Predicting survivability of high-grade serous ovarian cancer

Asha Mehta
Department of Biomedical Data Science

Anthony Tzen
Department of Computer Science

Abstract

Predicting the prognosis of high-grade serous ovarian cancer (HGSOC) patients continues to be a challenge, limiting treatment efficacy and optimal patient care. We used clinical and gene expression data from The Cancer Genome Atlas (TCGA) to predict survivability of HGSOC. A Cox proportional hazards model (c-index=0.58) and a random survival forests model (c-index=0.56) performed comparably when each used elastic net regularization to rank features. They both performed slightly better than the traditional baseline that used patient age and FIGO stage (c-index=0.53). These models indicate that gene expression data may hold more prognostic information than clinical data alone and might prove to be useful in the standard clinical setting.

Introduction

Approximately 1 in 78 women will develop ovarian cancer in her lifetime.[1] The most common, and the most deadly, form is high-grade serous ovarian cancer (HGSOC).[2] HGSOC is characterized by high clinical and genomic heterogeneity. Current gene expression biomarkers have only been tested in small samples and may not be robust.[3] Honing in on the correlation of specific gene expression patterns with survivability of HGSOC patients would enable more precise prognoses and more effective treatments. In this study, we measured the predictive power of genomic characteristics on survival time of high-grade serous ovarian cancer patients using existing machine learning methods. The input to our algorithm is transcriptome profiling data, where each feature corresponds with the biological activity of a known, significant gene. The output of our algorithm is the expected survival time of the patient. With this setup, we compared performance of two survival analysis methods, Cox regression and random survival forests, in predicting survival time of HGSOC patients.

Related Work

Currently, most ovarian cancer patients get the same treatment as no robust biomarkers have been found.[4] Because the average survival rate for HGSOC patients is, on average, 4 years with this treatment, though, scientists have begun tackling the problem of predicting survivability from multiple angles. Traditionally, only clinical data was used, as in a study examining survival outcomes of patients with both breast and ovarian cancer, using the Kaplan-Meier method for univariate analysis.[5] They found statistically different diagnosis times of the two cancers depending on the order in which the patient developed the conditions. However, the lack of genomic or other biological data limited their conclusions. More recently, though, genomic data has received substantial attention. One study used k-means clustering on gene expression data to divide HGSOC patients into three subtypes that were each run through ensemble classifiers and found that each subtype had different survival risks.[6] A second promising study applied a Cox proportional hazards model with elastic net regularization and 10-fold cross validation to gene expression data of 513 genes, resulting in a 101-gene signature for both 2- and 5-year overall survival.[7] Other studies have focused on introducing new types of data to elucidate the underlying biology of HGSOC. One study also used a Cox proportional hazards model, but applied it to computed-tomography (CT) images of pre-surgery tumors to predict 2-year overall survival. Another study reviewed the usefulness of circulating tumor DNA (ctDNA) in identifying molecular biomarkers.

Due to its widespread availability, genomic data is an ideal prognostic tool. However, most recent work using genomic data treats survivability as a discrete classification problem. This has produced promising results, but is not the most powerful use of this time-to-event data, both for biological discovery and patient care. Therefore, we chose to use survival analysis algorithms on genomic data, but treated survivability as a continuous variable.

Dataset and Features

For this study, we used the clinical and gene expression data of 373 patients diagnosed with ovarian cancer. The publicly available data was measured and collected as part of the Cancer Genome Atlas (TCGA) project.[8]

The gene expression data was available in the form of an mRNA expression profile per patient that includes the level of expression for each known, significant gene transcript in the patient's cells at the tumor site. Measurements were obtained by reading mRNA sequencing reads, identifying the reads via a standard mRNA analysis pipeline[9], and normalizing the read counts by the total number of reads and by the length of these reads. For this study, each gene's expression level, as measured via their normalized mRNA read counts, is a potential feature for our model.

Once this gene expression data was queried from the TCGA database, we further processed the data with feature standardization. In addition, because the vast majority of features had negligible variation across the dataset (i.e, features with zero value for all samples), variance thresholding was applied to select the 5% of features with the highest variances across the dataset. This yielded a basic starting dataset of 2900 gene expression features. This was still a significantly large number with respect to the 373 samples in the dataset, and thus we deployed a variety of additional model-specific feature selection techniques (see Methods section below).

