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

Supplementary material, transcriptome part

2025-02-23

# Supplementary Table

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 overall 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 overall model calibration. Low values are expected for a well calibrate 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 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: 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 overall model calibration. Low values are expected for a well calibrate model. Intergrated Brier score = 0.25 is expected for random predictions. | | | |

# Supplementary Figures

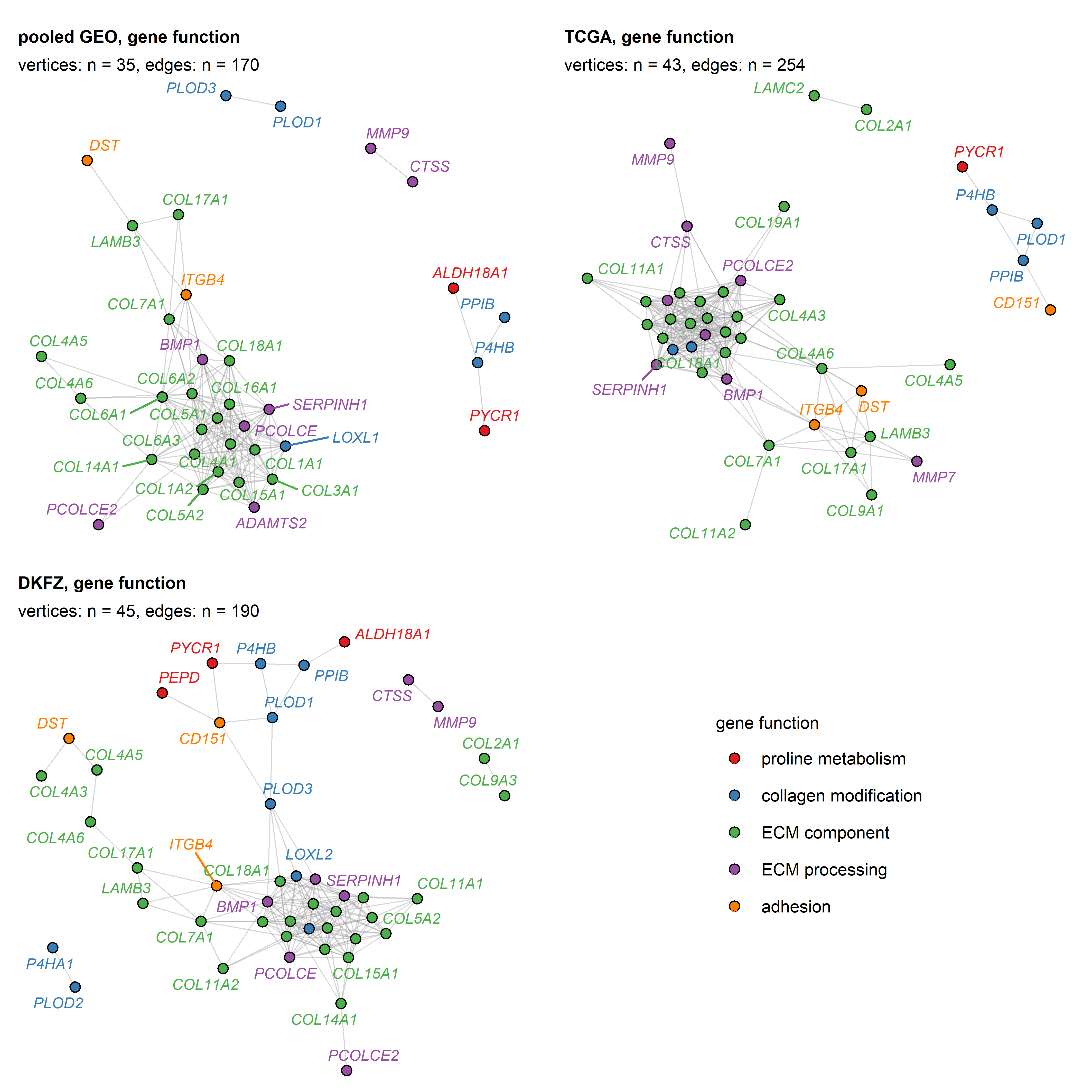


Figure 1: Co-expression networks of the collagen-related transcripts in PCA tissue.

