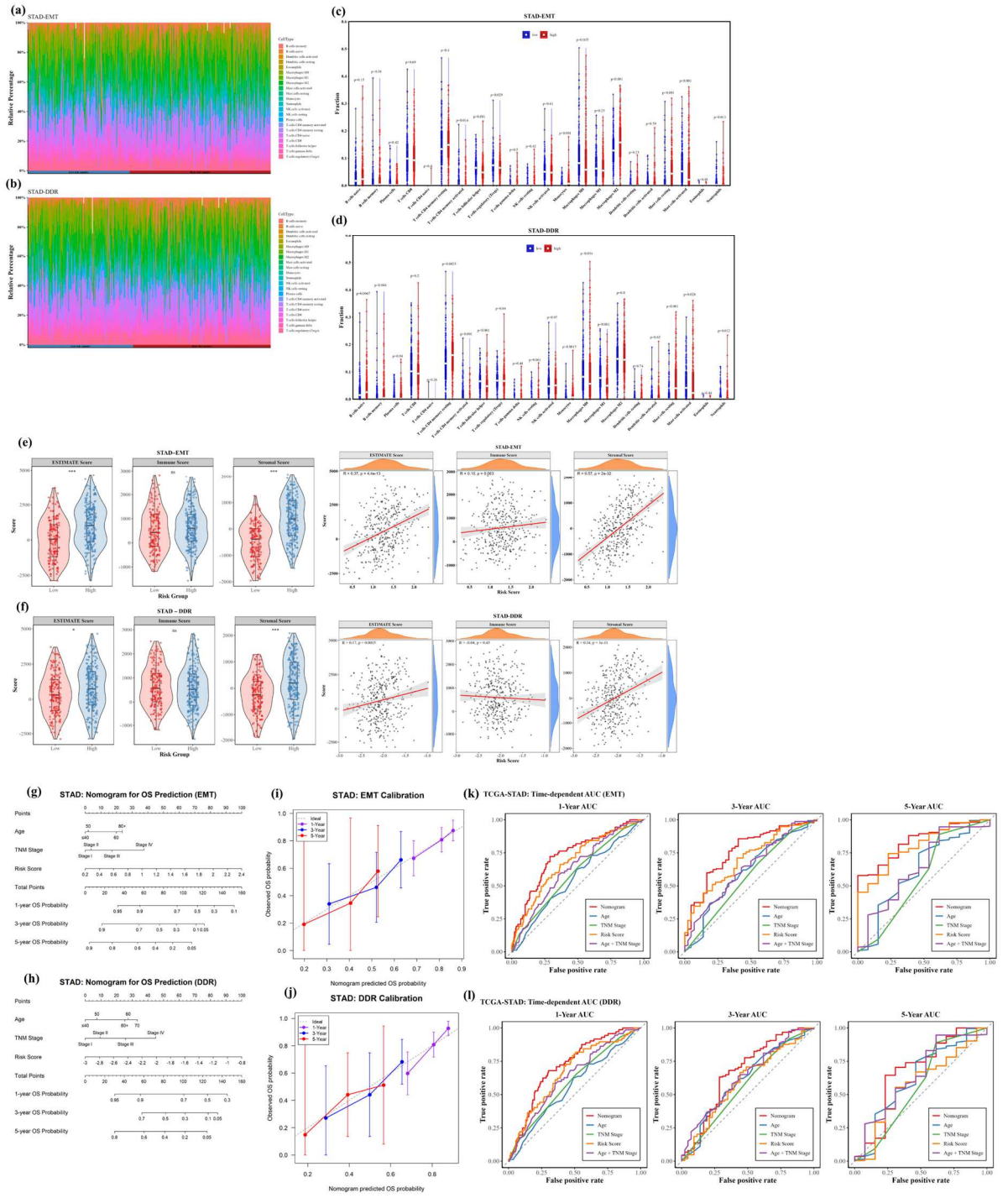


Figure 4.18. The immune cell infiltration analysis and nomogram construction and validation model for overall survival across high- and low-risk groups defined by EMT- and DDR-related signatures in the TCGA-STAD cohort. Bar chart plots of 22 immune cells infiltration across high- and low-risk groups of EMT- (a) and DDR-related (b) signatures. Raincloud plots between EMT- (c) and DDR-based (d) risk groups and 22 kinds of immune cells. (e)

ESTIMATE, stromal and immune scores between low-risk and high-risk in TCGA-STAD patients for EMT and DDR-related signatures. (f) Dual-margin scatterplots correlation between risk score and ESTIMATE-derived scores for EMT and DDR-related prognostic signatures in TCGA-STAD. Construction and validation of nomogram models provide survival predictions based on EMT- (g) and DDR-related (h) signatures. Calibration curves further demonstrated agreement between predicted and observed survival outcomes at all three OS for EMT- (i) and DDR-related (j) signature. Time-dependent ROC curves validate the predictive ability of the nomogram alone, age, TNM stage, risk score, and combined age + TNM stage at 1-, 3-, and 5-year OS for EMT- (k) and DDR-related (l) signatures.

In both TCGA-LUAD and TCGA-LUSC cohorts, for EMT-related signature, only macrophages M0 ($p < 0.001$, $p = 0.043$) showed high infiltration in high-risk group; instead for DDR-related signature, in TCGA-LUAD only neutrophils showed high infiltration in high-risk group, while in TCGA-LUSC no immune cell was significant in a high-risk group (see Figures 4.15–4.16c,d). Whereas the validation cohorts of LUAD revealed that T cell follicular helper ($p = 0.011$), NK cell resting ($p = 0.008$), macrophage M1 ($p = 0.030$), and mast cell activated ($p < 0.001$) were most enriched in high-risk group for EMT-related signature and T cells CD4 memory activated ($p < 0.001$), NK cells resting ($p = 0.026$), and mast cells activated ($p < 0.001$) for DDR-related signature (see Supplementary File 3, Figure S12b,e). However, in the LUSC-GSE30219 cohort, the enrichment of macrophages M1 ($p = 0.008$) was abundant in the high-risk group for EMT-related signature, and no significant enrichment was noted for the DDR-related signature (see Supplementary File 3, Figure S13b,e). Like other cancers, the same trend was observed for PAAD, where macrophage M0 ($p < 0.001$) was a highly enriched immune cell in high-risk groups for training (see Figure 4.17c,d) and validation cohort (see Supplementary File 3, Figure S14b,e). Interestingly, in STAD, macrophages M0 ($p < 0.05$) were enriched in low-risk group compared to high-risk group for both prognostic signatures. Whereas mast cell resting ($p < 0.001$) was most infiltrated in the high-risk group for training (see Figure 4.18c,d) as well as validation cohort (see Supplementary File 3, Figure S15b,e) for both prognostic signatures. Based on the results, it is reported that macrophages M0 ($p < 0.001$) showed high infiltration in high-risk groups for BRCA, COAD, LUAD, LUSC, and PAAD. In contrast, in STAD, mast cell resting ($p < 0.001$) was the most significant for both prognostic

signatures. Similarly, naïve B cells, B cell memory, plasma cells, and T cell follicular helper were most significant in low-risk groups in most cancers in training and validation cohorts.

To further evaluate the joint prognostic ability of the top six most significant immune cells and EMT- and DDR-related gene signatures in all six cancers, we performed Kaplan-Meier curves stratified by immune cell levels and risk score. Based on the degree of immune cell infiltration and median risk score values, the patients were divided into four clusters, including immune low/risk low (G1), immune high/risk low (G2), immune low/risk high (G3), and immune high/risk high (G4) for survival analysis. Across all six immune cell types, neutrophils and macrophages M2, presented the best prognostic association for EMT-related signature while neutrophils and naïve B cells showed the best prognostic ability in TCGA-BRCA for DDR-related prognostic signature, respectively (see Supplementary File 1, Figure S1a–d). While for BRCA-GSE20685, M0 and M1 macrophages showed the best prognostic ability for EMT-related signature (see Supplementary File 3, Figure S10c). For EMT-related signature, the best prognostic ability of macrophages M0 was found in TCGA-COAD, dendritic cells resting in TCGA-LUAD, plasma cells in TCGA-LUSC, macrophages M0 and eosinophils in TCGA-PAAD, and mast cells resting in TCGA-STAD, respectively (see Supplementary File 1, Figures S1–3(a–d)). Whereas in the validation cohort of COAD-GSE39582, M1 macrophages and resting dendritic cells presented the best prognostic ability, followed by resting mast cells in LUAD-GSE50081, activated dendritic cells in PAAD-GSE79668, and macrophages M2 with OS in STAD-GSE84433, respectively (see Supplementary File 3, Figures S11–15c). For the DDR-related prognostic signature, the optimal prognostic ability for OS was noted for activated dendritic cells in TCGA-COAD, mast cells activated in TCGA-LUAD, plasma cells in TCGA-LUSC, macrophages M0 in TCGA-PAAD, and resting mast cells in TCGA-STAD, respectively. In the validation cohorts, gamma delta T-cells showed the best prognostic ability in COAD-GSE39582, mast cell resting in LUAD-GSE50081, M0 macrophages in PAAD-GSE79668, and memory activatssed CD4 T cells in STAD-GSE84433 for DDR-related signature, respectively (see Supplementary File 3, Figures S11–15f). Results showed that in all cancer cohorts, the worst prognosis was observed in cluster G3 compared to other clusters, where immune cell type infiltration was low and the risk score was high for both prognostic signatures (see Supplementary File 1, Figures S1–3a,b,c,d).

To further explore the behaviour of risk scores based on stratification (low- and high-risk groups) in the tumour microenvironment, we applied the ESTIMATE algorithm to derive transcriptome-based ESTIMATE metrics. In TCGA-BRCA, the ESTIMATE, immune, and stromal scores are remarkably higher in low-risk group ($p < 0.001$) compared to high-risk group for EMT- and DDR-related signatures, suggesting low-risk samples tend to harbour more abundant stromal and immune infiltrates (see Figure 4.13e,f). The higher ESTIMATE score in the low-risk group indicated greater total non-tumour cellularity in these samples. The correlation analyses results showed the negative correlation for ESTIMATE, immune, and stromal scores for both signatures, which suggested that higher risk scores tend to harbour less non-tumour content on average (see Figure 4.13e,f). In TCGA-COAD, ESTIMATE and stromal and immune scores did not show a significant difference for both signatures. In contrast, EMT higher risk scores exhibited a positive significant correlation with ESTIMATE ($R = 0.64$, $p = 0.0033$), and stromal ($R = 0.69$, $p = 0.0011$) and immune scores ($R = 0.53$, $p = 0.019$) (see Figure 4.14e,f). TCGA-LUAD presented similar results to TCGA-BRCA, where ESTIMATE metrics were more enriched in low-risk groups for both signatures, and high-risk samples showed significant negative correlations (see Figure 4.15e,f).

Unlike in TCGA-LUAD, where EMT showed no significant relation to microenvironment, in TCGA-LUSC, both signatures showed a relationship with tumours that are richer in immune and stromal cells (see Figure 4.16e,f). Notably, high-risk groups in TCGA-LUSC exhibit higher ESTIMATE, immune, and stromal scores across signatures, indicating a greater non-tumour microenvironment content and increased inferred immune infiltration, particularly in high-risk groups. In TCGA-PAAD, only the immune score was statistically significant ($p < 0.05$) for the DDR signature, where low-risk samples exhibit more non-tumour microenvironment content (see Figure 4.17e,f). Likewise, for TCGA-LUSC, the results of TCGA-STAD were identical, with high-risk samples exhibiting greater non-tumour microenvironment content and higher inferred immune infiltration ($p < 0.001$). Interestingly, the immune score was non-significant for both signatures in TCGA-STAD, which does not appear to be driven by changes in immune cell infiltration, unlike other cancers (see Figure 4.18e,f).

4.8. Signature-Based Nomogram Provides Accurate Prognostic Estimates

We constructed a signature-based nomogram to predict the probability of 1-, 3-, and 5-year OS of the patients for clinical parameters such as age, risk score, and TNM stage in the training (TCGA) and validation cohorts (GEO). We also performed a calibration curve, ROC curve, DCA, C-index, and AIC to validate the discrimination, clinical usefulness, and calibration of the model in both the training and validation cohorts. The nomogram results reported that patients with higher risk scores have a poor prognosis in all cancers for both prognostic signatures in training (see Figures 4.13–4.18g,h) as well as in validation cohort (see Supplementary File 3, Figures S16–21a,e). This model provides powerful tools for estimating patient-specific outcomes in clinical parameters. The calibration plot results for both EMT- and DDR-based nomograms reported a strong correlation between predicted and observed survival probabilities at 1- and 3-year OS in the training set (see Figures 4.13–4.18i,j) and validation cohorts (see Supplementary File 3, Figures S16–21d,h). While 5-year survival predictions are less reliable, it might be due to fewer long-term survivals or increased variability. The results also showed that the DDR-based nomogram demonstrated better long-term accuracy with a 5-year prediction than the EMT-based nomogram (see Figures 4.13–4.18i,j).

The C-index and ROC results for nomogram for OS prediction revealed the optimal values of 0.780 and AUCs of 0.825, 0.817 and 0.766 at 1-, 3-, and 5-year OS for EMT-related signature, while for DDR-related signature the C-index and AUCs were noted as 0.783 and 0.837, 0.808, 0.772 at 1-, 3-, and 5-year OS in TCGA-BRCA cohort, respectively (see Table 4.5; Figure 4.13k,l). In TCGA-COAD, C-index for the nomogram was observed as 0.779 and 0.736 for EMT- and DDR-related signatures, and AUCs of 0.801, 0.813, and 0.755 for EMT-related signature, and 0.763, 0.743, and 0.651 for DDR-related signature at 1-, 3-, and 5-year OS prediction, respectively (see Table 4.5; Figure 4.14k,l). Similarly, in TCGA-LUAD and TCGA-LUSC, the C-index values of the nomogram models for OS prediction were 0.734 and 0.613 for the EMT-based signature, and 0.701 and 0.611 for the DDR-based signature, respectively. However, the AUCs for ROC were 0.798, 0.757, and 0.741 in TCGA-LUAD and 0.619, 0.678, and 0.657 in TCGA-LUSC cohort for EMT-based signature; for DDR-based signature, the AUCs were 0.762, 0.729, and 0.727 in TCGA-LUAD, and 0.621, 0.687, and 0.669 in TCGA-LUSC, respectively (see Table 4.5; Figures 4.15–4.16k,l). Furthermore, TCGA-PAAD and TCGA-STAD cohorts revealed the maximum C-index for nomogram OS prediction of 0.707

and 0.704 for EMT-based signature and 0.695 and 0.674 for DDR-based signature; the AUCs in TCGA-PAAD (0.769, 0.736, and 0.812 for EMT; 0.713, 0.796 and 0.741 for DDR) and TCGA-STAD (0.735, 0.766, and 0.837 for EMT; 0.717, 0.672 and 0.676 for DDR) at 1-, 3- and 5-year OS, respectively (see Table 4.5; Figures 4.17–4.18k,l).

Table 4.5. C-index of signature-based nomogram models for all six cancers in training and validation cohorts.

Model	C-index EMT	C-index DDR	C-index EMT	C-index DDR
	TCGA-BRCA		**BRCA-GSE20685**	
Nomogram COX PH	0.780632	0.783253	0.732489	0.714373
Age + TNM stage COX PH	0.759638	0.759638	0.689856	0.689856
Age + Risk score COX PH	0.699157	0.705026	0.646938	0.600275
TNM stage + Risk score COX PH	0.739427	0.737829	0.719874	0.703656
Age only COX PH	0.625229	0.625229	0.552994	0.552994
TNM stage only COX PH	0.683762	0.683762	0.664556	0.664556
Risk score only COX PH	0.658378	0.662325	0.627780	0.570683
TCGA-COAD			**COAD-GSE39582**	
Nomogram COX PH	0.779792	0.736645	0.699828	0.706325
Age + TNM stage COX PH	0.74583	0.74583	0.698577	0.698577
Age + Risk score COX PH	0.723032	0.659506	0.603233	0.611156
TNM stage + Risk score COX PH	0.744017	0.693981	0.631214	0.644343
Age only COX PH	0.604877	0.604877	0.587007	0.587007
TNM stage only COX PH	0.694162	0.694162	0.626964	0.626964
Risk score only COX PH	0.683386	0.607966	0.533510	0.546147
TCGA-LUAD			**LUAD-GSE50081**	
Nomogram COX PH	0.734811	0.701837	0.700195	0.679168
Age + TNM stage COX PH	0.679033	0.679033	0.626382	0.626382
Age + Risk score COX PH	0.688487	0.631328	0.658357	0.661392
TNM stage + Risk score COX PH	0.730862	0.698019	0.681986	0.649252
Age only COX PH	0.553423	0.553423	0.562866	0.562866
TNM stage only COX PH	0.66104	0.66104	0.595491	0.595491
Risk score only COX PH	0.680843	0.624618	0.624973	0.618903
TCGA-LUSC			**LUSC-GSE30219**	
Nomogram COX PH	0.613479	0.611892	0.708886	0.664456
Age + TNM stage COX PH	0.562219	0.562219	0.65252	0.65252
Age + Risk score COX PH	0.59557	0.583867	0.665119	0.612069
TNM stage + Risk score COX PH	0.608645	0.612169	0.649867	0.631963
Age only COX PH	0.523671	0.523671	0.570955	0.570955
TNM stage only COX PH	0.557883	0.557883	0.579576	0.579576
Risk score only COX PH	0.591888	0.583064	0.606764	0.595491
TCGA-PAAD			**PAAD-GSE79668**	
Nomogram COX PH	0.707593	0.695718	0.736292	0.707572
Age + TNM stage COX PH	0.57686	0.57686	0.657093	0.657093
Age + Risk score COX PH	0.710928	0.702382	0.709936	0.673077
TNM stage + Risk score COX PH	0.69245	0.67186	0.692776	0.616188
Age only COX PH	0.56228	0.56228	0.629006	0.629006
TNM stage only COX PH	0.527944	0.527944	0.556571	0.556571
Risk score only COX PH	0.697041	0.679201	0.666667	0.596154
TCGA-STAD			**STAD-GSE84433**	
Nomogram COX PH	0.70423	0.674151	0.631499	0.623043
Age + TNM stage COX PH	0.625921	0.625921	–	–
Age + Risk score COX PH	0.660191	0.635884	0.631499	0.623043
TNM stage + Risk score COX PH	0.686078	0.648341	–	–
Age only COX PH	0.53585	0.53585	0.583042	0.583042
TNM stage only COX PH	0.594473	0.594473	–	–

Risk score only COX PH	0.654579	0.627558	0.587404	0.58016

In the validation cohort of BRCA, COAD, LUAD, LUSC, PAAD, and STAD, the C-index of nomogram OS prediction was noted as 0.732, 0.699, 0.700, 0.708, 0.736 and 0.631 for EMT-related signature and 0.714, 0.706, 0.679, 0.664, 0.707, and 0.623 for DDR-related signature respectively (see Table 4.5). For EMT-related signature, the ROC results for nomogram revealed AUC values of 0.621, 0.778 and 0.788 for 1-, 3-, and 5-year OS in BRCA-GSE20685, 0.761, 0.738, and 0.736 in COAD-GSE39582, 0.687, 0.730, and 0.727 in LUAD-GSE50081, 0.661, 0.777, and 0.712 in LUSC-GSE30219, 0.824, 0.847, and 0.833 in PAAD-GSE79668, and 0.597, 0.651 and 0.669 in STAD-GSE84433, respectively (see Supplementary File 3, Figures S16–21c). Whereas for DDR-related signature, AUC values were noted as of 0.582, 0.780 and 0.774 for 1-, 3-, and 5-year OS in BRCA-GSE20685, 0.692, 0.724, and 0.714 in LUAD-GSE50081, 0.561, 0.720, and 0.685 in LUSC-GSE30219, 0.790, 0.820, and 0.823 in PAAD-GSE79668, and 0.557, 0.630 and 0.651 in STAD-GSE84433, respectively (see Supplementary File 3, Figures S16–21g). It was observed that the combined Age + TNM stage yielded higher AUC values for ROC at 1-, 3-, and 5-year OS, compared to single variables such as age and TNM stage. The results suggested strong predictive performance across 1-, 3-, and 5-year OS. The nomogram outperformed age, TNM stage, and their combination with EMT- and DDR-derived risk scores and emerged as a reliable prognostic tool for short- and long-term OS in all cancers.

Further, we performed the decision curve analysis (DCA) to assess the clinical utility of predictive models in training and validation cohorts. Based on results, it was determined that DCA confirmed the clinical value of signature-based nomograms in all six cancers and reported the higher net benefits than individual clinical parameters (age and TNM stage) across 1-, 3-, and 5-year OS predictions. It is concluded that the nomogram model performed well across cancers and has the potential to support personalized decision-making and improve patient outcomes (see Supplementary File 1, Figures S4–6). AIC analysis was used to find the goodness of fit for the model individually and in combination in the training and validation cohorts. The results reported that BRCA, COAD, and STAD have the lowest AIC values for nomograms, while in LUAD, LUSC, and STAD, the AIC values for the nomogram model were slightly higher than age + risk score, which significantly suggested a better balance between simplicity and accuracy of the model and avoided overlifting (see Supplementary File 2, Table S1-6).

Similarly, in the validation cohorts, the AIC values were lowest in BRCA, COAD, LUSC, PAAD, and STAD for the nomogram model, while in LUAD, the values were slightly higher than TNM stage + risk score, which indicated that our nomogram model was significantly optimal for prognostic EMT and DDR-related prognostic signatures (see Supplementary File 4, Table S1-6).

4.9. Expression and Survival Analysis of Prognostic EMT- and DDR-Related Signature Genes

To assess the prognostic significance of novel EMT- and DDR-related signatures, we investigated the expression profile between tumour and normal tissues and Kaplan-Meier curves. The results showed that EMT-related prognostic genes such as *GLIPR1, PDLIM4,* and *TFPI2* were significantly downregulated, and *SCG2* and *SDC1* were significantly upregulated genes in breast tumour tissue. While in DDR-related prognostic genes, *CCND2, DDB2,* and *TP53AIP1* were downregulated considerably, and *PMAIP1* and *RAD1* were upregulated in breast tumour tissue (see Supplementary File 1, Figure S7a,c). In COAD, *TIMP1* was upregulated and *CRLF1, FUCA1, PCOLCE2,* and *SGCG* were significantly downregulated genes for EMT-related signature, whereas *GADD45B* was a significantly downregulated gene for DDR-related signature in tumour tissues compared to normal tissues (see Supplementary File 1, Figure S8a,c). In LUAD, a panel of EMT-related genes including *PMEPA1, SERPINH1* and *SPOCK1* were significantly overexpressed and *ABI3BP, ID2, MGP,* and *PVR* showed downregulation in tumour tissue compared to normal tissues, while DDR drivers such as *CDC25C, CYCS* and *GADD45G* were upregulated and *CCND2* and *RRM2B* were downregulated in tumour tissue (see Supplementary File 1, Figure S9a,c).

Similarly, in LUSC cohort, *FSTL3, PTX3,* and *TGM2* were significantly downregulated genes in EMT-related panel, while *ABL1, CCND2,* and *GADD45G* were downregulated genes DDR-related signature (see Supplementary File 1, Figure S10a,c). Results indicated that *AREG, LAMA3, MMP3,* and *PVR* were significantly upregulated for EMT-related signature and *CDK1* for DDR-related signature in PAAD tumour tissues compared to normal (see Supplementary File 1, Figure S11a,c). Like other cohorts, in STAD cohort, *CDH6, CXCL1, DKK1, MATN3, NT5E,* and *SERPINE1* were significantly overexpressed in tumour tissues for EMT-related panel, and *BBC3, CASP3, CDK5, PML, SMC1A,* and *TNFRSF10B* for DDR-related panel (see

Supplementary File 1, Figure S12a,c). We further performed the Kaplan-Meier curves for prognostic genes in both signatures. It was noted that EMT-related signature genes such as *TFPI2* ($p < 0.0001$), *PDLIM4* ($p = 0.0082$), *GLIPR1* ($p = 0.0011$) and *EMP3* ($p = 0.0027$) exhibited worse prognosis for OS in BRCA, while *TP53AIP1* ($p < 0.0001$), *PMAIP1* ($p = 0.0025$), *DDB2* ($p = 0.0033$), *CCND2* ($p = 0.0015$), and *CASP9* ($p = 0.0002$) for DDR-related signature (see Supplementary File 1, Figure S7b,d). In COAD, only *FUCA1* ($p = 0.002$) showed worse OS for EMT-related signature, and for DDR-related signature, all genes showed better OS (see Supplementary File 1, Figure S8b,d). Similarly, in LUAD, the significant worse OS was observed in *RRM2B* ($p = 0.0007$), *CCND2* ($p = 0.0015$), *ID2* ($p = 0.0082$), *ABI3BP* ($p = 0.0005$) and *MGP* ($p = 0.0005$) for EMT-related genes, and *GADD45G* ($p < 0.0001$) for DDR-related signature (see Supplementary File 1, Figure S9b,d).

In the case of LUSC, all EMT-related signature genes presented better OS while *CCND2* ($p = 0.0027$) and *CHK2* ($p < 0.0001$) demonstrated worse OS outcomes for DDR-related signature (see Supplementary File 1, Figure S10b,d). In PAAD cohort, *PVR* ($p = 0.002$) and *MAGEE1* ($p = 0.0033$) showed the best overall survival in EMT-related signature, whereas *RAD52* ($p = 0.0130$), *CCND2* ($p < 0.0002$), and *IGFBP2* ($p = 0.0006$) had significant worse OS outcomes in DDR-related signature (see Supplementary File 1, Figure S11b,d). In STAD, *CXCL1* ($p = 0.0260$) showed the worst OS outcomes for EMT-related signature, while in DDR-related signature, all genes showed the worst OS outcomes ($p < 0.0001$) except *GADD45B* (see Supplementary File 1, Figure S12b,d). The results highlight the complex and mixed functions of EMT- and DDR-related signature genes in all six cancers and support their potential ability as therapeutic targets and prognostic biomarkers.

4.10. Low-Risk and High-Risk Groups Possess Distinct ssGSEA Scores Across Cancers

We performed the single-sample gene set enrichment analysis (ssGSEA) to illustrate the relationship between EMT and DDR scores. We compared the ssGSEA-derived EMT and DDR scores between high- and low-risk patient groups across all six cancers. Results showed that ssGSEA-derived EMT scores were generally higher in high-risk patients with COAD, LUAD, LUSC, PAAD, and STAD, whereas in the case of BRCA, the ssGSEA score was higher in the low-risk group. However, the most pronounced differences were noted in COAD, LUSC, and

STAD ($p < 0.001$) (see Figure 4.19a). The finding highlights a strong association between high-risk score and EMT activation, defining the role of EMT-related signature in driving the aggressive tumour phenotype.

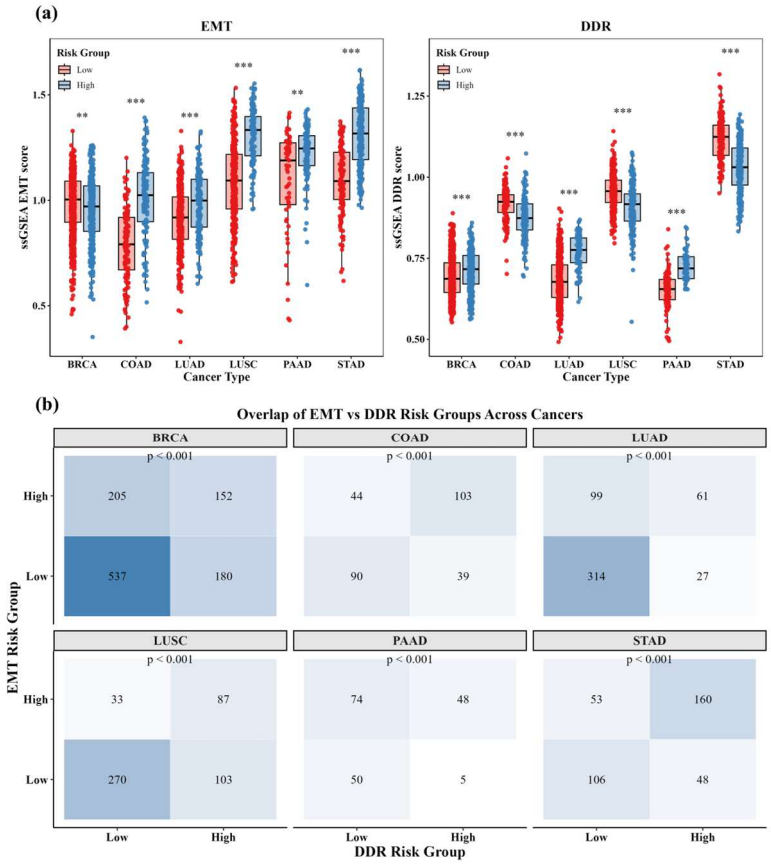


Figure 4.19. ssGSEA analysis for EMT and DDR-based risk groups across all six cancers. (a) Relationship between ssGSEA score and risk groups across cancers. (b) Overlapping of samples between EMT and DDR risk groups (low vs high) across cancers. *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$).

While in the case of DDR-related signature, BRCA, LUAD, and PAAD showed the most significant higher ssGSEA in high-risk groups ($p < 0.001$) compared to low-risk groups, COAD, LUSC, and STAD cohorts exhibited greater ssGSEA scores in low-risk groups (see Figure 4.19a). DDR scores are more varied in tumours, which might be due to the genomic instability and worse outcomes. The spread of scores indicates heterogeneity in cancer progression, reflecting the diverse molecular characteristics of individual tumours.

Furthermore, we conducted a contingency analysis to examine the overlap between EMT and DDR risk groups across various cancers. Each panel presented an overlap of stratifying patients between low and high-risk groups of EMT and DDR. In BRCA, the higher number of patients (537) overlapped between low-EMT and low-DDR groups ($p < 0.001$) (see Figure 4.19b). Similar results were observed for LUAD and LUSC, where the majority of patients were classified into low-risk groups based on both EMT and DDR-related signatures ($p < 0.001$) (see Figure 4.19b). While in COAD and STAD, a large number of patients overlapped between high-risk groups of both EMT and DDR-related signatures ($p < 0.001$) (see Figure 4.19b). Interestingly, in PAAD, the majority of patients (74) were classified into low-DDR and high-EMT risk groups (see Figure 4.19b). The results reported a significant association between DDR and EMT risk groups across cancers, indicating that the biological processes are closely related and co-regulated within these cancers and risk groups.

4.11. Drug Screening Correlation Analysis Identifies Putative Compounds Showing Anti-Tumoral Activity for EMT- and DDR-Related Prognostic Gene Signatures Cell Lines Across Cancers

We performed the Pearson correlation analysis to observe the relationship between IC_{50} of drugs and risk score of cell lines across cancers. It was observed that the risk score was negatively correlated with most of the drugs thus indicating that higher-risk patients were more sensitive (lower IC_{50}) to drugs in BRCA and COAD for EMT-related prognostic signature (see Figures 4.20–4.21a). While for LUAD, LUSC, PAAD and STAD, results showed positive correlation between risk score of patients and IC_{50} of most drugs which presented that high-risk groups were more resistant to drugs for EMT and DDR-related prognostic signatures (see Figure 4.22–4.25a,d). We further screened the top five positively and negatively correlated drugs with risk score across cancers. For EMT-related prognostic genes, it was observed that BDP.00009066 displayed the best positive correlation in BRCA with risk score ($R = 0.54$, $p = 0.1$), which suggested that samples in the high-risk group required a higher concentration of drug to achieve 50% growth inhibition (see Figure 4.20b). BDP-00009066 (BDP-9066) is a potent inhibitor of myotonic dystrophy-related Cdc42-binding kinases such as MRCKα and MRCKβ. Its mechanism is reported by occupy the ATP-binding cleft of MRCKβ which leads to blocking catalytic activity. Studies reported that BDP-9066 showed antiproliferative activity in haematological malignancies. By inhibiting phosphorylation of myosin light chain (MLC),

it suppresses the motility and invasion of SCC12 squamous cell carcinoma (Unbekandt et al., 2018).

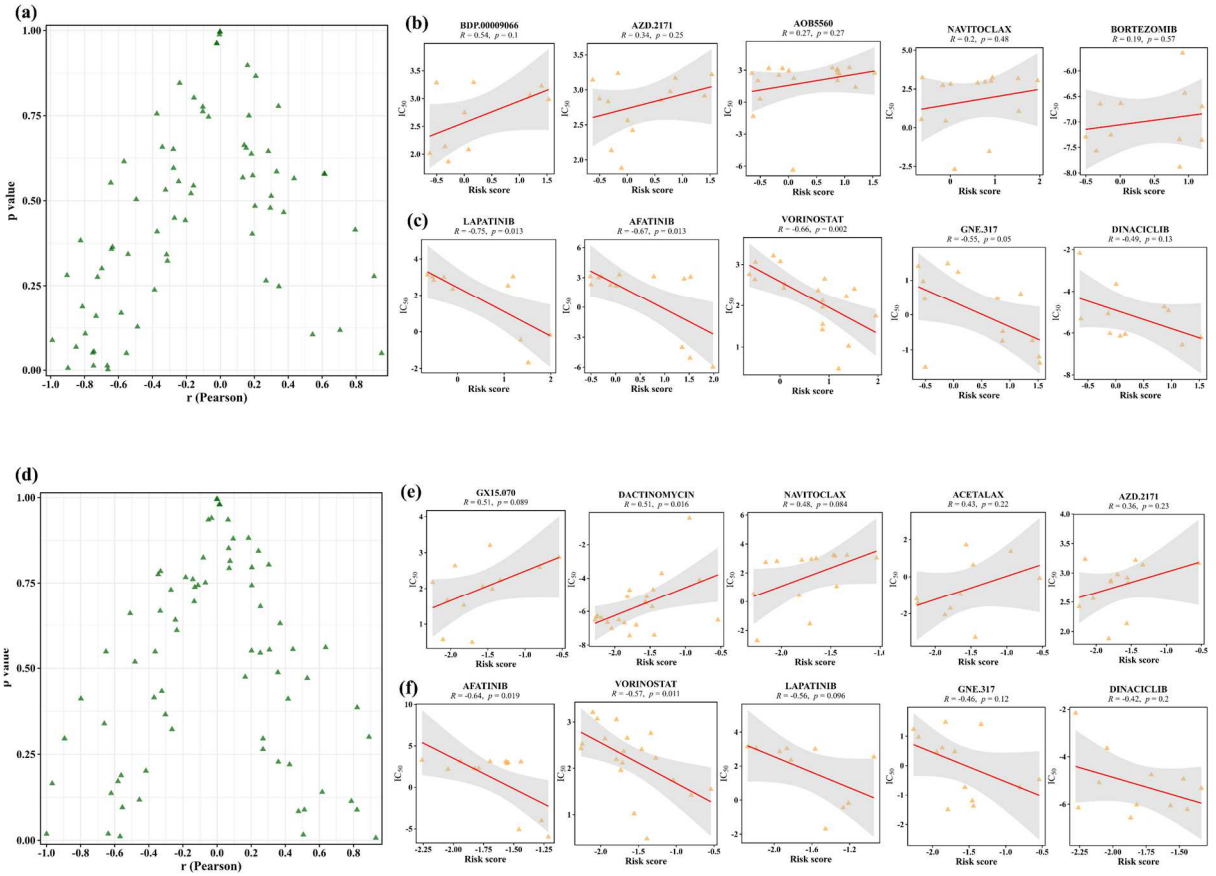


Figure 4.20. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in BRCA cell lines. (a, d) Correlation of EMT and DDR calculated risk score with IC_{50} of drugs in BRCA cell lines. (b, e) Pearson's positive correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes. (c, f) The Pearson's negative correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

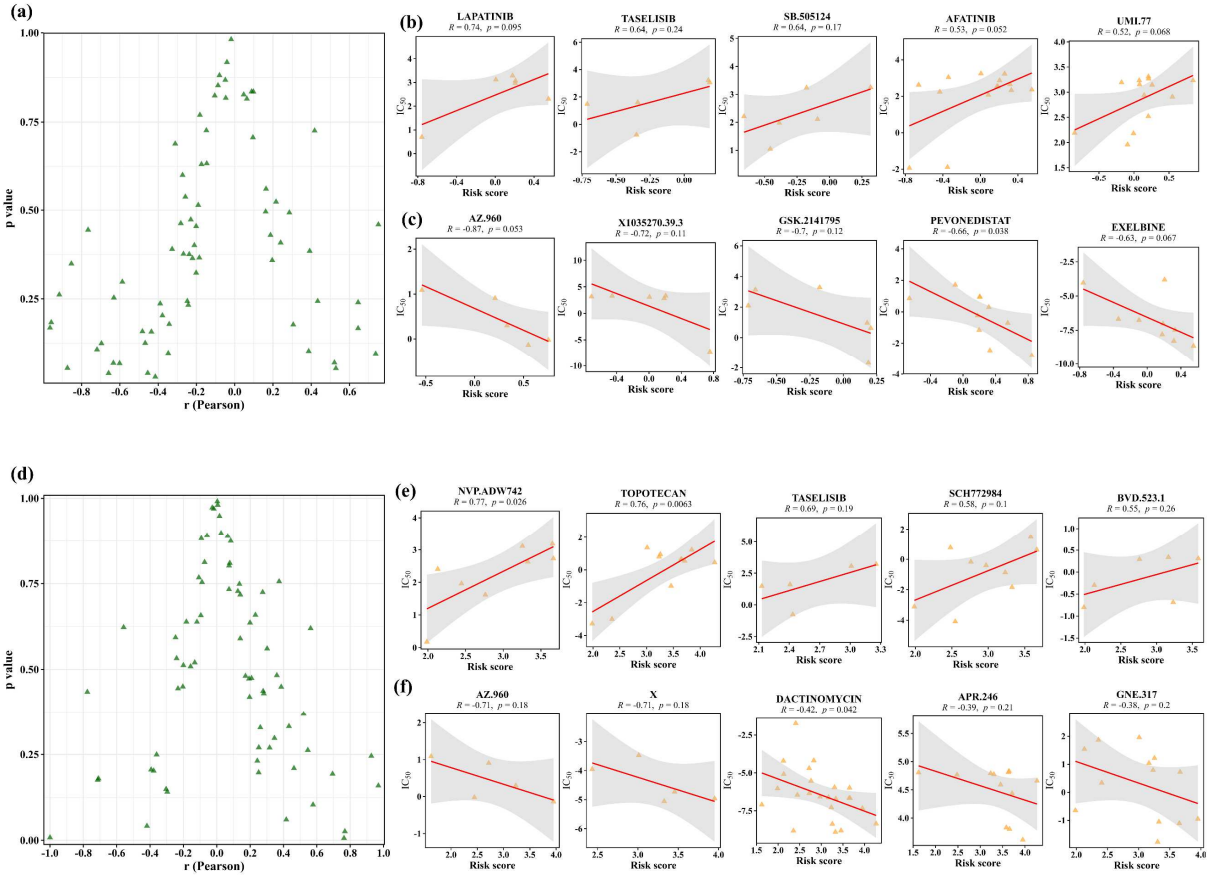


Figure 4.21. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in COAD cell lines. (a, d) Correlation of EMT and DDR calculated risk score with IC_{50} of drugs in COAD cell lines. (b, e) The Pearson's positive correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes. (c, f) The Pearson's negative correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

In COAD and LUAD, the optimal positive correlation was observed for Lapatinib (a tyrosine kinase inhibitor that targets EGFR and HER2 and blocks downstream of PI3K and MAPK signalling) (COAD: $R = 0.74$, $p = 0.095$; LUAD: $R = 0.83$, $p = 0.0058$) for EMT-related signature (see Figures 4.21–4.22b). Whereas in LUSC, PAAD and STAD, the drugs including Vorinostat (a pan-histone deacetylase inhibitor that increases acetylation of histone and non-histone proteins; $R = 0.92$, $p = 0.0031$), MST.312 (a synthetic telomerase inhibitor that block telomerase activity; $R = 0.52$, $p = 0.046$), and YK.4.279 (a small molecule that target the EWS-FLI1 fusion oncoprotein Ewing sarcoma; $R = 0.82$, $p = 0.026$) displayed the strongest and the most statistically significant positive correlation for EMT-related signature indicating that with

increasing the risk score, the concentration of doses also rises (see Figures 4.23–4.25b). Similarly, Lapatinib (R = -0.75, p = 0.013) showed the best statistically significant negative correlation in BRCA, Pevonedistat (First-in-class small molecule inhibitor of NEDD8-activating enzyme; R = -0.66, p = 0.038) in COAD, MST.312 (R = -0.43, p = 0.037) in LUAD, BMS.536924 (a selective ATP-competitive inhibitor of IGF-1R; R = -0.60, p = 0.21) in LUSC, Bortezomib (a reverse 26S proteasome inhibitor that disrupts degradation of key regulatory proteins; R = -0.49, p = 0.023) in PAAD, and BEZ235 (a dual PI3K/mTOR inhibitor that targets ATP-binding sites of class I PI3K isoforms and mTOR complexes; R = -0.95, p = 0.0036) in STAD, respectively, suggesting that high-risk scores require substantially lower drug concentration to achieve 50% growth inhibition (see Figures 4.20–4.25c).

