Development of a Novel Prognostic Score Combining Clinicopathologic Variables, Gene Expression and Mutationprofiles for Lung Adenocarcinoma

Guofeng Li
Shenzhen People's Hospital
Guangsuo Wang
Shenzhen People's Hospital
Yanhua Guo
Tongji university
Shixuan Li
Shenzhen People's Hospital
Youlong Zhang
HuaJia Biomedical Intelligence
Jialu Li (jialu.li@huajiabio.com)
HuaJia Biomedical Intelligence https://orcid.org/0000-0002-6411-6876
Bin Peng
Shenzhen People's Hospital

Technical innovations

Keywords: prognosis, gene expression profiles, lung adenocarcinoma, competing risks analysis, risk stratification, event-free survival, recurrence-free survival, integrative analysis

Posted Date: June 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-37856/v1

Version of Record: A version of this preprint was published on September 19th, 2020. See the published version at https://doi.org/10.1186/s12957-020-02025-0.

Development of a novel prognostic score combining clinicopathologic variables, gene expression and mutation profiles for lung adenocarcinoma

Guofeng Li[1], Guangsuo Wang[1], Yanhua Guo[2], Shixuan Li[1], Youlong Zhang[3], Jialu Li[3]*, Bin Peng[1]*

1 Department of Thoracic Surgery, Shenzhen People's Hospital, Second Clinical Medical College of Jinan University

2 Department of Thoracic Surgery, Shanghai Pulmonary Hospital, School of Medicine, Tongji University

3 Department of Biostatistics, HuaJia Biomedical Intelligence

Correspondence to:

Jialu Li, PhD, Department of Biostatistics, HuaJia Biomedical Intelligence, Shenzhen Overseas Chinese High-Tech Venture Park, Nanshan district, Shenzhen, China, 518057. Email: Jialu.li@huajiabio.com

Or to

Bin Peng, MD, Department of Thoracic Surgery, Shenzhen People's Hospital, Luohu district, Shenzhen, China, 518020. Email: 183672297@qq.com

Abstract

Background: Integrating phenotypic and genotypic information to improve prognostic prediction is under active investigation for lung adenocarcinoma (LUAD). In this study, we developed a new prognostic model for event-free survival (EFS) and recurrence-free survival (RFS) based on the combination of clinicopathologic variables, gene expression and mutation data.

Methods: We enrolled a total of 408 patients from the Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) project for the study. We pre-selected gene expression or mutation features, and constructed 14 different input feature sets for predictive model development. We assessed model performance with multiple evaluation metrics including the distribution of C-index on testing dataset, risk score significance and time-dependent AUC under competing risks scenario. We stratified patients into higher and lower-risk subgroups by the final risk score, and further investigated underlying immune phenotyping variations associated with the differential risk.

Results: The model integrating all three types of data achieved the best prediction performance. The resultant risk score provided a higher-resolution risk stratification than other models within pathologically-defined subgroups. The score could account for extra EFS-related variations that were not captured by clinicopathologic scores. Being validated for RFS prediction under a competing risks modeling framework, the score achieved a significantly higher time-dependent AUC as compared to that of the conventional clinicopathologic variables-based model (0.772 vs. 0.646, p-value< 0.001). The higher-risk patients were characterized with transcriptional aberrations of multiple immune-related genes, and a significant depletion of mast cells and natural killer cells.

Conclusions: We developed a novel prognostic risk score with improved prediction accuracy, using clinicopathologic variables, gene expression and mutation profiles as input, for LUAD. Such score was an significant predictor of both EFS and RFS.

Trail registration: This study was based on public open data from TCGA and hence the study objects were retrospectively registered.

Keywords

prognosis, gene expression profiles, lung adenocarcinoma, competing risks analysis, risk stratification, event-free survival, recurrence-free survival, integrative analysis.

Background

Lung cancer is the most frequently diagnosed cancer and the leading cause of cancer death, with a total of 2,093,876 new cases (11.6% of all cancers) and 1,761,007 deaths (18.4% of all cancers) reported worldwide in 2018[1]. LUAD is a major type of primary lung cancer which accounts for about 35% of all cases[2]. Improving survival of lung cancer is of high importance since the 5-year survival rate remains <15% and 10-year survival rate < 7%[3]. Currently, clinicopathologic factors including

American Joint Committee on Cancer (AJCC) tumor stage, tobacco smoking history and radiation therapy are used for prognostic analysis. However, whether the prediction performance of these clinicopathologic factors can be improved with phenotypic and genotypic profiles at gene level is still under investigation.

The high-throughput sequencing technology has made it possible a comprehensive interrogation of whole transcriptome and genome of tumor tissues at an increasingly reasonable cost[4, 5]. Previous studies focused on finding prognostic signatures based on gene expression[6-9] or mutation[10-12] for LUAD patients. For example, Li *et al.*[7] reported gene expression-based models with an average C-index of 0.604 on testing datasets from TCGA-LUAD in predicting overall survival (OS). Other studies using multiple types of input data made statistical inference on the significance of potential individual prognostic factors[13-16]. Two of these studies[15, 16] had shown clear benefit of combining genetic mutations and expression profiles in predicting OS and RFS at cross-validation level. In particular, they inferred that the genotype and expression data made around 5% and 50% relative contributions to explained variance of survival outcomes[16].

In this study, we aimed to improve prognostic prediction for LUAD by extensively integrating clinicopathologic variables, gene expression and mutation profiles as the input. We focused on the analysis of recurrence and death events as there exists minimal ambiguity in the database about the derivation of these outcomes[17]. We believe that our work will be informative for those who want to improve the precision treatment of LUAD.

Methods

Data

The study enrolled from the Cancer Genome Atlas lung adenocarcinoma (TCGA-LUAD) project[18] a total of 408 patients with relatively complete information in high-throughput DNA and RNA sequencing data, major clinicopathologic variables (at most 10 missing values was allowed) and follow-up data for recurrence or death events. The RNA expression was measured on a total of 60,483 genes and the somatic mutation was detected among 16,980 genes for each patient. The study cohort included a total of 8 clinicopathologic variables: age of initial diagnosis, gender, tobacco smoking status, AJCC tumor stage, adjuvant radiation treatment, adjuvant pharmaceutical treatment, history of other malignancies, and the anatomic position of tumor (Table S1). For missing value imputation, we used the mean estimate for continuous variables, and multinomial random sampling for categorical variables. The follow-up data included three types of events: recurrence, death and last follow-up, where the recurrence and death were defined as the composite event of interest in the EFS analysis. The last follow-up occurred before the events of interest were considered as the censoring event. In this cohort, 164 recurrence events, 45 dead events and 199 censoring events were observed.

