Improving the Prognosis of Colon Cancer Through Knowledge-Based Clinical-Molecular Integrated Analysis

Danyang Tong
Zhejiang University
Yu Tian
Zhejiang University
Qiancheng Ye
Zhejiang University
Jun Li
Zhejiang University School of Medicine Second Affiliated Hospital
Kefeng Ding
Zhejiang University School of Medicine Second Affiliated Hospital
Jingsong Li (ljs@zju.edu.cn)
Zhejiang University https://orcid.org/0000-0002-1064-637X

Research article

Keywords: colon cancer, clinical prognostic supplementary factor, clinical-molecular integrated analysis, pathway dysregulation scores

Posted Date: July 6th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-36892/v1

Version of Record: A version of this preprint was published at BioMed Research International on April 7th, 2021. See the published version at https://doi.org/10.1155/2021/9987819.

Abstract

Background

Colon cancer has high morbidity and mortality rates among cancers. Existing clinical staging systems cannot accurately assess the prognostic risk of colon cancer patients. Therefore, new prognostic factors are needed. In this study, a new pathway-based prognostic factor was discovered through a knowledge-based clinical-molecular integrated analysis.

Methods

A total of 374 samples from The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) dataset were used as the discovery set and 98 samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset were used as the validation set. After converting gene expression data into pathway dysregulation scores (PDS), the random survival forest and multicovariate Cox model were used to identify the best prognostic supplementary factors. Finally, a clinical prognostic model, a molecular prognostic model and a clinical-molecular integrated prognostic model were constructed to verify the supplementary effect of the discovered prognostic supplementary factor.

Results

The PDS of 14 pathways played important roles in prognostic prediction together with clinical prognostic factors through the random survival forest. Further screening through the multicovariate Cox model revealed that the PDS of pathway hsa00532 was the best clinical prognostic supplementary factor. The clinical-molecular integrated prognostic model constructed with clinical prognostic factors and the discovered prognostic factor was superior to the clinical prognostic model and molecular prognostic model in discriminative performance (C-indexes of 0.773, 0.746, and 0.619 in the discovery set and 0.893, 0.808, and 0.810 in the validation set, respectively). The Kaplan-Meier (KM) curves of patients grouped by PDS suggested that patients with a higher PDS had poorer prognosis, and stage II patients could be distinctly distinguished.

Conclusion

The PDS of pathway hsa00532 was a considerable clinical prognostic supplementary factor for colon cancer and may represent a potential prognostic marker for stage II colon cancer. The PDS calculation involves only 16 genes, which supports its potential clinical application prospects.

Background

Colon cancer is one of the top cancers in terms of incidence and mortality in both China and America [1, 2]. Recent global surveillance of cancer trends revealed that the age-standardized 5-year net survival of colon cancer ranges from approximately 15% to 75% in different countries, while for breast cancer or acute lymphoblastic leukemia in children, the 5-year net survival can be higher than 90% [3].

Currently, the American Joint Committee on Cancer (AJCC) tumor, node and metastasis (TNM) stage system is the most commonly used clinical staging tool for colon cancer. However, the accuracy of the staging system in assessing the prognostic risk of patients still needs to be improved, especially for stage II patients and stage IIIA patients [4]. In the recent 8th edition of the TNM staging system and the National Comprehensive Cancer Network (NCCN) Guidelines for colon cancer, molecular markers, including microsatellite instability (MSI), mismatch repair (MMR), RAS mutation and BRAF V6000E mutation, were recommended; however, the 8th edition of the TNM staging system for breast cancer already included the HER2 and ER status, which makes a large difference [5-7]. Therefore, to accurately analyze the prognosis of patients, we need to incorporate more prognostic factors in addition to clinical prognostic factors.

Incorporating molecular factors, such as gene expression data, would be a considerable option for improving the performance of cancer prognosis. However, in gene expression-based analyses of heterogeneous diseases, a single gene often provides weak information [8, 9]. A previous study claimed that multiple genes involved in the same biological process often have dysregulated performances. Therefore, the introduction of representative functional units, such as gene sets or pathways, may yield a more stable performance and may simultaneously provide certain biological annotations to improve the interpretability of the results [10-14]. These functional units, such as hypoxia-related genes or tumor microenvironment-related genes, can be obtained via prior knowledge, such as literature reviews, or directly obtained by analyzing a large number of genes and then used in subsequent prognostic analyses [15-17]. However, the gene set obtained directly by analyzing a large number of genes is not stable and will change with changes in the training sample; thus, using specific gene sets might be a good solution [18].

In recent years, machine learning methods have been widely used for cancer prognostic analysis. When performing prognostic analyses through machine learning methods, the introduction of prior knowledge, such as pathway information can further improve the performance of the model [19]. In most associated studies, molecular prognostic features are obtained by considering only the molecular features; therefore, new molecular features obtained through analysis may not be effectively combined with clinical features [20].

In this study, we conducted a knowledge-based clinical-molecular integrated analysis through a machine learning method, discovered new pathway-based molecular prognostic factors to supplement the clinical TNM staging system for colon cancer, and verified the improved performance of the clinical-molecular integrated prognostic models compared to that of clinical prognostic model.

Methods

Data sources and processing

Gene mRNA expression data of primary tumor and related clinical data in The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) project (discovery set) were obtained from cBioPortal

(http://www.cbioportal.org), and gene expression data for normal adjacent tissue in the TCGA-COAD were obtained from the UCSC Xena (http://xena.ucsc.edu/) as the reference set [21]. The mRNA sequence data of the discovery set and reference set used in this study were generated with the Illumina HiSeq 2000 platform and processed by the RNAseqV2 pipeline, which uses RNA-Seq by expectation maximization upper quartile (RSEM-UQ) for quantification. To validate the prognostic performance of the identified pathway-based factor, one independent dataset that offered identical clinical data and gene mRNA expression of primary tumors generated with a similar pipeline of colon cancer patients were obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (validation set) from LinkedOmics (http://linkedomics.org/cptac-colon/) [22]. The mRNA sequence data of the validation set used in this study were generated with the Illumina HiSeq 4000 platform and processed by the RNAseqV2 pipeline with RSEM-UQ for quantification. Both datasets can be used for an integrated analysis of clinical data and omics data. Therefore, this study used 374 cases in the TCGA-COAD data as the discovery set, 98 colon cancer cases in the CPTAC as the validation set, and 41 colon cancer normal adjacent tissue data from the TCGA as the reference set. The clinical data included T, N, and M stages and overall survival information. Other clinical prognostic factors, such as age and location, were not included because this study is focused on supplementing the clinical TNM staging system. All gene expression data were further log transformed for subsequent analysis. The following exclusion criteria were applied to the samples: containing Tis, N1c, or MX; lack of clear T, N and M stages; and invalid survival information. In gene expression data, genes without HUGO Gene Nomenclature Committee (HGNC) symbols and those with missing expression values or zero values were removed. The detailed information of the final dataset used for analysis is shown in Table 1.

Study protocol

In this study, the Kyoto Encyclopedia of Genes and Genomes (KEGG) human pathway database was introduced as prior knowledge, and a total of 327 prognostic gene sets with clear biological function definitions were obtained. Then, the best molecular prognostic factors from these pathway-based gene sets to supplement clinical prognosis were identified through a clinical-molecular integrated analysis by combining machine learning methods and survival analysis. The overall pipeline of the study is shown in Figure 1. Based on these pathway-based gene sets, this study converted gene expression data into a pathway dysregulation score (PDS), and used the pathway-based features obtained after the transformation and clinical features for subsequent analysis. Then, through the random survival forest and multicovariate Cox model, the molecular features that could be effectively combined with clinical features, and the best supplements for the prognosis of clinical features were screened as potential molecular prognostic factors. Finally, a clinical-molecular integrated prognosis model was constructed using clinical factors and the discovered molecular factors. Internal validation was performed on the discovery set and external validation was performed on the validation set to evaluate the prognostic improvement of the combination of the discovered supplementary prognostic factors and clinical prognostic factors in colon cancer overall survival.

