Survival Prediction of Breast Cancer Using Time-to-Event Model

Dongbing Han,[1,*] Shanshan Gong[2,†] and Chengyu Wang[3,‡]

[1, 2, 3]Computer Science Department, Columbia University, 116th and Broadway, 10025, NY, USA
[*]Corresponding author. Dongbing Han, dh3071@columbia.edu
[†]Corresponding author. Shanshan Gong, shanshan.gong@columbia.edu
[‡]Corresponding author. Chengyu Wang, cw3512@columbia.edu

Abstract

This study explored the prognostic capabilities of clinical attributes and mRNA genetic attributes in breast cancer survival, using data from the METABRIC database. The analysis utilized both univariate and multivariate Cox proportional hazards models to identify significant mRNA genetic attributes and their association with survival outcomes. The research discovered that integrating clinical attributes with mRNA expression data didn't notably improve survival prediction compared to using clinical attributes alone. These results indicate that incorporating mRNA data, despite its added complexity, doesn't enhance predictive accuracy beyond conventional clinical markers.

Key words: Survival prediction, Cox hazards model, mRNA, Time-to-event

Introduction

Time-to-event outcome prediction models are essential in biomedical research, providing personalized probabilities crucial for clinical decision-making. Various regression and machine learning techniques have been customized or developed to handle the inherent censoring in such data. This study examines the predictive power of clinical attributes and mRNA expression profiles in forecasting breast cancer survival, using data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database.

A 26-mRNA signature was identified and assessed using Cox proportional hazards models to determine its prognostic value in survival analysis, both independently and in conjunction with clinical attributes such as patient age, tumor size, and receptor status (ER, PR, HER2). Cox proportional hazards models, including both univariate and multivariate approaches, were employed for analysis. Additionally, Kaplan-Meier survival curves and log-rank tests were employed to compare survival outcomes across different risk categories.

The METABRIC dataset, a cornerstone in breast cancer research, provides a comprehensive repository of molecular and clinical information from over two thousand patients. It includes extensive clinical data (30 attributes), covering patient demographics, treatment histories, and survival outcomes, as well as genomic information (over 500 attributes) such as mRNA expression profiles and DNA mutations. Molecular subtypes, including ER, HER2, PR, and triple-negative/basal-like subtypes, are also incorporated in the clinical attributes, facilitating investigations into subtype-specific features and outcomes. With enriched pathological data and comprehensive survival information, the dataset empowers researchers to delve into the complexity of the disease and develop predictive models for patient outcomes.

Our results indicate that although the mRNA signature offers some prognostic insight, integrating it with clinical data does not notably improve prediction accuracy beyond what is achieved with clinical data alone. This lack of enhancement could be attributed to the complexity and potential overfitting associated with integrating high-dimensional genetic data. These findings emphasize the need to meticulously assess the additional value of genetic data within the framework of existing clinical predictors for breast cancer prognosis.

Access the METABRIC dataset publicly at the following link [1] and Access the code at the github[2]

Survival Analysis

This study employs a comprehensive set of statistical models to analyze the genetic attributes associated with the survival rates of breast cancer patients, integrating mRNA expression levels with clinical outcomes. The study rigorously followed established data analysis procedures to ensure the integrity and accuracy of the results. Each subsection details the particular analytical step or method utilized, outlining a systematic framework for our investigation.

[1] https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/
[2] https://github.com/hando189890/CBMF_4761_Bioinformatics_Project/

Survival Analysis

First, survival analysis focuses on analyzing time-to-event data, which is the duration until one or more events occur. In the context of medical research, this often refers to the time from the start of a study until the occurrence of a specific event, such as death or disease recurrence. Thus, it suits perfectly for this study. The primary challenge in survival analysis is handling censored data—instances where the patient's event outcome is unknown at the end of the study period. To address this, methods like the Univariate and Multivariate Cox regression models are utilized to explore the relationship between covariates and survival time while and estimators and metrics like the Kaplan-Meier estimator is used to estimate the survival functions.

Hazard Function

A central concept in this study's survival analysis is the hazard function, which provides a measure of the risk of the event occurring at a specific time, assuming survival until that time. The hazard function is crucial as it helps in understanding the instantaneous rate at which events happen, absent right-censoring. Cox regression models, both univariate and multivariate, focus on estimating the hazard ratio, which compares the hazard rates between groups of subjects. These ratios are pivotal in medical research as they facilitate the identification of risk factors and the evaluation of treatment effects. By incorporating covariates into the hazard function, Cox models allow researchers to adjust for multiple confounders simultaneously, providing a more nuanced understanding of the factors that influence survival. Moreover, these models do not assume a constant hazard ratio over time, which adds flexibility in dealing with complex survival data that might not follow a simple pattern.

