Astronomy & Astrophysics manuscript no. output
January 2, 2026

The miRNA Signature: Unveiling Breas Cancer Survival through High-Dimensional Cox Regression.

Matteo Bulgarelli, Giorgia Bertacchini

Università degli Studi di Modena e Reggio Emilia

January 2, 2026

ABSTRACT

Scrivere qui l'abstract

1. Introduction

Accurate survival prediction remains a cornerstone of precision oncology, providing the essential framework for personalized treatment strategies and clinical decision-making. In recent years, the integration of high-throughput sequencing with traditional clinical metadata has revolutionized our understanding of disease progression. While demographic and pathological variables—such as age, stage, and grade—offer a baseline for prognosis, they often fail to capture the underlying biological heterogeneity that dictates long-term patient outcomes.

The current State-of-the-Art (SOTA) in genomic survival analysis has largely been defined by the use of messenger RNA (mRNA) expression profiles. Though still a relatively evolving field, mRNA-based prognostic signatures have demonstrated significant success in stratifying patients into distinct risk groups, offering a more granular view of the molecular landscape than clinical data alone. However, the potential of other non-coding RNA species remains comparatively under-explored. MicroRNAs (miRNAs)—small, non-coding RNA molecules that post-transcriptionally regulate gene expression—present a compelling alternative. Due to their inherent stability in clinical samples and their role as master regulators of oncogenic pathways, miRNA sequencing (miRNA-Seq) data may hold untapped prognostic value that rivals or complements established mRNA-based models.

In this study, we investigate whether miRNA-Seq data, when integrated with standard clinical metadata, can achieve a level of predictive accuracy comparable to or exceeding current mRNA-centric approaches. To rigorously test this hypothesis, we implemented a graduated methodology consisting of three distinct analytical frameworks of increasing complexity.

2. Materials and Methods

2.1. Data Acquisition

The miRNA, mRNA and clinical data was all pulled from the GDC (Genomic Data Commons) Data Portal. We defined our cohort taking TCGA-BRCA for ductal and lobular neoplasms disease, with first diagnosis infiltrating duct carcinoma originating from breast. We then defined the filters:

- For miRNA we choose "miRNA-seq" as experimental strategy, "transcriptome profiling" as data category, "miRNA expression quantification" as data type, "open" access data, "tumor" as tissue type, and "primary" as tumor description
- For the clinical data we choose "clinical" as data category, "clincal supplement" for data type, and again with "open" access.
- For mRNA we defined the same filters as for the case of miRNA changing only "mRNA sequencing" for data category

The clinical files we used are XMLs, each one for a single patient, from which we extracted as features: age at initial diagnosis, vital-status (Dead/Alive), tumor stage, and follow-up data in terms of days to last follow-up in case or days to death. There were also some XML files of type OMF and txt files that we didn't include, and so we ended up with a total of 771 useful clinical files of mixed dead and alive patients. The reason we did not include OMF (Other Malignancy Form) is that they keep clinical information of our interest, and do not contain laboratory or transcriptomic data, such as mRNA or RNA-Seq.

For the miRNA data we took only the files ending with "mirnas.quantification.txt", because the remaining files were mainly logs, and we used as features: folder name, file name, read count and reads per million miRNA mapped. We ended up with 767 useful files, each with 1881 genes reads.

Finally for mRNA we extracted the files ending with "rna_seq.augmented_star_gene_counts.tsv", removing, as for the miRNA, the other log files, and choose as features for only the protein coding genes: gene name, gene id, unstranded, fpkm unstranded and tpm unstranded. We ended up with 787 files, each with around ~20k genes reads.

To build the 2 raw datasets we then joined the clinical data both with the miRNA and mRNA based on the folder name, which is referred to the case ID.

2.2. Data Analysis

For Data analysis we focused particularly on clinical data, where we saw that while tissues were sampled from different sites/hospitals, they were all sent to the same institute for analysis, which is the Nationalwide Children's Hospital, and that there is quite a big gap between alive and dead patients, since we have only 75 dead patients and 696 alive[1]. We also investigated the