Clinical data was used to characterize the outcome of each patient in our dataset. As is often the case with medical records, the data is partially censored; that is, the data only includes information about the patient from their diagnosis to the time of their last follow-up. Thus, the clinical data was processed to include two labels per patient: a binary event indicator indicating whether the patient was alive or dead, and a variable indicating the number of days between time of diagnosis and either the time of death or the time of their last follow-up. The patient's age and the FIGO stage of cancer at diagnosis was also extracted from the clinical dataset; these two factors acted as the dependent variables in our baseline experiments.

The same 70/30 train-test split was used for all experiments. Cox regression used five-fold cross-validation, while the random survival forest training set was further split into a 70/30 train-validation split for hyperparameter tuning.

Dataset	*Age at Index*	*FIGO stage*	*% Censored*
Training (n=196)	59.26 years (STD=10.81)	3 (STD=3.11)	35.7
Validation (n=66)	59.91 years (STD=11.79)	3 (STD=3.12)	36.3
Test (n=111)	59.59 years (STD=12.01)	3 (STD=3.04)	45.0

Methods

Background: Survival Data Analysis

One challenge in handling survival data such as the clinical dataset used in this study is that the data is partially censored, meaning it has incomplete follow-up information. This can occur because the patient dropped out of the study or because the event of concern did not occur during the study (in this case, the event of concern is a patient's death). A popular method for handling censored data is complete-case analysis, where only samples with no missing data are included in analyses.[10] However this makes the often incorrect assumption that missingness is uncorrelated with the event of interest. It also can decrease the sample size substantially.

In contrast to complete-case-analysis, time-to-event models are particularly suited to handling censored data in a sound way that reduces bias and preserves power. Generally, these models study the survival and hazard functions, and how they relate to each other and potential covariates[11]. The survival function $S(t)$ is the probability that the survival time is larger than a given time t, while the hazard function $h(t)$ is the probability of the event occurring at time t given that the event has not occurred yet.

$$S(t) = P(T > t) \qquad h(t|x) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t, x)}{\Delta t}$$

Cox Regression Models

The most popular of these models is Cox regression, which is a semi-parametric survival analysis method that relates survival time to multiple risk factors simultaneously. The model makes the proportional hazard assumption, modeling the hazard function as multiplicatively proportional to the dependent variables:

$$h(t|x) = h_0(t) \cdot \exp(\beta x)$$

where $h(t)$ is the expected hazard at time t and $h_0(t)$ is the baseline hazard (e.g. the hazard when all predictors are equal to zero), β is the model parameters, and x is the covariate vector.[12] The β parameter is estimated by maximizing the log partial likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \cdot \left(x^{(i)^T}\beta - \log\left(\sum_{j=1}^{n} \mathbf{1}\{y^{(j)} \ge y^{(i)}\} \cdot \exp\left(x^{(j)^T}\beta\right)\right)\right)$$

where $y^{(i)}$ is the time of death or censorship of sample i and where δ_i is equal to 0 if sample i is censored and 1 if not censored.

One drawback to the Cox regression model is that it uses a linear combination of the features to compute the features' multiplicative effect on the baseline hazard. Thus, to allow our model to learn non-linear relations, we implemented a Cox neural network with a single latent layer, setting its loss function to be the negative log partial likelihood.[13] In this way, the hazard function is modified:

$$h(t|x) = h_0(t) \cdot \exp\left(f_\theta(x)\right)$$

where $f_\theta(x)$ is the output of the latent layer of the neural network.

Another limitation of Cox regression is that it can become unreliable if the number of features is much larger than the number of samples. We thus used *sure-independence screening* (SIS) to further narrow down the feature set. The method ranks each feature by its marginal utility, which in our case is defined as the maximum partial likelihood of the single feature as modeled by Cox regression.[13]

Then, *penalization-based variable selection* was used to re-rank the top 400 features. This involved training a series of Cox regression models with increasing levels of elastic-net regularization; features whose coefficients remained nonzero in heavily regularized models are ranked higher in importance.[14,15] The process works best with a moderate number of features and so was used after SIS.[14]

Random Survival Forests

Random survival forests (RSF) is another survival analysis method that expands upon random forest (RF) methodology. In traditional RF, a bootstrap sample is used to "grow a tree." Then, a random set of covariates is chosen to split the tree. Repeating this process and averaging over individual tree's predictions can produce accurate and generalizable results. In RSF, the same approach is used, but splitting criteria uses both survival time and censoring status to maximize survival differences between the two daughter nodes.[16] We used the log-rank splitting rule in our experiments, which maximizes a standardized two-sample logrank test statistic.