Right Censoring

Survival analysis is particularly adept at handling right-censored data, which is a common challenge in medical research, like breast cancer studies. Right censoring occurs when the event of interest (such as death or disease recurrence) has not happened by the end of the study period or if the patient is lost to follow-up. This means that the exact time of the event is unknown, but it is known to occur after a certain time point.

Handling right-censored data appropriately is critical in breast cancer research because it ensures that all available information is utilized without introducing bias. Breast cancer studies often involve long-term follow-up periods, during which patients might drop out due to various reasons other than the cancer itself, such as moving away or opting out of the study. Moreover, many patients may still be alive at the end of the study, which is a positive outcome but complicates the statistical analysis because the full survival times are not known.

The Cox proportional hazards model is particularly suited for these types of data. It can effectively handle the incomplete data by calculating the hazard ratio between different patient groups while considering that some data points are censored. This model provides valuable insights into how different treatments or genetic markers influence survival, accounting for the fact that not every patient's event outcome is observed. In the context of breast cancer, where treatments and outcomes can vary widely among patients, using a statistical approach that accommodates right censoring is crucial for drawing accurate and meaningful conclusions from clinical trials and observational studies.

Methods

Univariate Cox Regression Analysis

Our initial analysis involved assessing 506 genetic attributes using univariate Cox regression models to correlate individual gene expression levels with time-to-event survival data. This step was crucial for feature selection, as it allowed us to focus on genes most likely to influence survival outcomes. To qualify for further analysis, genes needed to exhibit a p-value less than 0.000001 ($p < 0.000001$), ensuring a robust association with patient survival. We also computed the fold change of each gene, applying a log2 transformation to the ratio of mean expression levels between patient groups. Adjustments were made to prevent division by zero, thereby refining our understanding of gene regulation's role—whether upregulation or downregulation—in affecting survival rates. Furthermore, this analytical phase enabled the identification of potential biomarkers for early detection and therapeutic targets.

Multivariate Cox Regression Analysis

Following the identification of significant genes from the univariate analysis, a multivariate Cox regression analysis was conducted to further refine our understanding of genetic influences on breast cancer survival. This statistical model incorporated the selected genetic attributes as covariates, along with time-to-event data as the outcome variable. This step aimed to ascertain which genes maintain their prognostic value independently of other factors, thereby isolating the direct impact of specific genetic attributes on survival probabilities. Additionally, this phase helped us adjust for potential confounders, ensuring the reliability of our prognostic indicators while providing insights into the complex interactions between genes.

Risk Model

The significant mRNAs genetic attributes were utilized to construct a prognostic signature for predicting patient survival. A risk score model was established using a weighted approach, where regression coefficients and mRNA expression levels were factored into a risk formula. For each significant mRNAs $i \forall i \in [1, n]$, multiply the regression coefficient reflecting the contribution of each mRNA $coef_i$ with the value or expression level of each mRNA x_i.

$$\text{RiskScore} = \sum_{i=1}^{n} (coef_i * x_i)$$

This formula was then applied to compute the risk score for each patient, and then divide the patients into high-risk and low-risk groups based on the median value of the risk scores. The survival rates of these two groups were analyzed using the Kaplan-Meier method, which generated survival curves depicting the proportion of patients surviving over specified periods post-treatment. The survival experiences of the high-risk and low-risk groups were statistically compared using a log-rank test, revealing a significant difference in overall survival ($p < 0.001$). These findings underscore the utility of our risk model as an effective prognostic tool.

Statistical Analysis

This study further investigated the significant mRNAs to assess whether the risk score correlated with the clinical characteristics of the breast cancer patients. Thus a Chi-Square Test to determine the relationship between genetic attributes and clinical features was conducted. A significant result in this test indicates a meaningful distinction between patients with high-risk scores and those with low-risk scores for specific clinical attributes. They suggest that certain genetic markers are not only statistically associated with survival outcomes but are also linked to distinct clinical profiles. This supports the clinical relevance of our genetic profiling, indicating potential pathways through which genetic variations influence disease progression and treatment response. By identifying these associations, the study paves the way for more targeted therapeutic strategies, potentially guiding personalized treatment plans that consider both genetic and clinical factors. This tailored approach could lead to improved patient outcomes, as treatments could be better aligned with the individual genetic and clinical context of each patient.

