A Novel Prognostic Nomogram Based on Metastasis-Related Genes Signature for Predicting the Progression-free Interval in Patients With Papillary Thyroid Carcinoma

Rui Liu
Peking Union Medical College Hospital
Mengwei Wu
Peking Union Medical College Hospital
Zhen Cao
Peking Union Medical College Hospital
Xiaobin Li
Peking Union Medical College Hospital
Hongwei Yuan
Peking Union Medical College Hospital
Ziwen Liu (liuziwen@pumch.cn)
Peking Union Medical College Hospital

Research Article

Keywords: papillary thyroid carcinoma prognostic model, The Cancer Genome Atlas Program, metastasis-related genes, nomogram

Posted Date: January 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-140889/v1

Abstract

Background: The recurrence rate for papillary thyroid carcinoma (PTC) after surgery is high, which is a significant issue for patients regarding with low-grade malignancy. We built a novel predictive model with metastasis-related genes (MTGs) and relevant clinical parameters for predicting progression-free interval (PFI) after surgery for PTC.

Methods: We performed a bioinformatic analysis of integrated PTC datasets with the MTGs to identify differentially expressed MTGs (DE-MTGs). Then we generated PFI-related DE-MTGs and established a 14-gene signature using Lasso-Penalty regression. Finally, we established a signature and clinical parameters-based nomogram for predicting the PFI of PTC . We then validated the efficacy of the signature in marking off high risk patients; the nomogram's performance in predicting PFI was also evaluated with receiver operating characteristic (ROC) curve and Harrell's concordance index (C-index).

Results: We identified 155 DE-MTGs related to PFI in PTC. The functional enrichment analysis showed that the DE-MTGs were associated with important oncogenic process. Consequently, we found a novel 14-gene signature. The 14-gene signature could distinguish patients with poorer prognosis and predicted PFI accurately. The signature was a significant independent prognostic factor in PTC. Finally, we built a nomogram by including the signature and relevant clinical factors. Validation analysis showed that the nomogram’s efficacy was superior to the current clinical risk evaluating system in predicting the recurrence of PTC.

Conclusions: The 14-gene signature and nomogram were closely associated with PTC prognosis and may help clinicians improve the individualized prediction of PFI, especially for high-risk patients after surgery.

Background

Thyroid cancer (TC) has become the most commonly diagnosed endocrine tumor over the past decades[1]. Should the recent trends of TC prevail, it may become the fourth most common cancer in the United States by 2030[2]. The most common and least-aggressive histologic type of TC is papillary TC (PTC), comprising 80% of all cases. PTC is characterized by a favorable outcome after adequate surgical removal of the primary tumor and clinically significant lymph nodes[3]. However, one of the primary concerns after the initial surgery is persistent/recurrent disease, which is found in 5–20% of all cases of PTC[4]. If recurrence occurs, patients need to undergo re-operations, which could result in a higher risk of surgical complications[5]. Clinical predictive models such as the risk stratification of American Thyroid Association (ATA) have been widely used[5]. However, the clinical and pathological character-based models developed thus far do not reflect individual characteristics at the molecular level. Therefore, novel prognostic tools for guiding personalized surveillance, especially for patients with a high risk of recurrence, are urgently needed. The development a predictive model based on sensitive biomarkers would facilitate personalized monitoring, which in turn would reduce the possibility of advanced,

recurrent diseases in the postoperative follow-up period. Recently, progression in high-throughput sequencing has led to optimistic expectations about personalized medicine. Signatures based on biomarkers such as mRNA or lncRNA have great potential to predict cancer prognosis[6, 7]. These omics-based models can also reliably predict the prognosis of PTC[8, 9].

Lymph node metastasis (LNM) is one of the major causes of cancer recurrence[10]. Cells progressing through the metastatic cascade must employ a series of diverse cellular processes[11]. Local invasion and intravasation into the bloodstream, survival in the circulation, and growth in new organ environments are primary features of metastasis[12]. TC frequently metastasizes to the lymph nodes of the neck, and the rate of occult nodal metastases in PTC has been reported to be as high as 60–80%[13, 14]. In 60–75% of cases of TC recurrence, the phenomenon occurs in the cervical lymph nodes[15]. Hence, metastasis-related genes (MTGs) based predictive models may be closely related to the metastasis of PTC. Therefore, we previously searched the Human Cancer Metastasis Database (HCMDB), which curates 2183 potential MTGs based on more than 7,000 published pieces of literature[16]. We analyzed four datasets of PTC from Gene Expression Omnibus (GEO) and differentially expressed genes between PTC and normal samples. Then we identified differentially expressed MTGs (DE-MTGs) after the intersection with the experimentally supported MTGs derived from HCMDB. Finally, we proposed a 14-gene signature and constructed a nomogram with relevant clinical factors involved.

Methods

Obtain of TCGA-THCA data and clinical information

We used GDC API to download TCGA-THCA RNA sequencing data up to July 21, 2019, including 507 PTC cases and relevant follow-up information. Transcript per million (TPM) transformation followed by base-2 logarithm normalization was applied. After excluding cases in which PFI was ≤ 30 days, 488 PTC samples with complete follow-up information were finally included in the analysis. We also retrieved clinical and mutational data from the Cbioportal.

Integrated analysis and identification of DE-MTGs

We searched the GEO database to identify DEGs to obtain PTC datasets. The keywords for the searching included "Thyroid cancer", “Homo.sapiens” and "Thyroid carcinoma." Only datasets including PTC samples and normal thyroid samples based on the Affymetrix GPL570 genechip (Santa Clara, CA,USA) were included. Research focused on "cell lines," "xenografts," "poorly differentiated," and "undifferentiated" was excluded. Cases of childhood PTC, PTC in young adults, and radiation-induced PTC were also excluded. Eventually, four independent datasets (GSE29265, GSE33630[17], GSE35570[18], and GSE60542[19]) with 134 PTC tumor samples and 146 normal thyroid samples were enrolled. Especially for dataset GSE35570, PTC samples derived from Chernobyl radiation-exposed patients were excluded. Raw data were normalized using the RMAExpress software[20]. Probe names were transformed to official symbols based on Thermo Fisher Scientific Inc provided annotation file. If more than one probes to single gene symbol,, then the median value was replaced. DEG lists of four datasets were identified

independently with the R package “Limma” with $p < 0.05$, false discovery rate < 0.05, and |Log2FC| >1[21]. Reliable DEGs were then identified from the combination of differential analysis results from four datasets with the "RobustRankAggreg" package of the R software[22]. MTGs was downloaded from the HCMDB. After the intersection with the reliable DEGs, DE-MTGs were generated.

Functional enrichment analysis

We caried out functional enrichment analyses using the "clusterProfiler" package of the R software to explore the potential enriched function of the 155 DE-MTGs[23, 24]. The Benjamini and Hochberg method was used for FDR correction, define adjusted $p < 0.05$ as statistically significant.

Construction and verification of the novel 14-gene signature

The TCGA-THCA dataset was randomly divided into training and validation datasets equally. We used the univariate Cox regression model to identify the DE-MTGs that significantly associated ($p < 0.5$) with PFI in the training-set. Thirty-three DE-MTGs were further included. Then we further applied the LASSO regression analysis, to further select valid variables using the "glmnet" R package[25]. A 14-gene signature was found. Based on the optimal cut-off, the patients were then defined as low- or high-risk. The predictive efficacy of the 14-gene signature was then assessed with ROC curve, K-M analysis and C-index by the "timeROC" package and the "survcomp" package of the R software[26].

Verification of potential relationship between the malignancy and 14-gene signature

The expression pattern of MTG-based gene signature from three dataset (GSE29265, GSE33630, GSE76039[27]) was extracted, and the each sample’s gene risk score was generated to evaluate the potential relationship with malignancy in thyroid cancer. There were 9 ATC samples and 20 PTC samples in GSE29265; 11 ATC samples and 49 PTC samples in GSE33630; 20 ATC samples and 17 PDTC samples in GSE76039. P-value of < 0.05 as statistically significant. Anaplastic thyroid carcinoma (ATC), poorly differentiated thyroid carcinoma (PDTC).

Gene set enrichment analysis (GSEA) of the 14-gene signature

We explored the potential molecular alterations of the 14-gene signature by GSEA[28]. PTC samples (488) from the TCGA-THCA dataset were defined as low- or high-risk by the optimal cut-off value. GSEA v4.1 was then applied to found the biological alteration in high risk group. The gene sets included C2: KEGG, C5: GO, and C6: oncogenic signatures. FDR < 0.05 with |NES| > 1 were considered to indicate significant enrichment.

Independent prognostic parameters in PTC

We performed Cox regression analyses to find the correlated prognostic parameters in PTC. Clinical parameters included age, sex, BRAFV600E mutation, disease TNM stage, extrathyroidal extension, residual tumor, multifocality, anatomic sites of tumors, and histological type. The univariate analysis was

performed first, then the multivariate analysis. Factors with $p < 0.25$ were enrolled in the multivariate analysis to identify independent ones. A p-value of < 0.05 as statistically significant.

Construction of the novel nomogram

After collinearity diagnosis, a novel nomogram for predicting the 1-, 3-, and 5-year PFI of PTC was established incorporating independent and relevant clinical factors. We then evaluated the nomogram's predictive power with ROC curve, C-index, and calibration curve. According to the calculated points, patients were then defined as high or low risk due to the optimal cut-off by X-Tile. Bootstrap method with 1000 resamples generated the C-index. The calibration curve showed the actual and predicted PFI.

Statistical analysis

We used R v4.0.3 and GraphPad Prism 8.0.2 (GraphPad Software, San Diego, California USA) for statistical analysis. Comparison of survival curves were analyzed with Log-rank (Mantel-Cox) test. Comparison of continuous data were analyzed with unpaired t-test. A p-value of < 0.05 as statistically significant.

Results

Identification of reliable DE-MTGs

We conducted the research according to the flowchart shown in Fig. 1. In all, 587, 851, 1716, and 777 DEGs were classified from the GSE29265, GSE33630, GSE35570, and GSE60542, respectively, between PTC versus normal thyroid samples. 702 DEGs including 349 up- and 353 down-regulated were identified with the RRA (Supplementary Table 1). The top 20 up-regulated and down-regulated DEGs as shown in Fig. 2A. We downloaded a list including 1938 experimentally supported MTGs from HCMDB (Supplementary Table 2) to intersect with DEGs. Finally, 155 reliable DE-MTGs were identified, among which 98 were up-regulated and 57 were down-regulated (Fig. 2B, Supplementary Table 3).