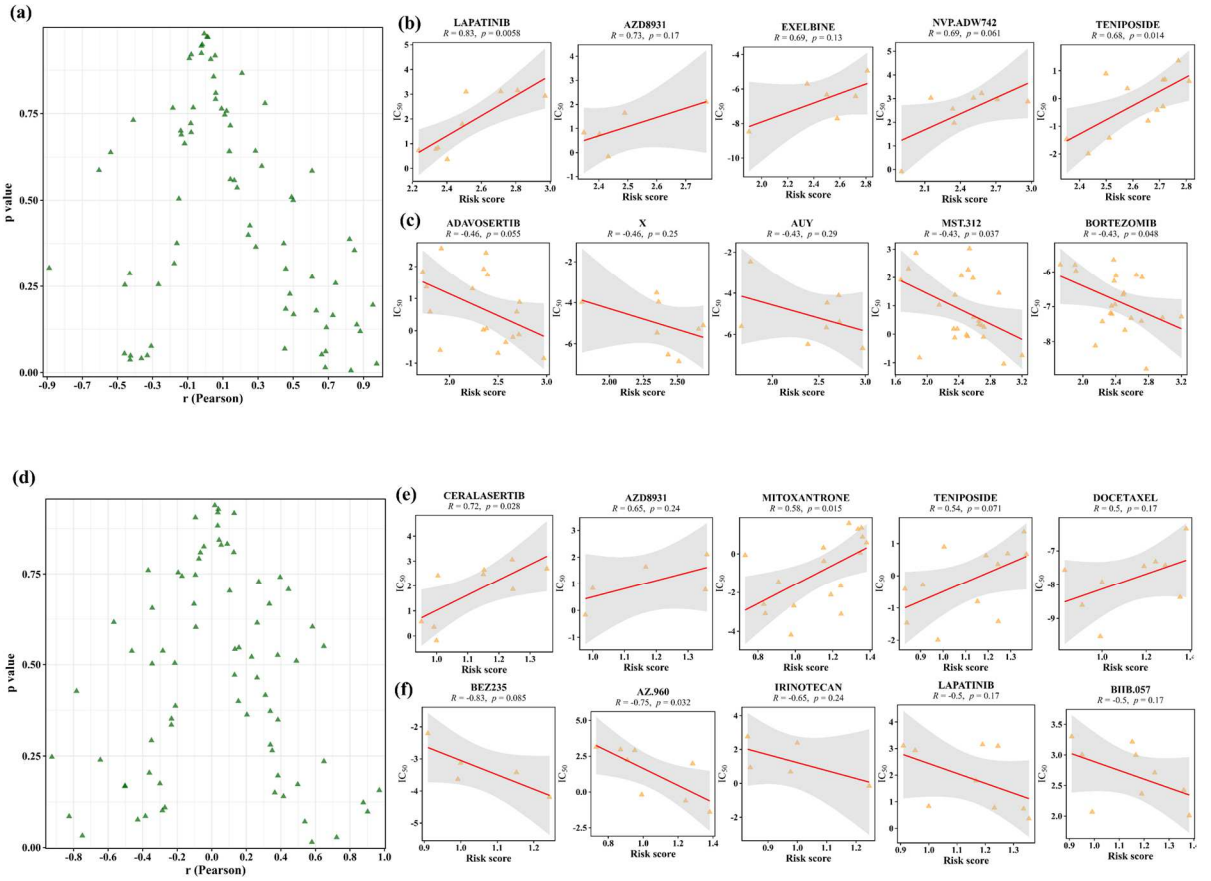


Figure 4.22. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in LUAD cell lines. (a, d) Correlation of EMT and DDR calculated risk score with IC_{50} of drugs in LUAD cell lines. (b, e) Pearson's positive correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes. (c, f) The

Pearson's negative correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

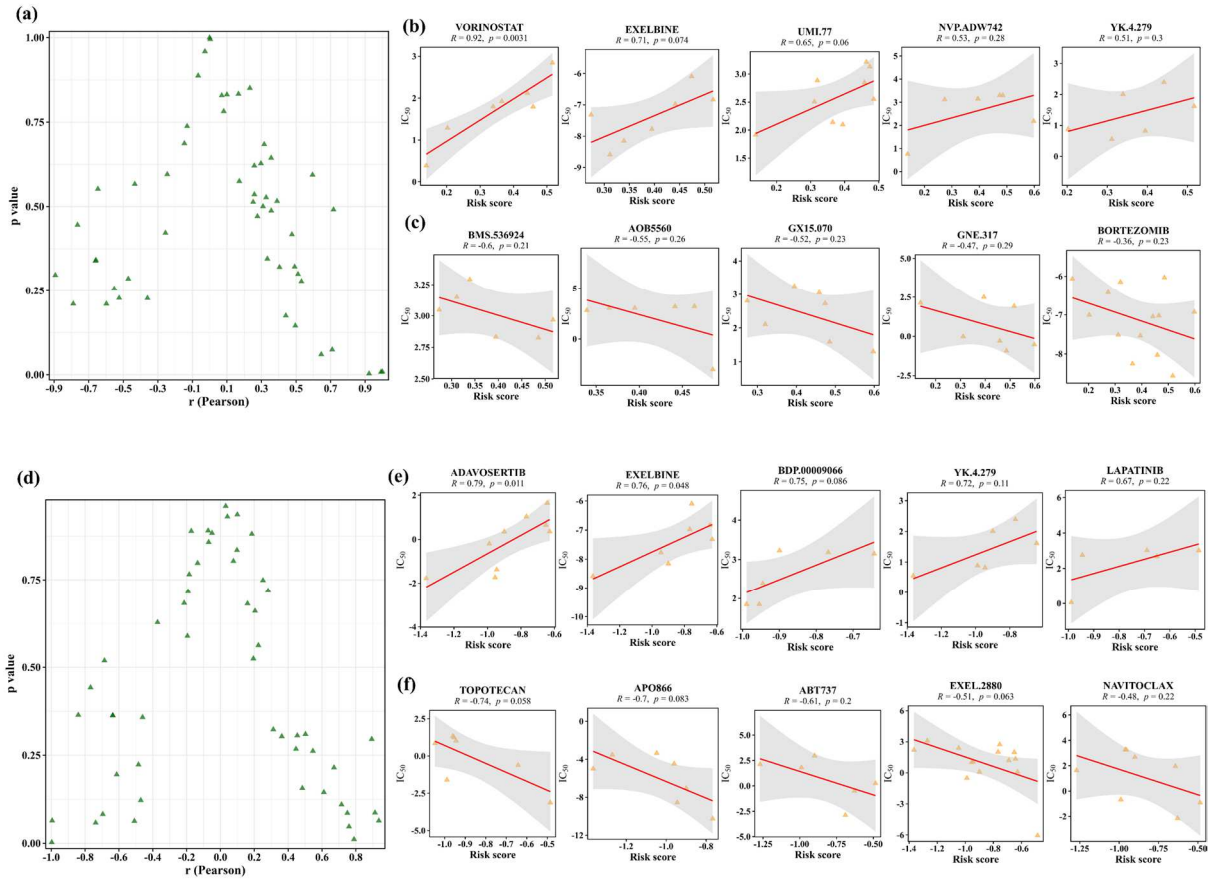


Figure 4.23. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in LUSC cell lines. (a, d) Correlation of EMT and DDR calculated risk score with IC_{50} of drugs in LUSC cell lines. (b, e) Pearson's positive correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes. (c, f) The Pearson's negative correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

For the DDR-related signature, Dactinomycin (a chromopeptide antibiotic that intercalates into DNA at GC-rich regions; $R = 0.51$, $p = 0.016$) showed the best statistically significant positive correlation in BRCA, NVP.ADW742 (a selective small molecule inhibitor of IGF-1R tyrosine kinase that prevents autophosphorylation; $R = 0.77$, $p = 0.026$) in COAD, Ceralasertib (a orally bioavailable inhibitor that blocks DNA damage checkpoint signalling; $R = 0.72$, $p = 0.028$) in LUAD, Adavosertib (First-in-class selective inhibitor of WEE1 kinase which disables G2/M

checkpoint; R = 0.79, p = 0.011) in LUSC, Vincristine (a vinca alkaloid that binds β-tubulin and inhibits microtubule polymerization; R = 0.84, p = 0.076) in PAAD, and AZ6102 (a potent dual inhibitor of tankyrase 1/2; R = 0.66, p = 0.00091) in STAD, respectively (see Figures 4.20–4.25e). While Afatinib (an irreversible ErbB family inhibitor and covalently binds to EGFR, HER2/4; R = -0.64, p = 0.019), Dactinomycin (R = -0.42, p = 0.042), BEZ235 (R = -0.83, p = 0.085), Topotecan (a topoisomerase I inhibitor that stabilizes the cleavable complex between DNA and topoisomerase I; R = -0.74, p = 0.058), GNE.317 (a brain-penetrant dual PI3K/mTOR inhibitor; R = -0.85, p = 0.033), and Podophyllotoxin-Bromide (R = -0.75, p = 0.0081) displayed the best negative correlation in BRCA, COAD, LUAD, LUSC, PAAD, and STAD, respectively (see Figures 4.20–4.25f).

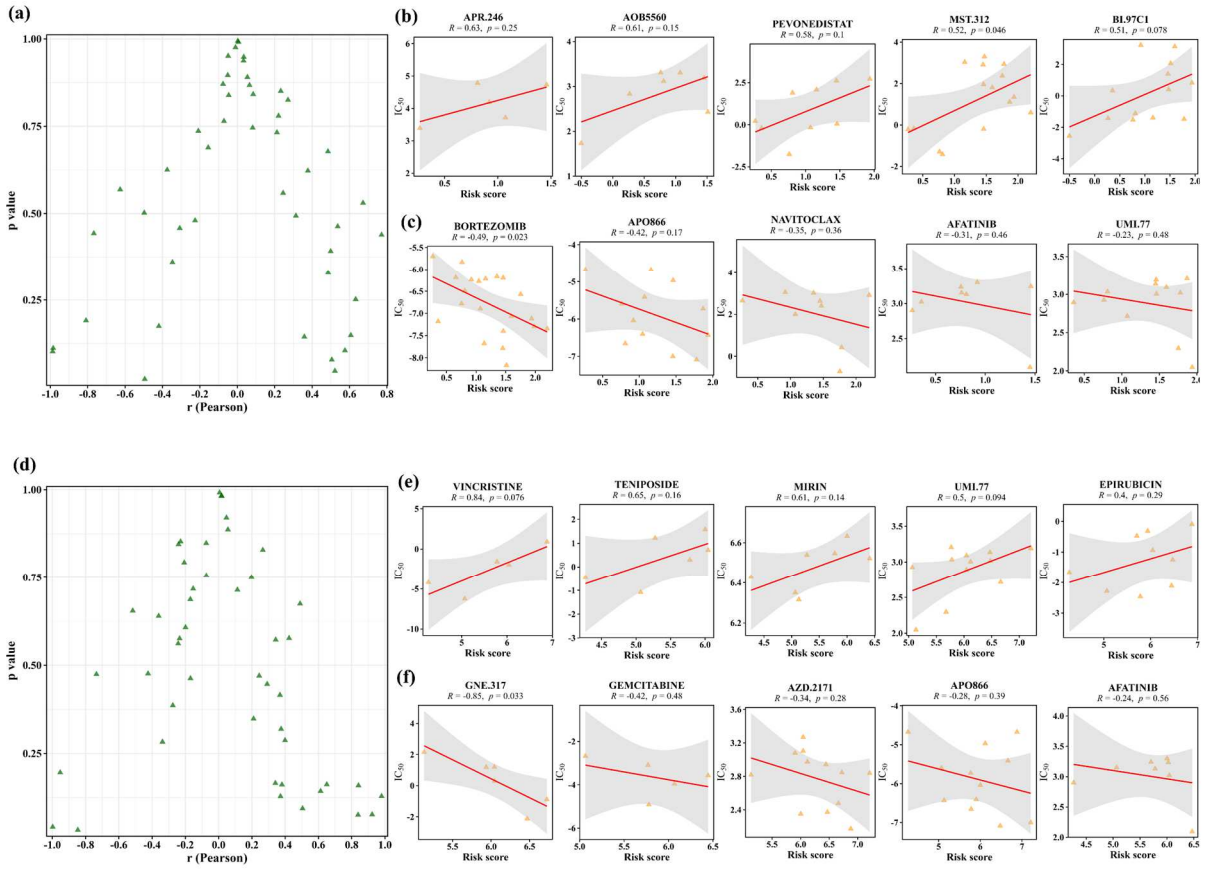


Figure 4.24. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in PAAD cell lines. (a, d) Correlation of EMT and DDR calculated risk score with IC_{50} of drugs in PAAD cell lines. (b, e) The Pearson's positive correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

(c, f) The Pearson's negative correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

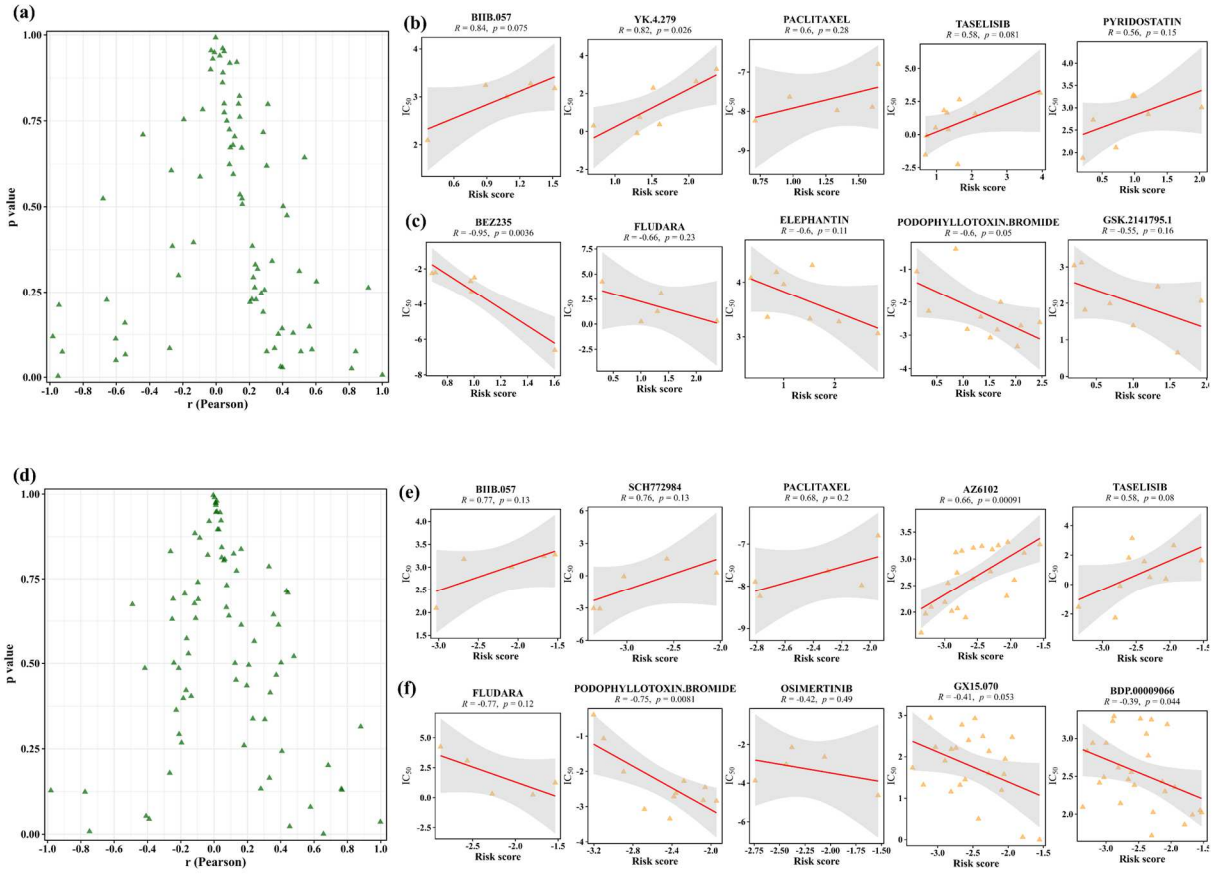


Figure 4.25. Drug resistance and sensitivity analysis of prognostic values of EMT and DDR gene signatures in STAD cell lines. (a, d) Correlation of EMT and DDR calculated risk score with IC_{50} of drugs in STAD cell lines. (b, e) Pearson's positive correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes. (c, f) The Pearson's negative correlation coefficient between the IC_{50} of the top five drugs and the risk score of EMT and DDR prognostic genes.

CHAPTER 5

DISCUSSION

In this study, we performed differential analysis to screen out the DGE by univariate Cox analysis based on the expression profile of six cancers between tumour and corresponding normal tissues. DGE were filtered out by applying specific threshold criteria via two methods: Cox regression and Kaplan-Meier survival analysis for both EMT and DDR sets across six solids tumours, including BRCA, COAD, LUAD, LUSC, PAAD, and STAD. Subsequently, the LASSO regression model was built on the overlapped genes for EMT and DDR individually and finally found the prognostic EMT- and DDR-related gene signatures. The patients were divided into high- and low-risk groups based on optimal cut-off values in each cancer for both prognostic signatures. Independent prognostic factors were identified using univariate and multivariate Cox regression analysis with clinicopathological characteristics. Then the relationship between clinical parameters and EMT- and DDR-related signature was determined, followed by KEGG and GO-BP analysis to find the top enriched pathway in novel EMT- and DDR-related signature for all six cancers. Immune cell types of infiltration score were also calculated using the CIBERSORTx algorithm and ESTIMATE (ESTIMATE, stromal, and immune scores) algorithms, and correlation was determined with respect to low- and high-risk groups of EMT- and DDR-related signatures in all cancers. Furthermore, the nomogram model was constructed and validated using calibration curves, DCA, C-index, and AIC values for DDR- and EMT-derived risk scores in all six cancers. Finally, drug screening correlation analysis was performed to find the positively and negatively correlated drugs based on prognostic EMT- and DDR-related gene signatures. The prognostic EMT- and DDR-related gene signatures were also validated in GSE20685, GSE39582, GSE50081, GSE30219, GSE79668, and GSE84433 cohorts.

It is known from literature that in EMT, cancer cells change their non-invasive to invasive state by losing epithelial polarized phenotype assets to gain migratory mesenchymal stem cell-like properties. This progression of EMT is characterized by loss of expression of epithelial markers such as E-Cadherin, p-120, and β-Catenin and gain of mesenchymal markers including N-Cadherin, *ZEB*, *SNAIL,* and *TWIST*, cell polarity, and tight junctions that are regulated by

various molecular pathways (Gottumukkala et al., 2024). On the other hand, we investigated another hallmark of cancer, DNA damage response (DDR), involves several DNA repair pathways that work collectively to help maintain the integrity of the genome from a wide range of challenges. Even under normal conditions, the DDR is constantly activated to manage the issues caused by mismatches or the complete collapse of the DNA replication fork (Jackson & Bartek, 2009). However, any failure to maintain genome integrity could lead to serious human consequences, including cancer and premature aging. Undoubtedly, several components of DDR pathways have been studied in detail, but the DDR is buffered by several factors, such as pathways or overlapping functions that can obscure the role of specific genes. These layers of regulation making it challenging to define their precise contribution in DNA repair processes (Cortez, 2019; Fielden et al., 2025). To combat this challenge, a comprehensive systemic approach is essential, particularly focused on mapping genetics in DDR. These kinds of interactions could help reveal novel biological relationships and new levels of regulation within genomic machinery.

Based on Cox, Kaplan-Meier, and LASSO regression model, the main finding of the study identified the six-gene novel EMT-signature in BRCA (*EMP3, GLIPR1, PDLIM4, SCG2, SDC1, TFPI2*) and COAD (*CRLF1, ENO2, FUCA1, PCOLCE2, SGCG, TIMP1*), nine-gene signature in LUAD (*ABI3BP, DKK1, GPC1, ID2, MGP, PMEPA1, PVR, SERPINH1, SPOCK1*), four-gene signature in LUSC (*FSTL3, PTX3, TGM2, TNFRSF12A*), ten-gene signature in PAAD (*AREG, CD44, DKK1, IGFBP2, ITGB5, LAMA3, MAGEE1, MMP3, PVR, SPP1*) and nine-gene signature in STAD (*CDH6, CXCL1, DKK1, GJA1, MATN3, NT5E, SERPINE1, SGCG, TFPI2*). While for DDR we identified novel six-gene DDR-related signature in BRCA (*CASP9, CCND2, DDB2, PMAIP1, RAD1, TP53AIP1*), two-gene signature in COAD (*ATM, GADD45B*), five-gene signature in LUAD (*CCND2, CDC25C, CYCS, GADD45G, RRM2B*), four-gene signature in LUSC (*ABL1, CCND2, CHEK2, GADD45G*), eight-gene signature in PAAD (*ATR, CCNB2, CCND2, CDK1, CDK6, NBN, PML, RAD52*), and seven-gene signature in STAD (*BBC3, CASP3, CDK5, GADD45B, PML, SMC1A, TNFRSF10B*).

In breast cancer, *CASP9* (which encodes a tumour suppressor caspase) is recognized as a hub gene in the inflammation, exhibiting a genetic variant that was associated with its mRNA expression level (M. Zhang et al., 2022). Based on RNA-sequencing data, the high expression

of *CASP9* is associated with a favourable OS role in patients with estrogen-receptor-positive (ER^+) breast cancer (Hoppe et al., 2025). A study revealed that *CCND2* (cyclin D2) is involved in the progression of the cell cycle: in particular, they stated that *CCND2* was significantly associated with colorectal tissues, and its upregulation led to poor OS (W. Zhang et al., 2022). *DDB2* (Damage-specific DNA binding protein 2) is known as a tumour suppressor gene and a key player in EMT, and promotes DNA repair in various cancers (Perucca et al., 2024). *PMAIP1* (encoded NOXA) is a key component of pro-apoptotic BCL2 family that induces apoptosis and directly associates with p53 transcriptional target and is often downregulated in cancer (Kuchur et al., 2021). *RAD1* is mainly involved in maintaining genome stability, and its mutation is associated with impaired checkpoint activation (Al-Zain et al., 2023). *TP53AIP1* (Tumour protein p53-Regulated apoptosis-inducing protein 1) is a key component of pro-apoptotic gene that is regulated by p53. In response to DNA damage, it promotes cell death; its downregulation is associated with cancer progression (Pan et al., 2025). In human cancer, *DKK1* (Dickkopf-1) is known as Wnt/β-catenin inhibitor and is associated with cancer fibroblast (Lee et al., 2025); *GPC1* (glypican) is a key player of glycosylphosphatidylinositol-anchored cell surface protein that is upregulated in pancreatic cancer (Li et al., 2023). *SNAI1* is an important transcriptional factor of EMT, which represses the activity of promoter genes that encode epithelial markers (Fan et al., 2025); *SPOCK1* is a Ca^{2+} binding proteoglycan that is involved in acts as an extracellular matrix in EMT (Slieker et al., 2023).

Among the identified prognostic EMT gene signatures across cancers, *SCG2* (Secretogranin II) (Yang et al., 2025), *TFPI2* (Tissue factor pathway inhibitor 2; a secreted Kunitz-type protease inhibitor) (Zhao et al., 2020), *CRLF1* (Cytokine receptor-like factor 1) (Li et al., 2021), *ENO2* (known as neuron-specific enolase; act as a neuroendocrine tumour marker) (Cai et al., 2024), *FUCA1* (Fucosidase, α-L-1; act as a tumour suppressor) (Zhao et al., 2024), *PCOLCE2* (Procollagen C-endopeptidase enhancer 2; involved in stroma remodelling) (Xie et al., 2025), *TIMP1* (Tissue inhibits of metalloproteinases 1; inhibitor of metalloproteinases) (Jackson et al., 2017), *DKK1* (Dickkopf-1; act as Wnt antagonist) (Zhuang et al., 2017), *SPOCK1* (SPARC/Osteonectin like 1; act as a secreted ECM glycoprotein) (Liu et al., 2023), *FSTL3* (Follistatin-like 3; act as a secreted glycoprotein) (Li et al., 2024), *PTX3* (Pentraxin 3; vital component of innate immune system) (Bonavita et al., 2015), *IGFBP2* (Insulin-like growth factor binding protein 2; act as a secreted protein that bind to IGF 1/2) (Li et al., 2020), *MMP3*

(Matrix metalloproteinases 3; act as a secreted protease and involved in ECM remodelling) (Gobin et al., 2019), *SPP1* (Osteopontin; known as secreted phosphoprotein and involved in pro-tumorigenic cytokine) (Su et al., 2024), *CXCL1* (C-X-C motif chemokine ligand 1; secreted chemokine and involved in neutrophil recruitment) (Feliciano, 2012), *MATN3* (Matrilin-3; secreted ECM glycoprotein) (Qin et al., 2025), and *SERPINE1* (known as plasminogen activator inhibor-1; a secreted serine protease inhibitor) (Zhu et al., 2025) are directly involved as soluble proteins. Some others prognostic EMT-related genes are identified as soluble proteins only after shedding/cleavage, including *GLIPR1* (Human glioma pathogenesis-related protein 1; only become soluble as *sGLIPR1* via proteolytic shedding) (Asojo et al., 2011), *SDC1* (Syndecan-1; only become soluble in sSDS1 domain when ectodomain is cleaved from membrane-bound protein) (Yao et al., 2019), *GPC1* (Glypican-1; only become soluble via proteolytic cleavage) (Quach et al., 2019), *PVR* (known as CD155; only soluble as sPVR domain and act an immune checkpoint regulator) (Stamm et al., 2018), *TNFRSF12A* (TWEAK receptor; only soluble as sFn14 domain via proteolytic shedding and act as a vital biomarker in cancers) (Xu et al., 2025) and *AREG* (Amphiregulin; act as a secreted growth factor; initially produced as transmembrane but then via shedding process become soluble protein) (M. S. Lee et al., 2016). In DDR-related prognostic signatures across cancers no gene were directly involved as soluble protein, rather only few genes act as soluble proteins under certain conditions such as *CYCS* (Cytochrome C), naturally found inside mitochondria but during apoptosis it could escapes into cell fluids (Liu et al., 1996), *GADD25A/GADD25B/GADD45G* (Growth arrest and DNA damage-inducible protein 45 alpha/beta/gamma), is only secreted during stressful conditions like cancer and inflammation (Palomer et al., 2024), *TNFRSF10B* (also known as Death receptor 5), which only becomes soluble in its sTNFRSF10B form via metalloprotease cleavage (Wagner et al., 2007). The accumulative data reveal that soluble proteins such as *TIMP1*, *DKK1*, *IGFBP2*, and *SPP1* have been reported in serum or plasma and are associated with cancer prognosis, and therapy response across cancers (Birgisson et al., 2010; Shen et al., 2012; Su et al., 2024). Based on our results, these prognostic soluble proteins may not only stratify patients but could also provide a non-invasive means of monitoring disease progression and treatment efficacy through liquid biopsy approaches.

Using TCGA and GEO (validation cohort) datasets, we identified the prognostic values of EMT- and DDR-related signatures. The results showed robust prognostic values, especially in

PAAD, which presented the highest AUC values for both models in training and validation cohorts. The findings indicate that our risk scores have the ability to predict prognostic models. We performed PCA and t-SNE algorithms for dimensionality reduction, and results confirmed that patients in low- and high-risk groups are distributed in various directions for both gene signatures across cancers, further confirming the prognostic ability of our models. Notably, univariate Cox and multivariate Cox regression analysis confirmed the risk score as an independent prognostic factor for both prognostic models across cancers in the training cohort while in the validation cohort the risk score was recognized as an independent prognostic factor in BRCA, LUAD, PAAD and STAD. In Kaplan-Meier survival analysis and clinicopathology subgroups for EMT- and DDR-related signatures, we found that the patients in the high-risk group showed the worst OS compared to the low-risk group in < 60, ≥ 60, TMN stage, N stage, and M stage across cancers in training and validation cohorts. Further, we identified the relationship between clinical parameters and EMT- and DDR-derived risk groups and found that high-risk scores possess more cases in BRCA, COAD, LUAD, and STAD, while in LUSC and PAAD, we found a converse pattern where more cases were reported in low-risk groups. The results are likely to be show that high-risk groups have advanced tumour stage (stage III-IV), distant metastasis (M1), tumour invasion (T3-T4), and lymph node (N1-N3), especially in BRCA, COAD, LUAD, and STAD cancers for both signatures. Our findings consistently demonstrated that high-risk groups that were stratified by EMT- and DDR-related signatures were associated with poor OS across cancers. A study documented that higher expression of *TNXB, MYOCD, ABI3BP, DNAH9*, and *SLIT3* was associated with poor OS in LUSC patients (Li et al., 2022).

We further identified the functional enrichment analysis to support the biological relevance of EMT- and DDR-related signatures. It was noted that EMT-related signature was primarily enriched in distinct pathways associated with epithelial differentiation, immune regulation, and extracellular matrix organization, which are key components of EMT drivers. Whereas in DDR-related signatures, cell regulation, chromosomal segregation, metabolic processes, and extracellular structure organization were the most enriched pathways, reflecting their role in genomic stability during cancer progression. These findings indicated that the identified novel EMT- and DDR-related prognostic signatures are not only statistically significant but also biologically meaningful. A study reported the GO and KEGG analysis based on the expression

of the *SLC35A2* prognostic gene and revealed that neuroactive ligand-receptor interaction, neuropeptide signalling pathway, and immunoglobulin complex were the most enriched pathways, respectively, in BC (Yan & Chen, 2025). Extracellular matrix organization (ECM) helped to stiffen and align collagen fibres at the proximal and distal invasive front, originating directional tracks that are associated with cancer cell local invasion and metastasis (Maiques et al., 2025). Chromosomal instability refers to the ability of cells to acquire errors in several chromosomal or structural abnormalities during cell division that contribute to genetic variation within tumours, which is considered a hallmark of cancer biology, such as progression, metastasis and immune evasion, and drug resistance (Chen et al., 2025).

The immune infiltration analysis suggested a strong association between risk scores stratified by EMT- and DDR-related signatures and tumour immune microenvironment. It was noted that high-risk groups were strongly infiltrated in immunosuppressive cells, including naïve B cells, macrophages M0, macrophages M2, and neutrophils, while the low-risk groups were highly infiltrated in plasma cells, macrophages M1, and memory B cells, reflecting their role in anti-tumour immunity. The findings recommended that immune profiles were not only distinct but also prognostically informative, suggesting that immune infiltration plays a significant role in modulating the risk associated with EMT and DDR signatures. The results of the ESTIMATE algorithm revealed that in BRCA and LUAD, low-risk samples consistently showed higher ESTIMATE, stromal, and immune scores ($p < 0.001$), indicating that these samples harbour richer stromal and immune infiltrates. While in LUSC and STAD, conversely, high-risk samples showed the higher ESTIMATE metrics score, suggesting the most aggressive lesions in the tissues. Our results are in agreement with the results of Lin et al. (Lin et al., 2020), who reported that high-risk showed the highest infiltration in macrophages M0/M1 and naïve B cells, whereas dendritic cells activated, resting dendritic cells and memory B cells exhibit strong infiltration in low-risk group in bladder cancer. Wang et al. (Wang et al., 2021) demonstrated that activated memory CD4+ T cells, macrophages M1, and regulatory T cells (Tregs) exhibited significant negative association with ECI (EMT-CYT Index), while, in contrast, macrophages M0/M1, B cell naïve, and resting memory CD4+T cells possessed positive association with ECI in the pan-cancer. M2 macrophages generally participate in the promotion of cell proliferation in EMT-linked processes, including tissue repair and wound healing, while macrophages are considered pro-inflammatory and induce anti-tumour immunity (Mantovani et al., 2002).

To further validate our results, we constructed a nomogram of EMT- and DDR-derived risk scores and clinical parameters to predict OS across six cancers for training (TCGA) and validation cohorts (GEO). The findings reveal that by combining clinical characteristics (age + TNM stage), this model provides better 1-, 3-, and 5-year OS estimation than individual factors. The nomogram consistently showed strong predictive performance by gaining a high AUC from ROC analysis and outperformed age, TNM stage, and risk score alone. DCA further validates the clinical utility of the model, presenting higher net benefits across threshold probabilities, suggesting the better performance of a model in decision-making in clinical settings. Yan and Chen (2025) constructed a comprehensive nomogram for 1-, 3-, 5, and 10-year OS in breast cancer and reported the robust alignment of clinical parameters with an ideal model, emphasizing the role of *SLC35A2* as a key prognostic contributor in BC patients. Huang et al. (Huang et al., 2025) constructed the nomogram model for male breast cancer and predicted the OS for 3-, 5- and 8-year probabilities. They find the C-index of 0.845, 0.803 and 0.798 in the train set and 0.775, 0.807, and 0.784 in the training set for 3-, 5- and 8-year OS, respectively. The DCA curves revealed the favourable net benefits in training as well as in trait sets.

We further calculated the EMT and DDR-related ssGSEA scores and found that the score was significantly higher in high-risk groups, particularly in LUAD, PAAD, and STAD ($p < 0.001$), emphasizing the role of EMT in immune evasion, metastasis, and therapy resistance. Interestingly, DDR-based calculated ssGSEA score showed a divergent trend across cancers. In COAD, LUSC, and STAD, high-risk group scores showed higher scores reflecting the hyperactivity of DNA repair that could be associated with survival under genomic stress. The lower score of high-risk groups in BRCA, LUAD, and PAAD could contribute to genomic instability and impaired repair capacity. Hence, the results reported the tissue-specific behaviour of DDR in cancer progression. However, the overlap results revealed that in BRCA, LUAD, and LUSC, the vast majority of patients fell into the low-EMT/low-DDR quadrant ($p < 0.001$), suggesting that these tumours with EMT programs also tend to have intact DNA damage response. While COAD and STAD, mirror image, where most patients fell into the high-EMT/high-DDR quadrant ($p < 0.001$), which indicated that EMT and DDR impairment work in a parallel way to drive tumour aggressiveness.

Furthermore, to capture functional chemosensitivity, we performed drug sensitivity correlation analysis across cancers for our prognostic EMT- and DDR-based risk scores with IC50 values

of a broad panel of drugs (Pearson's correlation). Our pharmacogenomic analysis yielded highly cancer-specific results, where high-risk groups in BRCA and COAD indicate hypersensitivity, whereas in LUAD, LUSC, PAAD, and STAD, they signify drug resistance. Based on the results, it is suggested that tissue-dependent relationships will be vital for exploiting EMT and DDR stratification to guide combination treatments. Our correlation results reported that IC_{50} of BDP-00009066, Lapatinib, Vorinostat, MST-312, and YK-4.279 showed a positive correlation with risk score for EMT- and DDR-related prognostic gene signatures across cancers. Jia et al. (Jia et al., 2022) reported that the results were positively correlated between risk score and IC_{50} of most antineoplastic drugs, indicating that high-risk scores were associated with drug resistance. They also found that IC_{50} values of carboplatin and 5-fluorouracil were positively correlated with the risk score. Similarly, drug sensitivity analysis conducted by Han et al. (Han et al., 2024) revealed selected IC_{50} values for drugs, including BDP-00009066, in low-risk group for colorectal cancer patients.

However, despite these insights, our study has several limitations. First, the study was based on retrospective transcriptomic datasets (TCGA and GEO); *in vitro* and *in vivo* validation is essential to clarify and understand the role of identified EMT- and DDR-related prognostic signatures. Second, the study lacks treatment data integration; the model does not account for therapeutic regimens such as chemotherapy or radiotherapy.