Analysis

As shown in Figure 1, the workflow in this study can be sketched in four parts: 1) data preprocessing, 2) feature integration and model development using the training set, 3) prediction and model evaluation using the testing set, and 4) exploration of molecular mechanisms related to differential prognostic risk.

Data preprocessing

We downloaded the gene expression and somatic mutation detection data generated by TCGA group[18] for this study. We used the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value to represent gene expression level. We restricted our analysis to genes with a FPKM summed across all samples greater than 500 and with non-zero expression in at least 200 patients. These genes were further filtered based on variance, in which genes with a standard error of log-transformed FPKM across samples greater than 0.4 were retained. This resulted in a total of 7,401 genes for model development. For gene mutation, we pooled single nucleotide variations (SNVs) and indels for analysis. We filtered out mutated genes, defined as those with at least one somatic mutation, occurred in fewer than 30 samples. A total of 271 genes passed such filtering. The distribution of the included clinicopathologic variables and genes with top 10 mutation frequencies were summarized in Table S1.

Model development

We applied univariate Cox regression model for feature pre-selection of the gene expression or mutation data. The model was fitted between each feature and EFS, and the importance of feature was determined by their ward test p-value. We set the p-value (unadjusted for multiple testing) cutoff as 3e-04 and 0.08 for gene expression and mutation, respectively. We then used lasso Cox regression[19] to develop the predictive model, using the pre-selected features as the input.

For model development, we randomly split the study cohort into training (285 patients) and testing (123 patients) sets. The input data for prediction model were prepared in following ways (Figure 1 & Table S2). In the first way, the features processed as described above were simply used as the input data without any further modification. In the second way, we used the univariate Cox model to narrow down the searching scope of gene expression and mutation data, respectively. These pre-selected genes were used as the input for prediction model development. In the third way, we performed a pairwise combination between the three types of predictors. We also included a combination that uses all three types of data as the input. Only the genes pre-selected in the second way were used here. In the fourth way, we added interaction terms to those prepared in the third way. The interaction features were generated by multiplication of any two or three types of features from the input. An interaction feature was included in predictive modeling only when its distribution is not severely unbalanced. A total of 14 different model input feature sets were constructed.

We used the one-standard-error rule[20] and the 3-fold cross validation method to find an optimized L1 penalty value for every input set using the training dataset. This penalty was finally used to develop a lasso Cox model using all training dataset.

Model evaluation

We used the C-index[21] as one criteria to evaluate the predictive performance of models on the testing dataset. We performed 1000 repetitions of model development and evaluation to mitigate the bias caused by data splitting. The score $X_{test}\beta$ was computed by multiplying the model coefficients and the features in testing dataset. This newly generated variable was analyzed in the Cox models to evaluate its significance related to EFS. Furthermore, we calculated $X\beta$ as the risk score to stratify patients within specifically defined subgroups. The Kaplan-Meier method was then used to analyze their event-free survival distribution.

We also used competing risks modeling to evaluate the significance of the risk score as a univariate predictor for RFS events. As sketched in Figure 3A, after initial treatment, one may progressed to recurrence (from 1 to 2) or to death (from 1 to 3). The death event would stop patients from having a recurrence, and thus posed a competing risk to recurrence. The competing risks model applied the cumulative incidence function (CIF) $I_k(t)$ to calculating the cumulative probability of each cause. The computational formula of CIF is given by:

$$I_k(t) = \Pr(T \leq t, D = k) = \int_0^t \lambda_k(t) S(s) ds$$

where $\lambda_k(t)$ is the hazard of cause k at time t, $S(t) = \exp\left[-\sum_{k=1}^{K} \int_0^t \lambda_k(s) ds\right]$ is the survival function. To incorporate covariates information, we used Fine and Gray method[22] to impose a proportional hazards assumption on the subdistribution hazard:

$$\bar{\lambda}_k(t|\mathbf{Z}) = \bar{\lambda}_{k,0}(t)\exp(\boldsymbol{\beta}_k^T \mathbf{Z})$$

where $\boldsymbol{\beta}$ is a vector of coefficients and $\mathbf{Z}$ is a matrix of covariates. Individuals who fail from another cause are remained in the risk set for $\bar{\lambda}_k(t)$ estimation[23]. The time-dependent AUC[24] was computed to evaluate the model fitting. The confidence interval was computed based on Blanche *et al.*[25]. All analyses were performed by R version 3.6 and packages including survival, glmnet, caret, cmprsk and riskRegression.

Exploration of underlying mechanisms

For differential gene expression analysis, we used the negative binomial generalized linear model with tag-wise dispersion in R package edgeR[26]. The raw count data was normalized by the TMM (the trimmed mean of M values)[27]. Only genes whose mean of counts was more than 15 reads and with non-zero count in every sample were retained for normalization. This resulted in a total of 15,653 genes used for downstream analysis. We performed gene sets enrichment analyses using multiple algorithms including GOseq, Enrichr, and GSEA[28-30].

We used CIBERSORT software to deconvolve the relative fractions of different immune cell types from the RNA sequencing data[31]. To infer the significance of enrichment of cell types between the higher and lower-risk patient subgroups, we used Wilcoxon rank-sum test to compute p-values. All p-values were corrected by Benjamini-Hochberg procedure to control the false discovery rate (FDR) and to obtain the adjusted p-values[32].

Result

Patient characteristics and feature processing

The study workflow was sketched in Figure 1. We enrolled a total of 408 patients with complete information in EFS data, major relevant clinicopathologic variables, and gene expression and mutation profiles. The median EFS time was 809 days (Figure S1; 95% CI: 692, 1018). We randomly split the data into model development dataset (285 patients) and testing dataset (123 patients) with a comparable censoring ratio (48.9% vs. 48.6%) for 1000 times. The average median EFS time for the development and testing dataset were 822 and 834 days, respectively. The distribution of included clinicopathologic variables of the cohort was summarized in Table S1. Only AJCC tumor stage and adjuvant treatment

were significantly associated with EFS.

For model input data integration, we prepared 5 sets of features selected from single type of data (single type features), 4 sets of features combined from different types of data (combined features), and 5 sets of features incorporated with interaction terms generated within (intra-type) each type of data or between (inter-type) different types of data (combined and interaction features). The size of each feature space was summarized in Table S2.

