Survival Analysis of Lung Cancer Data for Non-Invasive Cancer Detection

Alan Wu
Mentor: Dr. Subhajyoti De

Introduction

Cancer has always been one of the leading causes of death around the world. It is crucial that we detect the signals for cancer or tumor development early on. For the current common method to detect cancer progression, it is usually a painful and time-consuming process for the patient. Current methods consist of computerized tomography (CT) scans, X-ray, and molecular testing for cancer-specific markers (tumor-antigen testing). These methods are often invasive (sticking needle through skin) and are unable to detect cancers throughout the entire body. They also have mediocre accuracy and poor predictive capabilities, leading to greater chance of overdiagnosis and overtreatment of non-malignant disease. Further, methods that target specific cancers are not conclusive of detection of cancers in other parts of the body. The detection is generally limited to certain surfaces of the skin and cannot be conclusive of all parts of organs. Also, doctors often cannot accurately decipher the signals of cancer through x-ray pictures, or other types of pixelated analysis. Those types of digital evidence could be misleading due to the blurriness and fragmentation.

Although these methods are proven useful to some degree, we need to continue research to find a better method to make the detection more accurate and less invasive.

With the current in-practice methods consisting of approaches that are generally invasive, current research prospects in genomics include utilizing blood samples to analyze certain DNA fragments through liquid biopsy sequencing. Liquid biopsy is a blood test that is used to detect cancerous cell DNA and tumor cells. This research has led to a focus on the presence and count of circulating tumor DNA (ctDNA) to track the efficacy of curative treatment and prevalence of cancer in the body. Cell free DNA (cfDNA) has also been a biomarker of interest because certain signatures in the cfDNA can provide information regarding the tissue of origin of cancerous cells. However, the current limitations of analyzing this data include lack of data and also other experimental effects. There is a lot of variability in the ctDNA count during different stages of cancer and different areas of the body. For instance, Colon, breast, pancreas, and liver have large amounts of ctDNA while glioma, thyroid, renal cancer have low malignancies from the ctDNA.

Computational approaches to early detection include applying simple machine learning models such as linear, logistic regression and random forest algorithms. Major concerns in analyzing cfDNA data include batch effects. Batch effects represent disparity in sampling methods for data and different experimental conditions. These varied conditions of data collection could lead to misleading conclusions, mainly mistaking actual results for batch effect.

Thus, while these computational methods to current problems do exist, many are still in development, at the discretion of available data.

An improved processing of liquid biopsy, DELFI (DNA evaluation of fragments for early interception) has arisen as a method that is more sensitive than traditional approaches and would be able to detect tumor signatures earlier by having greater sensitivity to aggressive changes in the DNA fragmentation pattern. And once patients are diagnosed, we are most often interested in the survival status and the survival time of the patient with respect to different treatments. Using this DELFI sequencing technique as a metric for survival status and timeline, we may be able to develop a better estimate for a patient's survival based on treatment factors and the stage of cancer that they are experiencing. Our approach by studying the effects of this new sequencing technique can give valuable insight into how DELFI compares as an estimator to traditionally invasive procedures such as CT scans and MRIs. Although it is possible that this DELFI score will not have good predictive power for the progression of cancer, we will still be able to measure the biological impacts of traditional metrics such as the stage/severity of cancer and the type of treatment that patients receive.

Related Work

DELFI (DNA evaluation of fragments for early interception) has been validated as a method for early detection of cancer. Researchers in [1] have expanded on liquid biopsy of cell-free DNA by attempting to identify the origin and molecular features of the cfDNA using machine learning techniques. This study compared the DNA fragmentation profiles of healthy individuals to those with varying types of cancer (breast, colorectal, lung, ovarian, pancreatic, gastric or bile duct cancer) and found that the fragmentation patterns of healthy individuals were similar to those derived from white-blood cells while those with cancer had altered fragmentation patterns. The study then used the collected fragmentation patterns to predict the presence of different types of cancer using a gradient tree-boosted model and yielded an overall AUC value of 0.94 across 7 types of cancers. These results give us underlying information that

DELFI is useful in predicting a potential presence of cancer, but still leaves the question of whether or not it is good for monitoring the progression of cancer. Below is the workflow of the DELFI methodology.

