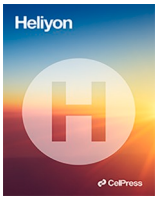
Heliyon 10 (2024) e38527

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Enhanced prognostic signature for lung adenocarcinoma through integration of adjacent normal and tumor gene expressions

Mingyue Hao [a,1], Dandan Li [b,1], Weihao Chen [c,1], Ming Xiong [a], Xinkun Wang [d,***], Yuanyuan Qiao [a,**], Wei Ma [a,*]

[a] Senior Department of Otolaryngology-Head & Neck Surgery, the Sixth Medical Center of PLA General Hospital, Beijing, 100048, China
[b] Department of Blood Transfusion, Sun Yat-sen Memorial Hospital of Sun Yat-Sen University, Guangzhou, 510515, China
[c] Department of Urology, the Third Medical Center of PLA General Hospital, Beijing, 100039, China
[d] Department of Radiology, the Fourth Medical Center of PLA General Hospital, Beijing, 100048, China

ARTICLE INFO

Keywords:
Prognosis biomarker
Combination
Robust prognosis signature
Lung adenocarcinoma

ABSTRACT

Background: Cancer prognosis-related signatures have traditionally been constructed based on gene expression profiles derived from tumor or normal tissues. However, the potential benefits of incorporating gene expression profiles from both tumor and normal tissues to improve signature performance have not been explored.

Methods: In this study, we developed three prognostic models for lung adenocarcinoma (LUAD) using gene expression profiles from tumor tissues, normal tissues, and a combination (COM) of both, sourced from The Cancer Genome Atlas (TCGA). To ensure comparability, the same workflow was followed for all three models.

Results: When applied to the TCGA LUAD dataset, the tumor-derived model exhibited the best overall performance, except in calibration analysis, where the normal-derived model performed better. The COM-derived model demonstrated intermediate performance. Validation on three independent test datasets revealed that the COM-derived model showed the best performance, while the normal-derived model showed the worst. In overall survival (OS) analysis, the low-risk group defined by the COM-derived model consistently exhibited longer mean survival times. The tumor-derived model did not consistently show this trend, and the normal-derived model produced opposite results. In discrimination analysis, no significant differences were observed. The COM-derived model demonstrated good discrimination ability for short periods, while the tumor-derived model performed better for longer periods. In calibration analysis, both the COM and tumor-derived models had similar absolute prediction errors, which were better than those of the normal-derived model. However, the tumor-derived model tended to underestimate survival

Abbreviations: AIC, Akaike Information Criterion; AUC, area under the curve; BIC, Bayesian Information Criterion; COM, combination; DFS, disease-free survival; GEO, Gene Expression Omnibus; K-M, Kaplan–Meier; LASSO, Least Absolute Shrinkage and Selection Operator; lncRNA, long non-coding RNA; LogTN, log2 ratio of tumor and normal; LUAD, lung adenocarcinoma; NTEAIT, new tumor events after initial treatment; PCA, Principal Component Analysis; PNSC, person neoplasm cancer status; PTOS, primary therapy outcome success; RNA-seq, RNA-sequencing; RS, Risk scores; SFRP1, secreted frizzled related protein 1; SPY, smoking pack years; SUM, sum of tumor and normal; TCGA, The Cancer Genome Atlas; timeROC, time-dependent receiver operating characteristic curve; TP73, tumor protein P73; TPM, transcripts per million.

* Corresponding author. the Sixth Medical Center of PLA General Hospital, NO.6 Fucheng road, Haidian, 100048, Beijing, China.
** Corresponding author. the Sixth Medical Center of PLA General Hospital, NO.6 Fucheng road, Haidian, 100048, Beijing, China.
*** Corresponding author. the Fourth Medical Center of PLA General Hospital, NO.51 Fucheng road, Haidian, 100048, Beijing, China.
E-mail addresses: wangxinkun301@163.com (X. Wang), qiaoyuan75@126.com (Y. Qiao), langmawei@bjmu.edu.cn (W. Ma).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e38527
Received 30 August 2023; Received in revised form 9 June 2024; Accepted 25 September 2024
Available online 26 September 2024

rates. The clinical feature analysis and validation in GSE229705 indicate that the risk score (RS) from the COM model is the most clinically significant. These results demonstrate that the COM model's RS aligns more closely with clinical data, maintaining stable performance and the strongest generalizability.

Conclusions: Overall, the COM-derived model demonstrated the best generalization ability. The superior performance of the tumor-derived model in the TCGA LUAD dataset might be due to overfitting. Our results suggest that appropriate combinations of gene expression data from tumor and normal tissues can enhance the predictive power of prognostic signatures.

1. Background

Cancer is a leading cause of global mortality [1]. In 2020, China accounted for approximately 24 % of new cancer cases and 30 % of cancer-related deaths worldwide [2]. An estimated 4,820,000 new cancer cases and 3,210,000 cancer-related deaths were reported in China [3]. Moreover, the incidence of cancer has been observed in younger populations in contemporary society [1]. Early cancer diagnosis significantly improves survival rates [4], and precise prognostication and appropriate therapeutic protocols contribute to enhanced quality of life and reduced mortality.

Numerous cancer prognosis biomarkers and signatures have been identified over the years. Initially, single-gene biomarkers were commonly employed for early-stage prognosis. For instance, alpha-fetoprotein (AFP) has been widely used for hepatocellular carcinoma (HCC) surveillance [5,6]. Prothrombin induced by vitamin K absence-II (PIVKA-II) has also been used for HCC surveillance due to its independence from AFP secretion [7]. The combination of AFP and PIVKA-II improved diagnostic accuracy slightly compared to either marker alone [8]. In colorectal cancer (CRC), *KRAS* mutations have been associated with increased risks of recurrence, metastasis, and worse overall survival (OS) [9]. Higher *STRIP2* expression levels were correlated with poor prognosis survival of lung adenocarcinoma (LUAD) [10]. Elevated plasma levels of miR-92a and miR-29a have been linked to advanced adenomas in CRC [11]. Specific genetic variations, such as the rs3740194 single nucleotide variation (SNV) in CUGBP Elav-Like Family Member 2 (*CELF2*) and the rs1131636 SNV, have been associated with poor OS of nasopharyngeal cancer [12,13].

In recent years, the development of high-throughput technologies has facilitated the emergence of omics-based multiple gene signatures and biomarkers. Examples include a 17-gene signature from the University of Arkansas for Medical Science that identifies patients with shorter OS and progression-free survival (PFS) in multiple myeloma [14]. SKY92, a 92-gene signature, has been developed to predict prognosis in multiple myeloma patients at any disease stage and treatment course [15]. For breast cancer, several classic prognostic gene-expression-based signatures, including Mammaprint (70-gene based), Oncotype DX (21-gene based), Rotterdam signatures (76-gene based), GGI-97 (97-gene based), PAM50 (50-gene based), and Endopredict (11-gene based) [16–21], have been marketed, with Mammaprint and PAM50 approved by the Food and Drug Administration (FDA). In nasopharyngeal cancer, a formula based on five microRNAs (miR-142–3p, miR-29c, miR-26a, miR-30e, and miR-93) has been developed to calculate OS-related risk scores (RS) [22]. High frequencies of methylation in six genes (WNT Inhibitory Factor 1 (*WIF1*), Ubiquitin C-Terminal Hydrolase L1 (*UCHL1*), Ras Association Domain Family Member 1 (*RASSF1A*), Cyclin A1 (*CCNA1*), Tumor Protein P73 (*TP73*), and Secreted Frizzled Related Protein 1 (*SFRP1*)) have been associated with shorter disease-free survival (DFS) in nasopharyngeal cancer [23]. In LUAD, prognostic signatures include a 15-gene signature based on ferroptosis-related genes [24], a five pyroptosis-related long non-coding RNA (lncRNA) prognostic risk signature [25], and a nine-glycolysis-related gene signature associated with metastasis and OS [26].

Most of these signatures have been developed using tumor tissue expression profiles, with a few derived from plasma. Tumor tissues are surrounded by normal tissues, which play a role in the tumor microenvironment and potentially influence cancer patient outcomes. Several studies have shown that tumor-adjacent normal tissue expression profiles can also be used to build prognostic risk signatures and may even outperform tumor profiles. For instance, an inflammatory gene signature in tumor-adjacent tissue has been identified as a strong predictor of disease progression in LUAD [27]. John R. Lamb et al. suggested that the survival genes in adjacent normal tissues undergo significant changes during tumorigenesis and mostly lose their predictive capability after tumor formation [28]. Stromal PDGFRb expression in prostate tumors and non-malignant prostate tissue could predict prostate cancer survival [29] The transcriptomes of tumor-adjacent normal tissues are more informative than those of tumors in predicting recurrence in colorectal cancer patients [30]. A study suggests that the genetic profile of tumor-adjacent normal tissue is more likely to be associated with recurrence than intratumoral gene expression following adjuvant chemoradiotherapy in patients with rectal cancer [31].

For complex diseases, single-gene biomarkers often lack sufficient predictive power. Multi-gene expression-based signatures have been developed using tumor tissue expression profiles or adjacent-normal tissue expression profiles, with a few derived from plasma. Integrating gene expression profiles of adjacent normal and tumor tissues in prognostic signature identification may improve predictive performance.

In this study, we tested this hypothesis in LUAD. We constructed three prognostic signatures based on tumor tissue, normal tissue, and a combination (COM) of both, using data from The Cancer Genome Atlas (TCGA) LUAD cohort followed the same workflow. The signatures were validated in four Gene Expression Omnibus (GEO, https://ncbi.nlm.nih.gov/geo/) datasets [32]. Our results demonstrate that the COM-derived signature exhibited superior stability.