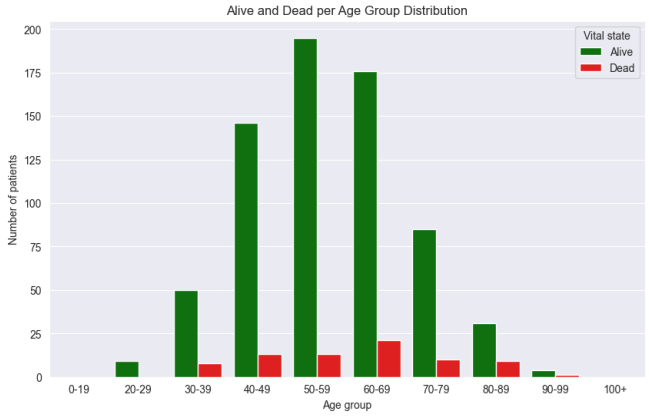

Fig. 1: Plot showing the ration between dead and alive patients grouped by age buckets of 10 years each

stages of cancer of the patients we worked with to try to understand what type of survivability results we may be expecting (See appendix A.1, A.2), and since we were interested mainly in dead patients and people with high days to last follow-up, we plotted also those distributions to better visualize the data (See appendix A.3, A.4).

2.3. Data Pre-processing

In the first place we performed one-hot encoding on the categorical cancer stage column, then we removed the outlier for what concerns the age, removing those patients that fall inside a scarcely represented range, picked as patients only those that are dead or have a 'days to last follow-up', where the threshold is set to the 25th percentile of the feature value: with these first steps we drastically reduced the samples of our datasets to around 300 samples both for miRNA and mRNA.

We proceeded to apply different normalization techniques based on the type of data, miRNA or mRNA, in order to produce different datasets to try to see if different normalization techniques improved the results achieved by the models. In particular for miRNA we made 2 datasets, the first obtained by normalizing the genetic features with just a log_2, and the second by normalizing them with quantile normalization (Bolstad et al. (2003)). For mRNA also we made 2 distinct datasets obtained by normalizing the genetic columns, one again by simply applying a log_2 and the other one by taking the already provided TPM normalized reads. In this way we produced 4 different datasets to test and confront the models on.

The last step before effectively saving the CSV files was to remove from each a first set of columns that had variance very low, using as threshold the 50th percentile variance value of the entire dataset, to perform a first simple filtering on features and reduce them. Since the very high dimensionality of the genetic columns in each dataset, especially the mRNA ones, we took a pretty high threshold to possibly light the job of the models that would process them, in order for them to concentrate on a smaller set of features.

Before passing each dataset through each model, genetic features and age at initial diagnosis of patients were also scaled with a Standard Scaler to achieve mean 0 and unitary standard deviation.

2.4. The 3 models framework

To investigate the prognostic utility of miRNA sequencing data compared to mRNA, we developed a multi-stage analytical framework consisting of three models of increasing architectural complexity. Each model was independently trained and evaluated on both data modalities to assess whether miRNA-based survival analysis achieves comparable or superior performance to established mRNA-based methods.

Penalized Cox Regression with Kernel PCA: Our baseline approach utilized an initial ElasticNet Cox model (Simon et al. (2011)) for feature selection, followed by Kernel Principal Component Analysis (KPCA) to handle non-linear structures during dimensionality reduction. The resulting components were then integrated into a final ElasticNet-penalized Cox model to predict patient survival outcomes. We choose to use an ElasticNet instead of a standard Lasso or Ridge penalty to overcome the 2 downsides these method have, which are: not being able to select more features than the number of samples of the dataset, and in case of Lasso, with a subset of highly correlated features, it would have randomly chosen one among the set. The ElasticNet overcomes these problems by combining the 2 penalties together by solving:

$$\arg\max_{\beta} \quad \log PL(\beta) - \alpha \left(r \sum_{j=1}^{p} |\beta_j| + \frac{1-r}{2} \sum_{j=1}^{p} \beta_j^2 \right)$$

Exploiting in this way the subset selection of Lasso, and the regularization strength of Ridge. In this first approach, after the appropriate part of data is scaled, only the genetic features are used first with an ElasticNet to perform a first step of feature selection by choosing the best penalization coefficient α: we propose 2 candidate values of alpha, one computed directly from the Randomized Cross Validation (**RCV alpha**) search we use to investigate values inside a certain range, and another one selected by a custom filter we defined (**filtered alpha**) that searches for the best alpha among the ones tested, that specifically keeps a number of active columns between 100 and 200. With the 2 alphas being selected we get the list of features with non 0 coefficient, and so the dataset is reduced to these features, which are then passed through the KPCA to further reduce them to 50 components. These components are then merged back with the clinical data, to form the actual set of data that is finally given to another ElasticNet to make the survival analysis predictions.