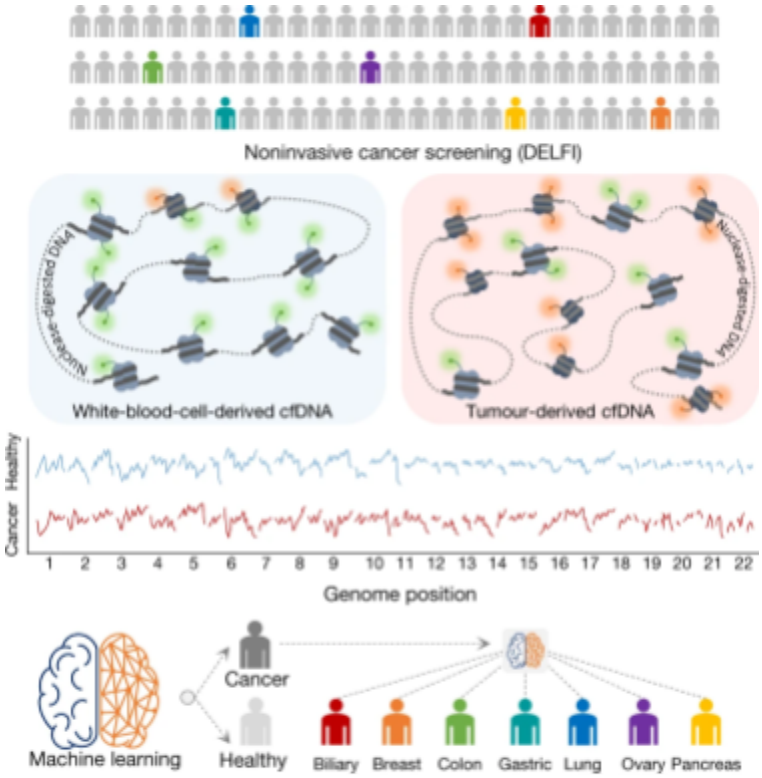

Figure 1: DELFI detection workflow

In [2], researchers applied the DELFI detection methodology to real-world lung cancer data. This study proved that DELFI yields good results on real-world data and compared to current methods of detection (CT scans and X-ray). They generated a DELFI score from DELFI detection that cross validated the results of whether or not the cfDNA profiles had characteristics of an individual with or without lung cancer which we will use in our analyses of this DELFI methodology. The data that we will use is also a subset taken from this study: n = 97 patients used to determine the relationship between prognosis and DELFI score. The DELFI scores in this subset ranged from 0.3-0.95 and included only patients that did indeed have some stage of lung cancer at the time of study. Given the nature of this data, we looked into survival analysis techniques to model the effect of DELFI score and survival time effectively.

Supplementary Data 6. DELFI score and survival analysis in the LUCAS cohort

Patient ID	Survival status	Days alive	Histology	Stage	1st line Oncological treatment	DELFI score	DELFI status
CGPLLU397P	1	1059	SCLC	I	Surgery	0.10	0
CGPLLU607P	1	1640	Adenocarcinoma	I	Surgery	0.53	1
CGPLLU637P	1	101	Squamous	IV	Palliative Chemotherapy/Radiation	0.82	1
CGPLLU538P	1	1228	Adenocarcinoma	IV	Palliative Chemotherapy/Radiation	0.24	0
CGPLLU349P	1	754	Adenocarcinoma	III	Chemotherapy/Radiation with curative intent	0.20	0
CGPLLU369P	1	542	Adenocarcinoma	IV	Chemotherapy/Radiation with curative intent	0.55	1
CGPLLU336P	1	723	Adenocarcinoma	III	Surgery	0.48	0
CGPLLU435P	0	2710	Adenocarcinoma	I	Surgery	0.31	0
CGPLLU398P	1	238	Adenocarcinoma	IV	Chemotherapy/Radiation with curative intent	0.63	1
CGPLLU629P	1	91	Adenocarcinoma	IV	Palliative Chemotherapy/Radiation	0.03	0
CGPLLU559P	1	283	Adenocarcinoma	IV	Surgery+adjuvant treatment	0.46	0
CGPLLU611P	1	245	Adenocarcinoma	IV	Surgery	0.16	0
CGPLLU314P	0	2766	Adenocarcinoma	IV	Chemotherapy/Radiation with curative intent	0.10	0
CGPLLU323P	0	2760	Squamous	I	Surgery	0.10	0
CGPLLU488P	1	660	Adenocarcinoma	III	Palliative Chemotherapy/Radiation	0.29	0
CGPLLU283P	1	425	Adenocarcinoma	IV	Palliative Chemotherapy/Radiation	0.75	1
CGPLLU373P	1	228	Squamous	III	Palliative Chemotherapy/Radiation	0.28	0
CGPLLU606P	1	1076	Adenocarcinoma	III	Palliative Chemotherapy/Radiation	0.48	0