Functional enrichment analysis

We analyzed potential function and pathway enrichment of the 155 DE-MTGs (Fig. 3A-D). In terms of BPs, the 155 DE-MTGs were mainly enriched in the cellular matrix organization, extracellular structure organization, and cell-substrate adhesion (Fig. 3A). In terms of CCs, the DE-MTGs identified were significantly enriched in the extracellular matrix, focal adhesion and membrane raft (Fig. 3B). In terms of MFs, the DE-MTGs identified were significantly enriched in receptor-ligand activity, signaling receptor activity, and so on. Pathway analysis further revealed that the 155 DE-MTGs mainly enriched in the proteoglycans in cancer, PI3K-Akt signaling pathways, and so on(Fig. 3D).

Identification and establishment of a 14-gene signature

A total of 488 TCGA PTC cases were enrolled in the PFI analysis. The clinical characteristics are as shown in Table 1. 33 PFI-related DE-MTGs were identified using the Cox proportional-hazards model (Fig. 4). A novel gene signature consisting of the following 14 DE-MTGs was constructed: enhancer of

zeste 2 polycomb repressive complex 2 subunits (EZH2), Kisspeptin-1 receptor (KISS1R), cellular retinoic acid-binding protein 2 (CRABP2), S100 calcium-binding protein A4 (S100A4), H19 imprinted maternally expressed transcript (H19), trefoil factor 3 (TFF3), DEP domain-containing mTOR interacting protein (DEPTOR), serum deprivation response protein (SDPR), aldehyde dehydrogenase 1 family member A1 (ALDH1A1), fibulin 5 (FBLN5), superoxide dismutase 3 (SOD3), LIF receptor subunit alpha (LIFR), FAM3 metabolism-regulating signaling molecule B (FAM3B), and angiotensin II receptor type 1 (AGTR1) (Figure S1). The formula to calculate the risk score was as follows: β1 × gene 1 expression + β2 × gene 2 expression + ····· + βn × gene N expression, where β is the corresponding correlation coefficient. The correlation coefficients were as follows: (-0.172507185), AGTR1; (-0.14855278), ALDH1A1; 0.091493337, CRABP2; 0.011858644, DEPTOR; 1.025547089, EZH2; (-0.602378679), FAM3B; (-0.401443236), FBLN5; 0.064515454, H19; 0.10018325, KISS1R; (-0.009465846), LIFR; 0.06333383, S100A4; (-0.023828277), SDPR; (-0.15551382), SOD3; and (-0.010656468), TFF3. The patients with the high-risk have shorter PFIs by using the Kaplan–Meier analysis with the training, validation, and entire TCGA datasets ($P < 0.0001$; Fig. 5A-C). The correlations between risk scores and recurrences are presented in Fig. 5D-F. In the training-set, the AUCs for PFI prediction were 0.8767 (1 year), 0.8133 (3 year), and 0.8458 (5 year), respectively (Fig. 5G). The C-index was 0.8445 (95% CI: 0.7792–0.9098). In the validation-set, the AUCs were 0.7525, 0.6483, and 0.6109 (Fig. 5H). The C-index was 0.6528 (95% CI: 0.5286–0.7770). In the total TCGA-set, AUCs were 0.8254, 0.7267, and 0.7098 (Fig. 5I). The C-index was 0.7481 (95% CI: 0.6743–0.8219). We also compared the efficacy of the 14-gene signature with the 5-gene signature reported by Wu et al. and the 7-gene signature described by Lin et al. Results showed that 14-gene signature had better prognostic value than the 5-gene signature (C-index, 0.7481 vs. 0.6806) and a comparable predictive value with the 7-gene signature (C-index, 0.7481 vs. 0.7425) (Supplementary Figs. 2A–F). Collectively, our results indicated that the 14-gene signature functions well in PFI forecast for PTC.

Table 1
Baseline characteristics of PTC patients in the TCGA-THCA dataset.

Clinical characters	Training dataset	Validation dataset	Entire TCGA dataset
N	244	244	488
Follow-up time (day)	1124.93 ± 984.86	1156.84 ± 1031.48	1140.89 ± 965.72
Risk score	-1.34 ± 1.39	-1.40 ± 1.39	-1.37 ± 1.39
Age	47.23 ± 15.62	47.64 ± 16.04	47.43 ± 15.83
PFI			
Progression-free	221	218	439
Progression	23	26	49
Sex			
Male	70	60	130
Female	174	184	358
Histological type			
Classical/usual	179	172	351
Follicular (≥ 99% follicular patterned)	51	50	101
Tall cell (≥ 50% tall cell features)	14	22	36
T			
T1	71	70	141
T2	84	77	161
T3	79	84	163
T4	9	12	21
TX	1	1	2
N			
N0	109	116	225
N1	30	27	57
N1a	39	47	86
N1b	42	28	70

Clinical characters	Training dataset	Validation dataset	Entire TCGA dataset
NX	24	26	50
M			
M0 and Mx	238	241	479
M1	6	2	8
NA	0	1	1
AJCC stage			
Stage I	126	147	273
Stage II	33	18	51
Stage III	55	55	110
Stage IV	30	22	52
NA	0	2	2
Residual tumor			
R0	186	185	371
Rx	13	17	30
R1	26	25	51
R2	2	2	4
NA	17	15	32
Extrathyroidal extension			
None	162	161	323
Minimal (T3)	67	63	130
Moderate or Advanced (T4)	8	10	18
NA	7	10	17
Multifocality			
Unifocal	136	124	260
Multifocal	103	115	218
NA	5	5	10
Anatomic site			

Clinical characters	Training dataset	Validation dataset	Entire TCGA dataset
Unilateral	192	187	379
Isthmus	10	12	22
Bilateral	41	40	81
NA	1	5	6

Verification of potential relationship between the undifferentiation and 14-gene signature

To explore whether the 14-gene signature is correlated with undifferentiation and malignant degree, we searched the expression pattern of MTG-based gene signature from three dataset (GSE 29265, GSE 33630, GSE 76039), and compared the gene risk scores between ATC/PDTC/PTC. In the datasets GSE 29265 and GSE 33630, risk scores were higher in ATC samples compared to PTC samples ($p < 0.005$, 0.0001, respectively), as shown in Fig. 6A–D. In the dataset GSE 76039, risk scores were higher in ATC samples compared to PDTC samples ($p < 0.05$), as shown in Fig. 6E, F.

GSEA

To seek the potential alteration underlying the 14-gene signature, we conducted GSEA in PTC from the TCGA-THCA (Fig. 7A–I). In the high-risk group samples, the molecular alterations were related to the homologous recombination, cell cycle, DNA replication and P53 signaling pathways. For the oncological signatures, a total of 34 terms, including the RB_P107_DN.V1_UP and Singh_Kras_Dependency_Signature were related. GSEA results are presented in Supplementary Table 4.

Prognosis-associated factors of the PFI in PTC

TCGA PTC Cases (406/488) with complete clinical information including age, sex, BRAFV600E mutation status, AJCC stage, extrathyroidal extension, residual tumor, multifocality, anatomic site of tumors, and the histological type, were enrolled to identify prognostic factors. The reasons for exclusion of each case from the analysis are described in Supplementary Table 5. Uni-Cox analysis showed that age, sex, T, M, N, TNM stage, extrathyroidal extension and risk score were significantly PFI-related ($p < 0.05$) (as shown in Table 2). Multi-Cox analysis showed that the gene score ($p < 0.0001$) and M ($p < 0.05$) were independent factors for prognosis (as shown in Table 3).

Table 2
Unadjusted univariate analysis of risk factors

Exposure	Statistics	HR	95% CI	P-value
Risk score	-1.3362 ± 1.3695	2.3997	1.8196 to 3.1648	< 0.0001
BRAF V600E				
Wildtype	196(48.28%)	1		
Mutant	210(51.72%)	1.2129	0.6570 to 2.2391	0.5373
Sex				
Male	103(25.37%)	1		
Female	303(74.63%)	0.5129	0.2751 to 0.9563	0.0357
Age				
≤55 years	289(71.18%)	1		
>55 years	117(28.82%)	2.2369	1.2160 to 4.1150	0.0096
Histological type				
Classical/usual	282(49.46%)	1		
Follicular (≥ 99% follicular patterned)	93(22.91%)	0.6749	0.2808 to 1.6221	0.3795
Tall Cell (≥ 50% tall cell features)	31(7.64%)	2.0703	0.8607 to 4.9800	0.1042
T				
T1	117(28.82%)	1		
T2	143(35.22%)	2.8638	0.9422 to 8.7040	0.0636
T3	131(32.27%)	4.6277	1.5809 to 13.5463	0.0052
T4	15(3.69%)	6.8996	1.7238 to 27.6170	0.0063
N stage				
N0	188(46.31%)	1		
NX	44(10.84%)	1.4074	0.4537 to 4.3656	0.554
N1	174(42.86%)	2.4772	1.2498 to 4.9101	0.0094
M				
M0&Mx	399(98.28%)	1		
M1	7(1.72%)	5.3685	1.6557 to 17.4068	0.0051
AJCC stage				

Exposure	Statistics	HR	95% CI	P-value
Stage I	231(56.9%)	1		
Stage II	46(11.33%)	1.2813	0.4282 to 3.8339	0.6576
Stage III	88(21.67%)	2.3246	1.1174 to 4.8360	0.024
Stage IV	41(10.1%)	4.1274	1.8153 to 9.3844	0.0007
Residual tumor				
R0	338(83.25%)	1		
Rx	24(5.91%)	0.9565	0.2297 to 3.9837	0.9513
R1	41(10.1%)	1.3082	0.5116 to 3.3455	0.5749
R2	3(0.74%)	3.1074	0.4228 to 22.8401	0.2653
Extrathyroidal extension				
None	285(70.2%)	1		
Minimal (T3)	108(26.6%)	1.963	1.0414 to 3.6999	0.037
Moderate or Advanced (T4)	13(3.2%)	2.9351	0.8774 to 9.8188	0.0805
Multifocality				
Unifocal	224(55.17%)	1		
Multifocal	182(44.83%)	1.2276	0.6663 to 2.2616	0.5107
Anatomic site				
Unilateral	325(80.05%)	1		
Isthmus	17(4.19%)	0.5271	0.0721 to 3.8535	0.5281
Bilateral	64(15.76%)	1.4741	0.6780 to 3.2051	0.3275

