A Molecule Based Nomogram Optimized the Prediction of Relapse in Stage I NSCLC

Rongrong Bian
Liuhe People's Hospital
Guorong Zhu
yancheng 3rd people's hospital
Feng Zhao
Taixing People's hospital
Rui Chen
Taixing People's Hospital
Wengji Xia
Xuzhou Central Hospital
lin wang (linwang1988@njmu.edu.cn)
Jiangsu Province Geriatric Institute: Jiangsu Province Geriatric Hospital https://orcid.org/0000-0001-6971-4894
Qiang Chen
Xuzhou Central Hospital

Research Article

Keywords: Recurrence, stage I non-small cell lung cancer, survival, nomogram

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-437687/v1

Abstract

Background:

Early-stage non-small cell lung cancer (NSCLC) is being diagnosed increasingly, and in 30% of diagnosed patients, recurrence will develop within 5 years. Thus, it is urgent to identify recurrence-related markers in order to optimize the management of patient-tailored therapeutics. The aim of the study was to develop a feasible tool to optimize the recurrence prediction of stage I NSCLC.

Methods:

The eligible datasets were downloaded from TCGA and GEO. In discovery phase, two algorithms, Least Absolute Shrinkage and Selector Operation and Support Vector Machine-Recursive Feature Elimination, were used to identify candidate genes. Recurrence associated signature was developed by penalized cox regression. The nomogram was constructed and further tested via two independent cohorts.

Results:

In this retrospective study, 14 eligible datasets and 7 published signatures were included. In discovery phase, 42 significant genes were highlighted as candidate predictors by two algorithms. A 13-gene based signature was generated by penalized cox regression categorized training cohort into high-risk and low-risk subgroups (HR = 8.873, 95% CI:4.228–18.480 $P < 0.001$). Furthermore, a nomogram integrating the recurrence related signature, age, and histology was developed to predict the recurrence-free survival in the training cohort, which performed well in the two external validation cohorts (concordance index: 0.737, 95%CI:0.732–0.742, $P < 0.001$; 0.666, 95%CI: 0.650–0.682, $P < 0.001$; 0.651, 95%CI:0.637–0.665, $P < 0.001$ respectively).

Conclusions:

The proposed nomogram is a promising tool for estimating recurrence free survival in stage I NSCLC, which might have tremendous value in guiding adjuvant therapy. Prospective studies are needed to test the clinical utility of the nomogram in individualized management of stage I NSCLC.

Introduction

With the adoption of low-dose spiral computed tomography screening for the high-risk individuals, the proportion of stage I non-small cell lung cancer (NSCLC) rise sharply. Surgical resection is the curative treatment for stage I NSCLC. However, local relapse and metastasis impede the therapeutic effect, and approximately 30% patients will suffer recurrence within 5 years of diagnosis(1, 2). Postoperative adjuvant therapy is a valid approach to prevent the relapse and metastasis, but the effect is still unclear

in stage I NSCLC(3, 4). Several large randomized studies failed to show a survival benefit from stage I NSCLC patients who received adjuvant systemic treatment after surgery due to side effects of therapy(5). Thus, in view of the high rate of relapse and the lack of useful biomarkers, a feasible prediction model is needed to predict the relapse free survival of stage I NSCLC patients and identify the high risk patients who may benefit from postoperative adjuvant therapy.

Plenty of studies have been published to identify clinical risk factors for survival or relapse(6). However, it is not enough to predict the relapse by only considering clinical risk factors, because stage I NSCLC is a heterogeneous disease, which comprising different subgroups with distinct molecular alterations(7). Expression profiling has showed great promise in getting insight of molecular mechanisms and biomarkers through analysis of thousands of genes(8, 9). And increasing molecular markers have been incorporated into clinical decision-making(10). Previous studies highlighted several relapse related signatures for stage I NSCLC, which showed molecular information and provided a more robust biomarker of relapse(11–13). Unfortunately, over fitting on small discovery data sets and lack of sufficient external validation impeded them to be widespread adopted into clinical practice. Therefore, developing and testing a relapse related signature in a large-scale study is needed, and available public gene expression data sets with relapse status brings the opportunity to identify more reliable signature for relapse. In addition, there is an increasing trend that combining molecular features and clinicopathologic features to predict the disease status or prognosis by recent investigations(14).

In this study, we aimed to develop and validate a predicting signature related to the relapse by multiple gene expression datasets, and leverage the clinical features to build a nomogram predicting the relapse of stage I NSCLC, which would improve capacity of relapse prediction and might have tremendous value in guiding the management of stage I NSCLC patients.

Methods

Study population and study design

Comprehensive search for eligible expression profile datasets related to recurrence of NSCLC were performed in Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). Datasets were filtered by the criteria such as containing stage I NSCLC patients, recurrence status, and relapse-free survival time. After primary filtration, we excluded those patients who had received neoadjuvant therapy, adjuvant chemotherapy, or other pharmaceutical therapy in each dataset. Tumors were pathologically classified as stage I according to the seventh edition of the TNM classification. After removing patients with no clinical or subtype information, demographic and clinical characteristics of the patients were recorded. Simultaneously, searching for published prognostic signatures related to recurrence were also conducted in Pubmed, Embase, and MEDLINE with the key words including "relapse NSCLC", "recurrence NSCLC", " relapse stage I NSCLC ", " recurrence stage I NSCLC ", "relapse lung cancer", "recurrence lung cancer", "relapse stage I lung cancer", and “recurrence stage I lung cancer”.

Three steps were designed to develop and test the recurrence-related nomogram for NSCLC patients with stage I. In the discovery phase, eligible datasets and published prognostic signatures were used to screen significant genes related to recurrence by two algorithms (LASSO and SVM-RFE), which obtained the candidate features for the further selection. In the training phase, patients from TCGA were included in training cohort due to the completed clinical records and respectively large sample size. Penalized Cox regression was performed to develop a recurrence signature in training cohort, which was subsequently tested in two independent cohorts. Then, a nomogram was constructed by combining recurrence signature and significant clinical features. In the validation phase, two independent GEO datasets were selected to test the performance of the nomogram. The study design was displayed by a flow chat in Fig. 1. Recurrence-free survival was defined as the time from the date of diagnosis to the date of recurrence or last follow-up. The schedule of follow-up and examination were described by the corresponding studies. Forest plot analyses were performed using R package, “meta” (https://github.com/guido-s/meta http://meta-analysis-with-r.org). A heterogeneity test for the combined HR was carried out using the I_2 statistic.