Linear PCA and Deep Learning: In the second configuration, we employed standard linear Principal Component Analysis (PCA) as an initial dimensionality reduction step. PCA performs a linear orthogonal projection that preserves maximal variance in the data, yielding a computationally efficient and interpretable low-dimensional representation. These reduced features were subsequently used as input to two DeepSurv architectures (Katzman et al. (2018)), featuring 3-layer and 5-layer network configurations, respectively, enabling the modeling of complex non-linear relationships between genomic features and patient mortality beyond the capacity of linear survival models.

DeepSurv is a neural extension of the classical Cox proportional hazards model in which the linear risk function is replaced by a feed-forward neural network. Specifically, the network learns a non-linear risk score $f_\theta(\mathbf{x})$ from the input covariates $\mathbf{x}$, while preserving the proportional hazards assumption. Model parameters are optimized by minimizing the negative log

partial likelihood of the Cox model, defined as

$$\mathcal{L}_{\text{Cox}}(\theta) = -\sum_{i:\delta_i=1}\left(f_\theta(\mathbf{x}_i) - \log\sum_{j\in\mathcal{R}_i}\exp\left(f_\theta(\mathbf{x}_j)\right)\right),$$

where δ_i denotes the event indicator and $\mathcal{R}_i$ is the risk set at time t_i. This formulation naturally accommodates right-censored survival data and yields relative risk estimates directly comparable to those obtained from traditional Cox regression, while allowing for the learning of non-linear covariate effects and higher-order feature interactions.

Variational Autoencoder (CoxVAE) and DeepSurv: In the final configuration, linear dimensionality reduction was replaced by a non-linear generative approach. We implemented a Cox-regularized variational autoencoder (CoxVAE) to learn a probabilistic latent representation by optimizing a variational lower bound while incorporating survival-related constraints. In contrast to PCA, which captures only linear correlations, the variational autoencoder models complex non-linear dependencies in the genomic feature space and enforces a structured latent distribution. After convergence, the encoder was frozen and used to extract low-dimensional embeddings, which were subsequently provided as input to the 3-layer and 5-layer DeepSurv networks for final survival estimation.

2.5. Evaluation

The main metric used to evaluate the models was the Concordance index (Harrell et al. (1982)) which is a rank-based metric that measures a model's ability to correctly stratify patients by their predicted risk of an event. In survival analysis, it represents the probability that a patient with a higher predicted risk will experience the event sooner than one with a lower risk, effectively handling right-censored data. For this study, the C-index serves as the primary validation tool to determine if miRNA-based features can achieve a discriminatory power comparable to mRNA. By focusing on relative risk ranking rather than absolute time-to-event, it provides a robust benchmark for evaluating the prognostic signal across our three modeling architectures.

Another metric we used, this time only for the final validation of the models is the Brier Score. It's a metric used to evaluate the accuracy of probabilistic survival predictions at specific time points. Unlike the C-index, which focuses on ranking patients, the Brier Score measures the calibration of a model by calculating the mean squared difference between the predicted probability of survival and the actual status (1 if alive, 0 if deceased). When integrated over time, we get the Integrated Brier Score (IBS), which provides a global assessment of model performance across the entire study duration. A score of 0 represents a perfect prediction, while a score of 0.25 typically indicates a model that is no more informative than a random guess.

3. Results

For what concerns the Penalized Cox model together with the KernelPCA (see appendix table B.1 for numeric results), based on the results shown, we can draw several insights regarding the impact of data type (miRNA vs. mRNA), normalization methods, and the feature selection process (RCV alpha vs. filtered alpha).