Table 3
Multivariate Cox regression analysis of risk factors

Exposure	Non-adjusted	95% CI	P-value	Adjusted	95% CI	P-value
Risk score	2.5861	1.7783 to 3.7610	< 0.0001	2.2299	1.5897 to 3.1277	< 0.0001
BRAF V600E				NA		
Wildtype	1					
Mutant	1.0786	0.5142 to 2.2624	0.8413			
Sex						
Male	1			1		
Female	0.8711	0.4097 to 1.8518	0.7198	0.7857	0.3800 to 1.6244	0.5151
Age						
Age						
≤55 years	1			1		
>55 years	1.9436	0.7166 to 5.2713	0.1917	1.6746	0.6685 to 4.1946	0.2711
Histological type						
Classical/usual	1			1		
Follicular (≥ 99% follicular patterned)	1.1864	0.4000 to 3.5190	0.758	1.3513	0.4745 to 3.8482	0.5729
Tall Cell (≥ 50% tall cell features)	1.5592	0.5525 to 4.3997	0.4014	1.7742	0.6595 to 4.7730	0.2561
T						
T1	1			1		
T2	2.9719	0.9088 to 9.7181	0.0716	2.2124	0.6969 to 7.0235	0.1779
T3	2.8755	0.7558 to 10.9401	0.1213	2.3733	0.6515 to 8.6457	0.1901
T4	0.4602	0.0269 to 7.8826	0.5923	0.623	0.0488 to 7.9511	0.7157
N stage						
N0	1			1		

Exposure	Non-adjusted	95% CI	P-value	Adjusted	95% CI	P-value
N1	2.0739	0.8799 to 4.8877	0.0954	2.1684	0.9547 to 4.9248	0.0644
NX	1.8035	0.5100 to 6.3774	0.3601	1.5408	0.4565 to 5.2008	0.4861
M						
M0&Mx	1			1		
M1	6.6338	1.2311 to 35.7467	0.0277	4.5052	1.0240 to 19.8218	
AJCC stage						
Stage I	1			1		
Stage II	1.0086	0.2403 to 4.2334	0.9907	1.2826	0.3548 to 4.6365	0.7042
Stage III	0.9048	0.3325 to 2.4618	0.8447	0.9534	0.3496 to 2.6005	0.9258
Stage IV	0.9736	0.2620 to 3.6181	0.9682	1.0311	0.3025 to 3.5144	0.961
Residual tumor				NA		
R0	1					
Rx	0.6786	0.1176 to 3.9169	0.6647			
R1	0.6375	0.2278 to 1.7835	0.391			
R2	3.9719	0.2331 to 67.6756	0.3404			
Extrathyroidal extension						
None	1			1		
Minimal (T3)	1.2717	0.4860 to 3.3275	0.6242	1.2626	0.5063 to 3.1483	0.617
Moderate or Advanced (T4)	3.5438	0.2107 to 59.6001	0.3796	1.9871	0.1687 to 23.3994	0.5852
Multifocality				NA		
Unifocal	1					
Multifocal	2.4904	1.0945 to 5.6667	0.0296			

Exposure	Non-adjusted	95% CI	P-value	Adjusted	95% CI	P-value
Anatomic site				NA		
Unilateral	1					
Isthmus	0.5152	0.0652 to 4.0687	0.5293			
Bilateral	1.1138	0.4317 to 2.8737	0.8236			

Construction and validation of the predictive nomogram

A nomogram for predicting the AUCs of PTC was built with a stepwise Cox regression model with the complete clinical information of the 406 patients (Fig. 8A). Parameters including gene risk score, age, T, N, M, and histological type were also incorporated in the nomogram. For the histological type, the tall cell variant was defined as “aggressive”[29]. The calibration curve showed that the nomogram functions well in predicting the PFIs (Fig. 8B). The AUCs for 1-, 3-, and 5-year PFIs were 0.9078 (95% CI, 0.8450, 0.9706), 0.7684 (0.6717, 0.8650), and 0.7543 (0.6372, 0.8713), respectively (Fig. 8C). The C-index was 0.7924 (95% CI, 0.7284, 0.8564). Further, patients with a higher nomogram points were associated with a significantly shorter PFI (Fig. 8D). We also compared the nomogram's prediction performance with the ATA risk stratification provided in clinical information. The AUCs of the nomogram were 0.9075, 0.7685, and 0.7544, respectively. The C-index was 0.7897 (95% CI, 0.7239–0.8555). AUCs for ATA risk stratification were 0.7201, 0.6656, and 0.6828. The C-index of ATA risk was 0.7850 (95% CI, 0.6742–0.8957) (Fig. 9A–C). The results showed that efficacy of the novel nomogram was better than ATA risk stratification system. Further, we built a visualized calculator using the “DynNom” package of R[30].

Clinical correlation of the novel nomogram

Next, we analyzed the correlation between the nomogram and clinical parameters. In groups divided by age, patients aged over 55 years had lower nomogram points than younger patients (Fig. 9D). In terms of ATA recurrence risk, patients with intermediate/high risk had lower nomogram points than those with low risk (Fig. 9E). Patients in stage III and IV had lower nomogram points in terms of tumor stage than those in earlier stage(Fig. 9F). The nomogram points of patients with LNM(+) were lower than those with LNM(-) (Fig. 9G). In terms of anatomic site and primary focality, the differences were not statistically significant ($P > 0.05$) (Fig. 9H, I). In terms of age, ATA risk, AJCC stage, and N stage, the differences were statistically significant ($P < 0.05$). In order to conveniently calculate the risk of recurrence in the specified time period, we conducted an online graphical calculator which is open accessed (https://liuruisurgeon.shinyapps.io/PTC_MTgs_Signature_Nomogram/). The interface and instruction were as shown in Fig. 10.

Discussion

Most patients with PTC achieve a relatively good prognosis. However, persistent disease or recurrences are observed in 5%-20% of patients, which can be associated with severe complications following re-operation[31]. For patients with a low risk of recurrence, prolonged thyroid stimulating hormone (TSH) suppression therapy may cause multiple adverse effects such as osteoporosis or osteopenia and cardiac comorbidities like atrial fibrillation[32]. Considering the excellent prognosis and high recurrence rate, development of novel diagnostic tools with high sensitivity and specificity seems to have greater clinical significance than exploration of neoadjuvant therapies. Traditional staging systems such as the ATA risk stratification system allow evaluation of recurrence risk with a stratified population rather than individualized risk, which indicates that a group of patients sharing the same clinical and pathological characteristics would have the same chance of recurrence[33]. However, the biological mechanisms underlying PTC progression are highly complex and heterogeneous and require a more accurate and personalized prediction model based on biomarkers at the molecular level. Therefore, specified gene signatures would predict the metastatic and recurrent potential of tumors effectively.

The incidence of PTCs has been continually increasing; however, the mortality rate has not changed substantially, which is probably because the majority of PTCs diagnosed incidentally are low-risk papillary thyroid microcarcinomas (PTMCs). Except for tumors with high-risk features such as extrathyroidal extension, clinically evident LNM(+)and special aggressive types, active surveillance appears to be safe[34] and can replace immediate surgery for low-risk PTC[35]. In general, active surveillance begins when patients are diagnosed with low-risk PTC by ultrasound examination of fine-needle aspiration biopsy (FNAB). Since PTCs involve biological mechanisms, the decision to perform active surveillance is based on gene signatures determined using biological tests followed by FNAB, which would be safer than assessments based on simple clinical and pathological characteristics, since patients with a higher gene risk score but with a low risk of clinical features would be treated more rationally.

PTC patients with cervical LNM are usually at a high risk of recurrence and have a poorer prognosis as the incidence of PTC has increased rapidly in recent years[36]. Therefore, LNM is a significant reason for locally advanced and recurrent diseases, which motivated us to focus on differentially expressed MTGs derived from HCMDB, which annotated about 2,000 potential MTGs based on more than 7,000 published pieces of literature. We identified 33, then reduced the variables to 14 DE-MTGs that were PFI-related of PTC. A novel 14-gene signature was then established and proved to be an independent prognostic factor of PTC. The high-risk patients were with a significantly shorter PFI than those with low risk. Among the 14 genes, CRABP2, EZH2, KISS1R, and S100A4 were upregulated and associated with shorter PFI (HR > 1), whereas AGTR1, ALDH1A1, DEPTOR, FAM3B, FBLN5, LIFR, SDPR, SOD3, and TFF3 were downregulated and associated with better PFI (HR < 1). In the identified 14 genes, several were previously proved to be associated with PTC progression through experiments. For example, extracellular S100A4 mediates human TC cell migration through the response of RAGE/Dia-1 signaling system[37]; overexpression of cancer stem cell markers, including ALDH1A1, in PTC was associated with a shorter PFI during follow-up[38]; ablation of estrogen receptor β decreases and suppresses PTC tumor growth, while the estrogen receptor β-H19 positive feedback loop has an influential role in PTC stem cell preservation[39]. Thus,

these genes have the potential to predict metastasis and recurrence in PTC. GO enrichment analysis showed that DE-MTGs were enriched in cell adhesion, cellular matrix organization, and cell-substrate junction, consistent with the definition of MTGs, which have been proven to be associated with cancer metastasis. To our knowledge, patients with ATC and PDTC only have a mean survival after diagnosis of 0.5 and 3.2 years, respectively, and undifferentiation is a major reason for the highly malignant degree[40]. The results of significant higher gene risk scores in ATC samples partly confirms our conjecture. Besides, we explored the potential molecular alteration by the 14-gene signature using GSEA. GSEA, which is based on careful consideration of all differential genes' role, can help reveal the complex behavior of genes in the condition of health and disease more accurately whereas traditional strategies including KEGG or GO are focused on identifying individual genes that exhibit differences[41]. Multiple alterations of gene expression in the high-risk group were involved in tumor biology processes, such as homologous recombination, cell cycle, and P53 signaling pathways. Thus, the potential mechanisms underlying patients' poor PFI in the high-risk group could be elucidated. However, further explorations are needed.

Nomograms are widely used since the ability to present the numerical probability of a particular clinical event by integrating prognostic variables[42]. Nomograms including a risk score based on gene signatures and clinicopathological parameters can predict prognosis more precisely after surgery. Moreover, numerical results are more comfortable for patients to understand than the traditional staging system. As described before, traditional staging systems cannot provide an individualized risk, which is consistent with the result that the novel nomogram was better than ATA risk stratification in efficacy of predicting the PFI in PTC. To our knowledge, the prognostic gene signature based on these 14 MTGs and the relevant nomogram has not been reported before. The limited number of genes made it practical and economically feasible than whole-genome sequencing.