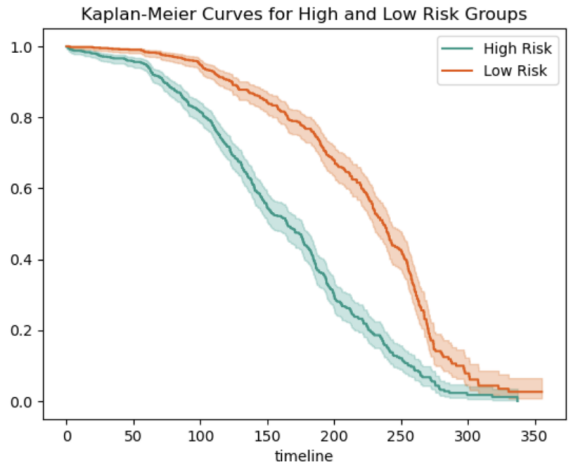

Fig. 1. Kaplan-Meier Curve. Survival curves depicting the proportion of patients surviving over specified periods for High Risk and Low Risk groups

Results

Identification of Significant mRNA Genetic Attributes Associated with Survival

First, by constructing a Univariate Cox regression analysis to identify mRNAs whose expression was strongly associated with overall survival, specifically retaining those with a p-value less than 0.000001 ($p < 0.000001$). From Table 1, genes such as MYC, CHEK2, and AKT1 exhibited p-values below this threshold and were thus considered significantly related to patient survival.

From a biological perspective, the relevance of these findings is supported by existing knowledge. The overexpression of MYC oncogene in breast cancer is known to contribute to increased tumor growth, aggressiveness, and poor prognosis through its role in regulating cell proliferation and apoptosis. Similarly, mutations in the CHEK2 tumor suppressor gene have been linked to a heightened risk of breast cancer, attributable to its integral role in DNA damage response and repair processes that preserve genomics stability. Additionally, the AKT1 gene, part of the PI3K/AKT signaling pathway, is often altered in breast cancer. This pathway's disruption promotes cell survival, proliferation, and chemotherapy resistance, thereby influencing patient outcomes and response to treatment.

Construction of Risk Model

A multivariate Cox regression analysis was performed to identify genes with independent prognostic significance using selected genetic attributes and time-to-event survival data as shown in Table 2.

The analysis revealed that the genes KMT2C and LAMA2 had regression coefficients less than zero, and their corresponding HRs were also less than 1, categorizing these genes as protective factors. This indicates that lower expression levels of these mRNAs are associated with better survival outcomes, suggesting their role in tumor suppression or slower disease progression.

Conversely, the gene MYC was identified as a risk factor, with a regression coefficient greater than zero and an HR greater than 1. This suggests that higher expression levels of MYC are associated with poor survival outcomes, reflecting its potential role in tumor progression or aggressiveness. The upregulation of this high-risk mRNA is thus correlated with poor overall survival, highlighting its prognostic significance.

Following the identification of these key genetic markers, a risk score for each patient was calculated based on their gene expression profiles. The survival curves for the two groups—categorized as high-risk and low-risk based on their risk scores—were plotted to visually assess the impact of the genetic attributes on survival outcomes. The resulting survival curves, illustrated in Figure 1, display a marked difference between the two groups, with the high-risk group showing significantly shorter survival times compared to the low-risk group.

To statistically validate these observations, a Log-Rank test was performed. The test produced a p-value of 3.79×10^{-38}, which strongly suggests a significant difference in survival times between the high-risk and low-risk groups. This substantial p-value confirms the effectiveness of the risk model in discriminating between different survival outcomes based on the genetic profile of the patients.

These results underscore the utility of integrating specific genetic markers into prognostic models for breast cancer. By identifying and quantifying the influence of protective and risk factors like KMT2C, LAMA2, and MYC, this model facilitates a deeper understanding of the genetic underpinnings of breast cancer prognosis. Moreover, the significant stratification of patients based on risk scores indicates that such genetic profiling can be instrumental in guiding clinical decisions and tailoring treatment strategies to individual patient profiles, ultimately aiming to improve therapeutic outcomes.

