Prognostic and biological relevance of collagen-related genes in prostate cancer

Supplementary material, transcriptome part

2025-02-25

# Supplementary Methods for transcriptome analyses

## Software

The analysis was done with R version 4.2.3 (R Foundation).

Tabular data were handled with the packages *tidyverse* (1), *rlang* (2) and [*trafo*](https://github.com/PiotrTymoszuk/trafo). Text data were handled with *stringi* (3). For transcormation of numeric variables and variable distribution analyses, package [*microViz*](https://github.com/PiotrTymoszuk/microViz) was used.

Import of the TCGA and DKFZ data sets from the cBioportal repository was accomplished with in-house-developed R scripts. Transcriptome data sets from the Gene Expression Omnibus were fetched with the *GEOquery* package (4). Gene and probe annotation was accomplished with the *AnnotationDbi* (5) and *org.Hs.eg.db* packages (6). ComBat adjustment of transcript expression levels for batch effects (7) was accomplished with [*htGLMNET*](https://github.com/PiotrTymoszuk/htGLMNET) package. Training/test cohort split was accomplished with *caret* (8).

For descriptive statistics, statistical hypothesis testing, effect size calculation, and correlation analyses, packages [*ExDA*](https://github.com/PiotrTymoszuk/ExDA) and [*fastTest*](https://github.com/PiotrTymoszuk/fastTest) were used. Co-expression network analyses were performed with packages *igraph* (9) and [*graphExtra*](https://github.com/PiotrTymoszuk/graphExtra).

Multi-parameter modeling of biochemical relapse (BCR) free survival and overall survival was done with the packages *glmnet* (10) and [*htGLMNET*](https://github.com/PiotrTymoszuk/htGLMNET) (Ridge, Elastic Net and LASSO cox regression), *survivalsvm* (11) (Support Vector Machines [SVM]), *rfsrc* (12,13) (Random Forest), *gbm* (14) (Gradient Boosted Machines [GBM]), *survival* (15), *surviminer* (16),  
[*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions), and *caret* (8). Univariable analysis of relapse-free survival was accomplished with packages [*kmOptimizer*](https://github.com/PiotrTymoszuk/kmOptimizer'), *survival*, *survminer* and [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions).

For visualization of the results, the packages *ggplot2* (1) (scatter, dot, bar and Forest plots), *survminer*, [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions), [*kmOptimizer*](https://github.com/PiotrTymoszuk/kmOptimizer') (Kaplan-Meier plots, plots of Brier scores), and [*graphExtra*](https://github.com/PiotrTymoszuk/graphExtra) (network plots) were used. Figures and tables were created with the packages *cowplot* (17) and *flextable* (18), respectively.

Parts of the manuscript and Supplementary Material were written in *rmarkdown* environment (19) with *bookdown* package (20), and rendered as Word files with *knitr* (21). Figure and table objects in the *rmarkdown* documents were managed with [*figur*](https://github.com/PiotrTymoszuk/figur).

## Data sources and data import

Transcriptome cohorts were selected from studies deposited at cBioportal and Gene Expression Omnibus (GEO) with the following criteria: biochemical relapse and biochemical relapse-free survival or death and overall survival information, and expression data for collagen-related genes investigated in our previous collagen project (22).

The [TCGA](https://www.cbioportal.org/study/summary?id=prad_tcga_pan_can_atlas_2018) prostate cancer data set (23,24) and [DKFZ](https://www.cbioportal.org/study/summary?id=prostate_dkfz_2018) data set (25) consisted of normalized RNA sequencing data for 493 and 118 cancer samples, respectively, with accompanying clinical information and were obtained from the cBioportal repository with in-house-developed R scripts.

The [GSE16560](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16560) (26) (microarray, Human 6k Transcriptionally Informative Gene Panel for DASL, n = 281 cancer samples), GSE54460(<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54460>) (27) (RNA sequencing, n = 106 cancer samples), [GSE70768](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70768) (28) (Illumina HumanHT-12 V4.0 expression beadchip microarray, n = 125), [GSE70769](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70769) (28) (Illumina HumanHT-12 V4.0 expression beadchip microarray, n = 94), and [GSE220095](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE220095) (29) (RNA sequencing, n = 176) data sets with normalized whole-transcriptome and basic clinical information were fetched from GEO with the getGEO() function from the *GEOquery* package.

Whole transcriptome expression levels were transformed with the and function for, respectively, microarray and RNA sequencing data sets, and subjected to adjustment for batch cohort effects with the ComBat algorithm (7) (function multi\_process(), package [*htGLMNET*](https://github.com/PiotrTymoszuk/htGLMNET)). Subsequently, the ComBat-adjusted log2-transformed expression data and clinical information from GSE54460, GSE70768, GSE70769, and GSE220095 data sets was merged into a ‘pooled GEO cohort’. The pooled GEO cohort, TCGA, and DKFZ cohorts were used in co-expression network analyses, as well as uni- and multi-parameter modeling of BCR-free survival. In the multi-parameter modeling of BCR-free survival, the pooled GEO cohort was used as the training collective, while the TCGA and DKFZ cohorts were used for model validation. The GSE16560 cohort with all-case mortality information and survival times was used for multi-parameter modeling of overall survival. The models of overall survival were developed in a training portion of the GSE16560 cohort with 2/3 of cancer samples, and validated in the remaining tumor specimens. The training/test cohort split was done with function createDataPartition() from *caret* package.

Characteristics of the GSE16560, pooled GEO, TCGA, and DKFZ cohort are presented in **Supplementary Table S1**. Characteristics of the GEO data sets that constituted the pooled GEO cohort are listed in **Table 2**. Characteristics of the training and test subsets of the GSE16560 cohort are summarized in **Table 3**. The investigated collagen-related genes with their functional classifications are listed in **Table 4**. ## Descriptive statistics, statistical hypothesis testing, effect size and statistical significance

Numeric values were referenced as medians with interquartile ranges and ranges. Qualitative variables were presented as percentages and counts of their categories within the complete observation set. Descriptive statistics were computed with function explore() from [*ExDA*](https://github.com/PiotrTymoszuk/ExDA) package.

Differences in numeric variables between two groups were investigated by Mann-Whitney test with r effect size statistic. Differences in numeric variables between three or more groups were assessed by Kruskal-Wallis test with effect size statistic. Differences in distribution of categorical variables were assessed by test with Cramer’s V effect size statistic. Statistical hypothesis testing for differences between analysis groups were investigated with function compare\_variables() from [*ExDA*](https://github.com/PiotrTymoszuk/ExDA) package and the tool set of [*fastTest*](https://github.com/PiotrTymoszuk/fastTest) package.

Effect sizes were interpreted as follows (30):

* r, weak: 0.1 - 0.3, moderate: 0.3 - 0.5, large: 0.5
* : weak: 0.02 - 0.13, moderate: 0.13 - 0.26, large: 0.26
* Cramer’s V, weak: 0.1 - 0.3, moderate: 0.3 - 0.5, large: 0.5

P values were corrected for multiple testing with the false discovery rate method (FDR) (31) within each analysis step and cohort. Effects were considered statistically significant for FDR-corrected p values < 0.05.

## Co-expression networks of the collagen-related genes

Co-expression networks of the collagen-related transcripts were constructed as weighted undirected graphs (9) in the pooled GEO, TCGA, and DKFZ cohorts with as\_iGraph() function from [*graphExtra*](https://github.com/PiotrTymoszuk/graphExtra) package. The network edges were defined by pairwise correlations between ComBat-adjusted -transformed levels of the transcripts with Spearman’s 0.5. Isolated vertices of the networks, i.e. transcripts lacking correlations with other transcripts with 0.5, were removed (function prune\_degree(), package [*graphExtra*](https://github.com/PiotrTymoszuk/graphExtra)). The following vertex importance statistics were computed for the co-expression networks with summary() method from [*graphExtra*](https://github.com/PiotrTymoszuk/graphExtra) package called for the graph objects: (1) degree as a measure of connectivity (number of network neighbors, i.e. correlations of the expression levels with 0.5), (2) hub score as a metric of overall correlation strength (eigenvector of the network’s affinity matrix), (3) betweenness as a measure of connectivity and centrality (minimum/maximum-scaled number of the shortest paths between the vertex pairs that pass through the vertex of interest), and (4) transitivity as a measure of local network density (probability that the neighbor vertices of a vertex are connected with each other). The vertex importance statistics are summarized in **Supplementary Table S5**. The co-expression networks were visualized with the Kamada-Kawai algorithm by calling plot() method for the co-expression graph objects (package [*graphExtra*](https://github.com/PiotrTymoszuk/graphExtra), modified implementation of *ggnetwork*) (32).

## Univariable analysis of BCR-free survival

The univariable analysis of BCR-free survival for the collagen-related genes was done in the pooled GEO, TCGA, and DKFZ cohorts. For each of the collagen-related transcripts, cancer patients were classified as high and low expressors by cutoffs of ComBat-processed -transformed expression levels which corresponded to the largest difference in biochemical relapse-free survival between the expression strata. The minimal size of the expression strata size was set to 25% of the observations; the cutoff search criterion was defined by minimal p value of Mentel-Henszel test. The classification was done with find\_cutoff() function from [*kmOptimizer*](https://github.com/PiotrTymoszuk/kmOptimizer)) package.

BCR risk in the high expressors as compared with the low expressors for each of the collagen-related transcripts was modeled by univariable Cox proportional hazard regression (function coxph() and as\_coxex(), packages *survival* and [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions)) (15). Hazard ratios (HR) for BCR risk in the high as compared with low expressors with 95% confidence intervals served as the inference statistics of the models. The inference p values (), were corrected for multiple testing with the FDR method. Accuracy of the model was assessed by Harrell’s concordance index (33) (high values are characteristic for high concordance between the predicted and observed BCR risk),  
model calibration and confidence of predictions was gauged by Integrated Brier Score (low values are expected for well calibrated and highly confident models) (34). The inference and model fit metrics were computed by calling summary() method from [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions) for the model objects. Genes found to be associated with unfavorable (pFDR < 0.05, HR > 1) or favorable BCR prognosis (pFDR < 0.05, HR < 1) in all three cohorts were considered as standalone markers of BCR prognosis listed in **Supplementary Table S6**.

## Multi-parameter modeling of BCR-free survival

Multi-parameter modeling of BCR-free survival was done in the pooled GEO, TCGA, and DKFZ cohorts with six machine learning algorithms: RIDGE regularized Cox proportional hazard regression (10,35), Elastic Net regularized Cox proportional hazard regression (10,35,36), LASSO regularized Cox proportional hazard regression (10,35,37), Support Vector Machines (SVM) (11,38), Random Forest (12,13,39), and Gradient Boosted Machines (GBM) (14,40–42). Tuning, training, and predictions of the models were handled by in-house developed wrappers around the respective R implementations of the algorithms in packages *glmnet*, *survivalsvm*, *rfsrc*, and *gbm*.

BCR-free survival was modeled with the following explanatory variables:

1. ComBat-adjusted -transformed expression levels of the collagen-related transcripts in the tumor tissue provided as first and second order terms (all algorithms)
2. clinical predictors of BCR risk, pathological tumor stage, ISUP grade, and ComBat-adjusted expression levels of the PSA-coding transcript, *KLK3* as a surrogate of blood PSA concentrations (43,44) (GBM algorithm)
3. ComBat-adjusted -transformed expression levels of the collagen-related transcripts (first and second order), pathological tumor stage, ISUP grade, and ComBat-adjusted expression levels of *KLK3* (GBM algorithm)

Numeric explanatory variables, i.e. first- and second-order ComBat-adjusted transcript levels were normalized by conversion to Z-scores prior to modeling (function zScores(), package [*microViz*](https://github.com/PiotrTymoszuk/microViz)). Because of differences in follow-up times between the training and validation cohorts, the BCR-free survival times were minimum/maximum scaled in each of the cohort (function minMax(), package [*microViz*](https://github.com/PiotrTymoszuk/microViz)).