There were limitations to our study. First, the primary source of RNA sequencing data and clinical information was the TCGA program, in which the source of samples were from North American people. When applying the model to patients from different countries or regions, possible deviations or biases should be accounted for. Second, due to the lack of large independent dataset of PTC, we validated the nomogram’s power on the TCGA dataset itself. Future validation of external datasets with complete follow-up data is necessary. Furthermore, some essential clinical information, such as N stage condition, was missing or uncertain (Nx), which would attenuate the prognostic model's predictive power. Finally, we compared with the 2009 ATA risk guidelineto evaluate predict power since the available TCGA program did not include the latest ATA risk stratification. Further comparison is required to validate the nomogram's efficacy with the newest ATA risk system.

Conclusion

We built a novel 14-gene signature, then established a nomogram combining the signature and relevant clinical and pathological factors for predicting the PFI of PTC. The novel nomogram was better than the

current stratification system. It may be helpful for individualized active and postoperative surveillance strategies.

Abbreviations

AUC, area under the curve; PTC, papillary thyroid carcinoma; ATC, anaplastic thyroid cancer; PDTC, poorly differentiated thyroid cancer; DE-MTGs, differentially expressed MTGs; PFI, progression-free interval; HCMDB, Human Cancer Metastasis Database; ROC, receiver operating characteristic; C-index, concordance index; GO, Gene Ontology; BPs, biological processes; CCs, cellular components; MFs, molecular functions; KEGG, Kyoto Encyclopedia of Genes and Genomes; ATA, American Thyroid Association; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas Program; EMT, epithelial-to-mesenchymal transition; RRA, robust rank aggregation; AJCC, American Joint Committee on Cancer; ANOVA, one-way analysis of variance.

Declarations

Acknowledgments

We thank Dr.Wei Ge for her helpful suggestions on the manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

None.

Availability of data and materials

All the datasets are obtained from the TCGA (https://portal.gdc.cancer.gov/) and Cbioportal database (http://www.cbioportal.org/). The databases are open accessed and available for public.

Funding

This research was supported by the Nature Science Foundation of Beijing [grant number: 7202164], CAMS Innovation Fund for Medical Sciences (CIFMS) [grant number: 2016-12M-3-005], and CAMS Innovation Fund for graduate students [grant number: 2019-1002-44].

Authors' contributions

RL and MW designed the study and obtained the data; RL and ZC analyzed and interpreted the data: RL and XL wrote the manuscript; HY and ZL revised and approved the manuscript.

References

1. La Vecchia C, Malvezzi M, Bosetti C, Garavello W, Bertuccio P, Levi F, et al. Thyroid cancer mortality and incidence: a global overview. Int J Cancer. 2015;136:2187–95.
2. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res. 2014;74:2913–21.
3. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in Thyroid Cancer Incidence and Mortality in the United States, 1974–2013. Jama-J Am Med Assoc. 2017;317:1338–48.
4. Bilimoria KY, Bentrem DJ, Ko CY, Stewart AK, Winchester DP, Talamonti MS, et al. Extent of surgery affects survival for papillary thyroid cancer. Ann Surg. 2007;246:375–84.
5. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid Off J Am Thyroid Assoc. 2016;26:1–133.
6. Wu M, Li X, Zhang T, Liu Z, Zhao Y. Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. Front Oncol. 2019;9. doi:10.3389/fonc.2019.00996.
7. Du Y, Gao Y. Development and validation of a novel pseudogene pair-based prognostic signature for prediction of overall survival in patients with hepatocellular carcinoma. BMC Cancer. 2020;20:887.
8. Wu M, Yuan H, Li X, Liao Q, Liu Z. Identification of a Five-Gene Signature and Establishment of a Prognostic Nomogram to Predict Progression-Free Interval of Papillary Thyroid Carcinoma. Front Endocrinol. 2019;10. doi:10.3389/fendo.2019.00790.
9. Lin P, Guo Y, Shi L, Li X, Yang H, He Y, et al. Development of a prognostic index based on an immunogenomic landscape analysis of papillary thyroid cancer. Aging. 2019;11:480–500. doi:10.18632/aging.101754.
10. Steeg PS. Targeting metastasis. Nat Rev Cancer. 2016;16:201–18.
11. Karamanou K, Franchi M, Vynios D, Brézillon S. Epithelial-to-mesenchymal transition and invadopodia markers in breast cancer: Lumican a key regulator. Semin Cancer Biol. 2020;62:125–33.
12. Ko J, Winslow MM, Sage J. Mechanisms of small cell lung cancer metastasis. EMBO Mol Med. 2020;:e13122.
13. Wada N, Duh Q-Y, Sugino K, Iwasaki H, Kameyama K, Mimura T, et al. Lymph Node Metastasis From 259 Papillary Thyroid Microcarcinomas. Ann Surg. 2003;237:399–407. doi:10.1097/01.SLA.0000055273.58908.19.

14. Asimakopoulos P, Shaha AR, Nixon IJ, Shah JP, Randolph GW, Angelos P, et al. Management of the Neck in Well-Differentiated Thyroid Cancer. Curr Oncol Rep. 2020;23:1. doi:10.1007/s11912-020-00997-6.
15. Watkinson JC, Franklyn JA, Olliff JFC. Detection and surgical treatment of cervical lymph nodes in differentiated thyroid cancer. Thyroid Off J Am Thyroid Assoc. 2006;16:187–94.
16. Zheng G, Ma Y, Zou Y, Yin A, Li W, Dong D. HCMDB: the human cancer metastasis database. Nucleic Acids Res. 2018;46 Database issue:D950–5. doi:10.1093/nar/gkx1008.
17. Dom G, Tarabichi M, Unger K, Thomas G, Oczko-Wojciechowska M, Bogdanova T, et al. A gene expression signature distinguishes normal tissues of sporadic and radiation-induced papillary thyroid carcinomas. Br J Cancer. 2012;107:994–1000.
18. Handkiewicz-Junak D, Swierniak M, Rusinek D, Oczko-Wojciechowska M, Dom G, Maenhaut C, et al. Gene signature of the post-Chernobyl papillary thyroid cancer. Eur J Nucl Med Mol Imaging. 2016;43:1267–77.
19. Tarabichi M, Saiselet M, Trésallet C, Hoang C, Larsimont D, Andry G, et al. Revisiting the transcriptional analysis of primary tumours and associated nodal metastases with enhanced biological and statistical controls: application to thyroid cancer. Br J Cancer. 2015;112:1665–74.
20. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma Oxf Engl. 2003;19:185–93.
21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
22. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012;28:573–80. doi:10.1093/bioinformatics/btr709.
23. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics J Integr Biol. 2012;16:284–7.
24. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
25. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33:1–22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/. Accessed 12 Dec 2020.
26. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011;27:3206–8. doi:10.1093/bioinformatics/btr511.
27. Landa I, Ibrahimpasic T, Boucai L, Sinha R, Knauf JA, Shah RH, et al. Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. J Clin Invest. 2016;126:1052–66.
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50. doi:10.1073/pnas.0506580102.

29. Cartwright S, Fingeret A. Contemporary evaluation and management of tall cell variant of papillary thyroid carcinoma. Curr Opin Endocrinol Diabetes Obes. 2020;27:351–7.
30. Jalali A, Alvarez-Iglesias A, Roshan D, Newell J. Visualising statistical models using dynamic nomograms. PloS One. 2019;14:e0225253.
31. Wong H, Wong KP, Yau T, Tang V, Leung R, Chiu J, et al. Is there a role for unstimulated thyroglobulin velocity in predicting recurrence in papillary thyroid carcinoma patients with detectable thyroglobulin after radioiodine ablation? Ann Surg Oncol. 2012;19:3479–85.
32. Schmidbauer B, Menhart K, Hellwig D, Grosse J. Differentiated Thyroid Cancer-Treatment: State of the Art. Int J Mol Sci. 2017;18.
33. Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, et al. Revised American Thyroid Association Management Guidelines for Patients with Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid. 2009;19:1167–214.
34. Ito Y, Miyauchi A, Oda H. Low-risk papillary microcarcinoma of the thyroid: A review of active surveillance trials. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol. 2018;44:307–15.
35. Saravana-Bawan B, Bajwa A, Paterson J, McMullen T. Active surveillance of low-risk papillary thyroid cancer: A meta-analysis. Surgery. 2020;167:46–55.
36. Yu X, Song X, Sun W, Zhao S, Zhao J, Wang Y-G. Independent Risk Factors Predicting Central Lymph Node Metastasis in Papillary Thyroid Microcarcinoma. Horm Metab Res Horm Stoffwechselforschung Horm Metab. 2017;49:201–7.
37. Medapati MR, Dahlmann M, Ghavami S, Pathak KA, Lucman L, Klonisch T, et al. RAGE Mediates the Pro-Migratory Response of Extracellular S100A4 in Human Thyroid Cancer Cells. Thyroid Off J Am Thyroid Assoc. 2015;25:514–27.
38. Kim HM, Koo JS. Immunohistochemical Analysis of Cancer Stem Cell Marker Expression in Papillary Thyroid Cancer. Front Endocrinol. 2019;10:523.
39. Li M, Chai H-F, Peng F, Meng Y-T, Zhang L-Z, Zhang L, et al. Estrogen receptor β upregulated by lncRNA-H19 to promote cancer stem-like properties in papillary thyroid carcinoma. Cell Death Dis. 2018;9:1120.
40. Landa I, Ibrahimpasic T, Boucai L, Sinha R, Knauf JA, Shah RH, et al. Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. J Clin Invest. 2016;126:1052–66.
41. Bild A, Febbo PG. Application of a priori established gene sets to discover biologically important differential expression in microarray data. Proc Natl Acad Sci U S A. 2005;102:15278. doi:10.1073/pnas.0507477102.
42. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in Oncology – More than Meets the Eye. Lancet Oncol. 2015;16:e173–80. doi:10.1016/S1470-2045(14)71116-7.