CHAPTER 6

CONCLUSIONS

In the present study, we established and validated robust prognostic gene signatures related to EMT and DDR across six major cancer types (BRCA, COAD, LUAD, PAAD and STAD). The findings confirmed that EMT- and DDR-derived risk scores serve as independent prognostic factors in multivariate models across these cancers. A strong correlation was observed between clinicopathological characteristics and the EMT- and DDR-derived risk scores, where higher tumour stage was consistently associated with a higher risk score. Functional enrichment analysis revealed that the EMT-related signature was significantly enriched not only in classical EMT pathways such as ECM organisation but also in immune-related signalling pathways. In contrast, the DDR-related signature was closely related to cell cycle regulation, genomic stability, and cellular signalling in BRCA. Both prognostic EMT- and DDR-related signatures demonstrated strong and consistent prognostic values, with high-risk groups exhibiting greater tumour aggressiveness and distinct immune cell infiltration patterns. The clinically integrated nomogram model, which combined risk scores with standard clinical variables (age and TNM stage), demonstrated superior predictive accuracy for 1-, 3-, and 5-year OS compared to individual factors. The nomogram consistently demonstrated strong predictive performance, yielding a high AUC from the ROC analysis. The drug sensitivity analysis indicated that EMT and DDR-based risk scores could predict divergent chemosensitivity profiles in a cancer-specific manner. The validation cohorts reported similar results to the training cohorts and confirmed the prognostic ability of identified EMT- and DDR-related gene signatures with OS. Furthermore, the findings highlight a critical role of EMT- and DDR-derived prognostic signatures in cancer progression, tumour microenvironment, and drug sensitivity. Based on results, we concluded that EMT- and DDR-related prognostic gene signatures represent effective biomarkers with potential as therapeutic targets in the treatment of these six cancers. However, further functional studies are warranted to validate these signatures in pre-clinical models, aiming to identify clinical applications of these signatures in the individualised treatment of BRCA, COAD, LUAD, LUSC, PAAD, and STAD.

BIBLIOGRAPHY

Al-Zain, A. M., Nester, M. R., Ahmed, I., & Symington, L. S. (2023). Double-strand breaks induce inverted duplication chromosome rearrangements by a DNA polymerase δ-dependent mechanism. *Nature communications*, *14*(1), 7020.

Amodio, V., Vitiello, P., Bardelli, A., & Germano, G. (2024). DNA repair-dependent immunogenic liabilities in colorectal cancer: opportunities from errors. *British Journal of Cancer*, *131*(10), 1576-1590.

Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C., & Vazquez, F. (2025). The present and future of the Cancer Dependency Map. *Nature Reviews Cancer*, *25*(1), 59-73.

Asojo, O. A., Koski, R. A., & Bonafé, N. (2011). Structural studies of human glioma pathogenesis-related protein 1. *Biological Crystallography*, *67*(10), 847-855.

BeLow, M., & Osipo, C. (2020). Notch signaling in breast cancer: a role in drug resistance. *Cells*, *9*(10), 2204.

Bieging, K. T., Mello, S. S., & Attardi, L. D. (2014). Unravelling mechanisms of p53-mediated tumour suppression. *Nature Reviews Cancer*, *14*(5), 359-370.

Birgisson, H., Nielsen, H. J., Christensen, I. J., Glimelius, B., & Brünner, N. (2010). Preoperative plasma TIMP-1 is an independent prognostic indicator in patients with primary colorectal cancer: a prospective validation study. *European journal of cancer*, *46*(18), 3323-3331.

Boehm, J. S., Garnett, M. J., Adams, D. J., Francies, H. E., Golub, T. R., Hahn, W. C., Iorio, F., McFarland, J. M., Parts, L., & Vazquez, F. (2021). Cancer research needs a better map. *Nature*, *589*(7843), 514-516.

Bonavita, E., Gentile, S., Rubino, M., Maina, V., Papait, R., Kunderfranco, P., Greco, C., Feruglio, F., Molgora, M., & Laface, I. (2015). PTX3 is an extrinsic oncosuppressor regulating complement-dependent inflammation in cancer. *cell*, *160*(4), 700-714.

BIBLIOGRAPHY

Cai, J., Yang, Y., Zhang, L., Fang, Y., Zhang, Y., Tan, M., Zhang, J., Tang, C., Ren, H., & Wang, L. (2024). Investigation of ENO2 as a promising novel marker for the progression of colorectal cancer with microsatellite instability-high. *BMC cancer*, *24*(1), 573.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N., & Murray, B. A. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, *48*(6), 607-616.

Chang, Y., Huang, Z., Quan, H., Li, H., Yang, S., Song, Y., Wang, J., Yuan, J., & Wu, C. (2023). Construction of a DNA damage repair gene signature for predicting prognosis and immune response in breast cancer. *Frontiers in oncology*, *12*, 1085632.

Chen, J., Song, Y., Li, M., Zhang, Y., Lin, T., Sun, J., Wang, D., Liu, Y., Guo, J., & Yu, W. (2021). Comprehensive analysis of ceRNA networks reveals prognostic lncRNAs related to immune infiltration in colorectal cancer. *BMC cancer*, *21*(1), 255.

Chen, X., Agustinus, A. S., Li, J., Dibona, M., & Bakhoum, S. F. (2025). Chromosomal instability as a driver of cancer progression. *Nature Reviews Genetics*, *26*(1), 31-46.

Chi, M., Liu, J., Mei, C., Shi, Y., Liu, N., Jiang, X., Liu, C., Xue, N., Hong, H., & Xie, J. (2022). TEAD4 functions as a prognostic biomarker and triggers EMT via PI3K/AKT pathway in bladder cancer. *Journal of Experimental & Clinical Cancer Research*, *41*(1), 175.

Chowdhury, D., Mistry, A., Maity, D., Bhatia, R., Priyadarshi, S., Wadan, S., Chakraborty, S., & Haldar, S. (2024). Pan-cancer analyses suggest kindlin-associated global mechanochemical alterations. *Communications Biology*, *7*(1), 372.

Cortez, D. (2019). Replication-coupled DNA repair. *Molecular cell*, *74*(5), 866-876.

Cui, Y., Wang, X., Zhang, L., Liu, W., Ning, J., Gu, R., Cui, Y., Cai, L., & Xing, Y. (2022). A novel epithelial-mesenchymal transition (EMT)-related gene signature of predictive value for the survival outcomes in lung adenocarcinoma. *Frontiers in oncology*, *12*, 974614.

Dai, J., Jiang, M., He, K., Wang, H., Chen, P., Guo, H., Zhao, W., Lu, H., He, Y., & Zhou, C. (2021). DNA damage response and repair gene alterations increase tumor mutational burden and promote poor prognosis of advanced lung cancer. *Frontiers in oncology*, *11*, 708294.

Dai, W., Xiao, Y., Tang, W., Li, J., Hong, L., Zhang, J., Pei, M., Lin, J., Liu, S., & Wu, X. (2021). Identification of an EMT-related gene signature for predicting overall survival in gastric cancer. *Frontiers in Genetics*, *12*, 661306.

de Assis, J. V., Coutinho, L. A., Oyeyemi, I. T., & Oyeyemi, O. T. (2022). Diagnostic and therapeutic biomarkers in colorectal cancer: a review. *American journal of cancer research*, *12*(2), 661.

De Rosa, C., De Rosa, V., Tuccillo, C., Tirino, V., Amato, L., Papaccio, F., Ciardiello, D., Napolitano, S., Martini, G., & Ciardiello, F. (2024). ITGB1 and DDR activation as novel mediators in acquired resistance to osimertinib and MEK inhibitors in EGFR-mutant NSCLC. *Scientific reports*, *14*(1), 500.

Den Hollander, P., Maddela, J. J., & Mani, S. A. (2024). Spatial and Temporal relationship between epithelial–mesenchymal transition (EMT) and stem cells in cancer. *Clinical chemistry*, *70*(1), 190-205.

Diehl, F., Schmidt, K., Choti, M. A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., & Szabo, S. A. (2008). Circulating mutant DNA to assess tumor dynamics. *Nature medicine*, *14*(9), 985-990.

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., & Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer*, *17*(2), 79-92.

Ding, K., He, Y., Wei, J., Fu, S., Wang, J., Chen, Z., Zhang, H., Qu, Y., Liang, K., & Gong, X. (2022). A score of DNA damage repair pathway with the predictive ability for chemotherapy and immunotherapy is strongly associated with immune signaling pathway in pan-cancer. *Frontiers in Immunology*, *13*, 943090.

Drew, Y., Zenke, F. T., & Curtin, N. J. (2025). DNA damage response inhibitors in cancer therapy: lessons from the past, current status and future implications. *Nature Reviews Drug Discovery*, *24*(1), 19-39.

Fan, C., Wang, Q., Krijger, P. H., Cats, D., Selle, M., Khorosjutina, O., Dhanjal, S., Schmierer, B., Mei, H., & de Laat, W. (2025). Identification of a SNAI1 enhancer RNA that drives cancer cell plasticity. *Nature communications*, *16*(1), 2890.

Feliciano, P. (2012). CXCL1 and CXCL2 link metastasis and chemoresistance. *Nature Genetics*, *44*(8), 840-840.

Ferlay J, E. M., Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. (2024). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today, accessed [02 05 2025]. In.

Fielden, J., Siegner, S. M., Gallagher, D. N., Schröder, M. S., Dello Stritto, M. R., Lam, S., Kobel, L., Schlapansky, M. F., Jackson, S. P., & Cejka, P. (2025). Comprehensive interrogation of synthetic lethality in the DNA damage response. *Nature*, 1-10.

Gandhi, L., Rodríguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., Domine, M., Clingan, P., Hochmair, M. J., & Powell, S. F. (2018). Pembrolizumab plus chemotherapy in metastatic non–small-cell lung cancer. *New England Journal of Medicine*, *378*(22), 2078-2092.

Georgakopoulos-Soares, I., Chartoumpekis, D. V., Kyriazopoulou, V., & Zaravinos, A. (2020). EMT factors and metabolic pathways in cancer. *Frontiers in oncology*, *10*, 499.

Gobin, E., Bagwell, K., Wagner, J., Mysona, D., Sandirasegarane, S., Smith, N., Bai, S., Sharma, A., Schleifer, R., & She, J.-X. (2019). A pan-cancer perspective of matrix metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. *BMC cancer*, *19*(1), 581.

Golan, T., Hammel, P., Reni, M., Van Cutsem, E., Macarulla, T., Hall, M. J., Park, J.-O., Hochhauser, D., Arnold, D., & Oh, D.-Y. (2019). Maintenance olaparib for germline

BRCA-mutated metastatic pancreatic cancer. *New England Journal of Medicine*, *381*(4), 317-327.

Gottumukkala, S. B., Ganesan, T. S., & Palanisamy, A. (2024). Comprehensive molecular interaction map of TGFβ induced epithelial to mesenchymal transition in breast cancer. *NPJ systems biology and applications*, *10*(1), 53.

Groelly, F. J., Fawkes, M., Dagg, R. A., Blackford, A. N., & Tarsounas, M. (2023). Targeting DNA damage response pathways in cancer. *Nature Reviews Cancer*, *23*(2), 78-94.

Guinney, J., Dienstmann, R., Wang, X., De Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., & Angelino, P. (2015). The consensus molecular subtypes of colorectal cancer. *Nature medicine*, *21*(11), 1350-1356.

Han, J., Wang, Q., Li, S., Yang, J., Qiu, Z., & Fu, W. (2024). Comprehensive analysis of basement membrane-related gene based on single-cell and bulk RNA sequencing data to predict prognosis and evaluate immune characteristics in colorectal cancer. *Environmental Toxicology*, *39*(6), 3367-3380.

Han, P., Guda, C., & Liu, Q. (2025). An efficient epithelial-mesenchymal transition-related gene signature for predicting the survival of patients with lung adenocarcinoma. *Translational Cancer Research*, *14*(8), 4989.

Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer discovery*, *12*(1), 31-46.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, *144*(5), 646-674.

Haynes, B., Saadat, N., Myung, B., & Shekhar, M. P. (2015). Crosstalk between translesion synthesis, Fanconi anemia network, and homologous recombination repair pathways in interstrand DNA crosslink repair and development of chemoresistance. *Mutation Research/Reviews in Mutation Research*, *763*, 258-266.

He, J., & Jia, Y. (2021). The evolution and advances of biomarker use in clinical trials for breast cancer treatment—a narrative review. *Translational Breast Cancer Research*, *2*(1).

Heitmeir, B., Deniz, M., Janni, W., Rack, B., Schochter, F., & Wiesmüller, L. (2022). Circulating tumor cells in breast cancer patients: a balancing act between stemness, EMT features and DNA damage responses. *Cancers*, *14*(4), 997.

Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature*, *553*(7689), 446-454.

Hoppe, R., Winter, S., Lo, W.-Y., Michailidou, K., Bolla, M. K., Keeman, R., Wang, Q., Dennis, J., Lush, M., & Kalari, K. R. (2025). Lessons learned from a candidate gene study investigating aromatase inhibitor treatment outcome in breast cancer. *npj Breast Cancer*, *11*(1), 18.

Huang, A., Li, D., Fan, Z., Chen, J., Zhang, W., & Wu, W. (2025). Long-term trends in the incidence of male breast cancer and nomogram for predicting survival in male breast cancer patients: a population-based epidemiologic study. *Scientific reports*, *15*(1), 2027.

Jackson, H. W., Defamie, V., Waterhouse, P., & Khokha, R. (2017). TIMPs: versatile extracellular regulators in cancer. *Nature Reviews Cancer*, *17*(1), 38-53.

Jackson, S. P., & Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, *461*(7267), 1071-1078.

Janjigian, Y. Y., Shitara, K., Moehler, M., Garrido, M., Salman, P., Shen, L., Wyrwicz, L., Yamaguchi, K., Skoczylas, T., & Bragagnoli, A. C. (2021). Nivolumab plus chemotherapy versus chemotherapy as first-line treatment for advanced gastric cancer/gastroesophageal junction cancer/oesophageal adenocarcinoma (CheckMate 649): a multicentre, randomised, open-label, phase 3 trial. *Lancet (London, England)*, *398*(10294), 27.

Jia, X., Chen, B., Li, Z., Huang, S., Chen, S., Zhou, R., Feng, W., Zhu, H., & Zhu, X. (2022). Identification of a four-gene-based SERM signature for prognostic and drug sensitivity prediction in gastric cancer. *Frontiers in oncology*, *11*, 799223.

Jones, W. D., Michener, C. M., Biscotti, C., Braicu, I., Sehouli, J., Ganapathi, M. K., & Ganapathi, R. N. (2020). RNA immune signatures from pan-cancer analysis are

prognostic for high-grade serous ovarian cancer and other female cancers. *Cancers*, *12*(3), 620.

Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, A. V., Neale, R. E., Tempero, M., Tuveson, D. A., & Hruban, R. H. (2016). Pancreatic cancer. *Nature reviews Disease primers*, *2*(1), 1-22.

Kopetz, S., Grothey, A., Yaeger, R., Van Cutsem, E., Desai, J., Yoshino, T., Wasan, H., Ciardiello, F., Loupakis, F., & Hong, Y. S. (2019). Encorafenib, binimetinib, and cetuximab in BRAF V600E–mutated colorectal cancer. *New England Journal of Medicine*, *381*(17), 1632-1643.

Kreis, J., Aybey, B., Geist, F., Brors, B., & Staub, E. (2023). Stromal signals dominate gene expression signature scores that aim to describe cancer-intrinsic stemness or mesenchymality characteristics. *bioRxiv*, 2023.2008. 2025.554747.

Kubota, Y., & Shitara, K. (2024). Zolbetuximab for Claudin18. 2-positive gastric or gastroesophageal junction cancer. *Therapeutic Advances in Medical Oncology*, *16*, 17588359231217967.

Kuchur, O., Kuchur, P., Kuzmina, D., Zavirsky, A., & Shtil, A. (2021). Differential Regulation of BBC3/PUMA and PMAIP1/Noxa in Ionizing Radiation: the Role of p53. *Cell and Tissue Biology*, *15*(6), 544-553.

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., & Laheru, D. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine*, *372*(26), 2509-2520.

Lee, H.-J., Li, C.-F., Ruan, D., Powers, S., Thompson, P. A., Frohman, M. A., & Chan, C.-H. (2016). The DNA damage transducer RNF8 facilitates cancer chemoresistance and progression through twist activation. *Molecular cell*, *63*(6), 1021-1033.

Lee, M. S., McGuffey, E. J., Morris, J. S., Manyam, G., Baladandayuthapani, V., Wei, W., Morris, V. K., Overman, M. J., Maru, D. M., & Jiang, Z.-Q. (2016). Association of CpG

island methylator phenotype and EREG/AREG methylation and expression in colorectal cancer. *British Journal of Cancer*, *114*(12), 1352-1361.

Lee, S., Ricci, B., Tran, J., Eul, E., Ye, J., Ren, Q., Clever, D., Wang, J., Wong, P., & Haas, M. S. (2025). Stroma-derived Dickkopf-1 contributes to the suppression of NK cell cytotoxicity in breast cancer. *Nature communications*, *16*(1), 1183.

Li, H., Zheng, N., Guo, A., Tang, W., Li, M., Cao, Y., Ma, X., Cao, H., Ma, Y., & Wang, H. (2024). FSTL3 promotes tumor immune evasion and attenuates response to anti-PD1 therapy by stabilizing c-Myc in colorectal cancer. *Cell Death & Disease*, *15*(2), 107.

Li, N., Quan, A., Li, D., Pan, J., Ren, H., Hoeltzel, G., de Val, N., Ashworth, D., Ni, W., & Zhou, J. (2023). The IgG4 hinge with CD28 transmembrane domain improves VHH-based CAR T cells targeting a membrane-distal epitope of GPC1 in pancreatic cancer. *Nature communications*, *14*(1), 1986.

Li, P., Yuan, H., Kuang, X., Zhang, T., & Ma, L. (2022). Network module function enrichment analysis of lung squamous cell carcinoma and lung adenocarcinoma. *Medicine*, *101*(47), e31798.

Li, T., Forbes, M. E., Fuller, G. N., Li, J., Yang, X., & Zhang, W. (2020). IGFBP2: integrative hub of developmental and oncogenic signaling network. *Oncogene*, *39*(11), 2243-2257.

Li, Y., Xun, J., Wang, B., Ma, Y., Zhang, L., Yang, L., Gao, R., Guan, J., Liu, T., & Gao, H. (2021). miR-3065-3p promotes stemness and metastasis by targeting CRLF1 in colorectal cancer. *Journal of Translational Medicine*, *19*(1), 429.

Lin, J., Yang, J., Xu, X., Wang, Y., Yu, M., & Zhu, Y. (2020). A robust 11-genes prognostic model can predict overall survival in bladder cancer patients based on five cohorts. *Cancer cell international*, *20*, 1-14.

Liu, X., Kim, C. N., Yang, J., Jemmerson, R., & Wang, X. (1996). Induction of apoptotic program in cell-free extracts: requirement for dATP and cytochrome c. *cell*, *86*(1), 147-157.

Liu, Y., Han, T., Wu, J., Zhou, J., Guo, J., Miao, R., Xu, Z., Xing, Y., Bai, Y., & Hu, D. (2023). SPOCK1, as a potential prognostic and therapeutic biomarker for lung adenocarcinoma, is associated with epithelial-mesenchymal transition and immune evasion. *Journal of Translational Medicine*, *21*(1), 909.

Lüönd, F., Sugiyama, N., Bill, R., Bornes, L., Hager, C., Tang, F., Santacroce, N., Beisel, C., Ivanek, R., & Bürglin, T. (2021). Distinct contributions of partial and full EMT to breast cancer malignancy. *Developmental cell*, *56*(23), 3203-3221. e3211.

Maiques, O., Sallan, M. C., Laddach, R., Pandya, P., Varela, A., Crosas-Molist, E., Barcelo, J., Courbot, O., Liu, Y., & Graziani, V. (2025). Matrix mechano-sensing at the invasive front induces a cytoskeletal and transcriptional memory supporting metastasis. *Nature communications*, *16*(1), 1394.

Mantovani, A., Sozzani, S., Locati, M., Allavena, P., & Sica, A. (2002). Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes. *Trends in immunology*, *23*(11), 549-555.

Meyer-Schaller, N., Cardner, M., Diepenbruck, M., Saxena, M., Tiede, S., Lüönd, F., Ivanek, R., Beerenwinkel, N., & Christofori, G. (2019). A hierarchical regulatory landscape during the multiple stages of EMT. *Developmental cell*, *48*(4), 539-553. e536.

Miller, R. E., Leary, A., Scott, C., Serra, V., Lord, C. J., Bowtell, D., Chang, D., Garsed, D. W., Jonkers, J., & Ledermann, J. (2020). ESMO recommendations on predictive biomarker testing for homologous recombination deficiency and PARP inhibitor benefit in ovarian cancer. *Annals of Oncology*, *31*(12), 1606-1622.

Moyret-Lalle, C., Prodhomme, M. K., Burlet, D., Kashiwagi, A., Petrilli, V., Puisieux, A., Seimiya, H., & Tissier, A. (2022). Role of EMT in the DNA damage response, double-strand break repair pathway choice and its implications in cancer treatment. *Cancer Science*, *113*(7), 2214-2223.

Network, C. G. A. R. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, *489*(7417), 519.

Network, C. G. A. R. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202.

Noordermeer, S. M., & van Attikum, H. (2019). PARP inhibitor resistance: a tug-of-war in BRCA-mutated cells. *Trends in cell biology*, *29*(10), 820-834.

Oxnard, G. R., Arcila, M. E., Sima, C. S., Riely, G. J., Chmielecki, J., Kris, M. G., Pao, W., Ladanyi, M., & Miller, V. A. (2011). Acquired resistance to EGFR tyrosine kinase inhibitors in EGFR-mutant lung cancer: distinct natural history of patients with tumors harboring the T790M mutation. *Clinical cancer research*, *17*(6), 1616-1622.

Palomer, X., Salvador, J. M., Griñán-Ferré, C., Barroso, E., Pallàs, M., & Vázquez-Carrera, M. (2024). GADD45A: With or without you. *Medicinal Research Reviews*, *44*(4), 1375-1403.

Pan, X., Du, X., & Jia, S. (2025). DNA methyltransferase DNMT3A inhibits TP53AIP1 expression and promotes cervical cancer development and metastasis. *Cytotechnology*, *77*(2), 1-16.

Patel, J. N., Ersek, J. L., & Kim, E. S. (2015). Lung cancer biomarkers, targeted therapies and clinical assays. *Translational lung cancer research*, *4*(5), 503.

Peng, Y., Li, H., Fu, Y., Guo, S., Qu, C., Zhang, Y., Zong, B., & Liu, S. (2022). JAM2 predicts a good prognosis and inhibits invasion and migration by suppressing EMT pathway in breast cancer. *International immunopharmacology*, *103*, 108430.

Perucca, P., Bassi, E., Vetro, M., Tricarico, A., Prosperi, E., Stivala, L. A., & Cazzalini, O. (2024). Epithelial-to-mesenchymal transition and NF-kB pathways are promoted by a mutant form of DDB2, unable to bind PCNA, in UV-damaged human cells. *BMC cancer*, *24*(1), 616.

Puisieux, A., Brabletz, T., & Caramel, J. (2014). Oncogenic roles of EMT-inducing transcription factors. *Nature cell biology*, *16*(6), 488-494.

Qin, C., Qin, H., Xie, H., Li, Y., Bi, A., Liao, X., Yang, K., Lu, C., Peng, T., & Zhu, G. (2025). The Role of MATN3 in Cancer Prognosis and Immune Infiltration Across Multiple Tumor Types. *Journal of Cancer*, *16*(5), 1519.

Quach, N. D., Kaur, S. P., Eggert, M. W., Ingram, L., Ghosh, D., Sheth, S., Nagy, T., Dawson, M. R., Arnold, R. D., & Cummings, B. S. (2019). Paradoxical role of glypican-1 in prostate cancer cell and tumor growth. *Scientific reports*, *9*(1), 11478.

Reck, M., Rodríguez-Abreu, D., Robinson, A. G., Hui, R., Csőszi, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., & Cuffe, S. (2016). Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer. *New England Journal of Medicine*, *375*(19), 1823-1833.

Renner, G., Janouskova, H., Noulet, F., Koenig, V., Guerin, E., Bär, S., Nuesch, J., Rechenmacher, F., Neubauer, S., & Kessler, H. (2016). Integrin α5β1 and p53 convergent pathways in the control of anti-apoptotic proteins PEA-15 and survivin in high-grade glioma. *Cell Death & Differentiation*, *23*(4), 640-653.

Russell, R., Perkhofer, L., Liebau, S., Lin, Q., Lechel, A., Feld, F. M., Hessmann, E., Gaedcke, J., Güthle, M., & Zenke, M. (2015). Loss of ATM accelerates pancreatic cancer formation and epithelial–mesenchymal transition. *Nature communications*, *6*(1), 7677.

Sengodan, S. K., Sreelatha, K., Nadhan, R., & Srinivas, P. (2018). Regulation of epithelial to mesenchymal transition by BRCA1 in breast cancer. *Critical reviews in oncology/hematology*, *123*, 74-82.

Shen, Q., Fan, J., Yang, X.-R., Tan, Y., Zhao, W., Xu, Y., Wang, N., Niu, Y., Wu, Z., & Zhou, J. (2012). Serum DKK1 as a protein biomarker for the diagnosis of hepatocellular carcinoma: a large-scale, multicentre study. *The lancet oncology*, *13*(8), 817-826.

Shi, Y., Wang, J., Huang, G., Zhu, J., Jian, H., Xia, G., Wei, Q., Li, Y., & Yu, H. (2022). A novel epithelial–mesenchymal transition gene signature for the immune status and prognosis of hepatocellular carcinoma. *Hepatology International*, *16*(4), 906-917.

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: a cancer journal for clinicians*, *74*(1), 12-49.

Skoulidis, F., & Heymach, J. V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nature Reviews Cancer*, *19*(9), 495-509.

Slieker, R. C., Donnelly, L. A., Akalestou, E., Lopez-Noriega, L., Melhem, R., Güneş, A., Abou Azar, F., Efanov, A., Georgiadou, E., & Muniangi-Muhitu, H. (2023). Identification of biomarkers for glycaemic deterioration in type 2 diabetes. *Nature communications*, *14*(1), 2533.

Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C., & Lordick, F. (2020). Gastric cancer. *The Lancet*, *396*(10251), 635-648.

Sok, C. P., Polireddy, K., & Kooby, D. A. (2023). Molecular pathology and protein markers for pancreatic cancer: relevance in staging, in adjuvant therapy, in determination of minimal residual disease, and follow-up. *Hepatobiliary Surgery and Nutrition*, *13*(1), 56.

Solomon, B. J., Mok, T., Kim, D.-W., Wu, Y.-L., Nakagawa, K., Mekhail, T., Felip, E., Cappuzzo, F., Paolini, J., & Usari, T. (2014). First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *New England Journal of Medicine*, *371*(23), 2167-2177.

Srinivasan, D., Balakrishnan, R., Chauhan, A., Kumar, J., Girija, D. M., Shrestha, R., Shrestha, R., & Subbarayan, R. (2025). Epithelial–Mesenchymal Transition in Cancer: Insights Into Therapeutic Targets and Clinical Implications. *MedComm*, *6*(9), e70333.

Stamm, H., Klingler, F., Grossjohann, E.-M., Muschhammer, J., Vettorazzi, E., Heuser, M., Mock, U., Thol, F., Vohwinkel, G., & Latuske, E. (2018). Immune checkpoints PVR and PVRL2 are prognostic markers in AML and their blockade represents a new therapeutic option. *Oncogene*, *37*(39), 5269-5280.

Su, X., Liang, C., Chen, R., & Duan, S. (2024). Deciphering tumor microenvironment: CXCL9 and SPP1 as crucial determinants of tumor-associated macrophage polarity and prognostic indicators. *Molecular Cancer*, *23*(1), 13.

Sukumar, J., Gast, K., Quiroga, D., Lustberg, M., & Williams, N. (2021). Triple-negative breast cancer: promising prognostic biomarkers currently in development. *Expert review of anticancer therapy*, *21*(2), 135-148.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

Thiery, J. P., Acloque, H., Huang, R. Y., & Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *cell*, *139*(5), 871-890.

Tripathi, D., Gupta, P. K., & Kulkarni, S. (2023). Epithelial mesenchymal transition: the ultimate driver of cancer on difficult Paths. *Journal of Cancer Immunology*, *5*(1), 40-60.

Unbekandt, M., Belshaw, S., Bower, J., Clarke, M., Cordes, J., Crighton, D., Croft, D. R., Drysdale, M. J., Garnett, M. J., & Gill, K. (2018). Discovery of potent and selective MRCK inhibitors with therapeutic effect on skin cancer. *Cancer research*, *78*(8), 2096-2114.

Vaughn, C. P., ZoBell, S. D., Furtado, L. V., Baker, C. L., & Samowitz, W. S. (2011). Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer. *Genes, Chromosomes and Cancer*, *50*(5), 307-312.

Wagner, K. W., Punnoose, E. A., Januario, T., Lawrence, D. A., Pitti, R. M., Lancaster, K., Lee, D., von Goetz, M., Yee, S. F., & Totpal, K. (2007). Death-receptor O-glycosylation controls tumor-cell sensitivity to the proapoptotic ligand Apo2L/TRAIL. *Nature medicine*, *13*(9), 1070-1077.

Wang, G., Xu, D., Zhang, Z., Li, X., Shi, J., Sun, J., Liu, H.-Z., Li, X., Zhou, M., & Zheng, T. (2021). The pan-cancer landscape of crosstalk between epithelial-mesenchymal transition and immune evasion relevant to prognosis and immunotherapy response. *NPJ precision oncology*, *5*(1), 56.

Wang, H., Bierie, B., Li, A. G., Pathania, S., Toomire, K., Dimitrov, S. D., Liu, B., Gelman, R., Giobbie-Hurder, A., & Feunteun, J. (2016). BRCA1/FANCD2/BRG1-driven DNA repair stabilizes the differentiation state of human mammary epithelial cells. *Molecular cell*, *63*(2), 277-292.

Wu, R., Ma, R., Duan, X., Zhang, J., Li, K., Yu, L., Zhang, M., Liu, P., & Wang, C. (2023). Identification of specific prognostic markers for lung squamous cell carcinoma based on tumor progression, immune infiltration, and stem index. *Frontiers in Immunology*, *14*, 1236444.

Wu, X., Cai, L., Liu, Y., Wang, B., Xia, T., & Liu, Z. (2025). Multi-omics analysis of parthanatos related molecular subgroup and prognostic model development in stomach adenocarcinoma. *PLoS One*, *20*(9), e0332988.

Xie, Y., Guan, S., Li, Z., Cai, G., Liu, Y., Li, G., Huang, P., & Lin, M. (2025). Identification of a metabolic-immune signature associated with prognosis in colon cancer and exploration of potential predictive efficacy of immunotherapy response. *Clinical and Experimental Medicine*, *25*(1), 46.

Xiong, X., Zheng, L.-W., Ding, Y., Chen, Y.-F., Cai, Y.-W., Wang, L.-P., Huang, L., Liu, C.-C., Shao, Z.-M., & Yu, K.-D. (2025). Breast cancer: pathogenesis and treatments. *Signal Transduction and Targeted Therapy*, *10*(1), 49.

Xu, Q., Fan, G., & Shao, S. (2025). Role of TNFRSF12A in cell proliferation, apoptosis, and proinflammatory cytokine expression by regulating the MAPK and NF-κB pathways in thyroid cancer cells. *Cytokine*, *186*, 156841.

Yan, R., & Chen, T. (2025). SLC35A2 is a novel prognostic biomarker and promotes cell proliferation and metastasis via Wnt/β-catenin/EMT signaling pathway in breast cancer. *Scientific reports*, *15*(1), 130.

Yang, J., Antin, P., Berx, G., Blanpain, C., Brabletz, T., Bronner, M., Campbell, K., Cano, A., Casanova, J., & Christofori, G. (2020). Guidelines and definitions for research on epithelial–mesenchymal transition. *Nature reviews Molecular cell biology*, *21*(6), 341-352.

Yang, X., Huang, R., Fang, M., He, Y., Xie, J., Liu, X., Zhang, C., Lou, Q., Deng, M., & Xiong, W. (2025). Secretogranin 2 binds LILRB4 resulting in immunosuppression. *Nature Immunology*, 1-14.

Yao, W., Rose, J. L., Wang, W., Seth, S., Jiang, H., Taguchi, A., Liu, J., Yan, L., Kapoor, A., & Hou, P. (2019). Syndecan 1 is a critical mediator of macropinocytosis in pancreatic cancer. *Nature*, *568*(7752), 410-414.

Yj, B. (2010). Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastrooesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet*, *376*, 687-697.

Zhang, M., Wu, K., Wang, M., Bai, F., & Chen, H. (2022). CASP9 as a prognostic biomarker and promising drug target plays a pivotal role in inflammatory breast cancer. *International Journal of Analytical Chemistry*, *2022*(1), 1043445.

Zhang, P., Li, Q., Zhang, Y., Wang, Q., Yan, J., Shen, A., & Hu, B. (2023). Identification of a novel gene signature with DDR and EMT difunctionalities for predicting prognosis, immune activity, and drug response in breast cancer. *International Journal of Environmental Research and Public Health*, *20*(2), 1221.

Zhang, P., Wei, Y., Wang, L., Debeb, B. G., Yuan, Y., Zhang, J., Yuan, J., Wang, M., Chen, D., & Sun, Y. (2014). ATM-mediated stabilization of ZEB1 promotes DNA damage response and radioresistance through CHK1. *Nature cell biology*, *16*(9), 864-875.

Zhang, W., Wang, B., Lin, Y., Yang, Y., Zhang, Z., Wang, Q., Zhang, H., Jiang, K., Ye, Y., & Wang, S. (2022). hsa_circ_0000231 Promotes colorectal cancer cell growth through upregulation of CCND2 by IGF2BP3/miR-375 dual pathway. *Cancer cell international*, *22*(1), 27.

Zhang, Y., Sun, Q., Liang, Y., Yang, X., Wang, H., Song, S., Wang, Y., & Feng, Y. (2024). FAM20A: a potential diagnostic biomarker for lung squamous cell carcinoma. *Frontiers in Immunology*, *15*, 1424197.

Zhang, Y., Zhang, X., Huang, X., Tang, X., Zhang, M., Li, Z., Hu, X., Zhang, M., Wang, X., & Yan, Y. (2023). Tumor stemness score to estimate epithelial-to-mesenchymal transition (EMT) and cancer stem cells (CSCs) characterization and to predict the prognosis and immunotherapy response in bladder urothelial carcinoma. *Stem Cell Research & Therapy*, *14*(1), 15.

Zhao, D., Qiao, J., He, H., Song, J., Zhao, S., & Yu, J. (2020). TFPI2 suppresses breast cancer progression through inhibiting TWIST-integrin α5 pathway. *Molecular Medicine*, *26*(1), 27.

Zhao, S., Sun, J., Chang, Q., Pang, S., Zhang, N., Fan, Y., & Liu, J. (2024). CTCF-activated FUCA1 functions as a tumor suppressor by promoting autophagy flux and serum α-L-fucosidase serves as a potential biomarker for prognosis in ccRCC. *Cancer cell international*, *24*(1), 327.

Zhao, W., Steinfeld, J. B., Liang, F., Chen, X., Maranon, D. G., Jian Ma, C., Kwon, Y., Rao, T., Wang, W., & Sheng, C. (2017). BRCA1–BARD1 promotes RAD51-mediated homologous DNA pairing. *Nature*, *550*(7676), 360-365.

Zhu, C., Liu, H., Li, Z., Shi, Y., Zhao, J., Bai, Y., Chen, Q., & Li, W. (2025). Prognostic significance and therapeutic potential of SERPINE1 in head and neck squamous cell carcinoma. *Cancer Medicine*, *14*(2), e70605.

Zhuang, X., Zhang, H., Li, X., Li, X., Cong, M., Peng, F., Yu, J., Zhang, X., Yang, Q., & Hu, G. (2017). Differential effects on lung and bone metastasis of breast cancer by Wnt signalling inhibitor DKK1. *Nature cell biology*, *19*(10), 1274-1285.

ANNEXURES

Annexure I – Publication

A.Hazafa, F. Campana, V. Rosa, F. Iommelli, E. Lonardo, C. M. D. Corte, A. Petrillo, A. Filippelli, V. Conti, V. D. Vecchio, M. L. Noce, F. Papaccio. Identification of Novel Epithelial-Mesenchymal Transition (EMT) and DNA Damage Repair (DDR)-related Prognostic Gene Signatures as Predictive Biomarkers in Six Cancer Cohorts (Submitted to Multi-Journal Submission Cell Press).

Annexure II – Appendices

Appendix A — Supplementary Files 1

Table S1. Gene symbols and Entrez ID of EMT and DDR gene sets, downloaded from MSigDB.