**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 Fruchterman-Reingold 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.*

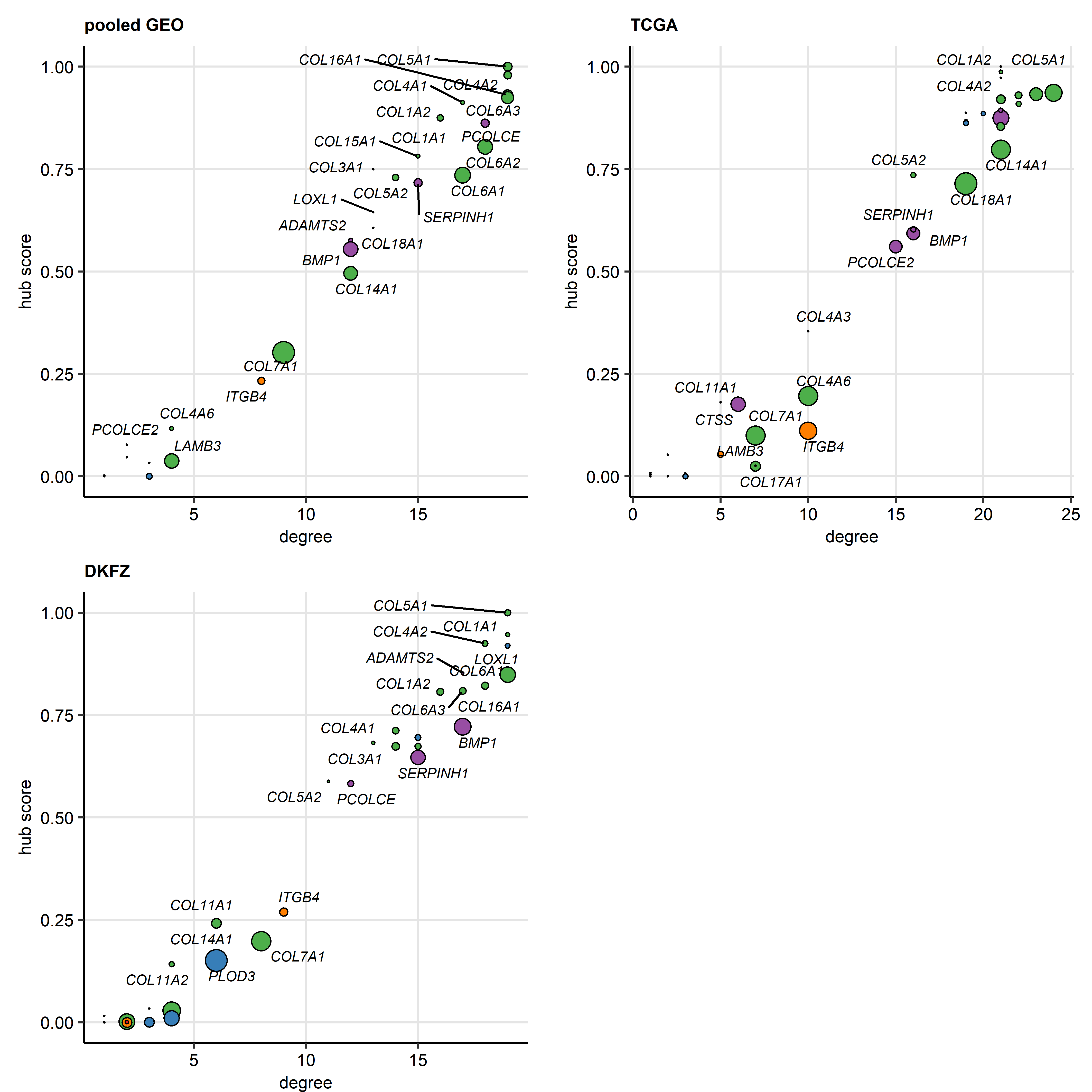


Figure 2: Vertex importance statistics for co-expression networks of the collagen-related transcripts in PCA.