2. Materials and method

2.1. Data collection

The preprocessed RNA-sequencing (RNA-seq) data of the LUAD cohort from TCGA were obtained from the UCSC Cancer Browser (UCSC Xena, https://xenabrowser.net/datapages/) [33]. The cohort comprised 517 tumor samples and 59 normal samples, with 58 tumor-normal paired samples. Corresponding clinical data were also retrieved. After matching the clinical and RNA-seq data, a total of 576 samples with both RNA-seq and clinical information were included. The length of human species genes was obtained from the BioMart tool of the Ensembl database (Ensembl, https://asia.ensembl.org/index.html) [34]. The RSEM normalized counts data were used along with gene length information to transform the data into transcripts per million (TPM) expression data. For validation, appropriate cohorts were searched in the GEO database, based on two criteria: (1) availability of OS time and living status data, and (2) a minimum of 15 tumor-normal paired samples. Three cohorts were identified for validation: GSE81089 (199 tumor samples, 19 normal samples, 19 tumor-normal paired samples), GSE102287 (32 tumor samples, 34 normal samples, 25 tumor-normal paired samples), and GSE31210 (226 tumor samples, 20 normal samples, 15 tumor-normal paired samples). GSE81089 cohorts were obtained through high throughput sequencing, while GSE102287 and GSE31210 were based on the Affymetrix Human Genome U133 Plus 2.0 Array (GEO accession number GPL570) platforms. Due to the original expression profiles following a negative binomial distribution, we log2 transformed the data (log (TPM+1,2)). Then the four datasets were integrated and normalized using ComBat batch correction through the SVA R package. In the subsequent analysis, the batch-corrected expression profile data of these four datasets were used. During the review process after submission, a study titled "Inflammation in the Tumor-Adjacent Lung as a Predictor of Clinical Outcome in Lung Adenocarcinoma" was published. The data uploaded to the GEO database for this study is GSE229705, which includes sequencing data for 143 paired LUAD tumor and normal samples, totaling 246 samples. These samples contain prognostic data on progression and recurrence within five years. To better validate our hypothesis, we also used this dataset for validation. This dataset was not batch-corrected with the other datasets, and based on the normalized counts data, we calculated TPM and performed a log2 transformation. All data used in this study were publicly accessible, and ethics committee approval was not required. The publication guidelines and policies of the TCGA and GEO databases were strictly adhered to.

2.2. Gene selection model construction

To ensure comparability, model construction was based on only 58 tumor-normal paired samples. Two methods were employed to combine tumor and normal data: the sum of tumor and normal (SUM, Equation (1)) and the log2 ratio of tumor and normal (LogTN, see Equation (2)). Univariable Cox regression analysis was performed for each gene to explore OS-related genes in both tumor and normal samples. For the COM, genes with a univariable Cox regression *P*-value lower than 0.05 in both tumor and normal samples were initially selected. Then, genes with a consistent sign of the coefficient between tumor and normal were chosen for the SUM, while genes with opposite signs of the coefficient were selected for the LogTN. The selected genes from the SUM and LogTN methods were combined as candidate genes for the COM model. Similarly, for the tumor and normal models, the top 50 significant genes from the univariable Cox regression analysis were selected as candidate genes. Model construction for tumor, normal, and COM followed the same workflow. To reduce dimensionality and select genes from the candidate gene set, Least Absolute Shrinkage and Selection Operator (LASSO) Cox regression was employed. Stepwise regression was then used to build the OS prognosis models. LASSO-Cox regression was performed using the SIS R package, while stepwise regression was performed using the "step" function in R.

2.3. Model validation

Model performance validation was carried out using three GEO cohorts: GSE31210, GSE81089, and GSE102287. To compare the performance of the three models, survival rates, discrimination, and calibration analyses were conducted. Survival curve analysis was calculated using the Kaplan–Meier (K-M) method, and differences between survival curves were assessed using the log-rank test via the survival R package. For discrimination analysis, the time-dependent receiver operating characteristic curve (timeROC) was employed to evaluate the discrimination of models using the timeROC R package. The area under the curve (AUC) was used to assess discrimination, and differences between ROC curves were tested using the "compare" function in the timeROC R package. Calibration analysis was performed using the survival R package, evaluating calibration performance based on the absolute difference values between predicted and calculated survival rates. In addition, we used the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare the three models. We also applied the three models to a new dataset, GSE229705, to assess their effectiveness in predicting tumor recurrence and progression.

2.4. Statistical analysis

Continuous variables were analyzed using the Student's t-test, while Fisher's exact test was used for categorical variables. All statistical analyses were performed using R language version 4.1.1, a free language environment for statistical computing and graphics available at https://www.r-project.org/[35].

$$SUM = Tumor + Normal \tag{1}$$

$$LogTN = \log 2\left(\frac{\widehat{2}Tumor + 1}{\widehat{2}Normal + 1}\right) \quad (2)$$

3. Result

3.1. Characteristics of the four cohorts

In the TCGA LUAD cohort, there are 517 tumor samples (502 with OS), 59 normal samples (59 with OS), and 58 paired samples (58 with OS). In the GSE81089 dataset, there are 199 tumor samples (198 with OS), 19 normal samples (19 with OS), and 19 paired samples (19 with OS). In the GSE102287 dataset, there are 32 tumor samples (32 with OS), 34 normal samples (34 with OS), and 25 paired samples (25 with OS). In GSE31210, there are 226 tumor samples (226 with OS), 20 normal samples (15 with OS), and 15 paired samples (15 with OS). The demographic characteristics of the four cohorts are summarized in Table 1. In the GSE229705 dataset, there are 246 samples, including 123 paired samples. For progression, 68 individuals experienced progression within 5 years, 45 did not experience progression within 5 years, and the status of others is either unknown or their follow-up time is less than 5 years. For recurrence, 58 individuals did not have recurrence within 5 years, 23 experienced recurrence within 5 years, and the status of others is either unknown or their follow-up time is less than 5 years.

Before using ComBat, principal component analysis (PCA) showed that the four datasets clustered separately (Supplementary Fig. 1A). After batch correction, the four datasets clustered together uniformly (Supplementary Fig. 1B).

3.2. Identification of candidate genes

For the SUM combination, genes with a Cox regression *P*-value <0.05 and a consistent risk trend in both tumor and normal samples of LUAD were selected. Specifically, genes were first filtered by Cox regression *P*-values lower than 0.05 in both tumor and normal samples. Subsequently, genes with the same sign of the Cox regression coefficients in tumor and normal were selected, resulting in 38 genes. For the LogTN combination, genes with a Cox regression *P*-value <0.05 and an opposite risk trend in tumor and normal samples of LUAD were selected. The first step was identical to the SUM combination, and in the second step, 8 genes with opposite signs of the Cox regression coefficients in tumor and normal were selected, leading to a total of 46 candidate genes found in the COM model and summarized in Supplementary Table 1. As shown in Fig. 1, the combination approach decreased the *P*-value of the Cox regression. For the SUM combination, in 35 of the 38 candidate genes, the *P*-values were lower than those of both tumor and normal for each gene (Figs. 1A and 35 genes with SUM minimum). Similarly, for the LogTN combination, all 8 candidate genes showed consistent results for each gene (Fig. 1B, all of the 8 genes with LogTN minimum). This indicates that the SUM and LogTN combinations can enhance the prognostic prediction ability of individual genes. For tumor and normal, the top 50 significant Cox regression genes (identified by the lowest *P*-value) were selected as candidate genes.

3.3. Construction of three prognosis models based on candidate genes of tumor, normal, and COM in TCGA LUAD

To ensure comparability, only 58 tumor-normal paired samples in the TCGA LUAD cohort were used to construct the prognosis models, following the same workflow for each. LASSO Cox regression was employed to identify overall survival prognostic genes from the candidate gene sets. The selected genes were then used to build the primary Cox regression models, followed by stepwise regression to generate the final models. The parameter selection of LASSO Cox during gene selection for the three models is shown in Supplementary Fig. 2. For the COM model, a 7-gene-based model was built, including 6 genes from SUM (*APOD*, *FUT2*, *TUBE1*, *C1orf105*, *HIST1H3I*, and *OPALIN*) 1 gene from LogTN (*COX6B2*) (Fig. 2A). As for tumor and normal, 15-gene-based and 9-gene-based models were constructed, respectively (Fig. 2B and C). RS were calculated for tumor, normal, and COM using each respective model. For COM, $RS_{com} = -0.3156 \times (APOD_n + APOD_t)$ - $0.2852 \times (FUT2_n + FUT2_t)$ - $1.459 \times (TUBE1_n + TUBE1_t) + 2.835 \times (C1orf105_n + C1orf105_t) + 1.294 \times (HIST1H3I_n + HIST1H3I_t) + 12.52 \times (OPALIN_n + OPALIN_t) + 1.8 \times \log2\ ((2\hat{}COX6B2_t + 1)/(2\hat{}COX6B2_n + 1))$. Here, the

Table 1
Demographic characteristics of the four cohorts.

	TGCA LUAD	GSE102287	GSE81089	GSE31210
Samples	576	66	218	246
Tumor samples	517	32	199	226
Normal samples	59	34	19	20
Paired samples	58	25	19	15
Patients	516	41	199	227
Age (mean ± SD)	65.4 (55.4–75.3)	60.9 (50.1–71.7)	67.9 (60.2–75.6)	59.6 (52.2–66.9)
Gender (female/male)	277/239	23/18	103/96	121/106
OS event (%)	183 (57.0 %)	26 (63.4 %)	93 (46.7 %)	35 (15.5)
OS time, (days, mean ± SD)	909.4 (13.5–1805.3)	1596.4 (141.9–3050.9)	1196.5 (542.6–1850.4)	1722.9 (1033.7–2412.2)
OS event in paired samples (%)	26 (44.8)	14 (56.0)	8 (42.0)	1 (6.67)

OS: overall survival; SD: standard deviation.