Comparison of integrated prognostic models

We first compared the prediction accuracy of models developed based on single type of input features. The performance of models based on clinicopathologic variables was the best with a C-index of 0.624±0.028 on the testing set (Figure S2A and Table S3). To assess the effectiveness of feature pre-selection, we compared models using features with or without univariate Cox analysis. For gene expression, the mean of testing C-index increased by 0.029 with univariate pre-selection (p-value< 0.001), while for mutation profiles, the mean of prediction accuracy increased by 0.013 (p-value< 0.001). These indicated the benefit of feature pre-selection step.

We next compared prognostic models that integrate different types of input data. The best prediction model was the model combining three types of input data, which achieved a significantly higher mean C-index (0.639±0.033) on the testing data as compared to the clinicopathologic model (p-value< 0.001; Figure S2B and Table S3). We then assessed whether the inter-type and intra-type interaction covariates can improve the prediction accuracy. Adding interaction covariates had limited benefits on prediction power (Figure S3C and Table S3). The final data integration we presented in this study was thus the one combining three types input variables without interactions.

Assessment of significance of the prognostic risk score

A successful application of a prognostic model requires a risk score that can be readily computed for clinical use. We therefore selected an individual model with a C-index (0.638) close to the mean of final data integration as described above, and calculated the linear combination of coefficients and features from the model as the event-risk score (or mathematically $X\beta$). We named this score as the mul-score (Table S4).

To further evaluate the significance of mul-score as compared to that of the cln-score (the risk score computed by clinicopathologic variables-based model), we fitted Cox proportional hazard models by setting the score as the single covariate on the testing set. For a fair comparison, we computed the cln-score from a model with a C-index (0.622) also close to the mean of clinicopathologic variables-based data integration. The p value of mul-score coefficient was more significant than that of cln-score in such univariate modeling (Table S5). When a multivariate Cox model was fitted using the two scores as covariates, only the mul-score was still statistically significant (Table S5). This suggested that the mul-score could capture extra EFS-related information that was not considered by cln-score.

We then investigated the risk stratification effectiveness of the two risk scores within specific groups of patients (Table S6). We found that the mul-score was not only significantly associated with EFS in each group, but also showed a higher level of relevance than that of the cln-score, as reflected by the fitting

p-values. We set the score median within each group as the threshold to stratify for the higher- and lower-risk subgroups. The mul-score generated a more striking stratification within each set of patients as compared to the cln-score (Figure 2, Figure S3, Table S6). For example, for stage IA subgroup, the mul-score identified a higher-risk subgroup with a median EFS time 18 months earlier than that of the cln-score (702 vs. 1255 days). On the other hand, for stage IIB, the mul-score revealed a significantly lower-risk subgroup who would develop an event 19 months later than that of the cln-score (1146 vs. 578 days).

RFS analysis

We next assessed the significance of proposed risk score as a predictor for RFS. We used competing risks models for the assessment because a total of 45 death events without recurrence observed in the cohort can act as the competing event for recurrence risk analysis (Figure 3A). Ignoring the effect from such competing event could lead to an over-estimation of recurrence risk (Figure 3B) since those who died without recurrence were still considered as having possibility of developing recurrence. The cumulative probabilities of these two types of events were shown in Figure 3C. Most failure events of both causes occurred before about day 2,000, and the failure rate became lower after then for recurrence while unchanged for direct death. The mean time-dependent AUC of mul-score for RFS prediction was significantly higher than that of the cln-score (Figure 3D; 0.772 vs. 0.646, p-value< 0.001).

Differential risk-related immune phenotyping variations

To explore the variations of immune phenotyping associated with differential prognostic risk, we performed differential gene expression analysis between the higher-risk (n=204) and lower-risk (n=204) patients as determined by mul-score. A total of 7,243 genes was differentially expressed, and two GOs were identified as significantly enriched: ficolin-1-rich granule (GO:0101002, adjusted p-value: 0.020) and ficolin-1-rich granule lumen (GO:1904813, adjusted p-value: 0.026).

For the immune phenotyping analysis, a total of 604 differentially expressed genes described above was immune-related according to a curated list generated from the immunology database and analysis portal (ImmPort)[33]. We further identified that 12 of them were also on the list of differentially expressed genes detected within stage IIB and within IIIA subgroups (Figure 2 & Table S7). As expected, the tumor suppressor gene *ACVR1B* was significantly down-regulated in the higher-risk patients. However, another gene *MSR1,* previously being reported as tumor suppressor in leukemia[34], was up-regulated, suggesting the complexity of tumor immune microenvironment in a high prognostic risk scenario.

We then used computational deconvolution methods to investigate the variations of immune cell compositions between the two risk subgroups. As shown in Figure S4, the inferred relative fractions of immune cell compositions varied both within and across risk subgroups (Figure S4A). In higher-risk patients, we observed a significant depletion of mast cells and activated natural killer cells (Figure S5B), indicating a transformation of innate immunity in LUAD tumor microenvironment from activating to suppressive status.

Discussion

Our study developed and validated a new prognostic model integrating clinicopathologic variables, gene expression and mutation data as the input. We used testing datasets to show that the model achieved a higher level of accuracy of EFS prediction than models based on any other input data integrations. Adding interaction covariates to prognostic models showed limited benefits on improving prediction power. We further compared at the level of risk score computed from these models. The univariate model fitting p values of the score indicated that the one generated from the best combinatorial model captured a wider spectrum of EFS-related variations. Moreover, the proposed score provided a higher-resolution EFS stratification for pathologically-defined subgroup of patients, and showed superior time-varying RFS prediction power than conventional clinicopathologic methods (mean time-dependent AUC= 0.772 vs. 0.646, p-value< 0.001). The higher-risk subgroup determined by the score was characterized with RNA expression aberration of multiple immune-associated genes and depletion of natural killer cells and mast cells.

Integrating different types of data is an effective way to improve prognostic prediction. In this study, the model integrating clinicopathologic variables, gene expression and mutation achieved the best performance in multiple evaluation metrics. Our conclusions were consistent with previous studies. For example, Chen *et al.*[13] integrated two micro-RNAs, two mRNAs and two DNA methylation sites as prognostic factors associated with OS, and they achieved a more significant risk stratification within pathologically-defined subgroups. Song *et.al* showed that, by integrating genetic mutations and expression profiles with clinicopathologic variables, the prediction of both OS and RFS showed the highest cross-validation accuracy among all the models in the TCGA-LUAD data[15]. Besides, Dong *et al.*[14] found that by adding DNA methylation and gene expression biomarkers to a model using only clinical data as the input, the AUCs improved by 18.3% and 16.4% in discovery and validation phases for early-stage LUAD patients, respectively. We extended some of these studies by introducing more types of input data integrations and stricter evaluation criteria. The resultant model not only showed improved prediction for EFS on the testing dataset, but also demonstrated its significance as a predictor for RFS under a bias-corrected competing risks modeling framework.

