# Dynamic Cobra Survival Analysis - Deep Learning and Statistical Methods

A Project Report Submitted

for the Course

## MA691 Advanced Statistical Algorithms

*by*

**Akshat Gupta** (Roll No. 180123002)

**Shubham Gandhi** (Roll No. 180123046)

**Shreyank Snehal** (Roll No. 180123045)

**Harsh Vardhan Singh Yadav** (Roll No. 180123014)

**Raj Kumar Roul** (Roll No. 180122037)

under the supervision of **Asst. Prof. Arabin Kumar Dey**

*to the*

**DEPARTMENT OF MATHEMATICS**

**INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI**

**GUWAHATI - 781039, INDIA**

*November 2021*

**Disclaimer:** This work is for learning purposes only. The work cannot be used for publications or commercial products without mentor's consent.

# ABSTRACT

Cystic fibrosis is an inherited disorder that causes severe damage to the lungs, digestive system and other organs in the body. There is no cure yet for this disease and thus foretelling the life span or survival rate of people affected by it is a great area of interest in clinical research. In this project, we analyse the CF dataset which is a longitudinal data which was collected by the UK Cystic Fibrosis Registry. We use multiple regression strategies to predict the survival function and then employ COBRA to combine all these methods to predict the survival function. We compare the results obtained by various methods and then see how COBRA is more efficient than all the other methods taken at once.

# 1 Problem Formulation

Currently available risk prediction methods are limited in their ability to deal with complex, heterogeneous, and longitudinal data such as that available in primary health care records. Survival analysis informs our understanding of the relationships between the (distribution of) first hitting times of events of interest (such as death, onset of a certain disease, etc.) and the covariates, and enables us to issue corresponding risk assessments for such events. Clinicians use survival analysis to make screening decisions or to prescribe treatments, while patients use the information about their clinical risks to adjust their lifestyles in order to mitigate such risks. Since the Cox proportional hazard model was first introduced, a variety of methods have been developed for survival analysis, ranging from statistical models to deep learning techniques.

A key limitation of existing survival models is that they utilize only a small fraction of the available longitudinal (repeated) measurements of biomarkers and other risk factors. In particular, even though biomarkers and other risk factors are measured repeatedly over time, survival analysis is typically based on the last available measurement. This represents a severe limitation, since the evolution of biomarkers and risk factors has been shown to be informative in predicting the onset of disease and various risks.

# 2 Cystic Fibrosis

Cystic Fibrosis (CF), which is the most common genetic disease in Caucasian populations, gives rise to different forms of dysfunction involving the respiratory and gastrointestinal systems, which primarily lead to progressive respiratory failure. Forced expiratory volume (FEV1), and its development, is a crucial biomarker in assessing the severity of CF as it allows clinicians to describe the progression of the disease and to anticipate the occurrence of respiratory failures. Therefore, to provide a better understanding of disease progression, it is essential to incorporate longitudinal measurements of biomarkers and risk factors into a model. Rather than discarding valuable information recorded over time, this allows us to make better risk assessments on the clinical events.

# 3 Survival Analysis Methods

In many health care studies, the main outcome under assessment is the time to an event of interest. The generic name for the time is survival time, although it may be applied to the time 'survived' from complete remission to relapse or progression as equally as to the time from diagnosis to death. If the event occurred in all individuals, many methods of analysis would be applicable. However, it is usual that at the end of follow-up some of the individuals have not had the event of interest, and thus their true time to event is unknown. Further, survival data are rarely normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. It is these features of the data that make the special methods called survival analysis necessary.

The following methods are implemented in our analysis of survival curves for CF-Patient:

- **Dynamic-DeepHit**, learns on the basis of the available longitudinal measurements, a data-driven distribution of first hitting times of competing events.

- **Kaplan Meier Estimate**, which predicts weather the patient will survive for at least one year.

- **Random Survival Forests**, a random survival forest (RSF) is an assemble of trees method for analysis of right censored time-to-event data and an extension of Brieman's random forest method.

- **Cox-Time**, a new method for time-to-event prediction, which is proposed by extending the Cox proportional hazards model with neural networks.

- **MTLR method**, which directly models the survival function, against Cox regression and Aalen regression as representatives of these survival analysis models.

- **COBRA Aggregation**, which takes distance between prediction of these above models to predict dynamic survival probabilities, to outperform our baseline KMF model.

## 3.1 Survival Function

The survival probability of a subject at time $\tau^*$, conditioned on the history of longitudinal measurements $\mathcal{X}^*$ can be derived by

$$\begin{aligned} S\left(\tau^* \mid \mathcal{X}^*\right) &\triangleq P\left(T > \tau^* \mid \mathcal{X}^*, T > t^*_{J^*}\right) \\ &= 1 - \sum_{k \neq \varnothing} F_k\left(\tau^* \mid \mathcal{X}^*\right) \end{aligned}$$