Step 1: PDS calculation

Among the pathway-based approaches, two methods, PARADIGM and Pathifier, are widely used to estimate the pathway dysregulation information in a particular sample [23, 24]. However, PARADIGM requires pathway mechanisms and is inappropriate for complex or incomplete pathways. Pathifier requires only expression data of genes involved in each pathway and is more suitable for this study. The PDS quantifies the biological difference of a specific pathway between a diseased sample and normal samples with a numeric value range from 0 to 1, and it is transformed from the gene expression data by the R package Pathifier [23]. The pathway information was obtained from KEGG through the R package KEGGREST (version 1.26.1). Previous studies have confirmed that the PDS calculated by this method can effectively characterize the abnormality of the pathway [14, 23, 25, 26]. The PDS in each sample indicates the distance of deviation between the projection of a specific pathway and the projection of normal samples on the principle component curve. The calculation of the PDS consists of the following two steps:

A: Construction of the one-dimension principle component curve for a specific pathway. The principle component curve is a nonparametric nonlinear extension of the principle component analysis that avoids the assumptions of dependence in the data [27]. The curve pathed through the middle of all sample points in a p-dimensional space, where each dimension is the expression of a specific gene in a specific pathway, and each point in the space represents a sample. The curve is defined as a principle curve if , where is a single parameter whose variation traces all the points along the data and X is the data matrix of *n* samples with *p* genes involved in the specific pathway. The principle component curve is built through iterations of scatterplot smoothing.

B: Measuring the deviation of one diseased sample *d* from normal adjacent samples. After the principle component curve is built, each sample, including both the diseased sample and the normal adjacent sample, is projected onto the curve at the nearest points and . The centroid of is defined as reference point *r*. Therefore, the PDS of sample *d* could be represented by the distance of point to the reference point *r* along the curve.

In this study, the PDS of 327 human pathways from KEGG was calculated based on this method.

Step 2: Discovery of prognostic factors for clinical supplementation

In this study, the random survival forest was used to screen prognostic factors that could supplement clinical prognosis and then the multicovariate Cox model was used to identify prognostic factors that could be the best supplementary factors for clinical prognostic factors.

The random survival forest is an ensemble tree-based method used to analyze right-censored survival data [28]. The nonparametric random survival forest model can assess nonlinear effects of variables and explore complex interactions between variables. In addition, variables in the random survival forest model

that do not have prognostic ability can be filtered by variable importance. The variable selection through the random survival forest consists of the following two steps:

1. In the random survival forest, variables with an importance greater than 0 are screened and recorded.
2. Considering the existence of random processes, step A would be repeated 100 times to generate a matrix of variables with a variable importance value greater than 0. The prognostic factors that were recorded as important prognostic factors multiple times were regarded as important prognostic factors.

First, in this study, the clinical prognostic factors and all molecular factors (e.g., the PDS) were used as variables of the random survival forest. Variables that showed positive prognostic power more than 90 times were selected as the initial important prognostic factors. Next, the initial important prognostic factors were screened again. Here, the variable selection procedure was repeated 10 times. In each repetition, variables that showed positive prognostic power over 95 times were recorded as important prognostic factors. Finally, molecular factors that were recorded as important prognostic factors in all 10 repetitions were regarded as the final important prognostic factors screened by the random survival forest. The numbers of trees that offer the lowest error rate were chosen for both modeling steps of the random survival forest.

To identify the best molecular factor for clinical prognostic supplementation, multicovariate Cox models were constructed with clinical prognostic factors and different combinations of different molecular prognostic factors among the final important molecular factors. The models in which molecular factors showed no statistical significance of prognostic importance (with a p value of the covariate larger than 0.05) were excluded. The discrimination performance of the remaining models was measured by the bias-corrected concordance index (C-index). Comparisons of the discrimination performance were performed to identify the best prognostic supplementary factors that could offer a high bias-corrected C-index with the least genes used in the analysis.

Step 3: Construction of the clinical-molecular integrated prognostic model

The clinical prognostic factors T, N and M stages and the molecular prognostic factors identified as supplementary prognostic factors were used as covariates to construct a multicovariate Cox model as the clinical-molecular integrated prognostic model. The formula of this model is as follows:

$$h(t) = h(0)\exp(\alpha_1(T\ stage) + \alpha_2(N\ stage) + \alpha_3(M\ stage) + \sum \beta_n M_n)$$

where *h(t)* is the risk of death at time t, *h(0)* is the baseline risk, α is the regression coefficient of clinical prognostic factors, β is the regression coefficient of molecular prognostic factors and *M* is the molecular prognostic factor discovered in Step 2. In addition, corresponding clinical prognostic factors and

molecular prognostic factors were used to construct the corresponding clinical prognostic model and molecular prognostic model to evaluate the improvements of prognostic performance between the clinical-molecular integrated prognostic model and clinical prognostic model. Finally, a nomogram was constructed based on the clinical-molecular integrated prognostic model to predict the 3-year colon cancer overall survival because the longest follow up time of validation set was 44 months which makes it impossible to validate the 5-year prognosis of our model in the validation set.

Step 4: Assessment of the improvement of the clinical-molecular integrated prognostic model compared to the clinical prognostic model

First, according to the distribution of the PDS of corresponding molecular factors discovered, patients in the discovery set were divided into different groups based on the highest degree of differentiation of survival curves, and an observation of patients in the validation set grouped with the same threshold was made. These findings could provide a direct observation of the relevance of the discovered molecular prognostic factors and survival.

Second, based on the clinical prognostic factors and discovered molecular prognostic factors, one clinical prognostic model, one molecular prognostic model and one clinical-molecular integrated prognostic model were constructed with the discovery set. Internal validation through bootstrapping with 200 iterations on the discovery set and external validation through bootstrapping with 200 iterations on the validation set were used to assess the discrimination performance of these models. At the same time, because the mean survival time of patients with metastasis was shorter than that of patients without metastasis, the performance of the prognostic model might be affected. Therefore, one clinical prognostic model, one molecular prognostic model and one clinical-molecular integrated prognostic model were constructed with the same clinical prognostic factors and molecular prognostic factors on nonmetastatic patients in the discovery set. Internal validation through bootstrapping with 200 iterations on nonmetastatic patients in the discovery set and external validation through bootstrapping with 200 iterations on nonmetastatic patients in the validation set were used to assess the discrimination performance of these models for nonmetastatic patients.

Finally, to compare the prognostic performance of directly using gene expression data and using converted PDS in this study, genes involved in the pathways were combined with clinical prognostic factors in the clinical-molecular integrated prognostic model. Comparisons between the gene-based integrated prognostic model and pathway-based integrated prognostic model were conducted.

In this study, a Cox model was used as the constructed prognostic model and the C-index was used to evaluate the discriminative performance of the models. Harrell's C-index was chosen to evaluate the overall discriminative performance of the model, while Uno's C-index, which is free of censoring, was chosen to evaluate the discriminative performance of the model at the 3-year time point [29, 30]. Because a multicovariate Cox model was used, the bias-corrected C-index which overcomes the problem of overfitting was obtained through bootstrapping while evaluating the overall discriminative performance and used for internal validation. For external validation, stratified bootstrapping was used to ensure that

in each iteration step, the resampled data included the event. Wilcoxon signed-rank test between the 200 C-indexes generated from the 200 iterations of bootstrapping procedure were performed to quantify the discriminative difference of the C-index between different models. External validation of the models included both the overall discriminative performance and discriminative performance at the 3-year time point.