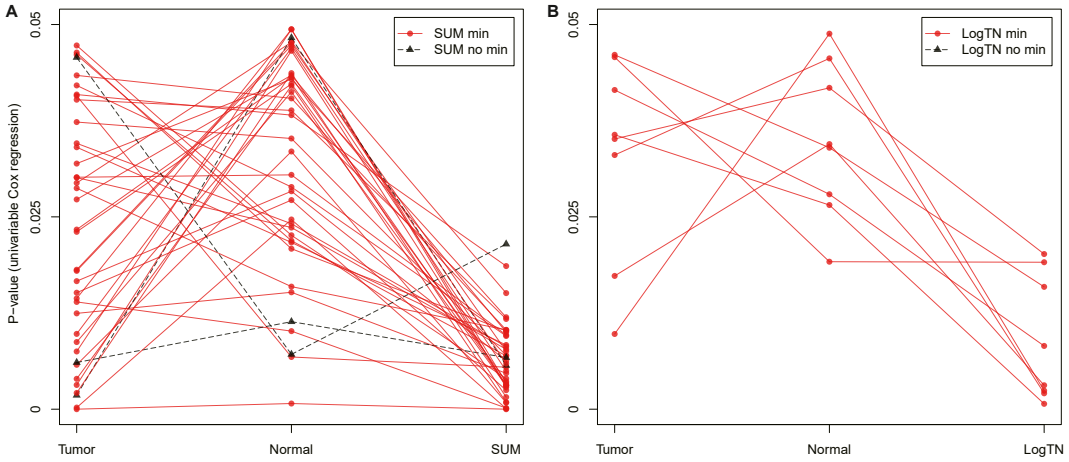


Fig. 1. Influence of combination on prognosis effect in TCGA LUAD. A: 38 candidate gene of SUM combination; B: 6 candidate gene of LogTN combination.

subscript "t" denotes tumor tissue expression, and "n" denotes normal tissue expression. For tumor, $RS_t = -6.221 \times USP4_t + 11.97 \times AIPL1_t + 8.408 \times DSC1_t - 5.343 \times OPRM1_t - 5.292 \times CFL1_t - 3.26 \times RCHY1_t - 1.279 \times LIN7B_t + 4.584 \times MAFK_t - 5.739 \times SLC11A2_t - 4.846 \times ZNF844_t - 6.065 \times RIMS3_t - 1.659 \times BANK1_t + 2.373 \times PSMC5_t - 2.932 \times FBXO8_t - 1.636 \times NTSR_t$. For normal, $RS_n = 61.82 \times ABCG8_n + 35.64 \times TDRG1_n + 74.7 \times C14orf23_n + 27.61 \times NEUROG2_n + 13.27 \times C3orf22_n - 10.29 \times C9orf173_n + 3.338 \times CYP2C18_n - 4.043 \times CABYR_n - 6.585 \times TDH_n$. The 58 patients were divided into two groups (29 in the high-risk group and 29 in the low-risk group) based on the median value of the RS (Fig. 3). Survival analysis between the two groups for the three models is shown in Supplementary Fig. 3. For all three models, the high-risk group had lower overall survival (P-value <0.001). The mean and median overall survival times are summarized in Table 2. Discrimination analysis using timeROC is presented in Fig. 4. TimeROC analysis was performed at three time points: 1 year, 2 years, and 3 years. There was no statistically significant difference between the three models at these time points (Table 3 and Fig. 4D). Calibration analysis was also performed and presented in Fig. 5. The correlation between predicted survival probability and actual survival probability for the three models is depicted in Fig. 5A, B, and 5C. The absolute difference between predicted and actual survival probability for each model is shown in Fig. 5D. COM and normal showed better prediction power than tumor (P-value = 0.0043 and 0.0013). However, there was no difference between COM and normal (P-value = 0.36).

Survival analysis indicates that the three models perform similarly in the TCGA LUAD cohort, which is expected since all three models were built based on the TCGA LUAD data. In terms of event discrimination ability, there is no difference between the three models. However, in terms of survival rate prediction ability, the tumor model performs worse.

3.4. Validation and comparison of the three models in GSE102287, GSE81089, and GSE31210 with paired samples

The three constructed models were applied to the GSE102287, GSE81089, and GSE31210 cohorts for performance validation. To ensure comparability, only tumor-normal paired samples were used. RS were calculated for each model and cohort, and samples were divided into two groups based on the median value of the risk score. OS analysis of the three cohorts is presented in Fig. 6. In GSE102287, the high-risk score group of COM had lower overall survival than the low-risk score group (P-value = 0.016, Fig. 6A). There was no difference between the high-risk score group and the low-risk score group divided by the tumor model (P-value = 0.37, Fig. 6B) or the normal model (P-value = 0.85, Fig. 6C). In GSE81089, none of the three models could distinguish different overall survival time groups (Fig. 6D, E, and 6F). Notably, the normal model displayed an opposite prediction (Fig. 6F). According to the Cox regression model, the high-risk score group should have a lower survival probability than the low-risk score group. However, in the normal model, the high-risk score group had a higher survival probability and longer overall survival time than the low-risk score group (P-value = 0.078, Fig. 6F). In GSE31210, due to the limited number of samples (15 paired samples) and events (just one event), there was no statistically significant difference between the high-risk and low-risk groups (Fig. 6G, H, and 6I). Both the tumor and normal models exhibited opposite predictions. Overall survival analysis of the three models in the three cohorts is summarized in Table 4. From the overall survival results of the three datasets, we can see that although the COM model was significant in only one dataset, its performance was very stable, with high-risk groups consistently showing lower survival rates.

TimeROC analysis requires sufficient samples; thus, it was only performed in GSE102287, as shown in Fig. 7 and Table 5. No statistically significant difference was observed (Table 5). From Fig. 7D, we can see that the COM-derived RS exhibited better discrimination performance in the short time period, while the tumor-derived RS exhibited better performance in the long time period.

Since GSE31210 had only one event, calibration analysis was performed in GSE102287 and GSE81089, as presented in Fig. 8. In GSE102287, the COM model showed the best overall survival probability prediction (COM versus tumor: P-value = 0.87, COM versus normal: P-value <0.001, Fig. 8D). The tumor model showed better prediction ability than the normal model (P-value = 0.002). However, both the tumor and normal models underestimated the survival probability at almost every point (Fig. 8F and G). In

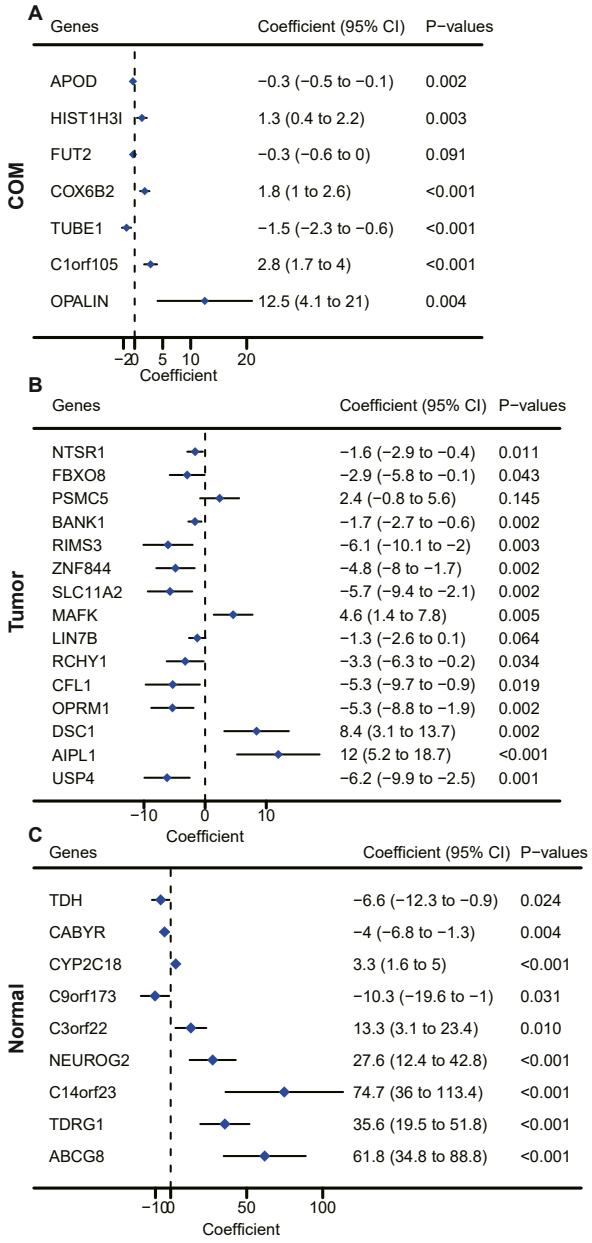


Fig. 2. Forest plot of the three models based TCGA LUAD. A: COM; B: Tumor; C: Normal.

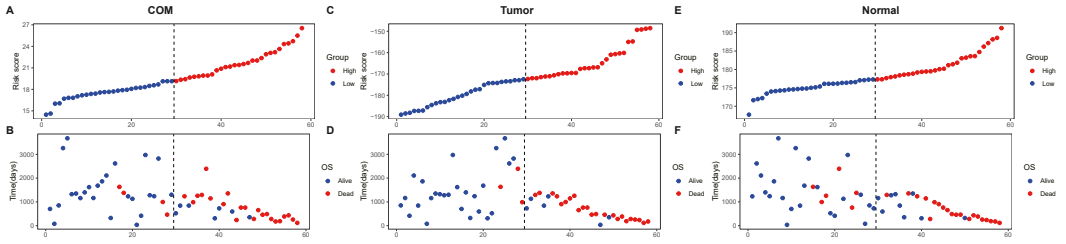


Fig. 3. The high- and low-risk groups divided by the median value of risk score of COM, tumor and normal, and the distribution of the OS and OS status between them in TCGA LUAD. A: COM; B: Tumor; C: Normal.