- **mRNA vs. miRNA Performance**: In general, mRNA datasets outperform miRNA datasets in terms of predictive accuracy. The mRNA datasets (both log_2 and tpm) consistently achieve higher mean C-index values, reaching up to 0.923 and 0.987 (max) with filtered alpha. In contrast, miRNA datasets peak at a mean of 0.882. For IBS (Integrated Brier Score) lower values indicate better calibration and the mRNA models generally show lower IBS values (mean ≈ 0.074–0.086) compared to miRNA (mean ≈ 0.081–0.127), suggesting mRNA features provide a more reliable survival probability estimation. While mRNA datasets demonstrate superior overall performance, the high C-index achieved by miRNA (particularly in the log_2 filtered configuration) confirms that microRNA signatures remain robust and clinically relevant predictors of survivability, likely reflecting critical regulatory layers not fully captured by the transcriptome alone.
- **Impact of Feature Filtering (RCV alpha vs. filtered alpha)**: The "filtered alpha" approach leads to a significant and consistent improvement across almost all metrics and datasets. In the miRNA log_2 dataset, the mean C-index jumps from 0.686 to 0.882. A similar trend is visible in mRNA tpm, moving from 0.719 to 0.923. A further improvement can be seen in terms of stability, where the standard deviation (std) of the C-index typically decreases when moving from RCV to filtered alpha (e.g., in mRNA log_2 it drops from 0.063 to 0.040), indicating that the filtered feature set produces more robust and less volatile results.
- **Normalization Comparison (Within the same base)**: For miRNA, log_2 normalization is superior to Quantile normalization when using filtered alpha (C-index 0.882 vs 0.717). Interestingly, with RCV alpha, they perform similarly, but log_2 scales much better after filtering. Both mRNA normalizations perform exceptionally well. However, mRNA tpm achieves the highest overall mean C-index (0.923) and the lowest mean IBS (0.074) under the filtered alpha configuration, making it the top-performing setup in this study.
- **Error and Variance (std)**: The miRNA Quantile dataset appears to be the most "unstable" or difficult for the model to learn from when filtered, as it is the only case where the C-index standard deviation actually increases (from 0.087 to 0.123) and the mean performance remains relatively low. The mRNA log_2 dataset with filtered alpha shows the lowest variance in C-index (0.040), suggesting this combination is highly reproducible.

Results for the PCA+DeepSurv (see appendizx table C.1) architecture show again different outcomes that are not always completely coherent. As the complexity of the head increases from 3 to 5 layers, only a marginal improvement can be seen: the mRNA tpm dataset shows the most notable improvement, with the mean C-index rising from 0.706 to 0.753, but conversely for miRNA with quantile normalization there is a decrease in mean performance, suggesting maybe that a deeper model may lead only to overfitting or optimization difficulties.

Consistently with previous observations, mRNA datasets maintain a higher performance ceiling compared to miRNA, especially when the architecture is deeper, even though for IBS, mRNA actually exhibits slightly worse values, compared for example to miRNA with log_2.

Different normalization technique does not seem to particularly impact in the case of mRNA, while there's a considerable change in the case of miRNA between log_2 and quantile. A final consideration is that more depth does not always correspond to an improvement also in terms of stability.

Finally results for the VAE+DeepSurv architecture (see appendix table)

4. Discussion

5. Conclusions

A brief summary of the primary takeaway. Does miRNA work for survival analysis? Is the increased model complexity worth it for clinical use?

References

Bolstad, B., Irizarry, R., Åstrand, M., & Speed, T. 2003, Bioinformatics, 19, 185, doi: 10.1093/bioinformatics/19.2.185

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. 1982, JAMA, 247 18, 2543. https://api.semanticscholar.org/CorpusID:23344910

Katzman, J., Shaham, U., Cloninger, A., et al. 2018, BMC Medical Research Methodology, 18, doi: 10.1186/s12874-018-0482-1

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. 2011, Journal of Statistical Software, 39, doi: 10.18637/jss.v039.i05