	EMT		DDR	
Sr. No.	**Gene symbol**	**Gene EZIDs**	**Gene symbol**	**Gene EZIDs**
1.	*ABI3BP*	25890	*ABL1*	25
2.	*ACTA2*	59	*AKT1*	207
3.	*ADAM12*	8038	*APAF1*	317
4.	*ANPEP*	290	*ATM*	472
5.	*APLP1*	333	*ATR*	545
6.	*AREG*	374	*ATRIP*	84126
7.	*BASP1*	10409	*BAX*	581
8.	*BDNF*	627	*BBC3*	27113
9.	*BGN*	633	*BID*	637
10.	*BMP1*	649	*BRCA1*	672
11.	*CADM1*	23705	*CASP3*	836
12.	*CALD1*	800	*CASP8*	841
13.	*CALU*	813	*CASP9*	842
14.	*CAP2*	10486	*CCNB1*	891
15.	*CAPG*	822	*CCNB2*	9133
16.	*CCN1*	3491	*CCNB3*	85417
17.	*CCN2*	1490	*CCND1*	595
18.	*CD44*	960	*CCND2*	894
19.	*CD59*	966	*CCND3*	896
20.	*CDH11*	1009	*CCNE1*	898
21.	*CDH2*	1000	*CCNE2*	9134
22.	*CDH6*	1004	*CDC25A*	993
23.	*COL11A1*	1301	*CDC25C*	995
24.	*COL12A1*	1303	*CDK1*	983
25.	*COL16A1*	1307	*CDK2*	1017
26.	*COL1A1*	1277	*CDK4*	1019
27.	*COL1A2*	1278	*CDK5*	1020
28.	*COL3A1*	1281	*CDK6*	1021
29.	*COL4A1*	1282	*CDKN1A*	1026
30.	*COL4A2*	1284	*CDKN1B*	1027
31.	*COL5A1*	1289	*CHEK1*	1111
32.	*COL5A2*	1290	*CHEK2*	11200
33.	*COL5A3*	50509	*CREB1*	1385
34.	*COL6A2*	1292	*CYCS*	54205
35.	*COL6A3*	1293	*DDB2*	1643
36.	*COL7A1*	1294	*E2F1*	1869
37.	*COL8A2*	1296	*FANCD2*	2177
38.	*COLGALT1*	79709	*FAS*	355
39.	*COMP*	1311	*GADD45A*	1647
40.	*COPA*	1314	*GADD45B*	4616
41.	*CRLF1*	9244	*GADD45G*	10912
42.	*CTHRC1*	115908	*H2AX*	3014
43.	*CXCL1*	2919	*HUS1B*	135458
44.	*CXCL12*	6387	*MDM2*	4193
45.	*CXCL6*	6372	*MRE11*	4361
46.	*CXCL8*	3576	*MYC*	4609
47.	*DAB2*	1601	*NBN*	4683
48.	*DCN*	1634	*PIDD1*	55367
49.	*DKK1*	22943	*PMAIP1*	5366
50.	*DPYSL3*	1809	*PML*	5371
51.	*DST*	667	*PRKDC*	5591
52.	*ECM1*	1893	*RAD1*	5810
53.	*ECM2*	1842	*RAD17*	5884

54.	*EDIL3*	10085	*RAD50*	10111
55.	*EFEMP2*	30008	*RAD51*	5888
56.	*ELN*	2006	*RAD52*	5893
57.	*EMP3*	2014	*RAD9A*	5883
58.	*ENO2*	2026	*RB1*	5925
59.	*FAP*	2191	*RFC1*	5981
60.	*FAS*	355	*RPA2*	6118
61.	*FBLN1*	2192	*RRM2B*	50484
62.	*FBLN2*	2199	*SESN1*	27244
63.	*FBLN5*	10516	*SFN*	2810
64.	*FBN1*	2200	*SMC1A*	8243
65.	*FBN2*	2201	*TLK1*	9874
66.	*FERMT2*	10979	*TLK2*	11011
67.	*FGF2*	2247	*TNFRSF10B*	8795
68.	*FLNA*	2316	*TP53*	7157
69.	*FMOD*	2331	*TP53AIP1*	63970
70.	*FN1*	2335		
71.	*FOXC2*	2303		
72.	*FSTL1*	11167		
73.	*FSTL3*	10272		
74.	*FUCA1*	2517		
75.	*FZD8*	8325		
76.	*GADD45A*	1647		
77.	*GADD45B*	4616		
78.	*GAS1*	2619		
79.	*GEM*	2669		
80.	*GJA1*	2697		
81.	*GLIPR1*	11010		
82.	*GPC1*	2817		
83.	*GPX7*	2882		
84.	*GREM1*	26585		
85.	*HTRA1*	5654		
86.	*ID2*	3398		
87.	*IGFBP2*	3485		
88.	*IGFBP3*	3486		
89.	*IGFBP4*	3487		
90.	*IL15*	3600		
91.	*IL32*	9235		
92.	*IL6*	3569		
93.	*INHBA*	3624		
94.	*ITGA2*	3673		
95.	*ITGA5*	3678		
96.	*ITGAV*	3685		
97.	*ITGB1*	3688		
98.	*ITGB3*	3690		
99.	*ITGB5*	3693		
100.	*JUN*	3725		
101.	*LAMA1*	284217		
102.	*LAMA2*	3908		
103.	*LAMA3*	3909		
104.	*LAMC1*	3915		
105.	*LAMC2*	3918		
106.	*LGALS1*	3956		
107.	*LOX*	4015		
108.	*LOXL1*	4016		
109.	*LOXL2*	4017		
110.	*LRP1*	4035		
111.	*LRRC15*	131578		
112.	*LUM*	4060		
113.	*MAGEE1*	57692		
114.	*MATN2*	4147		
115.	*MATN3*	4148		
116.	*MCM7*	4176		
117.	*MEST*	4232		

118.	*MFAP5*	8076
119.	*MGP*	4256
120.	*MMP1*	4312
121.	*MMP14*	4323
122.	*MMP2*	4313
123.	*MMP3*	4314
124.	*MSX1*	4487
125.	*MXRA5*	25878
126.	*MYL9*	10398
127.	*MYLK*	4638
128.	*NID2*	22795
129.	*NNMT*	4837
130.	*NOTCH2*	4853
131.	*NT5E*	4907
132.	*NTM*	50863
133.	*OXTR*	5021
134.	*P3H1*	64175
135.	*PCOLCE*	5118
136.	*PCOLCE2*	26577
137.	*PDGFRB*	5159
138.	*PDLIM4*	8572
139.	*PFN2*	5217
140.	*PLAUR*	5329
141.	*PLOD1*	5351
142.	*PLOD2*	5352
143.	*PLOD3*	8985
144.	*PMEPA1*	56937
145.	*PMP22*	5376
146.	*POSTN*	10631
147.	*PPIB*	5479
148.	*PRRX1*	5396
149.	*PRSS2*	5645
150.	*PTHLH*	5744
151.	*PTX3*	5806
152.	*PVR*	5817
153.	*QSOX1*	5768
154.	*RGS4*	5999
155.	*RHOB*	388
156.	*SAT1*	6303
157.	*SCG2*	7857
158.	*SDC1*	6382
159.	*SDC4*	6385
160.	*SERPINE1*	5054
161.	*SERPINE2*	5270
162.	*SERPINH1*	871
163.	*SFRP1*	6422
164.	*SFRP4*	6424
165.	*SGCB*	6443
166.	*SGCD*	6444
167.	*SGCG*	6445
168.	*SLC6A8*	6535
169.	*SLIT2*	9353
170.	*SLIT3*	6586
171.	*SNAI2*	6591
172.	*SNTB1*	6641
173.	*SPARC*	6678
174.	*SPOCK1*	6695
175.	*SPP1*	6696
176.	*TAGLN*	6876
177.	*TFPI2*	7980
178.	*TGFB1*	7040
179.	*TGFBI*	7045
180.	*TGFBR3*	7049
181.	*TGM2*	7052

182.	*THBS1*	7057
183.	*THBS2*	7058
184.	*THY1*	7070
185.	*TIMP1*	7076
186.	*TIMP3*	7078
187.	*TNC*	3371
188.	*TNFAIP3*	7128
189.	*TNFRSF11B*	4982
190.	*TNFRSF12A*	51330
191.	*TPM1*	7168
192.	*TPM2*	7169
193.	*TPM4*	7171
194.	*VCAM1*	7412
195.	*VCAN*	1462
196.	*VEGFA*	7422
197.	*VEGFC*	7424
198.	*VIM*	7431
199.	*WIPF1*	7456
200.	*WNT5A*	7474

Table S2. List of packages used in the study.

Package	Key function
limma	To pull out DEGs
edgeR	To normalize the counts
survival	For Kaplan-Meier and Cox analysis
survminer	Kaplan-Meier curves with risk table and p value
glmnet	To perform LASSO regression analysis
ggplot2	To build plots
grid	Fine-tune annotations
patchwork	To combine plots
ggpubr	To add automatic significance
VennDiagram	For Ven diagrams
pheatmap	For heatmaps
Rtsne	To perform t-distributed stochastic neighbor embedding (t-SNE)
RColorBrewer	For palettes
igraph	To build correlation graphs
timeROC	Time-dependent ROC
ggsignif	Bar plot or jitter plots
ggdist	To build raincloud + boxplots + jitter overlay
gghalves	To build raincloud plots
rms	To build nomogram plots
rmda	DCA for net benefits vs. threshold
clusterProfiler	For enrichment analysis
org.Hs.eg.db	For gene mapping
enrichplot	Latest enrichment plots
DOSE	To gene ontology enrichment dot plots
GSVA	To calculate ssGSEA score
estimate	To calculate ESTIMATE metrics score

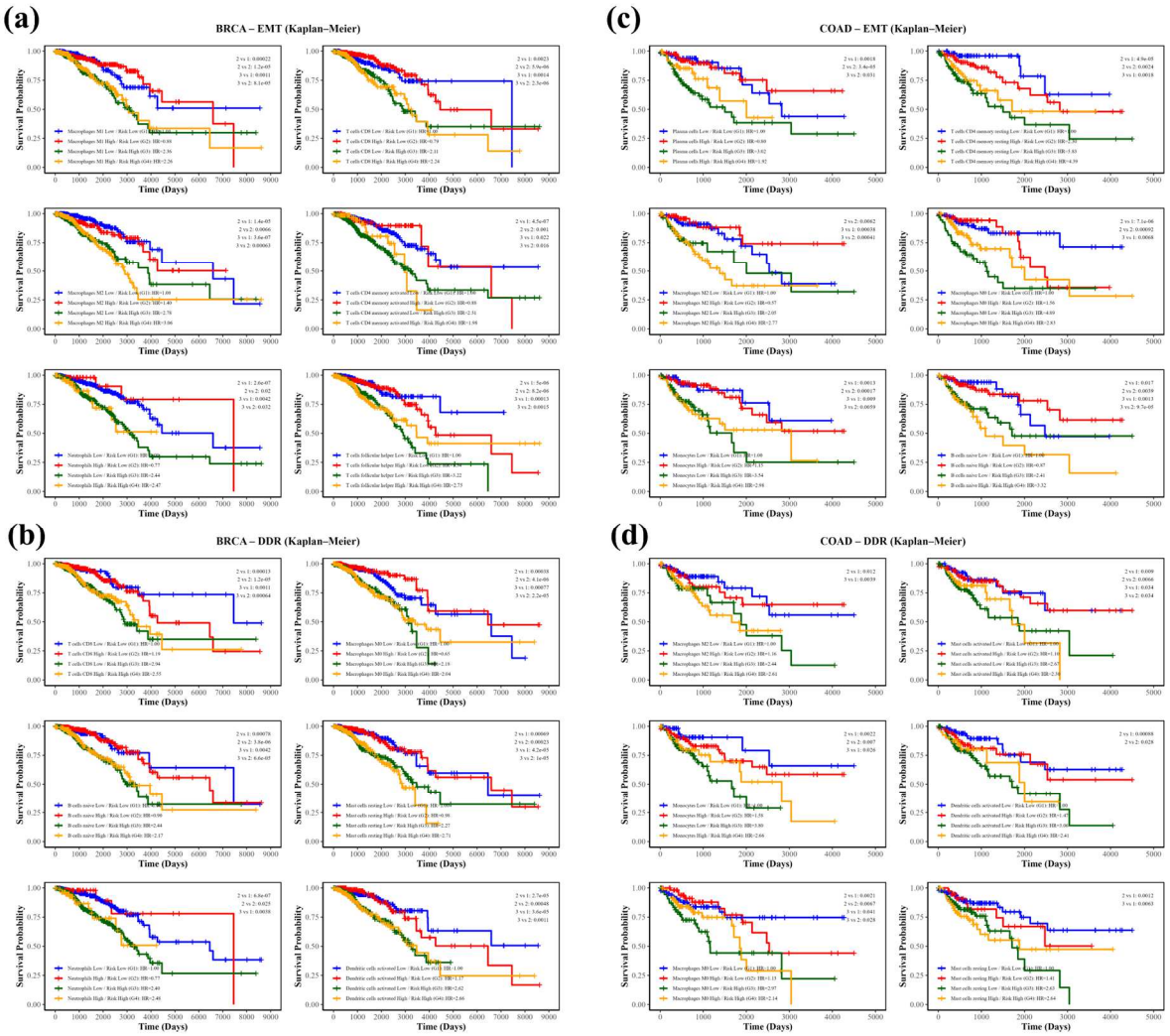


Figure S1. Combined Kaplan–Meier analysis of top six kinds of significant immune cells and risk scores for EMT- and DDR-related gene signatures in TCGA-BRCA (a, b) and TCGA-COAD (c, d).

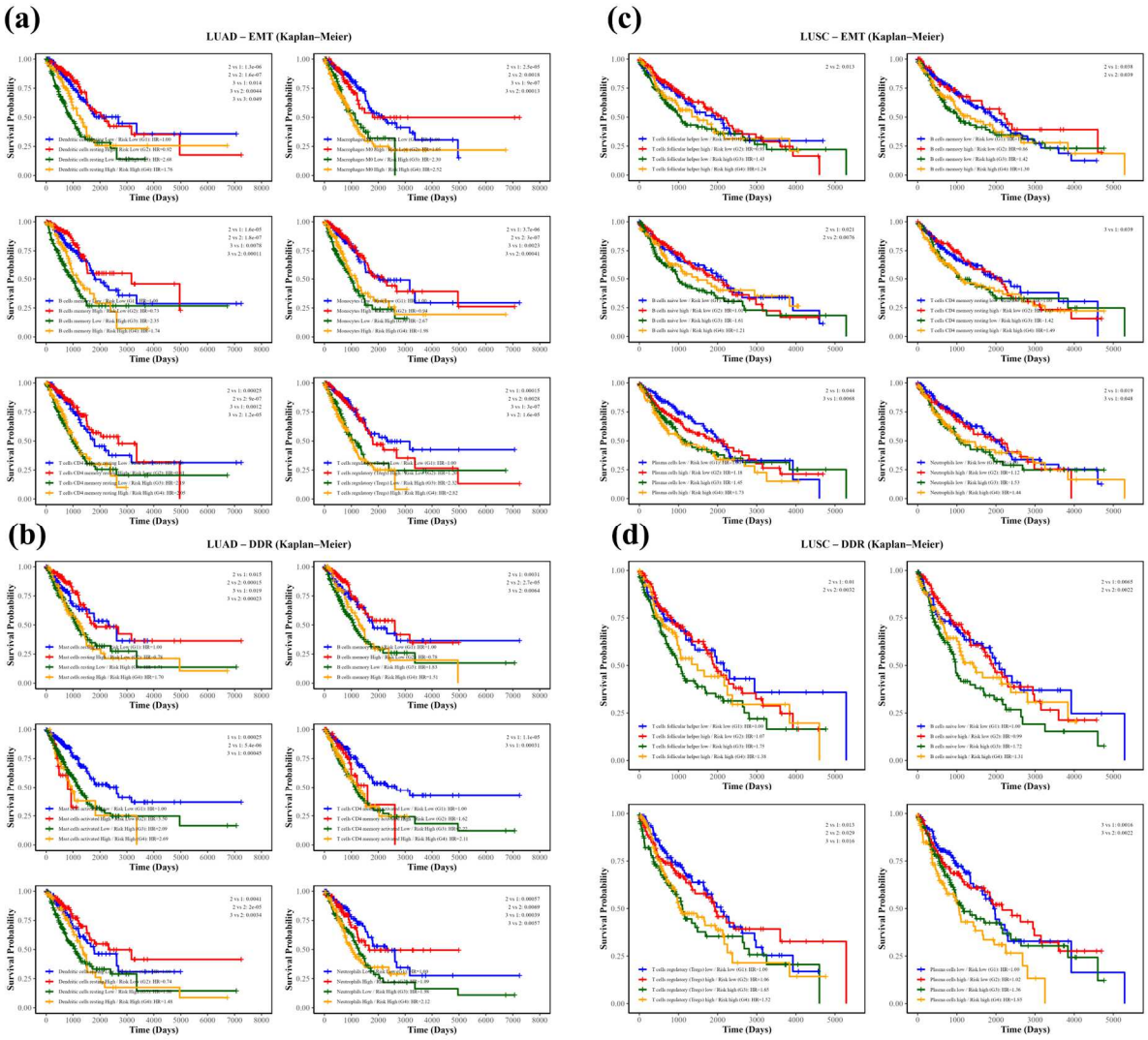


Figure S2. Combined Kaplan–Meier analysis of top six kinds of significant immune cells and risk scores for EMT- and DDR-related gene signatures in TCGA-LUAD (a, b) and TCGA-LUSC (c, d).

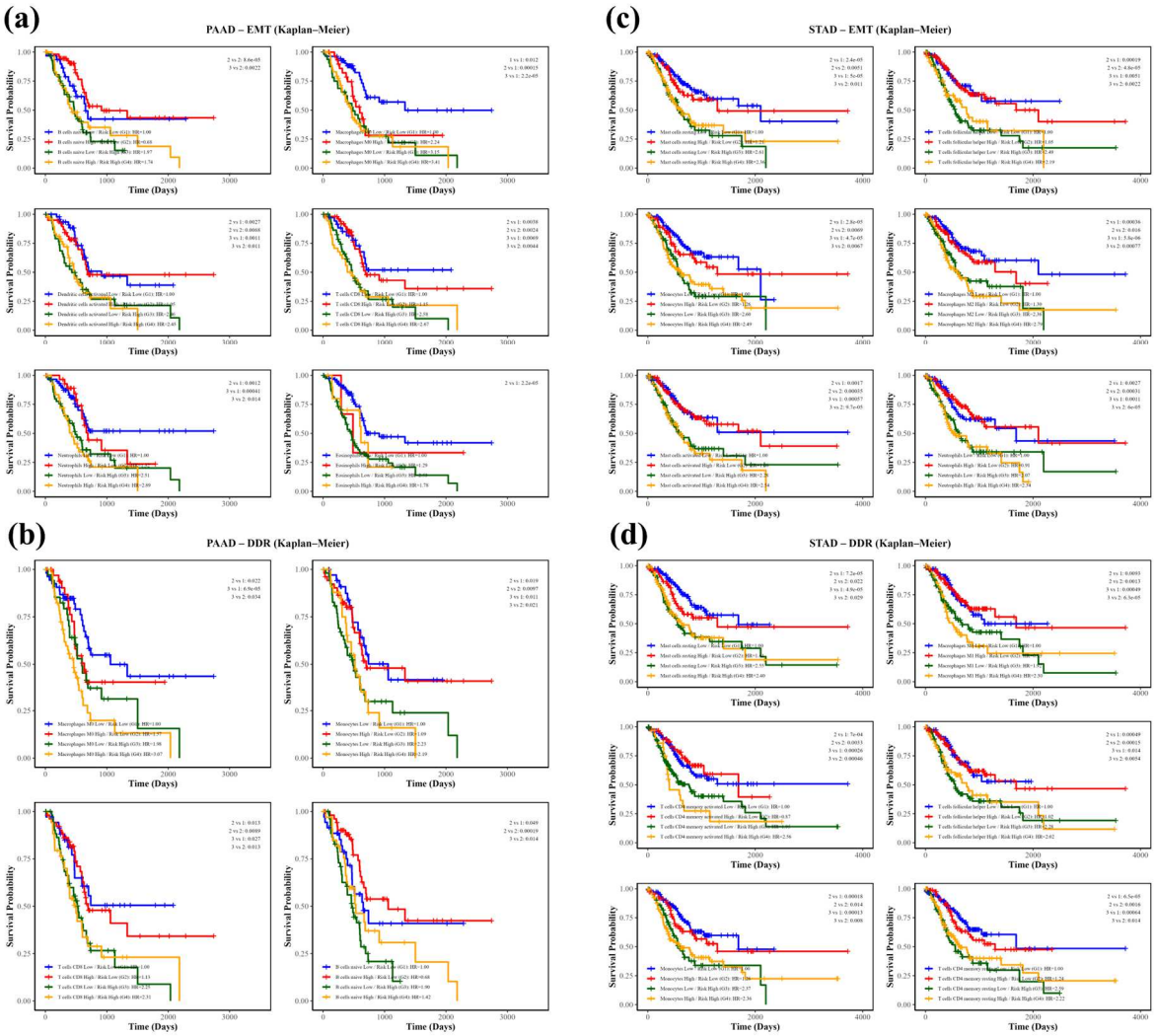


Figure S3. Combined Kaplan–Meier analysis of top six kinds of significant immune cells and risk scores for EMT- and DDR-related gene signatures in TCGA-PAAD (a, b) and TCGA-STAD (c, d).

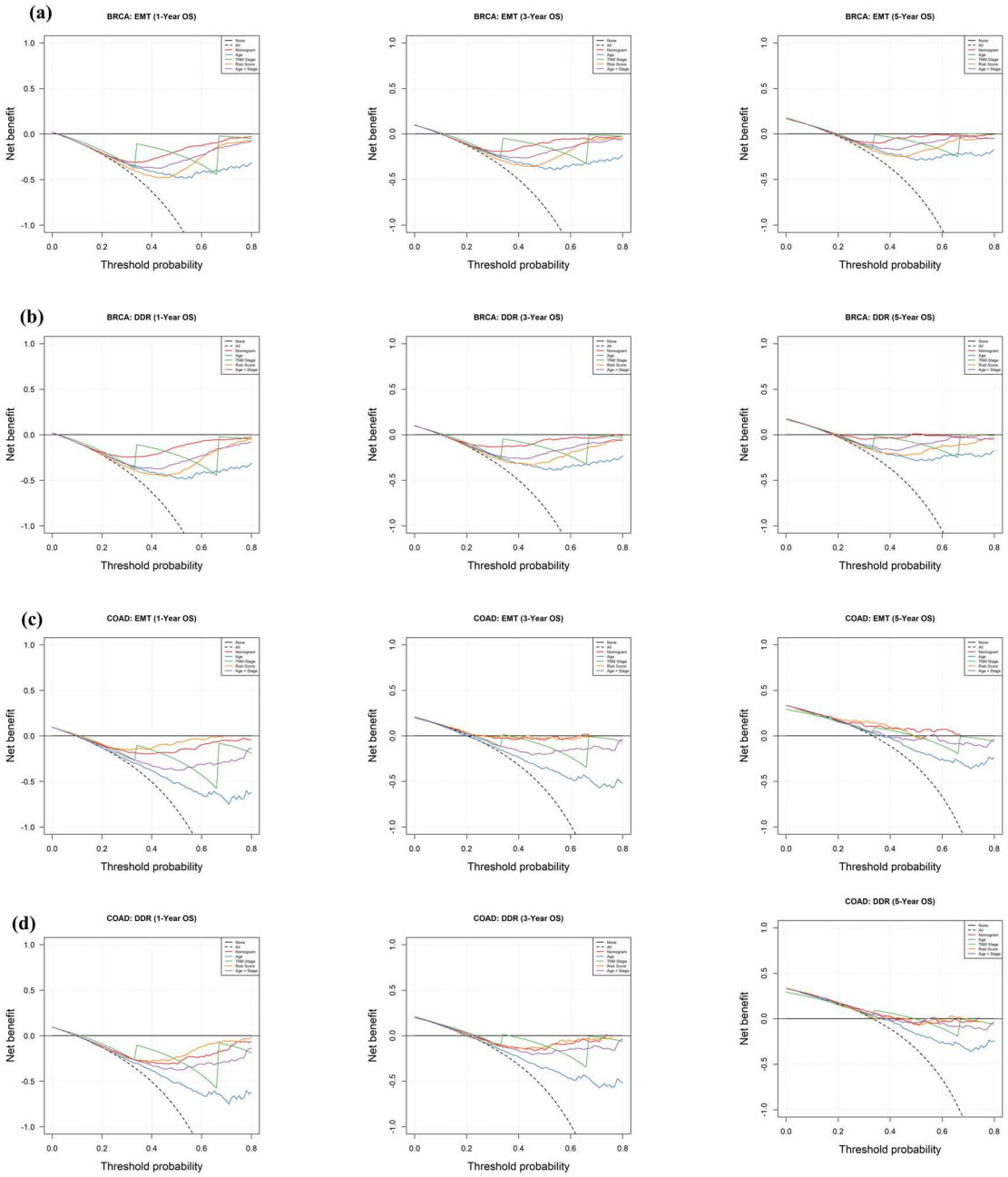


Figure S4. DCA curve analysis for different clinical parameters such as age, TNM stage, risk score, age + TNM stage, and nomograms at OS prediction in BRCA (a for EMT-related signature; b for DDR-related signature) and COAD cohorts (c for EMT-related signature; d for DDR-related signature).

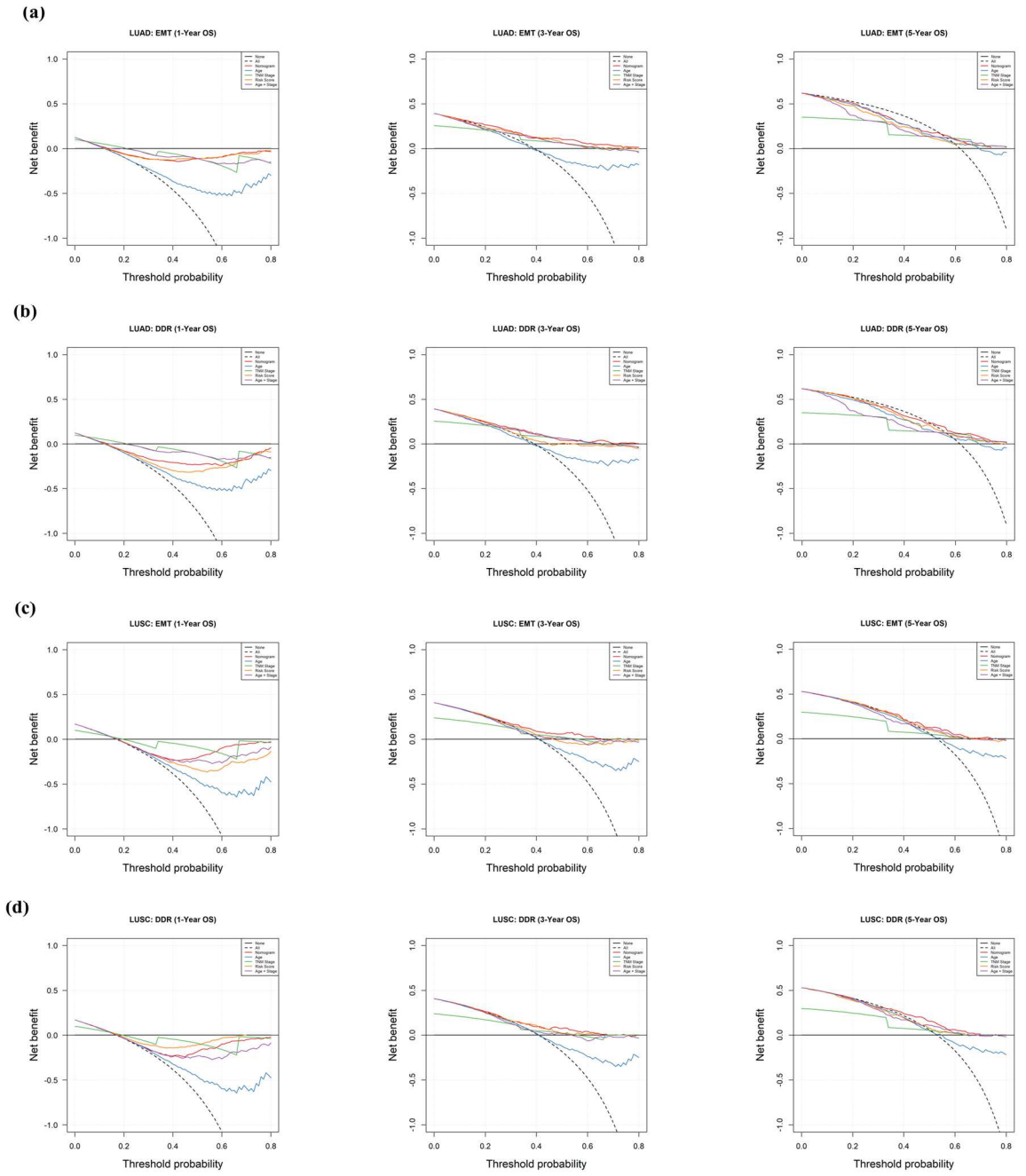


Figure S5. DCA curve analysis for different clinical parameters such as age, TNM stage, risk score, age + TNM stage, and nomograms at OS prediction in LUAD (a for EMT-related signature; b for DDR-related signature) and LUSC cohorts (c for EMT-related signature; d for DDR-related signature).

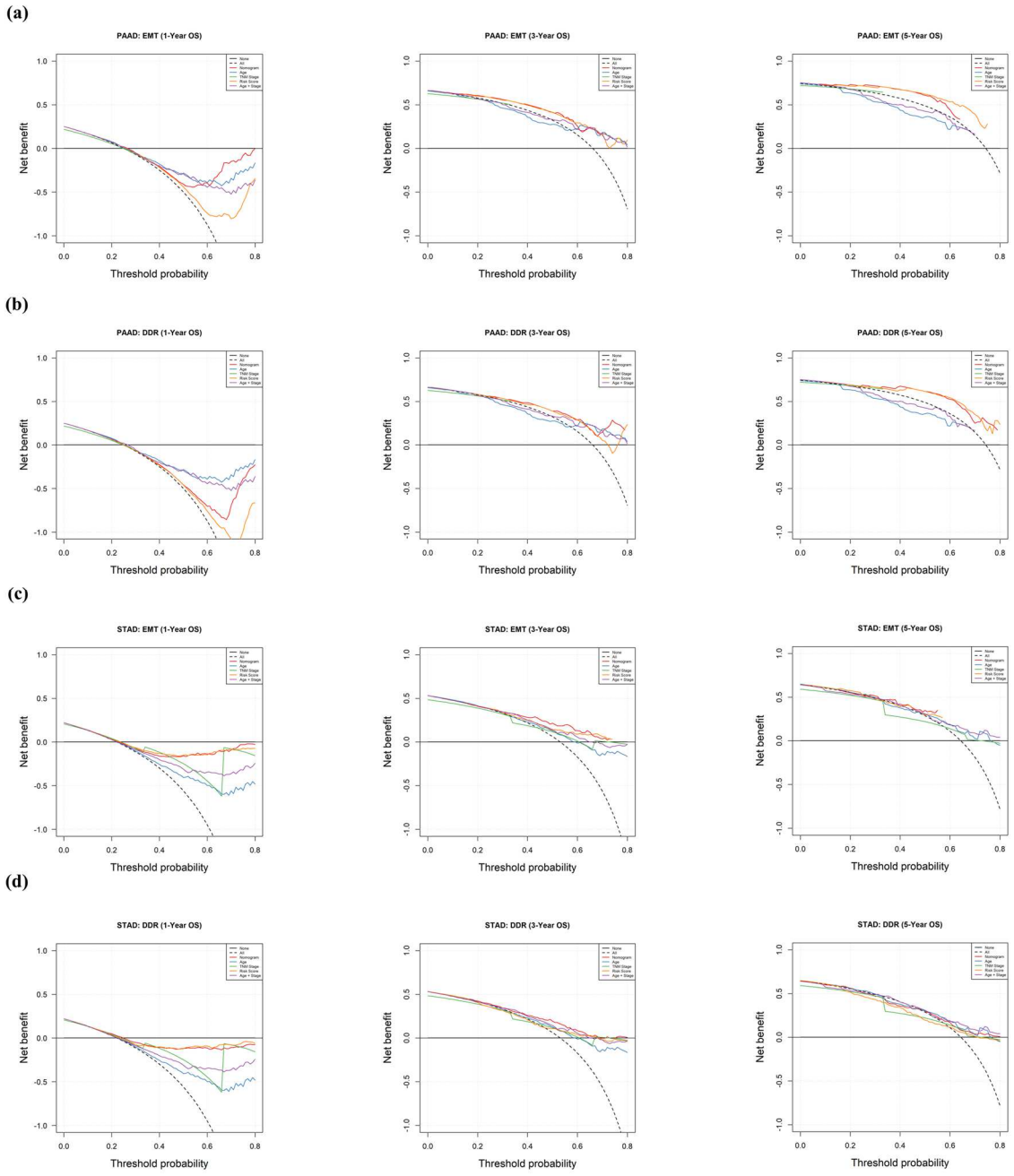


Figure S6. DCA curve analysis for different clinical parameters such as age, TNM stage, risk score, age + TNM stage, and nomograms at OS prediction in PAAD (a for EMT-related signature; b for DDR-related signature) and STAD cohorts (c for EMT-related signature; d for DDR-related signature).

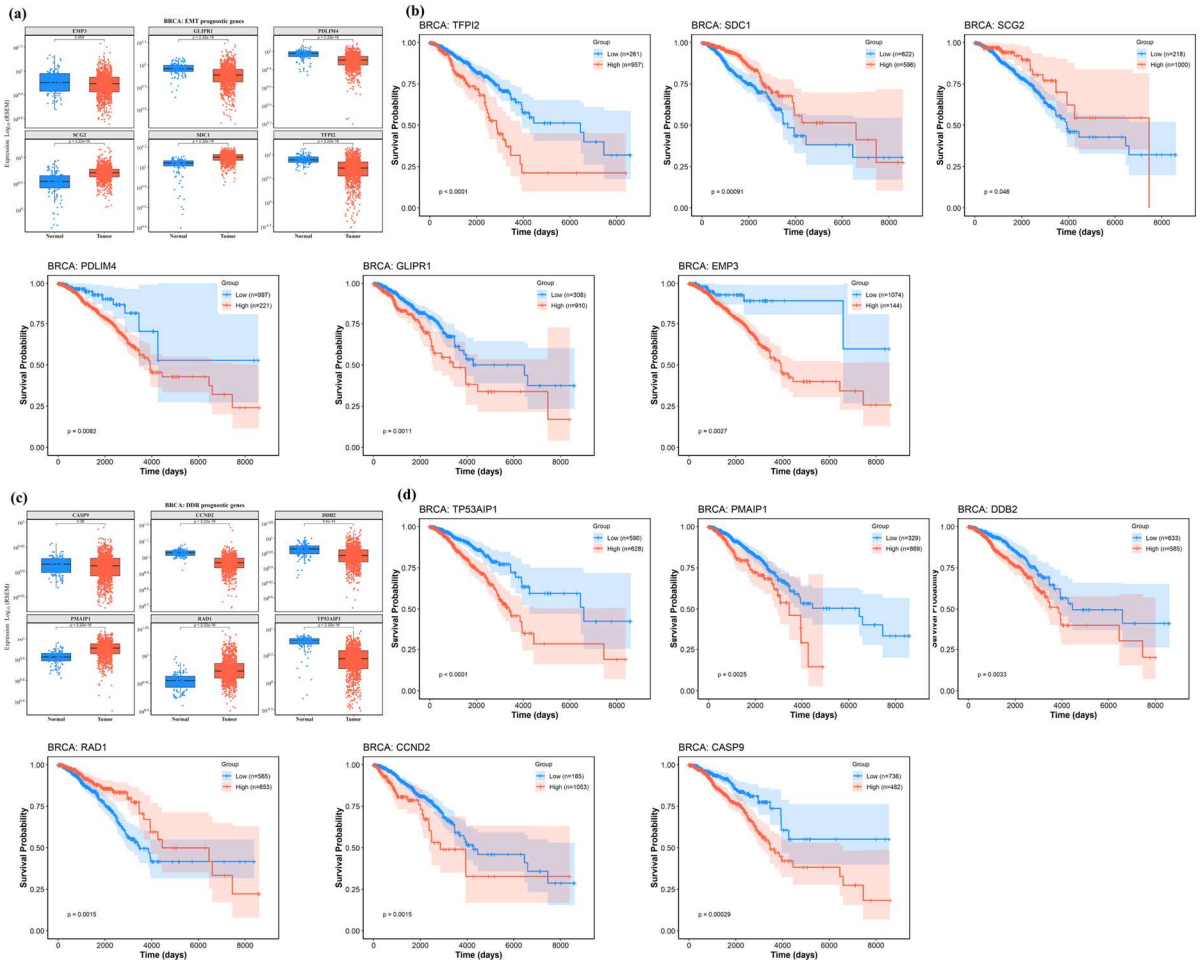


Figure S7. Expression profiles and survival analysis of identified prognostic EMT (a, b) and DDR (c, d) related signatures in BRCA cohort.

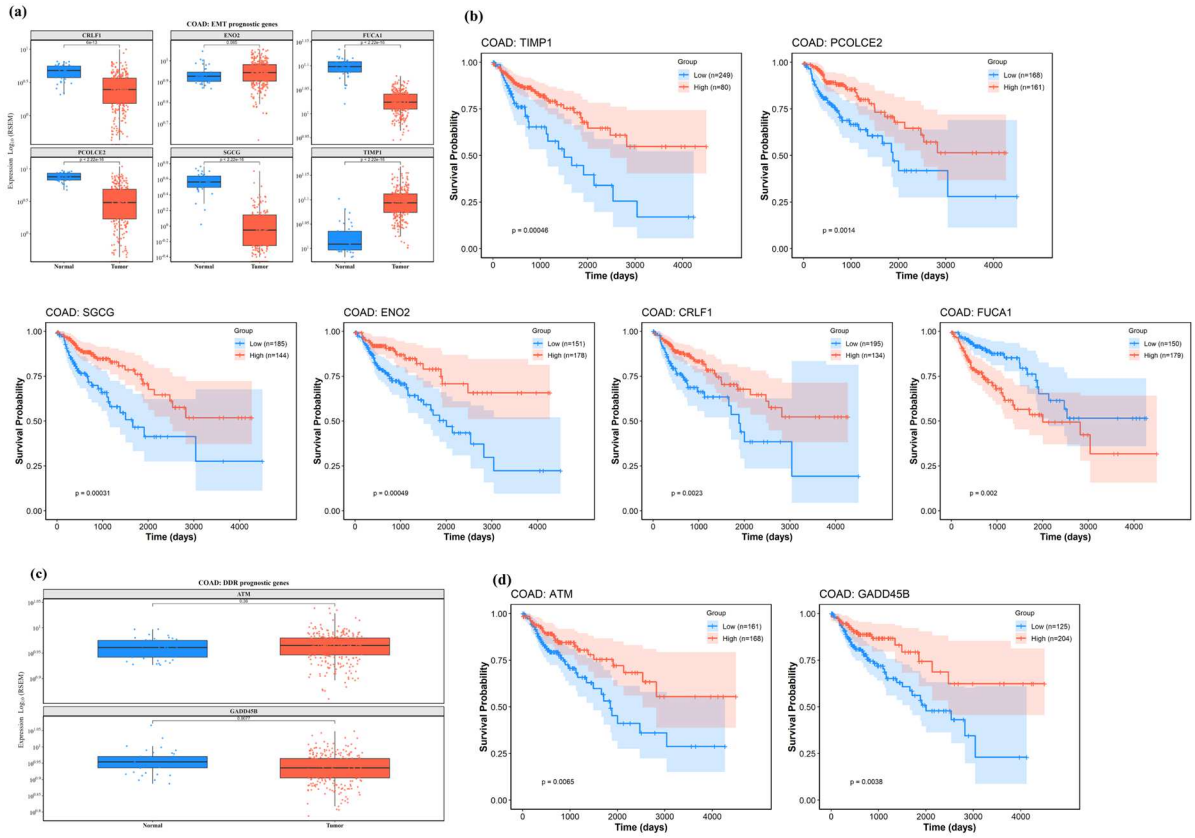


Figure S8. Expression profiles and survival analysis of identified prognostic EMT (a, b) and DDR (c, d) related signatures in TCGA-COAD cohort.