All statistical analyses were performed using R statistical software (version 3.5.3). Construction of Cox models and the nomogram and internal validation of Harrell's C-index and calibration plot were performed with the rms R package. External validation of Harrell's C-index was performed with the Hmisc R package. Uno's C-index was calculated with the survC1 R package. The Wilcoxon signed-rank test was performed with the stats R package. The random survival forest was performed with the randomForestSRC R package.

Results

Discovery of clinical prognostic supplementary factors

Though the random survival forest, a total of 14 pathways were screened as potential clinical prognostic supplementary factors as shown in Table 2. After further screening through the multicovariate Cox model, 27 combinations of different pathways were found to have significant prognostic effects in the integrated models. After comparing the bias-corrected C-indexes of these 27 different clinical-molecular integrated models, we found that models that included the pathway hsa0032 tended to have better discriminative performance as shown in Table 3. In addition, the prognostic model that incorporated only pathway hsa00532 and clinical prognostic factors had a considerable performance, while the incorporation of other pathways alone showed no significant effect on prognosis in the integrated models. Therefore, we consider the PDS of pathway has00532 to be the best potential clinical prognostic supplementary factor among these 14 pathways. In this study, 16 genes were included in the analysis: XYLT1, XYLT2, B4GALT7, B3GALT6, B3GAT3, CSGALNACT1, CSGALNACT2, CHSY1, CHPF, CHPF2, DSE, CHST11, CHST12, CHST3, CHST15 and CHST14. The other 4 genes, CHSY3, CHST13, CHST7 and UST, were removed during data preprocessing. These 16 genes were used for gene-based model construction in subsequent analyses.

Observation of the distribution of the PDS of pathway hsa00532 in the discovery set suggested that it approximately obeyed a normal distribution as shown in Figure 2A. Therefore, patients in the discovery set were divided into a high-PDS group and low-PDS group. A threshold of 0.6779 was found to most clearly separate these two groups on the Kaplan-Meier (KM) curve, while several peaks at approximately less than 0.5 of the density distribution led us to separate patients into three groups according to thresholds of 0.5 and 0.6779 as shown in Figure 2B. Detailed KM curves of patients in the discovery set with different TNM stages are shown in Figure 2 C-F. Patients in the validation set were also divided into two groups with the same threshold 0.6779, and the corresponding KM curves are shown in Figure 3 B-F.

Constructed knowledge-based clinical-molecular integrated prognostic model

With the identified knowledge-based prognostic factor, the PDS of pathway hsa00532 and clinical prognostic factors T, N and M stage, our knowledge-based clinical-molecular integrated prognostic model was built. To assess the improvement of our model compared to clinical prognostic model, a corresponding clinical prognostic model based on T, N and M stage and molecular prognostic model based on the PDS of pathway has00532 were constructed. The multicovariate Cox model was used to determine the regression coefficients of the models, with the coefficients of the knowledge-based clinical-molecular integrated prognostic model summarized in Table 4 and regression coefficients of the other models summarized in Additional file 1. A corresponding nomogram that predicts 3-year overall survival was constructed and is shown in Figure 4.

Assessment of the prognostic models

The discriminative performance of different models was measured with both Harrell's C-index for overall performance and Uno's C-index for performance at specific time points. In the internal validation, our model outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.773 vs 0.746, $p<0.0001$) and the molecular prognostic model (0.773 vs 0.619, $p<0.0001$) as shown in Figure 5A. In the external validation, our model again outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.893 vs 0.808, $p<0.0001$) and the molecular prognostic model (0.893 vs 0.810, $p<0.0001$) as shown in Figure 5B.

The prognostic performance of our model was assessed by Uno's C-index and calibration plot at the 3-year time point. The 3-year Uno's C-index in the discovery set suggested that our model has the best discriminative performance compared to the clinical model (0.793 vs 0.762, $p<0.0001$) and molecular model (0.793 vs 0.619, $p<0.0001$) as shown in Figure 5C, whereas in the validation set, the comparison results were 0.899 vs 0.816 ($p<0.0001$) compared to the clinical model and 0.899 vs 0.816 ($p<0.0001$) compared to the molecular model as shown in Figure 5D. The calibration plot of these models also showed that our model has a superior calibration performance compared with the clinical model at 3-year time point as shown in Figure 5E and Figure 5F.

Because the mean survival time of metastatic patients is shorter than that of nonmetastatic patients, the prognostic performance of the prognostic models may be affected. Therefore, the clinical prognostic model, molecular prognostic model and clinical-molecular integrated prognostic model were constructed with the same prognostic factors used for nonmetastatic patients. The same assessments were performed on these models for nonmetastatic patients, with nonmetastatic patients in the discovery set and validation set. In the internal validation, the integrated model outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.712 vs 0.665, $p<0.0001$ for Harrell's C-index, 0.763 vs 0.709, $p<0.0001$ for the 3-year Uno's C-index) and the molecular prognostic model (0.712 vs 0.655, $p<0.0001$ for Harrell's C-index, 0.763 vs 0.659, $p<0.0001$ for the 3-year Uno's C-index) as

shown in Figure 6A and Figure 6C. In the external validation, the integrated model again outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.824 vs 0.720, p<0.0001 for Harrell's C-index, 0.829 vs 0.743, p<0.0001 for the 3-year Uno's C-index) and the molecular prognostic model (0.824 vs 0.791, p<0.0001 for Harrell's C-index, 0.829 vs 0.799, p<0.0001 for the 3-year Uno's C-index) as shown in Figure 6B and Figure 6D.

Pathway-based model is superior to the gene-based model

Previous studies have claimed that the introduction of representative functional units should improve gene expression-based studies [8-13]. Therefore, genes involved in pathway hsa00532 were used to construct a gene-based clinical-molecular integrated prognostic model. The regression coefficients of the gene-based model suggested that only two genes (CSGALNACT1 and DSE) were prognostically related when combined with clinical factors, and they are were summarized in Table S3 of Additional file 1. Compared to our knowledge-based integrated model, the C-indexes of the gene-based integrated model in the discovery set were lower (0.721 vs 0.773 (p<0.0001) for the bias-corrected C-index, and 0.783 vs 0.793 (p<0.0001) for the 3-year Uno's C-index in the discovery set, and 0.825 vs 0.893 (p<0.0001) for Harrell's C-index and 0.826 vs 0.899 (p<0.0001) for the 3-year Uno's C-index in the validation set as shown in Figure 5A, Figure 5C, Figure 5B and Figure 5D, respectively). These results suggest that the pathway-based integrated model is superior to the gene-based integrated model in discriminative performance because the gene-based integrated model might include too many redundant prognostic factors.

Discussion

Through knowledge-based clinical-molecular integrated analysis by the random survival forest and multicovariate Cox model, this study successfully discovered the PDS of pathway hsa00532 was the best clinical prognostic supplementary factor. The results of internal validation and external validation suggested that the knowledge-based clinical-molecular integrated prognostic model had the best discriminative performance and an improved calibration performance than the clinical prognostic model for the discovery set, and our model was better than the other models for the validation set as well. In addition, the pathway-based models were superior to the gene-based model, which indicates that the incorporation of pathway information can make more use of the expression information of genes involved in a pathway.

The pathway hsa00532 is named Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate on the KEGG website, and it is related to the biosynthesis of chondroitin sulfate and dermatan sulfate. Previous studies have indicated that the dermatan sulfate chain is different between colon cancer and normal colonic mucosa, and chondroitin sulfate is associated with tumor metastasis [31-34]. However, the PDS of pathway hsa00532 showed no relevance to the metastasis status, with a Pearson correlation coefficient of 0.04 for the discovery dataset. In addition, pathway hsa00532 showed considerable supplementary power in the models for both all-stage and nonmetastatic colon cancer

patients, while the metastasis status and the PDS of pathway hsa00532 were regarded as independent significant prognostic factors in the constructed model. One possible explanation for this finding is that although the PDS in this study is generated from gene expression data, a series of regulatory and expression biology processes from the transcriptome is still needed generate the actual pathway products. Two genes CSGALNACT1 and DSE involved in pathway hsa00532 would be potential markers for colon cancer prognosis because only these two genes showed a potential prognostic effect in the gene-based integrated model. Further validation is required to validate the prognostic effect of these two genes as currently published papers have not mentioned them in conjunction with colon cancer prognosis.