Selection (tuning) of the optimal set of parameters controlling the model behavior such as regularization, number of learners, or kernel function, was accomplished by minimizing model errors or maximizing concordance indexes for out-of-bag predictions or out-of-fold predictions in cross-validation in the pooled GEO cohort. The tuning criteria and the optimal values of the tunable parameters are summarized in **Supplementary Table S7**. The survival models were trained in the pooled GEO cohort for the optimal combinations of the tunable parameters.

For the pooled GEO training cohort, and the TCGA and DKFZ validation collectives, predictor scores were returned by calling predict() method for the modeling objects. Next, univariable Cox proportional hazard regression models were constructed with the BCR-free survival as the response and each of the predictor scores as the sole explanatory variable (functions coxph() and as\_coxex(), packages *survival* and [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions)) (15). The proportional hazard assumption of the Cox proportional hazard models was checked with the summary(type = 'assumptions') method, which implements the genuine cox.zph() algorithm (45). The Cox proportional hazard models were used for evaluation of machine learning accuracy with Harrell’s concordance index (33) and assessment of global calibration with Integrated Brier Score (34). Those performance statistics were computed by calling summary(type = 'fit') method from [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions) package for the Cox proportional hazard modeling objects (**Supplementary Table S8**). To assess calibration and confidence of BCR risk prediction of the machine learning models, we plotted Brier Scores defined as a mean square difference between the predicted BCR probability and the 0/1-coded observed BCR event (34,46) for unique survival time points. With this graphical method we evaluated predictions made by the best-performing GBM models (**Supplementary Figure S6B**). In addition, ability of the machine learning models to assign PCA patients to low-, intermediate-, and high-risk groups was investigated by comparing BCR-free survival between patients stratified by tertiles of the predictor scores with Kaplan-Meier plots and Peto-Peto test (functions calibrate.coxex(), plot.calibrator(), and surv\_pvalue(), packages [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions) and *survminer*) as presented for the GBM models in **Figure 2B** and **Supplementary Figure 7**.

Importance of the explanatory variables for the predictive performance of the GBM models in the training pooled GEO cohort were measured with the relative influence method (40) and expressed as gradient of the sum of squared errors attributed to particular variables in the learner ensemble (, **Figure 2C**, **Supplementary Figure S6**).

In order to assess the confounding effect of single studies that constituted the pooled GEO data set on results of modeling of BCR-free survival with the best performing GBM algorithm, three nested Cox proportional hazard models were constructed with the pooled GEO data: (1) a model with study as the sole explanatory variable, (2) a model with normalized (Z-score) GBM linear predictor score as the sole explanatory variable, and (3) an additive full mode with study and normalized GBM linear predictor score as the explanatory variables. Hazard ratio estimates for the study and GBM predictor terms with 95% confidence intervals were extracted from the models with summary() method from the [*coxExtensions*](https://github.com/PiotrTymoszuk/coxExtensions) package. Contribution of the data set and GBM predictor terms to the additive model were explored by likelihood ratio test (LRT, method anova(), package *stats*). As presented in **Supplementary Figure S5**, adjustment for the study effect did not change the hazard ratio estimate of the GBM term in comparison of inference statistics of the GBM predictor-only and the additive Cox model. Additionally, the highly significant result of the LRT for the GBM predictor term and the non-significant LRT result for the study term indicate that the GBM predictor but not the study effect contributed substantially to the additive model. Collectively, these results let us conclude that prediction of BCR-free survival in the pooled GEO cohort is independent from differences in BCR-free survival between studies that constitute the pooled collective.

## Multi-parameter modeling of overall survival

Insufficient number of death cases in the TCGA cohort (n = 10) precluded any serious modeling of overall survival. To overcome this limitation, we resorted to the GSE16560 watchful waiting study cohort with long observation times and reliable overall survival data for n = 281 PCA patients with 206 deceased individuals (26). This collective was randomly spit into a training and test subset in a 2:1 ratio, and the training portion was used for tuning and training of machine learning models of overall survival employing the algorithms described for the BCR-free survival modeling. The explanatory factors were first- and second-order terms of ComBat-adjusted -transformed levels of the collagen-related transcript in the PCA tissue. Those expression variables were converted to Z-scores prior to modeling. The tuning and training process was analogical to the modeling of BCR-free survival (for the best combinations of tunable parameters, see **Supplementary Table S9**). The models were evaluated in the training and test subset of the GSE16560 cohort by computing concordance indexes and Integrated Brier Scores, visual assessment of Brier scores for unique time points, and comparison of overall survival between tertiles of predictor scores as described for the models of BCR-free survival. The model performance metrics are summarized in **Supplementary Table S10**.

## Data and code availability

Publicly available data sets were analyzed. Formatted data sets used for analyses will be made available upon request to the corresponding author. The R analysis pipeline for the transcriptome data is available from GitHub (<https://github.com/PiotrTymoszuk/BCR_collagen>).

# Supplementary Tables

Table 1: Characteristic of the analyzed cohorts. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Qualitative variables are presented as percentages of categories within the complete observation set.

| **Variablea** | **pooled GEO** | **TCGA** | **DKFZ** | **GSE16560** |
| --- | --- | --- | --- | --- |
| Cancer samples, N | 501 | 493 | 118 | 281 |
| Age at diagnosis, years | 62 [IQR: 57 - 67] range: 41 - 93 n = 124 | 61 [IQR: 56 - 66] range: 41 - 78 n = 493 | 48 [IQR: 46 - 49] range: 32 - 52 n = 118 | 74 [IQR: 69 - 79] range: 51 - 91 n = 281 |
| PSA at diagnosis | 7.9 [IQR: 5.7 - 13] range: 1 - 280 n = 492 | 7.4 [IQR: 5.1 - 12] range: 1.6 - 87 n = 187 | 8.1 [IQR: 5.9 - 23] range: 1.9 - 740 n = 116 |  |
| Clinical tumor stage | T1: 52% (n = 103) T2: 36% (n = 72) T3: 12% (n = 25) n = 200 |  |  |  |
| pT stage | T1: 2.9% (n = 14) T2: 56% (n = 272) T3: 38% (n = 181) T4: 3.1% (n = 15) n = 482 | T1: 0% (n = 0) T2: 38% (n = 186) T3: 60% (n = 290) T4: 2.1% (n = 10) n = 486 | T1: 0% (n = 0) T2: 64% (n = 74) T3: 30% (n = 35) T4: 6% (n = 7) n = 116 |  |
| pN stage | N0: 89% (n = 246) N1: 11% (n = 30) n = 276 | N0: 81% (n = 342) N1: 19% (n = 78) n = 420 |  |  |
| pM stage | M0: 86% (n = 32) M1: 14% (n = 5) n = 37 |  |  |  |
| ISUP grade | ISUP1: 18% (n = 87) ISUP2: 69% (n = 343) ISUP3+: 13% (n = 65) n = 495 | ISUP1: 9.1% (n = 45) ISUP2: 50% (n = 245) ISUP3+: 41% (n = 203) n = 493 | ISUP1: 11% (n = 13) ISUP2: 74% (n = 87) ISUP3+: 15% (n = 18) n = 118 | ISUP1: 30% (n = 83) ISUP2: 42% (n = 117) ISUP3+: 29% (n = 81) n = 281 |
| Surgical margins | negative: 65% (n = 205) positive: 35% (n = 108) n = 313 |  |  |  |
| Extracapsular extension | 46% (n = 42) n = 91 |  |  |  |
| Death |  | 2% (n = 10) n = 493 |  | 73% (n = 206) n = 281 |
| Overall survival, months |  | 30 [IQR: 17 - 48] range: 0.76 - 170 n = 493 |  | 100 [IQR: 52 - 150] range: 6 - 270 n = 281 |
| Biochemical relapse | 40% (n = 194) n = 487 | 19% (n = 93) n = 493 | 23% (n = 24) n = 105 |  |
| Biochemical relapse-free survival, months | 51 [IQR: 22 - 82] range: 0 - 170 n = 485 | 26 [IQR: 14 - 45] range: 0.76 - 170 n = 493 | 36 [IQR: 13 - 49] range: 0.5 - 76 n = 105 |  |
| aPSA: prostate-specific antigen; pT stage: pathological tumor stage; pN stage: pathological lymph node stage; pM stage: pathological metastasis stage; ISUP grade: grading according to the International Society of Urological Pathology. | | | | |

Table 2: Characteristic of GEO data sets which constitute the pooled GEO cohort. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Qualitative variables are presented as percentages of categories within the complete observation set.

| **Variablea** | **GSE54460** | **GSE70768** | **GSE70769** | **GSE220095** | **Significanceb** | **Effect sizeb** |
| --- | --- | --- | --- | --- | --- | --- |
| Cancer samples, N | 106 | 111 | 92 | 176 |  |  |
| PSA at diagnosis | 7.2 [IQR: 5.5 - 13] range: 1.8 - 73 n = 103 | 7.8 [IQR: 6 - 10] range: 3.2 - 24 n = 110 | 8 [IQR: 5.9 - 11] range: 1.5 - 120 n = 90 | 8.2 [IQR: 5.6 - 14] range: 1 - 120 n = 176 | ns (p = 0.54) | η² = -0.0018 |
| Clinical tumor stage |  | T1: 56% (n = 62) T2: 30% (n = 33) T3: 14% (n = 16) n = 111 | T1: 46% (n = 41) T2: 44% (n = 39) T3: 10% (n = 9) n = 89 |  | ns (p = 0.11) | V = 0.15 |
| pT stage | T1: 13% (n = 14) T2: 70% (n = 73) T3: 16% (n = 17) T4: 0.95% (n = 1) n = 105 | T1: 0% (n = 0) T2: 31% (n = 34) T3: 68% (n = 76) T4: 0.9% (n = 1) n = 111 | T1: 0% (n = 0) T2: 53% (n = 47) T3: 47% (n = 42) T4: 0% (n = 0) n = 89 | T1: 0% (n = 0) T2: 66% (n = 117) T3: 26% (n = 46) T4: 7.4% (n = 13) n = 176 | p < 0.001 | V = 0.31 |
| pN stage |  | N0: 91% (n = 82) N1: 8.9% (n = 8) n = 90 | N0: 100% (n = 18) N1: 0% (n = 0) n = 18 | N0: 87% (n = 146) N1: 13% (n = 22) n = 168 | ns (p = 0.18) | V = 0.11 |
| ISUP grade | ISUP1: 10% (n = 11) ISUP2: 75% (n = 80) ISUP3+: 14% (n = 15) n = 106 | ISUP1: 15% (n = 17) ISUP2: 77% (n = 85) ISUP3+: 8.1% (n = 9) n = 111 | ISUP1: 23% (n = 21) ISUP2: 61% (n = 55) ISUP3+: 16% (n = 14) n = 90 | ISUP1: 20% (n = 36) ISUP2: 68% (n = 120) ISUP3+: 11% (n = 20) n = 176 | ns (p = 0.088) | V = 0.11 |
| Biochemical relapse | 52% (n = 55) n = 106 | 17% (n = 19) n = 111 | 49% (n = 45) n = 92 | 43% (n = 75) n = 176 | p < 0.001 | V = 0.26 |
| Biochemical relapse-free survival, months | 49 [IQR: 18 - 77] range: 0 - 170 n = 106 | 30 [IQR: 17 - 49] range: 1 - 65 n = 111 | 58 [IQR: 19 - 80] range: 0.36 - 100 n = 92 | 75 [IQR: 46 - 110] range: 0.66 - 130 n = 176 | p < 0.001 |  |
| aPSA: prostate-specific antigen; pT stage: pathological tumor stage; pN stage: pathological lymph node stage; pM stage: pathological metastasis stage; ISUP grade: grading according to the International Society of Urological Pathology. | | | | | | |
| bCategorical variables: χ² test with Cramer's V effect size statistic. Numeric variables: Kruskal-Wallis test with η² effect size statistic. Biochemical relapse-free survival: Peto-Peto test. P values corrected for multiple testing with the false discovery rate method. | | | | | | |