Table 2
Overall survival analysis of risk score derived from three models in TCGA LUAD.

		COM	Tumor	Normal
Mean ST (days)	High	841	765	785
	Low	3063	3163	2813
95 % LCI of mean ST	High	614	591	607
	Low	2535	2655	2269
95 % UCI of mean ST	High	1452	1037	1052
	Low	3674	3674	3588
median ST	High	761	760	761
	Low	–	–	–
95 % LCI of median ST	High	460	460	460
	Low	–	2393	2393
95 % UCI of median ST	High	1258	1258	–
	Low	–	–	–
Status (alive/dead)	High	7/22	6/23	9/20
	Low	25/4	26/3	23/6

ST: survival time; LCI: lower 95 % confidence interval; UCI: upper 95 % confidence interval.

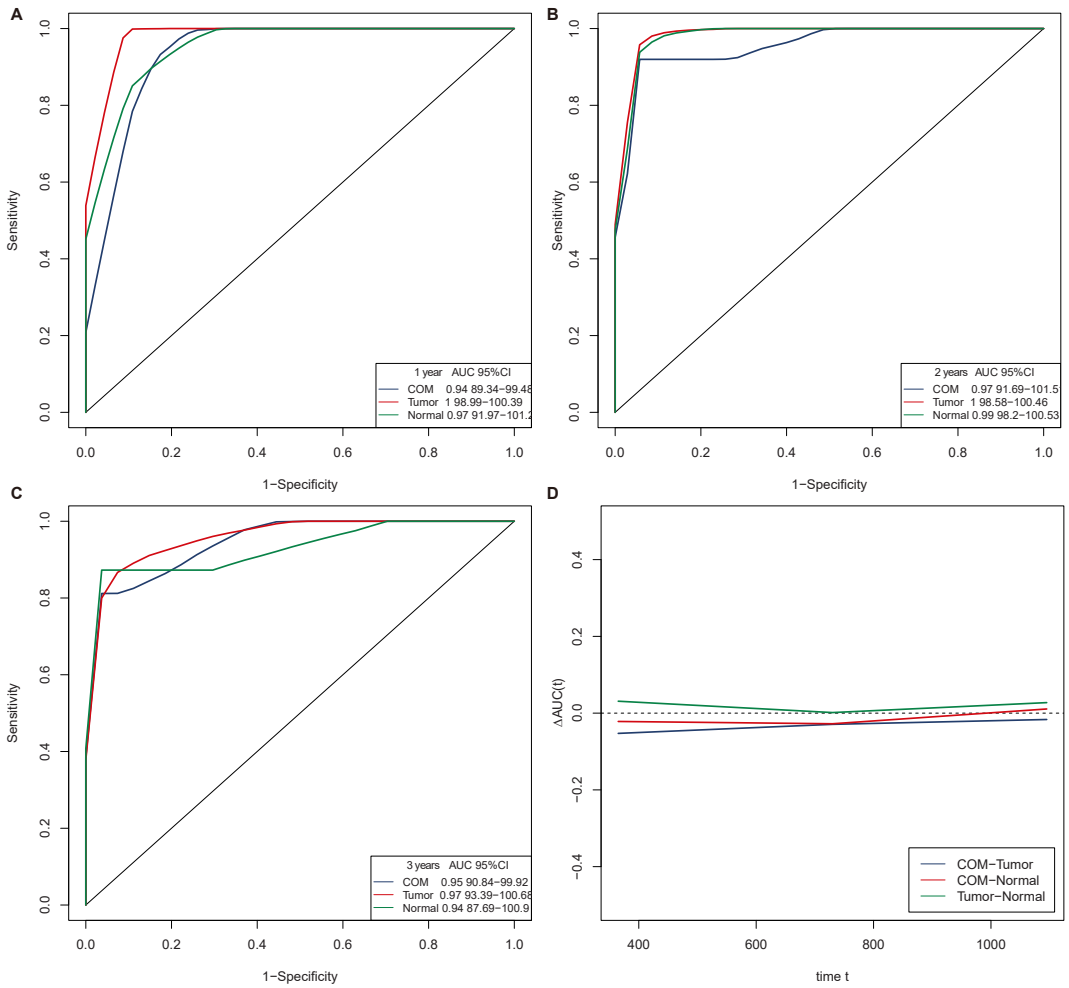


Fig. 4. Time-dependent discrimination ability of the three models in TCGA LUAD. A: ROC curve of the three model at 1 year; B: ROC curve of the three model at 2 years; C: ROC curve of the three model at 3 years; D: Difference of AUC between three models at different time points.

GSE81089, both the tumor and COM models exhibited better prediction than the normal model (tumor versus normal: *P*-value = 0.034, COM versus normal: *P*-value = 0.072, Fig. 8H). Although tumor exhibited better performance than COM in terms of values (*P*-value = 0.69, Fig. 8H), it still underestimated survival probability similarly to the normal model (Fig. 8F and G). In terms of survival

Table 3
P-values of ROC curve between three models at 1, 2, 3 years in TCGA LUAD.

Time points	Nromal vs Tumor	COM vs Normal	COM vs Tumor
1 year	0.26	0.55	**0.07**
2 years	0.87	0.37	0.33
3 years	0.53	0.77	0.48

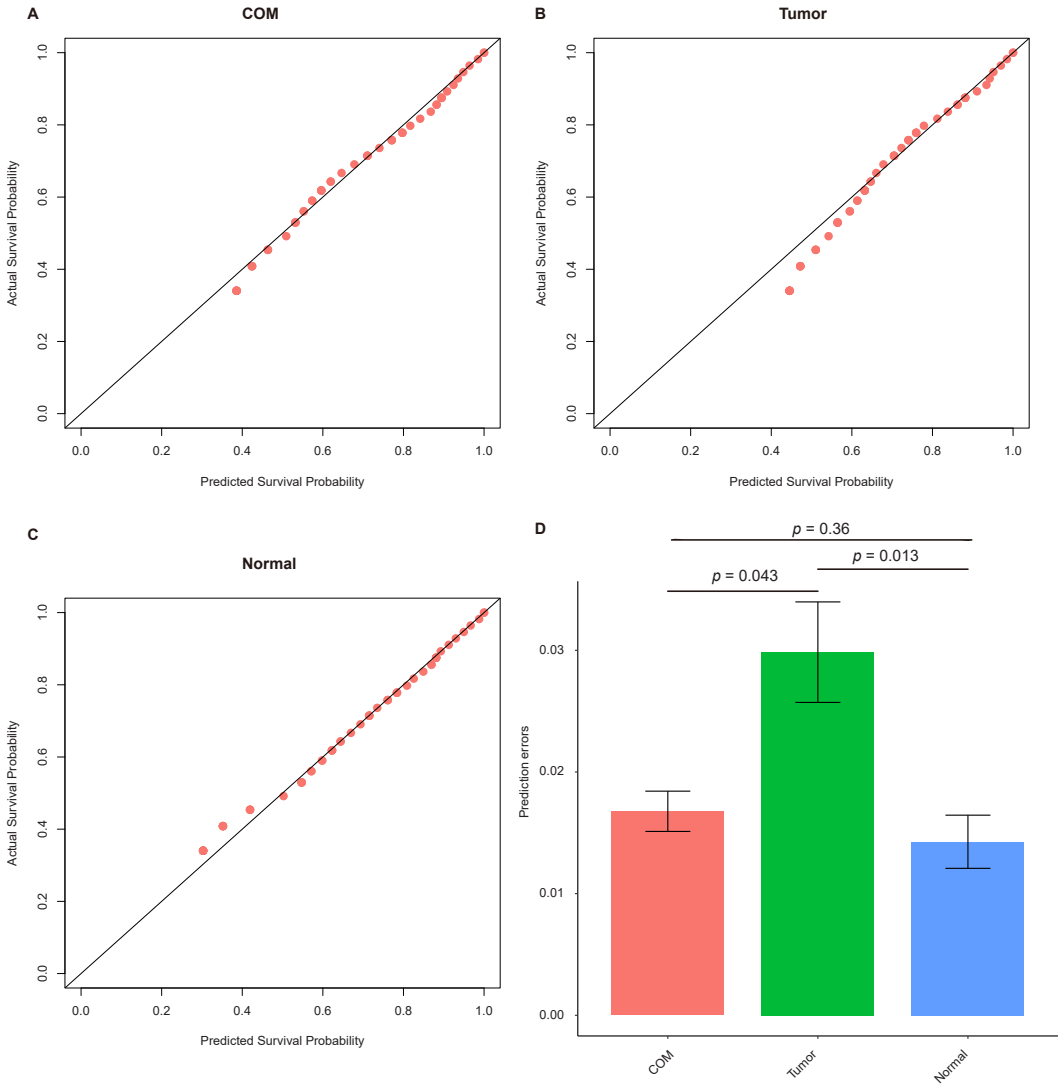


Fig. 5. Calibration analysis of the three models in TCGA LUAD. A, B, C: Plot of predicted and actual survival probability in the models of COM (A), tumor (B), and normal (C); D: The absolute difference between predicted and actual survival probability in the three models.

probability prediction, the tumor and COM models performed similarly and were significantly better than the normal model, but the tumor model tended to underestimate survival probability.