Validation of the mRNA-focused genetic attributes

Statistical analysis revealed that patients with high-risk scores differed significantly from those with low-risk scores in several key aspects, such as the 3-gene classifier subtype and estrogen receptor (ER) status, among others (Figure 2). Specifically, the Chi-Square Test indicated that the distribution of the 3-gene classifier subtypes—which categorize patients based on the genetic expression of HER2, estrogen, and progesterone receptors—varied significantly between the two risk groups.

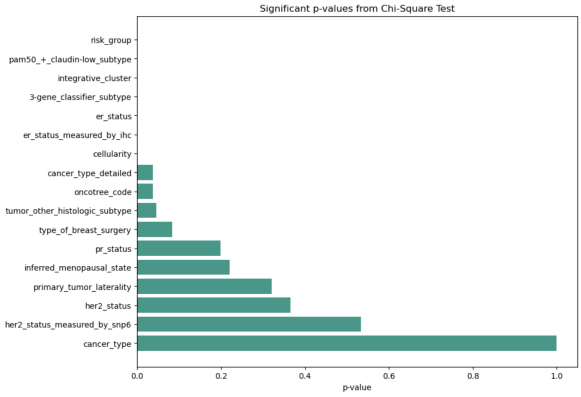

Fig. 2. Chi-Square Test. Bar graph depicting that clinical attributes for patients with high risk scores were significantly different from those with low risk scores.

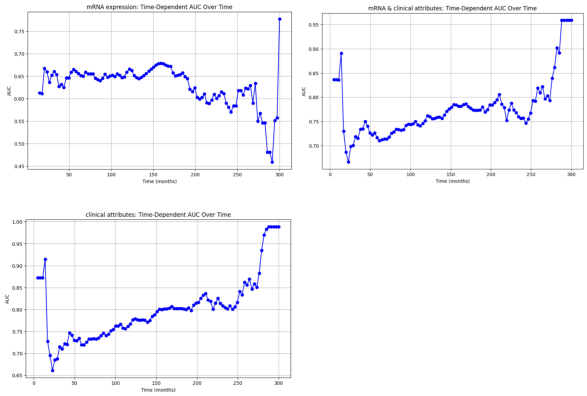

Fig. 3. Time-dependent AUC metrics. AUC curve of the model for 26-mRNA signature alone, 26-mRNA signature and clinical attributes, clinical attributes alone.

Patients in the high-risk group were more likely to be classified under subtypes associated with poorer prognosis.

Similarly, a significant difference was observed in the ER status, with the high-risk group showing a higher proportion of ER-negative status, which is generally linked with a more aggressive disease course and a lower response rate to hormonal therapies. These differential patterns underscore the potential of mRNA genetic attributes to not only predict survival outcomes but also to delineate distinct biological behaviors within breast cancer.

These findings suggest that the mRNA genetic attributes can serve as independent prognostic indicators in breast cancer. The ability of these genetic markers to stratify patients into clinically meaningful risk groups supports their use in refining prognostic models. Moreover, this stratification can guide the development of tailored treatment strategies, potentially enhancing therapeutic efficacy by aligning interventions more closely with the genetic and molecular landscape of individual tumors. The integration of such biomarkers into clinical practice could therefore significantly impact decision-making processes in oncology, optimizing treatment approaches and improving patient outcomes.

AUC Analysis of models with different input

In survival analysis, the time-dependent Area Under the Curve (AUC) enhances the traditional AUC by integrating the time dimension crucial in survival data. This metric gauges the model's capability to accurately differentiate between patients who are likely to experience an event and those who are less likely at specific time points across the study period. A perfect model would score a 1.0, indicating flawless discrimination, whereas a score below 0.5 would suggest that the model performs no better than random chance, rendering it ineffective.

Figure 3 illustrates that the 26-mRNA signature achieves a mean AUC of approximately 0.65, indicating a moderate ability of the mRNA attributes to predict survival outcomes. This performance, while not exceptional, demonstrates that genetic information contributes predictive value to the model.

When assessing the combination of 26-mRNA and clinical attributes, the results are comparable to those obtained using only clinical data, with both models averaging an AUC of around 0.78. This similarity in performance suggests a plateau in predictive improvement when integrating mRNA data with clinical variables. The lack of incremental benefit when combining these datasets could be due to the overlap in the information content. Clinical variables such as tumor stage, receptor status, or other biomarkers may already encapsulate the effects of underlying gene expressions. These clinical markers often reflect the physiological manifestations of genetic activity, potentially making additional genomic data redundant.