**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 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 collage-related transcript, point color codes for functional classification of the transcript.*

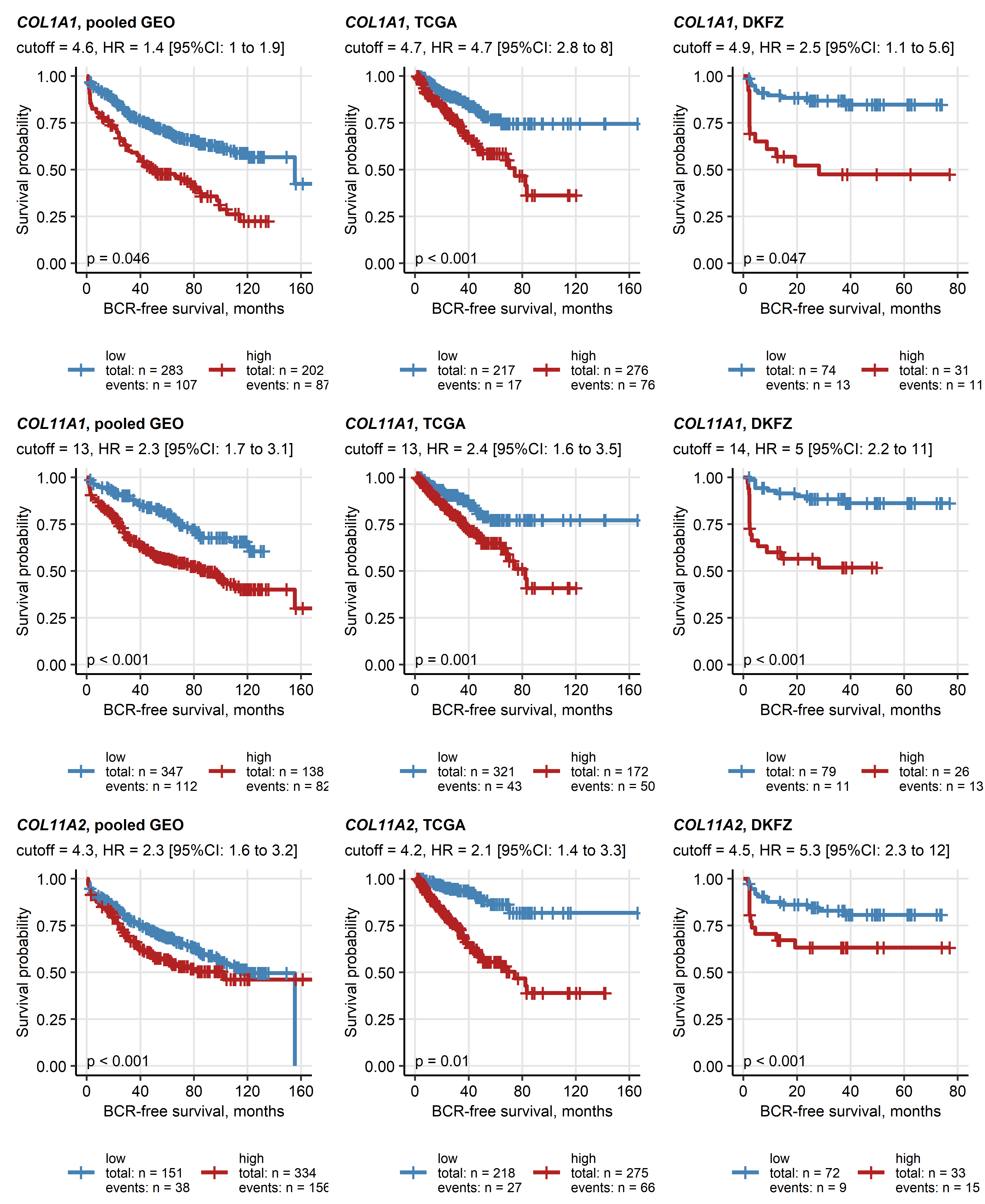


Figure 3: Top three unfavorable standalone collagen-related transcriptional markers of BCR-free survival in PCA.