The distribution of the PDS of pathway hsa00532 in the discovery set approximately obeyed a normal distribution, with several peaks at approximately less than 0.5 as shown in Figure 2A. Observation of the KM curves based on patient groups divided by thresholds of 0.5 and 0.6779 suggested that patients with higher PDSs had worse survival, which was also the conclusion drawn from nonmetastatic patients as shown in Figure 2B-E. The KM curves for stage I, II, III and IV patients divided by a threshold of 0.6779 are shown in Figure 2C-F. The results suggested that stage II patients could be distinctly distinguished and that the PDS of pathway hsa00532 could be a potential biomarker in separating high-risk stage II colon cancer patients. In addition, the KM curves for stage I, II, III and IV patients divided by a threshold of 0.6779 in the validation set are shown in Figure 3C-F. These results further confirmed that a higher PDS of pathway hsa00532 suggests an increased risk of colon cancer.

There are still limitations in this study. The clinical prognostic factors in this study involved only the T, N, and M stages, while in the actual clinical treatment of colon cancer, there are many other factors need to be considered, such as the patient's physical condition and the chemotherapy or radiotherapy regimen. In addition, due to the short follow-up time of the validation set, it was not possible to further validate the performance of our model on long-term prognosis. Through cooperation with local hospitals, we can collect more real world follow-up patients, and sequence their tumor samples to generate more molecular data. Therefore, further validation of our model could be conducted. In addition, the involvement of more clinical prognostic factors in clinical-molecular analysis could make more detailed and specific supplement to clinical prognosis. In addition, considering that the PDS of pathway hsa00532 can effectively distinguish the risk of stage II patients, further research and validation should be performed with more data. After further validation with real data, further research related to the PDS of pathway has00532, such as immunohistochemistry and other methods appropriate for clinical use should be conducted. Current study used only gene expression data from the transcriptome, and the addition of other types of omics data, such as genome or epigenome data, may further improve the accuracy of molecular features and better supplement the clinical prognosis. However, with the current technology, how to balance the improvement in discriminative performance and the cost of sequencing remains to be considered.

Conclusion

In conclusion, this study discovered that the PDS of pathway hsa00532 can be used as a supplementary prognostic factor for the three clinical prognostic factors T, N, and M stages. The clinical-molecular integrated prognostic model constructed with these three clinical prognostic factors and the discovered molecular prognostic factor is superior to the clinical prognostic model, molecular prognostic model or gene-based integrated prognostic model in discriminative performance. In addition, the PDS of pathway hsa00532 showed a significant ability to distinguish high risk stage II colon cancer patients and is a potential prognostic marker. The PDS calculation of pathway hsa00532 involves only 16 genes; therefore, it has good prospects for clinical use after further verification with real data.

Abbreviations

AJCC	**American Joint Committee on Cancer**
TNM	Tumor, Node, and Metastasis
NCCN	National Comprehensive Cancer Network
MSI	Microsatellite Instability
MMR	Mismatch Repair
TCGA	The Cancer Genome Atlas
COAD	Colon Adenocarcinoma
CPTAC	Clinical Proteomic Tumor Analysis Consortium
FPKM-UQ	fragments per kilobase of transcript per million mapped reads upper quartile
RSEM	RNA-Seq by expectation maximization
HGNC	HUGO Gene Nomenclature Committee
SD	Standard deviation
KEGG	Kyoto Encyclopedia of Genes and Genomes
PDS	pathway dysregulation score
C-index	concordance index
KM	Kaplan-Meier
CI	confidence interval
SE	standard error

Declarations

Ethics approval and consent to participate: The TCGA dataset used and analyzed in this study are unrestricted-access, which are available without any permission request through the cBioPortal

(http://www.cbioportal.org) and UCSC Xena (http://xena.ucsc.edu/). We definitely followed the National Institutes of Health Genomic Data Sharing Policy as well as the National Cancer Institution Genomic Data Sharing Policy in this study. The CPTAC dataset used and analyzed in this study are available through the LinkedOmics (http://linkedomics.org/cptac-colon/). These data are not limited for publication as they meet one of the freedom-to-publish criteria requested by the CPTAC, which is “A global analysis publication paper has been published on that tumor type or sample set”.

Consent for publication: Not applicable.

Availability of data and material: The dataset analyzed during the current study is available in the cBioPortal, http://www.cbioportal.org; generated by the National Cancer Institute CPTAC and available in the LinkedOmics, http://linkedomics.org/cptac-colon/; and available in the UCSC Xena, http://xena.ucsc.edu/.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by the National Natural Science Foundation of China [No. 81771936 to JSL, No. 81801796 to YT], Major Scientific Project of Zhejiang Lab [2018DG0ZX01 to JSL], and the National Key Research and Development Program of China [No. 2018YFC011690 to JSL], which provided financial support in the design of study, analysis of data and writing the manuscript; the National Natural Science Foundation of China [No. 81672916 to JL] and Fundamental Research Funds for the Central Universities [2020QNA5031 to YT], which provide financial support in the interpretation of the data.

Author contributions: DT, YT and JL contributed to the conception of the study. DT, YT and QY performed the data preparation. DT and YT performed the data analyses and wrote the manuscript. JSL and KD provided critical revisions. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [No. 81771936, No. 81801796 and No. 81672916], Major Scientific Project of Zhejiang Lab [2018DG0ZX01], the National Key Research and Development Program of China [No. 2018YFC011690] and Fundamental Research Funds for the Central Universities [2020QNA5031]. In addition, we acknowledge the TCGA Research Network and all staff members involved in the TCGA Project, the CPTAC Project, the cBioPortal Project and the UCSC Xena Project for providing the data.

References

1. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA: a cancer journal for clinicians. 2016;66(2):115-32.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA-Cancer J Clin. 2019;69(1):7-34.

3. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet. 2018;391(10125):1023-75.
4. Li J, Guo BC, Sun LR, Wang JW, Fu XH, Zhang SZ, et al. TNM staging of colorectal cancer should be reconsidered by T stage weighting. World journal of gastroenterology. 2014;20(17):5104-12.
5. Benson AB, Venook AP, Cederquist L, Chan E, Chen Y-J, Cooper HS, et al. Colon Cancer, Version 1.2017, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network. 2017;15(3):370-98.
6. Giuliano AE, Edge SB, Hortobagyi GN. Eighth Edition of the AJCC Cancer Staging Manual: Breast Cancer. Annals of surgical oncology. 2018;25(7):1783-5.
7. Weiser MR. AJCC 8th Edition: Colorectal Cancer. Annals of surgical oncology. 2018;25(6):1454-5.
8. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genet. 2003;34(3):267-73.
9. 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. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-50.
10. Bild AH, Yao G, Chang JT, Wang QL, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006;439(7074):353-7.
11. van den Akker EB, Passtoors WM, Jansen R, van Zwet EW, Goeman JJ, Hulsman M, et al. Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein- protein interaction modules as robust markers of human aging. Aging Cell. 2014;13(2):216-25.
12. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. The Lancet. 2018;391(10125):1023-75.
13. Ma SG, Kosorok MR, Huang JA, Dai Y. Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. BMC Med Genomics. 2011;4:10.
14. Fa BT, Luo CW, Tang Z, Yan YT, Zhang Y, Yu ZS. Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma. EBioMedicine. 2019;44:250-60.
15. Jiang H, Du J, Gu JM, Jin LG, Pu Y, Fei BJ. A 65-gene signature for prognostic prediction in colon adenocarcinoma. Int J Mol Med. 2018;41(4):2021-7.
16. Lee JH, Jung S, Park WS, Choe EK, Kim E, Shin R, et al. Prognostic nomogram of hypoxia-related genes predicting overall survival of colorectal cancer-Analysis of TCGA database. Sci Rep. 2019;9:9.