Our study also suggested new therapeutic opportunities for the higher-risk patients. For example, we found that gene *ACVR1B,* a tumor suppressor[35], was down-regulated in the higher-risk patients. The enhancement of tumor suppressing function of *ACVR1B* might be one possible strategy for treating high-risk LUAD patients. On the other hand, HSP90AB1 overexpression was found associated with poor prognosis both in our study and other studies[36, 37]. Whether inhibiting such expression could improve current standard therapies for higher-risk subgroup might warrants further studies. Also, we discovered that the exhaustion of innate immunity components were correlated with prognostic risk. It has been reported that the interaction between mast cells and natural killer cells is critical for anti-viral defense[38]. Whether there exists similar interactions important for anti-LUAD effect warrants experimental investigations.

Our proposed score included the expression profile of 13 genes and somatic mutation profile of 10 genes. We recognized that this is a relatively large panel of testing which involves both expression and mutation measurements. However, there already exists multi-panel testing technologies that can be readily translated for the score. For example, a 21-gene expression panel (Oncotype DX) based on qRT-PCR platform has already been made for clinical use to inform breast cancer treatment[39]. For

mutational testing, a 324 gene panel (FoundationOne CDx)[40] based on next generation sequencing (NGS) platform was approved for clinical genetic testing by FDA recently. We thus think the score has potential to be cost-effective with these multi-panel testing technologies.

Our study has limitations. First, the combinatorial models developed in this study were based on features already selected by models based on individual type of input data, using the same training dataset. This introduced more overfitting and could possibly cause the failure of selecting truly important combinatorial models. Second, the feature pre-selection method remains to be improved. We performed univariate Cox analysis to pre-select important features, and this method only provided a slight improvement for models based on single-type variables. Third, the EFS outcome we defined in this study included death from any causes. We recognized that including death not related to lung cancer could bias EFS estimates, but such detail information was not available from TCGA clinical dataset. We mitigated this by further evaluating the score on RFS analysis. Fourth, all analyses were performed on TCGA-LUAD dataset. More external validations should be made before considering clinical translation of the score.

Conclusions

In summary, our study proposed a novel prognostic risk score integrating clinicopathologic variables, gene expression and mutation data for LUAD. The score was useful for both EFS and RFS analyses.

List of abbreviations

LUAD, lung adenocarcinoma; TCGA, The Cancer Genome Atlas; AJCC, American Joint Committee on Cancer; EFS, event-free survival; OS, overall survival; RFS recurrence-free survival; FPKM, Fragments Per Kilobase of transcript per Million mapped reads; CIF, cumulative incidence function; mul-score, the risk score computed by the best combinatorial model; cln-score, the risk score computed by clinicopathologic variables-based model; lasso, least absolute shrinkage and selection operator; I2, the interaction features generated by multiplication of any two features within each type of data or between different types of data; I3, I2 and the interaction features generated by multiplication of any three features between different types of data.

Declarations

Ethics approval and consent to participate

No ethics approval was required for this work. All data in this study are publicly available.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed for the current study are available in the TCGA-LUAD repository: https://portal.gdc.cancer.gov/projects/TCGA-LUAD.

Competing interests

The authors have no competing interests to declare.

Funding
This work is supported by the authors and an internal R&D funding from HuaJia Biomedical Intelligence.

Author's contributions
(I) Conception and design: Guofeng Li, Jialu Li, Bin Peng; (II) Administrative support: Jialu Li, Bin Peng; (III) Collection and assembly of data: All authors; (IV) Data analysis and interpretation: Guofeng Li, Youlong Zhang, Jialu Li, Bin Peng; (V) Manuscript writing: All authors; (VI) Final approval of manuscript: All authors.

Acknowledgements
Not applicable.

Author's information
Guofeng Li, 4625094@qq.com
Guangsuo Wang, 908611104@qq.com
Yanhua Guo, 1789394540@qq.com
Shixuan Li, 371434395@qq.com
Youlong Zhang, youlong.zhang@huajiabio.com
Jialu Li, jialu.li@huajiabio.com
Bin Peng, 183672297@qq.com

Reference

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA Cancer J Clin* 2018, **68:**394-424.
2. Myers DJ, Wallen JM: **Cancer, Lung Adenocarcinoma.** In *StatPearls.* Treasure Island (FL); 2020
3. Crino L, Weder W, van Meerbeeck J, Felip E, Group EGW: **Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up.** *Ann Oncol* 2010, **21 Suppl 5:**v103-115.

4. Reuter JA, Spacek DV, Snyder MP: **High-throughput sequencing technologies.** *Mol Cell* 2015, **58:**586-597.

5. Schwarze K, Buchanan J, Taylor JC, Wordsworth S: **Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature.** *Genet Med* 2018, **20:**1122-1130.

6. Ma B, Geng Y, Meng F, Yan G, Song F: **Identification of a Sixteen-gene Prognostic Biomarker for Lung Adenocarcinoma Using a Machine Learning Method.** *J Cancer* 2020, **11:**1288-1298.

7. Li Y, Ge D, Gu J, Xu F, Zhu Q, Lu C: **A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies.** *BMC Cancer* 2019, **19:**886.

8. Xie H, Xie C: **A Six-Gene Signature Predicts Survival of Adenocarcinoma Type of Non-Small-Cell Lung Cancer Patients: A Comprehensive Study Based on Integrated Analysis and Weighted Gene Coexpression Network.** *Biomed Res Int* 2019, **2019:**4250613.

9. Shi X, Tan H, Le X, Xian H, Li X, Huang K, Luo VY, Liu Y, Wu Z, Mo H, et al: **An expression signature model to predict lung adenocarcinoma-specific survival.** *Cancer Manag Res* 2018, **10:**3717-3732.

10. Shi J, Hua X, Zhu B, Ravichandran S, Wang M, Nguyen C, Brodie SA, Palleschi A, Alloisio M, Pariscenti G, et al: **Somatic Genomics and Clinical Features of Lung Adenocarcinoma: A Retrospective Study.** *PLoS Med* 2016, **13:**e1002162.

11. Cho HJ, Lee S, Ji YG, Lee DH: **Association of specific gene mutations derived from**

machine learning with survival in lung adenocarcinoma. *PLoS One* 2018, **13:**e0207204.

12. La Fleur L, Falk-Sorqvist E, Smeds P, Berglund A, Sundstrom M, Mattsson JS, Branden E, Koyi H, Isaksson J, Brunnstrom H, et al: **Mutation patterns in a population-based non-small cell lung cancer cohort and prognostic impact of concomitant mutations in KRAS and TP53 or STK11.** *Lung Cancer* 2019, **130:**50-58.