Figure 2: LUCAS cohort lung cancer data

The data includes several features in which the survival status is 0, meaning the patient has not yet experienced the survival status event, or in this data, death. In this scenario, we cannot use regular analysis as there are a significant number of samples that have a survival status of 0. In these types of situations, we need to utilize survival analysis to analyze censored samples. From [3] these samples would be right-censored, as the timeline variable, survival time, is greater than the study time, or they had not experienced the status variable when the study had ended. The most traditional approach for this situation would be to estimate the hazard function or the survival function based on the data. There are several models in this field of survival analysis to estimate the probability of the status event (binary variable) occurring based on the time-to-event variable. Primarily they are used for preliminary analysis and to determine feature relationships with the status event variable. In our approach, we want to utilize these methods for preliminary analysis to determine feature significance and then for multivariate analysis to predict status event and time-to-event.

Methodology

Workflow

The approach to analyze the significance of DELFI score in the data includes a standard exploratory analysis of the dataset to determine any anomalies in features and also the general trend of cancer data. Then we conducted typical survival analysis such as fitting the data to the kaplan meier model, comparing categorical features using a log-rank test, and modeling initial cox proportional hazards models to measure feature significance when combined in one model. These initial methods all compare the relationship between the target variable (days alive) and the feature of interest. After initial feature

analysis, we will use several survival models to predict the time-to-event, in our event the survival_time variable based on the features in the data.

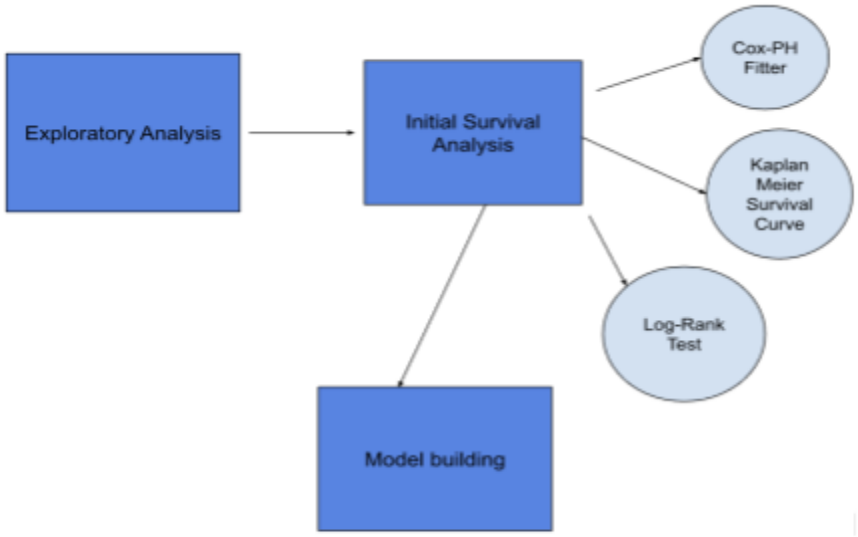

Figure 3: Full analysis worklflow

There are several methods in which the survival function (probability of survival vs timeline variable) can be predicted. In the survey [4], writers summarized several different methods. In selecting methods for our approaches, we wanted to survey a large range of methods. Thus, we started at traditional statistical algorithms such as the non-parametric Kaplan Meier model and extended that to regularized versions of the Cox Proportional Hazards model. To evaluate some more complex models, which may be able to capture the complexity of potentially correlated intermediate features, we chose the RSF (Random Survival Forest) model and also the Survival SVM (Support Vector Machine) model due to their ability to directly predict survival times given a training set of data with censored and uncensored samples.

Metrics in survival analysis also vary greatly. Results from [5] showed that there are varying metrics that work well for different types of outcomes for this type of analysis. In most situations, the concordance index (C-index) works well to measure the pairwise relationships between the binary event outcome (death event) and the time-to-event variable (days alive). Since we are attempting to predict the days alive variable as a measure of survival, another metric that has proven to be successful in many situations is the log weighted mean absolute error (MAE). The log weighted mae provides a traditional metric known for typical machine learning tasks while also adjusting it for censored samples with weighting. The logarithm of the MAE allows us to compare large differences in the absolute error of days alive predictions.