Appendix A: Data Analysis plots

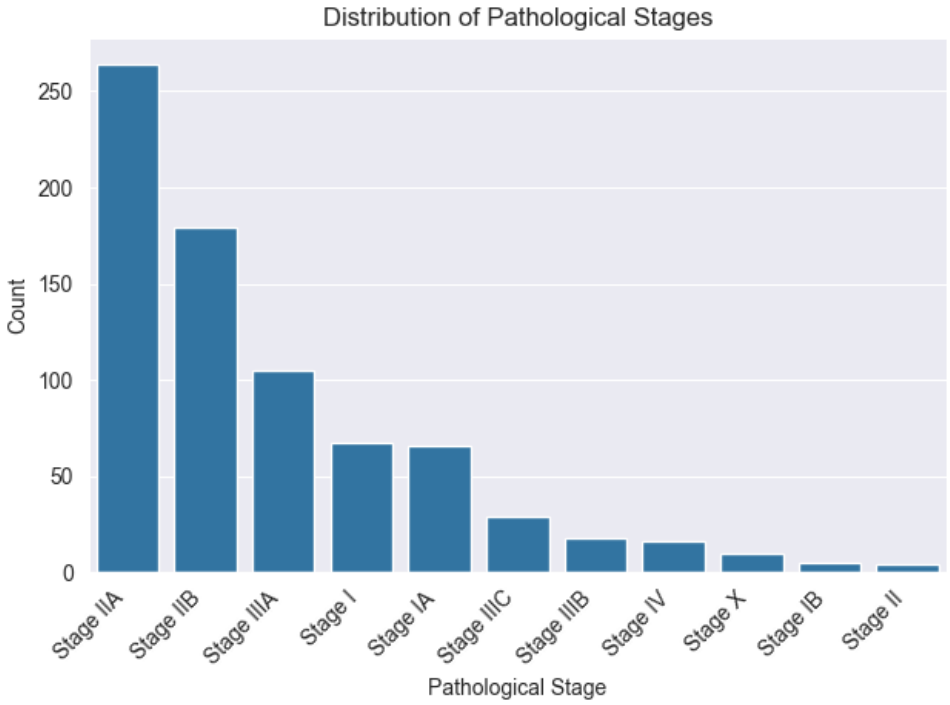

Fig. A.1: Distribution of cancer stages

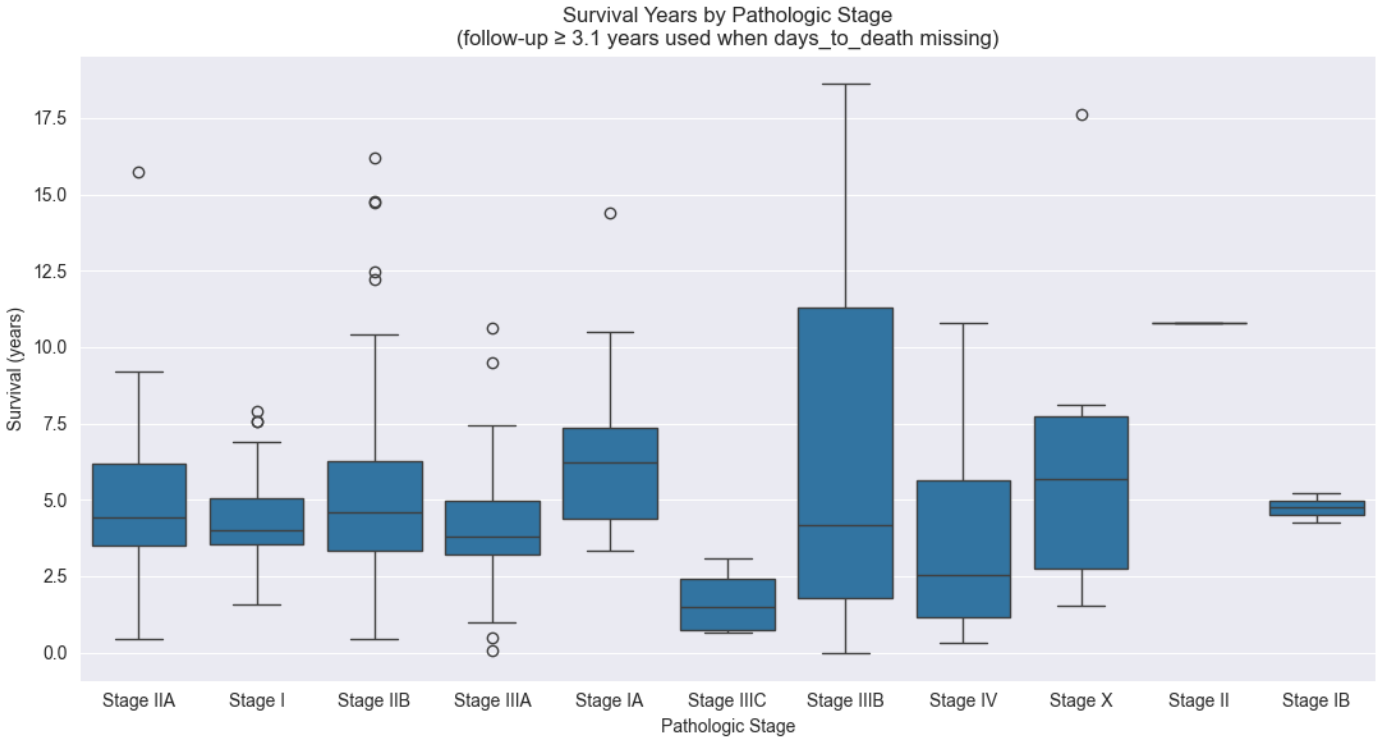

Fig. A.2: Survival time for dead patients and patients with high last days to follow-up (<3.1 years