Conclusion

In this research, we have validated a 26-mRNA signature that demonstrates significant potential as a biomarker for predicting breast cancer prognosis. The robust statistical analysis highlights its utility in stratifying patients into high-risk and low-risk categories, facilitating more personalized treatment strategies in clinical settings. This targeted approach could lead to more efficient resource allocation and improved patient outcomes by enabling oncologists to tailor treatments based on individual risk profiles.

Furthermore, our findings reveal meaningful correlations between mRNA expression levels and critical clinical features such as the 3-gene classifier subtype and estrogen receptor status. These correlations underscore the capacity of molecular diagnostics to enhance prognosis predictions in breast cancer, moving beyond traditional clinical parameters.

However, the clinical implementation of this 26-mRNA signature requires further validation across diverse patient populations. It is also crucial to integrate deeper biological insights to refine feature selection and apply advanced survival models that accommodate the complexities of the data. Additionally, the cost-effectiveness of such genetic testing must be evaluated in light of healthcare economic constraints.

In conclusion, this study advances our understanding of the genetic foundations of breast cancer and sets the stage for the application of these insights in clinical practice. The potential for improved patient management through enhanced risk prediction and personalized treatment plans marks a significant advancement in oncology. Ongoing research, validation, and interdisciplinary collaboration are essential to fully realize the potential of genetic profiling in enhancing cancer care.

References

1. H. Guo, C. Li, X. Su, and X. Huang. A Five-mRNA Expression Signature to Predict Survival in Oral Squamous Cell Carcinoma by Integrated Bioinformatic Analyses. *Genetic Testing and Molecular Biomarkers*, 25(8):517–527, 2021.
2. N. Ma, L. Si, M. Yang, M. Li, and Z. He. A highly expressed mRNA signature for predicting survival in patients with stage I/II non-small-cell lung cancer after operation. *Scientific Reports*, 11(1):5855, 2021.
3. Navodini Wijethilake, Dulani Meedeniya, Charith Chitraranjan, Indika Perera Survival prediction and risk estimation of Glioma patients using mRNA expressions *arXiv preprint arXiv:2011.00659*, 2020
4. Raktim Kumar Mondol, Ewan K.A. Millar, Arcot Sowmya, Erik Meijering BioFusionNet: Deep Learning-Based Survival Risk Stratification in ER+ Breast Cancer Through Multifeature and Multimodal Data Fusion *arXiv preprint arXiv:2402.10717*, 2024

Table 1. Significance of mRNAs in Univariate Cox Regression Analysis

mRNA	p-value	mRNA	p-value	mRNA	p-value
chek2	9.0×10^{-7}	mlh1	5.9×10^{-11}	rb1	2.0×10^{-26}
myc	9.9×10^{-22}	cdkn1a	2.0×10^{-7}	e2f5	6.3×10^{-15}
jak1	3.9×10^{-29}	stat2	3.1×10^{-12}	stat3	1.4×10^{-11}
adam10	8.0×10^{-15}	adam17	4.6×10^{-18}	ctbp1	4.9×10^{-11}
dtx3	4.4×10^{-13}	ep300	1.0×10^{-9}	jag2	2.1×10^{-8}
maml3	6.8×10^{-13}	ncstn	2.1×10^{-9}	notch3	2.2×10^{-9}
nrarp	1.2×10^{-8}	psen1	1.7×10^{-30}	rbpj	2.1×10^{-7}
acvr2b	6.4×10^{-7}	akt1	2.8×10^{-7}	akt1s1	4.6×10^{-15}
atr	1.1×10^{-7}	bmp6	2.4×10^{-12}	casp10	4.7×10^{-11}
casp6	1.0×10^{-11}	casp7	7.8×10^{-7}	casp8	2.4×10^{-25}
cxcl8	4.8×10^{-7}	cxcr1	1.9×10^{-9}	eif4e	7.8×10^{-12}
eif5a2	9.7×10^{-10}	fgf1	1.4×10^{-10}	folr2	2.9×10^{-16}
hif1a	6.5×10^{-13}	hras	5.2×10^{-7}	itgav	2.0×10^{-10}
kit	2.4×10^{-7}	map2k1	2.6×10^{-16}	map2k2	1.4×10^{-13}
map3k1	1.6×10^{-8}	mapk14	2.2×10^{-27}	mapk6	9.4×10^{-11}
mapk9	6.2×10^{-13}	nfkb2	4.9×10^{-9}	pdgfb	9.7×10^{-11}
pdpk1	1.8×10^{-19}	pik3r2	5.5×10^{-8}	plagl1	6.9×10^{-7}
rheb	3.0×10^{-15}	rps6ka1	7.9×10^{-9}	smad1	6.1×10^{-9}
smad2	1.7×10^{-8}	smad3	1.2×10^{-8}	smad4	1.5×10^{-12}
smad7	1.2×10^{-15}	tgfbr2	1.8×10^{-15}	tsc1	5.3×10^{-15}
tsc2	6.2×10^{-18}	arid1a	9.0×10^{-15}	cbfb	1.5×10^{-11}
kmt2c	6.0×10^{-19}	fn1	2.3×10^{-9}	map4	9.8×10^{-12}
afdn	2.0×10^{-11}	arid5b	3.2×10^{-11}	bap1	4.2×10^{-15}
birc6	3.9×10^{-8}	chd1	2.6×10^{-10}	kdm6a	1.3×10^{-9}
lama2	8.8×10^{-7}	ldlrap1	1.0×10^{-7}	ncoa3	2.9×10^{-8}
nf2	7.5×10^{-8}	nras	4.2×10^{-8}	prkcz	7.5×10^{-9}
setd1a	3.1×10^{-12}	setdb1	8.6×10^{-10}	sf3b1	6.3×10^{-36}
siah1	1.8×10^{-12}	sik1	7.0×10^{-10}	ubr5	5.8×10^{-7}
akr1c4	6.6×10^{-7}	cdk8	1.7×10^{-12}	cdkn2c	6.8×10^{-15}
cyp21a2	2.4×10^{-8}	hsd17b11	4.7×10^{-28}	hsd17b12	2.7×10^{-11}
hsd17b6	3.1×10^{-10}	hsd17b7	3.8×10^{-9}	hsd3b7	6.6×10^{-20}
ncoa2	3.4×10^{-8}	ran	5.7×10^{-22}	sdc4	1.7×10^{-8}
tnk2	8.9×10^{-19}				