Table 3: Characteristic of the training and test subsets of the GSE16560 cohort. For modeling of overall survival, the GSE16560 cohort was randomly split into a training and a test subset in a 2:1 ratio. Numeric characteristics of the subsets are presented as medians with interquartile ranges (IQR) and ranges. Qualitative variables are presented as percentages of categories within the complete observation set.

| **Variablea** | **Training** | **Test** | **Significanceb** | **Effect sizeb** |
| --- | --- | --- | --- | --- |
| Cancer samples, N | 188 | 93 |  |  |
| Age at diagnosis, years | 74 [IQR: 69 - 79] range: 51 - 91 n = 188 | 74 [IQR: 70 - 79] range: 57 - 91 n = 93 | ns (p = 0.96) | η² = -0.0029 |
| ISUP grade | ISUP1: 28% (n = 53) ISUP2: 42% (n = 79) ISUP3+: 30% (n = 56) n = 188 | ISUP1: 32% (n = 30) ISUP2: 41% (n = 38) ISUP3+: 27% (n = 25) n = 93 | ns (p = 0.96) | r = 0.043 |
| Death | 73% (n = 138) n = 188 | 73% (n = 68) n = 93 | ns (p = 0.96) | r = 0.003 |
| Overall survival, months | 100 [IQR: 48 - 150] range: 6 - 250 n = 188 | 100 [IQR: 54 - 140] range: 7 - 270 n = 93 | ns (p = 0.73) |  |
| aISUP grade: grading according to the International Society of Urological Pathology. | | | | |
| bCategorical variables: χ² test with Cramer's V effect size statistic. Numeric variables: Kruskal-Wallis test with η² effect size statistic. Overall survival: Peto-Peto test. P values corrected for multiple testing with the false discovery rate method. | | | | |

Table 4: Collagen-related genes and their classification.

| **Functional classification** | **Gene symbol** | **Entrez ID** |
| --- | --- | --- |
| proline metabolism | *ALDH18A1* | 5832 |
| *PEPD* | 5184 |
| *PYCR1* | 5831 |
| collagen modification | *LOX* | 4015 |
| *LOXL1* | 4016 |
| *LOXL2* | 4017 |
| *P4HA1* | 5033 |
| *P4HA2* | 8974 |
| *P4HB* | 5034 |
| *PLOD1* | 5351 |
| *PLOD2* | 5352 |
| *PLOD3* | 8985 |
| *PPIB* | 5479 |
| ECM component | *COL11A1* | 1301 |
| *COL11A2* | 1302 |
| *COL14A1* | 7373 |
| *COL15A1* | 1306 |
| *COL16A1* | 1307 |
| *COL17A1* | 1308 |
| *COL18A1* | 80781 |
| *COL19A1* | 1310 |
| *COL1A1* | 1277 |
| *COL1A2* | 1278 |
| *COL2A1* | 1280 |
| *COL3A1* | 1281 |
| *COL4A1* | 1282 |
| *COL4A2* | 1284 |
| *COL4A3* | 1285 |
| *COL4A5* | 1287 |
| *COL4A6* | 1288 |
| *COL5A1* | 1289 |
| *COL5A2* | 1290 |
| *COL6A1* | 1291 |
| *COL6A2* | 1292 |
| *COL6A3* | 1293 |
| *COL7A1* | 1294 |
| *COL9A1* | 1297 |
| *COL9A2* | 1298 |
| *COL9A3* | 1299 |
| *LAMA3* | 3909 |
| *LAMB3* | 3914 |
| *LAMC2* | 3918 |
| ECM processing | *ADAMTS2* | 9509 |
| *BMP1* | 649 |
| *CTSS* | 1520 |
| *MMP13* | 4322 |
| *MMP7* | 4316 |
| *MMP9* | 4318 |
| *PCOLCE* | 5118 |
| *PCOLCE2* | 26577 |
| *SERPINH1* | 871 |
| adhesion | *CD151* | 977 |
| *DST* | 667 |
| *ITGA6* | 3655 |
| *ITGB4* | 3691 |

Table 5: Vertex importance statistics for co-expression networks of the collagen-related transcripts. The co-expression networks were built for in the pooled GEO, TCGA, and DKFZ cohorts for transcripts with pairwise correlation of expression levels with Spearman's rho >= 0.5. Metrics of importance of the vertices of the co-expression networks, degree, hub score, betweenness, and transitivity, were computed. Top five vertices with the largest hub score per cohort are presented. The full table is available as a supplementary Excel file.

| **Cohort** | **Gene symbol** | **Functional classification** | **Degreea** | **Hub scoreb** | **Betweennessc** | **Transitivityd** |
| --- | --- | --- | --- | --- | --- | --- |
| DKFZ | *COL5A1* | ECM component | 19 | 1.00 | 0.0550 | 0.72 |
| *COL1A1* | ECM component | 19 | 0.95 | 0.0180 | 0.73 |
| *COL4A2* | ECM component | 18 | 0.92 | 0.0510 | 0.74 |
| *LOXL1* | collagen modification | 19 | 0.92 | 0.0320 | 0.71 |
| *ADAMTS2* | ECM processing | 17 | 0.85 | 0.0046 | 0.76 |
| *COL6A1* | ECM component | 19 | 0.85 | 0.4700 | 0.66 |
| TCGA | *COL5A1* | ECM component | 21 | 1.00 | 0.0000 | 0.86 |
| *COL1A2* | ECM component | 21 | 0.99 | 0.0120 | 0.84 |
| *COL4A2* | ECM component | 21 | 0.97 | 0.0000 | 0.86 |
| *COL16A1* | ECM component | 24 | 0.94 | 0.5900 | 0.69 |
| *COL15A1* | ECM component | 22 | 0.93 | 0.0730 | 0.80 |
| *COL6A3* | ECM component | 23 | 0.93 | 0.3200 | 0.69 |
| pooled GEO | *COL5A1* | ECM component | 19 | 1.00 | 0.1300 | 0.74 |
| *COL4A2* | ECM component | 19 | 0.98 | 0.0970 | 0.74 |
| *COL16A1* | ECM component | 19 | 0.93 | 0.1900 | 0.74 |
| *COL6A3* | ECM component | 19 | 0.92 | 0.2700 | 0.72 |
| *COL4A1* | ECM component | 17 | 0.91 | 0.0160 | 0.83 |
| aDegree: number of network neighbors, i.e. correlations with Spearman's ρ >= 0.5. | | | | | | |
| bHub score: eigenvector of the affinity matrix; a metric of the overall correlation strength. | | | | | | |
| cBetweenness: minimum/maximum scaled number of the shortest paths between pairs of network vertices which pass through the vertex of interest; measure of connectivity and centrality. | | | | | | |
| dTransitivity: probability that the adjacent vertices of a vertex are connected; measure of the local densitiy of the co-expression network. | | | | | | |

Table 6: Univariable analysis of biochemical relapse (BCR) free survival for the collagen-realted transcripts in the pooled GEO, TCGA, and DKFZ cohort. PCA patients were classified as high and low expressors of the collagen-related transcripts by expression cutoffs corresponding to the largest differences in BCR-free survival between the expression strata. BCR risk of high as compared with low expressors was modeled by univariable Cox proportional hazard regression. P values were corrected for multiple testing with the false discovery rate method. The modeling results are shown for the collagen-related genes associated with significant differences in BCR-free survival in all three investigated cohorts. The full analysis is available as a supplementary Excel file.

| **Association with BCR risk** | **Gene symbol** | **Cohort** | **log2 expression cutoff** | **Low expressors, N** | **High expressors, N** | **HR, 95%CIa** | **Significance** | **C-indexb** | **IBSc** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| unfavorable | *COL11A1* | pooled GEO | 4.3 | total: n = 151 BCR: n = 38 | total: n = 334 BCR: n = 38 | 2.3 [3.2 to 1.6] | p < 0.001 | 0.59 | 0.22 |
| TCGA | 4.2 | total: n = 218 BCR: n = 27 | total: n = 275 BCR: n = 27 | 2.1 [3.3 to 1.4] | p = 0.01 | 0.58 | 0.17 |
| DKFZ | 4.5 | total: n = 72 BCR: n = 9 | total: n = 33 BCR: n = 9 | 5.3 [12 to 2.3] | p < 0.001 | 0.71 | 0.12 |
| *COL11A2* | pooled GEO | 4.6 | total: n = 283 BCR: n = 107 | total: n = 202 BCR: n = 107 | 1.4 [1.9 to 1] | p = 0.046 | 0.55 | 0.22 |
| TCGA | 4.7 | total: n = 217 BCR: n = 17 | total: n = 276 BCR: n = 17 | 4.7 [8 to 2.8] | p < 0.001 | 0.67 | 0.18 |
| DKFZ | 4.9 | total: n = 74 BCR: n = 13 | total: n = 31 BCR: n = 13 | 2.5 [5.6 to 1.1] | p = 0.047 | 0.62 | 0.16 |
| *COL1A1* | pooled GEO | 13.0 | total: n = 347 BCR: n = 112 | total: n = 138 BCR: n = 112 | 2.3 [3.1 to 1.7] | p < 0.001 | 0.59 | 0.20 |
| TCGA | 13.0 | total: n = 321 BCR: n = 43 | total: n = 172 BCR: n = 43 | 2.4 [3.5 to 1.6] | p = 0.001 | 0.60 | 0.17 |
| DKFZ | 14.0 | total: n = 79 BCR: n = 11 | total: n = 26 BCR: n = 11 | 5 [11 to 2.2] | p < 0.001 | 0.69 | 0.15 |
| *COL3A1* | pooled GEO | 11.0 | total: n = 259 BCR: n = 81 | total: n = 226 BCR: n = 81 | 2 [2.7 to 1.5] | p < 0.001 | 0.59 | 0.20 |
| TCGA | 13.0 | total: n = 359 BCR: n = 58 | total: n = 134 BCR: n = 58 | 1.6 [2.5 to 1.1] | p = 0.049 | 0.55 | 0.19 |
| DKFZ | 13.0 | total: n = 79 BCR: n = 13 | total: n = 26 BCR: n = 13 | 3.5 [7.8 to 1.6] | p = 0.0091 | 0.64 | 0.14 |
| *COL5A2* | pooled GEO | 8.4 | total: n = 283 BCR: n = 96 | total: n = 202 BCR: n = 96 | 1.6 [2.2 to 1.2] | p = 0.0036 | 0.57 | 0.21 |
| TCGA | 8.4 | total: n = 267 BCR: n = 38 | total: n = 226 BCR: n = 38 | 1.7 [2.6 to 1.1] | p = 0.029 | 0.56 | 0.18 |
| DKFZ | 8.6 | total: n = 73 BCR: n = 12 | total: n = 32 BCR: n = 12 | 3.1 [6.9 to 1.4] | p = 0.018 | 0.64 | 0.15 |
| *LOXL1* | pooled GEO | 7.3 | total: n = 228 BCR: n = 76 | total: n = 257 BCR: n = 76 | 1.5 [2.1 to 1.2] | p = 0.0098 | 0.56 | 0.21 |
| TCGA | 7.3 | total: n = 216 BCR: n = 27 | total: n = 277 BCR: n = 27 | 2 [3.1 to 1.3] | p = 0.017 | 0.58 | 0.18 |
| DKFZ | 7.6 | total: n = 74 BCR: n = 12 | total: n = 31 BCR: n = 12 | 3 [6.6 to 1.3] | p = 0.02 | 0.63 | 0.15 |
| *SERPINH1* | pooled GEO | 9.3 | total: n = 278 BCR: n = 95 | total: n = 207 BCR: n = 95 | 1.6 [2.1 to 1.2] | p = 0.0076 | 0.55 | 0.21 |
| TCGA | 9.5 | total: n = 337 BCR: n = 50 | total: n = 156 BCR: n = 50 | 1.9 [2.9 to 1.3] | p = 0.017 | 0.56 | 0.18 |
| DKFZ | 9.5 | total: n = 79 BCR: n = 10 | total: n = 26 BCR: n = 10 | 6.6 [15 to 2.9] | p < 0.001 | 0.72 | 0.14 |
| favorable | *COL4A6* | pooled GEO | 6.8 | total: n = 310 BCR: n = 145 | total: n = 175 BCR: n = 145 | 0.49 [0.67 to 0.35] | p < 0.001 | 0.58 | 0.21 |
| TCGA | 6.2 | total: n = 170 BCR: n = 47 | total: n = 323 BCR: n = 47 | 0.46 [0.69 to 0.31] | p = 0.0034 | 0.60 | 0.22 |
| DKFZ | 5.8 | total: n = 29 BCR: n = 15 | total: n = 76 BCR: n = 15 | 0.15 [0.35 to 0.063] | p < 0.001 | 0.72 | 0.11 |
| *DST* | pooled GEO | 9.5 | total: n = 362 BCR: n = 156 | total: n = 123 BCR: n = 156 | 0.58 [0.82 to 0.4] | p = 0.0087 | 0.55 | 0.21 |
| TCGA | 9.1 | total: n = 233 BCR: n = 54 | total: n = 260 BCR: n = 54 | 0.58 [0.88 to 0.38] | p = 0.029 | 0.58 | 0.21 |
| DKFZ | 8.1 | total: n = 29 BCR: n = 13 | total: n = 76 BCR: n = 13 | 0.27 [0.6 to 0.12] | p = 0.0054 | 0.66 | 0.14 |
| *LAMB3* | pooled GEO | 7.8 | total: n = 355 BCR: n = 152 | total: n = 130 BCR: n = 152 | 0.68 [0.95 to 0.48] | p = 0.046 | 0.54 | 0.22 |
| TCGA | 7.8 | total: n = 308 BCR: n = 72 | total: n = 185 BCR: n = 72 | 0.48 [0.79 to 0.3] | p = 0.017 | 0.59 | 0.20 |
| DKFZ | 6.6 | total: n = 33 BCR: n = 15 | total: n = 72 BCR: n = 15 | 0.21 [0.47 to 0.089] | p = 0.0012 | 0.69 | 0.13 |
| *PCOLCE2* | pooled GEO | 7.6 | total: n = 353 BCR: n = 158 | total: n = 132 BCR: n = 158 | 0.6 [0.86 to 0.42] | p = 0.015 | 0.55 | 0.22 |
| TCGA | 7.5 | total: n = 338 BCR: n = 74 | total: n = 155 BCR: n = 74 | 0.5 [0.82 to 0.3] | p = 0.029 | 0.58 | 0.21 |
| DKFZ | 7.0 | total: n = 56 BCR: n = 18 | total: n = 49 BCR: n = 18 | 0.32 [0.81 to 0.13] | p = 0.033 | 0.63 | 0.15 |
| aHarazd ratio (HR) with 95% confidence intervals (CI). | | | | | | | | | |
| bC-index: Harrel's concordance index, a measure of model accuracy. High values are expected for a highly accurate model; concordance index = 0.5 is expected for random predictions. | | | | | | | | | |
| cIBS: Integrated Brier Score, a metric of model calibration. Low values are expected for a well calibrate model. Intergrated Brier score = 0.25 is expected for random predictions. | | | | | | | | | |