Statistical analysis

The expression profile datasets were normalized and qualified using the R package (“affy”). Gene IDs were mapped to genes using the corresponding annotation files. Multiple probes corresponding to the same gene ID were averaged to one measurement for each sample. Differential expression genes between recurrence samples and non-recurrence samples were generated by “limma” package of R software with adjusted false discovery rate less than 0.05 and fold-change greater than 2. Recurrence related genes were identified by cox univariable analysis with significant *P* values. Significant genes from each dataset and genes from these prognostic signatures were submitted to the Least Absolute Shrinkage and Selector Operation (LASSO) algorithm in order to select candidate features with optimal lambda value, which was defined by 10 fold cross-validation (R package, glmnet). Meanwhile, another method called Support Vector Machine-Recursive Feature Elimination (SVM-RFE) was used for selection (R package, e1071). Then, a recurrence related signature was developed from all candidate features selected by two algorithms using penalized Cox analysis. The nomogram was developed by “rms” package of R software. The variable included into nomogram were identified by backward wise step method and the clinical significance. The concordance index (C-index) were used to evaluate the discrimination of the nomogram, and calibration plots revealed the accuracy of the nomogram. “SurvivalROC” package of R software was used to draw the receiver operating characteristic (ROC) curves. All steps were conducted in R version 3.3.3 software. The cut-off values were defined by X-tile software with the highest $\chi 2$ value in each set(15).

Results

Characteristics of patients

Retrospectively analysis of the gene expression profiles related to recurrence identified 14 eligible cohorts, including 13 microarray datasets from GEO (https://www.ncbi.nlm.nih.gov/geo/) and 1 RNA-Seq dataset from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/). A total of 1414 stage I NSCLC patients were included from 14 published cohorts retrieved by the systematic search. The characteristics of these datasets were listed in Table 1. We didn't integrate these datasets together due to different platforms and different racial groups and segregated them into three phrases. TCGA cohort with the largest sample size and complete clinical records were assigned into the training phase, and we selected another 2 individual datasets with moderate sample size and matching clinical records for independent validation, named GEO50081 and GEO30219. In addition, 7 published prognostic signatures related to recurrence were identified, which were also incorporated into the discovery phase (Table S1).

The median recurrence-free survival time of the patients in TCGA was 508 days (range, 2 to 6812 days). During follow-up, 19.9% of the patients (70 of 351) developed recurrence. The median follow-up time of the patients in the testing cohort were 1729 days and 1935 days (range, 44 and 4393 days and 30 days and 7680 days, respectively). 25% and 24.6% of the patients from two validation cohort showed relapse. In TCGA cohort, the 1- and 5-year RFS rates were 67% and 8.8%, respectively. And the 1- and 5-year OS rates were 85.5% and 43.5% in GEO50081 and 87.3% and 52.8% in GSE30219, respectively.

Construction and validation of the recurrence signature

We conducted a univariable Cox regression analysis to identify 1058 genes associated with recurrence-free survival from differential expression genes in each dataset. After analyzing the published recurrence associated signatures, we obtained 66 significant genes and incorporated them with significant genes from microarray datasets. LASSO analysis identified 36 candidate genes related to recurrence by the optimal lambda which was determined through 10-times cross validations (Fig. 2A &Fig. S1A). Simultaneously, another algorithm, named SVM-RFE, selected 13 candidate genes through ranking the features based on their weights and eliminating the feature with the lowest weight (Fig. 2B &Fig. S1B). Combining the results from two algorithms, 42 candidate genes were selected, including 3 overlapped genes (Table S2). Then, we submitted them into penalized Cox analysis, and a risk signature was built by a 13-mRNA with the corresponding coefficients in the training cohort (Fig. 2C). A cluster plot showed the expression profile of the 13 mRNAs (Fig. 2D). Among the 13-mRNAs, *DUSP4, LDOC1, NRIP3, B3GNT7*, and *CCBP2* showed positive effect in predicting the recurrence, while the rest 8 genes showed negative effect in prediction, including *PCSK6, TPSB2, HSF4, CPNE7, CAPN12, GABRE, HLF*, and *AGER* (Table S3).

After calculating the risk score of each patient by recurrence related signature, we dichotomized patients in two group at the median. Kaplan–Meier (KM) survival analysis showed distinct different survival between high risk group and low risk group (Hazard Ratio (HR): 8.837, 95% confidence interval: 4.228–18.480, $P < 0.001$) (Fig. 3A). Prediction ability of the signature was evaluated by time independent ROC curve, showing well performance (AUC:0.79 at 1-year recurrence-free survival).

The robustness of the 13-mRNA signature was further confirmed in two independent cohorts (GSE50081 and GSE30219). Survival curves were remarkably different between the high risk group and the low risk

group based on the signature, showing recurrence-free survival in 51.6% and 34.3% and 56.3% and 45%, respectively. (HR: 3.556, 95% CI: 1.587–7.966; *P* = 0.001 and HR: 2.586, 95% CI: 1.226–5.286; *P* = 0.007) (Fig. 3B&C). Time independent ROC curve was used to demonstrate the predictive capability of survival prediction in two datasets. The AUCs at the 5-year recurrence-free survival was 0.73 in GSE50081 cohort and 0.72 in GSE30219 cohort, respectively. In addition, recurrence associated signature was further validated in another three cohorts that contained the 13-mRNA expression information (GSE37745, GSE8894, and GSE31210) by a fixed effects meta-analysis model due to their limited sample size (Fig. S2). There was low heterogeneity or inconsistency across these cohorts (HR = 1.87, 95%CI:1.29–2.70, *P*< 0.001; Heterogeneity: I^2 = 0%, τ^2 = 0, *P* = 0.37), suggesting that these data are not the result of selection bias. A strong concordance was exhibited with previous results by meta-analysis, indicating that the recurrence pattern was a reliable and generalizable signature in predicting the recurrence of stage I NSCLC.