17. Zhou R, Zeng D, Zhang J, Sun H, Wu J, Li N, et al. A robust panel based on tumour microenvironment genes for prognostic prediction and tailoring therapies in stage I-III colon cancer. EBioMedicine. 2019;42:420-30.
18. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005;21(2):171-8.
19. Bae JM, Kim JH, Kwak Y, Lee DW, Cha Y, Wen X, et al. Distinct clinical outcomes of two CIMP-positive colorectal cancer subtypes based on a revised CIMP classification system. British journal of cancer. 2017;116(8):1012-20.
20. Chaudharyl K, Poirionl OB, Lu LQ, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clinical Cancer Research. 2018;24(6):1248-59.
21. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173(2):291-304.e6.
22. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell. 2019;177(4):1035-49 e19.
23. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(16):6388-93.
24. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu JC, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237-i45.
25. Huang SJ, Yee C, Ching T, Yu H, Garmire LX. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. PLoS Comput Biol. 2014;10(9): e1003851.
26. Liu C, Srihari S, Lal S, Gautier B, Simpson PT, Khanna KK, et al. Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. Molecular Oncology. 2016;10(1):179-93.
27. Hastie T, Stuetzle W. PRINCIPAL CURVES. J Am Stat Assoc. 1989;84(406):502-16.
28. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. RANDOM SURVIVAL FORESTS. Ann Appl Stat. 2008;2(3):841-60.
29. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. EVALUATING THE YIELD OF MEDICAL TESTS. JAMA-J Am Med Assoc. 1982;247(18):2543-6.
30. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine. 2011;30(10):1105-17.
31. Iida J, Meijne AM, Knutson JR, Furcht LT, McCarthy JB. Cell surface chondroitin sulfate proteoglycans in tumor cell adhesion, motility and invasion. Semin Cancer Biol. 1996;7(3):155-62.
32. Lee CM, Tanaka T, Murai T, Kondo M, Kimura J, Su W, et al. Novel chondroitin sulfate-binding cationic liposomes loaded with cisplatin efficiently suppress the local growth and liver metastasis of tumor

cells in vivo. Cancer Res. 2002;62(15):4282-8.

33. Fuster MM, Esko JD. The sweet and sour of cancer: glycans as novel therapeutic targets. Nature reviews Cancer. 2005;5(7):526-42.
34. Daidouji K, Takagaki K, Yoshihara S, Matsuya H, Sasaki M, Endo M. Neoplastic changes in saccharide sequence of dermatan sulfate chains derived from human colon cancer. Dig Dis Sci. 2002;47(2):331-7.

Tables

Table 1: Detailed information of the data used for analysis.

	Discovery Set	Validation Set	Reference Set
Characteristic	**TCGA-COAD**	**CPTAC**	**Normal Samples**
	(%)	**(%)**	**(%)**
Patients	374	98	41
Survival status	Alive: 293 (78.3)	Alive: 90 (91.8)	Alive: 29 (70.7)
	Dead: 81 (21.7)	Dead: 8 (8.2)	Dead: 12 (29.3)
Age*	Mean: 66.75	Mean: 65.43	Mean: 70.34
	SD: 12.73	SD: 11.56	SD: 13.23
	Range: 31-90	Range: 35-93	Range: 40-90
Gender	Male: 199 (53.2)	Male: 41 (41.8)	Male: 20 (48.8)
	Female: 175 (46.8)	Female: 57 (58.2)	Female: 21 (51.2)
Survival time	Min: 0.47	Min: 1	Min: 0
(Overall survival time in months)	Mean: 30.24	Mean: 27.96	Mean: 27.66
	Median: 24.27	Median: 30	Median: 24.37
	Max: 150.07	Max: 44	Max: 101.40
T stage	T1: 9 (2.4) T2: 65 (17.4)	T1: 0 (0) T2: 12 (12.2)	Not available
	T3: 258 (69.0)	T3: 73 (74.5)	
	T4: 42 (11.2)	T4: 13 (13.3)	
N stage	N0: 226 (60.4) N1: 84 (22.5)	N0: 52 (53.1) N1: 31 (31.6)	Not available
	N2: 64 (17.1)	N2: 15 (15.3)	
M stage	M0: 315 (84.2)	M0: 91 (92.9)	Not available
	M1: 59 (15.8)	M1: 7 (7.1)	
Number of genes	10887	10887	10887

*: The characteristic ‘Age’ refers to the age at initial diagnosis in the discovery set and reference set but the age at time of tissue procurement in the validation set.

SD: standard deviation

Table 2: Description of the pathways identified by the random survival forest.

KEGG Pathway ID	Pathway Name
hsa00450	Selenocompound metabolism - Homo sapiens (human)
hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate - Homo sapiens (human)
hsa02010	ABC transporters - Homo sapiens (human)
hsa04380	Osteoclast differentiation - Homo sapiens (human)
hsa04614	Renin-angiotensin system - Homo sapiens (human)
hsa04750	Inflammatory mediator regulation of TRP channels - Homo sapiens (human)
hsa04911	Insulin secretion - Homo sapiens (human)
hsa04971	Gastric acid secretion - Homo sapiens (human)
hsa04975	Fat digestion and absorption - Homo sapiens (human)
hsa05032	Morphine addiction - Homo sapiens (human)
hsa05133	Pertussis - Homo sapiens (human)
hsa05152	Tuberculosis - Homo sapiens (human)
hsa05167	Kaposi sarcoma-associated herpesvirus infection - Homo sapiens (human)
hsa05321	Inflammatory bowel disease (IBD) - Homo sapiens (human)

Pathway hsa00532 (shown in bold) was identified as the best prognostic supplementary factor.

Table 3: Bias-corrected C-indexes of the 27 different clinical-molecular integrated models.

Covariates Used in the Model	Bias-corrected Harrell's C-index (± 95% CI)
T, N, M, hsa00532, hsa04911, hsa05133, hsa05152	0.775 ± 0.0038
T, N, M, hsa00532	0.773 ± 0.0038
T, N, M, hsa00532, hsa04911, hsa05133	0.773 ± 0.0038
T, N, M, hsa02010, hsa05152, hsa05321	0.773 ± 0.0038
T, N, M, hsa02010, hsa05167, hsa05321	0.772 ± 0.0037
T, N, M, hsa00532, hsa04380, hsa04911, hsa05133	0.772 ± 0.0039
T, N, M, hsa00532, hsa05133, hsa05152	0.772 ± 0.0040
T, N, M, hsa00532, hsa04380, hsa04971, hsa05133	0.771 ± 0.0039
T, N, M, hsa00532, hsa05133, hsa05167	0.771 ± 0.0040
T, N, M, hsa00532, hsa04975, hsa05133	0.770 ± 0.0041
T, N, M, hsa02010, hsa04911, hsa05133, hsa05152	0.768 ± 0.0040
T, N, M, hsa02010, hsa05133, hsa05152	0.768 ± 0.0040
T, N, M, hsa00532, hsa04971, hsa05133	0.767 ± 0.0038
T, N, M, hsa00532, hsa04380, hsa05133	0.766 ± 0.0041
T, N, M, hsa02010, hsa04911, hsa05133	0.764 ± 0.0038
T, N, M, hsa02010, hsa05133, hsa05167	0.764 ± 0.0040
T, N, M, hsa00532, hsa04750, hsa05133	0.764 ± 0.0041
T, N, M, hsa02010, hsa04911	0.763 ± 0.0037
T, N, M, hsa02010, hsa04380, hsa04911, hsa05133	0.763 ± 0.0040
T, N, M, hsa05133, hsa05152	0.761 ± 0.0042
T, N, M, hsa02010, hsa04750, hsa05133	0.759 ± 0.0040
T, N, M, hsa00450, hsa04911	0.758 ± 0.0036
T, N, M, hsa04380, hsa05133	0.758 ± 0.0043
T, N, M, hsa04911, hsa05133	0.756 ± 0.0039
T, N, M, hsa00450, hsa04911, hsa05133, hsa05152	0.756 ± 0.0040
T, N, M, hsa04380, hsa04911, hsa05133	0.754 ± 0.0041
T, N, M, hsa05133, hsa05167	0.754 ± 0.0042

T, N, M	0.746 ± 0.0040

CI: confidence interval

Table 4: Regression coefficients of the knowledge-based clinical-molecular integrated prognostic model.