Concordance Index

To evaluate predictors for our right-censored survival data, we used the concordance index (also known as Harell's C-statistic) as our evaluation metric. This metric performs pairwise comparison on each comparable pair of samples and computes the fraction of pairs that are ordered correctly relative to each other. This allows for evaluation of performance on both censored and uncensored samples. By evaluating the ordering of predicted survival times, this index more accurately assesses the performance of time-to-event models than other evaluation metrics.

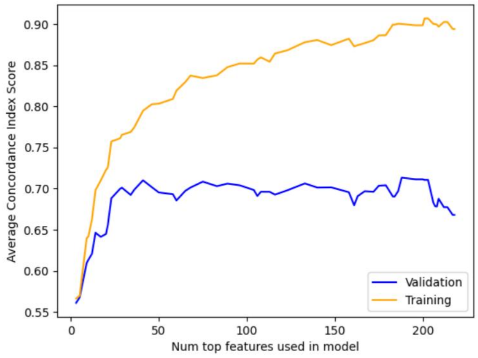
Experiments, Results, and Discussion

Cox Regression

Working under the Cox regression proportional hazards assumption, SIS was used to rank features. Then, penalization-based variable selection was used on the top 400 features to define candidate feature sets. For each set of top features, a model is trained and evaluated. Five-fold cross-validation was used to determine the best Elastic regularization factor to use, with the L1 ratio set to 0.9. The model with the best cross-validation score used the top 101 features and had a test concordance index of 0.572, which is comparable to published results[17] and is slightly better than our baseline.

We also repeated this process without SIS, using penalization-based variable selection on the 2900 features to see if this process selected more viable feature sets. The results were not significantly different, as the average cross-validation score plateaus at around 50 features. The simplest model with the best validation score (97 features) had a slightly better test concordance index of 0.583. Looking at features with the largest corresponding coefficients, we identified PSMB8 and ZC3H15 as two top predictive variables, which is aligned with what other studies of cancer prognosis have found.[18]

These Cox regression models tend to significantly overfit the training data, as the difference in validation and training scores quickly diverges as the number of selected features increases (Figure 1). The large gap between the validation and test scores also indicate significant bias towards the validation data, a bias that was likely introduced during feature selection since the feature selection process used both the training and validation data.



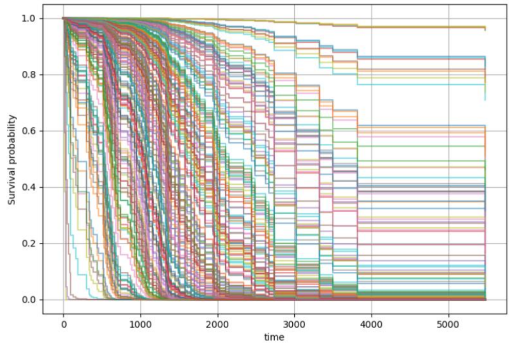


Figure 1 (left): Concordance index scores of each model trained in the penalized-based feature selection process.
Figure 2 (right): The survival functions for each test sample as predicted by our top-performing Cox regression model.

Cox Neural Network

We then experimented with a Cox neural network. The latent layer's nodes used tanh activation functions, as this was most successfully used by related studies.[17,19] The model used an Adam optimizer and stopped early if loss did not improve after 20 epochs.