Building a nomogram

Univariable analysis revealed that histology and risk score were two significant risk factors for recurrence-free survival in training cohort. After multivariable analysis, it remained as an independent risk factor (Table 2). For the purpose of clinically application, we developed a nomgram in the training cohort to individually predict the probability of recurrence by incorporating the recurrence related signature with age and histology selected by backwards wise step method with least AIC value (Fig. 4A). The Concordance index (C-index) of the nomogram was 0.737 with 95%CI ranged from 0.731 to 0.743 (*P*< 0.001), indicating the good discrimination ability. Calibration plots of 1-year, 3-year, and 5-year illustrated the excellent accuracy of the prediction in the training cohort (Fig. 4B). X-tile software was used to generate the optimal cut-off value by the highest χ2 value, which categorized patients into tertiles (low risk group: 0.528-2.6, medium risk group: 2.6–11.7, and high risk group: 11.7–13.6). KM survival analysis revealed significant distinctions between each risk subgroup (HR = 8.837, 95%CI:4.228–18.480, *P*< 0.001) (Fig. 4C).

Validation of the nomogram

To better validating the performance of the nomogram in predicting recurrence, two previously used cohorts were utilized for further testing. The favorable calibration plots indicated the nomogram retained the accuracy of prediction in the validation cohorts (Fig. 4D&E). The discriminable ability was assessed by C-index, which was 0.667 for the GSE50081 and 0.651 for the GSE30219 (95%CI: 0.652–0.682, *P* = 0.0002 and 95%CI: 0.637–0.665, *P* = 0.0004, respectively). Remarkably, the nomogram was separately applied to predict prognosis for stage IA and IA patients, and survival analysis yielded the significant distinctions between three subgroups, which was consistent with previous result (Fig. 5A&B). Furthermore, we utilized the tertiles method in the validation cohorts. In GSE50081, low risk group and medium risk group showed significant different survival compared to high risk group (Fig. 5C). But no statistical significance was achieved between low risk group and medium risk group (*P* = 0.857). While low risk group showed the better recurrence-free survival than medium and high risk subgroup in

GSE30219 (Fig. 5D). However, medium risk subgroup had the similar survival with high risk subgroup ($P = 0.939$).

Discussion

Recurrence of stage I NSCLC is a challenging clinical issue, which shortens the survival time and reduces the effect of surgery(16). Adjuvant therapy has been recommended for postoperative therapy for stage II or III NSCLC patients according to guidelines(3), however, it remains controversial in stage I NSCLC. Published clinical trials have not showed a consistent survival benefit due to drug toxicity and side effect(17). Thus, how to select the patients at high risk of recurrence is an unsolved problem in management of patients with stage I NSCLC.

Increasing investigations were devoted to seek the risk factors for recurrence of early stage NSCLC, which provided the clues for decision-making of postoperative management(18). Brandt et al identified that pT stage and lymphovascular invasion were correlated with distant recurrence and declined the disease-free survival(19). Yu and his colleagues developed and validated a radiomics model for prediction of clinical outcome, which benefited to the choice of treatment(20). In addition, numerous molecular signatures, based on the expression profiles, have been proposed to predict the recurrence-free survival in recent years, which reveal the heterogeneous differences between individuals. Noro et al proposed a two-gene prognostic classifier to predict the recurrence for lung squamous cell carcinoma after surgical resection(21). Shuta's study built a relapse-related molecular signature for lung adenocarcinomas to identify patients at high risk of relapse(22). However, as the problems of small sample size, different microarray platforms, heterogeneity from patients, and diverse range gene selection algorithms, few molecular signatures were broadly adopted in clinic for early stage lung cancer. Compared with previous studies, this study strengthened several aspects and made up the deficiencies. This novel nomogram maximized the potential of molecular biology and clinical factors. In addition, the reliability and robustness of the nomogram was tested in multi-scaled cohorts from different patients and platforms, which showed well performance. As a functional tool derived from molecular biomarkers and clinical variables, it may help optimize patient care by providing better prediction of recurrence, selecting patients for postoperative adjuvant therapy, and stratifying patients in prospective clinical trials. To our knowledge, this is the first study assessing recurrence for stage I NSCLC by combining molecular signature with clinical variables.

In our preliminary work, we found 14 expression profile datasets related to recurrence from GEO database and TCGA. By checking the clinical records and the sample size, we found that TCGA contained the largest sample size and completed clinical records, GSE31210 only involved lung adenocarcinoma samples, and either of GSE41271 or GSE68465 lacked of the expression values of *TPSB2* due to different microarray platforms. Thus, we assigned TCGA into the training cohort and selected GSE50081 and GSE30129 into the validation cohort due to moderate sample size and matching clinical records. The diverse racial group and wide geographic distribution of patients made themselves representativeness and generalizability, which enhanced the reliability of the model. Candidate genes were screened by two

routine algorithms in order to minimize the possibility of missing or ignoring key markers. L1 penalized Cox regression analysis, a broadly adopted method, was utilized to construct the 13-mRNA signature from candidate genes by yielding the corresponding coefficients(14, 23). Our 13-mRNA signature exhibited favorable discrimination in the training and the validation cohorts, with an AUC of 0.79, 0.73, and 0.72, respectively. The cutoff values of different datasets, used to define the high risk and low risk groups by recurrence associated signature, were determined by the corresponding median risk scores owing to different platforms. Published studies presented evidence supporting that a series of clinical variables, such as age, histology, and differentiation, were associated with recurrence-free survival of early stage NSCLC(10). Therefore, we considered the clinical variables and constructed a nomogram by incorporating clinical variables with our molecular signature to provide an easy-to-use tool for clinicians, showing good calibration and discrimination in the training and validation cohorts. Univariable regression analysis revealed that histology showed statistically significant results in the training cohorts. However, after adjusting with the 13-mRNA signature, it wasn't significant, which might be related to the respective small sample size. And our meta-analysis of the entire cohort revealed that age and histology were two key variable associated with recurrence-free survival (Fig. S3). Backward wise step method demonstrated that age, histology, and signature were eligible variables with the least AIC value, which could be incorporated into the nomogram. Validation analysis confirmed the reliability and generalization of the nomogram. Tertile stratified method allowed the remarkably distinctions between survival curves. Notably, we found no statistically significant differences between low and medium risk groups in GSE50081 and medium and high risk groups in GSE30219. This might be the lower samples in low risk group in GSE50081 and lower samples in medium risk group, which couldn't discriminate themselves from other groups.