Covariate	Coefficient ± SE	HR	95% CI	P
T Stage	-1.62 ± 1.42	0.20	0.012-3.18	0.25
T2	0.38 ± 1.02	1.47	0.20-10.79	0.71
T3	1.22 ± 1.05	3.38	0.43-26.37	0.25
T4				
N Stage	-0.01 ± 0.31	0.99	0.54-1.80	0.96
N1	0.67 ± 0.30	1.95	1.07-3.54	0.028
N2				
M Stage	1.02 ± 0.28	2.78	1.60-4.82	0.00029
M1				
Hsa00532*	2.86 ± 1.42	17.53	1.08-283.24	0.044

HR: hazard ratio

*Covariate Hsa00532 used in the model is the PDS of pathway has00532.

Figures

Figure 1

Overall pipeline of the study.

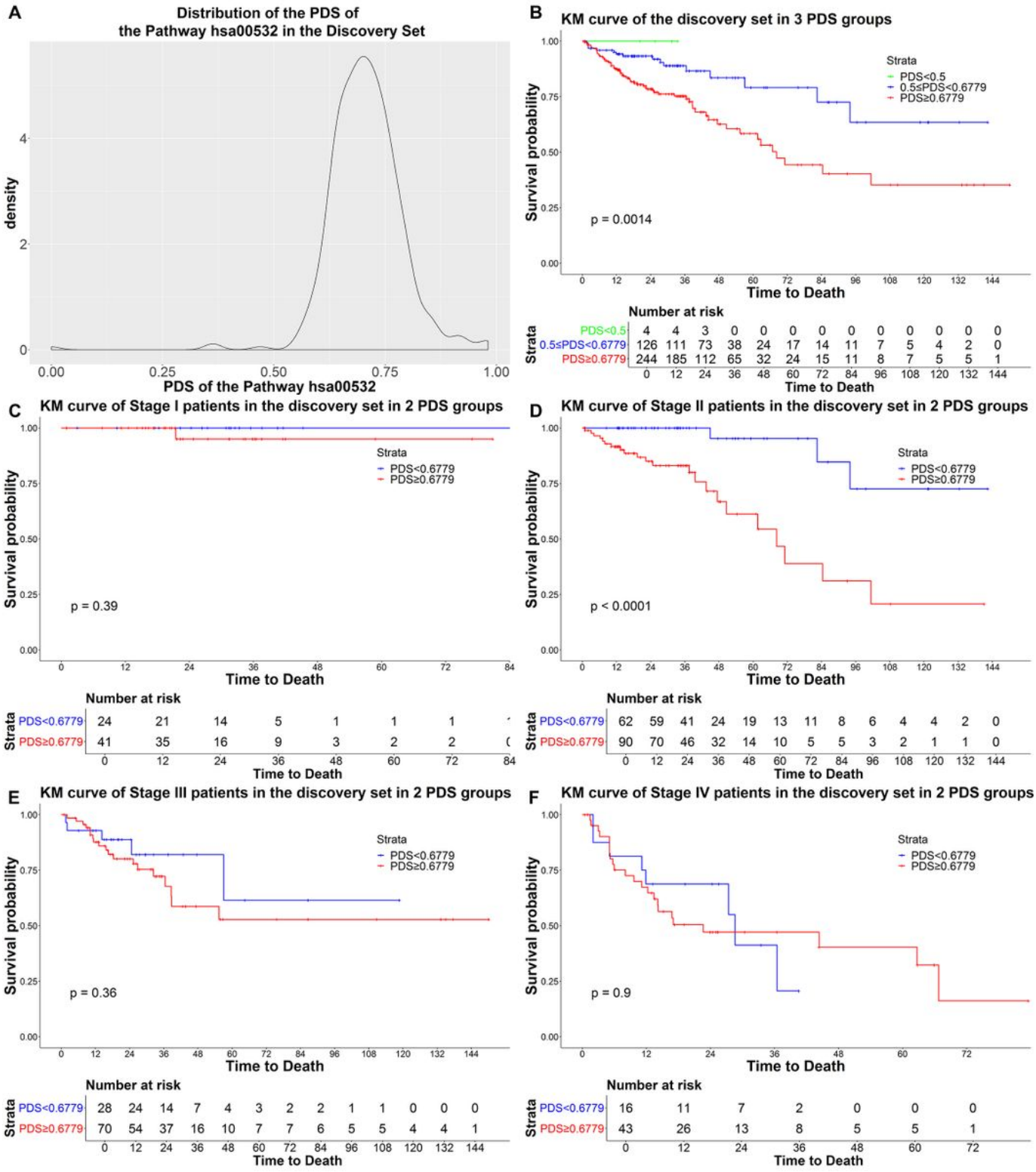

Figure 2

Observation of the PDS of pathway hsa00532 in the discovery set. A: Density distribution of the PDS of pathway hsa00532 in the discovery set. B: KM curve plotted based on three groups of patients in the discovery set divided by the PDS with thresholds of 0.5 and 0.6779. C: KM curve plotted based on two groups of stage I patients in the discovery set divided by the PDS with a threshold of 0.6779. D: KM curve plotted based on two groups of stage II patients in the discovery set divided by the PDS with a threshold

of 0.6779. E: KM curve plotted based on two groups of stage III patients in the discovery set divided by the PDS with a threshold of 0.6779. F: KM curve plotted based on two groups of stage IV patients in the discovery set divided by the PDS with a threshold of 0.6779.