**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 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 (p(HR = 0)) 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.*

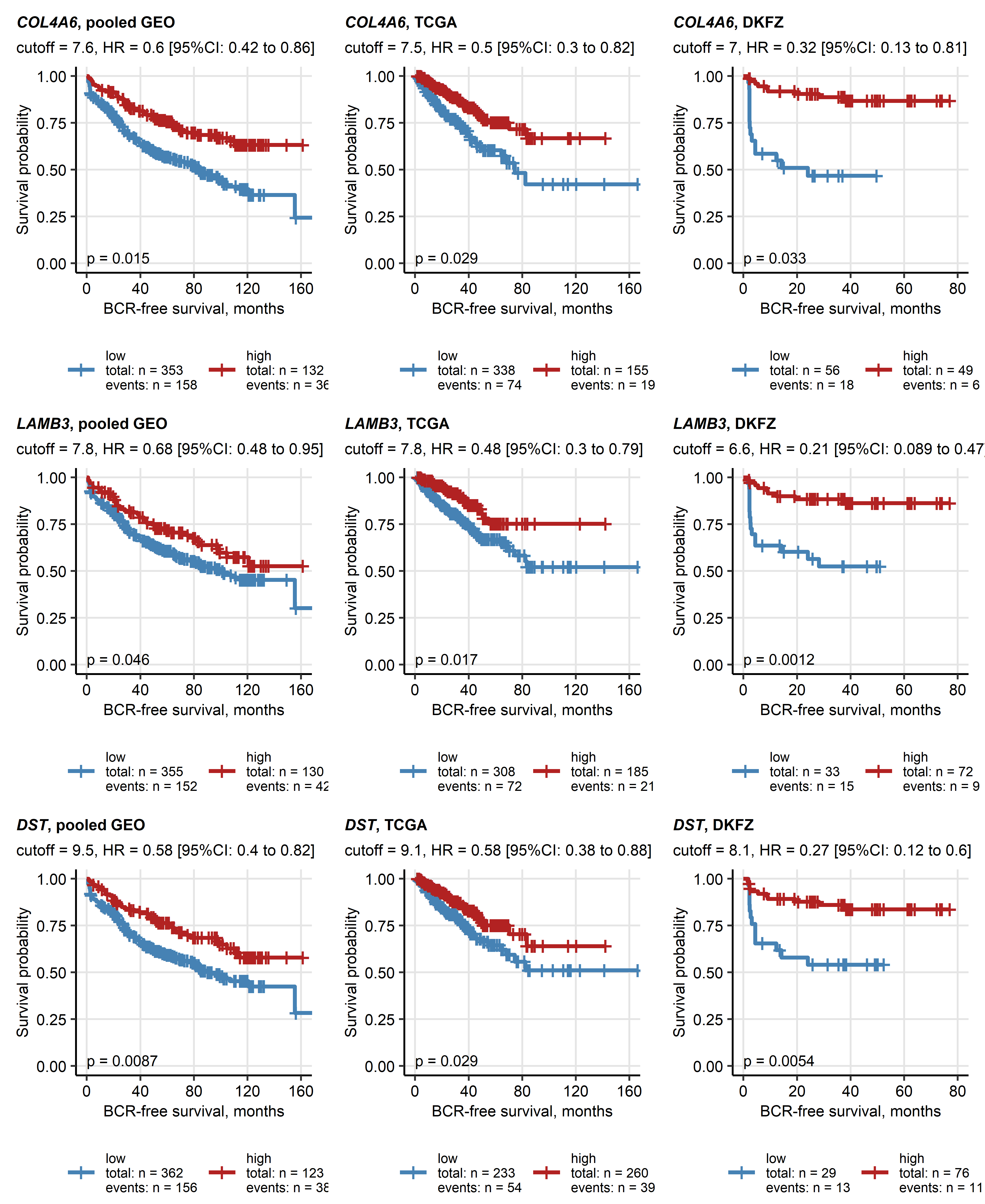


Figure 4: Top three favorable standalone collagen-related transcriptional markers of BCR-free survival in PCA.