There are some limitations of this study should be mentioned. First of all, this was a multi-scaled study based on the datasets from GEO and TCGA, but prospective studies are required to further validate our finding. In addition, several significant clinical variables were not recorded in some datasets, which hampered the accuracy of our model. Furthermore, different platforms of the datasets hindered the integrated analysis of these datasets, which reduced the power of the model.

Conclusion

In this study, we developed a reliable and robust nomogram which could be reproducible in ethnically and geographically diverse cohorts. It will provide classification of patients at high risk for recurrence is important for identifying those who may benefit from adjuvant chemotherapy and optimize the design of prospective clinical trials.

Abbreviations

NSCLC
non-small cell lung cancer
LASSO

Least Absolute Shrinkage and Selector Operation
SVM-RFE
Support Vector Machine-Recursive Feature Elimination

Declarations

Conflicts of interest: All authors declared no conflicts of interest.

Acknowledgment

This study was supported by the Xuzhou Science and Technology Bureau (No. KC20103), Provincial Commission of Health and Family Planning (No. H2018112), Cadre Health Project in Jiangsu Province (No. BJ17025).

Data Availability Statement:

I confirm that my data is available.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA: A Cancer Journal for Clinicians. 2019;69(1):7-34.
2. Schuchert MJ, Normolle DP, Awais O, Pennathur A, Wilson DO, Luketich JD, et al. Factors influencing recurrence following anatomic lung resection for clinical stage I non-small cell lung cancer. Lung Cancer. 2019;128:145-51.
3. Zhong WZ, Wang Q, Mao WM, Xu ST, Wu L, Shen Y, et al. Gefitinib versus vinorelbine plus cisplatin as adjuvant treatment for stage II-IIIA (N1-N2) EGFR-mutant NSCLC (ADJUVANT/CTONG1104): a randomised, open-label, phase 3 study. (1474-5488).
4. Kris MG, Gaspar LE, Chaft JE, Kennedy EB, Azzoli CG, Ellis PM, et al. Adjuvant Systemic Therapy and Adjuvant Radiation Therapy for Stage I to IIIA Completely Resected Non-Small-Cell Lung Cancers: American Society of Clinical Oncology/Cancer Care Ontario Clinical Practice Guideline Update. LID - 10.1200/JCO.2017.72.4401 [doi]. (1527-7755).
5. Osarogiagbon RU, Veronesi G, Fang W, Ekman S, Suda K, Aerts JG, et al. Early-Stage NSCLC: Advances in Thoracic Oncology 2018. (1556-1380).
6. Mao Q, Xia W, Dong G, Chen S, Wang A, Jin G, et al. A nomogram to predict the survival of stage IIIA-N2 non-small cell lung cancer after surgery. (1097-685X).
7. Jamal-Hanjani M, Wilson GA, Horswell S, Mitter R, Sakarya O, Constantin T, et al. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. LID - 10.1093/annonc/mdw037 [doi]. (1569-8041).
8. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. LID - 10.1038/nm.3850 [doi].

(1546-170X).

9. Xu W, Jia G, Davie JR, Murphy L, Kratzke R, Banerji S. A 10-Gene Yin Yang Expression Ratio Signature for Stage IA and IB Non-Small Cell Lung Cancer. (1556-1380).
10. Liang W, Zhang L, Jiang G, Wang Q, Liu L, Liu D, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. LID - 10.1200/JCO.2014.56.6661 [doi]. (1527-7755).
11. Kent MS, Mandrekar SJ, Landreneau R, Nichols F, Foster NR, DiPetrillo TA, et al. A Nomogram to Predict Recurrence and Survival of High-Risk Patients Undergoing Sublobar Resection for Lung Cancer: An Analysis of a Multicenter Prospective Study (ACOSOG Z4032). LID - 10.1016/j.athoracsur.2016.01.063 [doi]. (1552-6259).
12. Won YW, Joo J, Yun T, Lee GK, Han JY, Kim HT, et al. A nomogram to predict brain metastasis as the first relapse in curatively resected non-small cell lung cancer patients. LID - 10.1016/j.lungcan.2015.02.006 [doi]. (1872-8332).
13. Zhang Y, Sun Y, Xiang J, Zhang Y, Hu H, Chen H. A clinicopathologic prediction model for postoperative recurrence in stage Ia non-small cell lung cancer. LID - 10.1016/j.jtcvs.2014.02.064 [doi]. (1097-685X).
14. Qiu J, Peng B, Tang Y, Qian Y, Guo P, Li M, et al. CpG Methylation Signature Predicts Recurrence in Early-Stage Hepatocellular Carcinoma: Results From a Multicenter Study. LID - 10.1200/JCO.2016.68.2153 [doi]. (1527-7755).
15. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. (1078-0432).
16. Dai C, Shen J, Ren Y, Zhong S, Zheng H, He J, et al. Choice of Surgical Procedure for Patients With Non-Small-Cell Lung Cancer ≤ 1 cm or > 1 to 2 cm Among Lobectomy, Segmentectomy, and Wedge Resection: A Population-Based Study. LID - 10.1200/JCO.2015.64.6729 [doi]. (1527-7755).
17. Haratani K, Hayashi H, Chiba Y, Kudo K, Yonesaka K, Kato R, et al. Association of Immune-Related Adverse Events With Nivolumab Efficacy in Non-Small-Cell Lung Cancer. LID - 10.1001/jamaoncol.2017.2925 [doi]. (2374-2445).
18. Zhang Y, Zheng D, Xie J, Li Y, Wang Y, Li C, et al. Development and Validation of Web-Based Nomograms to Precisely Predict Conditional Risk of Site-Specific Recurrence for Patients With Completely Resected Non-small Cell Lung Cancer: A Multiinstitutional Study. (1931-3543).
19. Brandt WS, Bouabdallah I, Tan KS, Park BJ, Adusumilli PS, Molena D, et al. Factors associated with distant recurrence following R0 lobectomy for pN0 lung adenocarcinoma. (1097-685X).
20. Yu W, Tang C, Hobbs BP, Li X, Koay EJ, Wistuba, II, et al. Development and Validation of a Predictive Radiomics Model for Clinical Outcomes in Stage I Non-small Cell Lung Cancer. (1879-355X).
21. Noro R, Ishigame T, Walsh N, Shiraishi K, Robles AI, Ryan BM, et al. A Two-Gene Prognostic Classifier for Early-Stage Lung Squamous Cell Carcinoma in Multiple Large-Scale and Geographically Diverse Cohorts. (1556-1380).