Table 7: Selection of the optimal parameters of machine learning models of biochemical relapse (BCR) free survival in the pooled GEO training cohort.

| **BCR modela** | **Selection criterion** | **Tuned parameters** |
| --- | --- | --- |
| RIDGE Coxb | minimal deviance, repeated 10-fold cross-validation | λ = 0.41 |
| Elastic Net Coxb | minimal deviance, repeated 10-fold cross-validation | λ = 0.031 |
| LASSO Coxb | minimal deviance, repeated 10-fold cross-validation | λ = 0.014 |
| SVMb | maximal concordance index, repeated 10-fold cross-validation | SVM model type = vanbelle1 kernel = add\_kernel γ = 0.005 |
| Random Forestb | maximal concordance index, out-of-bag predictions | number of variables per tree, mtry = 4 splitting rule = logrank minimal node size = 5 number of splits = 5 |
| GBM, expressionb | minimal deviance, 10-fold cross-validation | number of decision trees = 500 shrinkage = 0.03 interaction depth = 3 minimal node size = 5 |
| GBM, clinicc | minimal deviance, 10-fold cross-validation | number of decision trees = 1000 shrinkage = 0.01 interaction depth = 3 minimal node size = 5 |
| GBM, clinic + expressiond | minimal deviance, 10-fold cross-validation | number of decision trees = 500 shrinkage = 0.05 interaction depth = 4 minimal node size = 5 |
| aSVM: Support Vector Machines; GBM: Gradient Boosted Machines. | | |
| bExplanatory factors: expression levels of the collagen-related transcripts. | | |
| cExplanatory factors: pathological tumor stage, ISUP grade, and log2 expression of the PSA coding gene KLK3 in the tumor tissue. | | |
| dExplanatory factors: pathological tumor stage, ISUP grade, and log2 expression of the PSA coding gene KLK3 in the tumor tissue, log2 expression of the collagen-related transcripts. | | |

Table 8: Performance of machine learning models at prediction of biochemical relapse (BCR) free survival.

| **BCR modela** | **Cohort type** | **Cohort** | **C-indexb** | **IBSc** |
| --- | --- | --- | --- | --- |
| RIDGE Coxb | training | pooled GEO | 0.74 | 0.170 |
| test | TCGA | 0.66 | 0.190 |
| DKFZ | 0.79 | 0.110 |
| Elastic Net Coxb | training | pooled GEO | 0.75 | 0.170 |
| test | TCGA | 0.64 | 0.190 |
| DKFZ | 0.78 | 0.120 |
| LASSO Coxb | training | pooled GEO | 0.75 | 0.170 |
| test | TCGA | 0.64 | 0.190 |
| DKFZ | 0.77 | 0.120 |
| SVMb | training | pooled GEO | 0.65 | 0.180 |
| test | TCGA | 0.61 | 0.200 |
| DKFZ | 0.75 | 0.120 |
| Random Forestb | training | pooled GEO | 0.94 | 0.058 |
| test | TCGA | 0.66 | 0.240 |
| DKFZ | 0.82 | 0.380 |
| GBM, expressionb | training | pooled GEO | 0.89 | 0.089 |
| test | TCGA | 0.66 | 0.180 |
| DKFZ | 0.81 | 0.110 |
| GBM, clinicc | training | pooled GEO | 0.75 | 0.190 |
| test | TCGA | 0.70 | 0.140 |
| DKFZ | 0.71 | 0.120 |
| GBM, clinic + expressiond | training | pooled GEO | 0.88 | 0.110 |
| test | TCGA | 0.67 | 0.160 |
| DKFZ | 0.79 | 0.110 |
| aSVM: Support Vector Machines; GBM: Gradient Boosted Machines. | | | | |
| bExplanatory factors: expression levels of the collagen-related transcripts. | | | | |
| cExplanatory factors: pathological tumor stage, ISUP grade, and log2 expression of the PSA coding gene KLK3 in the tumor tissue. | | | | |
| dExplanatory factors: pathological tumor stage, ISUP grade, and log2 expression of the PSA coding gene KLK3 in the tumor tissue, log2 expression of the collagen-related transcripts. | | | | |
| bC-index: Harrel's concordance index, a measure of model accuracy. High values are expected for a highly accurate model; concordance index = 0.5 is expected for random predictions. | | | | |
| cIBS: Integrated Brier Score, a metric of model calibration. Low values are expected for a well calibrated model. Intergrated Brier score = 0.25 is expected for random predictions. | | | | |

Table 9: Selection of the optimal parameters of machine learning models of overall survival (OS) in the training subset of the GSE16560 cohort. The models used log2 expression of the collagen-related genes as explanatory variables.

| **OS modela** | **Selection criterion** | **Tuned parameters** |
| --- | --- | --- |
| RIDGE Cox | minimal deviance, repeated 10-fold cross-validation | λ = 2.6 |
| Elastic Net Cox | minimal deviance, repeated 10-fold cross-validation | λ = 0.12 |
| LASSO Cox | minimal deviance, repeated 10-fold cross-validation | λ = 0.057 |
| SVM | maximal concordance index, repeated 10-fold cross-validation | SVM model type = vanbelle1 kernel = add\_kernel γ = 1 |
| Random Forest | maximal concordance index, out-of-bag predictions | number of variables per tree, mtry = 14 splitting rule = logrank minimal node size = 5 number of splits = 1 |
| GBM | minimal deviance, 10-fold cross-validation | number of decision trees = 500 shrinkage = 0.04 interaction depth = 2 minimal node size = 5 |
| aSVM: Support Vector Machines; GBM: Gradient Boosted Machines. | | |

Table 10: Performance of machine learning models at prediction of overall survival (OS) in the training and test subsets of the GSE16560 cohort. The models used log2 expression of the collagen-related genes as explanatory factors.

| **OS model** | **Cohort subset** | **C-indexb** | **IBSc** |
| --- | --- | --- | --- |
| RIDGE Cox | training | 0.70 | 0.130 |
| test | 0.59 | 0.180 |
| Elastic Net Cox | training | 0.69 | 0.120 |
| test | 0.58 | 0.170 |
| LASSO Cox | training | 0.69 | 0.120 |
| test | 0.58 | 0.170 |
| SVM | training | 0.64 | 0.110 |
| test | 0.53 | 0.180 |
| Random Forest | training | 0.89 | 0.039 |
| test | 0.50 | 0.140 |
| GBM | training | 0.73 | 0.110 |
| test | 0.54 | 0.180 |
| bC-index: Harrell's concordance index, a measure of model accuracy. High values are expected for a highly accurate model; concordance index = 0.5 is expected for random predictions. | | | |
| cIBS: Integrated Brier Score, a metric of model calibration. Low values are expected for a well calibrate model. Intergrated Brier score = 0.25 is expected for random predictions. | | | |