3.5. Validation and comparison of the tumor and normal models in GSE102287, GSE81089, and GSE31210 with all tumor and normal samples

Since in TCGA LUAD, GSE31210, and GSE81089, the normal samples and paired samples are essentially the same, the normal model is no longer analyzed in these three datasets. Therefore, we analyzed the performance of the tumor model across all tumor samples in the four datasets and the performance of the normal model across all normal samples in GSE102287, as shown in

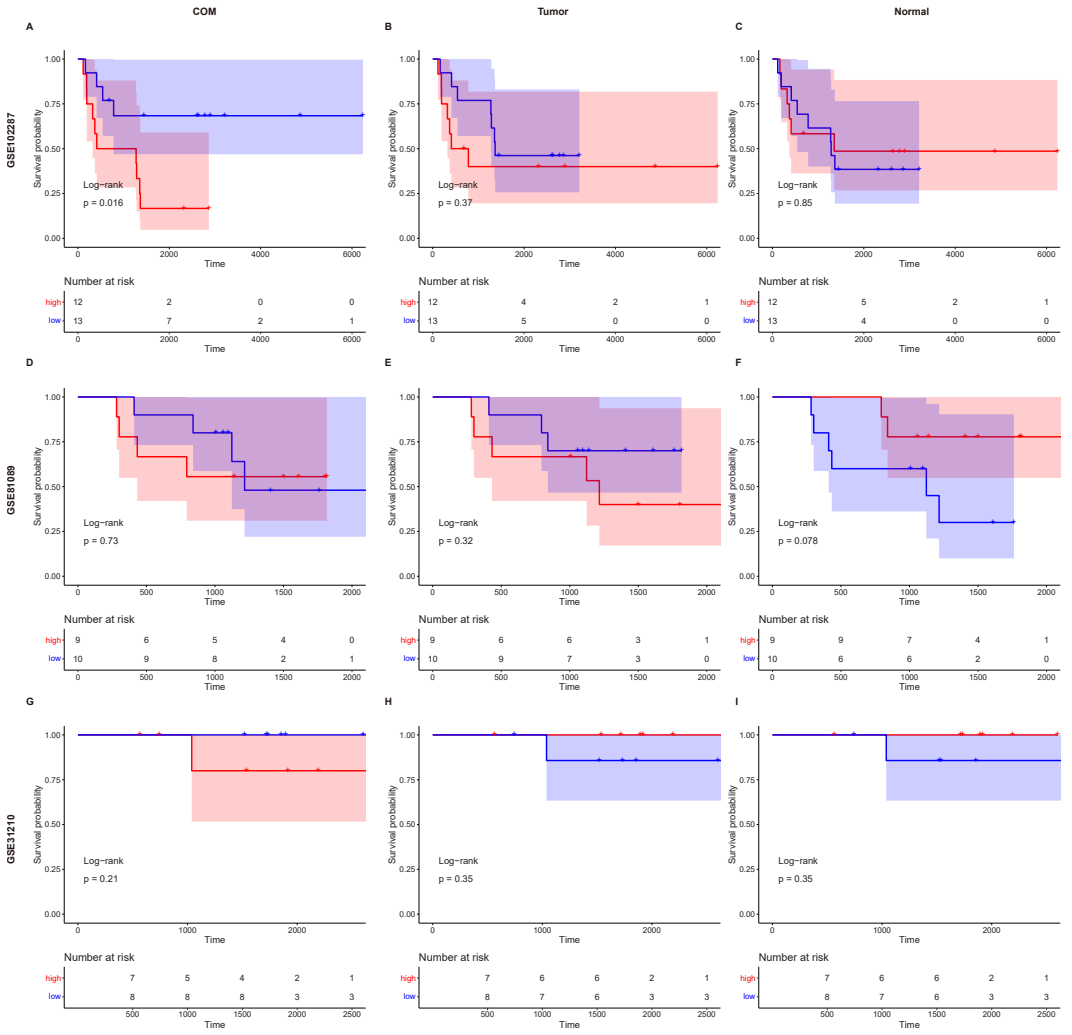


Fig. 6. Survival curve analysis of the three models in the three tested GEO datasets. A, B, C: Models of COM (A), tumor (B), and normal (C) validated in GSE102287; D, E, F: Models of COM (D), tumor (E), and normal (F) validated in GSE81089; G, H, I: Models of COM (G), tumor (H), and normal (I) validated in GSE31210.

Table 4
Overall survival analysis of the three models in the three cohorts.

		GSE102287	GSE81089	GSE31210
P-value (log-rank)	C	**0.016**	0.73	0.21
	T	0.37	0.32	0.35
	N	0.85	**0.078**	0.35
Mean ST (days)	High (C)	1613	1432	2420
	Low (C)	4418	1562	2765
	High (T)	2707	1311	2765
	Low (T)	3371	1755	2518
	High (N)	3286	1904	2765
	Low (N)	2859	1158	2518
Median ST (days)	High (C)	843	–	–
	Low (C)	–	1216	–
	High (T)	595	1216	–
	Low (T)	1366	–	–
	High (N)	1353	–	–
	Low (N)	1285	1123	–

ST: survival time; C: COM; T: tumor; N: normal.

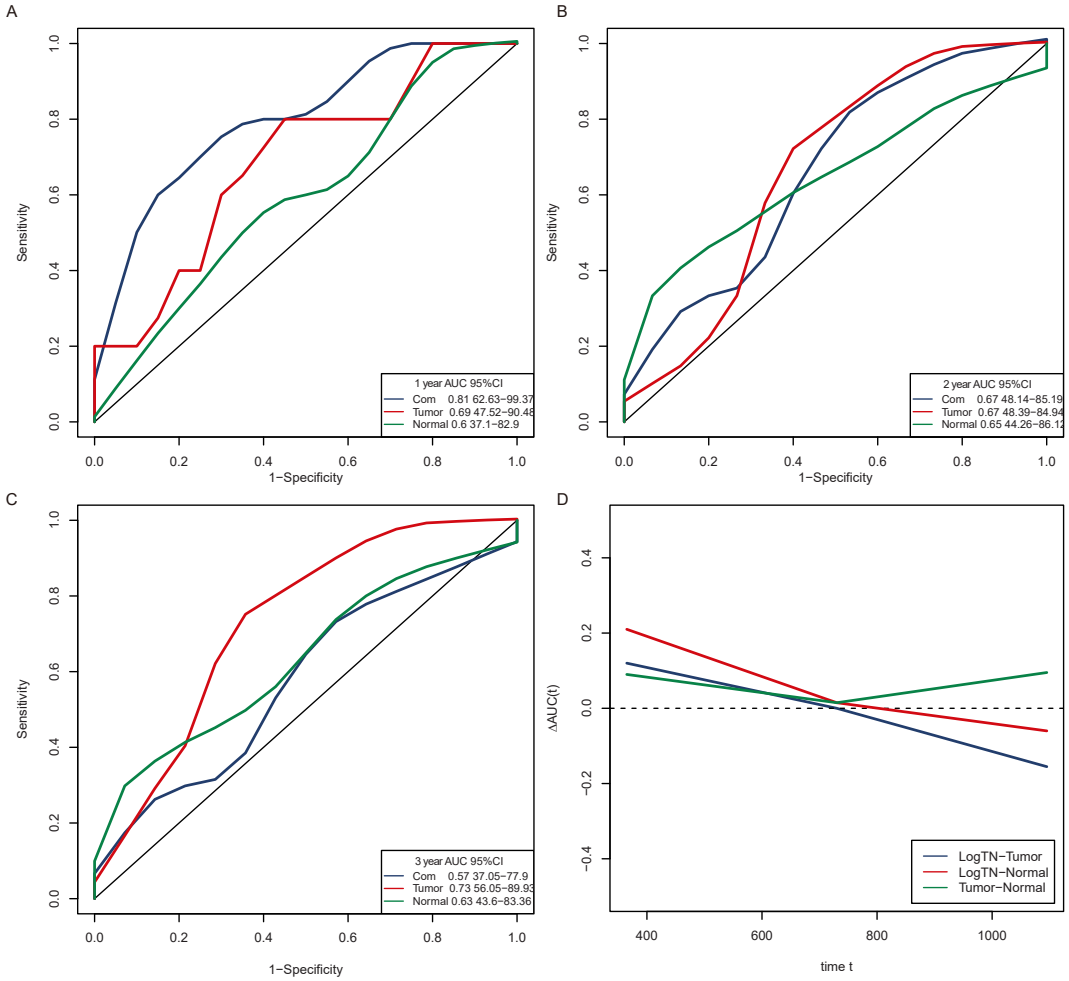


Fig. 7. Time-dependent discrimination ability of the three models in GSE102287. A: ROC curve of the three model at 1 year; B: ROC curve of the three model at 2 years; C: ROC curve of the three model at 3 years; D: Difference of AUC between three models at different time points.

Table 5
P-values of ROC curve between three models at 1, 2, 3 years in GSE102287.

Time points	Normal vs Tumor	COM vs Normal	COM vs Tumor
1 year	0.63	0.30	0.16
2 years	0.92	0.93	1.0
3 years	0.49	0.72	0.28

Supplementary Fig. 4. Except for the tumor model in the TCGA dataset, which demonstrated discrimination and predictive ability (Supplementary Fig. 4A), the tumor model showed no discrimination or predictive ability in the other datasets. Similar to its performance in paired samples, the tumor model tended to underestimate survival probability.

3.6. Effect of proportion of stage I-IV cases on model performance validation

Although model validation results indicate that the COM model is generally superior in stability and generalizability compared to the other two models, the overall validation effect was not satisfactory, and the other two models performed poorly in validation. Therefore, we examined whether the proportions of different pathological stages were similar across the four datasets. As shown in Supplementary Table 2, GSE31210 differs significantly from the other datasets in terms of pathological stage proportions. Fisher's exact test also revealed a significant difference in the proportion of pathological stages between the TCGA dataset and GSE31210 (Supplementary Table 3). Given that GSE31210, with its limited number of paired samples, is not the primary validation dataset, the poor validation results of the tumor and normal models are not due to differences in pathological stage proportions.