Model Building Workflow

To build and systematically test the different models, there are a few aspects to consider. First, we must preprocess the data to a format that will be ‘fair’ for all the models, meaning we need to have an even split of censored and uncensored data points, while also one-hot encoding multi-class categorical features. The most significant part of model building is tuning the hyperparameters such as the regularization rate for the models and also guaranteeing the split of censored to uncensored samples so that our models are trained fairly. Using a stratified k-fold on the time-to-event variable (death event) we can guarantee that there will always be a predetermined percentage of samples that are censored. This way, there can always be a valid c-index calculated from the specific split and will make our models train on relatively equal data.

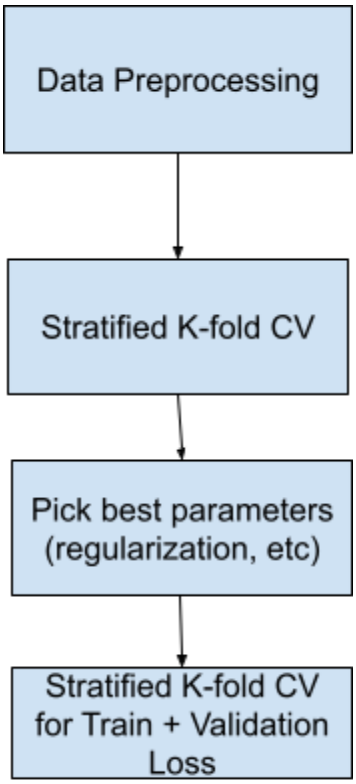

Figure 3: Model Building wowrkflow

Specifically, we use the models from the sksurv [6] including the ridge, LASSO, and elastic net regularized cox proportional hazards model and both survival SVM and RSF models. And we also use the metric calculations from [7] to evaluate each of our models for final cross validation scores. For the models that do not directly predict the days alive variable, we will take the median value of the calculated

population survival curve as the predicted survival time. This median value represents a 50% chance or toss up that the patient will die.

Feature Significance

To measure the feature significance of the models that have measurable coefficients (linear proportional cox models, accelerated failure time model), we can compare the magnitude and sign of the coefficients that contribute towards the final trained model. Coefficients that have greater magnitude (either positive or negative) have greater significance as they carry greater weight towards the final target result (predicted days alive).

Conclusion

Best Model

Model	Weighted MAE	C-Index	Log MAE
Kaplan Meier Fitter	769.704	0.500	1.378
Ridge Regularized Cox-PH	557.213	0.813	0.899
Lasso Regularized Cox-PH	581.493	0.825	0.910
ElasticNet Regularized Cox-PH	557.951	0.828	0.875
AFT	554.073	0.803	0.921
Survival SVM	483.828	0.831	0.838
RSF (Random Survival Forest)	592.314	0.819	0.902

Figure 4: Model CV validation scores

We discovered the best model was the survival SVM model with the linear kernel setting. While it is not surprising that the model with a relatively complex decision pattern was able to generalize well to unseen data, it is fairly surprising to see that the RSF performed the worst out of all the linear models. Given the small sample size (97), the RSF could have overfit the data and thus yielded worse results for the same training and test sets as the linear models did. These results demonstrate that the simple models generally could perform very well with just some regularization added to it. The best regularized linear model was the ElasticNet model, which yielded a weighted MAE of 557.951 and log weighted MAE of 0.899. These results are not so far from the best model that we determined. In context however, these models are not the best models that could have been obtained. This can also be attributed to the small sample size. In context, an MAE of ~483 days is very inconsistent for patients who could actually rely on these prediction times. To drop this value there could be a inclusion of more data and more time spent with tuning hyperparameters.

Feature Significance

In terms of the features, there are a few significant findings. The conclusions include the significance of DELFI in predicting the survival status and also the individual feature significance of the treatment given to the patient and the stages of cancer.