**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 (p(HR = 0)) 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.*

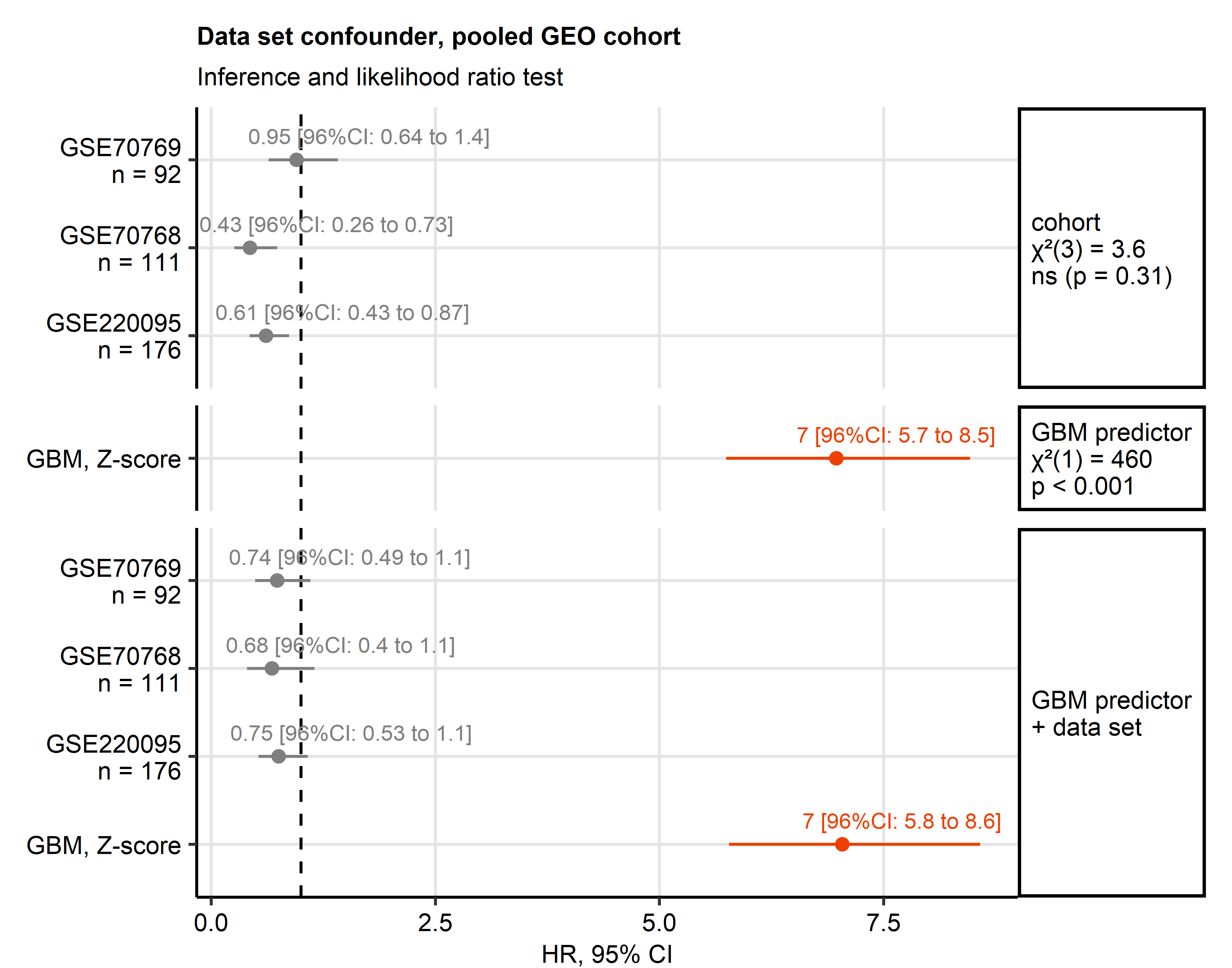


Figure 5: Investigation of confounding effect of the cohort on prediction of BCR-free survival by the GBM model in the pooled GEO cohort.