22. Tomida S, Takeuchi T, Shimada Y, Arima C, Matsuo K, Mitsudomi T, et al. Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. LID - 10.1200/JCO.2008.19.7053 [doi]. (1527-7755).
23. Li B, Cui Y, Diehn M, Li R. Development and Validation of an Individualized Immune Prognostic Signature in Early-Stage Nonsquamous Non-Small Cell Lung Cancer. LID - 10.1001/jamaoncol.2017.1609 [doi]. (2374-2445).

Tables

Table 1. Baseline information of TCGA and GEO datasets.

Datasets	Sample size	Platform	No. of stage I	Year	Country
Discovery step					
GSE41271	275	Illumina HumanWG-6 v3.0 expression beadchip	132	2013	USA
GSE7339	100	Hitachisoft AceGene Human Oligo Chip 30K Subset A	-	2007	Japan
GSE11969	163	Agilent Homo sapiens 21.6K custom array	78	2009	Japan
GSE13213	117	Agilent-014850 Whole Human Genome Microarray	61	2009	Japan
GSE5843	48	PRHU05-S1-0006 (PC Human Operon v2 21k)	48	2007	Australia
GSE8894	138	Affymetrix Human Genome U133 Plus 2.0 Array	88	2007	South Korea
GSE5123	51	GPL3877 PRHU05-S1-0006 (PC Human Operon v2 21k)	30	2006	Australia
GSE32863	116	Illumina HumanWG-6 v3.0 expression beadchip	34	2012	USA
GSE68465	462	GPL96 Affymetrix Human Genome U133 Plus 2.0 Array	114	2015	USA
GSE31210	246	GPL570 Affymetrix Human Genome U133 Plus 2.0 Array	162	2011	Japan
GSE37745	196	GPL570 Affymetrix Human Genome U133 Plus 2.0 Array	50	2012	Sweden
Training step					
TCGA	1124	Illumina Hiseq	351	2015	TCGA project
Testing step					
GSE50081	181	GPL570 Affymetrix Human Genome U133 Plus 2.0 Array	124	2013	Canada
GSE30219	307	GPL570 Affymetrix Human Genome U133 Plus 2.0 Array	142	2013	France

Table 2. Univariable and multivariable Cox regression analysis of training datasets.

Variables		Univariable		Multivariable	
	No. of cases	HR(95%CI)‡	*P* value	HR(95%CI)	*P* value
Training (TCGA)	351				
Age median(IQR¶)	68(61-74)	1.001(0.976-1.028)	0.925	1.011(0.985-1.037)	0.428
Sex (man vs women)	190/161	0.799(0.499-1.280)	0.351	0.932(0.578-1.502)	0.772
Smoking (smoker vs non-smoker)	233/118	1.003(0.611-1.648)	0.991	1.017(0.595-1.738)	0.952
Histology (non-squamous vs squamous cell carcinoma)	189/162	1.923(1.186-3.116)	0.008*	1.369(0.816-2.298)	0.234
Stage (IA vs IB)	191/160	0.999(0.622-1.606)	0.997	8.283(3.917-17.513)	0.966
Risk score (high vs low)	177/174	8.816(4.216-18.436)	<0.001*	0.989(0.603-1.623)	<0.001*
Validation1 (GSE50081)	124				
Age median(IQR)	71(63-76)	1.015(0.979-1.054)	0.416	1.006(0.970-1.043)	0.752
Sex (man vs women)	68/56	1.550(0.742-3.237)	0.243	1.482(0.679-3.232)	0.323
Smoking (smoker vs non-smoker)	39/71†	0.810(0.461-1.425)	0.465	0.634(0.362-1.111)	0.111
Histology (non-squamous vs squamous cell carcinoma)	91/33	1.052(0.484-2.286)	0.898	0.967(0.436-2.143)	0.934
Stage (IA vs IB)	46/78	0.336(0.138-0.820)	0.017*	0.343(0.138-0.848)	0.021*
Risk score (high vs low)	64/60	3.556(1.587-7.966)	0.002*	3.334(1.467-7.577)	0.004*
Validation2 (GSE30219)	142				
Age median(IQR)	62(55-70)	1.031(0.999-1.065)	0.060	1.024(0.990-1.060)	0.165
Sex (man vs women)	118/24	3.890(0.933-6.223)	0.062	1.057(0.404-2.764)	0.910
Smoking (smoker vs non-smoker)	-	-	-	-	-
Histology (non-squamous vs squamous cell carcinoma)	86/56	1.488(0.764-2.897)	0.242	1.789(0.878-3.646)	0.109

Stage (IA vs IB)	-	-	-	-	-
Risk score (high vs low)	71/71	2.521(1.233-5.152)	0.011*	2.883(1.353-6.140)	0.006*

¶: IQR = interquartile range; †: 14 cases without smoking status; *: significant *P* value; ‡: HR= hazard ratio, CI = confidence interval.