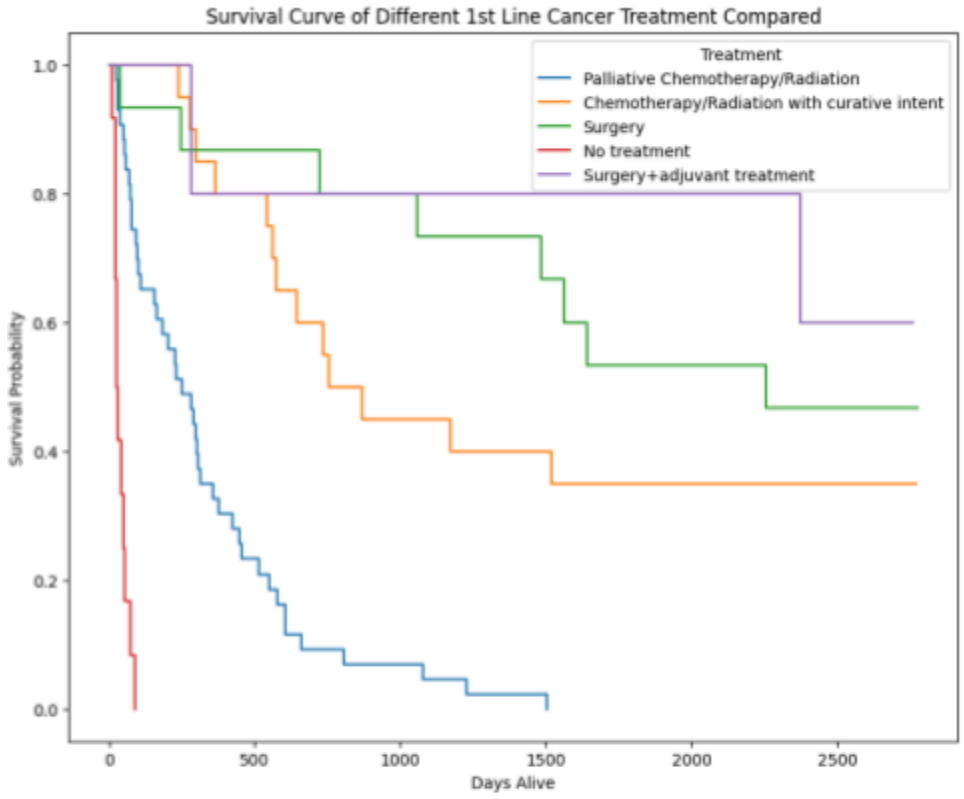

Figure 5: Survival curve of different treatments of cancer

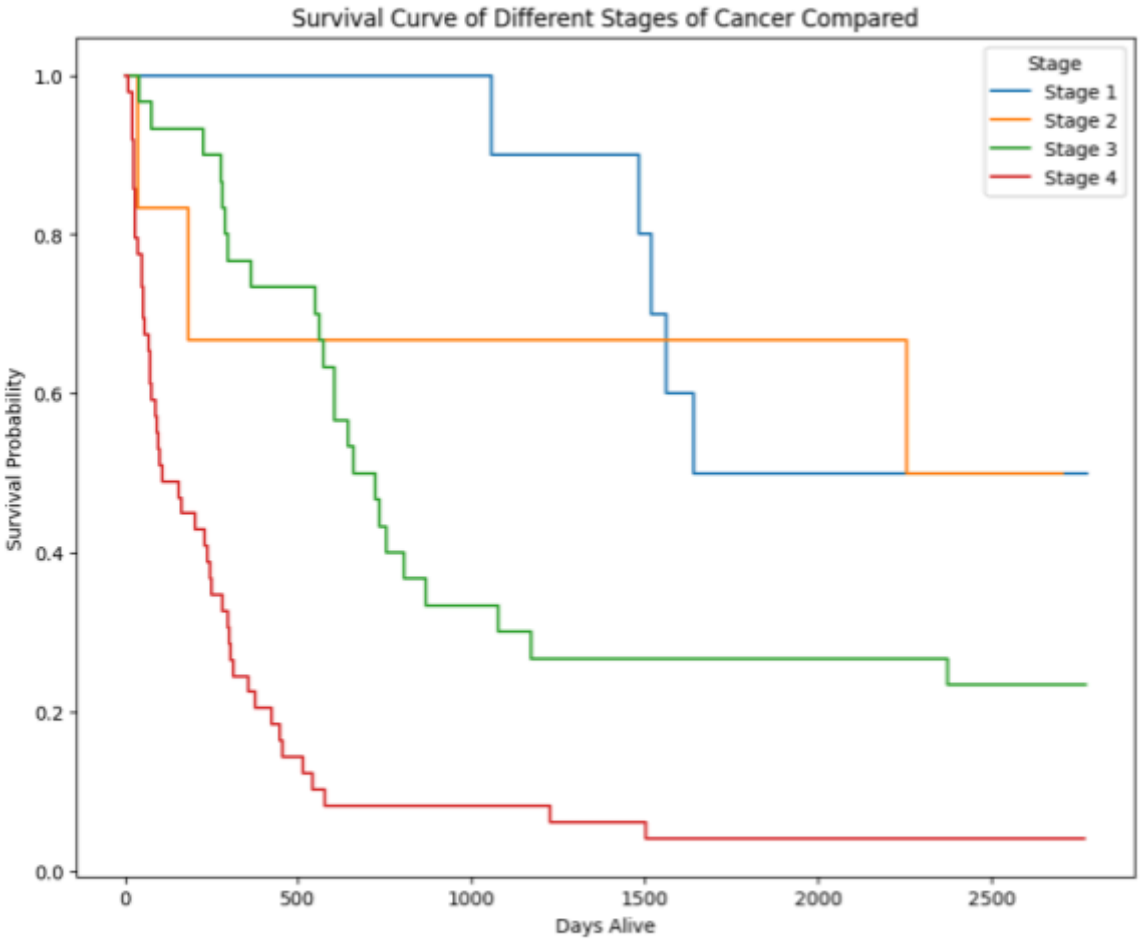

Figure 6: Survival curves of stage of cancer

For the treatment feature, it is evident that having no treatment has significant effects compared to any of the other features. The steep curve represents the severe dropoff in probability of survival in a short time frame, which is to be expected. Cancer patients that receive no treatment are expected to die quicker. Surprisingly, chemotherapy and radiation with curative intent performed the best, with the longest curve. The other treatments also performed much better than the no treatment category. However, between the other categories there seems to be some overlap, leading us to believe that there could be correlation between these other categories.

For the stage feature, it seems that it is consistent with what we already know about stage and cancers. Our original belief is that the stage of cancer is correlated with the survival time. The more aggressive the cancer, or the greater the stage, the shorter the patient will survive. This seems to be the case as well, as stage 4 curve has the steepest curve, and stage 1 having the longest curve, representing a higher survival probability as time goes on. It is interesting to note that stage 2 has correlation with the other stages, meaning it could be a tampering factor when fitting initial models.