Figures

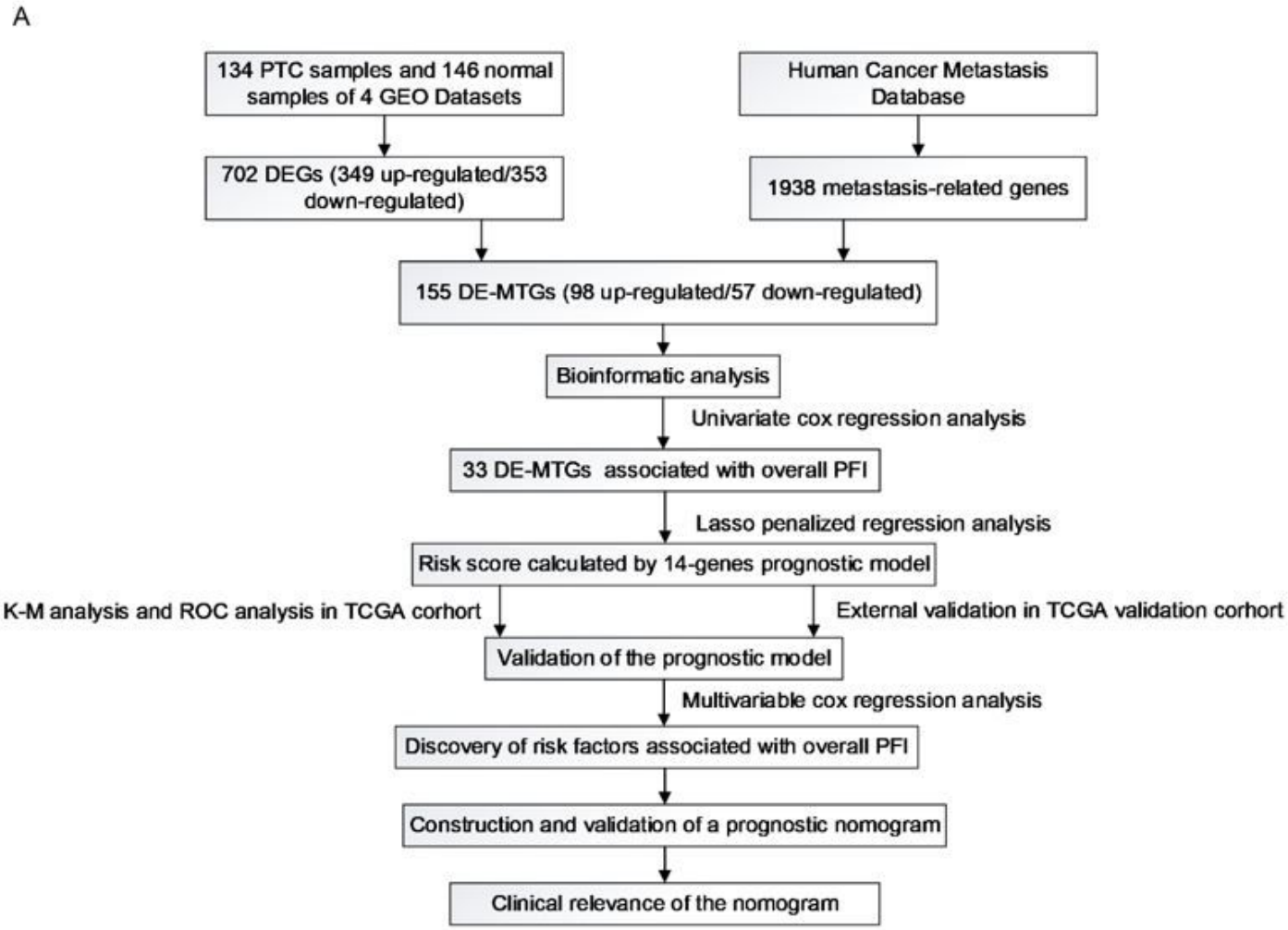

Figure 1

Flowchart describing the process of establishment and validation of the 14-gene signature and prognostic nomogram.

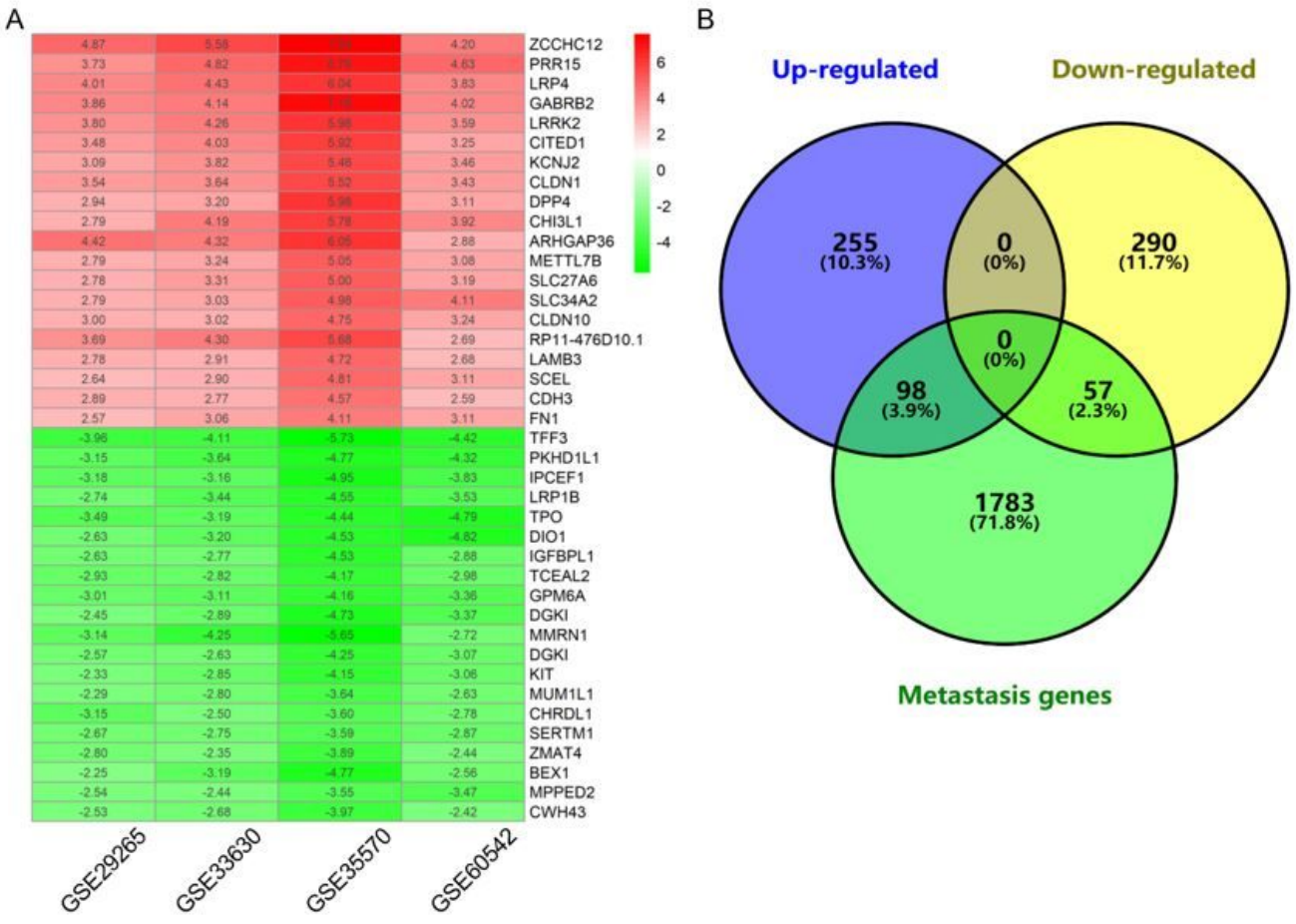

Figure 2

Identification of DE-MTGs in PTC. (A) Heatmap presenting the top 20 upregulated and downregulated DEGs in PTC after integrated analysis of the 4 GEO datasets using the RRA method. (B) 155 DE-MTGs, including 98 upregulated and 57 downregulated genes, were identified based on the intersection between GEO and THCA datasets.

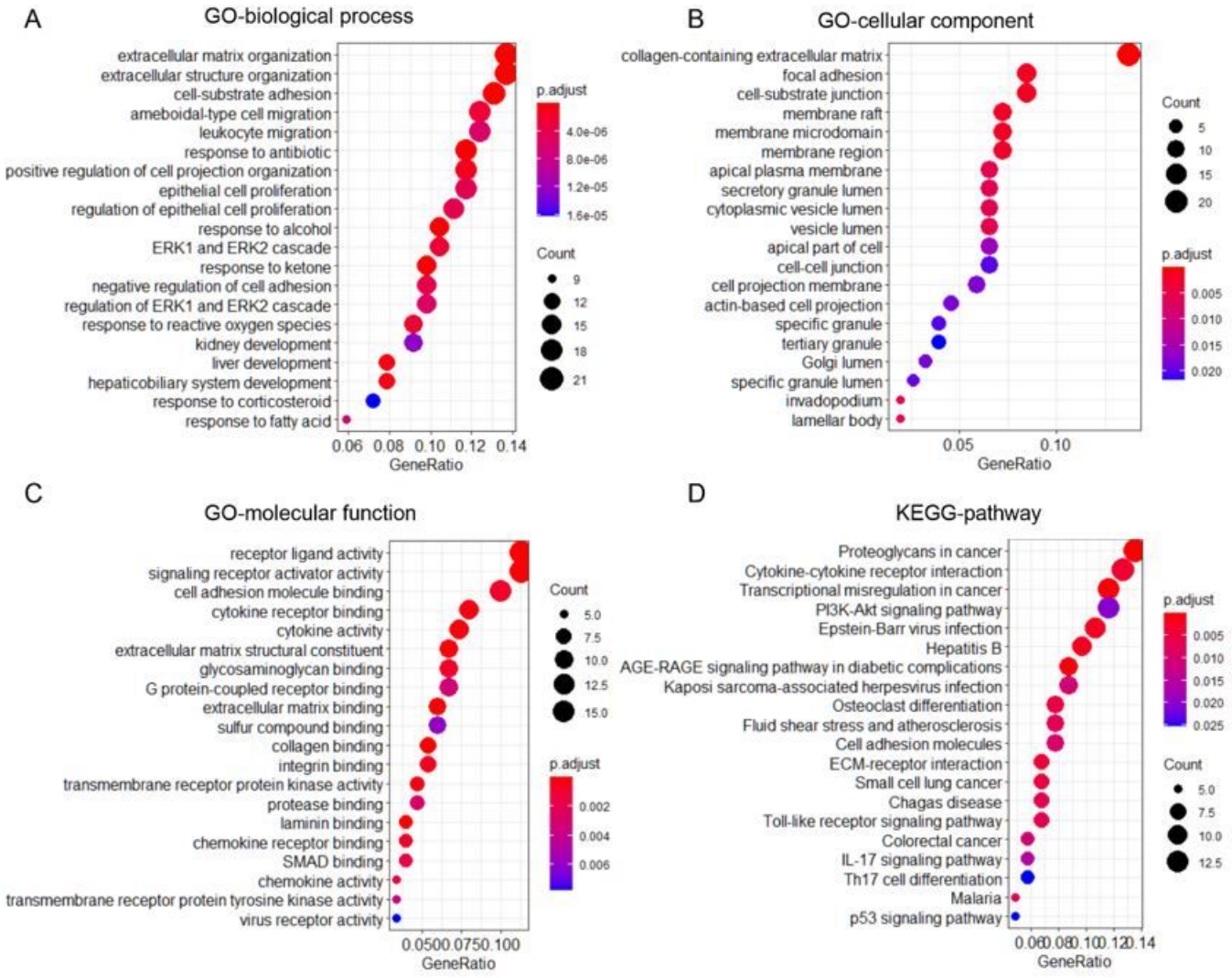

Figure 3

Functional enrichment analysis of the 155 DE-MTGs. (A) The top 20 enriched gene ontology (GO) biological process (BP) terms of the DE-MTGs. (B) The cellular components (CC) terms of the DE-MTGs. (C) The molecular function (MF) terms of the DE-MTGs. (D) The KEGG pathway terms of the DE-MTGs.