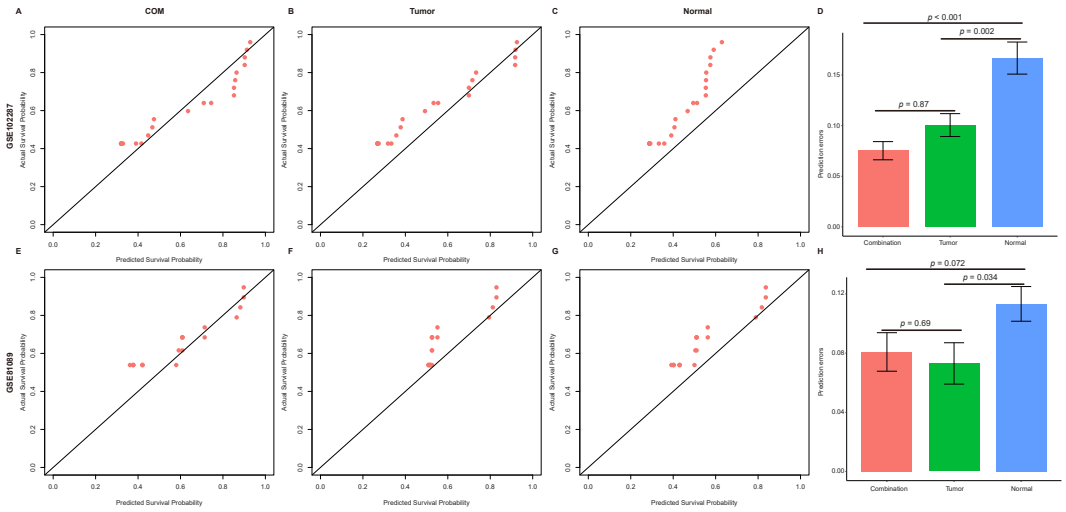


Fig. 8. Calibration analysis of the three models in GSE102287 and GSE81089. A, B, C: Plot for predicted and actual survival probability of COM (A), tumor (B), and normal (C) in GSE102287; D: The absolute difference between predicted and actual survival probability of the three models in GSE102287; E, F, G: Plot for predicted and actual survival probability of COM (A), tumor (B), and normal (C) in GSE81089; D: The absolute difference between predicted and actual survival probability of the three models in GSE81089.

3.7. Association analysis between risk score and clinical characteristics

The association between risk score and clinical characteristics was analyzed in both TCGA LUAD and GSE102287 cohorts. In TCGA LUAD, the COM-derived risk score was higher in males than in females at a significance level of *P*-value <0.1 (*P*-value = 0.060), consistent with known lung cancer statistics indicating higher mortality risk in males [1]. According to the 2023 cancer statistics, it was estimated that there would be 117,550 new cases of lung and bronchus cancer in males and 120,790 in females in 2023, with 67, 160 and 59,910 deaths, respectively [1]. Regarding pathological stage, both the COM and tumor models showed higher risk scores associated with higher-grade stages (*P*-value = 0.017 and *P*-value <0.001, respectively). The higher-risk score groups of both COM and tumor models had more new tumor events after initial treatment (NTEAIT) (*P*-value = 0.035 and *P*-value <0.001, respectively). Additionally, the high-risk score group exhibited more deaths in all three models (Table 6). In terms of person neoplasm cancer status (PNSC), the low-risk score group had more tumor-free samples in all three models, while for primary therapy outcome success (PTOS), the low-risk score group of both COM and tumor models had more successes (Table 6).

In GSE102287, the age of the lower-risk score group in the normal model was higher (*P*-value = 0.020, Table 7), which was a counterintuitive result. Only the risk score of the COM model demonstrated discrimination between alive and dead (*P*-value = 0.015) in terms of overall survival. Although there was a difference between the two groups in the normal model regarding smoking pack years (SPY) (*P*-value = 0.030), it contradicted the actual situation due to the tobacco epidemic's impact on lung cancer deaths [1]. The risk score of the COM model exhibited the best performance for SPY in both cohorts. While all three high-risk score groups had lower SPY in TCGA LUAD, the COM model demonstrated the smallest difference values (Table 6). Conversely, in GSE102287, the high-risk score group of COM had higher SPY, in contrast to the trends observed in both the tumor and normal models (Table 7).

These analyses of clinical characteristics in relation to risk scores indicate that the COM model's risk score aligns more closely with actual clinical data. While the Tumor and Normal models performed adequately in the training dataset (TCGA LUAD), they exhibited anomalous results (e.g., age) or even contradicted clinical reality (e.g., smoking pack years) in the testing dataset (GSE102287). The performance of the COM model remains the most stable and clinically consistent, with the strongest generalization capability.

3.8. Validation in GSE229705

After the initial submission of this study, new LUAD data from GSE229705 was published, which includes 123 tumor-normal paired samples. This dataset contains three types of information: progression type, progression within five years, and recurrence within five years. We selected the progression and recurrence within five years for analysis. We calculated TPM from the normalized counts, performed log2 transformation, and then applied the three risk score calculation formulas to obtain the risk scores for the three models. Comparison revealed that only the COM model's risk score had a discernible ability to differentiate between progression and recurrence (Fig. 9). Unlike the other datasets, GSE229705 was not batch-corrected with TCGA using ComBat, making it entirely new data. The COM model's strong performance on this new dataset was unexpected and impressive. This indicates that the COM model can indeed improve generalization capability.

Table 6
Association analysis between risk score and clinical characteristics in LUAD.

		COM			Tumor			Normal		
		High	Low	P-value	High	Low	P-value	High	Low	P-value
Gender	Female	13	21	**0.060**	16	18	0.790	15	19	0.424
	Male	16	8		13	11		14	10	
Age	Mean	66.4	66.1	0.915	66.4	67.0	0.575	65.7	66.7	0.730
Stage	I	10	19	**0.017**	7	22	**< 0.001**	13	16	0.430
	II-IV	19	9		22	6		16	12	
Status	Live	7	15	**< 0.001**	6	26	**< 0.001**	9	23	**< 0.001**
	Dead	22	4		23	3		20	6	
NTEAIT	Yes	13	4	**0.035**	15	2	**< 0.001**	12	5	0.143
	No	14	22		11	23		15	21	
PNCS	With tumor	17	2	**< 0.001**	19	0	**< 0.001**	16	3	**< 0.001**
	Tumor free	11	25		8	28		12	24	
PTOS	CR	11	22	**0.004**	9	24	**< 0.001**	13	20	**0.061**
	PD/SD	13	4		15	2		13	4	
SPY	Mean	39.0	41.1	0.812	36.7	42.6	0.493	37.5	42.1	0.584

NTEAIT: new tumor event after initial treatment; PNCS: person neoplasm cancer status; PTOS: primary therapy outcome success; SPY: smoking pack years. T-test was used for continuous variables and Fisher's exact test was used for categorical variables.

Table 7
Association analysis between risk score and clinical characteristics in GSE102287.

		COM			Tumor			Normal		
		High	Low	P-value	High	Low	P-value	High	Low	P-value
Gender	Female	5	10	0.111	8	7	0.688	9	6	0.226
	Male	7	3		4	6		3	7	
Age	Mean	61.6	59.2	0.633	58.4	62.1	0.449	54.6	65.7	**0.020**
Stage	I	7	7	1.00	6	8	0.695	6	8	0.695
	II-IV	5	6		6	5		6	5	
Status	Live	2	9	**0.015**	5	6	1.00	6	5	0.695
	Dead	10	4		7	7		6	8	
SPY	Mean	46.8	37.0	0.435	38.7	44.5	0.642	28.1	54.3	0.030

SPY: smoking pack years. T-test was used for continuous variables and Fisher's exact test was used for categorical variables.

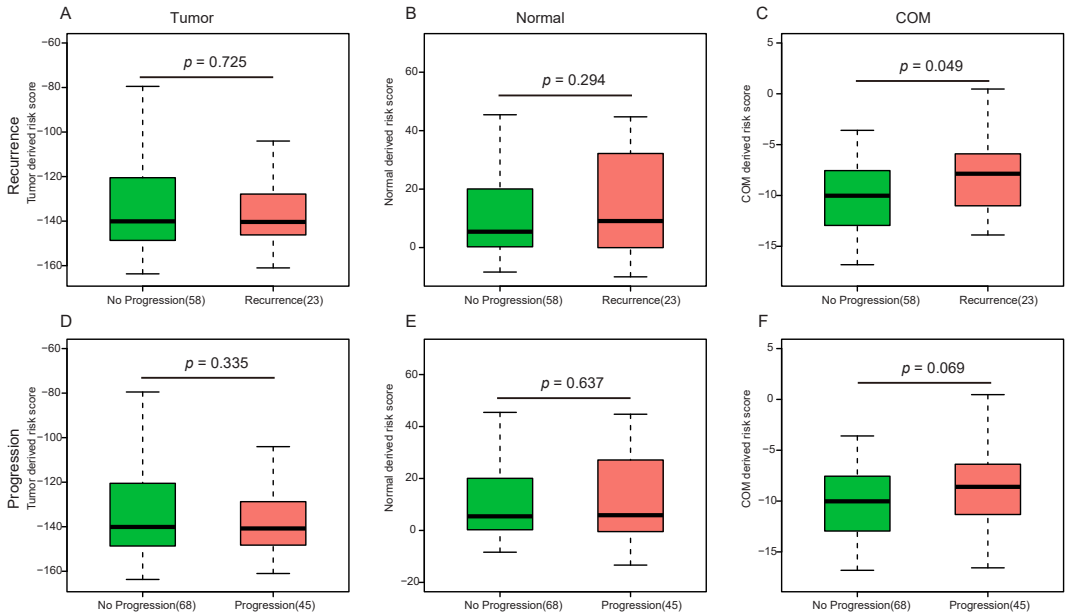


Fig. 9. Validation of the three models in the GSE229705.