Appendix D: VAE+DEepSurv results and plots

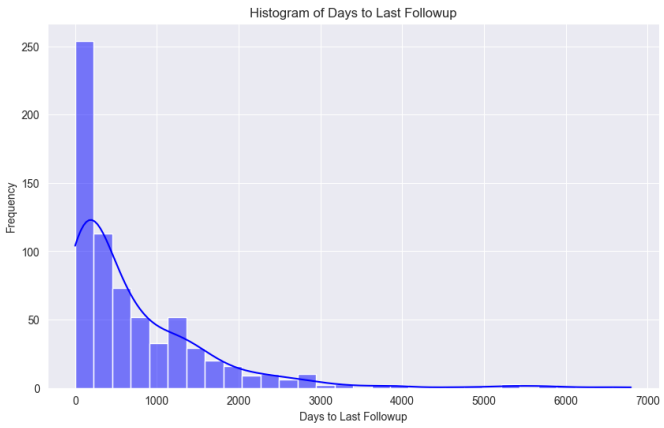

Fig. A.3: Distribution of days to last followup

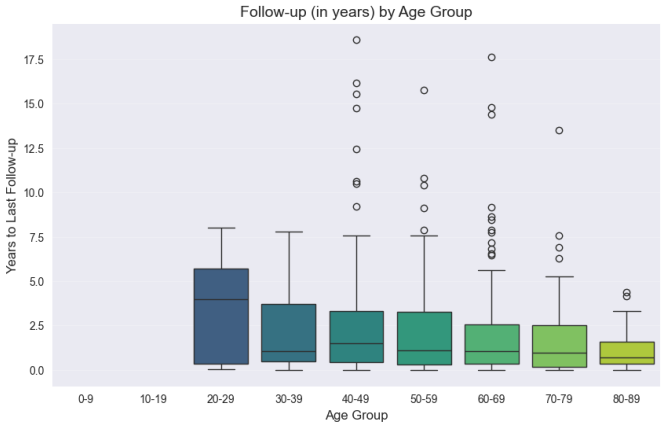

Fig. A.4: Years to last followup grouped by age in buckets of 10 years each

Appendix B: ElasticNet+KPCA tests and plots

Table B.1: Different datasets performances using ElasticNet and KernelPCA

		RCV alpha		filtered alpha	
Dataset	**Stat.**	**C-index**	**IBS**	**C-index**	**IBS**
miRNA log_2	mean	0.686	0.125	0.882	0.081
	std	0.123	0.044	0.066	0.031
	min	0.454	0.041	0.727	0.030
	max	0.904	0.183	0.958	0.139
miRNA quantile	mean	0.700	0.120	0.717	0.127
	std	0.087	0.036	0.123	0.038
	min	0.545	0.036	0.495	0.068
	max	0.841	0.179	0.865	0.186
mRNA log_2	mean	0.876	0.086	0.900	0.079
	std	0.063	0.026	0.040	0.022
	min	0.739	0.038	0.82	0.041
	max	0.961	0.120	0.961	0.111
mRNA tpm	mean	0.719	0.122	0.923	0.074
	std	0.090	0.029	0.042	0.020
	min	0.542	0.074	0.86	0.046
	max	0.843	0.178	0.987	0.111