# Supplementary Figures

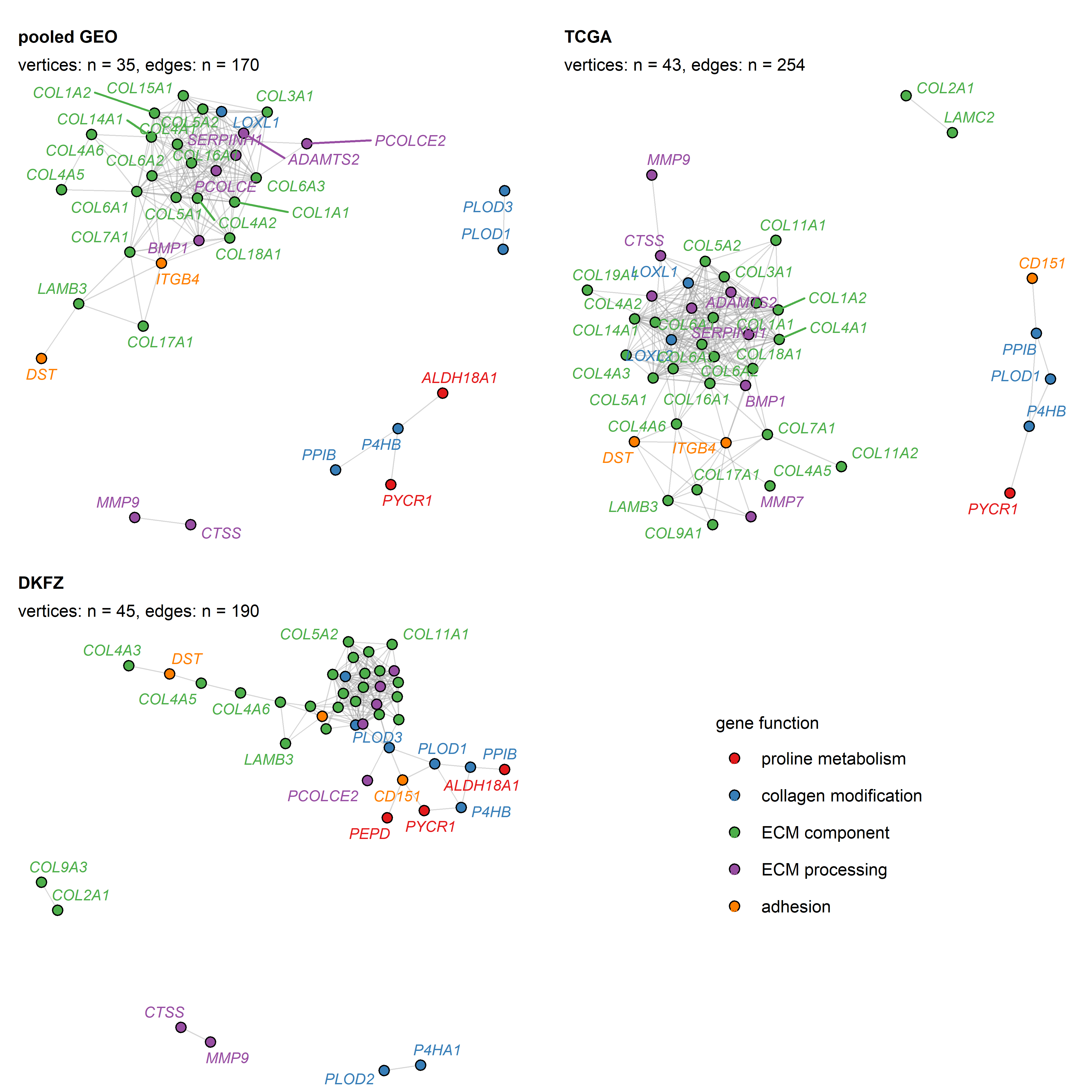


Figure 1: Co-expression networks of the collagen-related transcripts in PCA tissue.

**Supplementary Figure S1. Co-expression networks of the collagen-related transcripts in PCA tissue.**

*Co-expression networks of the collagen-related transcripts in PCA were constructed in the pooled GEO (cancer samples: n = 501), TCGA (n = 493), and DKFZ cohort (n = 118). The network edges were defined by pairwise correlations between ComBat-adjusted transcript levels in the cancer tissue with Spearman’s 0.5. Isolated vertices, i.e. transcripts without correlation partners, were removed.* *The co-expression networks were visualized with Kamada-Kawai algorithm. Each point represents a single transcript in the network; point color codes for functional classification of the transcript. Edges, i.e. correlations of expression levels with 0.5, are depicted as gray lines. Numbers of transcripts (vertices) and correlations with 0.5 (edges) are displayed in the plot captions.*

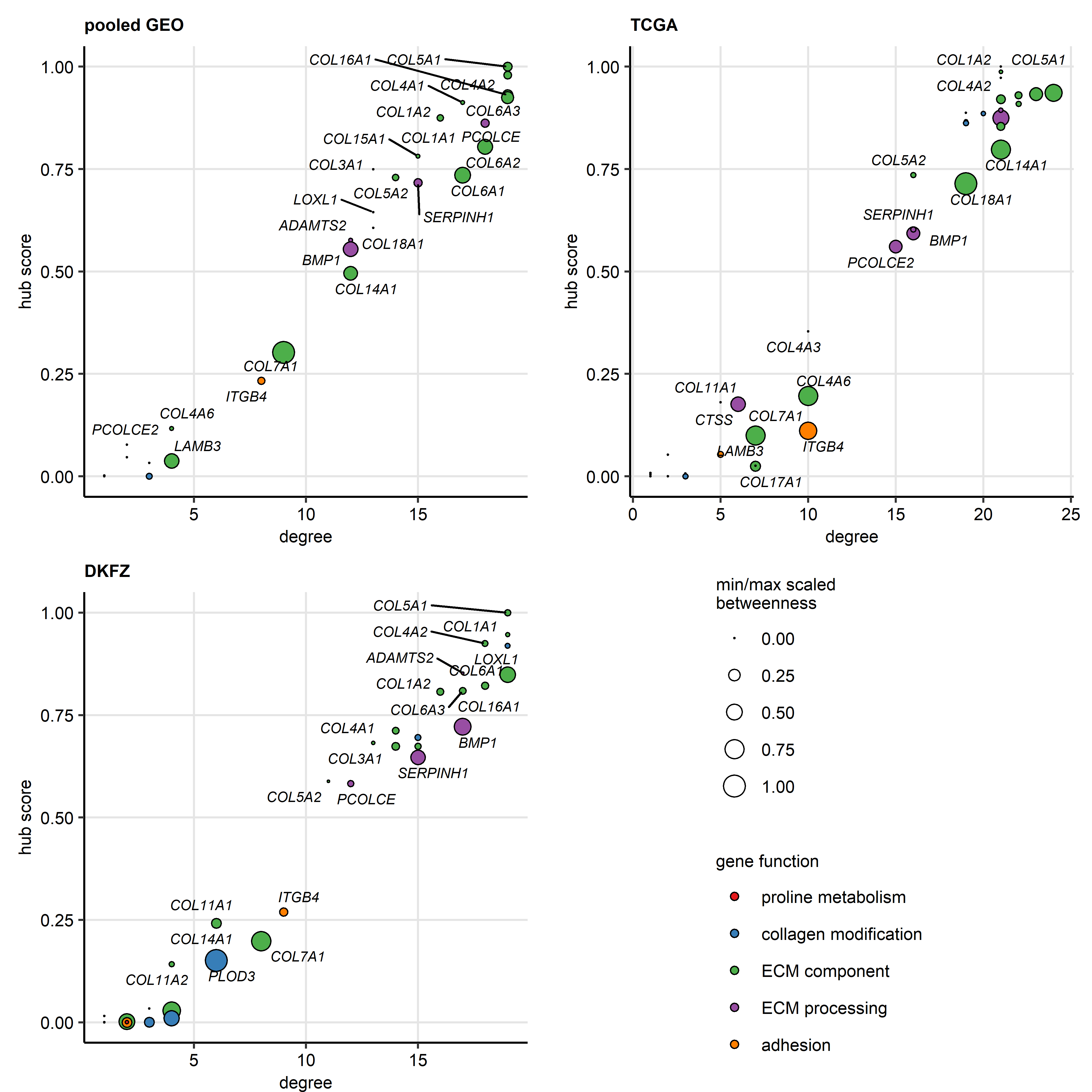


Figure 2: Vertex importance statistics for co-expression networks of the collagen-related transcripts in PCA.

**Supplementary Figure S2. Vertex importance statistics for co-expression networks of the collagen-related transcripts in PCA.**

*Co-expression networks of the collagen-related transcripts were constructed in the pooled GEO (cancer samples: n = 501), TCGA (n = 493), and DKFZ cohort (n = 118) as presented in Supplementary Figure S1.* *The following numeric statistics of vertex importance were computed for each collagen-related transcript: degree as a measure of connectivity (number of correlation partners with Spearman’s 0.5), hub score as a measure of overall correlation strength (eigenvector of the affinity matrix), and betweenness as a measure of connectivity and centrality (minimum/maximum-scaled number of the shortest paths between all pairs of vertices that pass through the vertex of interest).* *The vertex importance metrics for the collagen-related transcripts in the co-expression networks are shown in dot plots. Each point represents a single collagen-related transcript, point color codes for functional classification of the transcript.*

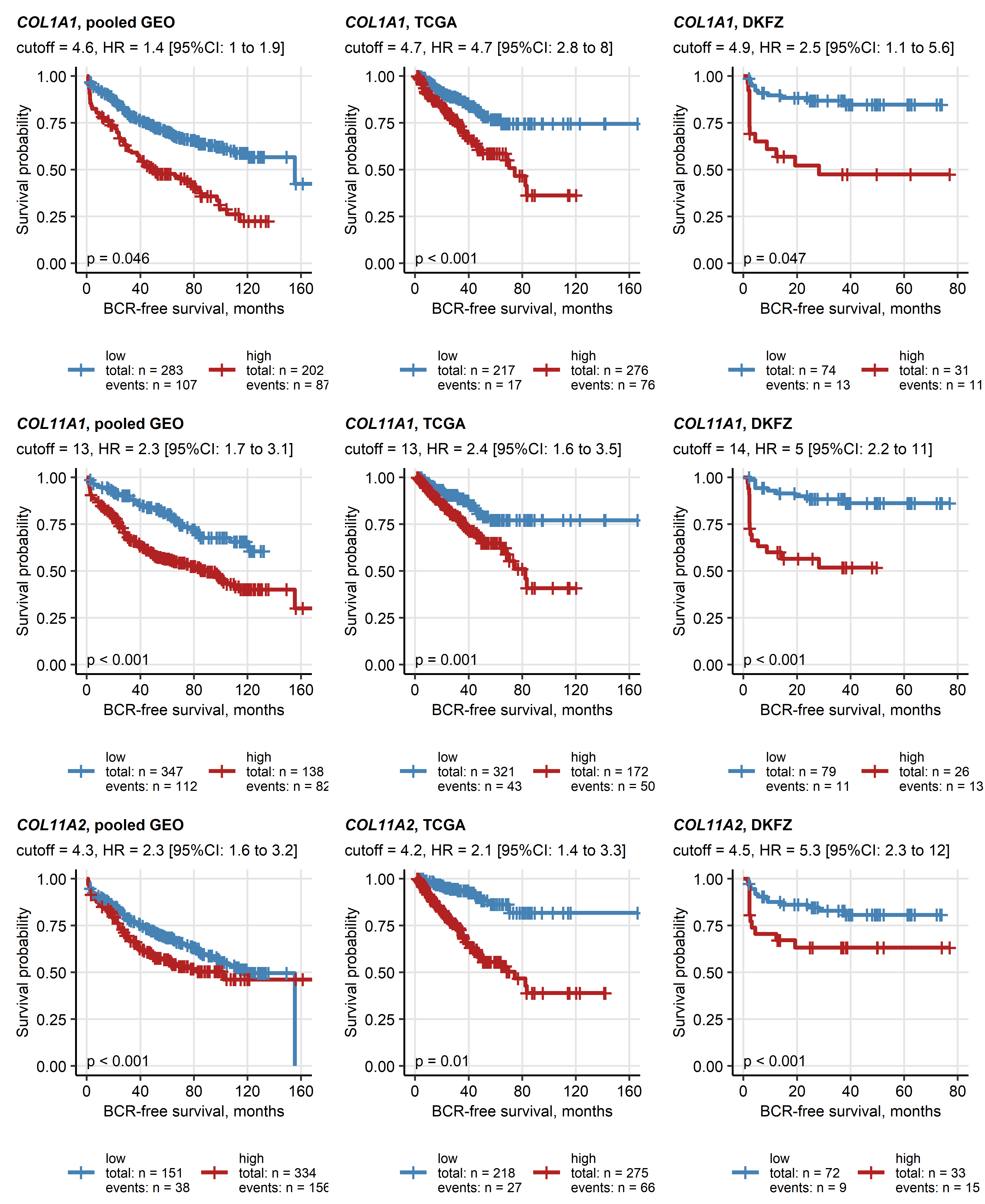


Figure 3: Top three unfavorable standalone collagen-related transcriptional markers of BCR-free survival in PCA.