13. Chen D, Song Y, Zhang F, Wang X, Zhu E, Zhang X, Jiang G, Li S, Chen C, Chen Y: **Genome-Wide Analysis of Lung Adenocarcinoma Identifies Novel Prognostic Factors and a Prognostic Score.** *Front Genet* 2019, **10:**493.

14. Dong X, Zhang R, He J, Lai L, Alolga RN, Shen S, Zhu Y, You D, Lin L, Chen C, et al: **Trans-omics biomarker model improves prognostic prediction accuracy for early-stage lung adenocarcinoma.** *Aging (Albany NY)* 2019, **11:**6312-6335.

15. Song Y, Chen D, Zhang X, Luo Y, Li S: **Integrating genetic mutations and expression profiles for survival prediction of lung adenocarcinoma.** *Thorac Cancer* 2019, **10:**1220-1228.

16. Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, Porta MG, Jadersten M, Dolatshad H, Verma A, Cross NC, Vyas P, et al: **Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes.** *Nat Commun* 2015, **6:**5901.

17. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.** *Cell* 2018, **173:**400-416

e411.

18. Cancer Genome Atlas Research N: **Comprehensive molecular profiling of lung adenocarcinoma.** *Nature* 2014, **511:**543-550.

19. Tibshirani R: **The lasso method for variable selection in the Cox model.** *Stat Med* 1997, **16:**385-395.

20. Hastie T, Robert Tibshirani, and Martin Wainwright.: **Statistical learning with sparsity: the lasso and generalizations.** *CRC press* 2015.

21. Harrell FE, Jr., Lee KL, Mark DB: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.** *Stat Med* 1996, **15:**361-387.

22. Fine JP, and Robert J. Gray: **A proportional hazards model for the subdistribution of a competing risk.** *Journal of the American statistical association* 1999, **94.446:**496-509.

23. Putter H, Fiocco M, Geskus RB: **Tutorial in biostatistics: competing risks and multi-state models.** *Stat Med* 2007, **26:**2389-2430.

24. Heagerty PJ, Lumley T, Pepe MS: **Time-dependent ROC curves for censored survival data and a diagnostic marker.** *Biometrics* 2000, **56:**337-344.

25. Blanche P, Dartigues JF, Jacqmin-Gadda H: **Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks.** *Stat Med* 2013, **32:**5381-5397.

26. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

27. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.

28. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biol* 2010, **11:**R14.

29. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinformatics* 2013, **14:**128.

30. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

31. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA: **Profiling Tumor Infiltrating Immune Cells with CIBERSORT.** *Methods Mol Biol* 2018, **1711:**243-259.

32. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research.** *Behav Brain Res* 2001, **125:**279-284.

33. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V, Smith T, Campbell J, et al: **ImmPort: disseminating data to the public for the future of immunology.** *Immunol Res* 2014, **58:**234-239.

34. Chen Y, Sullivan C, Peng C, Shan Y, Hu Y, Li D, Li S: **A tumor suppressor function of the Msr1 gene in leukemia stem cells of chronic myeloid leukemia.** *Blood* 2011, **118:**390-400.

35. Loomans HA, Arnold SA, Hebron K, Taylor CJ, Zijlstra A, Andl CD: **Loss of ACVRIB**

leads to increased squamous cell carcinoma aggressiveness through alterations in cell-cell and cell-matrix adhesion proteins. *Am J Cancer Res* 2017, **7:**2422-2437.

36. Liu K, Kang M, Li J, Qin W, Wang R: **Prognostic value of the mRNA expression of members of the HSP90 family in non-small cell lung cancer.** *Exp Ther Med* 2019, **17:**2657-2665.

37. Coskunpinar E, Akkaya N, Yildiz P, Oltulu YM, Aynaci E, Isbir T, Yaylim I: **The significance of HSP90AA1, HSP90AB1 and HSP90B1 gene polymorphisms in a Turkish population with non-small cell lung cancer.** *Anticancer Res* 2014, **34:**753-757.

38. Portales-Cervantes L, Dawod B, Marshall JS: **Mast Cells and Natural Killer Cells-A Potentially Critical Interaction.** *Viruses* 2019, **11**.

39. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, et al: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351:**2817-2826.

40. Ready N, Hellmann MD, Awad MM, Otterson GA, Gutierrez M, Gainor JF, Borghaei H, Jolivet J, Horn L, Mates M, et al: **First-Line Nivolumab Plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer (CheckMate 568): Outcomes by Programmed Death Ligand 1 and Tumor Mutational Burden as Biomarkers.** *J Clin Oncol* 2019, **37:**992-1000.