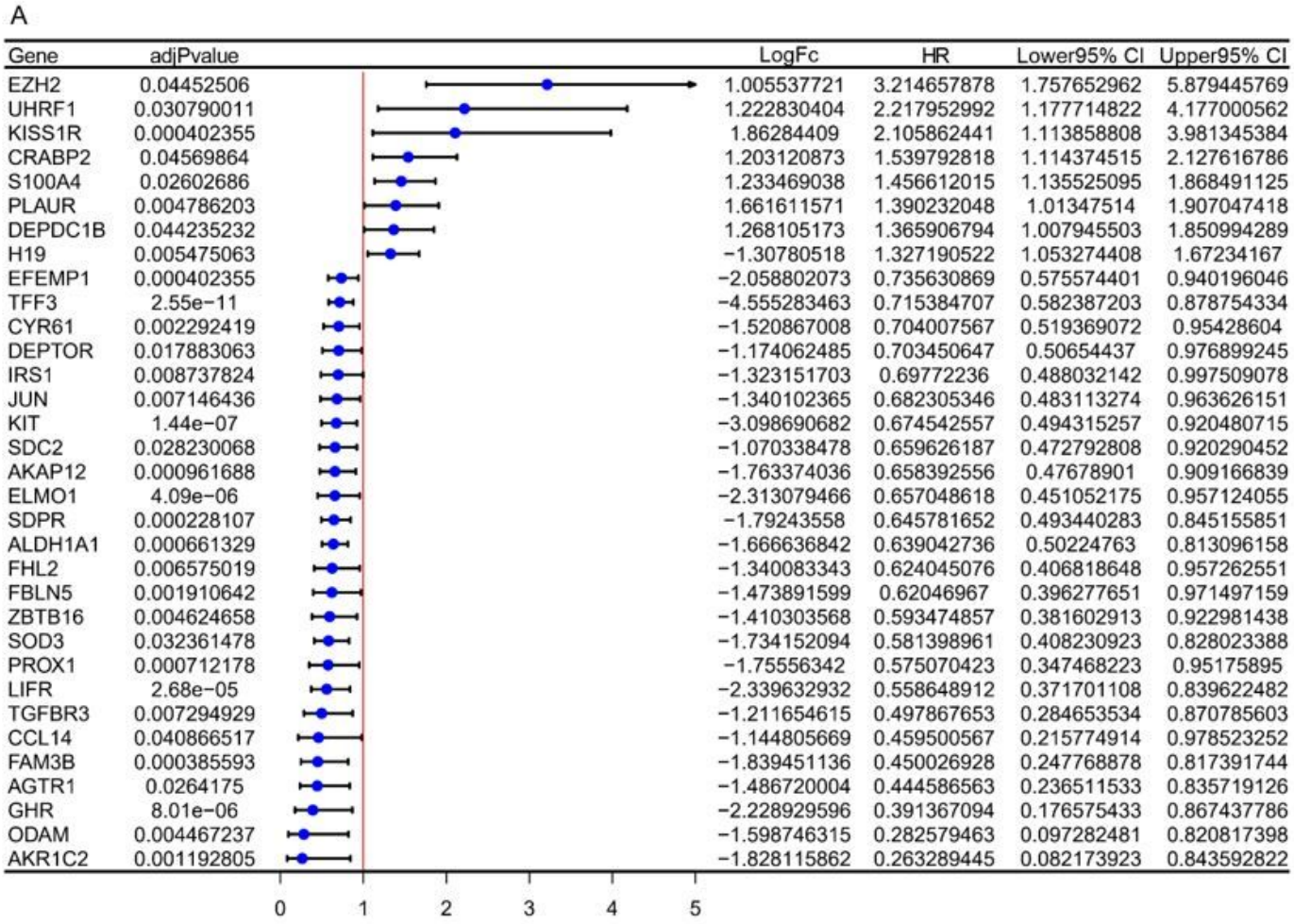

Figure 4

Differential expression level and hazard ratios (HR) of the 33 DE-MTGs. (A) Forest plot with hazard ratios (HR) representing the prognostic values of the 33 DE-MTGs that were PFI-related in PTC.

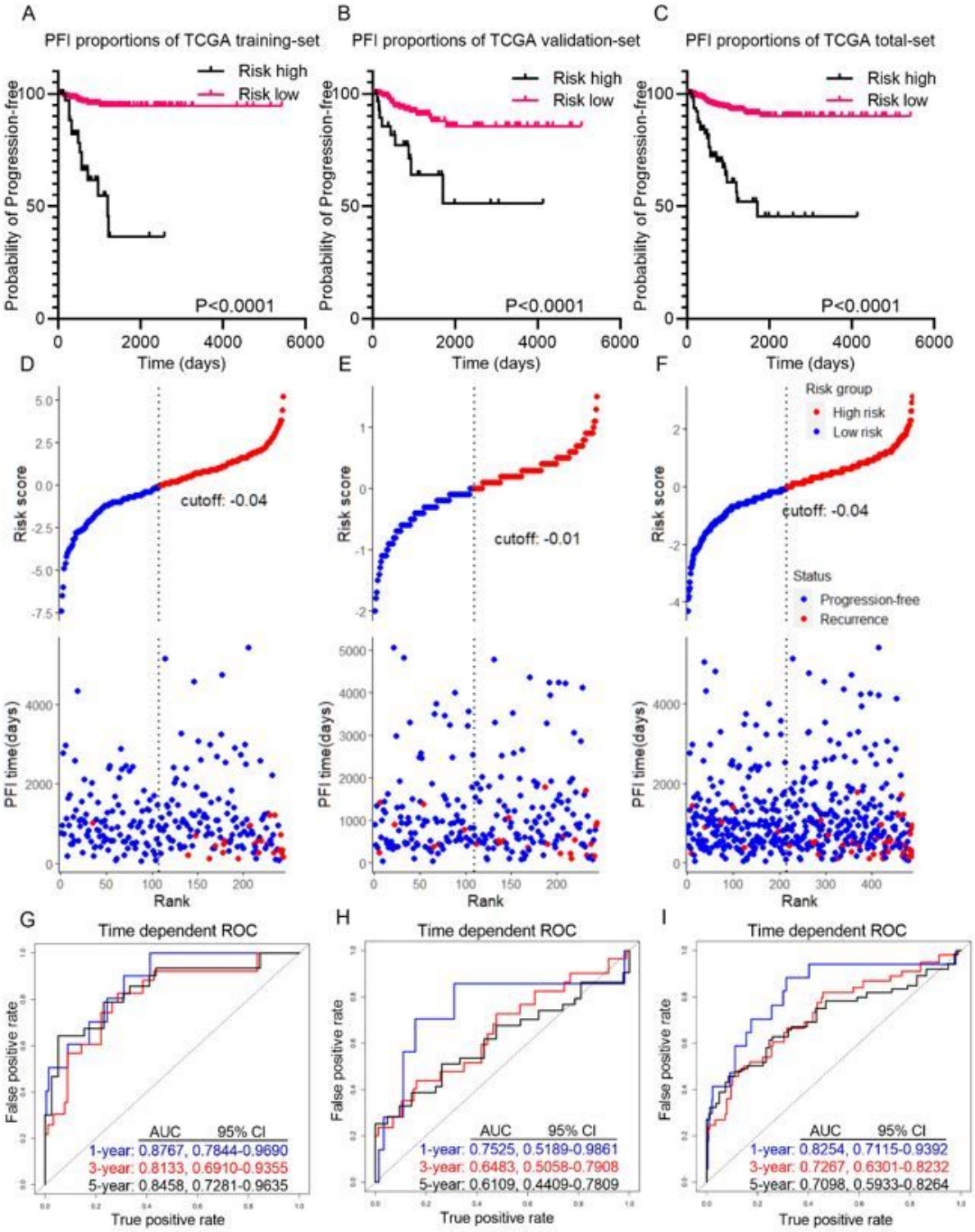

Figure 5

Evaluation of the efficacy of the 14-gene signature in the TCGA-THCA dataset. The dataset was randomly divided into the training set and the validation set equally. (A-C) K-M survival curves of the 14-gene signature represented PFI proportions at different timepoints. Patients from the training/validation/total sets are defined as “high risk” or “low risk” according to the optimal cut-off values by X-Tile software. (D-F) Relationship between the gene risk score (up) and recurrence status of patients of high/low-risk

(down) in the training/validation/total TCGA-THCA dataset. (G-I) Time-dependent ROC for the predictions of PFI for the 14-gene signature in the training/validation/total sets. ****P < 0.0001.