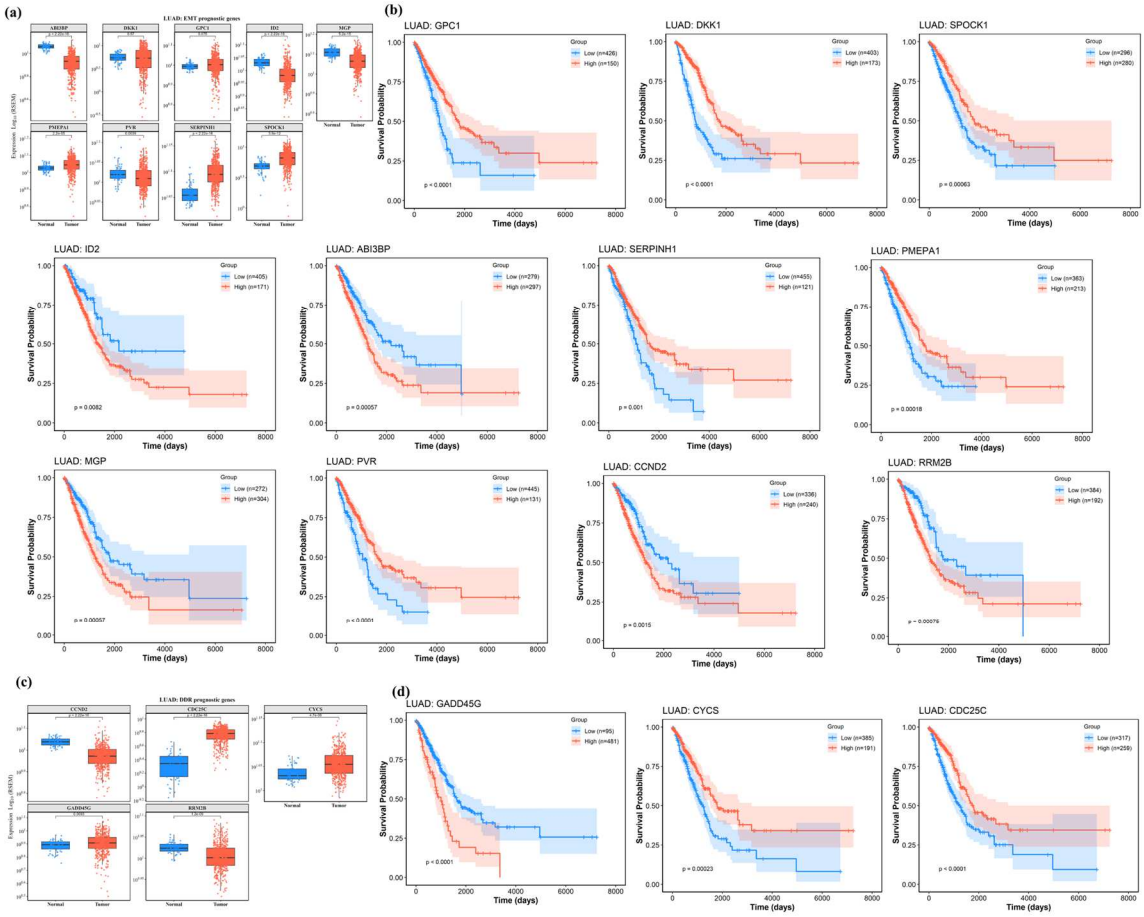


Figure S9. Expression profiles and survival analysis of identified prognostic EMT (a, b) and DDR (c, d) related signatures in LUAD cohort.

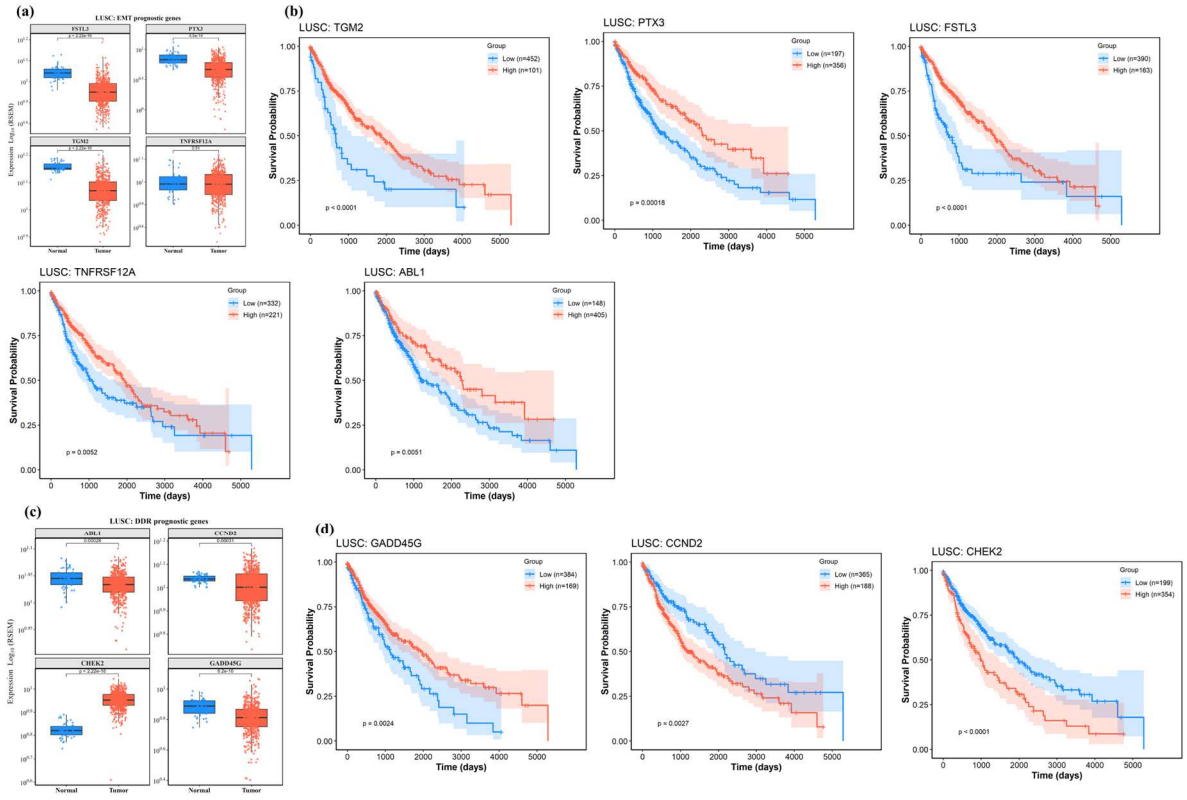


Figure S10. Expression profiles and survival analysis of identified prognostic EMT (a, b) and DDR (c, d) related signatures in LUSC cohort.

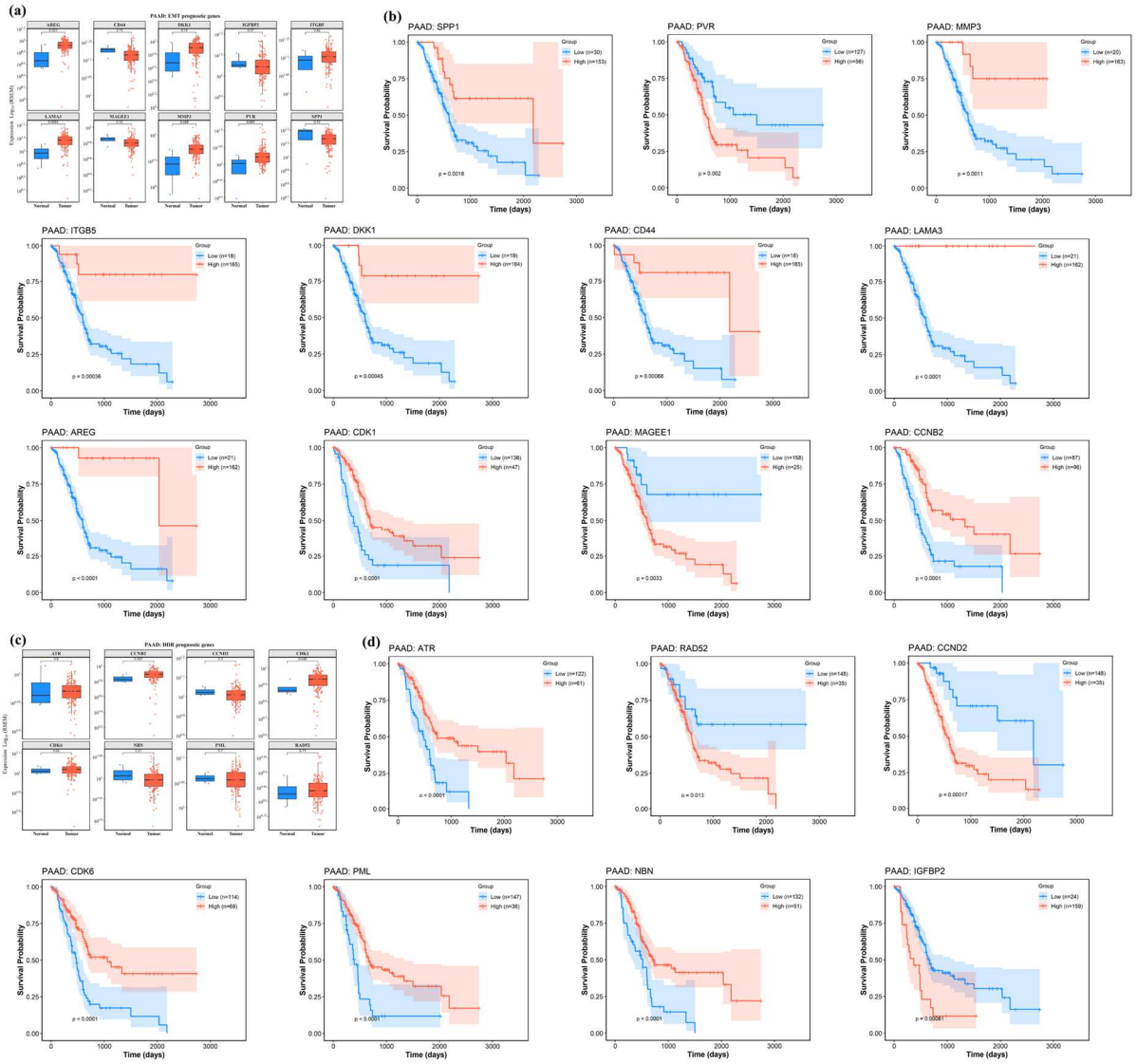


Figure S11. Expression profiles and survival analysis of identified prognostic EMT (a, b) and DDR (c, d) related signatures in PAAD cohort.

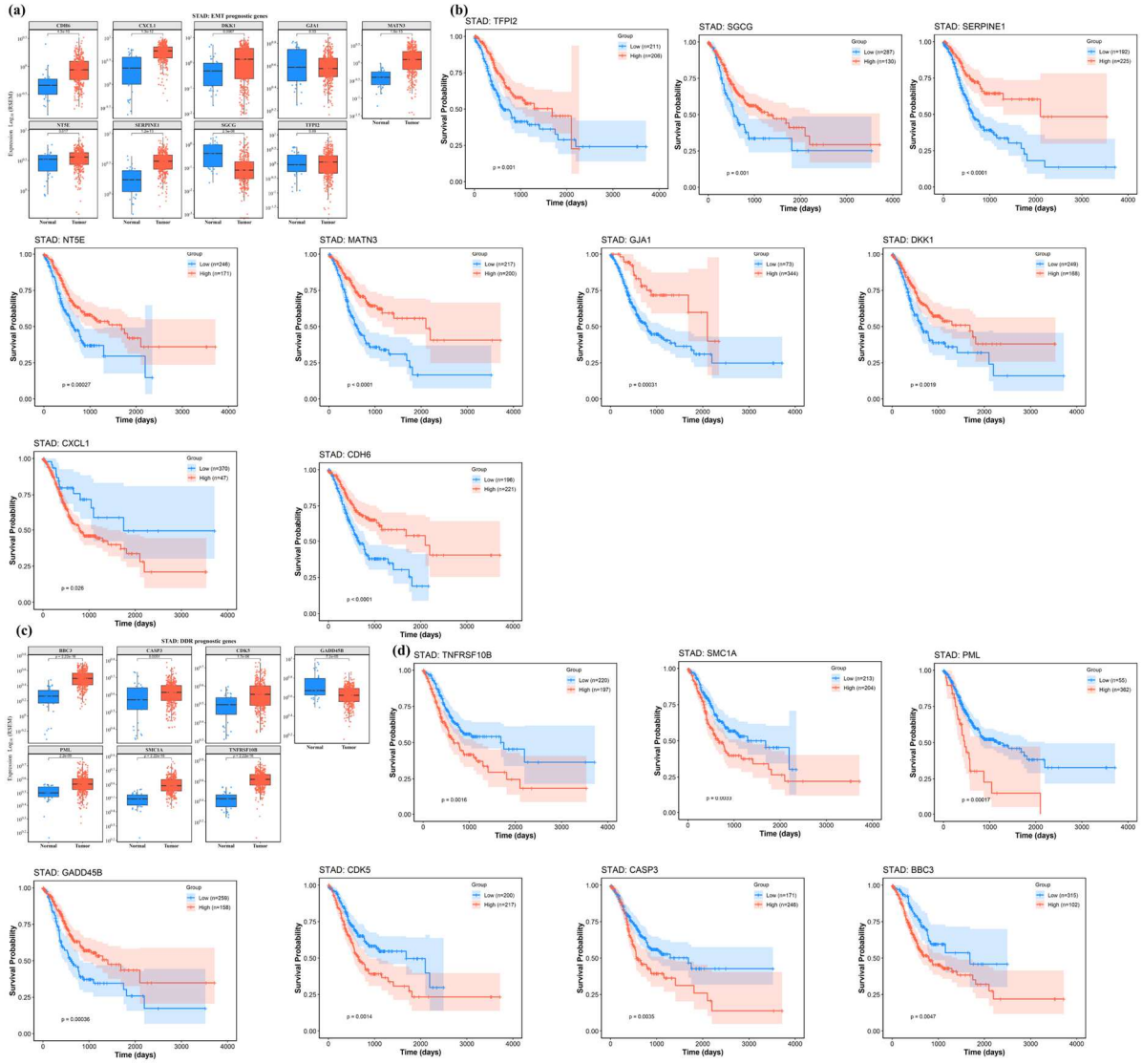


Figure S12. Expression profiles and survival analysis of identified prognostic EMT (a, b) and DDR (c, d) related signatures in STAD cohort.

Appendix B — Supplementary Files 2

Table S1. The Akaike Information Criterion (AIC) values individually and in combination in the training cohort of TCGA-BRCA for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.94 (0.58-1.54)	0.81	0.73 (0.45-1.17)	0.187	0.95 (0.60-1.50)	0.82	–	–	0.73 (0.47-1.15)	0.177	–	–	–	–
Age 50	0.69 (0.41-1.18)	0.179	0.68 (0.40-1.15)	0.149	0.69 (0.42-1.12)	0.133	–	–	0.70 (0.43-1.14)	0.147	–	–	–	–
Age 70	1.82 (1.10-3.01)	0.0194	1.77 (1.07-2.90)	0.0252	1.75 (1.08-2.85)	0.0235	–	–	1.68 (1.03-2.72)	0.036	–	–	–	–
Age 80+	3.84 (2.09-7.07)	0.0000149	3.62 (1.99-6.60)	0.0000259	3.81 (2.09-6.92)	0.0000117	–	–	3.57 (1.97-6.46)	0.0000257	–	–	–	–
TNM Stage II	1.56 (0.90-2.72)	0.114	1.79 (1.03-3.11)	0.0377	–	–	1.38 (0.79-2.38)	0.256	–	–	1.55 (0.90-2.68)	0.115	–	–
TNM Stage III	3.25 (1.82-5.81)	0.0000679	3.35 (1.88-5.96)	0.0000384	–	–	3.02 (1.70-5.37)	0.000163	–	–	3.02 (1.71-5.36)	0.000151	–	–
TNM Stage IV	8.16 (3.94-16.91)	1.59E-08	12.01 (5.87-24.57)	9.7E-12	–	–	8.24 (4.00-16.95)	1.01E-08	–	–	12.87 (6.34-26.13)	1.55E-12	–	–
Risk score	2.97 (2.06-4.29)	5.5E-09	–	–	2.91 (2.10-4.05)	1.75E-10	3.09 (2.15-4.43)	9.11E-10	–	–	–	–	2.88 (2.10-3.95)	5.71E-11
AIC	**1516.466668**	–	1547.389178	–	1649.129434	–	1537.661904	–	1685.400232	–	1572.128334	–	1671.268388	–

DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR | Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.79 (0.49-1.28)	0.343	0.73 (0.45-1.17)	0.187	0.80 (0.51-1.25)	0.325	–	–	0.73 (0.47-1.15)	0.177	–	–	–	–
Age 50	0.65 (0.38-1.10)	0.105	0.68 (0.40-1.15)	0.149	0.70 (0.43-1.15)	0.162	–	–	0.70 (0.43-1.14)	0.147	–	–	–	–
Age 70	1.79 (1.09-2.96)	0.0224	1.77 (1.07-2.90)	0.0252	1.78 (1.10-2.89)	0.0195	–	–	1.68 (1.03-2.72)	0.036	–	–	–	–
Age 80+	3.26 (1.78-5.97)	0.000126	3.62 (1.99-6.60)	0.0000259	3.42 (1.89-6.19)	0.0000494	–	–	3.57 (1.97-6.46)	0.0000257	–	–	–	–

TNM Stage II	1.57 (0.90-2.73)	0.113	1.79 (1.03-3.11)	0.0377	–	–	1.35 (0.78-2.33)	0.29	–	–	1.55 (0.90-2.68)	0.115	–	–
TNM Stage III	3.15 (1.77-5.62)	0.0000982	3.35 (1.88-5.96)	0.0000384	–	–	2.87 (1.61-5.09)	0.000324	–	–	3.02 (1.71-5.36)	0.000151	–	–
TNM Stage IV	12.45 (6.07-25.52)	5.91E-12	12.01 (5.87-24.57)	9.7E-12	–	–	12.92 (6.35-26.27)	1.59E-12	–	–	12.87 (6.34-26.13)	1.55E-12	–	–
Risk score	3.05 (2.09-4.44)	6.06E-09	–	–	2.80 (1.96-4.00)	1.49E-08	3.22 (2.20-4.69)	1.42E-09	–	–	–	–	2.91 (2.03-4.17)	5.71E-09
AIC	**1516.055544**	–	1547.389178	–	1655.884463	–	1537.885672	–	1685.400232	–	1572.128334	–	1678.251449	–

Table S2. The Akaike Information Criterion (AIC) values individually and in combination in the training cohort of TCGA-COAD for EMT- and DDR-related prognostic signatures.

EMT-related signature

Clinical parameter	EMT | Nomogram COX PH		EMT | Age + TNM stage COX PH		EMT | Age + Risk score COX PH		EMT | TNM stage + Risk score COX PH		EMT | Age only COX PH		EMT | TNM stage only COX PH		EMT | Risk score only COX PH	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.06 (0.38-2.97)	0.911	1.29 (0.46-3.59)	0.628	1.23 (0.48-3.19)	0.667	–	–	1.30 (0.50-3.35)	0.592	–	–	–	–
Age 50	0.90 (0.34-2.37)	0.83	0.84 (0.32-2.19)	0.718	1.45 (0.63-3.36)	0.383	–	–	1.38 (0.60-3.19)	0.449	–	–	–	–
Age 70	2.35 (1.09-5.04)	0.0284	2.42 (1.13-5.16)	0.0223	2.51 (1.23-5.11)	0.0115	–	–	2.22 (1.09-4.51)	0.0282	–	–	–	–
Age 80+	3.06 (1.32-7.14)	0.00941	3.03 (1.30-7.06)	0.0104	2.95 (1.31-6.61)	0.00873	–	–	2.67 (1.19-5.97)	0.0168	–	–	–	–
TNM Stage II	2.04 (0.60-6.93)	0.252	2.16 (0.64-7.36)	0.216	–	–	2.05 (0.61-6.93)	0.248	–	–	2.27 (0.67-7.66)	0.187	–	–
TNM Stage III	3.04 (0.88-10.46)	0.0781	3.82 (1.12-13.05)	0.0328	–	–	2.58 (0.76-8.79)	0.129	–	–	3.36 (1.00-11.35)	0.0509	–	–
TNM Stage IV	11.51 (3.24-40.87)	1.58E-04	15.16 (4.30-53.46)	2.34E-05	–	–	9.60 (2.73-33.71)	4.20E-04	–	–	11.74 (3.36-41.03)	1.15E-04	–	–
Risk score	3.39 (1.77-6.49)	2.26E-04	–	–	4.31 (2.41-7.73)	8.83E-07	3.28 (1.75-6.12)	1.96E-04	–	–	–	–	3.88 (2.19-6.88)	3.38E-06
AIC	**509.2597338**	–	520.8636146	–	624.8939876	–	514.445774	–	646.4465728	–	526.0530766	–	628.1507181	–

DDR-related signature

Clinical parameter	DDR | Nomogram COX PH		DDR | Age + TNM stage COX PH		DDR | Age + Risk score COX PH		DDR | TNM stage + Risk score COX PH		DDR | Age only COX PH		DDR | TNM stage only COX PH		DDR | Risk score only COX PH	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.24 (0.45-3.46)	0.675	1.29 (0.46-3.59)	0.628	1.25 (0.48-3.23)	0.645	–	–	1.30 (0.50-3.35)	0.592	–	–	–	–
Age 50	0.75 (0.28-1.99)	0.565	0.84 (0.32-2.19)	0.718	1.22 (0.52-2.84)	0.642	–	–	1.38 (0.60-3.19)	0.449	–	–	–	–
Age 70	2.45 (1.15-5.23)	0.0209	2.42 (1.13-5.16)	0.0223	2.52 (1.23-5.14)	0.0112	–	–	2.22 (1.09-4.51)	0.0282	–	–	–	–
Age 80+	2.98 (1.28-6.95)	0.0115	3.03 (1.30-7.06)	0.0104	2.63 (1.17-5.87)	0.0188	–	–	2.67 (1.19-5.97)	0.0168	–	–	–	–

TNM Stage II	1.97 (0.58-6.75)	0.28	2.16 (0.64-7.36)	0.216	–	–	2.10 (0.62-7.15)	0.235	–	–	2.27 (0.67-7.66)	0.187	–	–
TNM Stage III	3.15 (0.90-10.99)	0.0718	3.82 (1.12-13.05)	0.0328	–	–	2.84 (0.82-9.83)	0.1	–	–	3.36 (1.00-11.35)	0.0509	–	–
TNM Stage IV	11.54 (3.14-42.35)	2.27E-04	15.16 (4.30-53.46)	2.34E-05	–	–	9.24 (2.52-33.90)	7.94E-04	–	–	11.74 (3.36-41.03)	1.15E-04	–	–
Risk score	2.01 (0.90-4.47)	0.0878	–	–	3.32 (1.67-6.59)	6.12E-04	1.67 (0.78-3.57)	0.185	–	–	–	–	2.80 (1.46-5.35)	0.00184
AIC	**519.8901633**	–	520.8636146	–	636.229596	–	526.2849682	–	646.4465728	–	526.0530766	–	639.2272918	–

Table S3. The Akaike Information Criterion (AIC) values individually and in combination in the training cohort of TCGA-LUAD for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.02 (0.55-1.87)	0.961	1.29 (0.71-2.34)	0.403	1.23 (0.68-2.23)	0.49	–	–	1.56 (0.87-2.81)	0.138	–	–	–	–
Age 50	0.84 (0.54-1.29)	0.416	0.92 (0.60-1.42)	0.714	0.88 (0.58-1.35)	0.563	–	–	0.99 (0.65-1.52)	0.981	–	–	–	–
Age 70	1.17 (0.81-1.69)	0.41	1.19 (0.83-1.72)	0.348	1.25 (0.87-1.79)	0.234	–	–	1.28 (0.89-1.84)	0.176	–	–	–	–
Age 80+	1.36 (0.72-2.56)	0.342	1.42 (0.76-2.66)	0.269	1.50 (0.80-2.80)	0.206	–	–	1.51 (0.81-2.82)	0.196	–	–	–	–
TNM Stage II	2.13 (1.47-3.08)	6.14 E-05	2.30 (1.59-3.33)	8.93 E-06	–	–	2.09 (1.44-3.02)	8.98 E-05	–	–	2.27 (1.57-3.28)	1.20 E-05	–	–
TNM Stage III	2.74 (1.85-4.06)	5.01 E-07	3.21 (2.18-4.74)	3.73 E-09	–	–	2.78 (1.89-4.10)	2.38 E-07	–	–	3.26 (2.22-4.79)	1.71 E-09	–	–
TNM Stage IV	3.48 (1.97-6.14)	1.65 E-05	3.46 (1.98-6.04)	1.38 E-05	–	–	3.49 (2.00-6.06)	9.69 E-06	–	–	3.56 (2.05-6.17)	6.23 E-06	–	–
Risk score	2.91 (2.11-4.02)	9.34 E-11	–	–	3.23 (2.35-4.44)	5.67 E-13	2.87 (2.08-3.95)	1.09 E-10	–	–	–	–	3.20 (2.33-4.38)	4.79 E-13
AIC	**1799.476982**	–	1836.990477	–	1856.594352	–	1794.824375	–	1903.22415	–	1831.998534	–	1853.037326	–

DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR | Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.12 (0.61-2.04)	0.719	1.29 (0.71-2.34)	0.403	1.35 (0.75-2.43)	0.322	–	–	1.56 (0.87-2.81)	0.138	–	–	–	–
Age 50	0.88 (0.57-1.34)	0.542	0.92 (0.60-1.42)	0.714	0.93 (0.61-1.43)	0.751	–	–	0.99 (0.65-1.52)	0.981	–	–	–	–
Age 70	1.28 (0.89-1.86)	0.187	1.19 (0.83-1.72)	0.348	1.36 (0.95-1.96)	0.0959	–	–	1.28 (0.89-1.84)	0.176	–	–	–	–
Age 80+	1.46 (0.78-2.73)	0.241	1.42 (0.76-2.66)	0.269	1.56 (0.84-2.92)	0.162	–	–	1.51 (0.81-2.82)	0.196	–	–	–	–
TNM Stage II	2.24 (1.55-3.24)	1.67 E-05	2.30 (1.59-3.33)	8.93 E-06	–	–	2.20 (1.52-3.17)	2.73 E-05	–	–	2.27 (1.57-3.28)	1.20 E-05	–	–
TNM Stage III	2.77 (1.87-4.10)	3.55 E-07	3.21 (2.18-4.74)	3.73 E-09	–	–	2.86 (1.94-4.22)	1.07 E-07	–	–	3.26 (2.22-4.79)	1.71 E-09	–	–

TNM Stage IV	3.51 (2.00-6.18)	1.30 E-05	3.46 (1.98-6.04)	1.38 E-05	–	–	3.48 (2.00-6.06)	9.84 E-06	–	–	3.56 (2.05-6.17)	6.23 E-06	–	–
Risk score	2.76 (1.79-4.25)	4.34 E-06	–	–	3.00 (1.99-4.53)	1.76 E-07	2.62 (1.71-4.02)	1.00 E-05	–	–	–	–	2.88 (1.92-4.33)	3.32 E-07
AIC	**1817.667 212**	–	1836.990 477	–	1878.057 17	–	1814.175 729	–	1903.22 415	–	1831.998 534	–	1875.780 046	–

Table S4. The Akaike Information Criterion (AIC) values individually and in combination in the training cohort of TCGA-LUSC for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.45 (0.16-1.23)	0.118	0.52 (0.19-1.44)	0.209	0.43 (0.16-1.18)	0.102	–	–	0.50 (0.18-1.38)	0.181	–	–	–	–
Age 50	1.19 (0.75-1.89)	0.47	0.97 (0.61-1.53)	0.896	1.11 (0.70-1.75)	0.668	–	–	0.91 (0.58-1.44)	0.699	–	–	–	–
Age 70	1.11 (0.81-1.51)	0.52	1.12 (0.82-1.52)	0.473	1.03 (0.76-1.39)	0.867	–	–	1.04 (0.77-1.41)	0.789	–	–	–	–
Age 80+	1.26 (0.68-2.33)	0.457	1.32 (0.72-2.45)	0.369	1.19 (0.65-2.19)	0.577	–	–	1.24 (0.67-2.28)	0.494	–	–	–	–
TNM Stage II	1.25 (0.90-1.72)	0.181	1.18 (0.85-1.62)	0.325	–	–	1.23 (0.89-1.69)	0.21	–	–	1.15 (0.83-1.58)	0.395	–	–
TNM Stage III	1.61 (1.12-2.32)	0.00944	1.56 (1.09-2.25)	0.0152	–	–	1.57 (1.10-2.25)	0.0129	–	–	1.52 (1.07-2.17)	0.021	–	–
TNM Stage IV	3.34 (1.34-8.30)	0.00938	3.28 (1.32-8.13)	0.0105	–	–	3.24 (1.31-8.01)	0.0107	–	–	3.25 (1.31-8.02)	0.0107	–	–
Risk score	4.68 (2.31-9.52)	1.97E-05	–	–	4.47 (2.21-9.06)	3.23E-05	4.23 (2.13-8.42)	3.93E-05	–	–	–	–	4.08 (2.05-8.09)	5.86E-05
AIC	**2198.285135**	–	2214.69861	–	2224.456004	–	**2195.272821**	–	2239.980215	–	2210.412353	–	2220.801414	–

DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR | Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.55 (0.20-1.51)	0.248	0.52 (0.19-1.44)	0.209	0.54 (0.20-1.48)	0.231	–	–	0.50 (0.18-1.38)	0.181	–	–	–	–
Age 50	1.12 (0.70-1.77)	0.641	0.97 (0.61-1.53)	0.896	1.05 (0.67-1.66)	0.822	–	–	0.91 (0.58-1.44)	0.699	–	–	–	–
Age 70	1.17 (0.86-1.59)	0.331	1.12 (0.82-1.52)	0.473	1.08 (0.80-1.47)	0.61	–	–	1.04 (0.77-1.41)	0.789	–	–	–	–
Age 80+	1.30 (0.70-2.40)	0.405	1.32 (0.72-2.45)	0.369	1.22 (0.66-2.25)	0.517	–	–	1.24 (0.67-2.28)	0.494	–	–	–	–
TNM Stage II	1.26 (0.91-1.74)	0.16	1.18 (0.85-1.62)	0.325	–	–	1.24 (0.90-1.70)	0.198	–	–	1.15 (0.83-1.58)	0.395	–	–
TNM Stage III	1.62 (1.13-2.33)	0.00863	1.56 (1.09-2.25)	0.0152	–	–	1.57 (1.10-2.24)	0.0132	–	–	1.52 (1.07-2.17)	0.021	–	–

TNM Stage IV	3.56 (1.43-8.86)	0.00638	3.28 (1.32-8.13)	0.0105	–	–	3.42 (1.38-8.45)	0.0078	–	–	3.25 (1.31-8.02)	0.0107	–	–
Risk score	2.98 (1.86-4.79)	6.28E-06	–	–	2.86 (1.79-4.55)	9.97E-06	2.95 (1.85-4.72)	5.95E-06	–	–	–	–	2.87 (1.81-4.55)	7.72E-06
AIC	**2195.965992**	–	2214.69861	–	2222.274923	–	**2191.570054**	–	2239.980215	–	2210.412353	–	2216.972679	–

Table S5. The Akaike Information Criterion (AIC) values individually and in combination in the training cohort of TCGA-PAAD for EMT- and DDR-related prognostic signatures.

EMT-related signature

Clinical parameter	EMT | Nomogram COX PH		EMT | Age + TNM stage COX PH		EMT | Age + Risk score COX PH		EMT | TNM stage + Risk score COX PH		EMT | Age only COX PH		EMT | TNM stage only COX PH		EMT | Risk score only COX PH	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.84 (0.41-1.73)	0.634	0.89 (0.43-1.84)	0.755	0.84 (0.40-1.72)	0.626	–	–	0.89 (0.43-1.83)	0.751	–	–	–	–
Age 50	0.83 (0.44-1.55)	0.552	0.96 (0.51-1.80)	0.902	0.80 (0.42-1.49)	0.475	–	–	0.89 (0.48-1.65)	0.709	–	–	–	–
Age 70	1.00 (0.59-1.71)	0.992	1.15 (0.67-1.95)	0.616	1.00 (0.59-1.69)	0.986	–	–	1.25 (0.73-2.12)	0.417	–	–	–	–
Age 80+	2.71 (1.30-5.65)	0.00769	3.36 (1.62-6.98)	0.00112	2.73 (1.32-5.68)	0.00703	–	–	3.61 (1.75-7.46)	5.26E-04	–	–	–	–
TNM Stage II	0.86 (0.39-1.89)	0.711	2.21 (1.01-4.86)	0.0481	–	–	0.84 (0.38-1.83)	0.66	–	–	2.35 (1.08-5.13)	0.0317	–	–
TNM Stage III	0.49 (0.06-3.99)	0.502	1.10 (0.13-9.05)	0.93	–	–	0.44 (0.05-3.57)	0.439	–	–	1.06 (0.13-8.65)	0.96	–	–
TNM Stage IV	0.76 (0.16-3.73)	0.74	2.24 (0.45-11.02)	0.323	–	–	0.69 (0.14-3.34)	0.641	–	–	2.14 (0.44-10.44)	0.348	–	–
Risk score	5.00 (2.80-8.93)	5.13E-08	–	–	5.16 (2.91-9.16)	2.02E-08	5.01 (2.85-8.80)	2.02E-08	–	–	–	–	5.15 (2.95-9.00)	8.09E-09
AIC	**780.3477493**	–	817.4400355	–	**775.9280368**	–	780.9607668	–	821.2886307	–	819.8416171	–	776.9963673	–

DDR-related signature

Clinical parameter	DDR | Nomogram COX PH		DDR | Age + TNM stage COX PH		DDR | Age + Risk score COX PH		DDR | TNM stage + Risk score COX PH		DDR | Age only COX PH		DDR | TNM stage only COX PH		DDR | Risk score only COX PH	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.76 (0.37-1.58)	0.467	0.89 (0.43-1.84)	0.755	0.76 (0.37-1.57)	0.459	–	–	0.89 (0.43-1.83)	0.751	–	–	–	–
Age 50	0.99 (0.52-1.86)	0.967	0.96 (0.51-1.80)	0.902	0.98 (0.53-1.85)	0.962	–	–	0.89 (0.48-1.65)	0.709	–	–	–	–
Age 70	1.04 (0.61-1.77)	0.886	1.15 (0.67-1.95)	0.616	1.04 (0.61-1.78)	0.879	–	–	1.25 (0.73-2.12)	0.417	–	–	–	–
Age 80+	3.52 (1.68-7.41)	8.92E-04	3.36 (1.62-6.98)	0.00112	3.55 (1.69-7.46)	8.04E-04	–	–	3.61 (1.75-7.46)	5.26E-04	–	–	–	–
TNM Stage II	1.21 (0.56-2.64)	0.627	2.21 (1.01-4.86)	0.0481	–	–	1.18 (0.54-2.57)	0.672	–	–	2.35 (1.08-5.13)	0.0317	–	–

TNM Stage III	0.81 (0.10-6.74)	0.848	1.10 (0.13-9.05)	0.93	–	–	0.65 (0.08-5.31)	0.687	–	–	1.06 (0.13-8.65)	0.96	–	–
TNM Stage IV	1.35 (0.27-6.60)	0.714	2.24 (0.45-11.02)	0.323	–	–	1.20 (0.25-5.84)	0.818	–	–	2.14 (0.44-10.44)	0.348	–	–
Risk score	4.68 (2.76-7.96)	1.13E-08	–	–	4.84 (2.88-8.13)	2.51E-09	4.45 (2.64-7.50)	2.21E-08	–	–	–	–	4.62 (2.77-7.71)	4.37E-09
AIC	**775.9367781**	–	817.4400355	–	**771.1209051**	–	779.7294911	–	821.2886307	–	819.8416171	–	775.1293176	–

Table S6. The Akaike Information Criterion (AIC) values individually and in combination in the training cohort of TCGA-STAD for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.53 (0.26-1.09)	0.084	0.50 (0.24-1.02)	0.0572	0.54 (0.27-1.11)	0.0947	–	–	0.58 (0.28-1.18)	0.132	–	–	–	–
Age 50	0.56 (0.35-0.91)	0.0185	0.65 (0.41-1.05)	0.0778	0.67 (0.42-1.05)	0.0813	–	–	0.76 (0.48-1.20)	0.245	–	–	–	–
Age 70	1.13 (0.76-1.69)	0.539	1.17 (0.79-1.74)	0.431	1.08 (0.73-1.59)	0.694	–	–	1.10 (0.75-1.62)	0.617	–	–	–	–
Age 80+	1.13 (0.48-2.69)	0.778	0.98 (0.41-2.30)	0.956	1.10 (0.50-2.45)	0.811	–	–	0.95 (0.43-2.09)	0.889	–	–	–	–
TNM Stage II	1.13 (0.59-2.16)	0.715	1.33 (0.70-2.53)	0.379	–	–	1.11 (0.58-2.11)	0.759	–	–	1.28 (0.68-2.43)	0.443	–	–
TNM Stage III	1.73 (0.95-3.15)	0.0745	2.16 (1.19-3.90)	0.0108	–	–	1.61 (0.89-2.94)	0.117	–	–	2.01 (1.11-3.62)	0.0206	–	–
TNM Stage IV	3.32 (1.58-6.98)	0.00152	3.91 (1.88-8.12)	2.53E-04	–	–	2.90 (1.41-5.97)	0.00394	–	–	3.20 (1.57-6.54)	0.00142	–	–
Risk score	4.28 (2.79-6.55)	2.45E-11	–	–	4.35 (2.84-6.65)	1.22E-11	4.02 (2.64-6.12)	9.70E-11	–	–	–	–	4.07 (2.68-6.18)	4.99E-11
AIC	**1441.886788**	–	1484.36849	–	1523.124755	–	1446.631588	–	1567.278149	–	1486.77974	–	1523.205048	–

DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR | Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.47 (0.23-0.97)	0.0401	0.50 (0.24-1.02)	0.0572	0.46 (0.22-0.95)	0.036	–	–	0.58 (0.28-1.18)	0.132	–	–	–	–
Age 50	0.58 (0.36-0.93)	0.0246	0.65 (0.41-1.05)	0.0778	0.64 (0.40-1.02)	0.0616	–	–	0.76 (0.48-1.20)	0.245	–	–	–	–
Age 70	1.17 (0.79-1.74)	0.435	1.17 (0.79-1.74)	0.431	1.11 (0.75-1.63)	0.597	–	–	1.10 (0.75-1.62)	0.617	–	–	–	–
Age 80+	0.94 (0.40-2.23)	0.89	0.98 (0.41-2.30)	0.956	0.90 (0.40-1.98)	0.786	–	–	0.95 (0.43-2.09)	0.889	–	–	–	–
TNM Stage II	1.30 (0.69-2.48)	0.419	1.33 (0.70-2.53)	0.379	–	–	1.33 (0.70-2.52)	0.388	–	–	1.28 (0.68-2.43)	0.443	–	–