Fig. B.1: miRNA log_2 plots

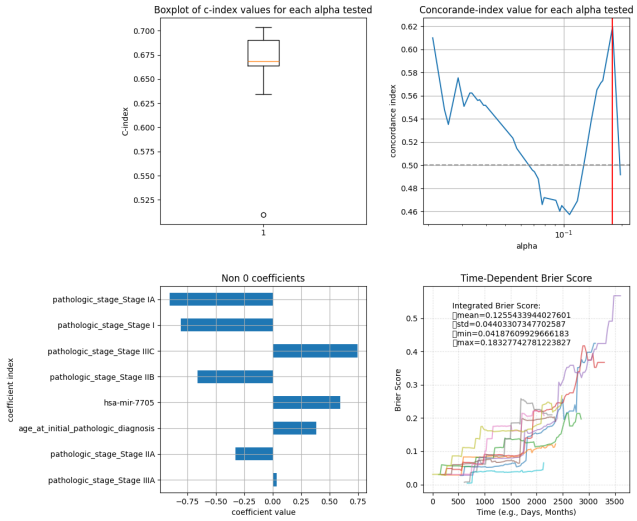

Fig. B.2: with RCV α

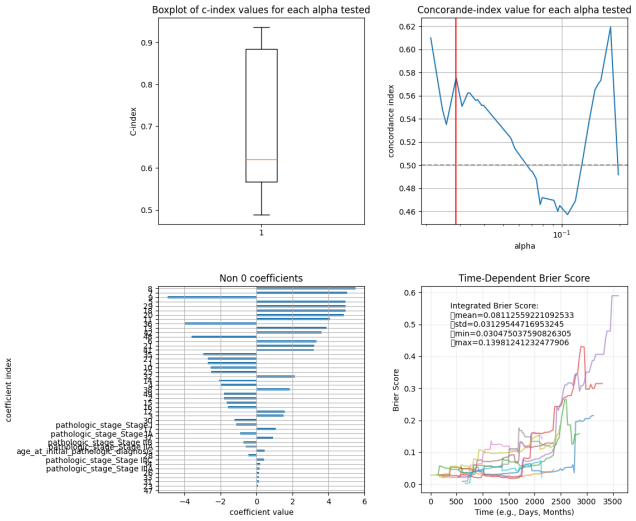

Fig. B.3: with filtered α

Fig. B.4: miRNA quantile plots

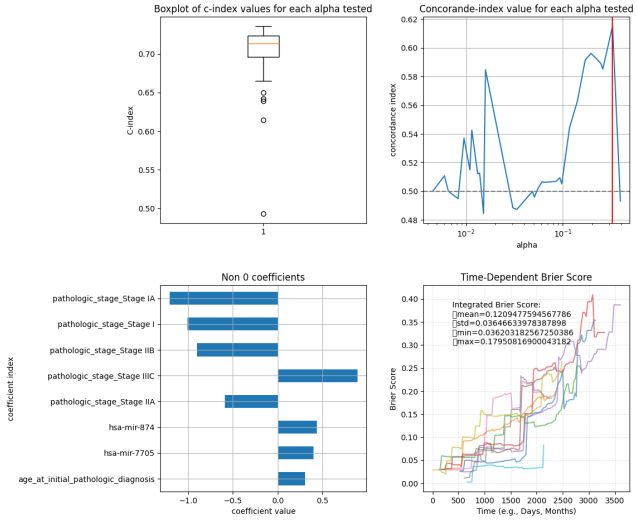

Fig. B.5: with RCV α

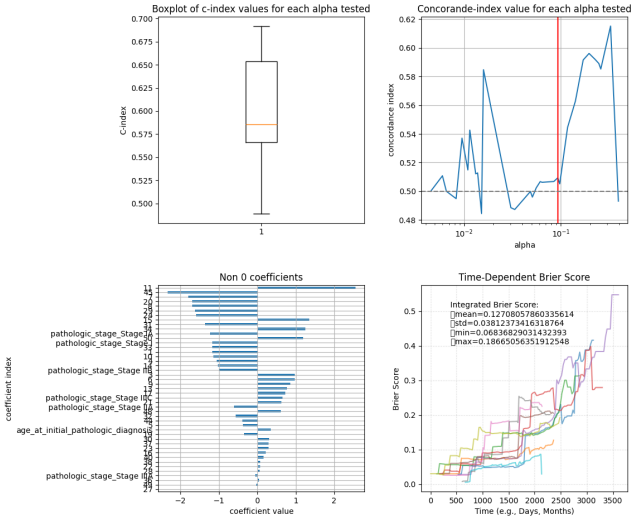

Fig. B.6: with filtered α

Fig. B.7: mRNA log_2 plots

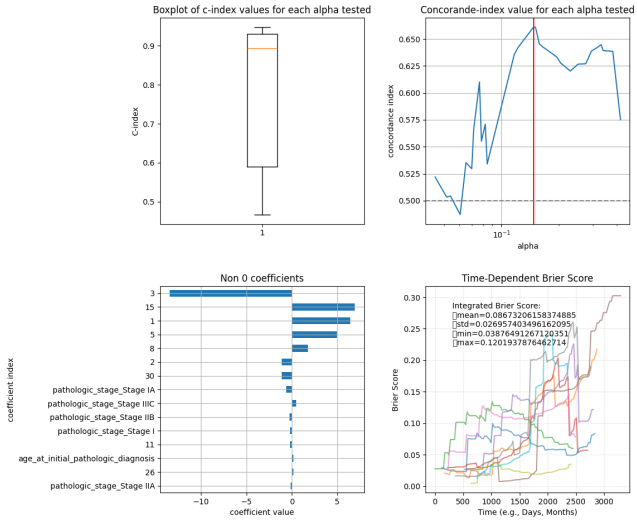

Fig. B.8: with RCV α

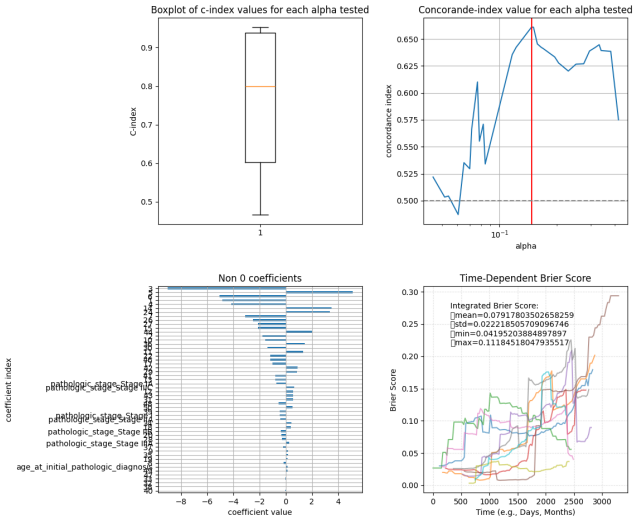

Fig. B.9: with filtered α

Fig. B.10: mRNA tpm plots

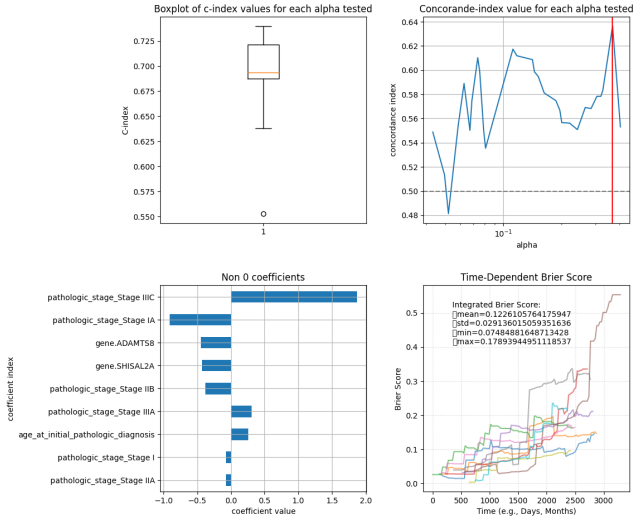

Fig. B.11: with RCV α

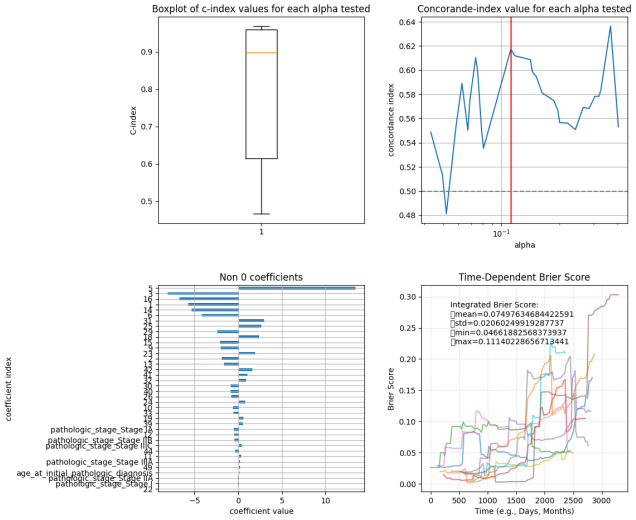

Fig. B.12: with filtered α

Appendix C: PCA+DeepSurv results and plots

Table C.1: Different datasets performances using linear PCA and DeepSurv networks

		3-layers		5-layers	
Dataset	**Stat.**	**C-index**	**IBS**	**C-index**	**IBS**
miRNA log_2	mean	0.679	0.130	0.695	0.130
	std	0.126	0.025	0.122	0.025
	min	0.511	0.098	0.500	0.094
	max	0.838	0.173	0.896	0.184
miRNA quantile	mean	0.656	0.142	0.612	0.135
	std	0.176	0.045	0.145	0.037
	min	0.375	0.074	0.429	0.059
	max	0.864	0.220	0.907	0.186
mRNA log_2	mean	0.684	0.155	0.701	0.147
	std	0.073	0.032	0.068	0.032
	min	0.585	0.104	0.602	0.125
	max	0.786	0.227	0.818	0.234
mRNA tpm	mean	0.706	0.158	0.753	0.151
	std	0.120	0.052	0.125	0.042
	min	0.468	0.117	0.550	0.102
	max	0.888	0.293	0.886	0.250