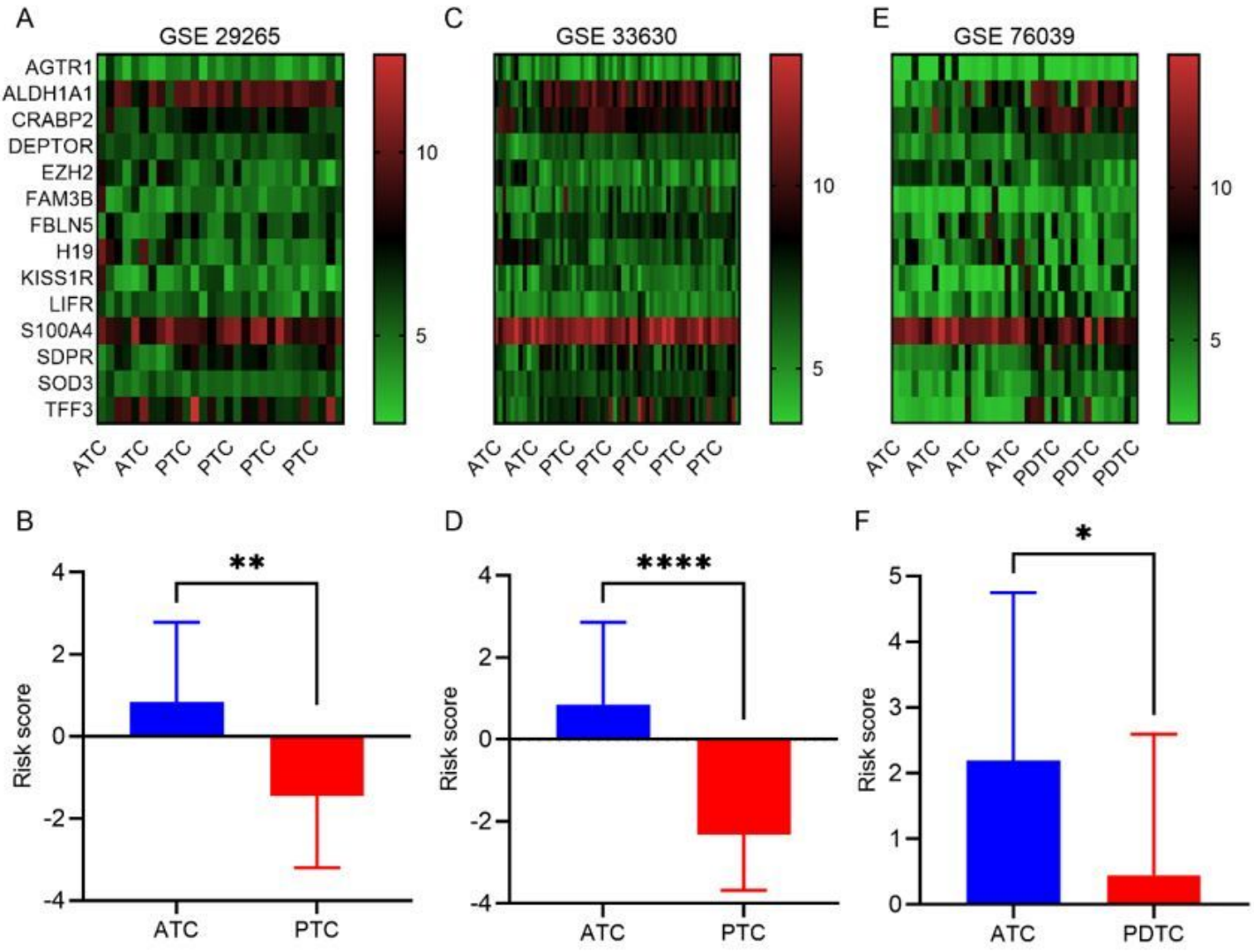

Figure 6

Potential relationship between the malignancy and 14-gene signature. The 14 genes expression pattern and gene risk score of samples from three GEO dataset (GSE 29265, GSE 33630, GSE 76039) were generated. (A, B) Expression pattern and gene risk score of ATC versus PTC samples from GSE 29265 (9 ATC and 20 PTC samples). (C, D) Expression pattern and gene risk score of ATC versus PTC samples from GSE 33630 (11 ATC and 49 PTC samples). (E, F) Expression pattern and gene risk score of ATC versus PDTC samples from GSE 76039 (20 ATC and 17 PDTC samples). ****P < 0.0001.

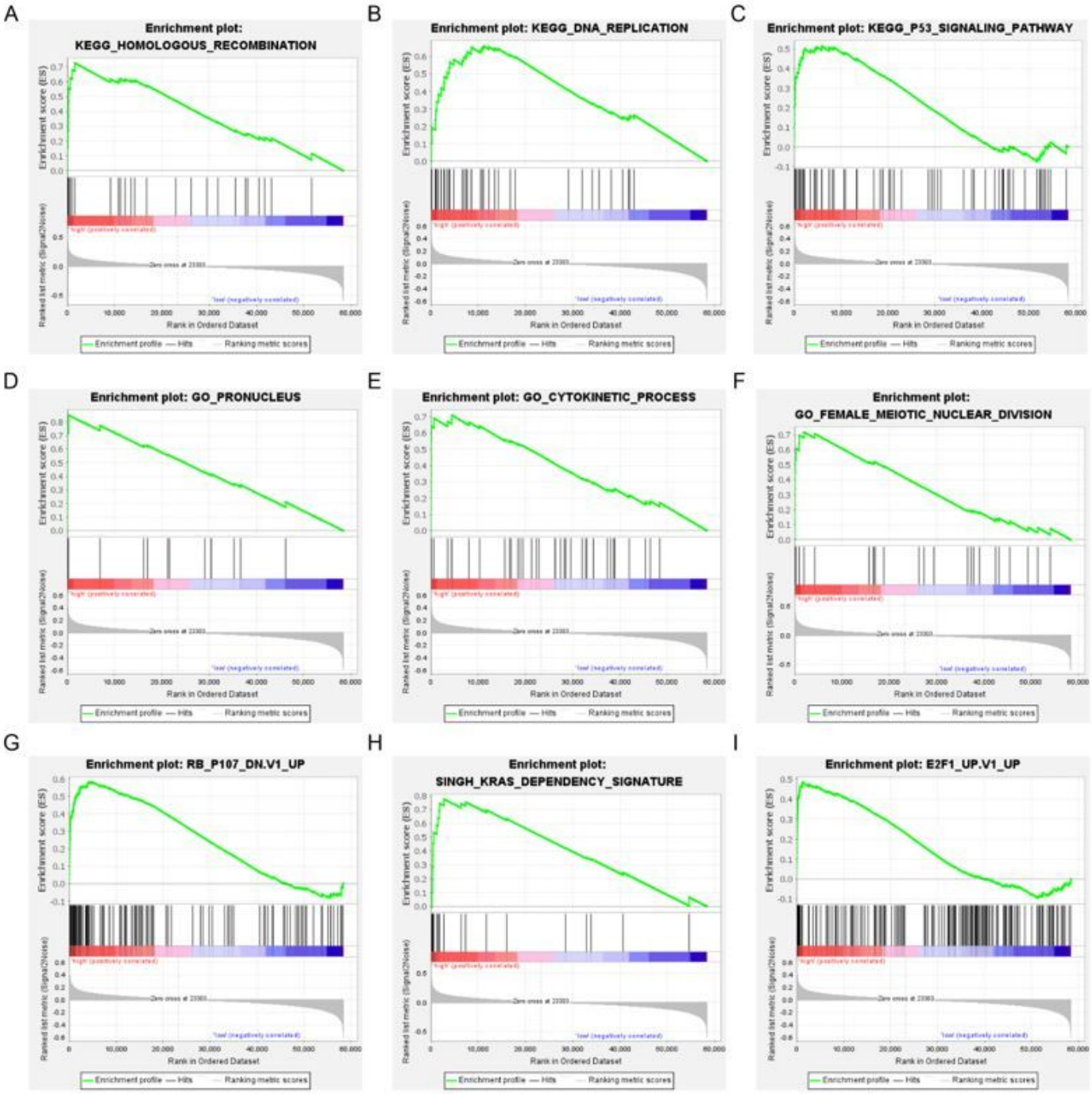

Figure 7

Gene set enrichment analysis (GSEA) analysis of the 14-gene signature. (A-I) Top signaling pathways, biological functions and oncogenic signatures significantly enriched in the high-risk group identified by GSEA.