**Figure 5. Investigation of confounding effect of the cohort 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 data sets 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 data sets as independent variables. Performance and inference of this data set-adjusted GBM model was compared with performance and inference of the initial GBM model.*

*Coefficient estimates expressed as hazard ratios (HR) with 95% confidence intervals of the data set-adjusted, the initial and the data set-only survival models are presented in a Forest plot. Significance of contribution of the cohort 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 data set.*

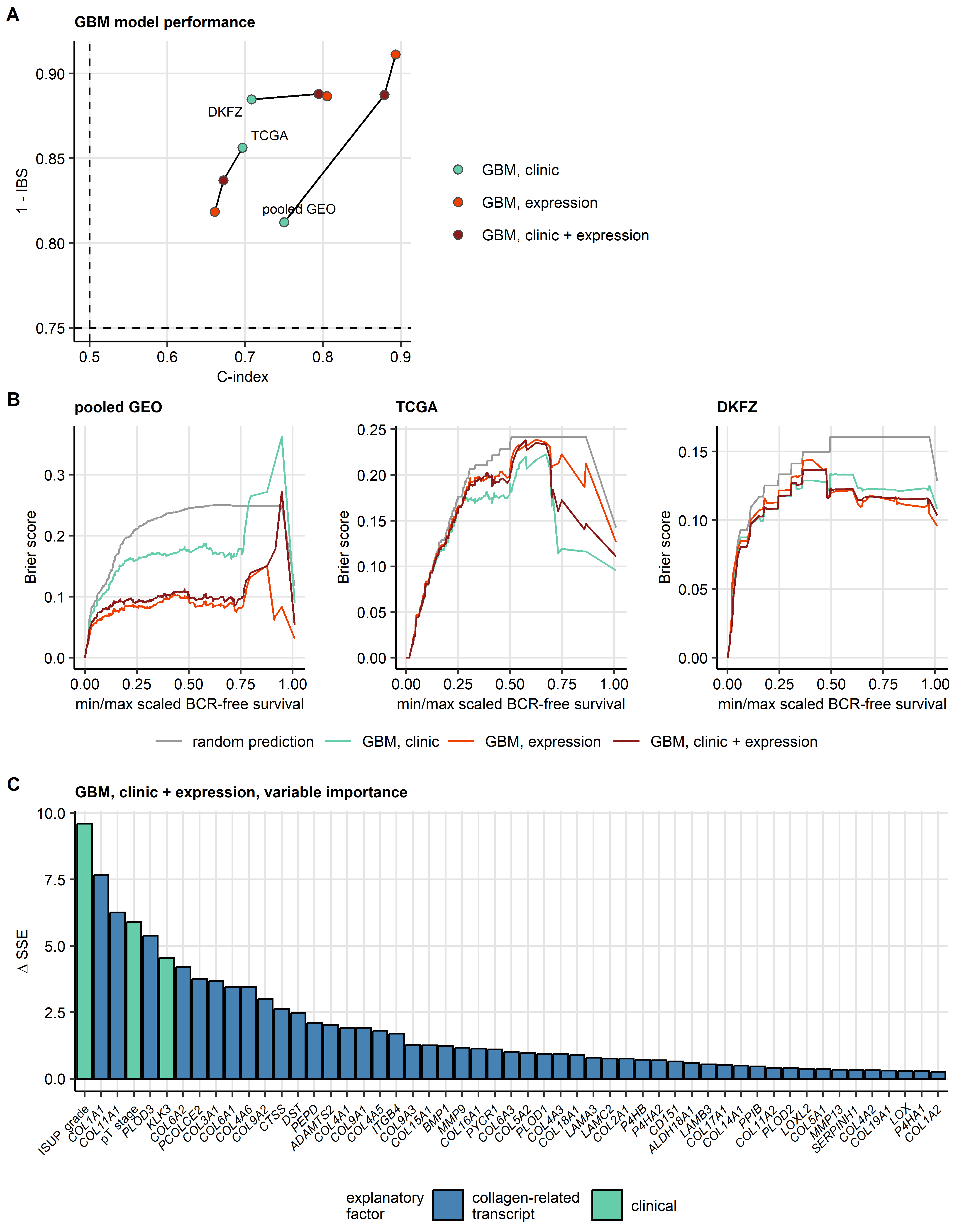


Figure 6: Modeling of BCR-free survival by GBM algorithm with clinical predictors and mRNA expression of the collagen-related genes.