**Supplementary Figure S3. Top three unfavorable standalone collagen-related transcriptional markers of BCR-free survival in PCA.**

*Prostate cancer (PCA) patients in the pooled GEO (total: n = 485, biochemical relapse [BCR]: n = 194), TCGA (total: n = 493, BCR: n = 93), and DKFZ cohort (total: n = 105, BCR: n = 24) were classified as high and low expressors for each of the 55 collagen-related transcripts with the expression cutoffs corresponding to the largest differences in survival between the high and low expressors.* *Collagen-related transcripts associated with unfavorable and favorable BCR prognosis were identified by univariable Cox proportional hazard regression as presented in Figure 1.* *COL1A1, COL11A1, and COL11A2 were identified as the strongest unfavorable BCR risk markers, as metered by mean Harrell’s concordance index (measure of model accuracy) in the investigated cohorts. Fractions of BCR-free patients in the high and low expression strata are visualized in Kaplan-Meier plots. The log2 expression cutoffs used for the high/low expressor classification, hazard-ratios (HR) of BCR risk in the high as compared with the low expression strata with 95% confidence intervals (95% CI) are displayed in the plot captions. P values corrected for multiple testing with the false discovery rate method are shown in the plots. Numbers of patients and BCR cases in the expression strata are indicated in the plot legends.*

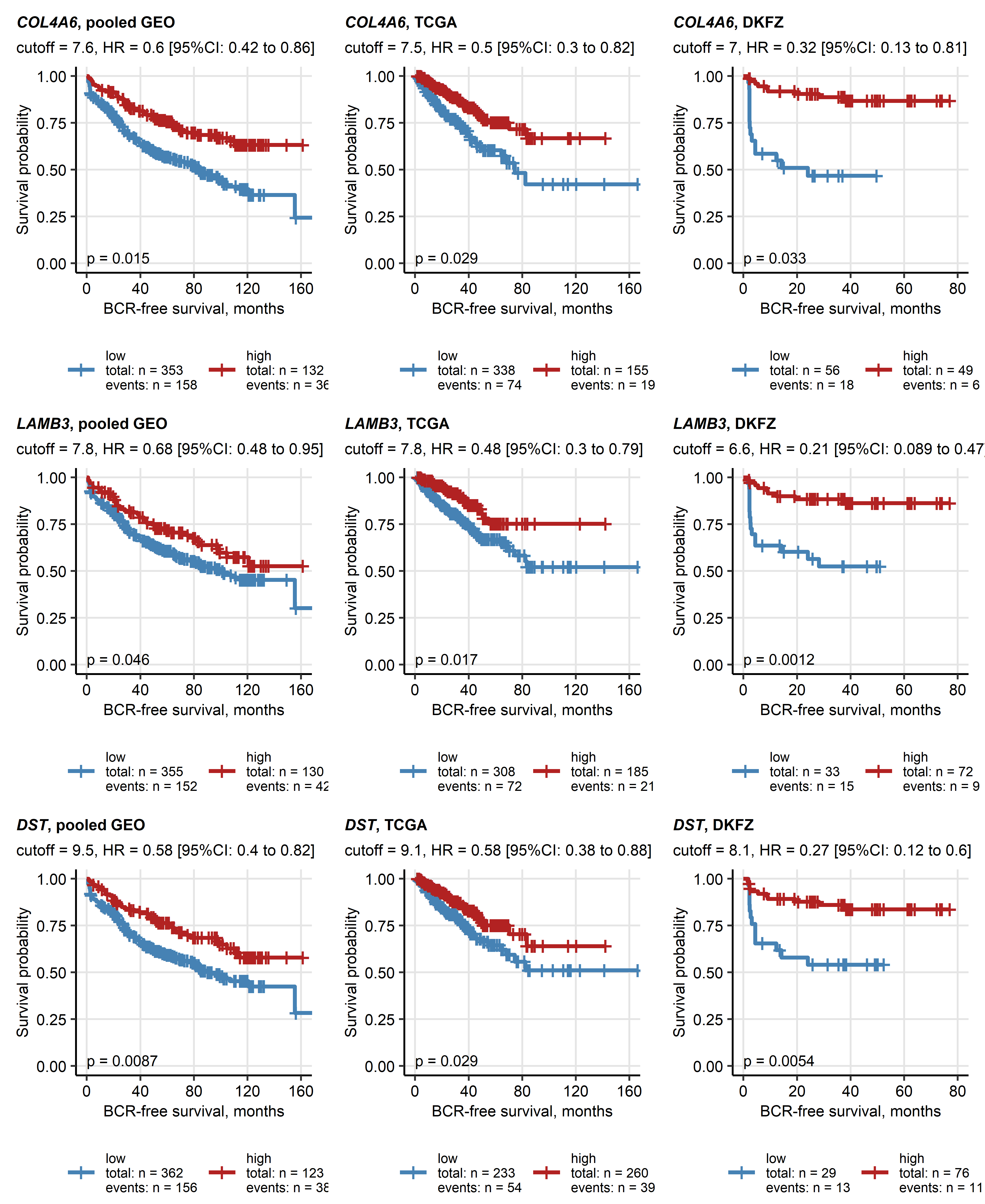


Figure 4: Top three favorable standalone collagen-related transcriptional markers of BCR-free survival in PCA.

**Supplementary Figure S4. Top three favorable standalone collagen-related transcriptional markers of BCR-free survival in PCA.**

*Prostate cancer (PCA) patients in the pooled GEO (total: n = 485, biochemical relapse [BCR]: n = 194), TCGA (total: n = 493, BCR: n = 93), and DKFZ cohort (total: n = 105, BCR: n = 24) were classified as high and low expressors for each of the 55 collagen-related transcripts with the expression cutoffs corresponding to the largest differences in survival between the high and low expressors.* *Collagen-related transcripts associated with unfavorable and favorable BCR prognosis were identified by univariable Cox proportional hazard regression as presented in Figure 1.* *COL4A6, LAMB3, and DST were identified as the strongest favorable BCR risk markers, as measured by mean Harrell’s concordance index (measure of model accuracy) in the investigated cohorts. Fractions of BCR-free patients in the high and low expression strata are visualized in Kaplan-Meier plots. The log2 expression cutoffs used for the high/low expressor classification, hazard-ratios (HR) of BCR risk in the high as compared with the low expression strata with 95% confidence intervals (95% CI) are displayed in the plot captions. P values corrected for multiple testing with the false discovery rate method are shown in the plots. Numbers of patients and BCR cases in the expression strata are indicated in the plot legends.*

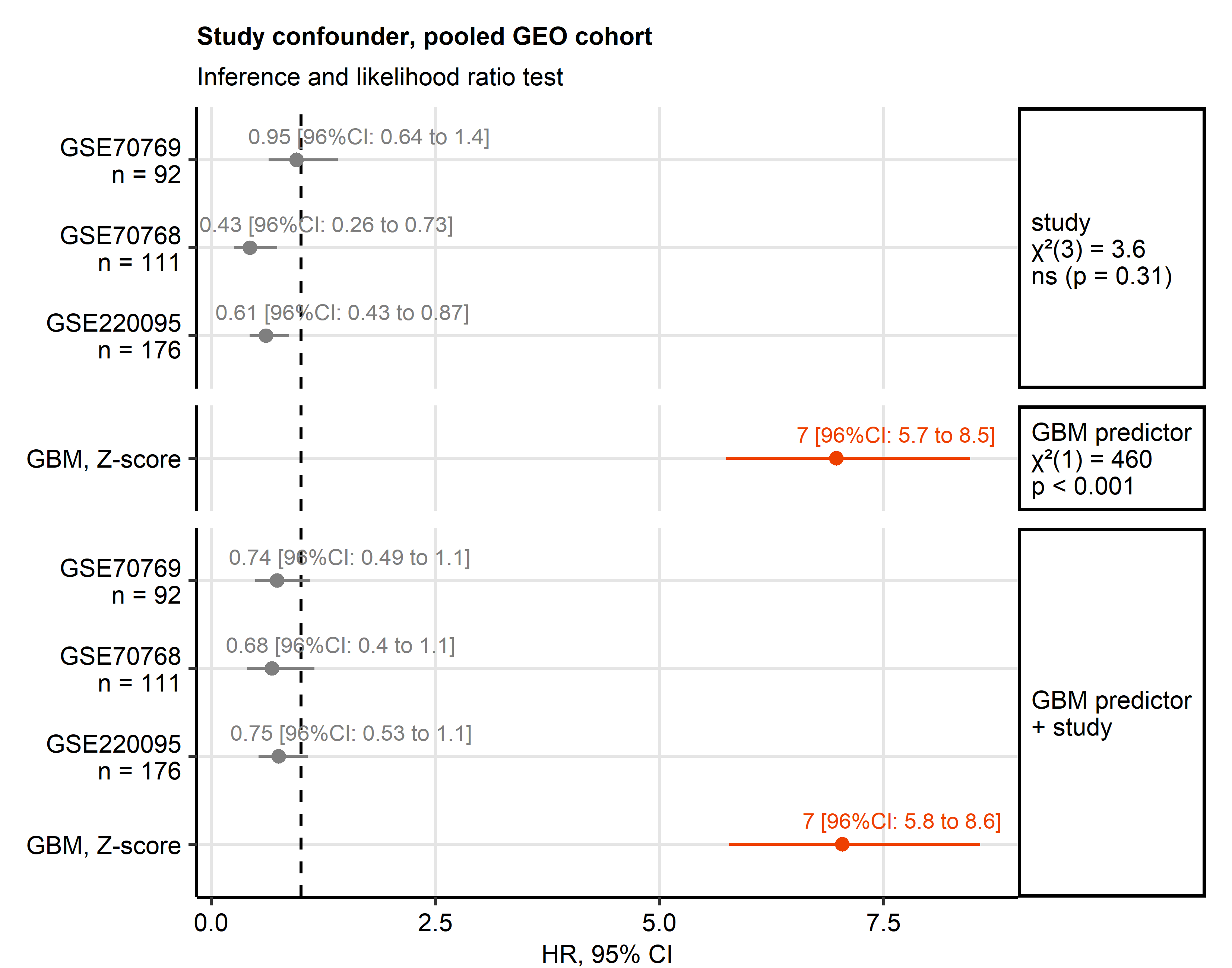


Figure 5: Investigation of the confounding study effect of the cohort on prediction of BCR-free survival by the GBM model in the pooled GEO cohort.

**Figure 5. Investigation of confounding study effect on prediction of BCR-free survival by the GBM model in the pooled GEO cohort.**

*Biochemical relapse (BCR) free survival was modeled by the GBM algorithm with ComBat-adjusted -transformed expression levels of the collagen-related genes as explanatory factors. The Gradient Boosted Machines (GBM) model was trained in the pooled GEO cohort (total: n = 485, BCR: n = 194).* *To investigate if and to which extend the BCR risk prediction depends on the confounding effect of single studies constituting the pooled GEO cohort (GSE54460, GSE70768, GSE70769, and GSE220095), we constructed a canonical Cox proportional hazard regression model of BCR-free survival with the GBM predictor score and assignment to the study as independent variables. Inference of this data set-adjusted GBM model was compared with inference of the initial GBM model.*

*Coefficient estimates expressed as hazard ratios (HR) with 95% confidence intervals of the study-adjusted, the initial and the study-only survival models are presented in a Forest plot. Significance of contribution of the study and GBM predictor score to the BCR-free survival prediction were assessed by likelihood ratio test (LRT), whose results are presented in the plot facets. Note the significant contribution of the GBM predictor score to the prediction of BCR-free survival and the non-significant effect of the study.*