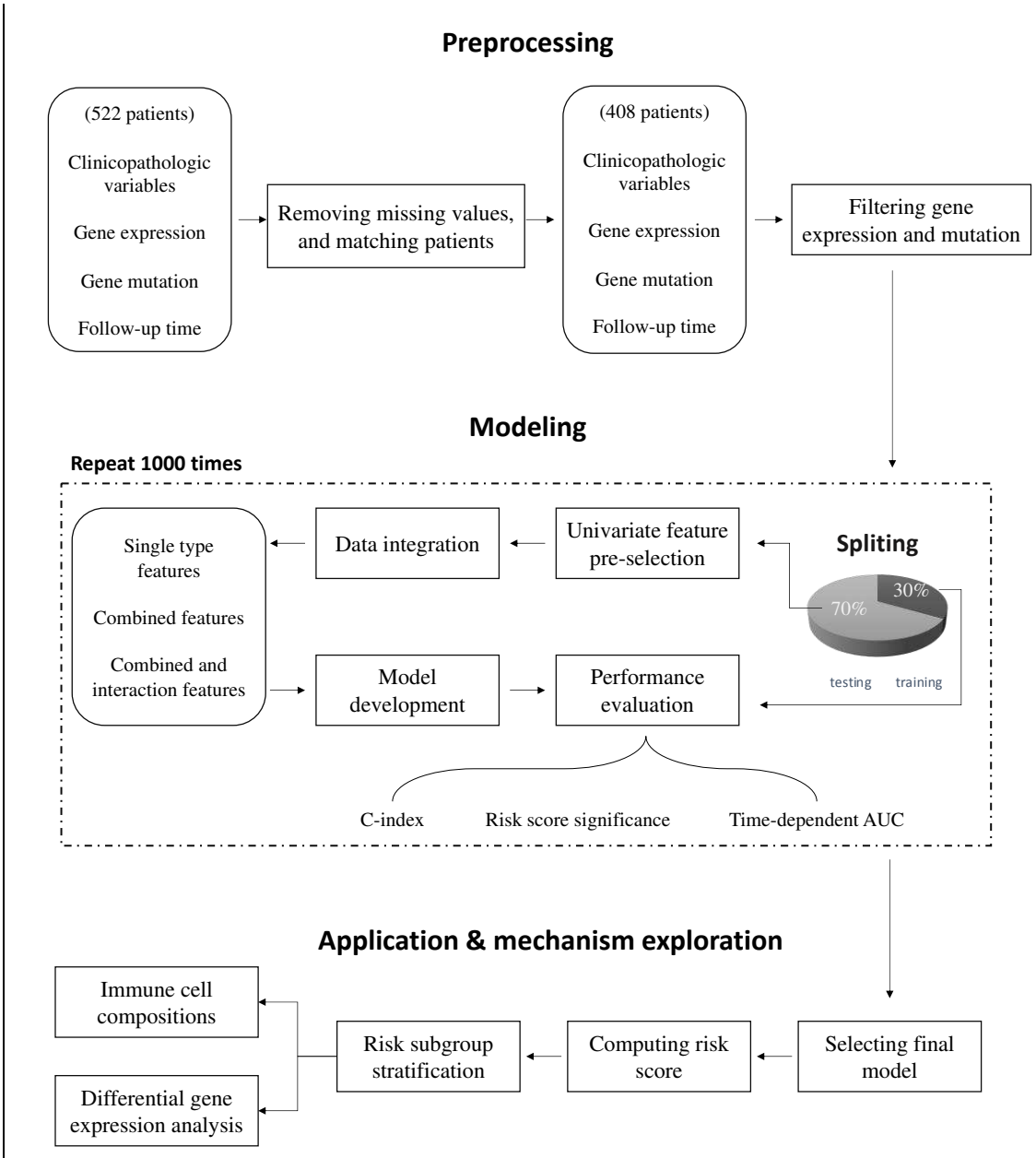
Preprocessing
(522 patients)
Clinicopathologic variables
Gene expression
Gene mutation
Follow-up time
Removing missing values, and matching patients
(408 patients)
Clinicopathologic variables
Gene expression
Gene mutation
Follow-up time
Filtering gene expression and mutation
Modeling
Repeat 1000 times
Single type features
Combined features
Combined and interaction features
Data integration
Univariate feature pre-selection
Spliting
30%
70%
testing
training
Model development
Performance evaluation
C-index
Risk score significance
Time-dependent AUC
Application & mechanism exploration
Immune cell compositions
Differential gene expression analysis
Risk subgroup stratification
Computing risk score
Selecting final model

Figure 1: Workflow used in this study.

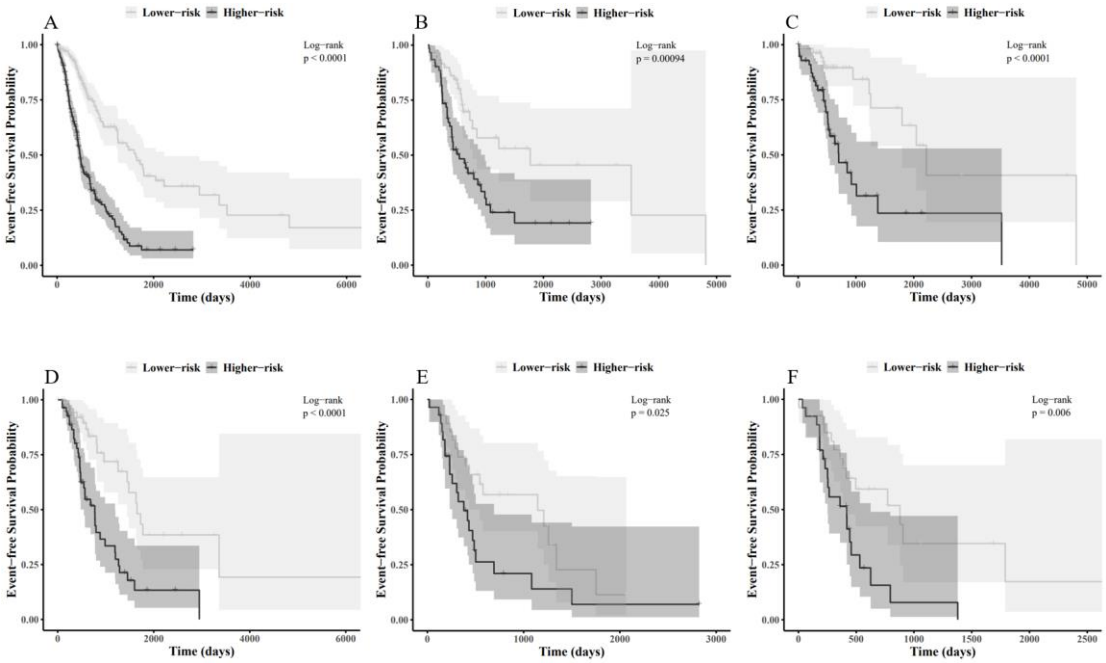

Figure 2: Kaplan-Meier curve of higher-risk and lower-risk subgroups as stratified by mul-score within different groups of patients. The sets from A to F are: A, all patients; B, testing set; C, patients in AJCC pathologic tumor stage IA; D, patients in AJCC pathologic tumor stage IB; E, patients in AJCC pathologic tumor stage IIB; F, patients in AJCC pathologic tumor stage IIIA.