3.9. Models assessment via AIC and BIC

AIC and BIC are commonly used methods to evaluate model complexity and goodness of fit. Generally, as the number of parameters in a model increases, the model's goodness of fit also improves. However, once the parameters reach a certain level, additional parameters may no longer provide useful information, causing the improvement in goodness of fit to slow down or even stop. According to the formulas for AIC and BIC, when the increase in goodness of fit does not offset the complexity added by more parameters, both AIC and BIC will increase, indicating the best model at this point. While adding more parameters typically enhances the model's goodness of fit, when these additional parameters fail to provide new information, the increase in goodness of fit reflects overfitting. AIC and BIC can mitigate some overfitting by correcting for model complexity and excluding parameters with little information, but they cannot entirely eliminate overfitting or assess model stability and generalization capability. The goodness of fit in AIC and BIC is based on the training data. If there is a substantial discrepancy between the training and validation data, AIC and BIC are insufficient. To avoid overfitting, it is essential to identify appropriate parameters that show minimal variation and have a similar distribution in both training and other datasets. Building a model with such parameters ensures less overfitting and stronger generalization capability, even if the goodness of fit is slightly lower compared to other parameters. Supplementary Table 4 presents the AIC and BIC values for the three models. It is evident that the COM model has the fewest parameters and the smallest goodness of fit, resulting in the highest AIC and BIC values. Since BIC places a greater weight on the number of parameters, the differences in BIC among the three models are smaller. Although the tumor model appears best based on AIC and BIC, the COM model demonstrates the strongest generalization capability. This indicates that the COM model's fewer parameters provide sufficient stable information, highlighting the severe overfitting present in the Tumor and Normal models.

4. Discussion

When the influence of a gene on overall OS was consistent in both tumor and normal tissues, the SUM combination could increase the effect power. Conversely, when the influence of a gene on OS was opposite in tumor and normal tissues, the LogTN combination could enhance the effect power. In this study, we constructed a COM-based overall survival prognosis model comprising 7 genes using the TCGA LUAD cohort. Among these genes, 6 (*APOD*, *FUT2*, *TUBE1*, *C1orf105*, *HIST1H3I*, and *OPALIN*) were derived from the SUM method, and 1 gene (*COX6B2*) was derived from the LogTN method. Additionally, we built two models based on tumor and normal tissues.

When the three models were applied to the basic cohort TCGA LUAD, the tumor-derived model showed the best performance in overall survival analysis. The difference in mean and median survival times between the high and low-risk score groups was the greatest in the tumor model. The tumor-derived risk score also showed the best discrimination power for time-independent overall survival status, as shown in Table 2. However, in calibration analysis, the tumor-derived model showed the worst overall survival rate prediction ability, while the normal model exhibited the best.

When the three models were used for validation in the three GEO cohorts, the COM-derived model showed the best performance, while the normal-derived model showed the worst. In the overall survival analysis of the three GEO cohorts, the high-risk score group of the normal model consistently had longer mean overall survival time and a higher overall survival rate. The tumor model exhibited a similar pattern in GSE31210. The low-risk score group of the COM model had a higher overall survival rate and longer mean overall survival time in all three GEO cohorts. In GSE102287, the overall survival in the high-risk score group of the COM model was notably lower. Only in GSE81089, the tumor model showed slightly better performance than the COM model. For discrimination, there was no statistically significant difference between the three models in the three GEO cohorts. However, Fig. 7D showed that the COM-derived risk score exhibited better discrimination ability in the short time period, while the tumor and normal-derived risk scores performed better in the long time period. Regarding calibration, in the three cohorts, the COM-derived model showed the best prediction performance. In GSE102287, the COM model produced the smallest prediction error. In GSE81089, although the tumor model produced the smallest prediction error, it still consistently underestimated the survival rates similar to the normal model.

When the tumor model was applied to all tumor samples, it showed some predictive ability only in the TCGA cohort and almost no predictive ability in the three GEO cohorts. Pathological stage analysis indicated that the poor performance of the model in the test datasets was not due to differences in pathological stages. The analysis of clinical characteristics and risk scores showed that the COM model's performance remained the most stable and consistent with actual clinical data, demonstrating the strongest generalization ability. The validation of the three models in GSE229705 further confirmed that the risk score from the COM model was the most clinically relevant, the most stable, and had the strongest generalization ability. Although AIC and BIC analysis suggested that the tumor model had the lowest scores and should be preferred, the actual performance of the COM model was superior.

From a holistic perspective, the COM-derived model demonstrated strong generalization ability. In the TCGA LUAD cohort, which was used for model construction, the tumor-derived model exhibited the best performance. The COM-derived model showed intermediate performance, while the normal-derived model had the worst performance, except for overall survival rate prediction. In the three GEO cohorts, the COM-derived model exhibited the best prediction power. The superior performance of the tumor-derived model in the TCGA LUAD cohort may have been due to overfitting. The combination of tumor and normal samples indeed increased the generalization ability of the model.

However, some limitations existed. First, the number of tumor-normal paired samples was insufficient. For example, in the TCGA LUAD cohort, there were 517 tumor samples, but only 58 paired samples were available. Second, the number of validation cohorts was limited. Although three validation cohorts were used in this study, GSE31210 was almost useless due to the limited number of samples and events. Third, the data quality and preprocessing of the four datasets differed, which could introduce bias.

Nevertheless, this study has demonstrated that the combination of tumor and normal samples in the COM model provides the best generalization ability, the most stable predictive power, and the greatest clinical relevance. Although the combination method used in this study was simple and the predictive performance did not significantly surpass that of the tumor and normal models, it offers a new approach for constructing prognostic models. We believe that better methods for integrating tumor and normal tissue data will emerge in the future, leading to more stable and generalizable models.

5. Conclusions

This study demonstrates that the integration of tumor and normal data enhances the predictive capability and generalizability of prognosis models compared to using tumor or normal data alone. In solid tumors, tumor tissues originate from and are surrounded by normal tissues. The adjacent normal tissues play a significant role in shaping the tumor microenvironment and can influence the clinical outcomes of cancer patients. Traditional prognostic signatures predominantly rely on tumor tissue data, with only a few depending on normal tissue, thus overlooking the valuable insights provided by the combination of both. As a result, the generalizability of tumor-derived or normal-derived prognostic signatures has been limited. Our findings offer a novel perspective for the development of improved prognosis prediction models, potentially leading to enhanced quality of life and increased survival rates for individuals with tumors.

Data availability statement

All data used in this study were open access.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31500756) and Innovation Cultivating Foundation of The Sixth Medical Center of Chinese PLA General Hospital (Grant No: CXPY202007).

CRediT authorship contribution statement

Mingyue Hao: Resources, Validation, Visualization, Writing – original draft. **Dandan Li:** Data curation, Formal analysis, Writing – original draft. **Weihao Chen:** Data curation, Validation, Writing – original draft. **Ming Xiong:** Data curation. **Xinkun Wang:** Project administration, Supervision, Writing – review & editing. **Yuanyuan Qiao:** Project administration, Supervision, Writing – review & editing. **Wei Ma:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e38527.

References

[1] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, CA A Cancer J. Clin. 73 (1) (2023) 17–48, 2023.
[2] W. Cao, H.D. Chen, Y.W. Yu, N. Li, W.Q. Chen, Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020, Chin Med J (Engl) 134 (7) (2021) 783–791.
[3] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, N. Li, W. Chen, Cancer statistics in China and United States, 2022: profiles, trends, and determinants, Chin Med J (Engl) 135 (5) (2022) 584–590.
[4] N. Hawkes, Cancer survival data emphasise importance of early diagnosis, Bmj 364 (2019) l408.
[5] P.R. Galle, F. Foerster, M. Kudo, S.L. Chan, J.M. Llovet, S. Qin, W.R. Schelman, S. Chintharlapalli, P.B. Abada, M. Sherman, A.X. Zhu, Biology and significance of alpha-fetoprotein in hepatocellular carcinoma, Liver Int. 39 (12) (2019) 2214–2229.
[6] F. Trevisani, P.E. D'Intino, A.M. Morselli-Labate, G. Mazzella, E. Accogli, P. Caraceni, M. Domenicali, S. De Notariis, E. Roda, M. Bernardi, Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of HBsAg and anti-HCV status, J. Hepatol. 34 (4) (2001) 570–575.
[7] F.A. Durazo, L.M. Blatt, W.G. Corey, J.H. Lin, S. Han, S. Saab, R.W. Busuttil, M.J. Tong, Des-gamma-carboxyprothrombin, alpha-fetoprotein and AFP-L3 in patients with chronic hepatitis, cirrhosis and hepatocellular carcinoma, J. Gastroenterol. Hepatol. 23 (10) (2008) 1541–1548.
[8] J. Ji, H. Wang, Y. Li, L. Zheng, Y. Yin, Z. Zou, F. Zhou, W. Zhou, F. Shen, C. Gao, Diagnostic evaluation of des-gamma-carboxy prothrombin versus α-fetoprotein for hepatitis B virus-related hepatocellular carcinoma in China: a large-scale, multicentre study, PLoS One 11 (4) (2016) e0153227.
[9] G.A. Margonis, G. Spolverato, Y. Kim, G. Karagkounis, M.A. Choti, T.M. Pawlik, Effect of KRAS mutation on long-term outcomes of patients undergoing hepatic resection for colorectal liver metastases, Ann. Surg Oncol. 22 (13) (2015) 4158–4165.