Figures

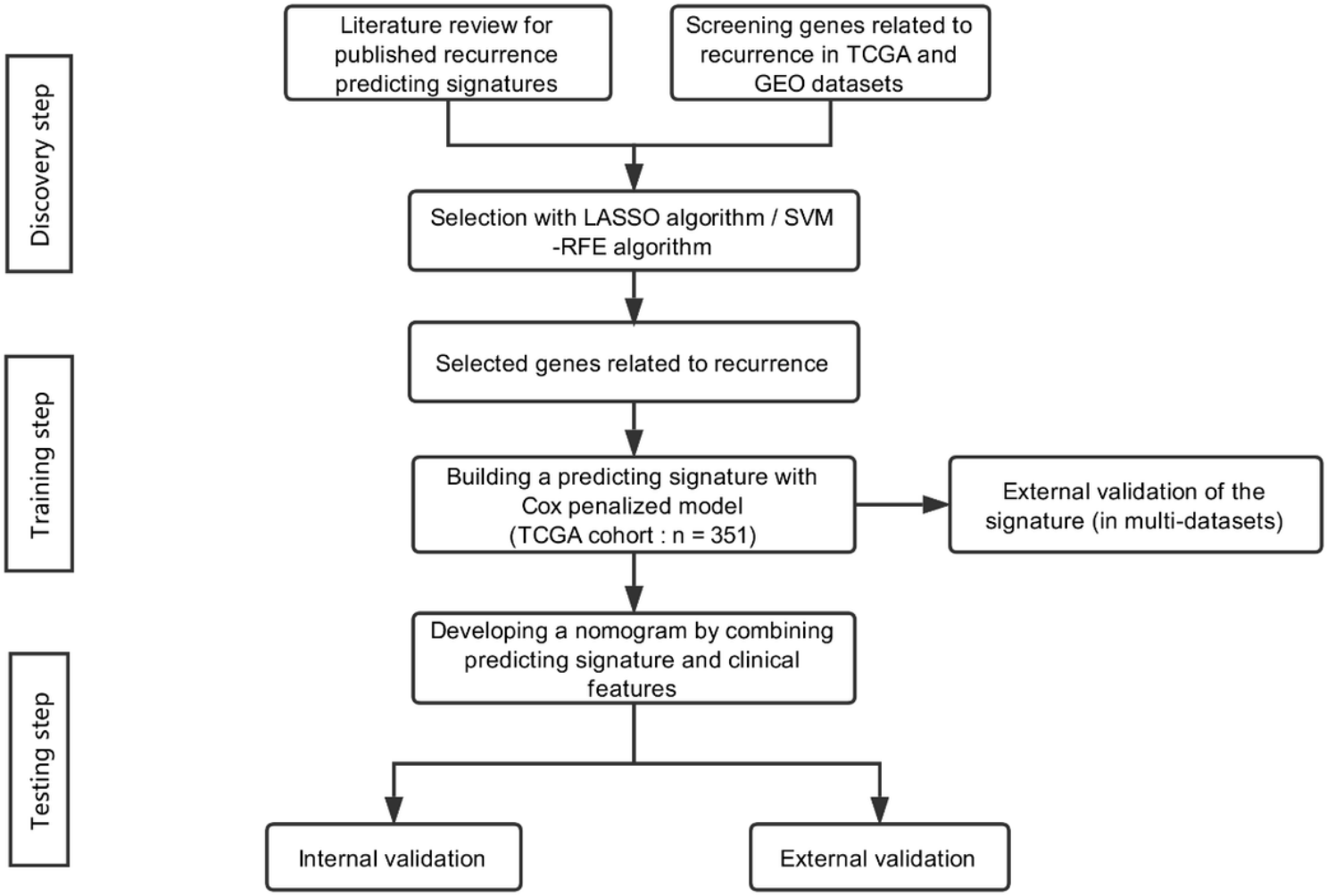

Figure 1

Study flowchart. LASSO, Least Absolute Shrinkage and Selector Operation; SVM-RFE, Support Vector Machine-Recursive Feature Elimination; TCGA, the Cancer Genome Atlas.

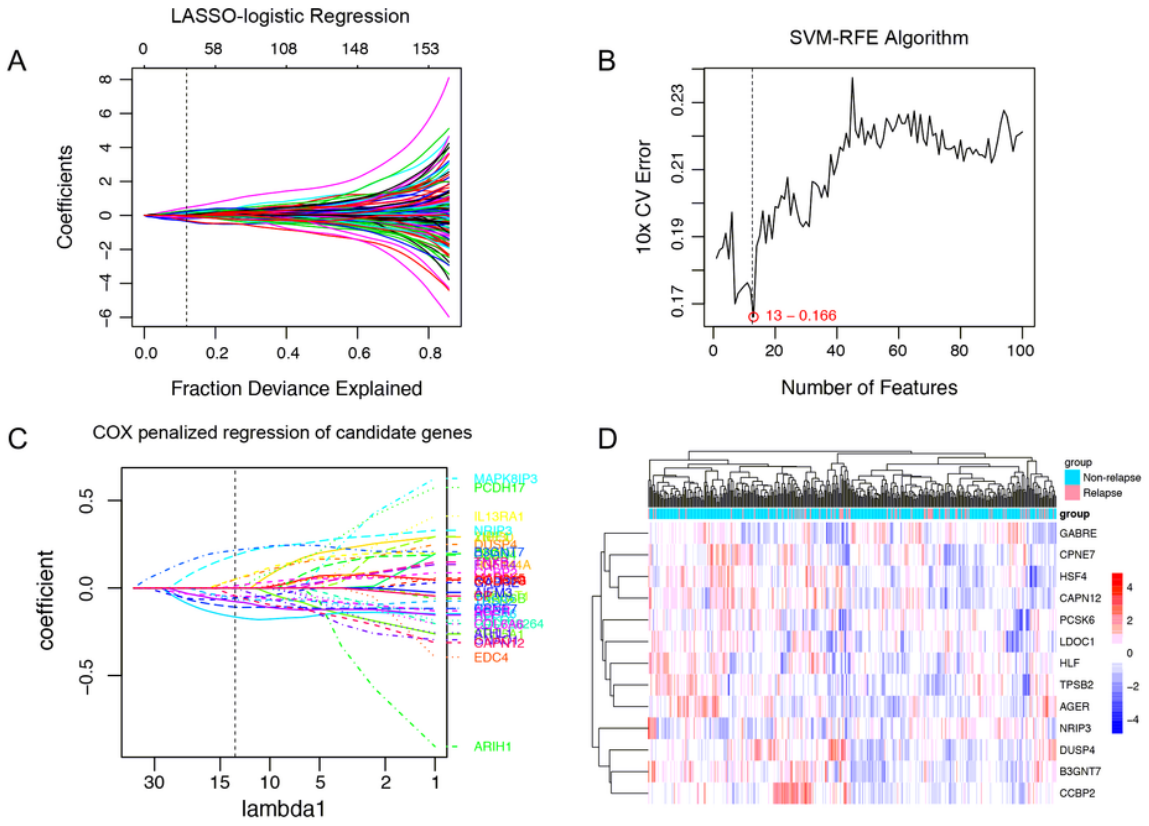

Figure 2

Two algorithms were used for feature selection. (A) LASSO and (B) SVM-RFE algorithms in the discovery cohort. (C) Penalized cox regression was used to identify the significant genes related to recurrence in training cohort. (D) Cluster analysis of incorporation of genes that were selected from previous algorithms in the discovery cohort. LASSO, Least Absolute Shrinkage and Selector Operation; SVM-RFE, Support Vector Machine-Recursive Feature Elimination.