Neural networks have more predictive power since they can learn more complex, non-linear relations among the feature variables. These models are thus more prone to overfitting; to mitigate this, the latent layer's kernel weights were regularized with L1 and L2 factors (elastic-net regularization). In addition, Gaussian noise (std=0.01) was added to the input, and dropout was used on the connections between the latent layer and the output node. Hyperparameters to configure thus included the number of nodes in the latent layer, L1 & L2 regularization factors, dropout rate and learning rate. Grid search was conducted to find the optimal hyperparameters. Each trial was evaluated with multiple runs of 3-fold cross-validation, as we found that cross-validations with more than 3 folds resulted in wildly varying results as the validation splits became too small to be representative of the entire dataset.

However, even with these regularization techniques, the neural network consistently overfits to the training data when using all 2900 features, with the best model's test concordance score being 0.498. We thus selected features via the feature selection processes described previously. This partially reintroduces the Cox proportional hazards assumption, but significantly improves the model performance. The best model used the 55 features and had a test concordance index of 0.580 (averaged over 10 trials). The selected hyperparameters were 55 latent nodes, L1 & L2 factors of 0.001 and 0 respectively, a dropout rate of 0.1, and a learning rate of 0.001.

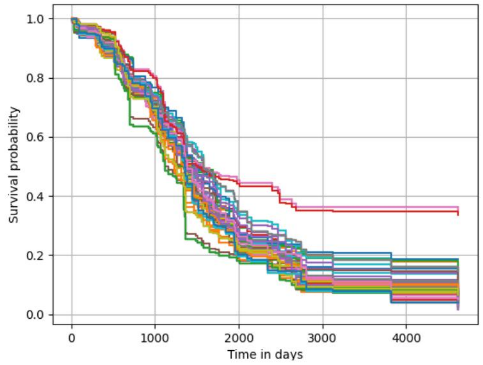
Interestingly, even when using the same features, the neural network model did not perform better than Cox regression; thus, it appears that there are no strong predictive factors that would otherwise be missed by the linear Cox regression model. Still, this model was able to perform at a comparable level while using fewer features than the best-performing Cox regression model.

Random Survival Forests

An RSF model was trained using only patient age and FIGO stage to produce a baseline concordance score of 0.534. The model was then trained with three different feature ranking methods. RSF models are especially sensitive to large datasets so more stringent feature ranking was required than for Cox regression. First, the variance thresholding described previously was used with a threshold set to include only the top 3% of features for a total of 1740 features, resulting in a concordance score of 0.482. Next, penalized-based variable selection identified the top 10 features from the top 5% of features selected by variance thresholding, resulting in a concordance score of 0.535. Finally, 5 genes were manually selected based on their high prognostic value reported by Millstein et al.,[20] producing a concordance score of 0.500.

Due to the large hyperparameter set size required for RSF, randomized search was used to identify optimal hyperparameters for each experiment, specifically the number of trees in the forest, the maximum depth of the tree, the maximum number of features to consider when looking for a split, the minimum number of samples required in a leaf, and the minimum number of samples required to split an internal node. The highest performing feature ranking experiment set the number of trees to 12, the maximum tree depth to 19.150, the maximum number of features to consider to 0.200, the minimum samples per leaf to 4, and the minimum samples to split to 0.485. The random number generator seed was manually set to 44.

When compared to the RSF baseline, only penalized-based variable selection produced comparable results. Both the baseline and penalized-based variable selection survival curves show no strong patterns in survival time (Figure 3 and 4). That being said, *TAP1* was identified as the primary gene contributing to survival time prediction here, which is in line with Millstein et al.'s results and points to this gene as a prognostic biomarker for HGSOC.



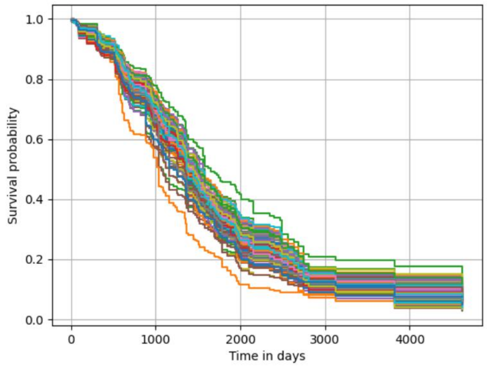


Figure 3 (left): The survival functions for each test sample as predicted by the baseline RSF model.
Figure 4 (right): The survival functions for each test sample as predicted by our top-performing RSF model.