TNM Stage III	2.03 (1.12-3.67)	0.0191	2.16 (1.19-3.90)	0.0108	–	–	1.99 (1.10-3.60)	0.0227	–	–	2.01 (1.11-3.62)	0.0206	–	–
TNM Stage IV	3.44 (1.66-7.14)	8.94E-04	3.91 (1.88-8.12)	2.53E-04	–	–	3.17 (1.55-6.49)	0.00161	–	–	3.20 (1.57-6.54)	0.00142	–	–
Risk score	3.48 (2.04-5.94)	4.90E-06	–	–	3.85 (2.29-6.48)	3.63E-07	2.95 (1.78-4.87)	2.64E-05	–	–	–	–	3.14 (1.93-5.13)	4.43E-06
AIC	**1465.235998**	–	1484.36849	–	1543.141459	–	1471.171409	–	1567.278149	–	1486.77974	–	–	1545.786918

Appendix C — Supplementary Files 3

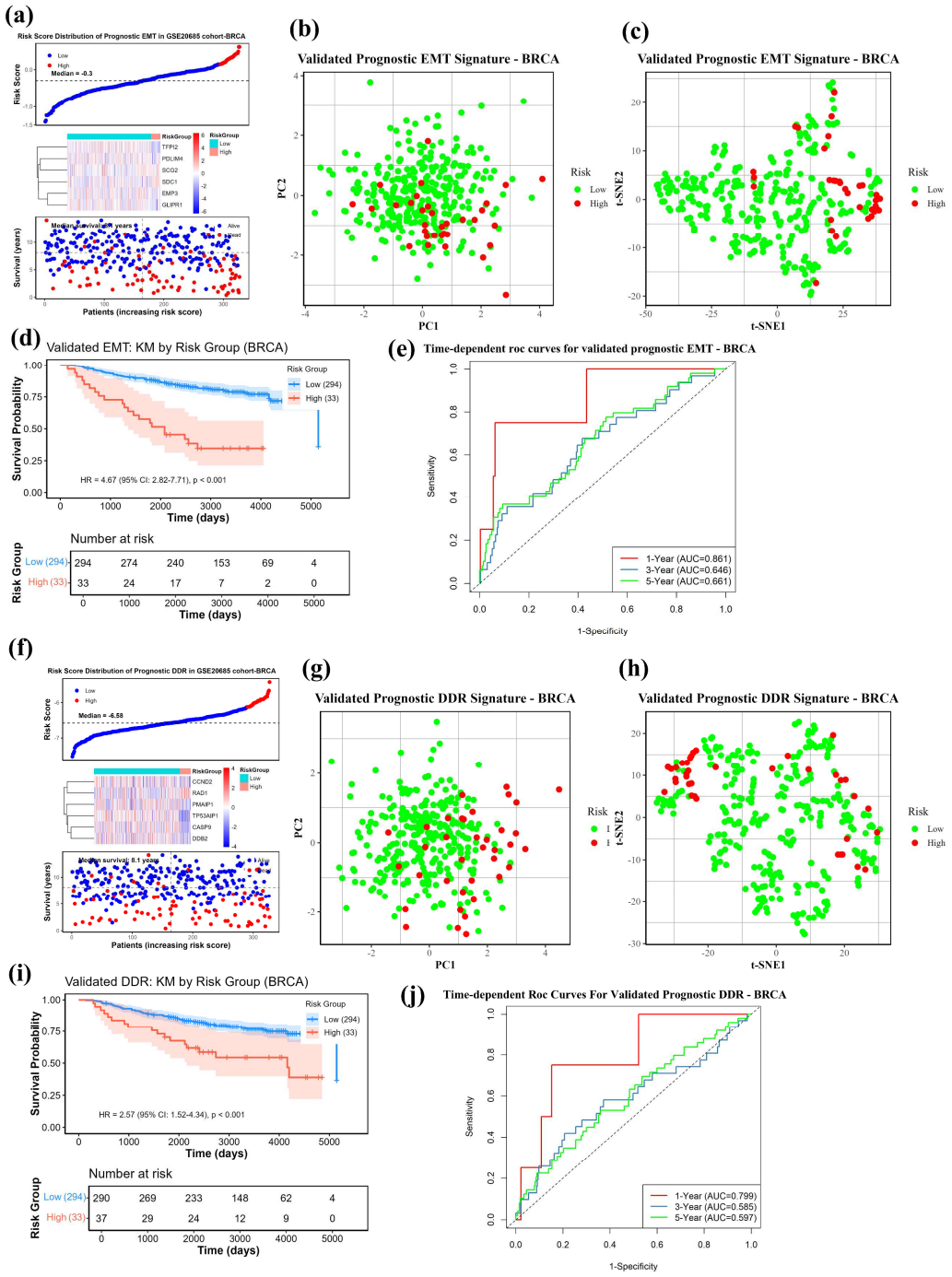


Figure S1. Risk score distribution, Kaplan-Meier estimates, tSNE, PC, and Time-dependent ROC curves in the validation cohort of GSE20685 (BRCA) for EMT and DDR prognostic gene signatures. Risk score analyses include distributions, survival status, and heatmaps of the (a) EMT and (f) DDR-related genes expression. The PCA plot of risk score in the validation cohort

of BRCA for (b) EMT and (g) DDR-related prognostic gene signatures. The t-SNE analysis of risk score in the validation cohort of BRCA for (c) EMT and (h) DDR-related gene signatures. Kaplan-Meier estimates for OS based on the (d) EMT and (i) DDR-related prognostic gene signatures. Time-dependent ROC curves for (e) EMT and (j) DDR-related gene signatures.

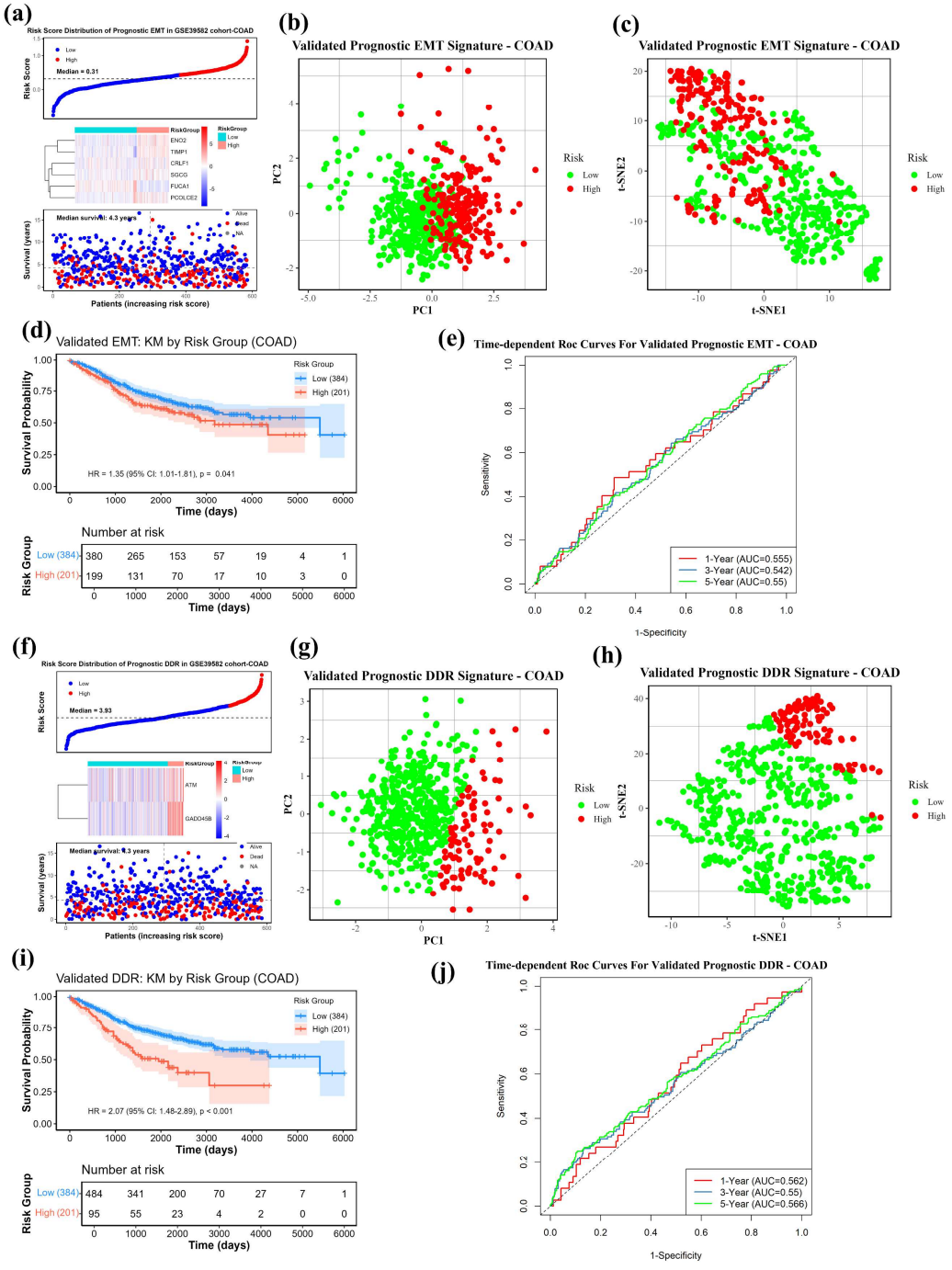


Figure S2. Risk score distribution, Kaplan-Meier estimates, tSNE, PC, and Time-dependent ROC curves in the validation cohort of GSE39582 (COAD) for EMT and DDR prognostic gene

signatures. Risk score analyses include distributions, survival status, and heatmaps of the (a) EMT and (f) DDR-related genes expression. The PCA plot of risk score in the validation cohort of COAD-GSE39582 for (b) EMT and (g) DDR-related prognostic gene signatures. The t-SNE analysis of risk score in the validation cohort of COAD-GSE39582 for (c) EMT and (h) DDR-related gene signatures. Kaplan-Meier estimates for OS based on the (d) EMT and (i) DDR-related prognostic gene signatures. Time-dependent ROC curves for (e) EMT and (j) DDR-related gene signatures.

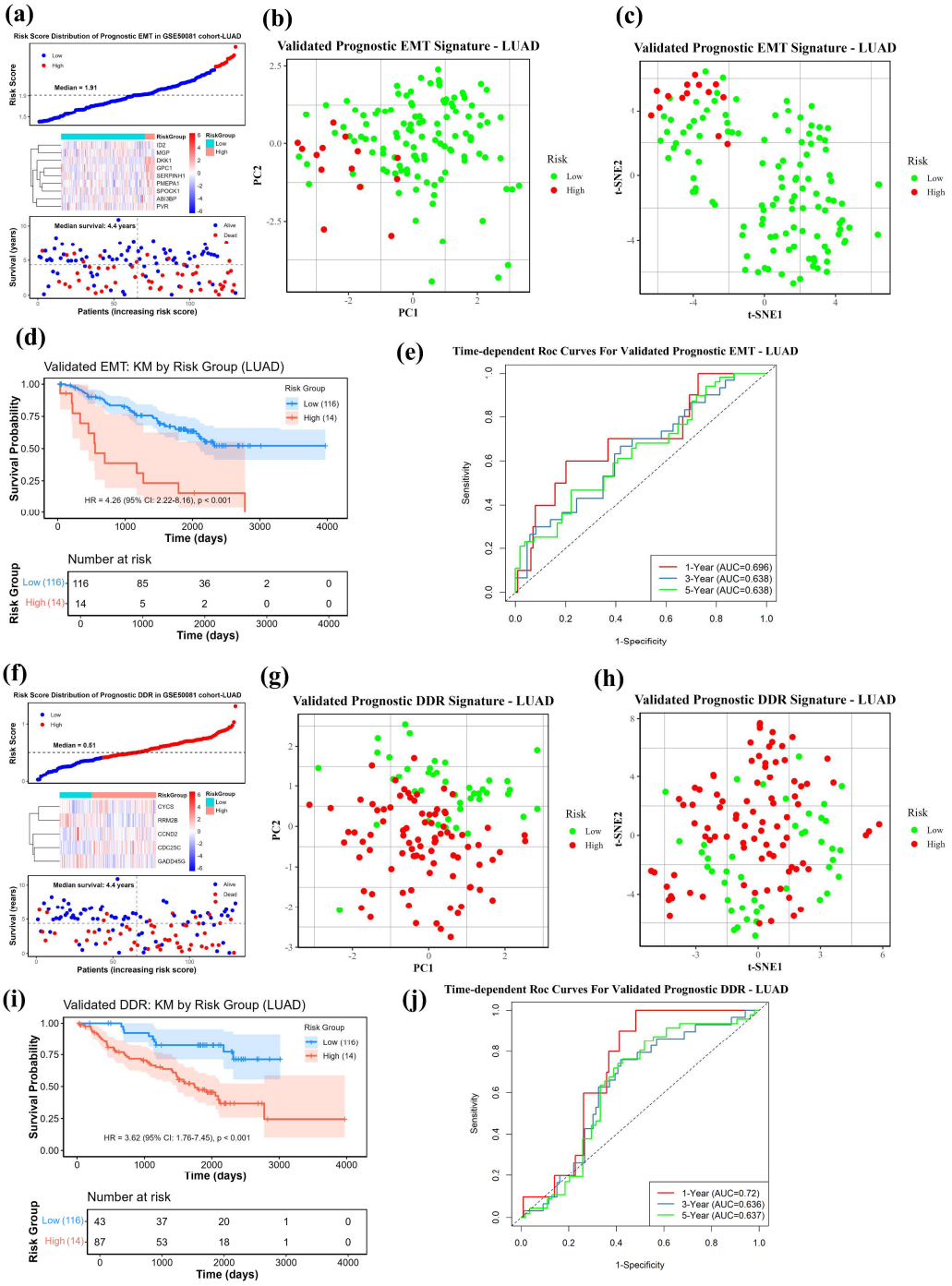


Figure S3. Risk score distribution, Kaplan-Meier estimates, tSNE, PC, and Time-dependent ROC curves in the validation cohort of GSE50081 (LUAD) for EMT and DDR prognostic gene signatures. Risk score analyses include distributions, survival status, and heatmaps of the (a) EMT and (f) DDR-related genes expression. The PCA plot of risk score in the validation cohort of LUAD-GSE50081 for (b) EMT and (g) DDR-related prognostic gene signatures. The t-SNE analysis of risk score in the validation cohort of LUAD-GSE50081 for (c) EMT and (h) DDR-

related gene signatures. Kaplan-Meier estimates for OS based on the (d) EMT and (i) DDR-related prognostic gene signatures. Time-dependent ROC curves for (e) EMT and (j) DDR-related gene signatures.

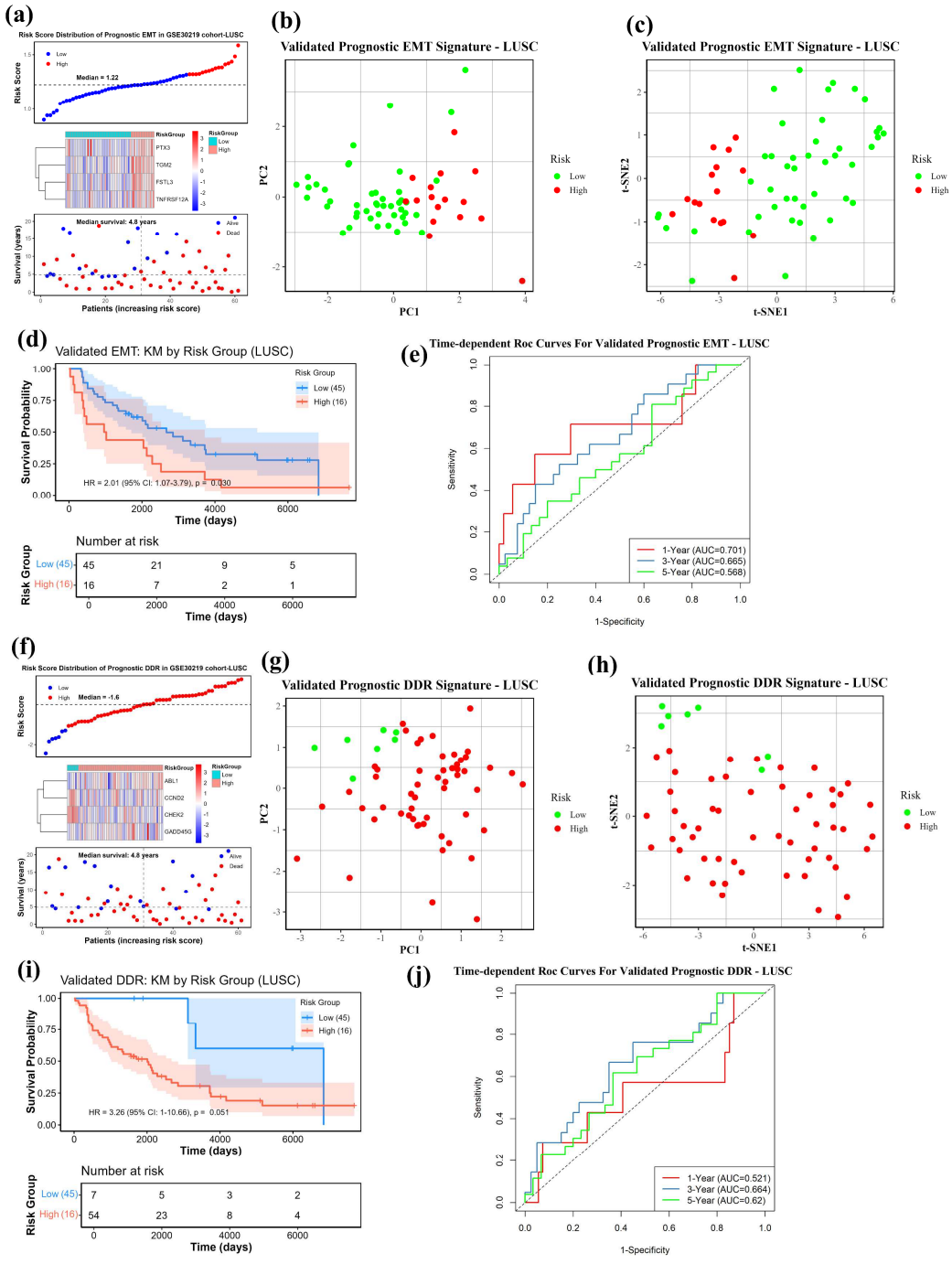


Figure S4. Risk score distribution, Kaplan-Meier estimates, tSNE, PC, and Time-dependent ROC curves in the validation cohort of GSE30219 (LUSC) for EMT and DDR prognostic gene signatures. Risk score analyses include distributions, survival status, and heatmaps of the (a)

EMT and (f) DDR-related genes expression. The PCA plot of risk score in the validation cohort of LUSC-GSE30219 for (b) EMT and (g) DDR-related prognostic gene signatures. The t-SNE analysis of risk score in the validation cohort of LUSC-GSE30219 for (c) EMT and (h) DDR-related gene signatures. Kaplan-Meier estimates for OS based on the (d) EMT and (i) DDR-related prognostic gene signatures. Time-dependent ROC curves for (e) EMT and (j) DDR-related gene signatures.

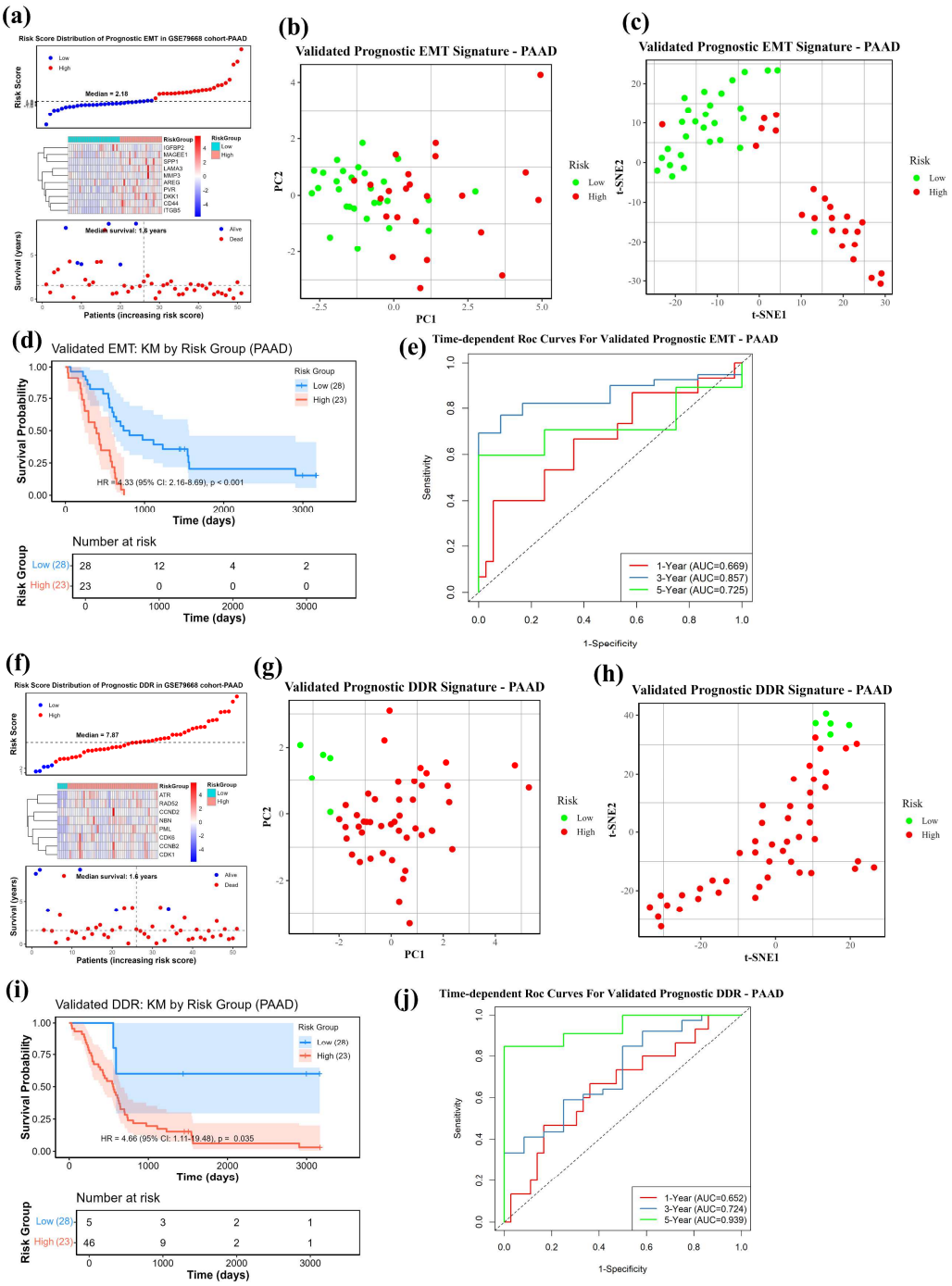


Figure S5. Risk score distribution, Kaplan-Meier estimates, tSNE, PC, and Time-dependent ROC curves in the validation cohort of GSE79668 (PAAD) for EMT and DDR prognostic gene signatures. Risk score analyses include distributions, survival status, and heatmaps of the (a) EMT and (f) DDR-related genes expression. The PCA plot of risk score in the validation cohort of PAAD-GSE79668 for (b) EMT and (g) DDR-related prognostic gene signatures. The t-SNE analysis of risk score in the validation cohort of PAAD-GSE79668 for (c) EMT and (h) DDR-related gene signatures. Kaplan-Meier estimates for OS based on the (d) EMT and (i) DDR-related prognostic gene signatures. Time-dependent ROC curves for (e) EMT and (j) DDR-related gene signatures.

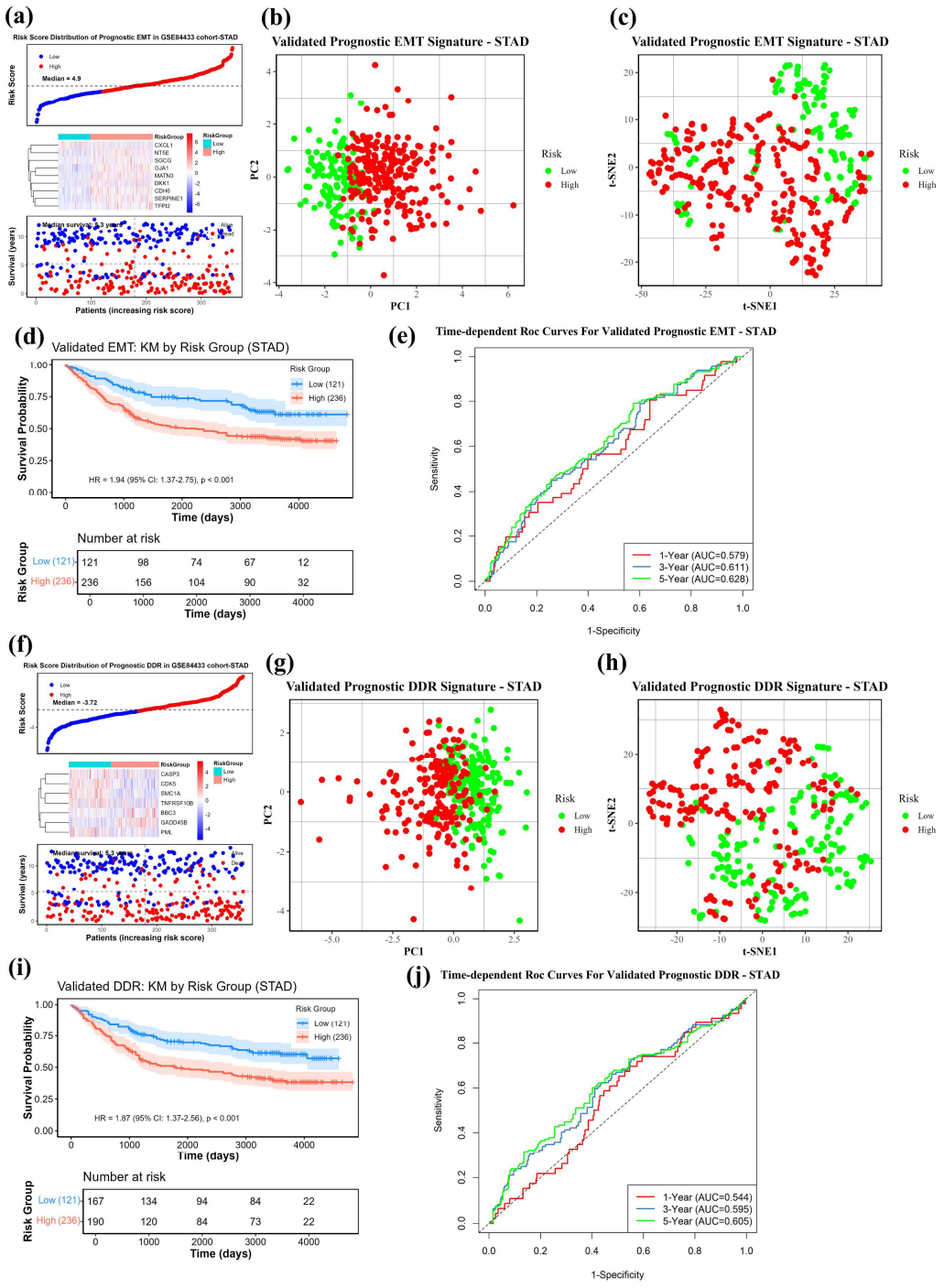


Figure S6. Risk score distribution, Kaplan-Meier estimates, tSNE, PC, and Time-dependent ROC curves in the validation cohort of GSE84433 (STAD) for EMT and DDR prognostic gene signatures. Risk score analyses include distributions, survival status, and heatmaps of the (a) EMT and (f) DDR-related genes expression. The PCA plot of risk score in the validation cohort of STAD-GSE84433 for (b) EMT and (g) DDR-related prognostic gene signatures. The t-SNE analysis of risk score in the validation cohort of STAD-GSE84433 for (c) EMT and (h) DDR-

related gene signatures. Kaplan-Meier estimates for OS based on the (d) EMT and (i) DDR-related prognostic gene signatures. Time-dependent ROC curves for (e) EMT and (j) DDR-related gene signatures.

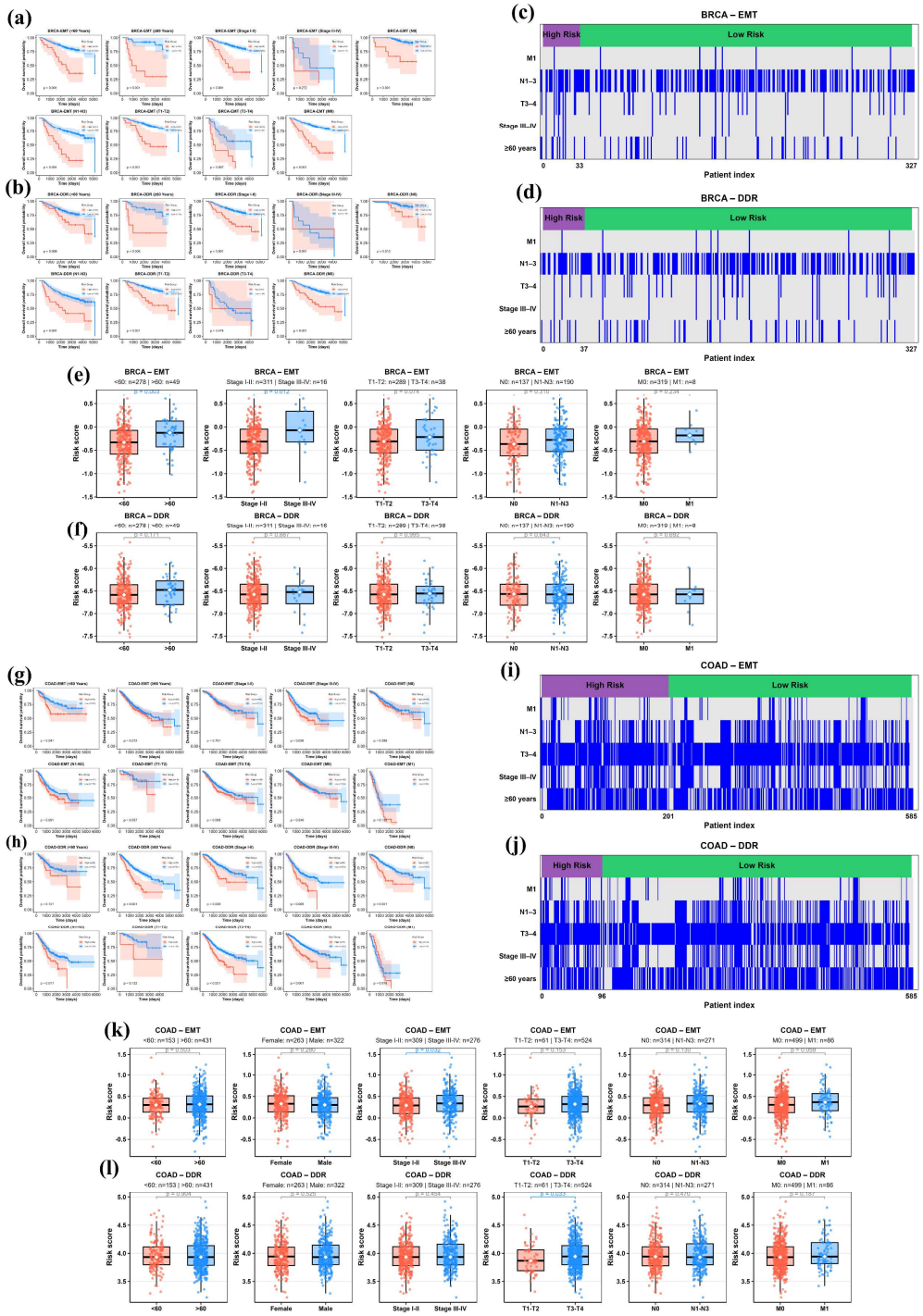


Figure S7. Clinicopathological validation of identified prognostic EMT- and DDR-related gene signatures in BRCA-GSE20685 and COAD-GSE39582 datasets. Kaplan-Meier estimates for OS based on the EMT- and DDR-related gene signatures in subgroups stratified according to

age, TNM stage, T, N, and M in BRCA-GSE20685 (a, b), and COAD-GSE39582 cohorts (g, h). Heatmaps of association between the risk score with clinical characteristics in BRCA-GSE20685 (c, d) and COAD-GSE39582 (i, j) cohort based on prognostic EMT- and DDR-related gene signatures. Relationship between risk score and clinicopathologic characteristics for EMT- and DDR-related prognostic gene signatures in BRCA-GSE20685 (e, f), and COAD-GSE39582 cohorts (k, l).

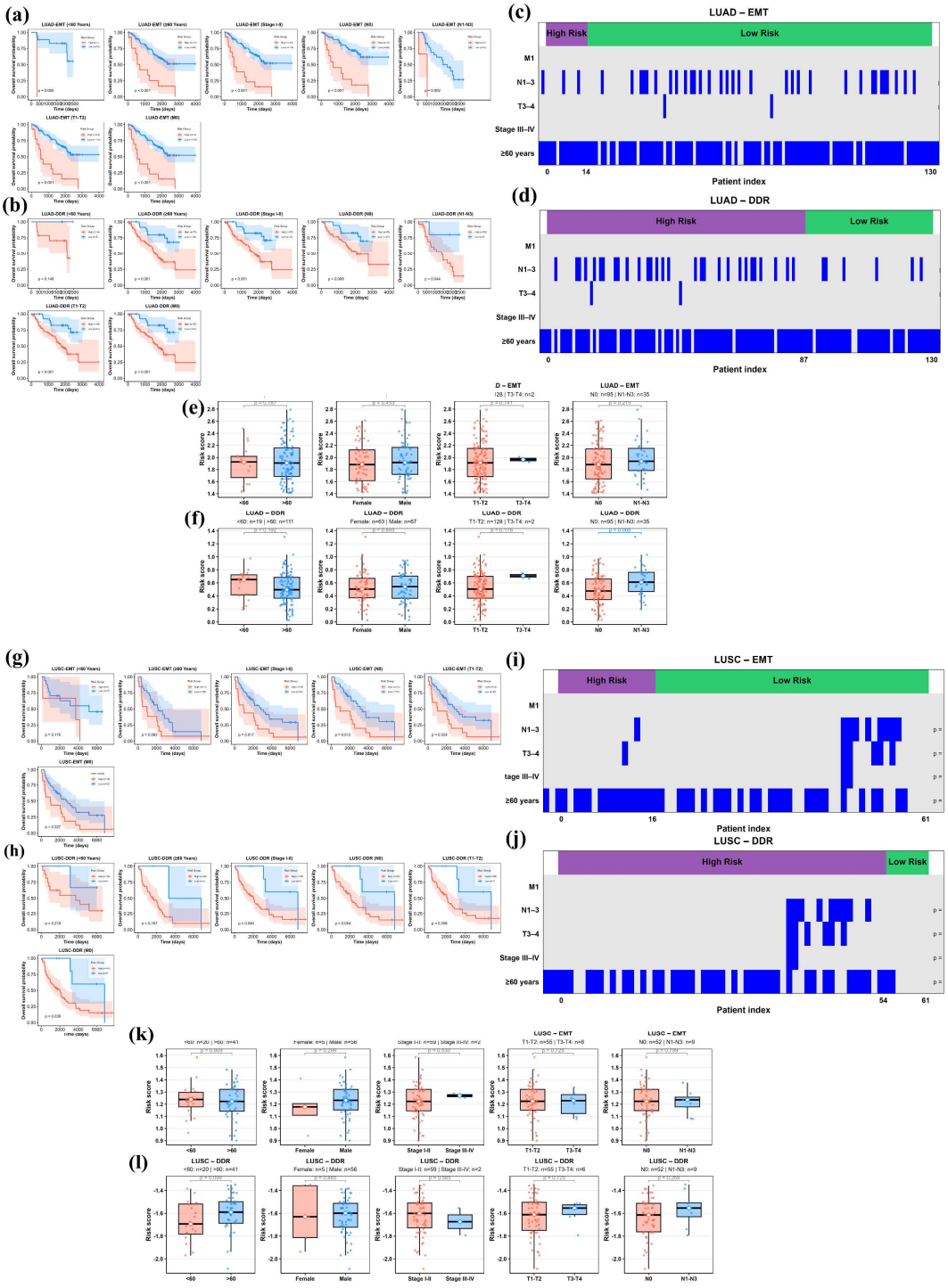


Figure S8. Clinicopathological validation of identified prognostic EMT- and DDR-related gene signatures in LUAD-GSE50081 and LUSC-GSE30219 datasets. Kaplan-Meier estimates for OS based on the EMT- and DDR-related gene signatures in subgroups stratified according to age, TNM stage, T, N, and M in LUAD-GSE50081 (a, b), and LUSC-GSE30219 cohorts (g, h). Heatmaps of association between the risk score with clinical characteristics in LUAD-

GSE50081 (c, d) and LUSC-GSE30219 (i, j) cohort based on prognostic EMT- and DDR-related gene signatures. Relationship between risk score and clinicopathologic characteristics for EMT- and DDR-related prognostic gene signatures in LUAD-GSE50081 (e, f), and LUSC-GSE30219 cohorts (k, l).