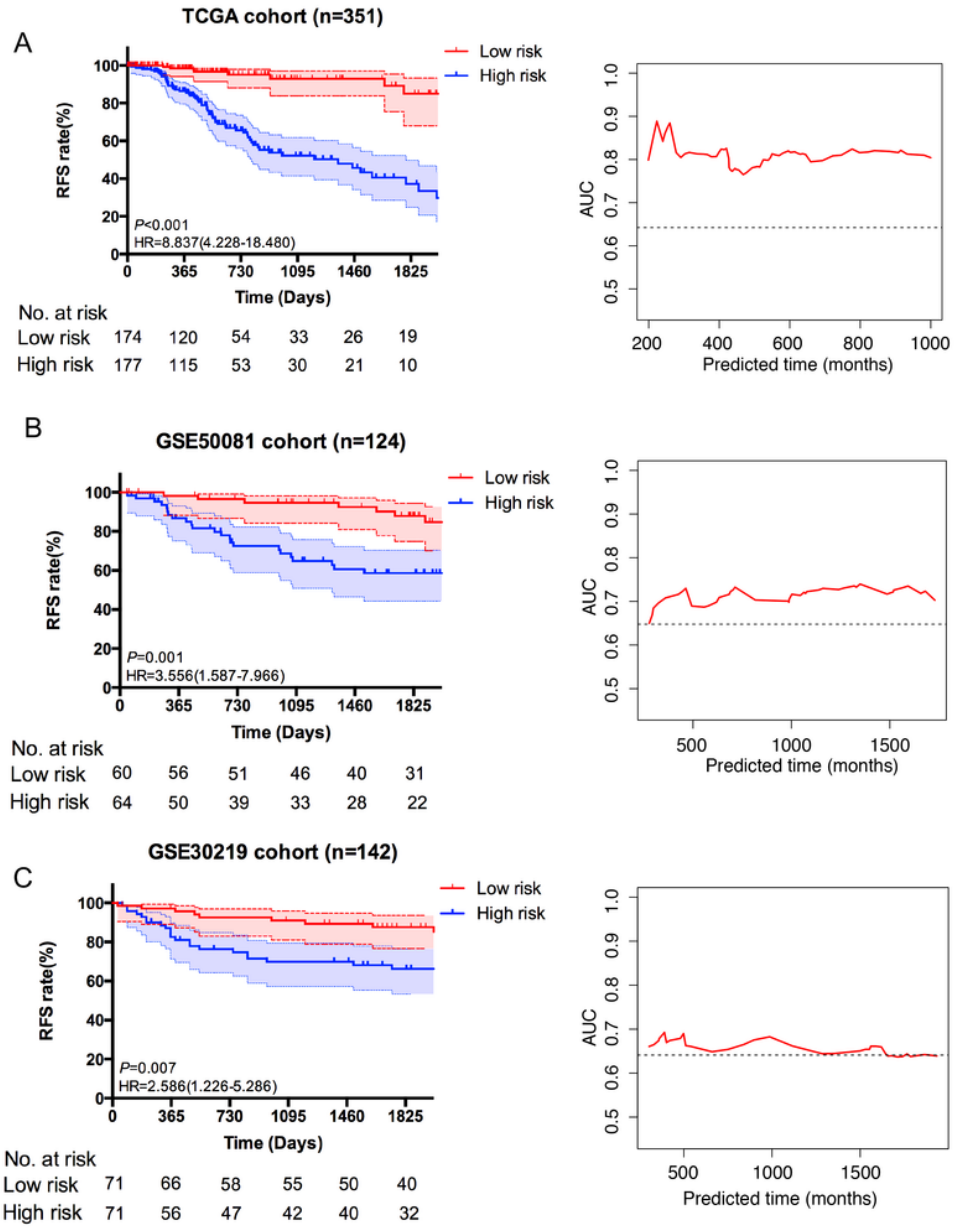

Figure 3

The performance of signature in predicting recurrence in the three cohorts: (A) training cohort, (B) external validation cohort 1, and (C) external validation cohort 2. HR, hazard ratio; AUC, area under the curve.