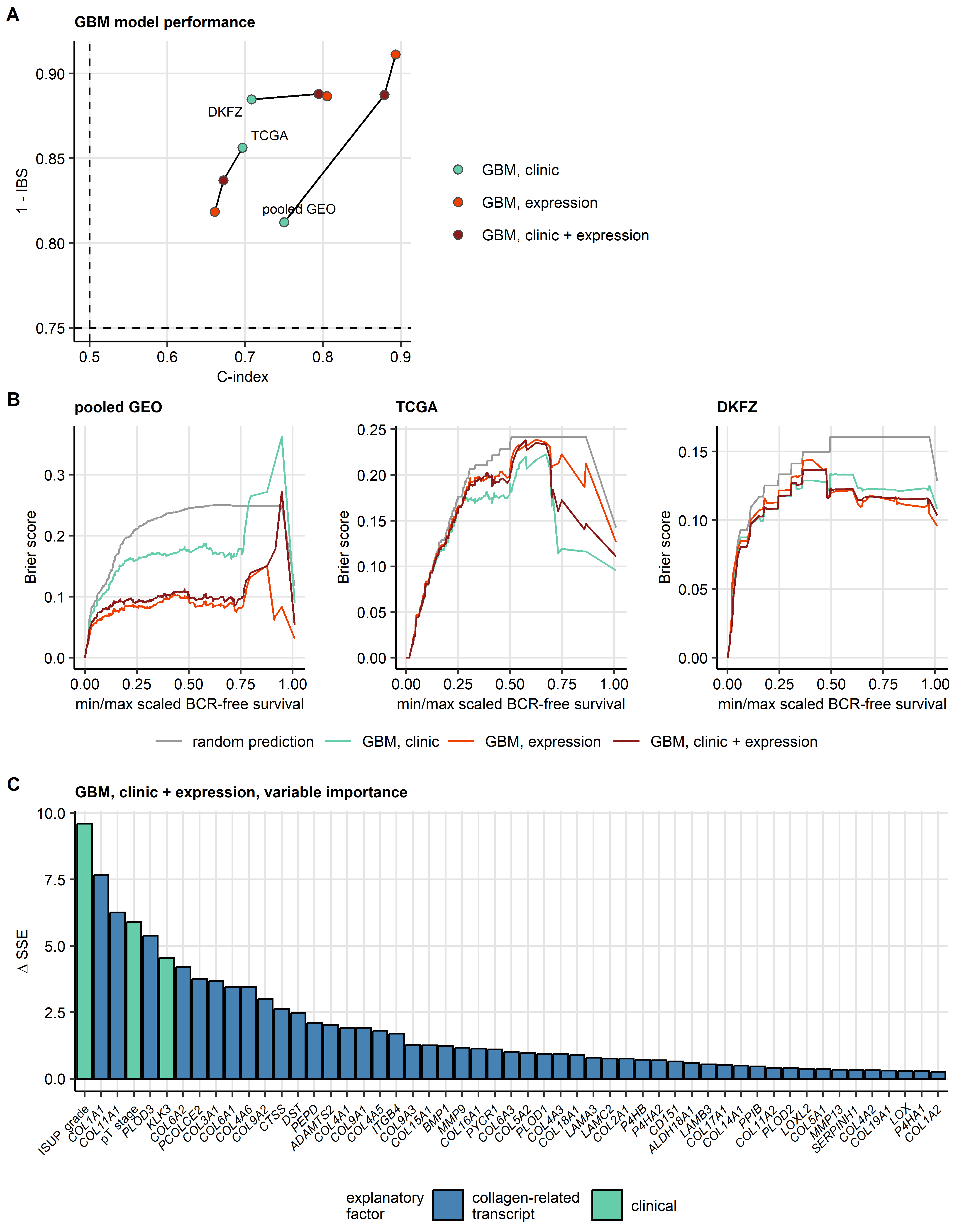


Figure 6: Modeling of BCR-free survival by GBM algorithm with clinical predictors and mRNA expression of the collagen-related genes.

**Figure 6. Modeling of BCR-free survival by GBM algorithm with clinical predictors and mRNA expression of the collagen-related genes.**

*Three Gradient Boosted Machines (GBM) models of biochemical relapse (BCR) free survival were trained in the pooled GEO cohort: a model with clinical predictors (age, pathological tumor stage, ISUP grade, and tumor tissue expression of KLK3 transcript coding for PSA protein), a model with ComBat-adjusted expression levels of the collagen-related transcripts, and a model with the clinical and transcriptomic predictors. Performance of the GBM models was evaluated in the pooled GEO training collective (total: n = 465, BCR: n = 182), and TCGA (total: n = 476, BCR: n = 89) and DKFZ (total: n = 98, BCR: n = 18) validation cohorts.*

*(A) Numeric metrics of survival model performance, Harrell’s concordance index (C-index, measure of model accuracy) and Integrated Brier Score (IBS, measure of calibration, low values are characteristic for well calibrated models), are presented in a dot plot. Each point represents a single GBM model, point color codes for the model type. Models in the same cohort are connected with lines. Note the substantial improvement of the GBM model accuracy and calibration upon inclusion of the collagen-related transcript expression levels in the explanatory variable set in the pooled GEO and DKFZ cohorts.*

*(B) Brier Scores for unique survival time points for the GBM models in the pooled GEO, TCGA, and DKFZ cohort. Low Brier Scores at a particular time point indicate accurate and confident prediction of the BCR risk. Note: gray lines represent random survival predictions by a null model.*

*(C) Importance of the explanatory variables for BCR risk prediction by the GBM model with the clinical and transcriptional predictors. The variable importance was computed as difference in sum of squared error () attributed to inclusion of a particular explanatory factor in the GBM learner ensemble. High values are characteristic for highly influential variables. values for the variables that contributed substantially to the BCR risk prediction ( > 0) are presented in a bar plot.*

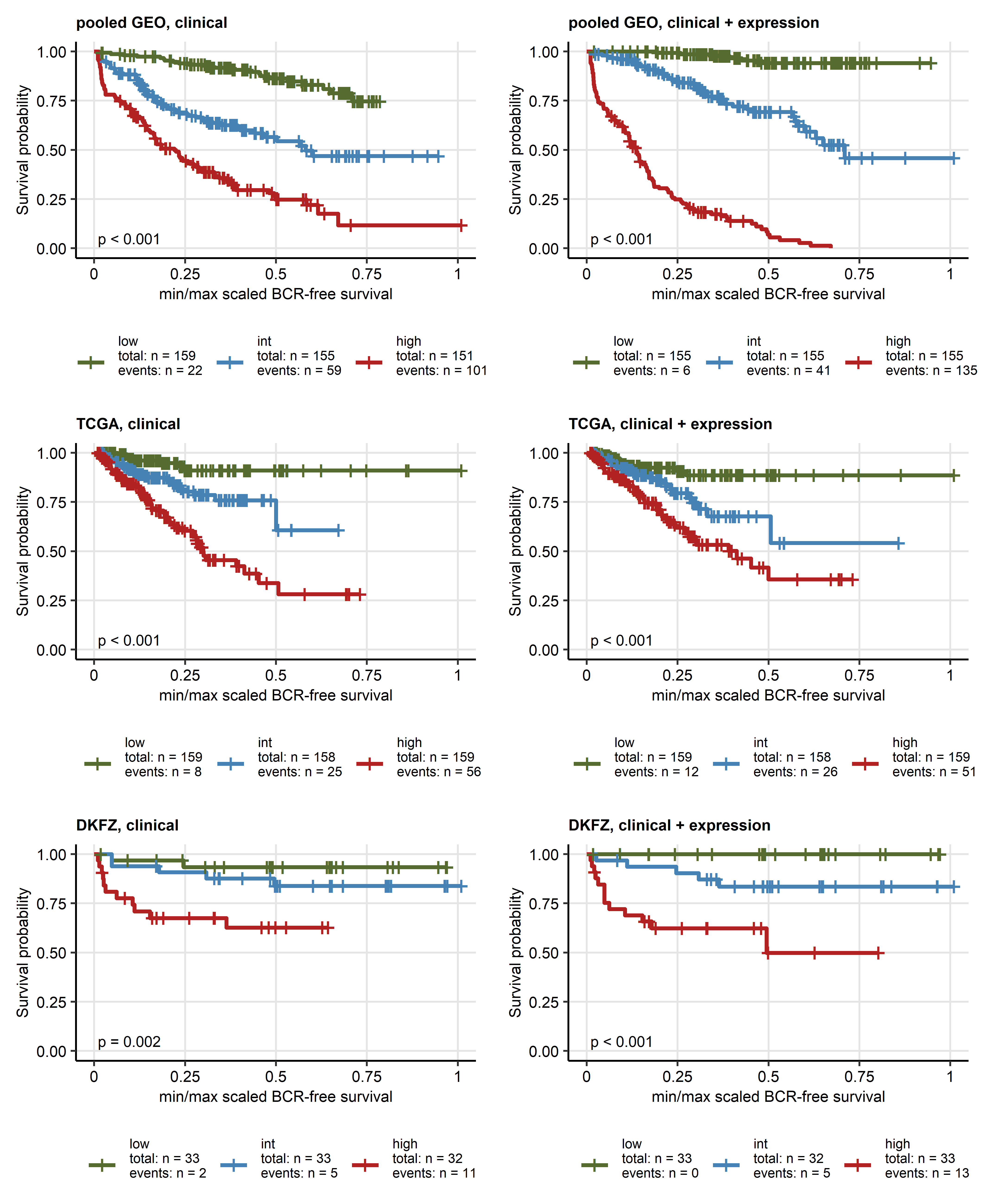


Figure 7: GBM modeling of BCR-free survival in PCA with clinical prediction and expression of collagen-related transcripts: survival in tertiles of predictor scores.

**Figure 7. GBM modeling of BCR-free survival in PCA with clinical prediction and expression of collagen-related transcripts: survival in tertiles of predictor scores.**

*Three Gradient Boosted Machines (GBM) models of biochemical relapse (BCR) free survival were trained in the pooled GEO cohort: a model with clinical predictors (age, pathological tumor stage, ISUP grade, and expression of KLK3 transcript coding for PSA protein), a model with ComBat-adjusted expression levels of the collagen-related transcripts, and a model with the clinical and transcriptomic predictors, as presented in Supplementary Figure S7.* *BCR-free survival in tertiles of the predictor scores of the clinical-only and the combined clinical - expression GBM models was visualized with Kaplan-Meier plots. Numbers of patients and BCR cases in the predictor score tertiles are indicated in the figure legend. P values for differences in survival between the tertiles obtained by Peto-Peto test are displayed in the plots. Note improved resolution of the survival curves for the combined clinical/expression model as compared with the clinical-only model in the pooled GEO and DKFZ cohort.*