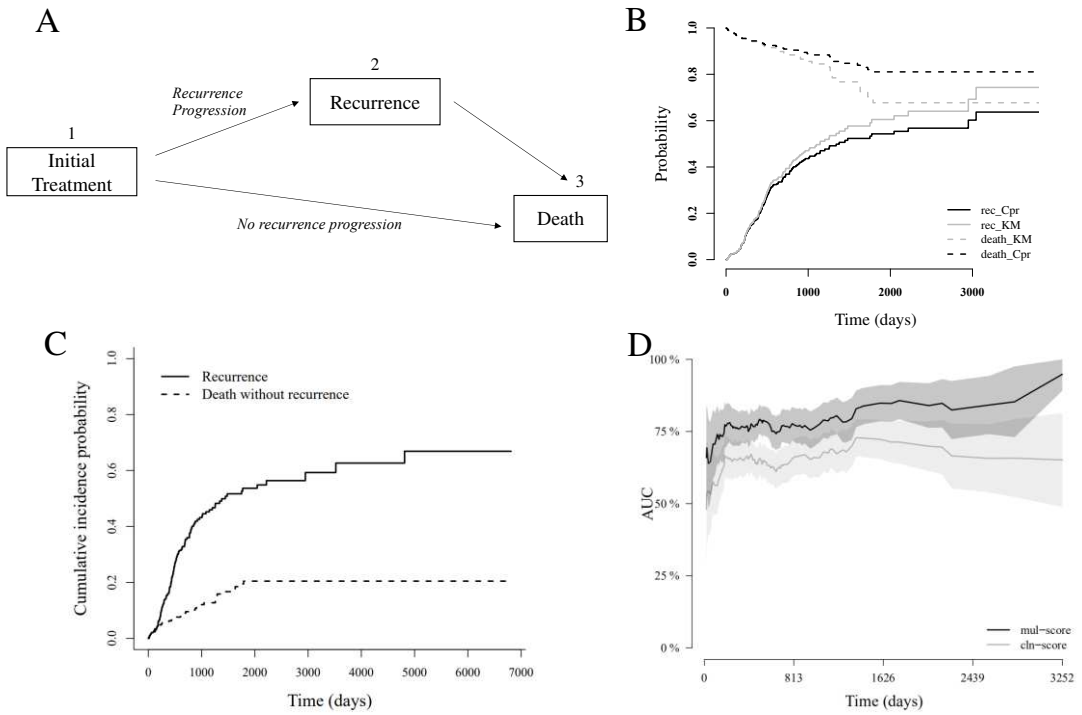

Figure 3: RFS analysis results. A, sketch of disease process for patients after initial treatment; B, estimated cumulative probabilities for recurrence, and estimated survival (1- cumulative probability) for death without recurrence, using either competing risks modeling (Cpr) or naïve KM modeling (KM); C, the cumulative probability of developing recurrence and death without recurrence; D, comparison of time-dependent AUCs of risk scores for RFS prediction.

Additional files

Additional file 1

Table S1: Summary of distribution and EFS association of clinicopathologic variables and genes with top 10 mutation frequencies used in the study.

Table S2: Description of input feature sets and the size of feature space.

Table S3: The summary of C-index, related to Figure S2.

Table S4: The name and coefficient of features of the selected combinatorial model.

Table S5: Summary of p values of the coefficients in models fitted with single or multiple risk scores on the testing datasets.

Table S6: Summary of model fitting results (likelihood ratio test) and median survival time (in days) of higher- and lower-risk subgroups within each sets of patients.

Table S7: Summary of 12 immune-related genes commonly identified as differentially expressed within all patients, stage IIB and IIIA subgroups.

Figure S1: The Kaplan-Meier curve for EFS of 408 patients enrolled in this study.

Figure S2: The C-index of models developed based on single type of data (A) combined feature sets (B) and models with interaction covariates (C). The white box summarized C-index for models selected from cross-validation (1 standard error rule), while the gray box for those from testing. Abbreviations can be referred to Table S2.

Figure S3: Kaplan-Meier curve of higher- and lower-risk subgroups stratified by cln-score within different sets of patients. The sets from A to F are: A, all patients; B, testing set; C, patients in AJCC pathologic tumor stage IA; D, patients in AJCC pathologic tumor stage IB; E, patients in AJCC pathologic tumor stage IIB; F, patients in AJCC pathologic tumor stage IIIA.

Figure S4: Barplot summary of inferred relative fractions of cell types (A) and volcano plot summary for the significance of difference in immune cellular compositions between the higher and lower-risk subgroup patients (B).

Figures

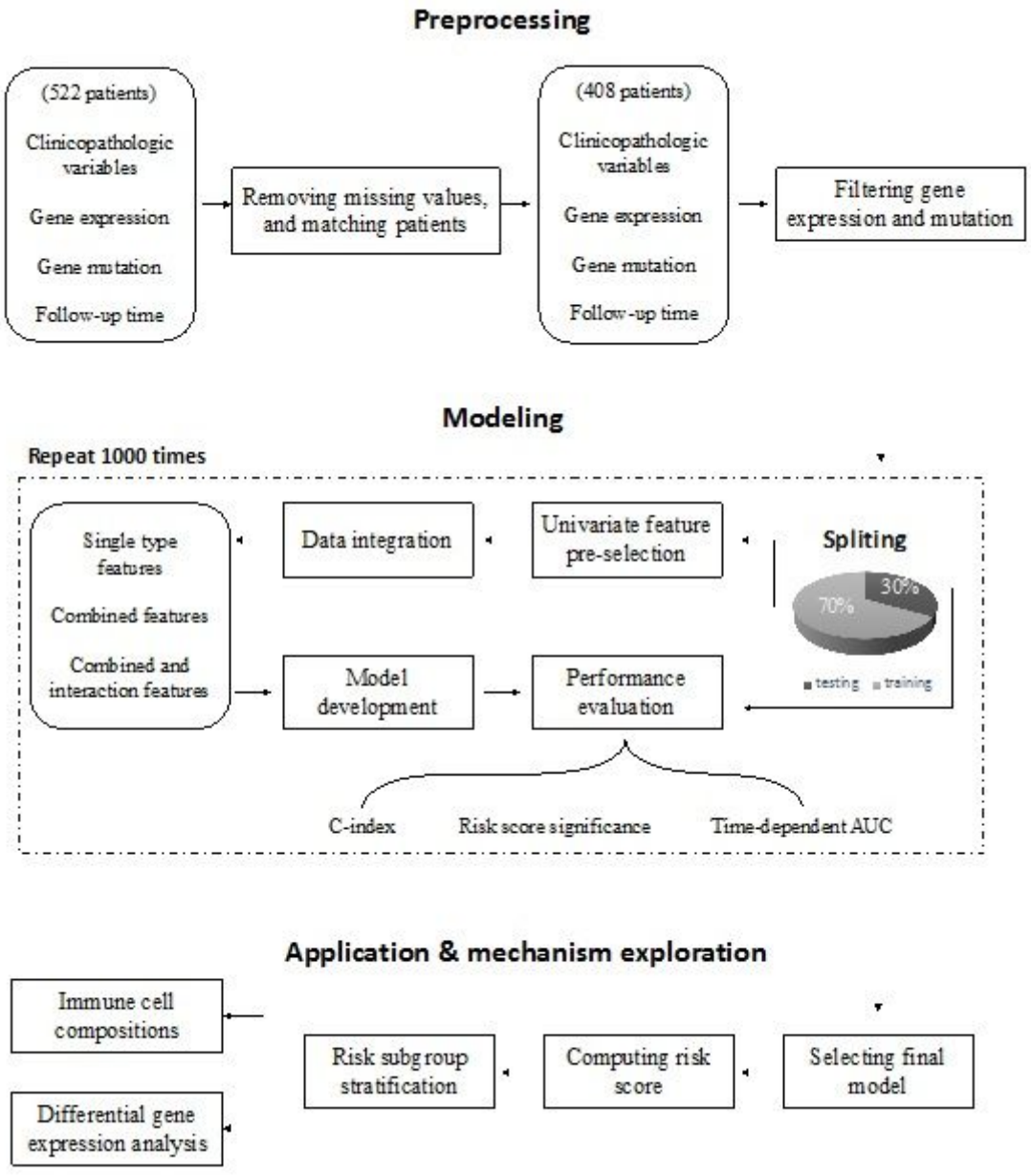

Figure 1

Workflow used in this study.