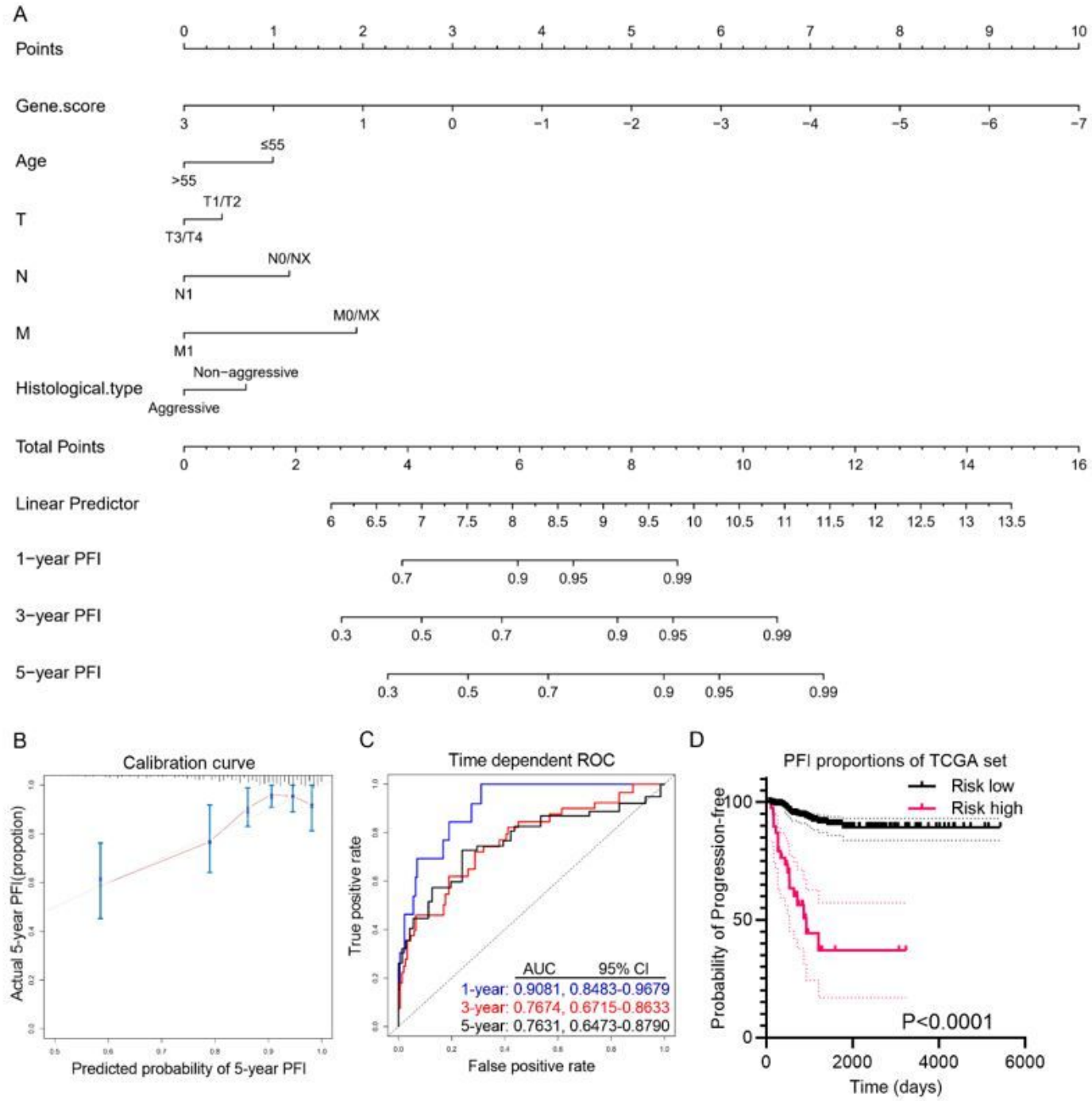

Figure 8

Construction and validation of the nomogram in predicting PFI of PTC in the TCGA-THCA dataset. (A) A nomogram based on the 14-gene signature and relevant clinical features for forecasting the PFI of PTC. (B) The calibration curve for internal validation of the nomogram. (C) The nomogram's prognostic efficacy using time-dependent ROC for predicting the 1-, 3- and 5-year PFI of PTC. (D) K–M survival curve of the nomogram represented PFI proportions at different timepoints. Patients from the TCGA-THCA dataset are put into two groups according to the optimal nomogram points cut-off value determined by X-Tile. ****P < 0.0001.

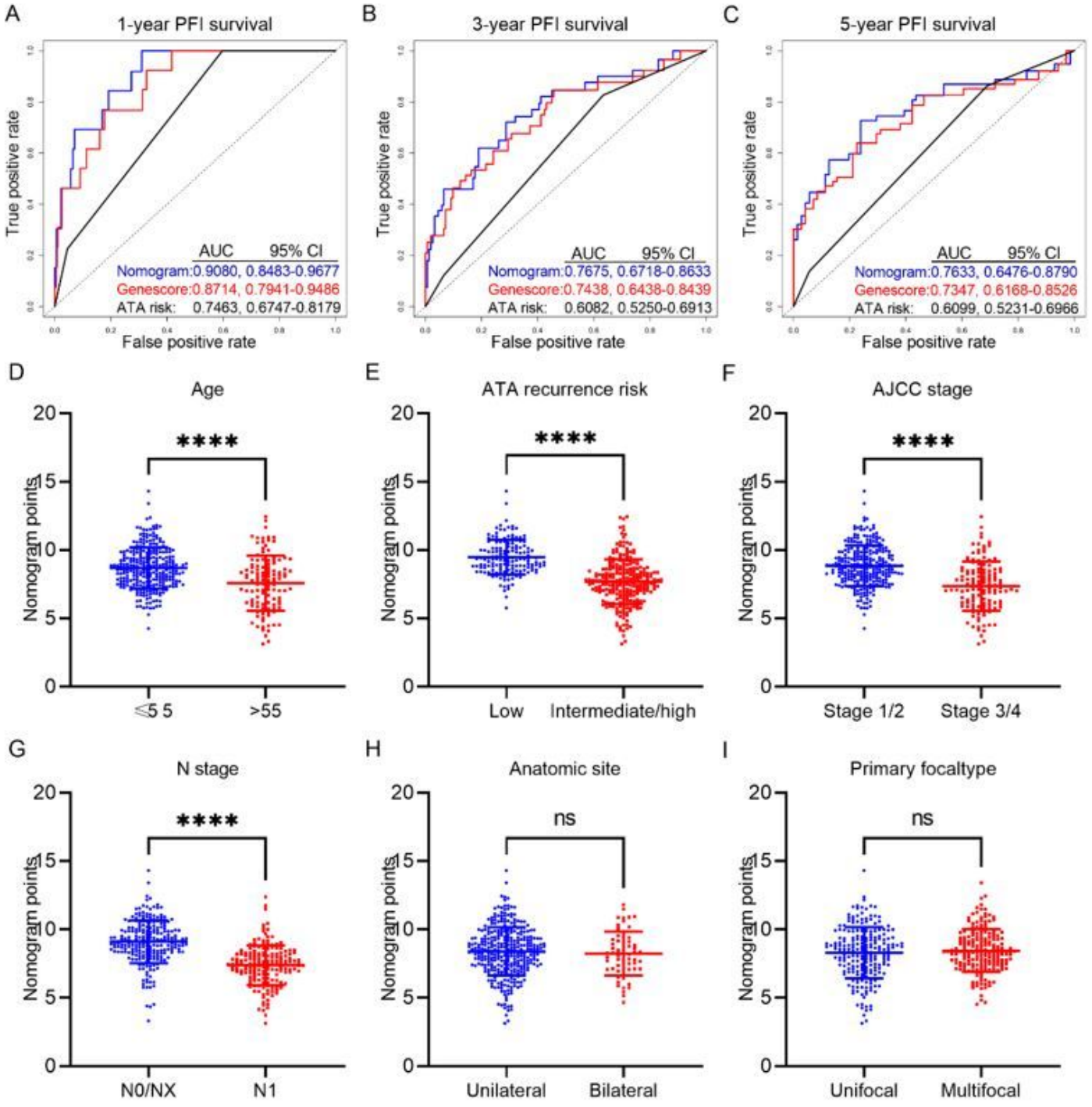

Figure 9

Clinical correlation of the nomogram. (A–C) Prognostic efficacy of the 14-gene signature, the nomogram, and the ATA risk stratification. (D-I) The distribution of the nomogram points according to different ages, ATA risks, AJCC stages, N stages, anatomic sites and primary focal types in the TCGA-THCA dataset. ****P < 0.0001.

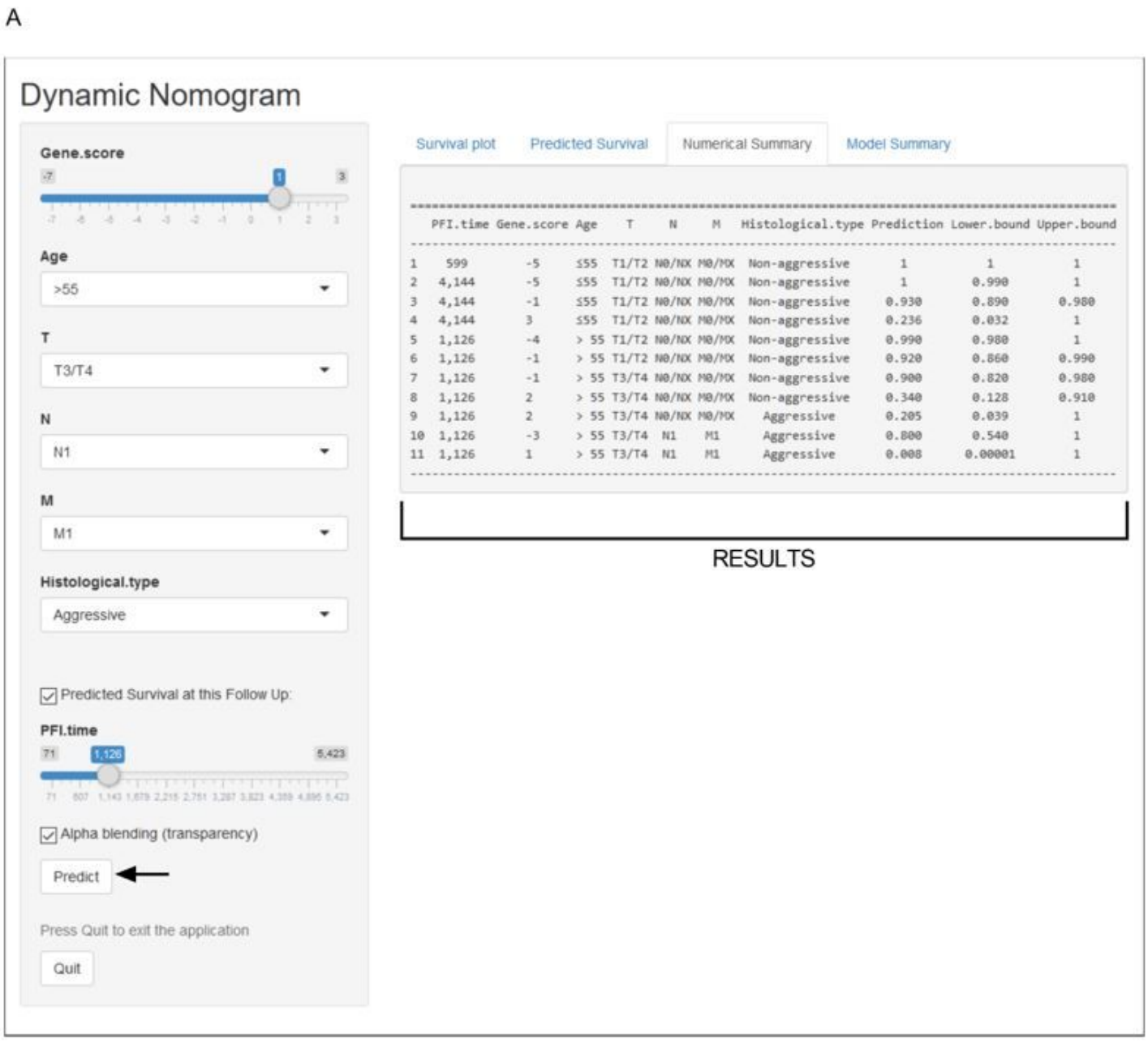

Figure 10

The online graphical calculator correlated with the nomogram. Click on the “predict” button (black arrow) after parameters selection including age, T, N, M, histological type, and time period of follow up, the numerical prediction will be shown in the right side of the interface. The website address is: https://liuruisurgeon.shinyapps.io/PTC_MTgs_Signature_Nomogram/.

Supplementary Files

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

- 20210108supplementaryfile.pdf