Model	*Feature Selection Process*	*Number of Features*	*Concordance Index Score*
Cox Regression	Clinical Baseline (Age + Stage)	2	0.572
	Variance Thresholding (top 5%) + Penalized-Based Selection	97	**0.583**
	Variance Thresholding (top 5%) + SIS + Penalized-Based Selection	101	0.572
Cox Neural Network	Variance Thresholding Only (top 5%)	2900	0.498
	Variance Thresholding (top 5%) + Penalized-Based Selection	55	**0.580**
	Variance Thresholding (top 5%) + SIS + Penalized-Based Selection	65	0.563
Random Survival Forests	Clinical Baseline (Age + Stage)	2	0.534
	Variance Thresholding Only (top 3%)	1740	0.482
	Variance Thresholding + Penalized-Based Selection	10	**0.535**
	Manually Selected Genes (TAP1, ZFHX4, CXCL9, FBN1, PTGER)	5	0.500

Additional Discussion

Ultimately, the linear Cox regression models performed the best with a test concordance index of 0.583, with Cox neural networks performing at about the same level. Thus, although RSFs and Cox neural networks have more potential predictive power (they do not make the proportional hazards assumption), they did not find more predictive, non-linear patterns in the survival data. This can be observed when comparing the survival curves predicted with these models (Figures 2, 3, and 4). RSF models have more flexibility in predicting survival curves,[21] but in our experiments, the RSF survival curves show no clear survival time patterns and have significantly less variation than the curves predicted by the Cox regression model.

We also note that the SIS process appears to be unnecessary in the feature selection process. While SIS ranking is easier to interpret and compute, SIS evaluates each feature alone and does not account for how the feature may relate to other features. Thus, it suffices to only use penalized-based selection when this option is computationally feasible.

Lastly, while our feature selection methods were crucial in identifying predictive features, we found that the process is highly sensitive to the specific dataset used for feature selection. To evaluate this, we re-ran penalized-based feature selection on a different training set of 196 samples and found that the top 55 features had only 16 features in common with the original top 55 features. This implied that most features in the dataset simply did not contain enough signal that generalized across the entire dataset.

Conclusion & Future Work

In this project, we experimented with Cox regression and RSF models in combination with multiple feature selection techniques to predict survival times of HGSOC patients from gene expression data. Our best-performing model is a standard Cox regression model trained on 97 features selected via penalized-based feature selection. Although the model is simpler than more powerful models such as RSF and Cox neural networks, its test concordance index score of 0.583 is comparable to published work[22] and is slightly better than our clinical baselines. In addition, the top features utilized by our Cox regression and RSF models corresponded with the expression levels of genes that have been biologically linked to ovarian or other cancer types.

Still, more improvement is necessary before applying these predictions in a clinical setting. We found that the feature selection methods we used are highly sensitive to the training dataset used, causing models to quickly overfit. Future work is necessary to better handle the complexity and noise in the dataset feature space. For example, some studies first clustered the data into subtypes and then trained a separate model for each.[23] The feature space itself could also be simplified via feature embedding methods such as PCA or by carefully incorporating prior biological knowledge when selecting specific features.

We also plan to continue to tune our RSF and Cox neural network models. As mentioned, these models can learn more complex patterns in the data if such patterns are present, but this generally requires more data and careful hyperparameter tuning to avoid overfitting. While we do want to reduce the dimensionality of the feature space, we also will look to enrich the feature space with new sources of data (such as the epigenetic data also available in TCGA) or by carefully adding specific higher-order features (such as gene co-expression levels). Altogether, these efforts will help further elucidate how to use such biological data for predicting the prognosis of HGSOC patients.