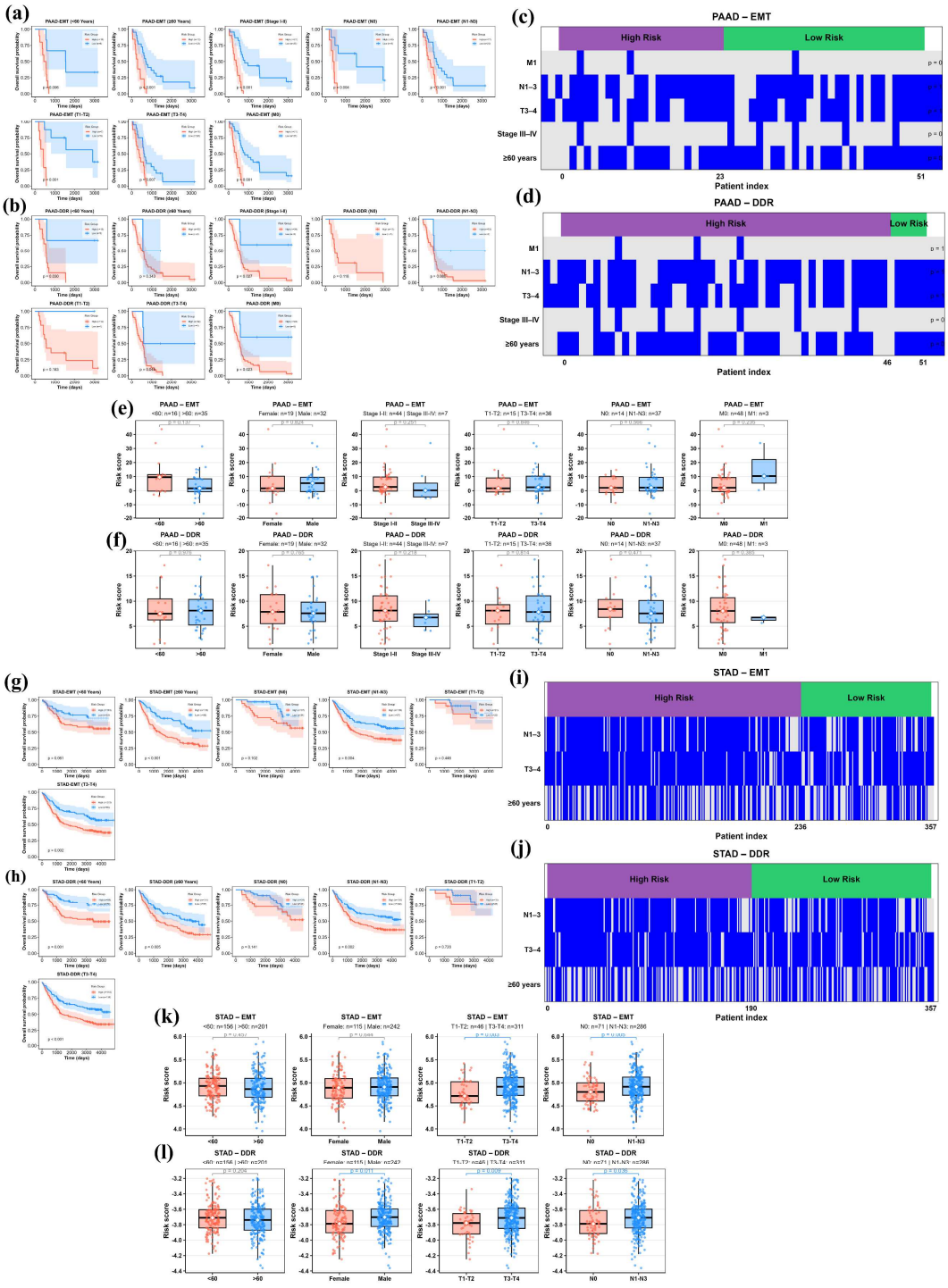


Figure S9. Clinicopathological validation of identified prognostic EMT- and DDR-related gene signatures in PAAD-GSE79668 and STAD-GSE84433 datasets. Kaplan-Meier estimates for OS based on the EMT- and DDR-related gene signatures in subgroups stratified according to age, TNM stage, T, N, and M in PAAD-GSE79668 (a, b), and STAD-GSE84433 cohorts (g, h). Heatmaps of association between the risk score with clinical characteristics in PAAD-GSE79668 (c, d) and STAD-GSE84433 (i, j) cohort based on prognostic EMT- and DDR-related gene signatures. Relationship between risk score and clinicopathologic characteristics for EMT- and DDR-related prognostic gene signatures in PAAD-GSE79668 (e, f), and STAD-GSE84433 cohorts (k, l).

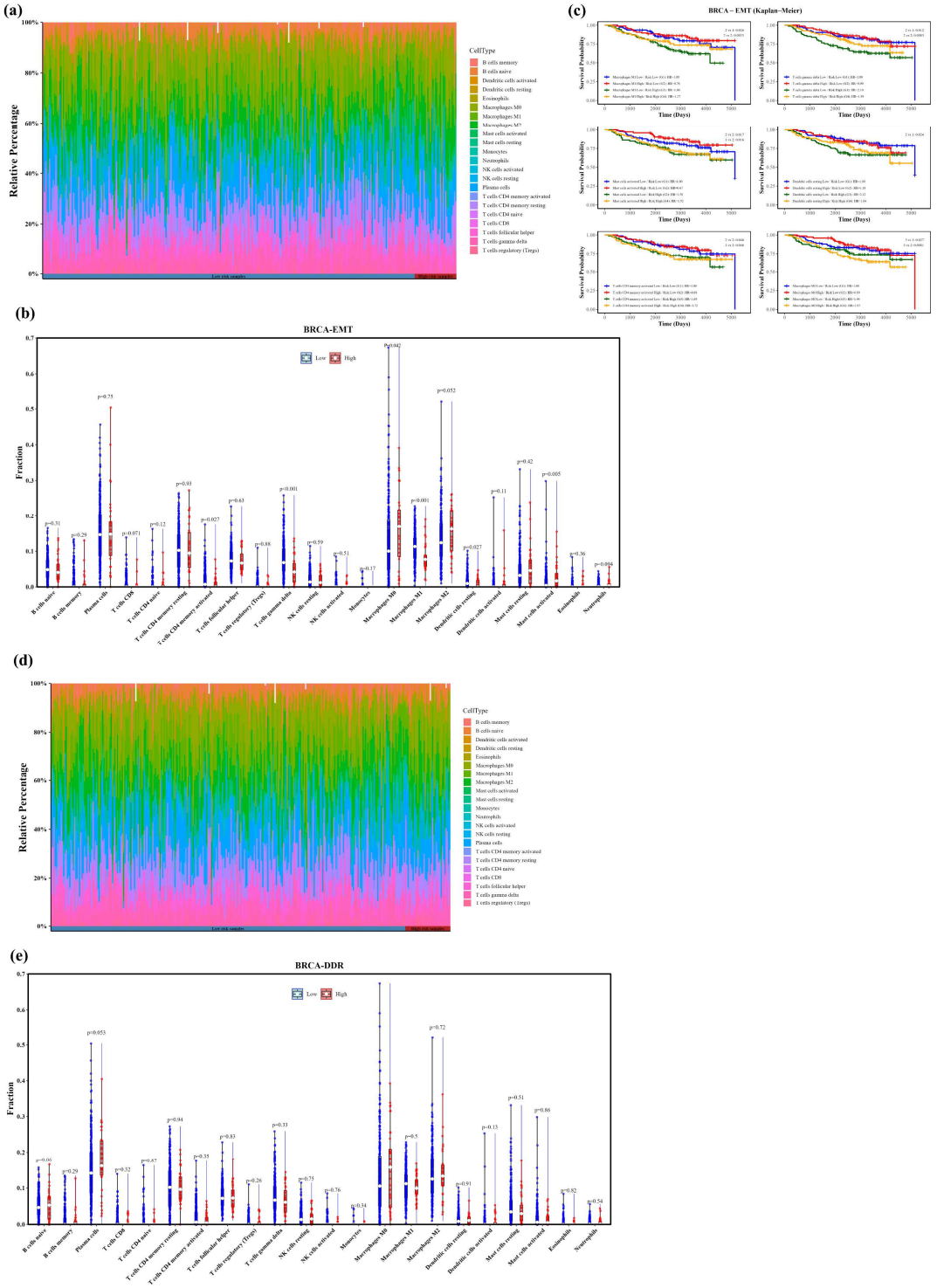


Figure S10. Immune infiltration analysis of prognostic EMT- and DDR gene signatures in the BRCA-GSE20685 validation cohort. The bar chart shows the infiltration of 22 types of immune cells in samples from high- and low-risk groups for (a) EMT- and (d) DDR-related gene signatures in the validation GSE20685 BRCA cohort. The difference of 22 kinds of immune cells between the low-risk group and the high-risk group of (b) EMT and (e) DDR-related gene

signatures ($p < 0.05$ indicated that it was statistically significant). (c) Combined Kaplan–Meier analysis of top significant immune cells and risk scores in EMT-related gene signature.

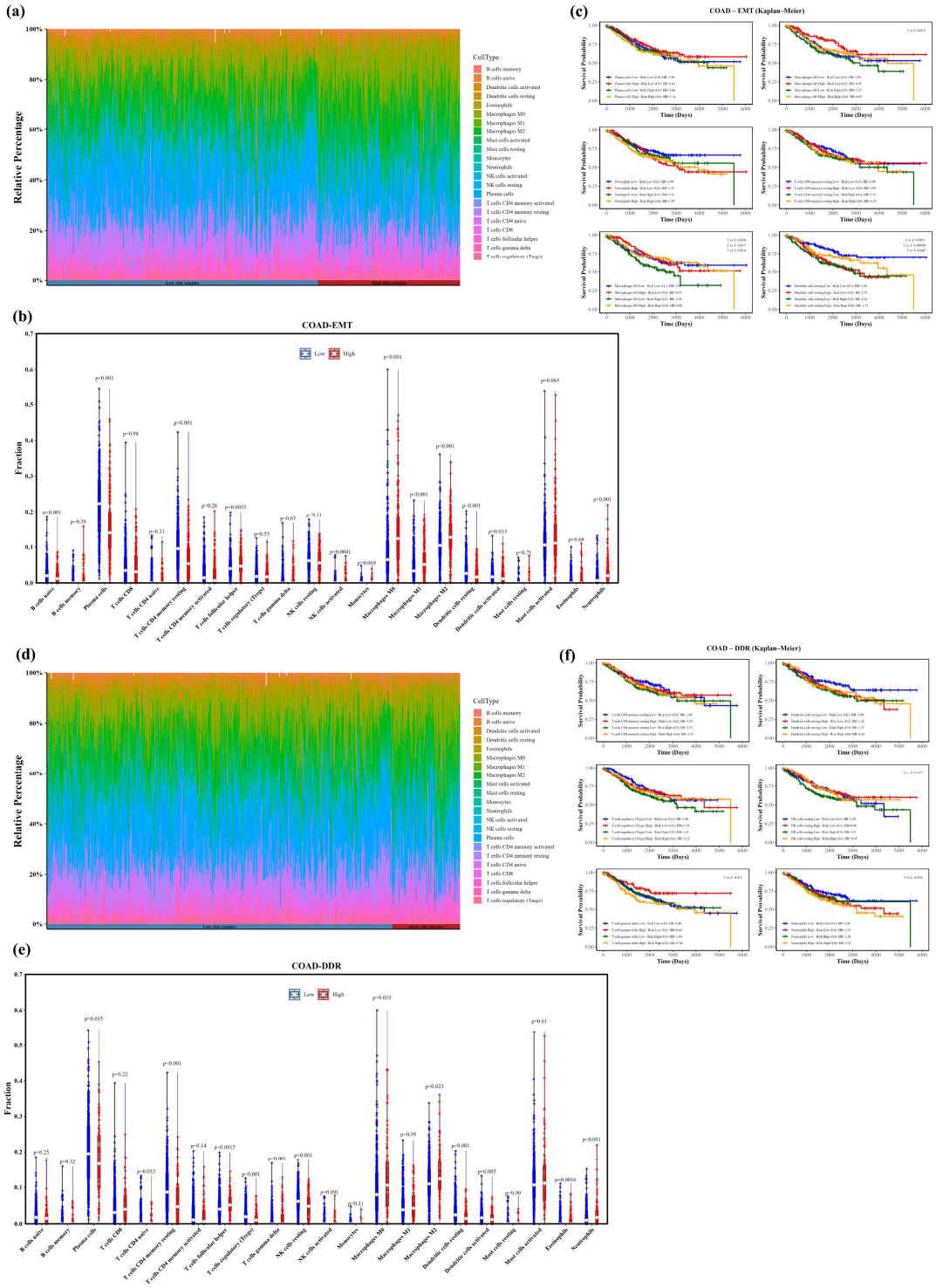


Figure S11. Immune infiltration analysis of prognostic EMT- and DDR gene signatures in the COAD-GSE39582 validation cohort. The bar chart shows the infiltration of 22 types of immune

cells in samples from high- and low-risk groups for (a) EMT- and (d) DDR-related gene signatures in the validation COAD-GSE39582 cohort. The difference in 22 kinds of immune cells between the low-risk group and the high-risk group, as indicated by (b) EMT- and (e) DDR-related gene signatures, was statistically significant ($p < 0.05$). Combined Kaplan–Meier analysis of top significant immune cells and risk scores in (c) EMT- and (f) DDR-related gene signature.

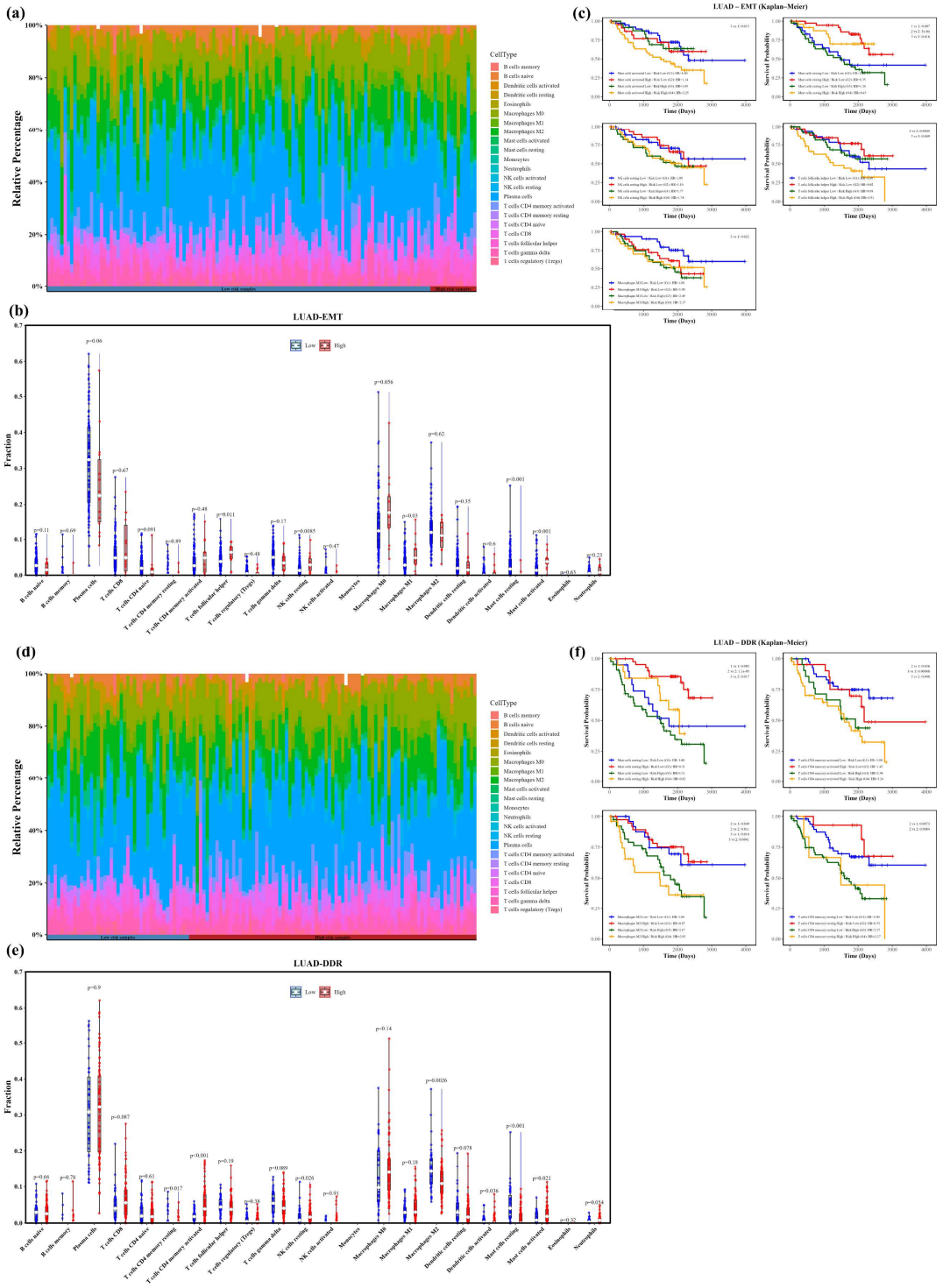


Figure S12. Immune infiltration analysis of prognostic EMT- and DDR gene signatures in the LUAD-GSE50081 validation cohort. The bar chart shows the infiltration of 22 types of immune cells in samples from high- and low-risk groups for (a) EMT- and (d) DDR-related gene signatures in the validation LUAD-GSE50081 cohort. The difference in 22 kinds of immune cells between the low-risk group and the high-risk group, as indicated by (b) EMT- and (e)

DDR-related gene signatures, was statistically significant ($p < 0.05$). Combined Kaplan–Meier analysis of top significant immune cells and risk scores in (c) EMT- and (f) DDR-related gene signature.

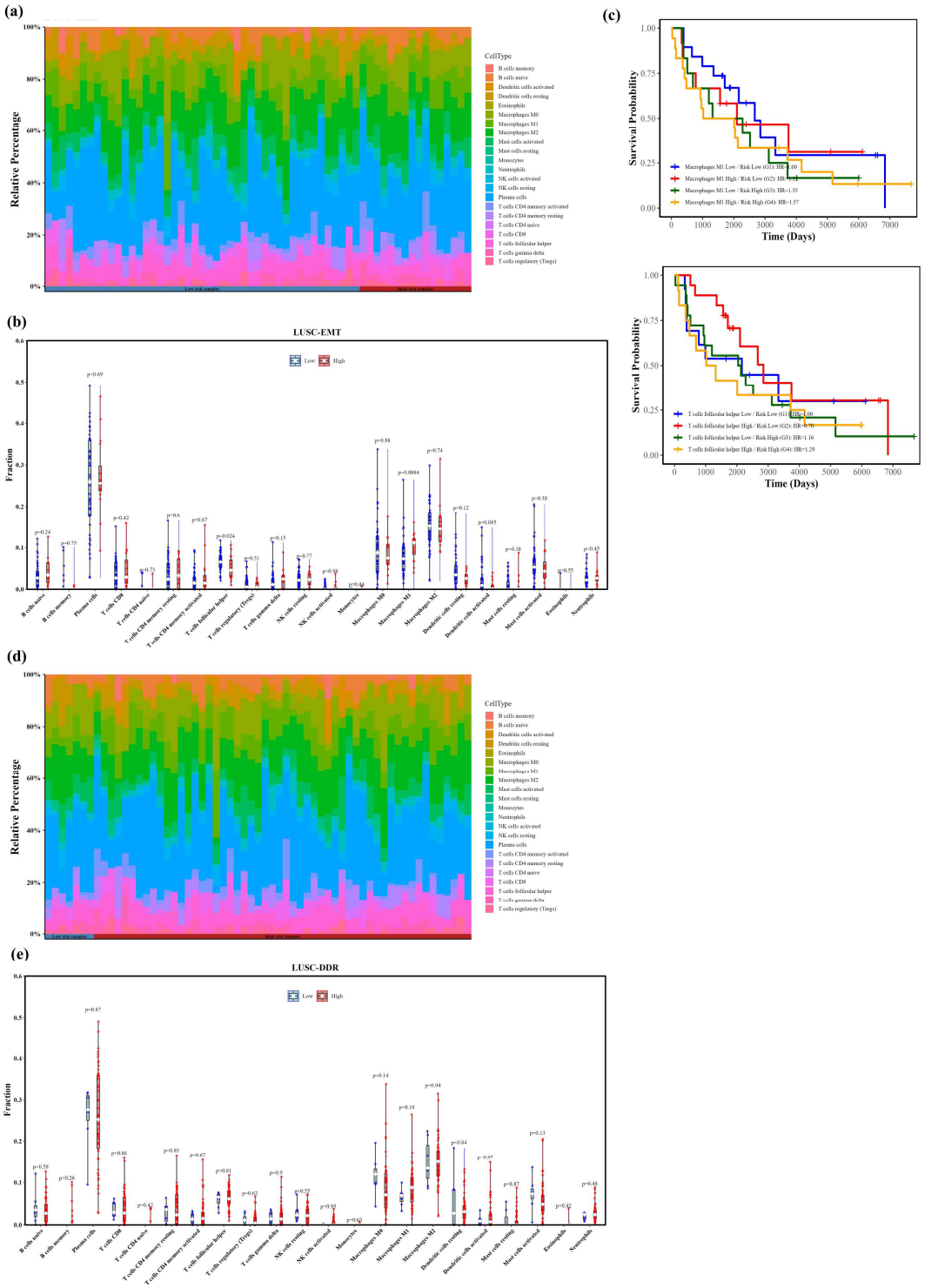


Figure S13. Immune infiltration analysis of prognostic EMT- and DDR gene signatures in the LUSC-GSE30219 validation cohort. The bar chart shows the infiltration of 22 types of immune

cells in samples from high- and low-risk groups for (a) EMT- and (d) DDR-related gene signatures in the validation LUSC-GSE30219 cohort. The difference in 22 kinds of immune cells between the low-risk group and the high-risk group, as indicated by (b) EMT- and (e) DDR-related gene signatures, was statistically significant ($p < 0.05$). Combined Kaplan–Meier analysis of top significant immune cells and risk scores in (c) EMT-related gene signature.

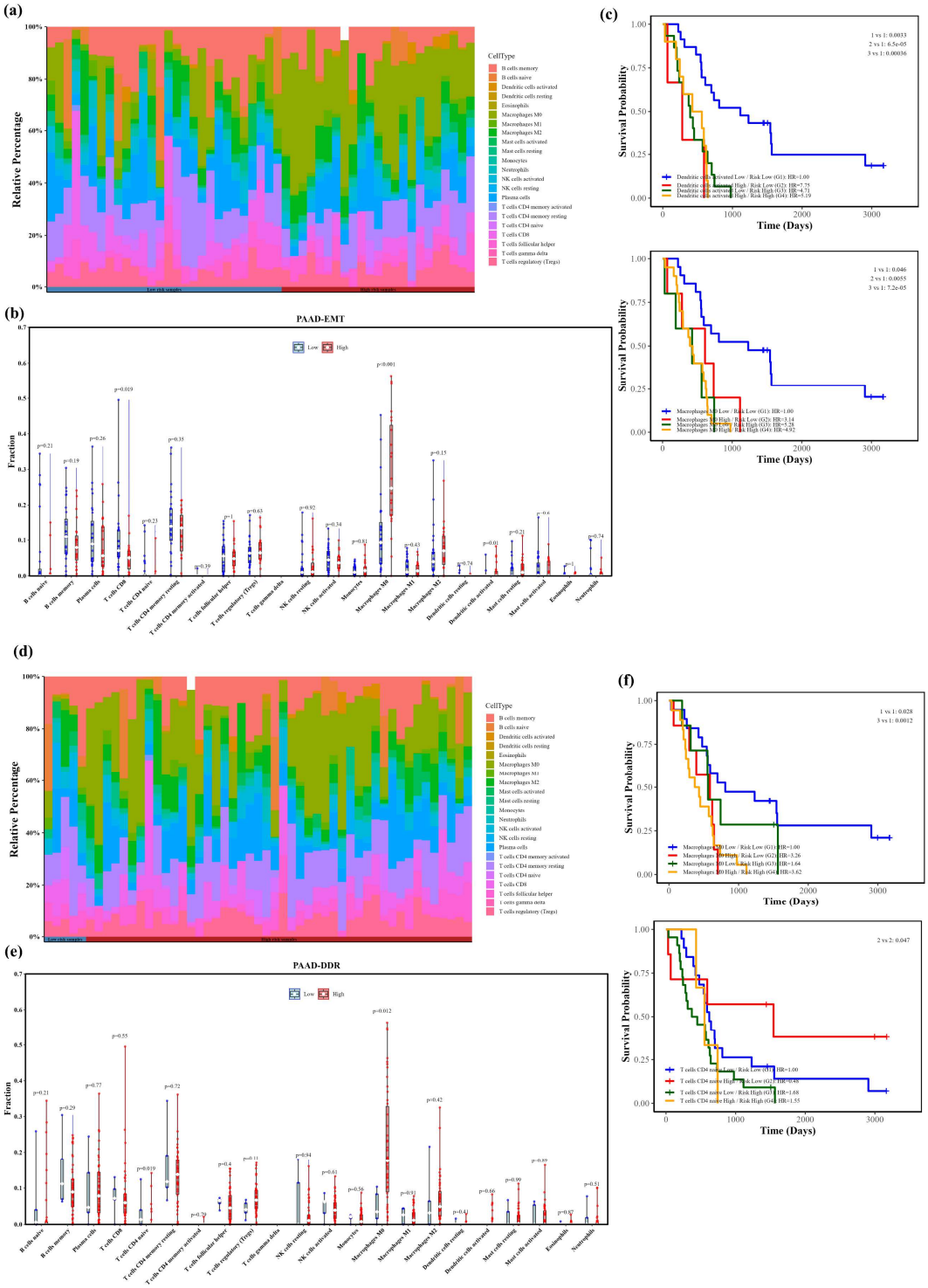


Figure S14. Immune infiltration analysis of prognostic EMT- and DDR gene signatures in the PAAD-GSE79668 validation cohort. The bar chart shows the infiltration of 22 types of immune cells in samples from high- and low-risk groups for (a) EMT- and (d) DDR-related gene signatures in the validation PAAD-GSE79668 cohort. The difference in 22 kinds of immune cells between the low-risk group and the high-risk group, as indicated by (b) EMT- and (e) DDR-related gene signatures, was statistically significant ($p < 0.05$). Combined Kaplan–Meier analysis of top significant immune cells and risk scores in (c) EMT- and (f) DDR-related gene signature.

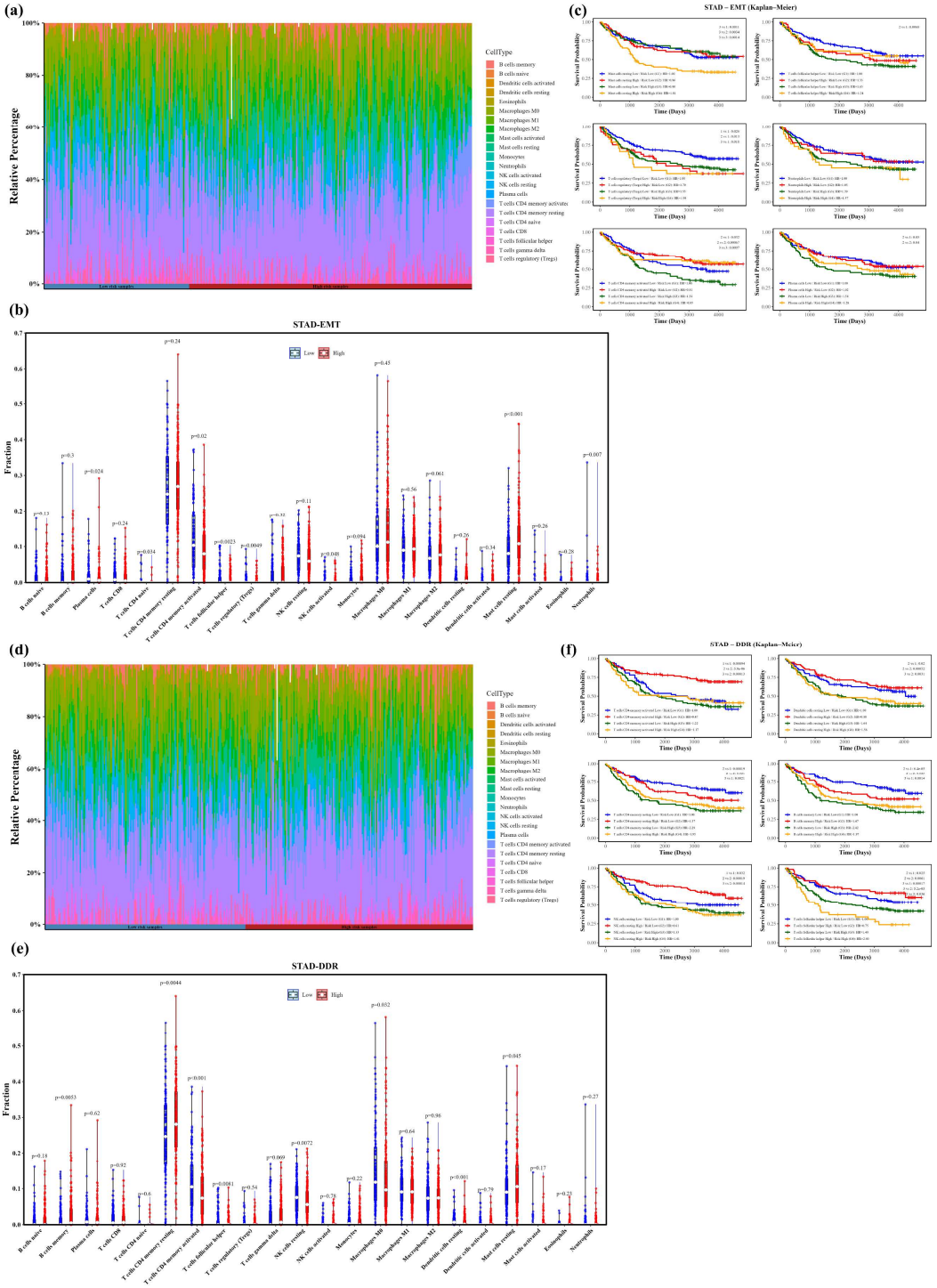


Figure S15. Immune infiltration analysis of prognostic EMT- and DDR gene signatures in the STAD-GSE84433 validation cohort. The bar chart shows the infiltration of 22 types of immune cells in samples from high- and low-risk groups for (a) EMT- and (d) DDR-related gene signatures in the validation STAD-GSE84433 cohort. The difference in 22 kinds of immune cells between the low-risk group and the high-risk group, as indicated by (b) EMT- and (e)

DDR-related gene signatures, was statistically significant ($p < 0.05$). Combined Kaplan–Meier analysis of top significant immune cells and risk scores in (c) EMT- and (f) DDR-related gene signature.

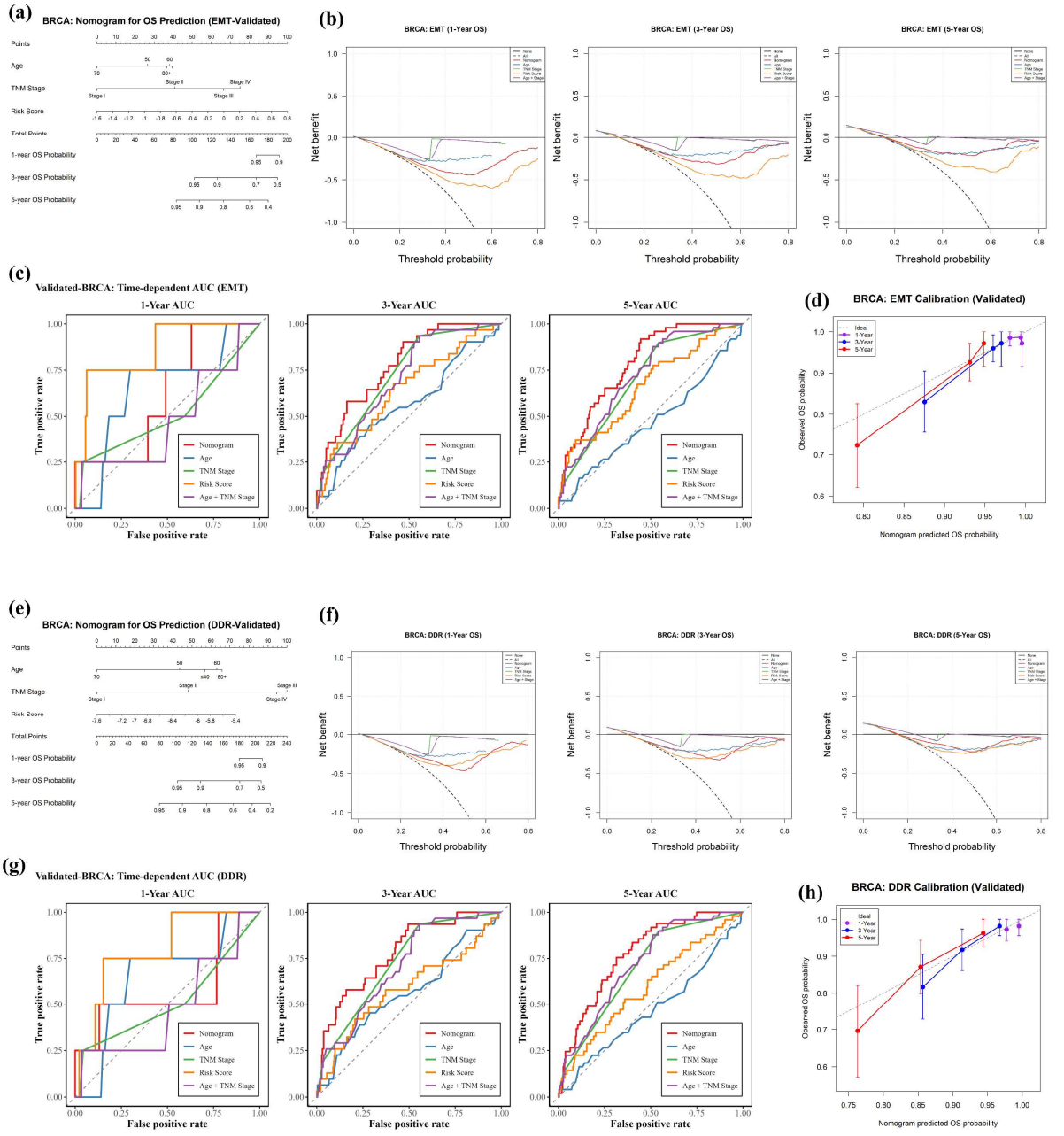


Figure S16. Validation of Nomogram model for prognostic EMT and DDR gene signatures in BRCA-GSE20685 cohort. Construction and validation of a nomogram model for OS prediction for (a) EMT- and (e) DDR-related prognostic gene signatures. DCA curves of the nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction in the BRCA-GSE20685 cohort for (b) EMT and (f) DDR. Time-dependent ROC curves of the

nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction for (c) EMT- and (g) DDR-related prognostic genes. Calibration curves between predicted and actual OS prediction for 1, 3, and 5 years for (d) EMT and (h) DDR.

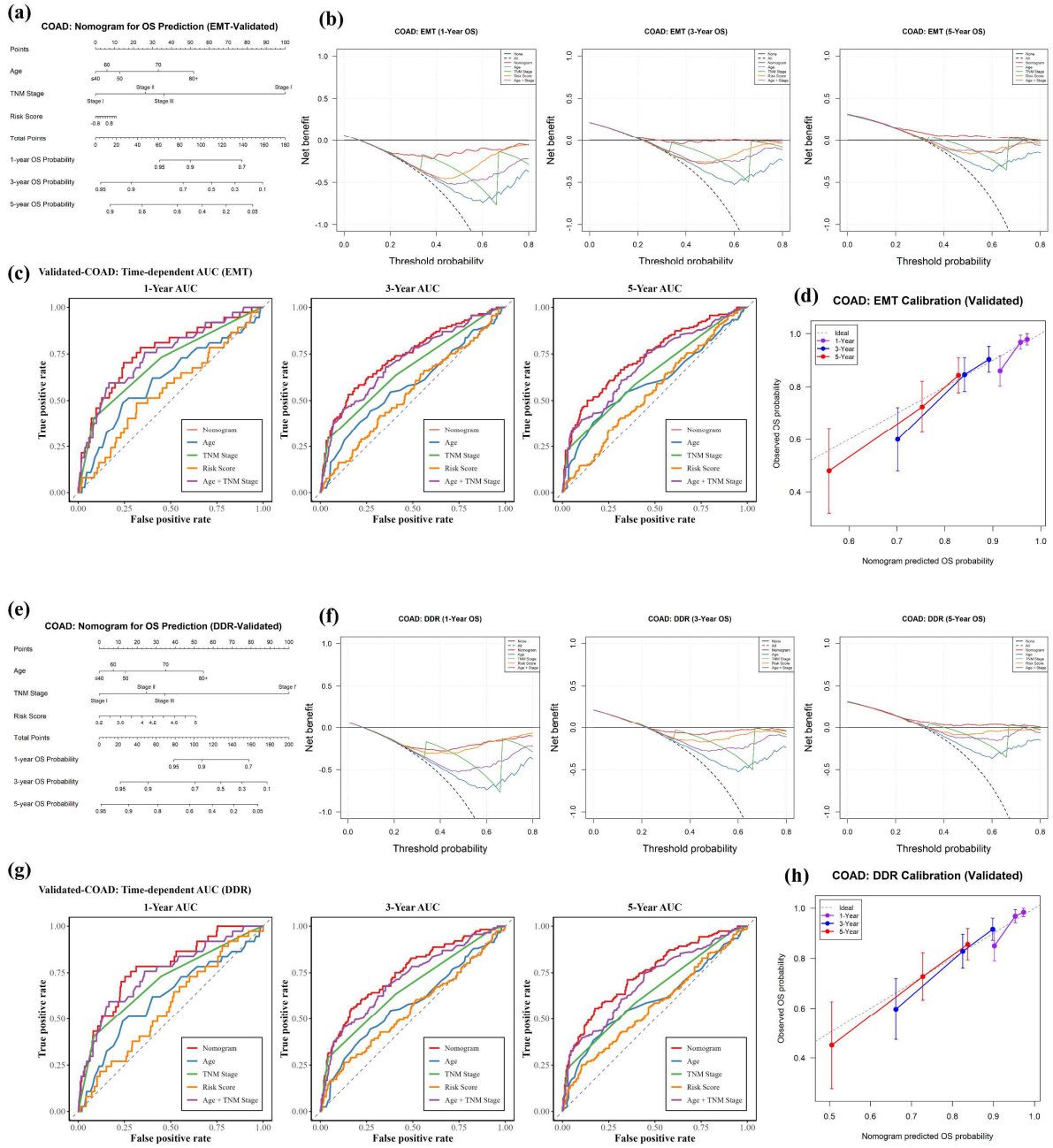


Figure S17. Validation of Nomogram model for prognostic EMT and DDR gene signatures in COAD-GSE39582 cohort. Construction and validation of a nomogram model for OS prediction for (a) EMT- and (e) DDR-related prognostic gene signatures. DCA curves of the nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction in the COAD-GSE39582 cohort for (b) EMT and (f) DDR. Time-dependent ROC curves of the

nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction for (c) EMT- and (g) DDR-related prognostic genes. Calibration curves between predicted and actual OS prediction for 1, 3, and 5 years for (d) EMT and (h) DDR.