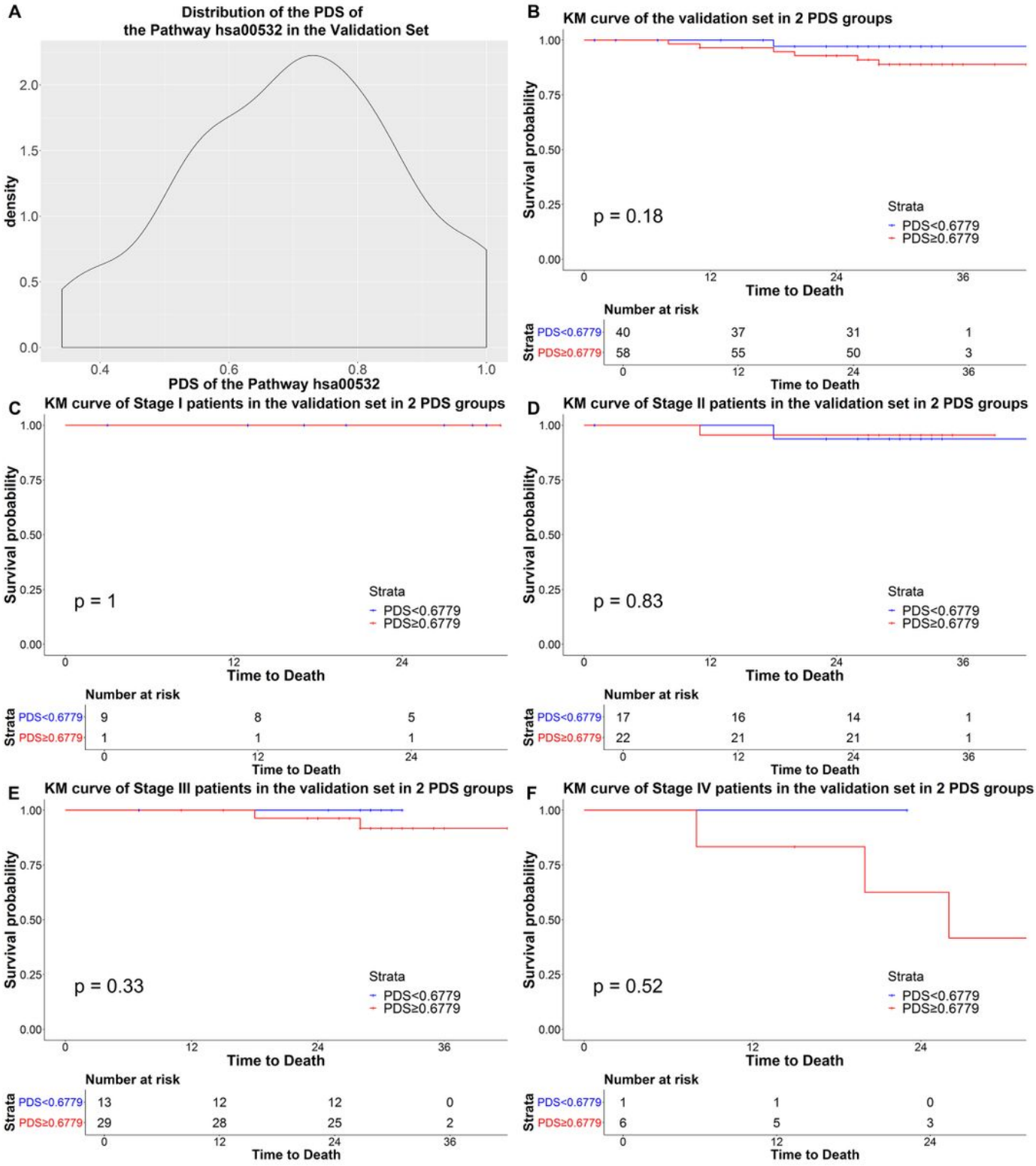

Figure 3

Observation of the PDS of pathway hsa00532 in the validation set. A: Density distribution of the PDS of pathway hsa00532 in the validation set. B: KM curve plotted based on two groups of patients in the

validation set divided by the PDS with a threshold of 0.6779. C: KM curve plotted based on two groups of stage I patients in the validation set divided by the PDS with a threshold of 0.6779. D: KM curve plotted based on two groups of stage II patients in the validation set divided by the PDS with a threshold of 0.6779. E: KM curve plotted based on two groups of stage III patients in the validation set divided by the PDS with a threshold of 0.6779. F: KM curve plotted based on two groups of stage IV patients in the validation set divided by the PDS with a threshold of 0.6779.

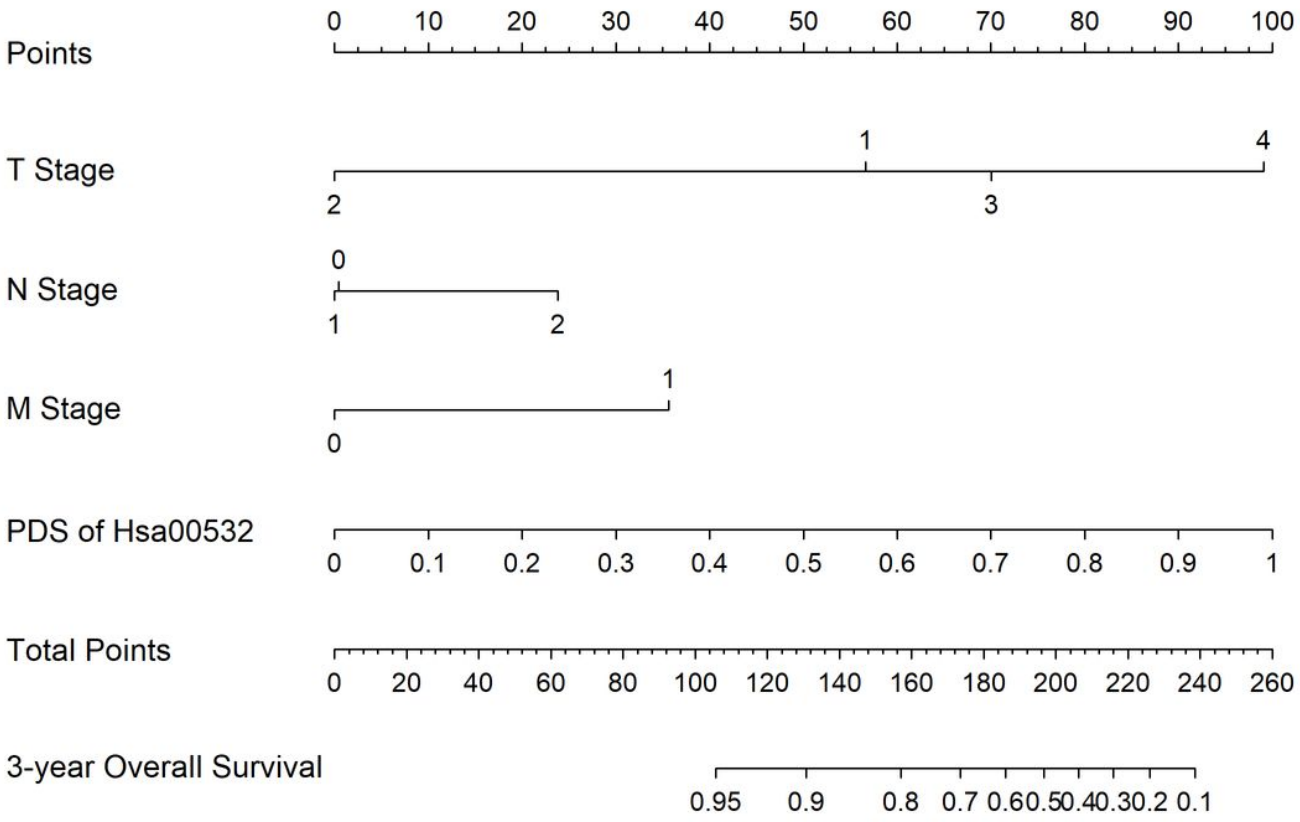

Figure 4

Nomogram predicting the 3-year overall survival of colon cancer patients.