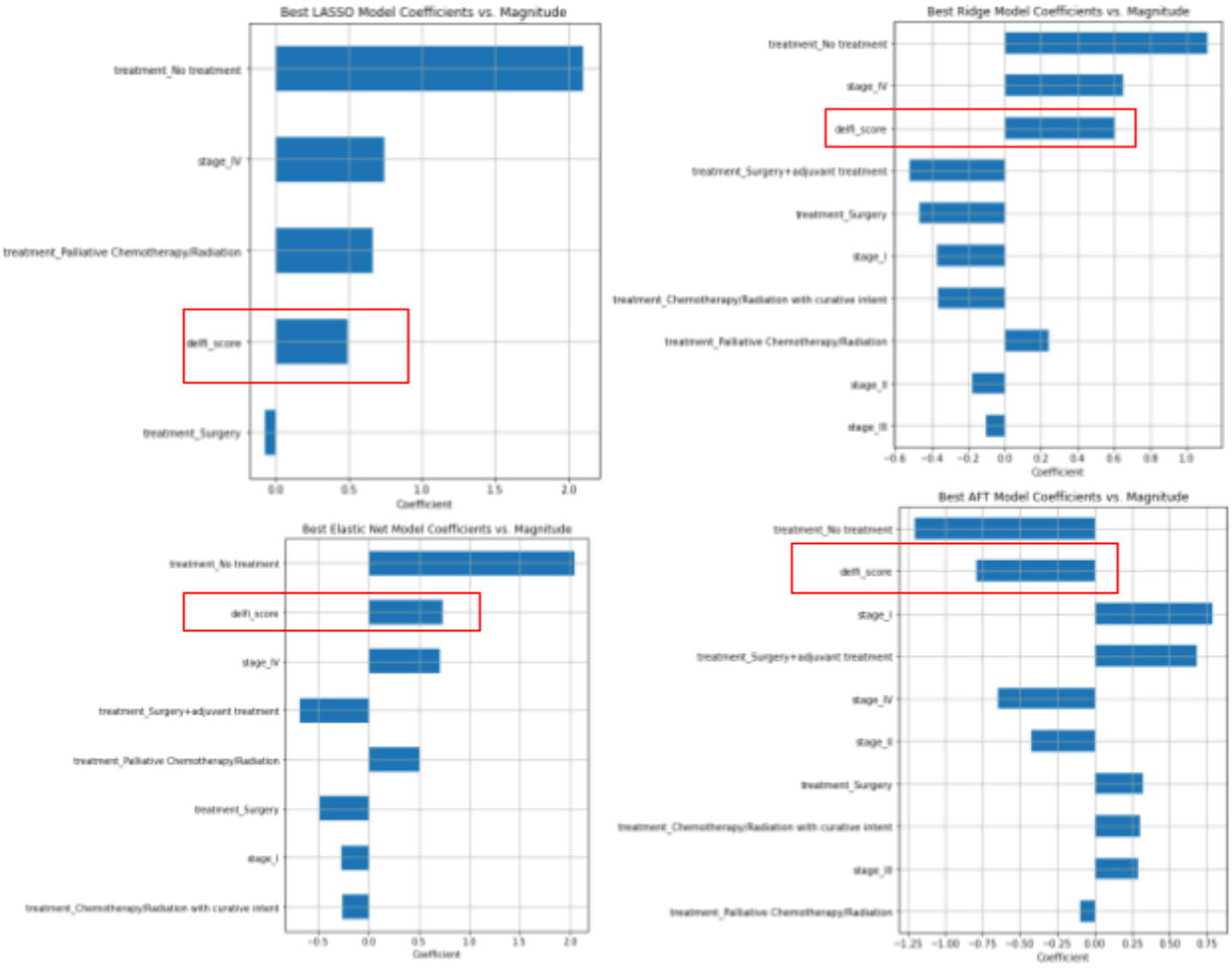

Figure 8: DELFI score significance of linear models. DELFI Magnitude, from left to right: 0.488, 0.603, 0.734, -0.783

When we take a look at the DEFLI score values for the different linear models, it is evident that the delfi_score variable is one of the largest magnitude variables, after no treatment variable in most cases.

The delfi_score variable is vastly significant in our analyses and was not regularized to 0 in even the lasso model, meaning it holds some significance in predicting the target survival time of the patient.

Biological Implications

Biologically, our conclusions are relatively consistent with what the medical industry already knows about cancers. In our scenario, we studied lung cancers. We confirmed existing knowledge that the stage of cancer is correlated with how long a patient lives. The more aggressive cancer, the shorter the patient lives. Furthermore, we confirmed biologically which treatments currently work best on patients. And this confirmed that chemotherapy, surgery, and radiation treatment tended to lengthen the survival curves of patients and thus gave them higher probability of survival after these first line oncological treatments. One major biological implication of the delfi score is that it is biologically significant in predicting the survival outcome and survival time of a lung cancer patient. It is evident that with greater delfi score, the patient will observe a shorter survival time. Overall, the delfi score is significant and can be used in future studies as a metric for whether or not a patient will survive or rather how long they will survive.

Bibliography

1. Cristiano, S., Leal, A., Phallen, J. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019). https://doi.org/10.1038/s41586-019-1272-6
2. Mathios, D., Johansen, J.S., Cristiano, S. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun 12, 5060 (2021). https://doi.org/10.1038/s41467-021-24994-w
3. Clark, T., Bradburn, M., Love, S. et al. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer 89, 232–238 (2003). https://doi.org/10.1038/sj.bjc.6601118
4. Wang, P., Li, Y., & Reddy, C. K. (2017). Machine Learning for Survival Analysis: A Survey. arXiv preprint arXiv:1708.04649.
5. Qi, S., Kumar, N., Farrokh, M., Sun, W., Kuan, L.-H., Ranganath, R., Henao, R., & Greiner, R. (2023). An Effective Meaningful Way to Evaluate Survival Models. https://arxiv.org/pdf/2306.01196
6. Pölsterl, S. (2020). scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research, 21, 1-5. Retrieved from https://www.jmlr.org/papers/volume21/20-729/20-729.pdf

7. Qi, Shi-Ang & Sun, Weijie & Greiner, Russ. (2024). SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. Proceedings of the AAAI Symposium Series. 2. 453-457. 10.1609/aaaiss.v2i1.27713.