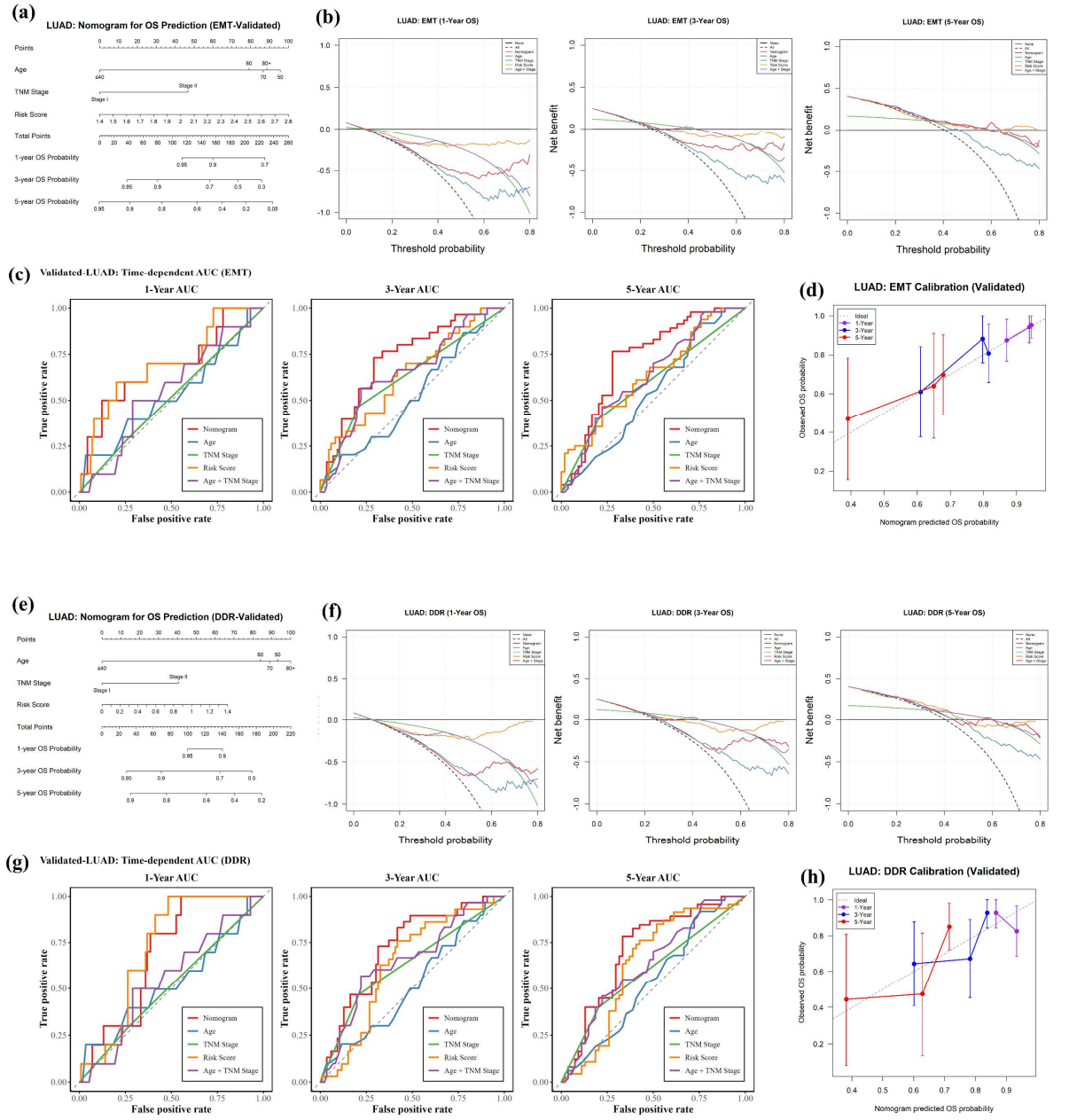


Figure S18. Validation of Nomogram model for prognostic EMT and DDR gene signatures in LUAD-GSE50081 cohort. Construction and validation of a nomogram model for OS prediction for (a) EMT- and (e) DDR-related prognostic gene signatures. DCA curves of the nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction in the LUAD-GSE50081 cohort for (b) EMT and (f) DDR. Time-dependent ROC curves of the

nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction for (c) EMT- and (g) DDR-related prognostic genes. Calibration curves between predicted and actual OS prediction for 1, 3, and 5 years for (d) EMT and (h) DDR.

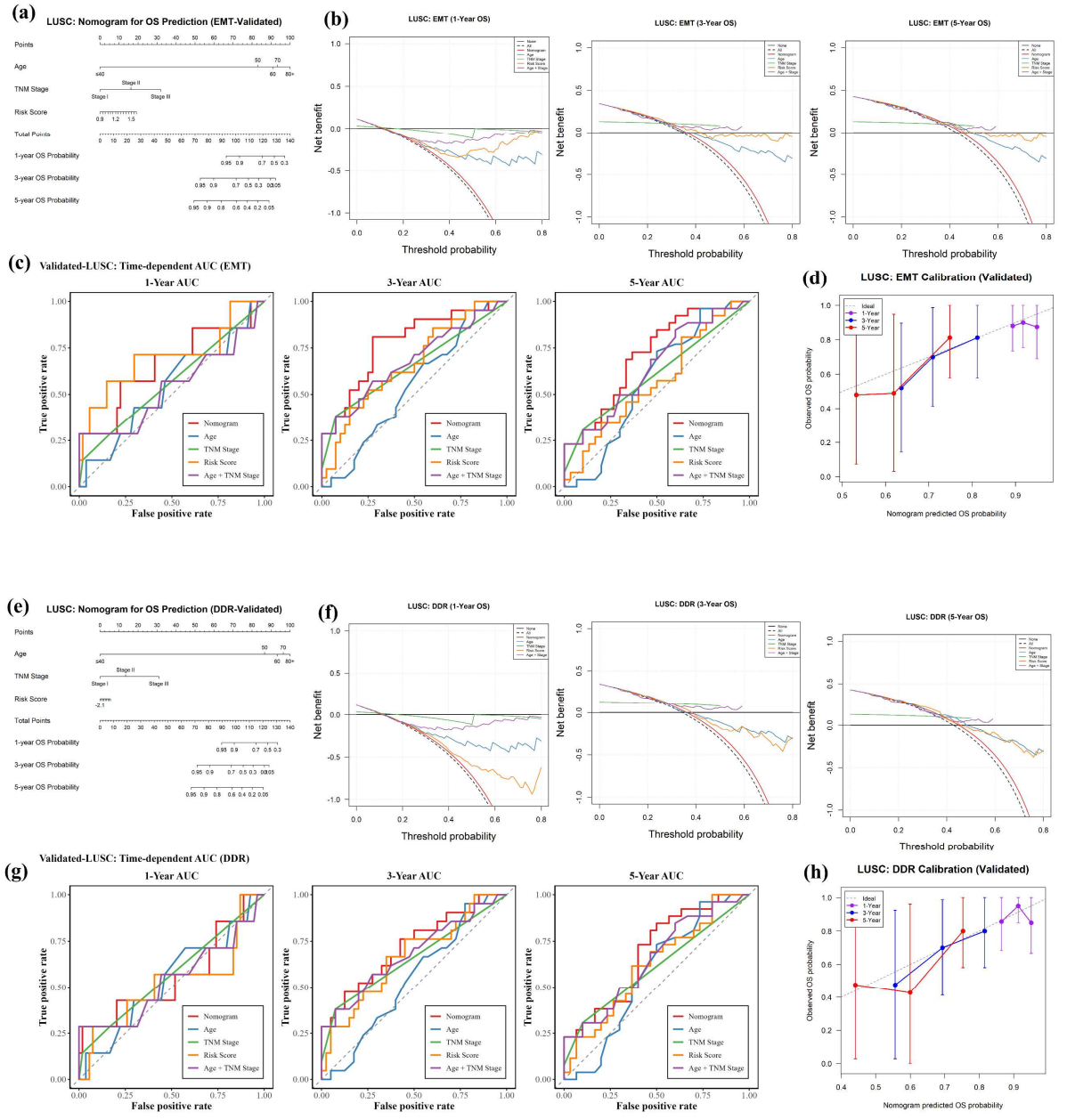


Figure S19. Validation of Nomogram model for prognostic EMT and DDR gene signatures in LUSC-GSE30219 cohort. Construction and validation of a nomogram model for OS prediction for (a) EMT- and (e) DDR-related prognostic gene signatures. DCA curves of the nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction in the LUSC-GSE30219 cohort for (b) EMT and (f) DDR. Time-dependent ROC curves of the nomogram, age, TNM stage, risk score, and age + TNM stage for 1, 3, and 5-year OS prediction

for (c) EMT- and (g) DDR-related prognostic genes. Calibration curves between predicted and actual OS prediction for 1, 3, and 5 years for (d) EMT and (h) DDR.

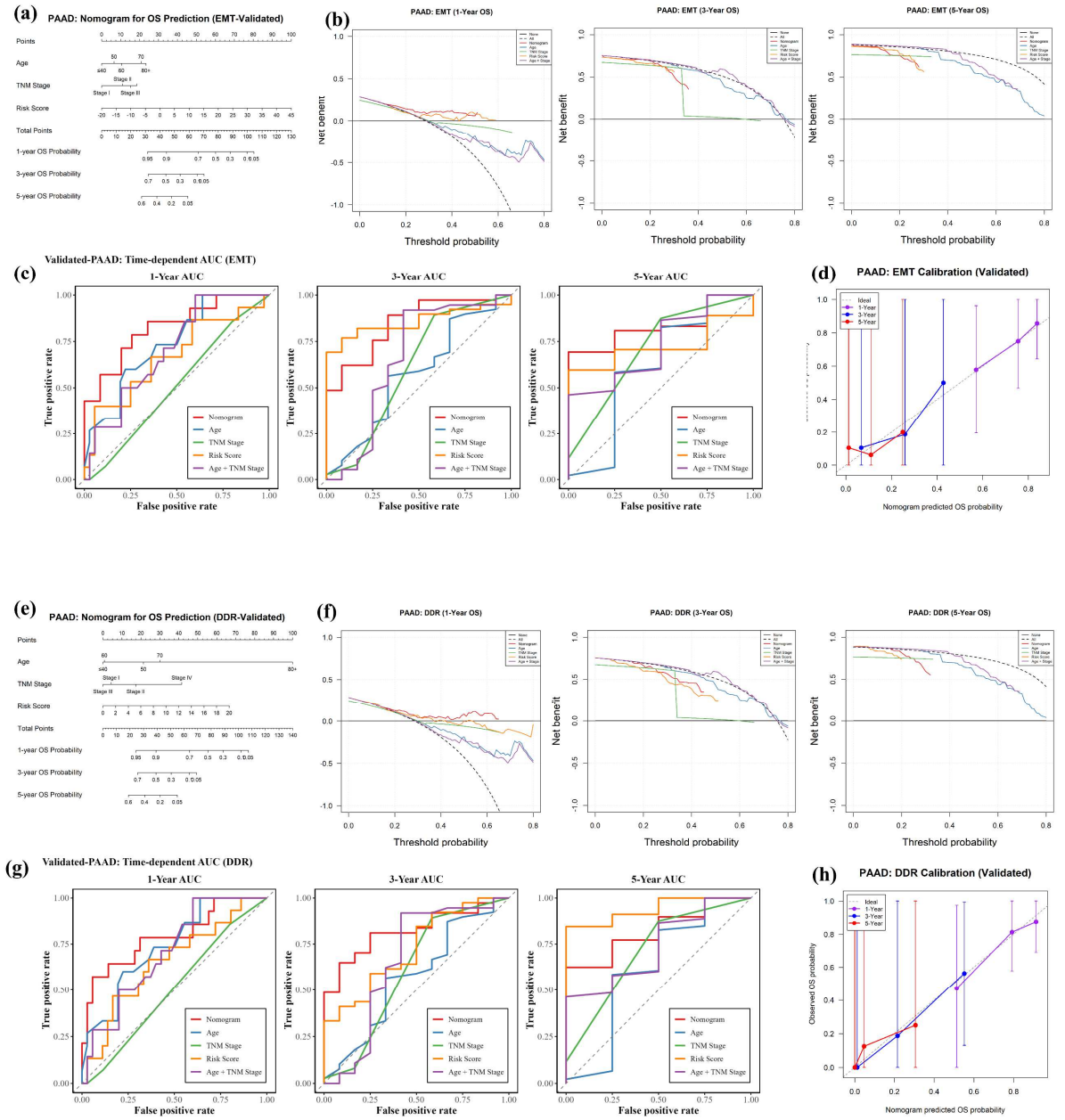


Figure S20. Validation of Nomogram model for prognostic EMT and DDR gene signatures in PAAD-GSE79668 cohort. Construction and validation of a nomogram model for OS prediction for (a) EMT- and (e) DDR-related prognostic gene signatures. DCA curves of the nomogram, age, TNM stage, risk score and age + TNM stage for 1, 3, and 5-year OS prediction in the PAAD-GSE79668 cohort for (b) EMT and (f) DDR. Time-dependent ROC curves of the nomogram, age, TNM stage, risk score and age + TNM stage for 1, 3, and 5-year OS prediction

for (c) EMT- and (g) DDR-related prognostic genes. Calibration curves between predicted and actual OS prediction for 1, 3, and 5-year for (d) EMT and (h) DDR.

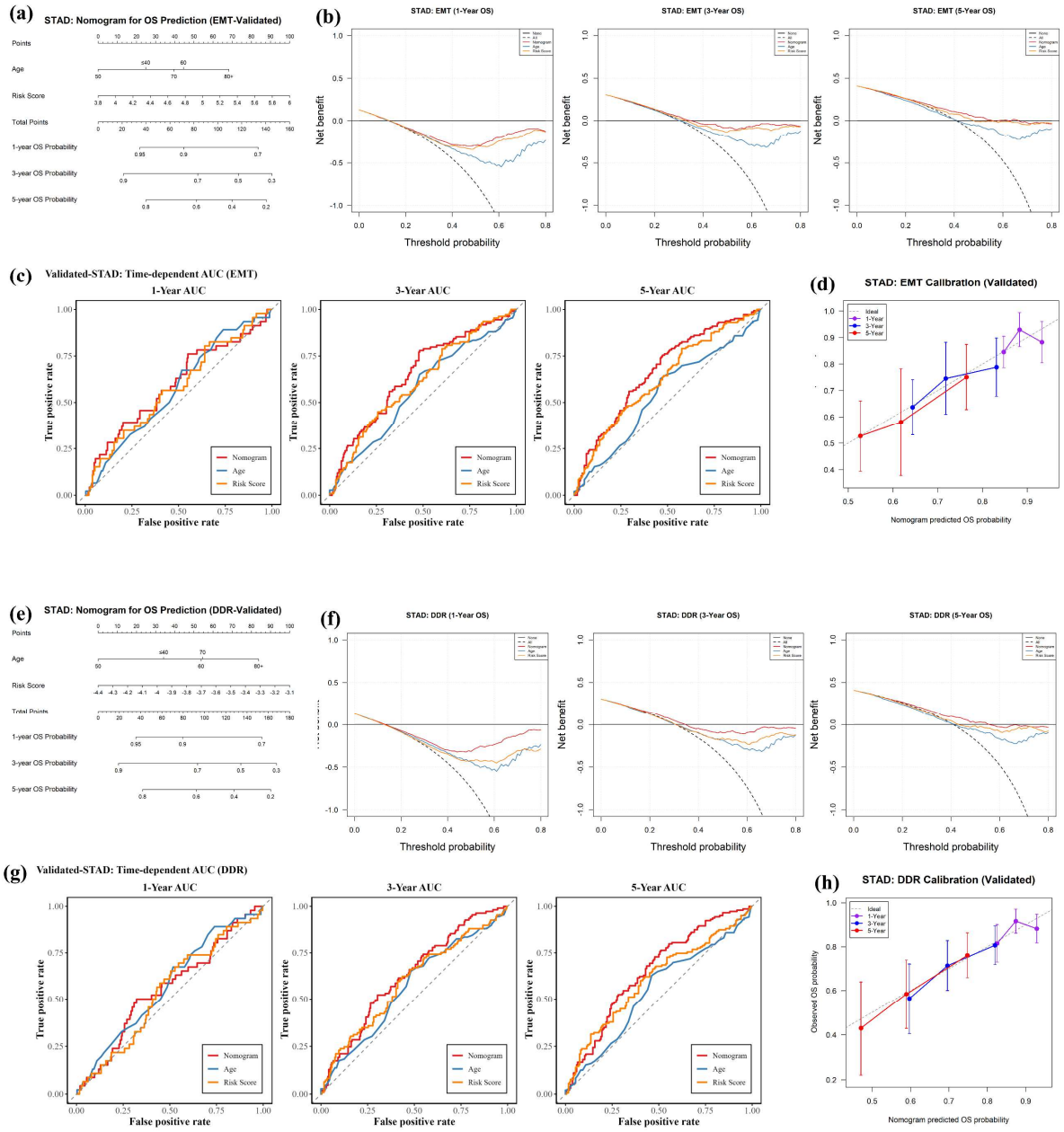


Figure S21. Validation of Nomogram model for prognostic EMT and DDR gene signatures in STAD-GSE84433 cohort. Construction and validation of a nomogram model for OS prediction for (a) EMT- and (e) DDR-related prognostic gene signatures. DCA curves of the nomogram, age, TNM stage, risk score and age + TNM stage for 1, 3, and 5-year OS prediction in the STAD-GSE84433 cohort for (b) EMT and (f) DDR. Time-dependent ROC curves of the nomogram, age, TNM stage, risk score and age + TNM stage for 1, 3, and 5-year OS prediction

for (c) EMT- and (g) DDR-related prognostic genes. Calibration curves between predicted and actual OS prediction for 1, 3, and 5-year for (d) EMT and (h) DDR.

Appendix D — Supplementary Files 4

Table S1. The Akaike Information Criterion (AIC) values individually and in combination in the validation cohort of BRCA-GSE20685 for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.47 (0.83-2.61)	0.192	1.35 (0.76-2.40)	0.309	1.37 (0.78-2.38)	0.274	–	–	1.34 (0.77-2.35)	0.3	–	–	–	–
Age 60	1.41 (0.62-3.23)	0.413	1.57 (0.69-3.59)	0.287	1.19 (0.53-2.71)	0.673	–	–	1.38 (0.61-3.14)	0.437	–	–	–	–
Age 70	0.45 (0.06-3.53)	0.448	0.40 (0.05-3.16)	0.385	0.52 (0.07-3.91)	0.524	–	–	0.57 (0.08-4.34)	0.591	–	–	–	–
Age 80+	1.35 (0.22-8.21)	0.745	2.06 (0.34-12.35)	0.429	3.27 (0.72-14.75)	0.124	–	–	6.66 (1.52-29.15)	0.0119	–	–	–	–
TNM Stage II	3.40 (1.90-6.05)	3.42E-05	3.27 (1.85-5.81)	4.99 E-05	–	–	3.20 (1.81-5.68)	6.57 E-05	–	–	3.14 (1.77-5.55)	8.54 E-05	–	–
TNM Stage III	7.33 (2.14-25.15)	0.00153	9.55 (2.84-32.12)	2.68 E-04	–	–	6.61 (2.34-18.68)	3.61 E-04	–	–	9.37 (3.40-25.84)	1.54 E-05	–	–
TNM Stage IV	9.47 (3.81-23.55)	1.33E-06	11.32 (4.57-28.07)	1.62 E-07	–	–	10.20 (4.13-25.19)	4.71 E-07	–	–	11.91 (4.84-29.31)	7.09 E-08	–	–
Risk score	3.48 (1.80-6.70)	2.03E-04	–	–	3.72 (1.96-7.06)	6.11 E-05	3.49 (1.81-6.73)	1.85 E-04	–	–	–	–	3.90 (2.06-7.39)	2.81 E-05
AIC	**862.4665**	–	875.22118	–	889.14117	–	**857.99255**	–	904.33645	–	871.10434	–	884.8385	–
DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR| Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.39 (0.78-2.47)	0.263	1.35 (0.76-2.40)	0.309	1.38 (0.79-2.40)	0.262	–	–	1.34 (0.77-2.35)	0.3	–	–	–	–
Age 60	1.62 (0.71-3.71)	0.253	1.57 (0.69-3.59)	0.287	1.37 (0.60-3.11)	0.45	–	–	1.38 (0.61-3.14)	0.437	–	–	–	–
Age 70	0.34 (0.04-2.76)	0.313	0.40 (0.05-3.16)	0.385	0.52 (0.07-3.92)	0.525	–	–	0.57 (0.08-4.34)	0.591	–	–	–	–
Age 80+	1.74 (0.30-10.15)	0.54	2.06 (0.34-12.35)	0.429	6.08 (1.38-26.70)	0.0168	–	–	6.66 (1.52-29.15)	0.0119	–	–	–	–

TNM Stage II	3.27 (1.84-5.81)	5.10E-05	3.27 (1.85-5.81)	4.99 E-05	–	–	3.12 (1.77-5.53)	9.05 E-05	–	–	3.14 (1.77-5.55)	8.54 E-05	–	–
TNM Stage III	11.76 (3.49-39.63)	7.05E-05	9.55 (2.84-32.12)	2.68 E-04	–	–	10.03 (3.63-27.71)	8.74 E-06	–	–	9.37 (3.40-25.84)	1.54 E-05	–	–
TNM Stage IV	10.28 (4.13-25.60)	5.52E-07	11.32 (4.57-28.07)	1.62 E-07	–	–	10.95 (4.44-27.01)	2.06 E-07	–	–	11.91 (4.84-29.31)	7.09 E-08	–	–
Risk score	2.26 (1.16-4.41)	0.0163	–	–	2.15 (1.14-4.06)	0.0179	2.12 (1.09-4.12)	0.0261	–	–	–	–	2.13 (1.13-4.03)	0.0193
AIC	**871.50691**	–	875.22118	–	900.78263	–	**868.19204**	–	904.33645	–	871.10434	–	898.34224	–

Table S2. The Akaike Information Criterion (AIC) values individually and in combination in the validation cohort of COAD-GSE39582 for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.73 (0.39-1.36)	0.321	0.73 (0.39-1.36)	0.319	0.82 (0.44-1.52)	0.523	–	–	0.81 (0.44-1.52)	0.518	–	–	–	–
Age 60	0.85 (0.51-1.40)	0.516	0.85 (0.51-1.41)	0.531	0.82 (0.50-1.36)	0.446	–	–	0.83 (0.50-1.36)	0.453	–	–	–	–
Age 70	1.66 (1.03-2.65)	0.0358	1.66 (1.04-2.66)	0.0342	1.36 (0.85-2.17)	0.194	–	–	1.36 (0.86-2.17)	0.191	–	–	–	–
Age 80+	2.62 (1.58-4.35)	1.87 E-04	2.64 (1.59-4.37)	1.71 E-04	2.36 (1.43-3.90)	7.95E-04	–	–	2.40 (1.45-3.96)	6.36 E-04	–	–	–	–
TNM Stage II	1.92 (0.77-4.75)	0.159	1.93 (0.78-4.77)	0.156	–	–	2.14 (0.86-5.29)	0.0998	–	–	2.17 (0.88-5.36)	0.093	–	–
TNM Stage III	2.45 (0.99-6.11)	0.0539	2.48 (1.00-6.17)	0.051	–	–	2.58 (1.04-6.42)	0.0411	–	–	2.63 (1.06-6.52)	0.037	–	–
TNM Stage IV	11.88 (4.63-30.47)	2.58 E-07	12.13 (4.75-30.98)	1.83 E-07	–	–	10.62 (4.15-27.19)	8.50 E-07	–	–	10.99 (4.32-27.95)	4.89 E-07	–	–
Risk score	1.12 (0.69-1.83)	0.652	–	–	1.43 (0.88-2.32)	0.149	1.16 (0.71-1.88)	0.552	–	–	–	–	1.50 (0.93-2.42)	0.1
AIC	**2170.7281**	–	2168.9321	–	2246.5522	–	2198.9304	–	2246.6641	–	2197.2848	–	2265.5503	–
DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR| Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.71 (0.38-1.33)	0.286	0.73 (0.39-1.36)	0.319	0.85 (0.45-1.58)	0.601	–	–	0.81 (0.44-1.52)	0.518	–	–	–	–
Age 60	0.85 (0.51-1.41)	0.535	0.85 (0.51-1.41)	0.531	0.87 (0.52-1.43)	0.573	–	–	0.83 (0.50-1.36)	0.453	–	–	–	–
Age 70	1.68 (1.05-2.69)	0.0299	1.66 (1.04-2.66)	0.0342	1.40 (0.88-2.24)	0.155	–	–	1.36 (0.86-2.17)	0.191	–	–	–	–
Age 80+	2.74 (1.65-4.55)	9.43 E-05	2.64 (1.59-4.37)	1.71 E-04	2.47 (1.50-4.09)	4.15E-04	–	–	2.40 (1.45-3.96)	6.36 E-04	–	–	–	–
TNM Stage II	1.84 (0.74-4.57)	0.187	1.93 (0.78-4.77)	0.156	–	–	2.09 (0.84-5.15)	0.111	–	–	2.17 (0.88-5.36)	0.093	–	–

TNM Stage III	2.33 (0.94-5.82)	0.0691	2.48 (1.00-6.17)	0.051	–	–	2.51 (1.01-6.23)	0.0475	–	–	2.63 (1.06-6.52)	0.037	–	–
TNM Stage IV	11.69 (4.57-29.92)	2.91 E-07	12.13 (4.75-30.98)	1.83 E-07	–	–	10.22 (4.00-26.09)	1.16 E-06	–	–	10.99 (4.32-27.95)	4.89 E-07	–	–
Risk score	2.00 (1.17-3.44)	0.0119	–	–	2.15 (1.21-3.82)	0.00922	1.77 (1.04-3.02)	0.0362	–	–	–	–	2.17 (1.24-3.80)	0.00697
AIC	**2164.7746**	–	2168.9321	–	2241.9953	–	2194.9978	–	2246.6641	–	2197.2848	–	2261.1854	–

Table S3. The Akaike Information Criterion (AIC) values individually and in combination in the validation cohort of LUAD-GSE50081 for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.19 (0.02-1.67)	0.135	0.20 (0.02-1.71)	0.141	0.17 (0.02-1.49)	0.111	–	–	0.18 (0.02-1.53)	0.116	–	–	–	–
Age 60	0.75 (0.28-2.03)	0.578	0.83 (0.31-2.22)	0.708	0.76 (0.28-2.03)	0.578	–	–	0.81 (0.30-2.18)	0.68	–	–	–	–
Age 70	0.86 (0.32-2.28)	0.758	0.92 (0.35-2.44)	0.867	0.81 (0.31-2.15)	0.674	–	–	0.89 (0.34-2.36)	0.816	–	–	–	–
Age 80+	0.88 (0.28-2.81)	0.835	0.97 (0.31-3.08)	0.957	0.96 (0.30-3.04)	0.944	–	–	1.07 (0.34-3.40)	0.902	–	–	–	–
TNM Stage II	2.22 (1.27-3.86)	0.00484	2.42 (1.39-4.21)	0.00178	–	–	2.29 (1.32-3.96)	0.00312	–	–	2.51 (1.45-4.36)	0.00101	–	–
Risk score	3.37 (1.40-8.08)	0.00649	–	–	3.60 (1.56-8.31)	0.00268	3.46 (1.43-8.35)	0.00581	–	–	–	–	3.73 (1.61-8.64)	0.00218
AIC	**461.01368**	–	466.34108	–	466.45777	–	**456.83283**	–	473.41523	–	462.33187	–	462.97994	–
DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR| Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.19 (0.02-1.62)	0.128	0.20 (0.02-1.71)	0.141	0.16 (0.02-1.38)	0.0954	–	–	0.18 (0.02-1.53)	0.116	–	–	–	–
Age 60	0.85 (0.32-2.28)	0.748	0.83 (0.31-2.22)	0.708	0.82 (0.30-2.18)	0.685	–	–	0.81 (0.30-2.18)	0.68	–	–	–	–
Age 70	0.93 (0.35-2.46)	0.882	0.92 (0.35-2.44)	0.867	0.89 (0.34-2.36)	0.815	–	–	0.89 (0.34-2.36)	0.816	–	–	–	–
Age 80+	1.13 (0.35-3.64)	0.837	0.97 (0.31-3.08)	0.957	1.30 (0.41-4.14)	0.662	–	–	1.07 (0.34-3.40)	0.902	–	–	–	–
TNM Stage II	2.06 (1.14-3.73)	0.0169	2.42 (1.39-4.21)	0.00178	–	–	2.24 (1.26-4.00)	0.00599	–	–	2.51 (1.45-4.36)	0.00101	–	–
Risk score	2.33 (0.80-6.77)	0.119	–	–	3.68 (1.33-10.18)	0.0121	2.01 (0.71-5.68)	0.187	–	–	–	–	3.05 (1.12-8.33)	0.0297
AIC	**465.9691**	–	466.34108	–	469.41188	–	**462.63451**	–	473.41523	–	462.33187	–	467.76859	–

Table S4. The Akaike Information Criterion (AIC) values individually and in combination in the validation cohort of LUSC-GSE30219 for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.00 (0.00-Inf)	0.997	0.00 (0.00-Inf)	0.997	0.00 (0.00-Inf)	0.997			0.00 (0.00-Inf)	0.997				
Age 60	1.79 (0.82-3.89)	0.143	1.65 (0.76-3.61)	0.207	1.81 (0.83-3.91)	0.134			1.67 (0.77-3.63)	0.192				
Age 70	2.15 (0.89-5.19)	0.0903	2.17 (0.90-5.26)	0.0856	1.67 (0.71-3.94)	0.242			1.72 (0.73-4.07)	0.217				
Age 80+	3.47 (0.92-13.11)	0.0669	2.89 (0.77-10.87)	0.117	2.68 (0.72-9.97)	0.14			2.28 (0.62-8.43)	0.215				
TNM Stage II	3.31 (1.42-7.72)	0.00551	2.98 (1.30-6.81)	0.00967			2.31 (1.03-5.22)	0.0433			1.90 (0.87-4.13)	0.108		
TNM Stage III	10.45 (2.15-50.77)	0.00363	10.80 (2.23-52.36)	0.00315			7.20 (1.57-32.99)	0.011			7.43 (1.63-33.91)	0.00965		
Risk score	7.48 (0.71-78.68)	0.0936			5.47 (0.58-52.02)	0.139	11.33 (0.93-137.89)	0.0569					6.95 (0.66-72.94)	0.106
AIC	**284.82873**		285.66833		291.67694		290.30977		291.87237		292.04327		293.52497	
DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR| Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.00 (0.00-Inf)	0.997	0.00 (0.00-Inf)	0.997	0.00 (0.00-Inf)	0.997	–	–	0.00 (0.00-Inf)	0.997	–	–	–	–
Age 60	1.66 (0.76-3.63)	0.205	1.65 (0.76-3.61)	0.207	1.66 (0.76-3.60)	0.201	–	–	1.67 (0.77-3.63)	0.192	–	–	–	–
Age 70	2.10 (0.86-5.14)	0.103	2.17 (0.90-5.26)	0.0856	1.63 (0.68-3.88)	0.269	–	–	1.72 (0.73-4.07)	0.217	–	–	–	–
Age 80+	2.77 (0.73-10.51)	0.134	2.89 (0.77-10.87)	0.117	2.13 (0.57-7.92)	0.258	–	–	2.28 (0.62-8.43)	0.215	–	–	–	–
TNM Stage II	2.79 (1.18-6.57)	0.0189	2.98 (1.30-6.81)	0.00967	–	–	1.79 (0.82-3.91)	0.146	–	–	1.90 (0.87-4.13)	0.108	–	–
TNM Stage III	10.76 (2.22-52.19)	0.0032	10.80 (2.23-52.36)	0.00315	–	–	7.41 (1.63-33.77)	0.00962	–	–	7.43 (1.63-33.91)	0.00965	–	–
Risk score	1.69 (0.27-10.76)	0.576	–	–	2.48 (0.43-14.16)	0.307	4.23 (0.68-26.25)	0.121	–	–	–	–	4.47 (0.76-26.25)	0.0977

AIC	**287.351 38**	–	285.668 33	–	292.787 9	–	291.534 05	–	291.872 37	–	292.043 27	–	293.315 08	–

Table S5. The Akaike Information Criterion (AIC) values individually and in combination in the validation cohort of PAAD-GSE79668 for EMT- and DDR-related prognostic signatures.

EMT-related signature														
Clinical parameter	**EMT | Nomogram COX PH**		**EMT | Age + TNM stage COX PH**		**EMT | Age + Risk score COX PH**		**EMT | TNM stage + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | TNM stage only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.68 (0.18-2.53)	0.562	0.43 (0.13-1.48)	0.181	0.35 (0.12-1.07)	0.0659	–	–	0.46 (0.16-1.34)	0.154	–	–	–	–
Age 60	1.23 (0.43-3.53)	0.698	0.69 (0.28-1.70)	0.417	0.90 (0.35-2.33)	0.835	–	–	0.61 (0.26-1.46)	0.268	–	–	–	–
Age 70	2.18 (0.77-6.18)	0.142	1.37 (0.55-3.46)	0.501	1.53 (0.60-3.90)	0.374	–	–	1.00 (0.42-2.39)	1	–	–	–	–
Age 80+	96.93 (6.37-1476.16)	9.94E-04	18.90 (1.59-225.28)	0.0201	56.95 (4.30-753.69)	0.00216	–	–	19.22 (1.62-228.30)	0.0192	–	–	–	–
TNM Stage II	1.89 (0.74-4.86)	0.186	2.30 (0.91-5.80)	0.077	–	–	1.67 (0.67-4.19)	0.271	–	–	2.27 (0.93-5.53)	0.071	–	–
TNM Stage III	2.37 (0.60-9.42)	0.221	1.31 (0.34-4.99)	0.693	–	–	3.02 (0.82-11.13)	0.0972	–	–	2.05 (0.57-7.43)	0.272	–	–
TNM Stage IV	2.93 (0.27-32.11)	0.38	5.41 (0.51-57.71)	0.162	–	–	1.46 (0.17-12.83)	0.734	–	–	2.70 (0.31-23.27)	0.365	–	–
Risk score	1.09 (1.05-1.13)	2.43E-05	–	–	1.07 (1.03-1.10)	5.75E-05	1.08 (1.04-1.11)	2.26 E-05	–	–	–	–	1.05 (1.02-1.07)	8.71E-04
AIC	**259.7171**	–	273.38923	–	275.74403	–	262.53496	–	287.98536	–	274.44817	–	261.46378	–
DDR-related signature														
Clinical parameter	**DDR | Nomogram COX PH**		**DDR| Age + TNM stage COX PH**		**DDR | Age + Risk score COX PH**		**DDR | TNM stage + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | TNM stage only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	0.35 (0.10-1.23)	0.101	0.43 (0.13-1.48)	0.181	0.46 (0.15-1.37)	0.164	–	–	0.46 (0.16-1.34)	0.154	–	–	–	–
Age 60	0.36 (0.14-0.98)	0.046	0.69 (0.28-1.70)	0.417	0.38 (0.15-0.97)	0.0426	–	–	0.61 (0.26-1.46)	0.268	–	–	–	–
Age 70	1.53 (0.59-3.98)	0.38	1.37 (0.55-3.46)	0.501	1.10 (0.45-2.70)	0.833	–	–	1.00 (0.42-2.39)	1	–	–	–	–
Age 80+	47.49 (3.72-607.05)	0.00298	18.90 (1.59-225.28)	0.0201	44.62 (3.55-560.78)	0.00327	–	–	19.22 (1.62-228.30)	0.0192	–	–	–	–
TNM Stage II	1.89 (0.71-5.05)	0.205	2.30 (0.91-5.80)	0.077	–	–	1.96 (0.78-4.88)	0.15	–	–	2.27 (0.93-5.53)	0.071	–	–

TNM Stage III	0.81 (0.20-3.27)	0.766	1.31 (0.34-4.99)	0.693	–	–	1.97 (0.54-7.17)	0.306	–	–	2.05 (0.57-7.43)	0.272	–	–
TNM Stage IV	6.23 (0.54-71.44)	0.142	5.41 (0.51-57.71)	0.162	–	–	2.93 (0.34-25.48)	0.33	–	–	2.70 (0.31-23.27)	0.365	–	–
Risk score	1.18 (1.08-1.28)	2.09E-04	–	–	1.17 (1.08-1.27)	2.15E-04	1.10 (1.02-1.19)	0.0101	–	–	–	–	1.11 (1.03-1.19)	0.00476
AIC	**261.46378**	–	273.38923	–	276.6128	–	270.03188	–	287.98536	–	274.44817	–	282.54838	–

Table S6. The Akaike Information Criterion (AIC) values individually and in combination in the validation cohort of STAD-GSE84433 for EMT- and DDR-related prognostic signatures.

EMT-related signature						
Clinical parameter	**EMT | Age + Risk score COX PH**		**EMT | Age only COX PH**		**EMT | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.63 (0.96-2.75)	0.0682	1.79 (1.06-3.02)	0.0289	–	–
Age 60	2.40 (1.52-3.79)	1.73E-04	2.38 (1.51-3.76)	1.95E-04	–	–
Age 70	2.18 (1.25-3.80)	0.00627	2.40 (1.38-4.17)	0.00203	–	–
Age 80+	3.81 (1.81-8.03)	4.25E-04	4.46 (2.13-9.33)	7.42E-05	–	–
Risk score	2.44 (1.54-3.86)	1.36E-04	–	–	2.58 (1.62-4.09)	5.99E-05
AIC	**1890.8224**	–	1902.9641	–	1904.2718	–
DDR-related signature						
Clinical parameter	**DDR | Age + Risk score COX PH**		**DDR | Age only COX PH**		**DDR | Risk score only COX PH**	
	HR (95% CI)	***P* value**	**HR (95% CI)**	***P* value**	**HR (95% CI)**	***P* value**
Age 40	1.75 (1.04-2.96)	0.0357	1.79 (1.06-3.02)	0.0289	–	–
Age 60	2.42 (1.53-3.82)	1.54E-04	2.38 (1.51-3.76)	1.95E-04	–	–
Age 70	2.45 (1.40-4.26)	0.00159	2.40 (1.38-4.17)	0.00203	–	–
Age 80+	3.94 (1.87-8.30)	3.00E-04	4.46 (2.13-9.33)	7.42E-05	–	–
Risk score	3.53 (1.71-7.28)	6.29E-04	–	–	3.74 (1.79-7.82)	4.52E-04
AIC	**1893.3601**	–	1902.9641	–	1907.8359	–