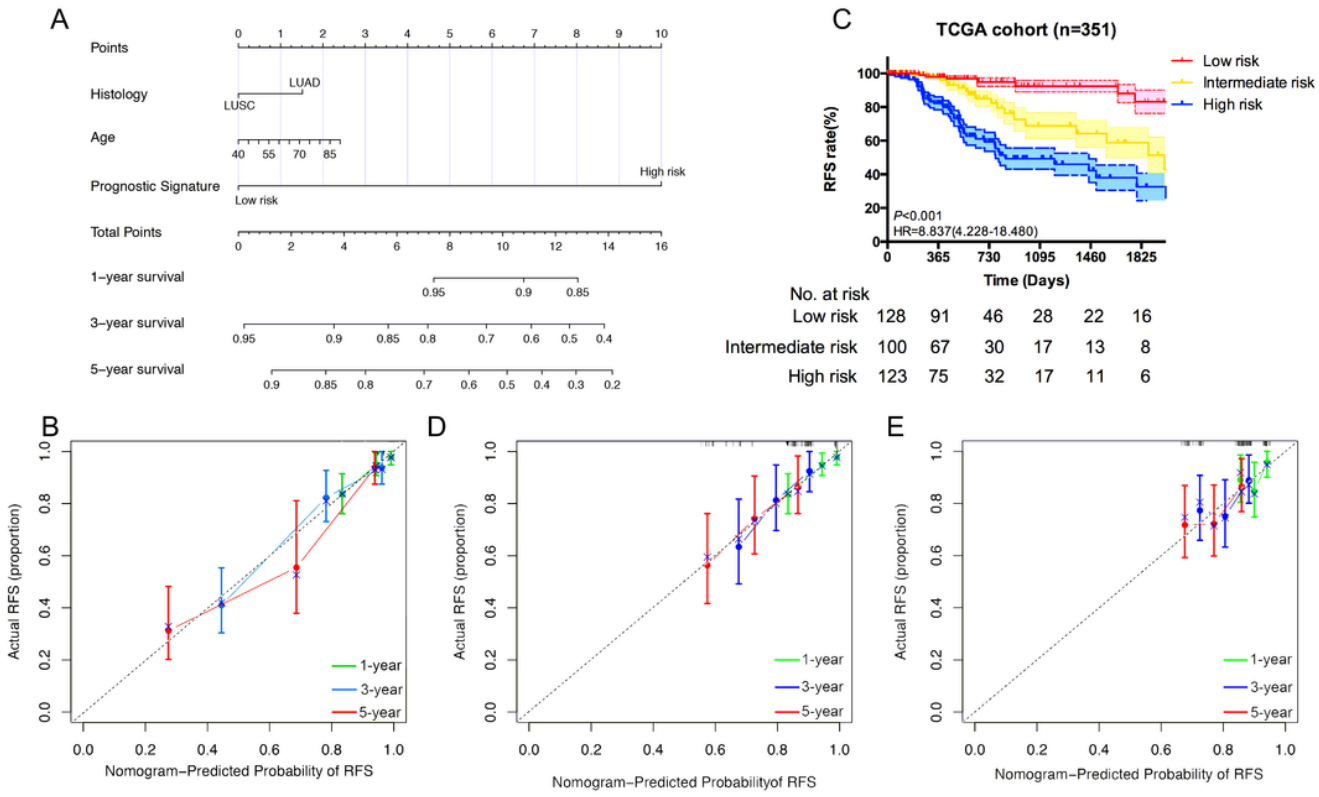

Figure 4

(A) Nomogram to predict the 1-year, 3-year, and 5-year RFS. Calibration curve for RFS nomogram model in (B) TCGA cohort, (D) external validation 1 cohort, and (E) external validation 2 cohort. The dashed represents the ideal nomogram, and the green represents the 1-year observed, blue represents the 3-year observed, and red represents the 3-year observed nomogram. RFS, recurrence-free survival. (C) Kaplan-Meier curves plotting RFS of different risk subgroups stratified with nomogram score in TCGA cohort.

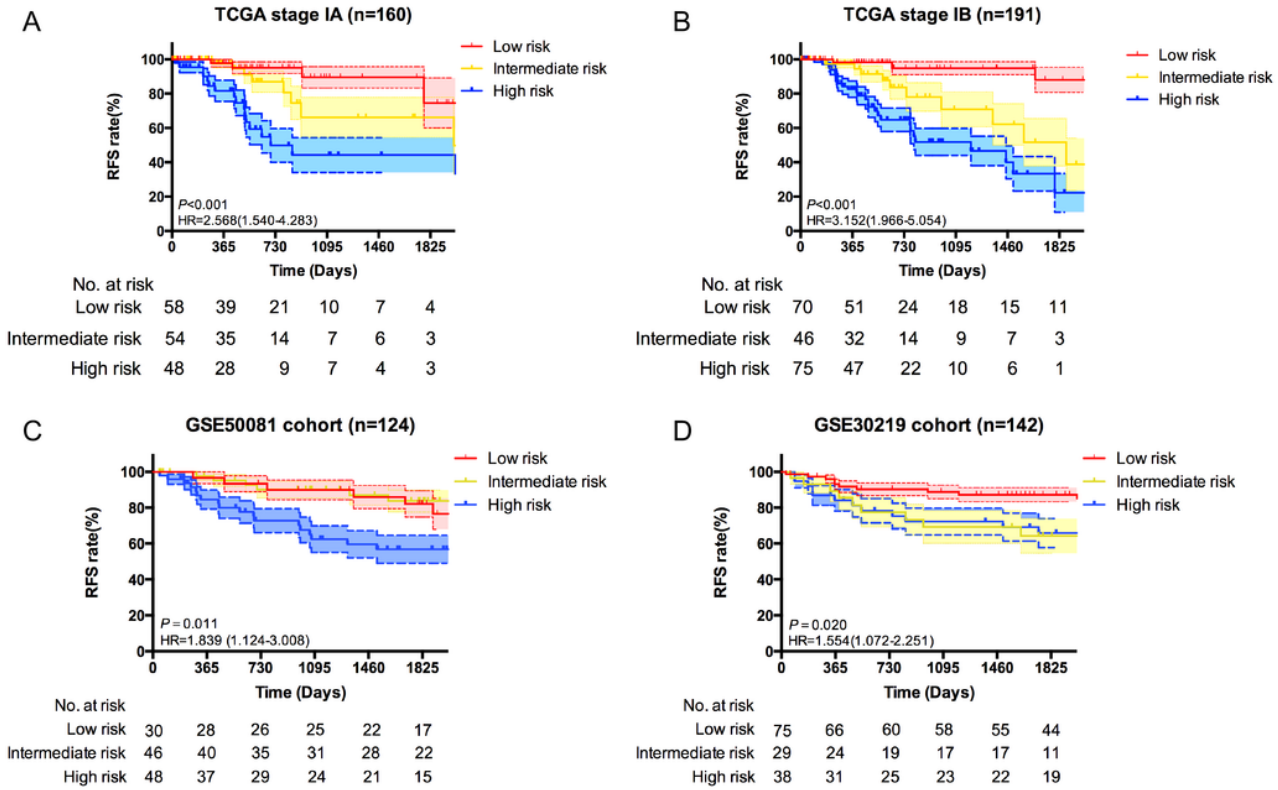

Figure 5

Subgroup and stratification analysis of the relapse-related nomogram. Subgroup analysis for stage IA (A)and stage IB (B) in TCGA cohort. (C&D) Kaplan-Meier curves plotting RFS of two validation cohorts for respective nomogram score categories.

Supplementary Files

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

- SupplymentalTables.docx
- figureS1.tif
- figureS2.tif
- FigureS3.tif