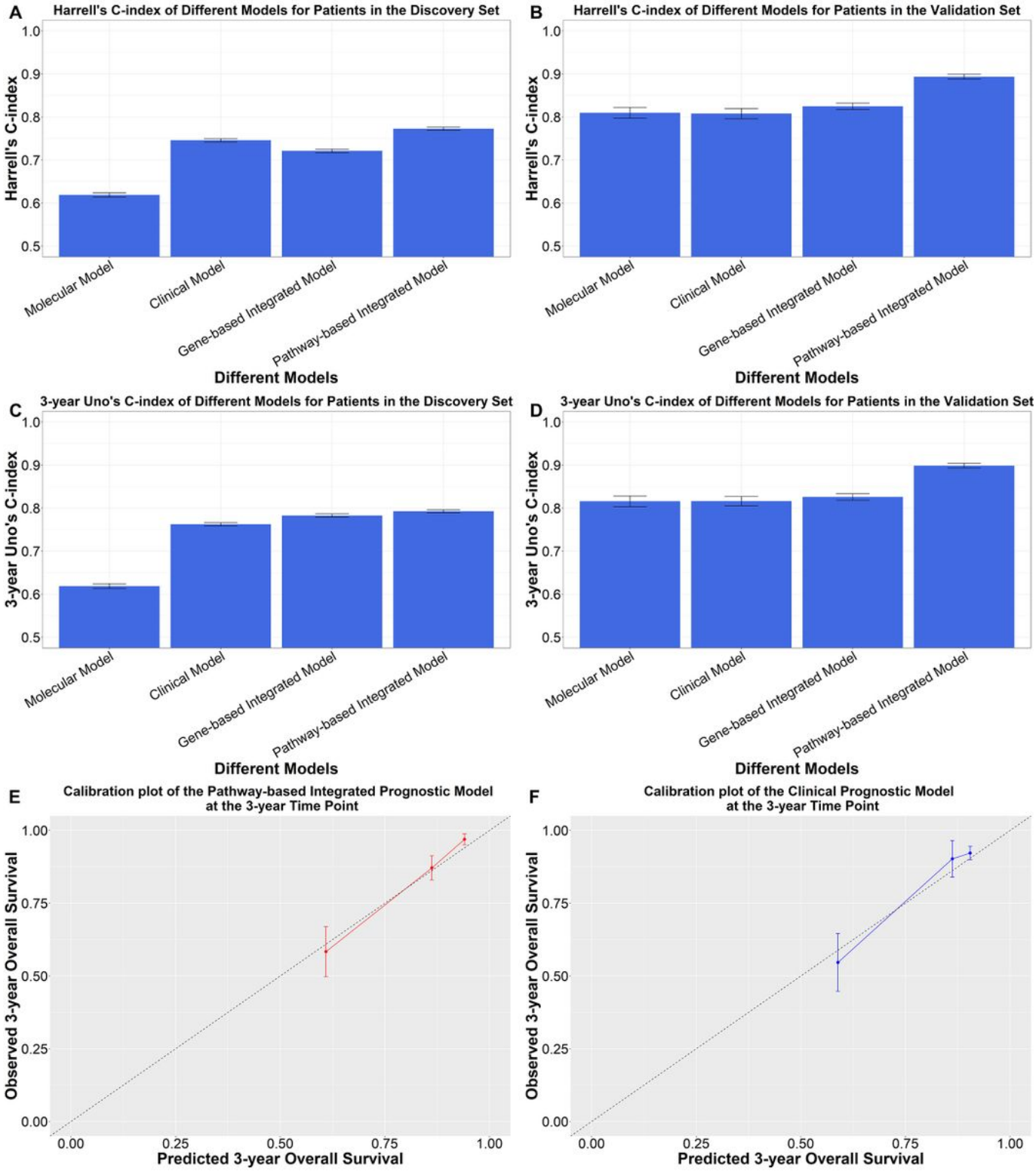

Figure 5

Prognostic performance of different models. A: Bias-corrected Harrell's C-index of our pathway-based integrated model and other models for the discovery set. B: Harrell's C-index of our pathway-based integrated model and other models for the validation set. C: 3-year Uno's C-index of our pathway-based integrated model and other models for the discovery set. D: 3-year Uno's C-index of our pathway-based

integrated model and other models for the validation set. E: calibration plot of our pathway-based integrated model at the 3-year time point. F: calibration plot of clinical model at the 3-year time point.

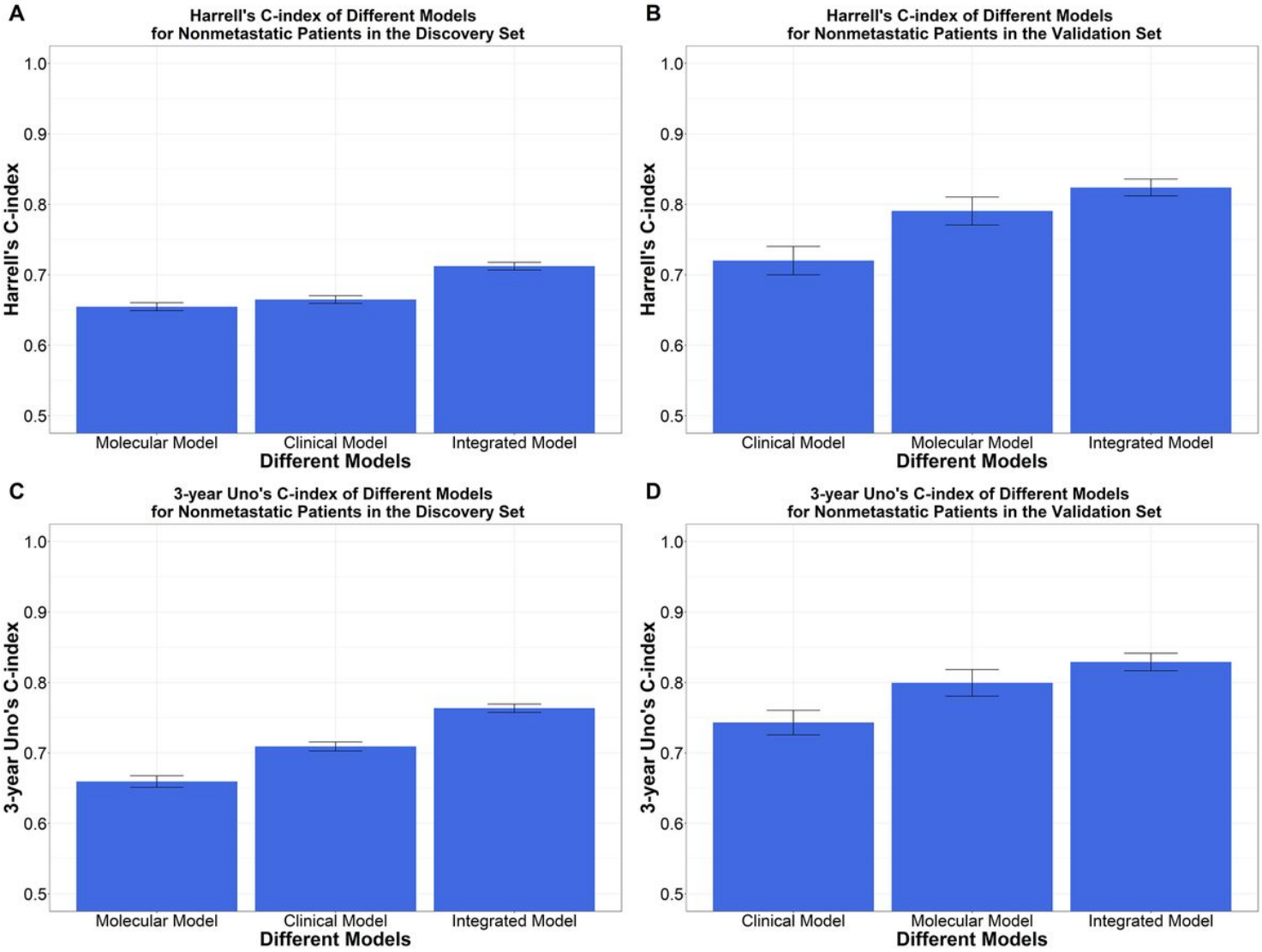

Figure 6

Prognostic performance of different models for nonmetastatic patients. A: Bias-corrected Harrell's C-index of our pathway-based integrated model and other models for nonmetastatic patients in the discovery set. B: Harrell's C-index of our pathway-based integrated model and other models for nonmetastatic patients in the validation set. C: 3-year Uno's C-index of our pathway-based integrated model and other models for nonmetastatic patients in the discovery set. D: 3-year Uno's C-index of our pathway-based integrated model and other models for nonmetastatic patients in the validation set.

Supplementary Files

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

- Additionalfile1.docx