**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, measured of model accuracy) and Integrated Brier Score (IBS, measure of overall 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 fit 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.*

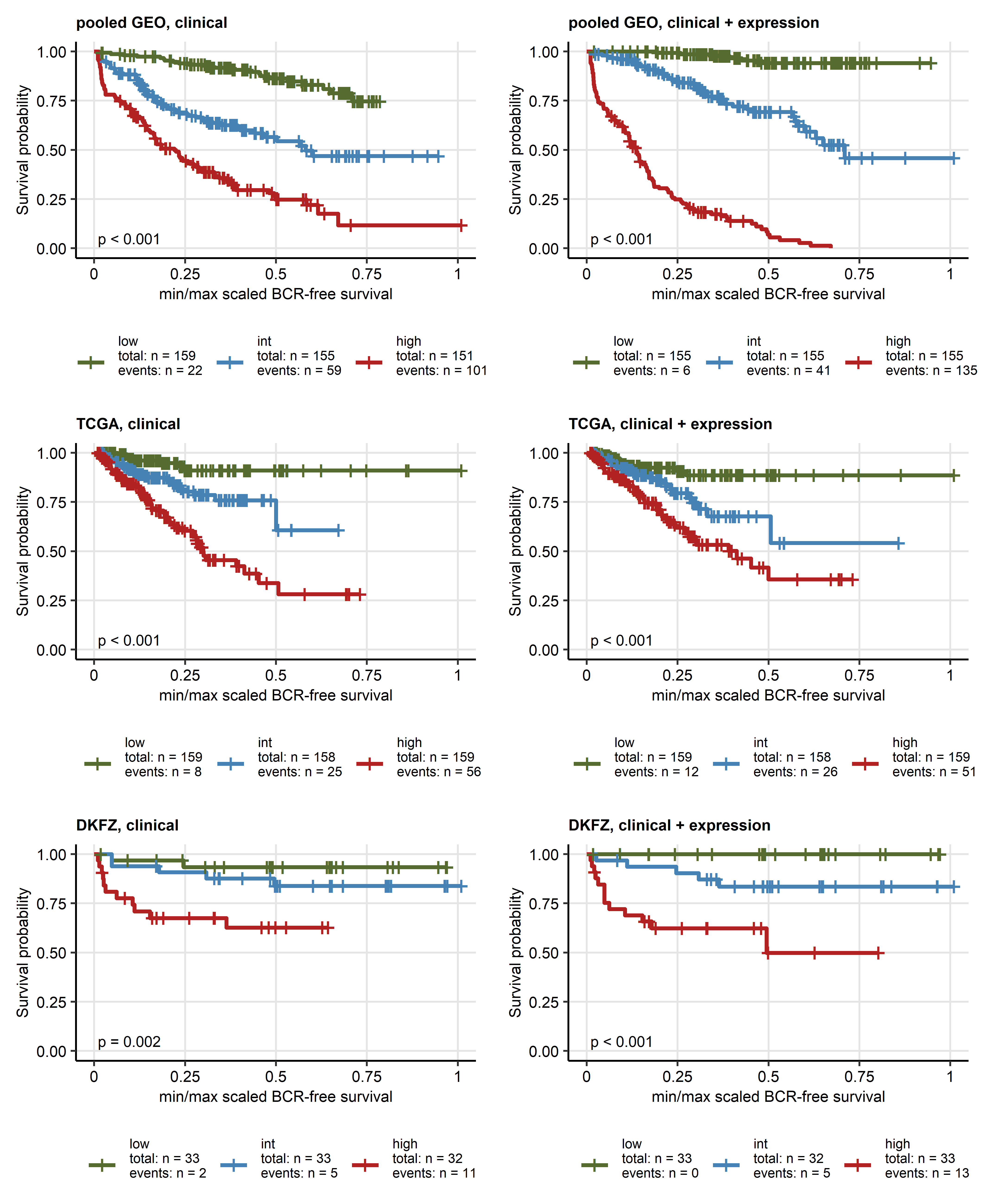


Figure 7: GBM modeling of BCR-free survival in PCA with clinical prediction and expression of collagen-related transcripts: survival in tertiles of predictor scores.