Github code: https://github.com/ashamehta/cs229_project/

Contributions

- Preprocessed gene mutations and gene expressions data (Anthony)
- Preprocessed clinical data (Asha)
- Researched survival analysis methods (Asha, Anthony)
- Cox regression analysis (Anthony leading, Asha assisting/learning how to write better code)
- Coding Cox regression runs, feature selection, hyperparameter tuning (Anthony)
- Neural net Cox regression (Anthony)
- Coding random survival forest runs, some feature selection, randomized search hyperparameter tuning (Asha)
- Final paper (Asha, Anthony)
- (Our third teammate withdrew from the class in mid-May)

References

[1] National Ovarian Cancer Coalition. What is ovarian cancer: Ovarian tumors and cysts. Retrieved April 15, 2021, from https://www.cancer.org/cancer/ovarian-cancer/about/what-is-ovarian-cancer.html

[2] Lisio, M. A., Fu, L., Goyeneche, A., Gao, Z. H., & Telleria, C. (2019). High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *International journal of molecular sciences*, *20*(4), 952. doi: 10.3390/ijms20040952

[3] Network, T. C. G. A. (2011). Integrated genomic analyses of ovarian carcinoma. Nature 490:292. doi: 10.1038/nature11453

[4] Fowler, Matthew. "Research Finds Better Method to Estimate Ovarian Cancer Survival Rates." Cancer Network, Cancer Network, 16 Oct. 2020, www.cancernetwork.com/view/research-finds-better-method-to-estimate-ovarian-cancer-survival-rates.

[5] Chen, Chang MD; Xu, Yali MD; Huang, Xin MD; Mao, Feng MD; Shen, Songjie MD; Xu, Ying MD; Sun, Qiang MD. Clinical characteristics and survival outcomes of patients with both primary breast cancer and primary ovarian cancer, Medicine: August 07, 2020 - Volume 99 - Issue 32. doi: 10.1097/MD.0000000000021560

[6] Gao, Yi-Cheng, et al. "An Ensemble Strategy to Predict Prognosis in Ovarian Cancer Based on Gene Modules." *Frontiers*, Frontiers, 5 Apr. 2019, doi.org/10.3389/fgene.2019.00366.

[7] Millstein, J.Bowtell, D. et al. "Prognostic gene expression signature for high-grade serous ovarian cancer." Annals of Oncology, Volume 31, Issue 9, 1240 - 1250

[8] see *3.*

[9] "mRNA Analysis Pipeline." *GDC Docs*, docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#fpkm.

[10] Carroll, Orlagh U., et al. "How Are Missing Data in Covariates Handled in Observational Time-to-Event Studies in Oncology? A Systematic Review." BMC Medical Research Methodology, vol. 20, no. 1, May 2020, p. 134. BioMed Central, doi:10.1186/s12874-020-01018-7.

[11] Harrell , Frank E. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer International Publishing, 2015., doi:10.1007/978-3-319-19425-7

[12] LaMorte, Wayne W. "Cox Proportional Hazards Regression Analysis." https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html. Accessed 7 May 2021.

[13] Pölsterl, Sebastian. "Survival Analysis for Deep Learning." July 2019. https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/.

[14] Fan, Jianqing, et al. "High-Dimensional Variable Selection for Cox's Proportional Hazards Model." Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown, Jan. 2010, pp. 70–86. projecteuclid.org, doi:10.1214/10-IMSCOLL606.

[15] Khan, Md Hasinur Rahaman, and J. Ewart H. Shaw. "Variable Selection for Survival Data with a Class of Adaptive Elastic Net Techniques." Statistics and Computing, vol. 26, no. 3, May 2016, pp. 725–41. Springer Link, doi:10.1007/s11222-015-9555-8.

[16] Ishwaran, Hemant, et al. "Random Survival Forests." The Annals of Applied Statistics, vol. 2, no. 3, Sept. 2008, pp. 841–60. Project Euclid, doi:10.1214/08-AOAS169.

[17] Ching T., et al. "Coxnnet: An artificial neural network method for prognosis prediction of high-throughput omics data." PLoS Computational Biology, vol 14, no. 4, April 2018. doi.org: 10.1371/journal.pcbi.1006076.

[18] Uhlen, et al. "A pathology atlas of the human cancer transcriptome." Science, vol. 357, issue 6352, Aug. 2017. The Human Protein Atlas, doi: 10.1126/science.aan2507.

[19] Hao, Jie, et al. "Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data." BMC Medical Genomics, vol 12, Dec. 2019. doi: 10.1186/s12920-019-0624-2.

[20] see *7*

[21] Weathers, Brandon, and Culter, Richard. "Comparison of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis." May 2017. https://digitalcommons.usu.edu/gradreports/927

[22] see *16, 17*

[23] see *6*