Source: Significant genetic attributes and corresponding p-Values of Univariate Cox Regression.

Table 2. Multivariate Cox Regression Analysis for Prognostic mRNAs

Gene	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%
hsd17b11	0.19	1.21	0.07	0.06	0.33	1.06	1.39
cdkn2c	0.08	1.08	0.05	-0.02	0.17	0.98	1.19
jak1	0.25	1.29	0.05	0.15	0.35	1.17	1.42
gsk3b	0.13	1.13	0.06	0.01	0.24	1.01	1.27
spry2	0.07	1.07	0.05	-0.03	0.18	0.97	1.19
lama2	-0.05	0.95	0.06	-0.17	0.06	0.85	1.07
kmt2c	-0.20	0.82	0.06	-0.31	-0.08	0.73	0.92
casp8	0.09	1.10	0.04	0.00	0.18	1.00	1.20
tgfbr2	-0.02	0.98	0.08	-0.18	0.15	0.83	1.16
map4	0.08	1.08	0.06	-0.04	0.19	0.96	1.21
abcb1	-0.05	0.95	0.05	-0.15	0.06	0.86	1.06
kit	0.07	1.07	0.05	-0.04	0.17	0.96	1.18
tsc2	-0.07	0.93	0.05	-0.18	0.03	0.84	1.03
pdgfra	-0.17	0.84	0.07	-0.30	-0.04	0.74	0.96
igf1	-0.05	0.95	0.06	-0.16	0.06	0.85	1.06
tnk2	-0.11	0.90	0.05	-0.20	-0.02	0.82	0.98
myc	0.16	1.17	0.05	0.06	0.25	1.06	1.29
stat5a	0.03	1.03	0.04	-0.05	0.11	0.95	1.12
smad4	0.05	1.05	0.05	-0.05	0.15	0.96	1.16
ccnd2	0.07	1.08	0.05	-0.03	0.18	0.97	1.20
rps6	-0.09	0.92	0.05	-0.18	0.01	0.83	1.01
pdgfb	-0.02	0.98	0.05	-0.11	0.08	0.89	1.08
jak2	-0.05	0.95	0.04	-0.14	0.03	0.87	1.03
rheb	-0.02	0.98	0.05	-0.12	0.08	0.89	1.09
ncoa3	-0.04	0.96	0.05	-0.13	0.05	0.88	1.05
akt1	0.07	1.07	0.05	-0.02	0.16	0.98	1.17

Source: Summary of Multivariate Cox Regression.