**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 expressin 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.*

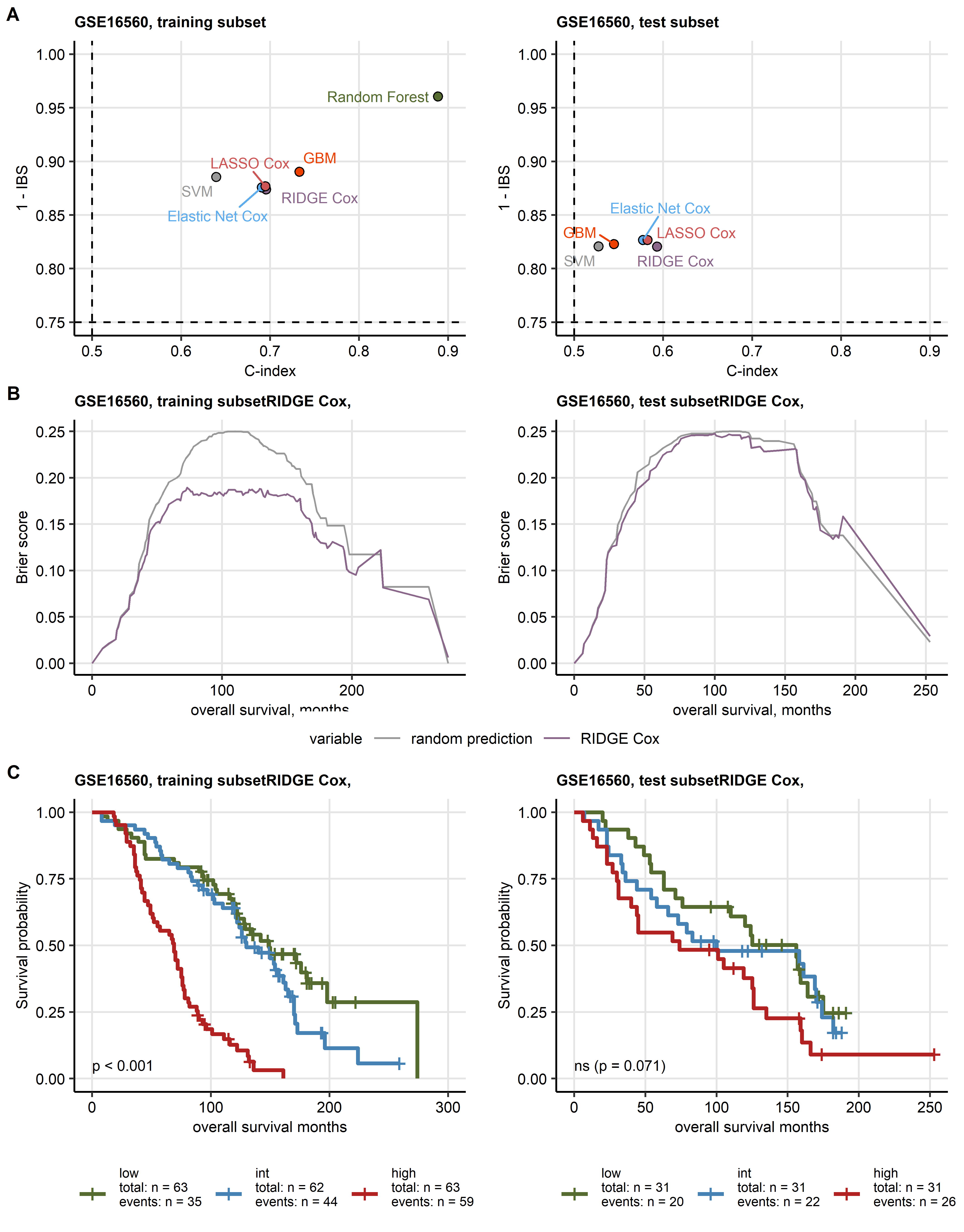


Figure 8: Modeling of overall survival in PCA with expression levels of the collagen-related transcripts as explanatory factors.

**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 overall 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