[10] J. Wang, Y. Yuan, L. Tang, H. Zhai, D. Zhang, L. Duan, X. Jiang, C. Li, Long non-coding RNA-TMPO-AS1 as ceRNA binding to let-7c-5p upregulates STRIP2 expression and predicts poor prognosis in lung adenocarcinoma, Front. Oncol. 12 (2022) 921200.
[11] Z. Huang, D. Huang, S. Ni, Z. Peng, W. Sheng, X. Du, Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer, Int. J. Cancer 127 (1) (2010) 118–126.
[12] Y.M. Guo, J.R. Chen, Y.C. Feng, M.L.K. Chua, Y. Zeng, E.P. Hui, A.K.C. Chan, L.Q. Tang, L. Wang, Q. Cui, H.Q. Han, C.L. Luo, G.W. Lin, Y. Liang, Y. Liu, Z.L. He, Y.X. Liu, P.P. Wei, C.J. Liu, W. Peng, B.W. Han, X.Y. Zuo, E.H.W. Ong, E.L.L. Yeo, K.P. Low, G.S. Tan, T.K.H. Lim, J.S.G. Hwang, B. Li, Q.S. Feng, X. Xia, Y.F. Xia, J. Ko, W. Dai, M.L. Lung, A.T.C. Chan, D.Y.M. Lo, M.S. Zeng, H.Q. Mai, J. Liu, Y.X. Zeng, J.X. Bei, Germline polymorphisms and length of survival of nasopharyngeal carcinoma: an exome-wide association study in multiple cohorts, Adv. Sci. 7 (10) (2020) 1903727.
[13] Y.M. Guo, M.X. Sun, J. Li, T.T. Liu, H.Z. Huang, J.R. Chen, W.S. Liu, Q.S. Feng, L.Z. Chen, J.X. Bei, Y.X. Zeng, Association of CELF2 polymorphism and the prognosis of nasopharyngeal carcinoma in southern Chinese population, Oncotarget 6 (29) (2015) 27176–27186.
[14] J.D. Shaughnessy Jr., F. Zhan, B.E. Burington, Y. Huang, S. Colla, I. Hanamura, J.P. Stewart, B. Kordsmeier, C. Randolph, D.R. Williams, Y. Xiao, H. Xu, J. Epstein, E. Anaissie, S.G. Krishna, M. Cottler-Fox, K. Hollmig, A. Mohiuddin, M. Pineda-Roman, G. Tricot, F. van Rhee, J. Sawyer, Y. Alsayed, R. Walker, M. Zangari, J. Crowley, B. Barlogie, A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1, Blood 109 (6) (2007) 2276–2284.
[15] E.H. van Beers, M.H. van Vliet, R. Kuiper, L. de Best, K.C. Anderson, A. Chari, S. Jagannath, A. Jakubowiak, S.K. Kumar, J.B. Levy, D. Auclair, S. Lonial, D. Reece, P. Richardson, D.S. Siegel, A.K. Stewart, S. Trudel, R. Vij, T.M. Zimmerman, R. Fonseca, Prognostic validation of SKY92 and its combination with ISS in an independent cohort of patients with multiple myeloma, Clin. Lymphoma, Myeloma & Leukemia 17 (9) (2017) 555–562.
[16] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R. M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (6871) (2002) 530–536.
[17] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F.L. Baehner, M.G. Walker, D. Watson, T. Park, W. Hiller, E.R. Fisher, D.L. Wickerham, J. Bryant, N. Wolmark, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med. 351 (27) (2004) 2817–2826.
[18] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, T. Jatkoe, E.M. Berns, D. Atkins, J.A. Foekens, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Lancet 365 (9460) (2005) 671–679.
[19] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M.J. Van de Vijver, J. Bergh, M. Piccart, M. Delorenzi, Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis, J. Natl. Cancer Inst. 98 (4) (2006) 262–272.
[20] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J.F. Quackenbush, I.J. Stijleman, J. Palazzo, J.S. Marron, A.B. Nobel, E. Mardis, T.O. Nielsen, M.J. Ellis, C.M. Perou, P.S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (2009) 1160–1167.
[21] M. Filipits, M. Rudas, R. Jakesz, P. Dubsky, F. Fitzal, C.F. Singer, O. Dietze, R. Greil, A. Jelen, P. Sevelda, C. Freibauer, V. Müller, F. Jänicke, M. Schmidt, H. Kölbl, A. Rody, M. Kaufmann, W. Schroth, H. Brauch, M. Schwab, P. Fritz, K.E. Weber, I.S. Feder, G. Hennig, R. Kronenwett, M. Gehrmann, M. Gnant, A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors, Clin. Cancer Res. 17 (18) (2011) 6012–6020.
[22] N. Liu, N.Y. Chen, R.X. Cui, W.F. Li, Y. Li, R.R. Wei, M.Y. Zhang, Y. Sun, B.J. Huang, M. Chen, Q.M. He, N. Jiang, L. Chen, W.C. Cho, J.P. Yun, J. Zeng, L.Z. Liu, L. Li, Y. Guo, H.Y. Wang, J. Ma, Prognostic value of a microRNA signature in nasopharyngeal carcinoma: a microRNA expression analysis, Lancet Oncol. 13 (6) (2012) 633–641.
[23] W. Jiang, N. Liu, X.Z. Chen, Y. Sun, B. Li, X.Y. Ren, W.F. Qin, N. Jiang, Y.F. Xu, Y.Q. Li, J. Ren, W.C. Cho, J.P. Yun, J. Zeng, L.Z. Liu, L. Li, Y. Guo, H.Q. Mai, M. S. Zeng, T.B. Kang, W.H. Jia, J.Y. Shao, J. Ma, Genome-wide identification of a methylation gene panel as a prognostic biomarker in nasopharyngeal carcinoma, Mol. Cancer Therapeut. 14 (12) (2015) 2864–2873.
[24] Z. Ren, M. Hu, Z. Wang, J. Ge, X. Zhou, G. Zhang, H. Zheng, Ferroptosis-related genes in lung adenocarcinoma: prognostic signature and immune, Drug resistance, mutation analysis, Front. Genet. 12 (2021) 672904.
[25] J. Song, Y. Sun, H. Cao, Z. Liu, L. Xi, C. Dong, R. Yang, Y. Shi, A novel pyroptosis-related lncRNA signature for prognostic prediction in patients with lung adenocarcinoma, Bioengineered 12 (1) (2021) 5932–5949.
[26] L. Zhang, Z. Zhang, Z. Yu, Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma, J. Transl. Med. 17 (1) (2019) 423.
[27] I. Dolgalev, H. Zhou, N. Murrell, H. Le, T. Sakellaropoulos, N. Coudray, K. Zhu, V. Vasudevaraja, A. Yeaton, C. Goparaju, Y. Li, I. Sulaiman, J.J. Tsay, P. Meyn, H. Mohamed, I. Sydney, T. Shiomi, S. Ramaswami, N. Narula, R. Kulicke, F.P. Davis, N. Stransky, G.A. Smolen, W.Y. Cheng, J. Cai, S. Punekar, V. Velcheti, D. H. Sterman, J.T. Poirier, B. Neel, K.K. Wong, L. Chiriboga, A. Heguy, T. Papagiannakopoulos, B. Nadorp, M. Snuderl, L.N. Segal, A.L. Moreira, H.I. Pass, A. Tsirigos, Inflammation in the tumor-adjacent lung as a predictor of clinical outcome in lung adenocarcinoma, Nat. Commun. 14 (1) (2023) 6764.
[28] J.R. Lamb, C. Zhang, T. Xie, K. Wang, B. Zhang, K. Hao, E. Chudin, H.B. Fraser, J. Millstein, M. Ferguson, C. Suver, I. Ivanovska, M. Scott, U. Philippar, D. Bansal, Z. Zhang, J. Burchard, R. Smith, D. Greenawalt, M. Cleary, J. Derry, A. Loboda, J. Watters, R.T. Poon, S.T. Fan, C. Yeung, N.P. Lee, J. Guinney, C. Molony, V. Emilsson, C. Buser-Doepner, J. Zhu, S. Friend, M. Mao, P.M. Shaw, H. Dai, J.M. Luk, E.E. Schadt, Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting, PLoS One 6 (7) (2011) e20090.
[29] C. Hägglöf, P. Hammarsten, A. Josefsson, P. Stattin, J. Paulsson, A. Bergh, A. Ostman, Stromal PDGFRbeta expression in prostate tumors and non-malignant prostate tissue predicts prostate cancer survival, PLoS One 5 (5) (2010) e10747.
[30] J. Kim, H. Kim, M.S. Lee, H. Lee, Y.J. Kim, W.Y. Lee, S.H. Yun, H.C. Kim, H.K. Hong, S. Hannenhalli, Y.B. Cho, D. Park, S.S. Choi, Transcriptomes of the tumor-adjacent normal tissues are more informative than tumors in predicting recurrence in colorectal cancer patients, J. Transl. Med. 21 (1) (2023) 209.
[31] S. Schneider, D.J. Park, D. Yang, A. El-Khoueiry, A. Sherrod, S. Groshen, O. Streeter, S. Iqbal, K.D. Danenberg, H.J. Lenz, Gene expression in tumor-adjacent normal tissue is associated with recurrence in patients with rectal cancer treated with adjuvant chemoradiation, Pharmacogenetics Genom. 16 (8) (2006) 555–563.
[32] The gene expression Omnibus database. https://ncbi.nlm.nih.gov/geo/. (Accessed 2 June 2022).
[33] The UCSC cancer browser. https://xenabrowser.net/datapages/. (Accessed 25 October 2019).
[34] The Ensembl database. https://asia.ensembl.org/index.html. (Accessed 28 August 2019).
[35] The R software. https://www.r-project.org/. (Accessed 24 October 2021).