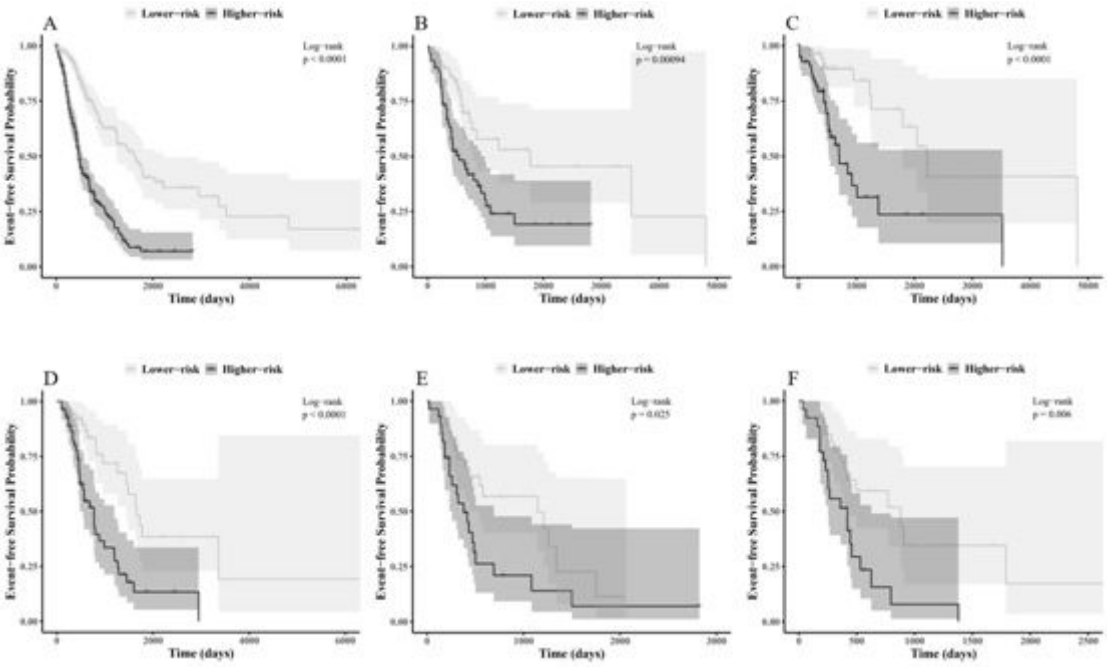

Figure 2

Kaplan-Meier curve of higher-risk and lower-risk subgroups as stratified by mul-score within different groups of patients. The sets from A to F are: A, all patients; B, testing set; C, patients in AJCC pathologic tumor stage IA; D, patients in AJCC pathologic tumor stage IB; E, patients in AJCC pathologic tumor stage IIB; F, patients in AJCC pathologic tumor stage IIIA.

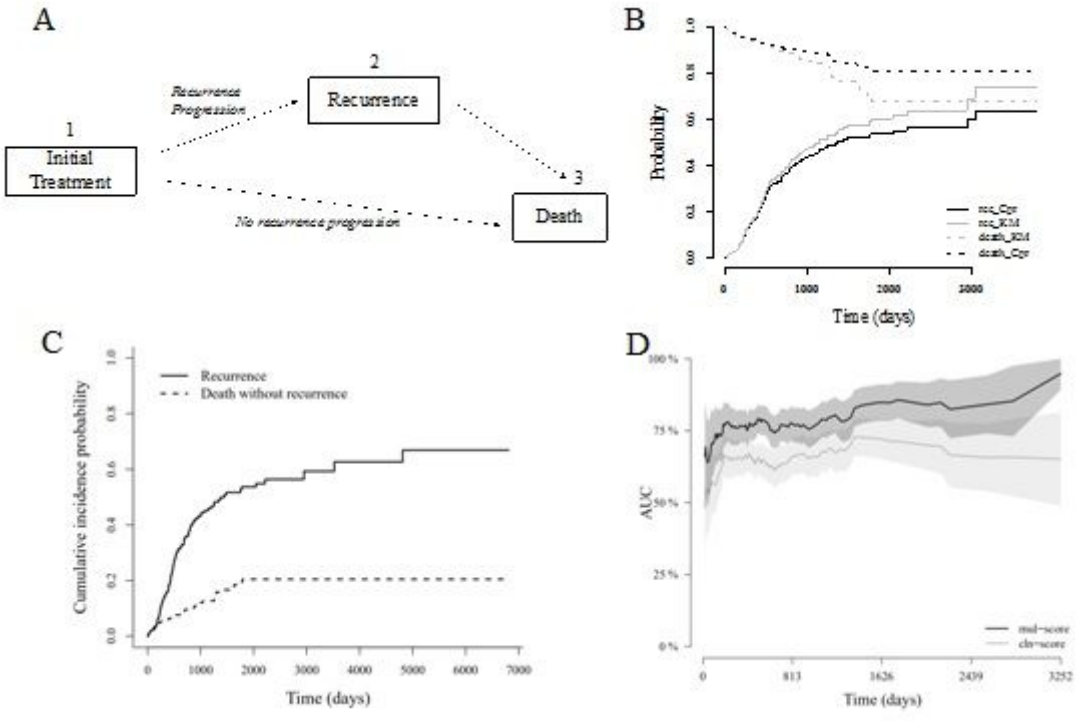

Figure 3

RFS analysis results. A, sketch of disease process for patients after initial treatment;B, estimated cumulative probabilities for recurrence, and estimated survival (1- cumulative probability) for death without recurrence, using either competing risksmodeling (Cpr) or naïve KM modeling (KM); C, the cumulative probability of developing recurrence and death without recurrence; D, comparison of time-dependent AUCs of risk scores for RFS prediction.

Supplementary Files

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

- Additionalfile.docx
- cover.docx