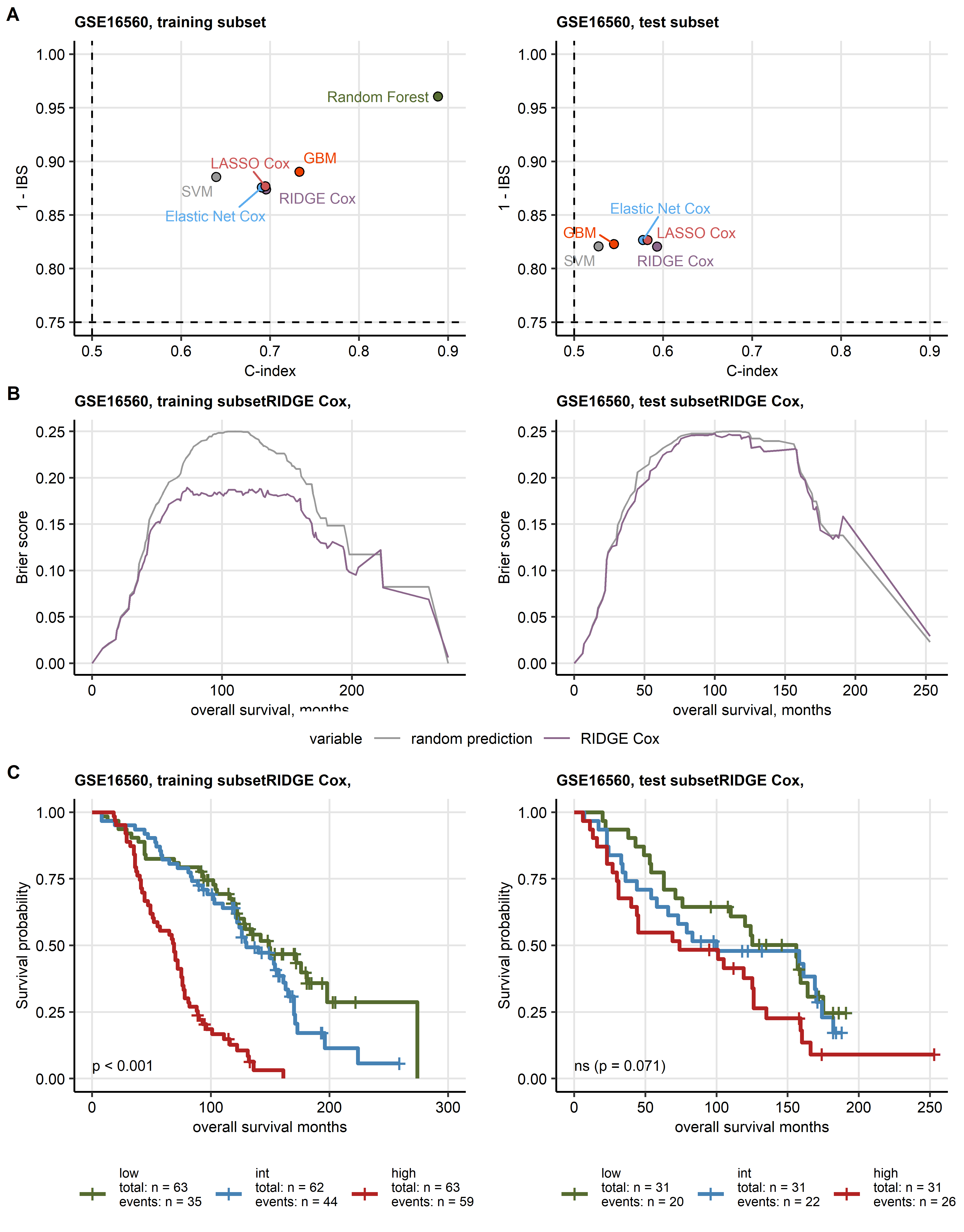


Figure 8: Modeling of overall survival in PCA with expression levels of the collagen-related transcripts as explanatory factors.

**Figure 8. Modeling of overall survival in PCA with expression levels of the collagen-related transcripts as explanatory factors.**

*Overall survival in prostate cancer (PCA) was modeled by multi-parameter machine learning models with ComBat-processed -transformed expression levels of 55 collagen-related transcripts as explanatory factors.* *The models were trained in the training subset of the GSE16560 cohort with six machine learning algorithms (RIDGE Cox regression, Elastic Net Cox regression, LASSO Cox regression, survival Support Vector Machines [SVM], survival Random Forest, and survival Gradient Boosted Machines [GBM]).* *Performance of the models was evaluated in the training (total patients: n = 188, deaths: n = 138) and test (total patients: n = 93, deaths: n = 68) subsets of the GSE16560 cohort.*

*(A) Metrics of performance of the survival models, Harrell’s concordance index (C-index, measure of model accuracy) and Integrated Brier Score (IBS, measure of model calibration, low values are characteristic for good calibration), are presented in dot plots. Dashed lines represent C-index and IBS values expected for random death risk prediction. Note the superior accuracy and calibration of the GBM model in the validation collectives.*

*(B) Brier Scores for unique survival time points for the best performing RIDGE Cox model. Low Brier Scores at a particular time point indicate accurate and confident prediction of the BCR risk. Note: gray lines represent random survival predictions by a null model.*

*(C) Overall survival of PCA patients stratified by tertiles of the predictor scores of the best performing RIDGE Cox model. Statistical significance of differences between the predictor score tertiles was determined by Peto-Peto test corrected for multiple testing wit the false discovery rate method. Fractions of alive patients and time after diagnosis are visualized in Kaplan-Meier plots. Numbers of patients and deaths in the predictor score tertiles are indicated in the legends. P values are displayed in the plots.*

# References

1. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. *Journal of Open Source Software* (2019) 4:1686. doi: [10.21105/joss.01686](https://doi.org/10.21105/joss.01686)

2. Henry L, Wickham Hadley. rlang: Functions for Base Types and Core R and ’Tidyverse’ Features. (2022) <https://cran.r-project.org/web/packages/rlang/index.html>

3. Gagolewski M, Tartanus B. Package ’stringi’. (2021) <https://cran.r-project.org/web/packages/stringi/index.html http://cran.ism.ac.jp/web/packages/stringi/stringi.pdf>

4. Sean D, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)* (2007) 23:1846–1847. doi: [10.1093/BIOINFORMATICS/BTM254](https://doi.org/10.1093/BIOINFORMATICS/BTM254)

5. Pagès H, Carlson M, Falcon S, Li N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. (2022) doi: [10.18129/B9.bioc.AnnotationDbi](https://doi.org/10.18129/B9.bioc.AnnotationDbi)

6. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. (2022) doi: [10.18129/B9.bioc.org.Hs.eg.db](https://doi.org/10.18129/B9.bioc.org.Hs.eg.db)

7. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* (2012) 28:882. doi: [10.1093/BIOINFORMATICS/BTS034](https://doi.org/10.1093/BIOINFORMATICS/BTS034)

8. Kuhn M. Building predictive models in R using the caret package. *Journal of Statistical Software* (2008) 28:1–26. doi: [10.18637/jss.v028.i05](https://doi.org/10.18637/jss.v028.i05)

9. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal* (2006) Complex Sy:1695. <https://igraph.org>

10. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* (2010) 33:1–22. doi: [10.18637/jss.v033.i01](https://doi.org/10.18637/jss.v033.i01)

11. Fouodo CJK, König IR, Weihs C, Ziegler A, Wright MN. Support vector machines for survival analysis with R. *R Journal* (2018) 10:412–423. doi: [10.32614/RJ-2018-005](https://doi.org/10.32614/RJ-2018-005)

12. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *https://doiorg/101214/08-AOAS169* (2008) 2:841–860. doi: [10.1214/08-AOAS169](https://doi.org/10.1214/08-AOAS169)

13. Ishwaran H, Kogalur UB. randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). (2022) <https://cran.r-project.org/web/packages/randomForestSRC/index.html>

14. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models. (2022) <https://cran.r-project.org/package=gbm>

15. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. 1st ed. New York: Springer Verlag. (2000).

16. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using ’ggplot2’. (2016) <https://cran.r-project.org/package=survminer>

17. Wilke CO. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 1st ed. Sebastopol: O’Reilly Media. (2019).

18. Gohel D. flextable: Functions for Tabular Reporting. (2022) <https://cran.r-project.org/web/packages/flextable/index.html>

19. Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J. rmarkdown: Dynamic Documents for R. (2022) <https://cran.r-project.org/web/packages/rmarkdown/index.html>

20. Xie Y. *Bookdown: Authoring books and technical documents with R Markdown*. (2016). doi: [10.1201/9781315204963](https://doi.org/10.1201/9781315204963)

21. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R. (2022) <https://cran.r-project.org/web/packages/knitr/index.html>

22. Heidegger I, Frantzi M, Salcher S, Tymoszuk P, Martowicz A, Gomez-Gomez E, Blanca A, Lendinez Cano G, Latosinska A, Mischak H, et al. Prediction of Clinically Significant Prostate Cancer by a Specific Collagen-related Transcriptome, Proteome, and Urinome Signature. *European urology oncology* (2024) doi: [10.1016/J.EUO.2024.05.014](https://doi.org/10.1016/J.EUO.2024.05.014)

23. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* (2018) 173:400–416.e11. doi: [10.1016/J.CELL.2018.02.052](https://doi.org/10.1016/J.CELL.2018.02.052)

24. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, Annala M, Aprikian A, Armenia J, Arora A, et al. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* (2015) 163:1011–1025. doi: [10.1016/j.cell.2015.10.025](https://doi.org/10.1016/j.cell.2015.10.025)

25. Gerhauser C, Favero F, Risch T, Simon R, Feuerbach L, Assenov Y, Heckmann D, Sidiropoulos N, Waszak SM, Hübschmann D, et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer cell* (2018) 34:996–1011.e8. doi: [10.1016/J.CCELL.2018.10.016](https://doi.org/10.1016/J.CCELL.2018.10.016)

26. Sboner A, Demichelis F, Calza S, Pawitan Y, Setlur SR, Hoshida Y, Perner S, Adami HO, Fall K, Mucci LA, et al. Molecular sampling of prostate cancer: A dilemma for predicting disease progression. *BMC Medical Genomics* (2010) 3: doi: [10.1186/1755-8794-3-8](https://doi.org/10.1186/1755-8794-3-8)

27. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, Gillespie T, Park JY, Nam RK, Sugar L, et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer research* (2014) 74:3228–3237. doi: [10.1158/0008-5472.CAN-13-2699](https://doi.org/10.1158/0008-5472.CAN-13-2699)

28. Ross-Adams H, Lamb A, Dunning M, Halim S, Lindberg J, Massie C, Egevad L, Russell R, Ramos-Montoya A, Vowler S, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine* (2015) 2:1133–1144. doi: [10.1016/j.ebiom.2015.07.017](https://doi.org/10.1016/j.ebiom.2015.07.017)

29. Schimmelpfennig C, Rade M, Füssel S, Löffler D, Blumert C, Bertram C, Borkowetz A, Otto DJ, Puppel SH, Hönscheid P, et al. Characterization and evaluation of gene fusions as a measure of genetic instability and disease prognosis in prostate cancer. *BMC cancer* (2023) 23: doi: [10.1186/S12885-023-11019-6](https://doi.org/10.1186/S12885-023-11019-6)

30. Cohen J. Statistical Power Analysis for the Behavioral Sciences. *Statistical Power Analysis for the Behavioral Sciences* (2013) doi: [10.4324/9780203771587](https://doi.org/10.4324/9780203771587)

31. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* (1995) 57:289–300. doi: [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)

32. Briatte F, Bojanowski M, Canouil M, Charlop-Powers Z, Fisher JC, Johnson K, Rinker T. ggnetwork: Geometries to Plot Networks with ’ggplot2’. (2021) <https://cran.r-project.org/package=ggnetwork>

33. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* (1996) 15:361–387. doi: [10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4)

34. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* (1999) 18:2529–2545. doi: [10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5](https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5)

35. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* (2011) 39:1–13. doi: [10.18637/JSS.V039.I05](https://doi.org/10.18637/JSS.V039.I05)

36. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2005) 67:301–320. doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)

37. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)* (1996) 58:267–288. doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)

38. Weston J, Watkins C. Multi-Class Support Vector Machines. (1998)

39. Breiman L. Random forests. *Machine Learning* (2001) 45:5–32. doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)

40. Friedman JH. Greedy function approximation: A gradient boosting machine. *https://doiorg/101214/aos/1013203451* (2001) 29:1189–1232. doi: [10.1214/AOS/1013203451](https://doi.org/10.1214/AOS/1013203451)

41. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* (2013) 7:63623. doi: [10.3389/FNBOT.2013.00021/BIBTEX](https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX)

42. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* (2002) 38:367–378. doi: [10.1016/S0167-9473(01)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)

43. Lunger L, Retz M, Bandur M, Souchay M, Vitzthum E, Jäger M, Weirich G, Schuster T, Autenrieth M, Kübler H, et al. KLK3 and TMPRSS2 for molecular lymph-node staging in prostate cancer patients undergoing radical prostatectomy. *Prostate cancer and prostatic diseases* (2021) 24:362–369. doi: [10.1038/S41391-020-00283-3](https://doi.org/10.1038/S41391-020-00283-3)

44. Boyukozer FB, Tanoglu EG, Ozen M, Ittmann M, Aslan ES. Kallikrein gene family as biomarkers for recurrent prostate cancer. *Croatian Medical Journal* (2020) 61:450. doi: [10.3325/CMJ.2020.61.450](https://doi.org/10.3325/CMJ.2020.61.450)

45. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* (1994) 81:515. doi: [10.2307/2337123](https://doi.org/10.2307/2337123)

46. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review* (1950) 78:1–3. doi: [10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2](https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2)