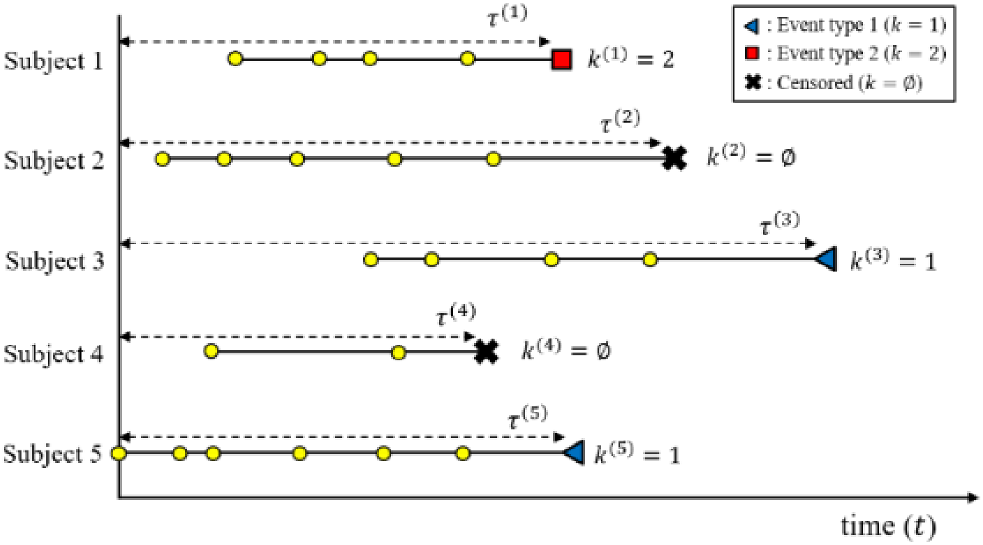
Here $F_k$ is the cumulative probability that death due to Cystic Fibrosis has occurred.

# 4 Dataset Description

We have used the CF dataset which is a longitudinal data which was collected by the UK Cystic Fibrosis Registry. This time-to-event (survival) data provides three pieces of information for each subject: i) observed covariates, ii) time-to-event(s), and iii) a label indicating the type of event (e.g., death or adverse clinical event) including right-censoring. Observed covariates include static (time-invariant) and time-varying covariates that are recorded for a period of time. We suppose that the longitudinal measurement times, event times, and censoring times are aligned based on a synchronization event, such as the entry to a clinical trial, the date of an intervention, and the onset of a condition.

For each subject, the dataset comprises of a sequence of longitudinal observations until time $t$, which include both static and time-varying covariates recorded at one or multiple times $t_j < t$. Covariates are not necessarily measured at regular time intervals and not every covariate is observed at each measurement (i.e., partially missing).

We assume that every subject experiences exactly one event among $K \geq 1$ possible events of interest. Survival data is often right-censored because events of interest are not always observed as subjects are lost to follow-up. The set of possible events is $K = \{\phi, 1, 2, \ldots, K\}$, with $\phi$ denoting the right-censoring.



In this illustration of survival data, yellow dots indicate the time at which longitudinal measurements are observed. $K = 1$ represents the event death due to Cystic Fibrosis while $K = 2$ represents death due to any other factors. The subjects having a black cross in the end are those for whom the data is censored.

Note: For our experiments we have $K = 2$ as censored data to simplify the model as most of the models don't support censorship in python.

## 5 COBRA: A combined regression strategy

COBRA (COmBined Regression Alternative) is a method which combines multiple weak learners. Given a set of preliminary estimators $r_1, \ldots, r_M$, it creates a prediction mapping for each weak learner on the training data. These are then used while predicting on the test data to find existing data points that are close to the considered point. The predictions corresponding to these data points are used to generate the final prediction by taking help of some summary metric. We have employed mean in our implementations.

# 6 Survival Machines

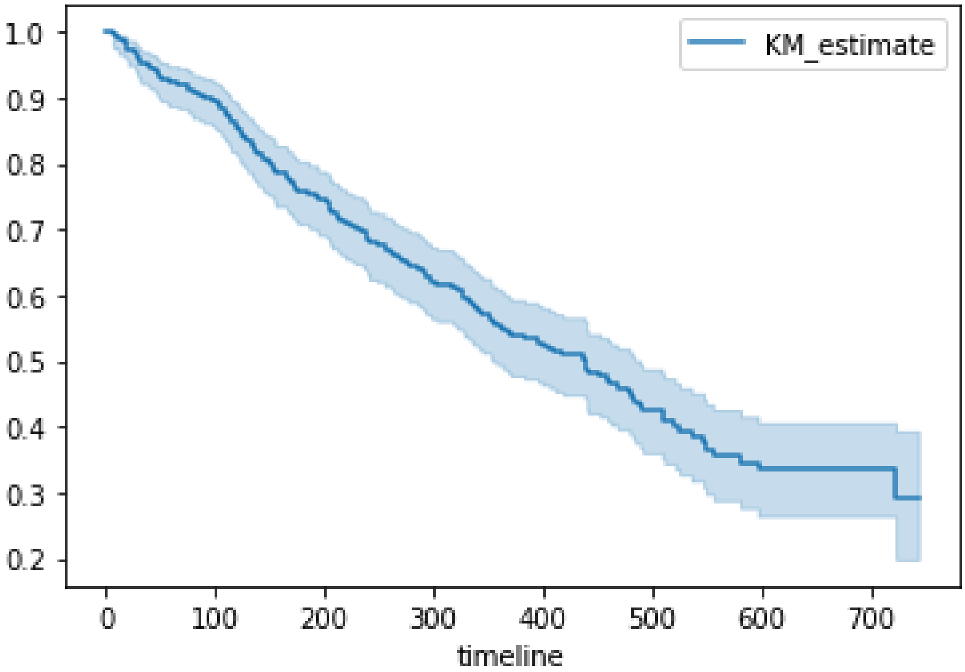
## 6.1 KMF

The Kaplan-Meier (KM) method is a popular method to analyze 'time-to-event' data. Within nephrology, the outcome variable in survival analyses is often all-cause mortality.

The estimator of the survival function S(t), the probability that life is longer than t is given by:

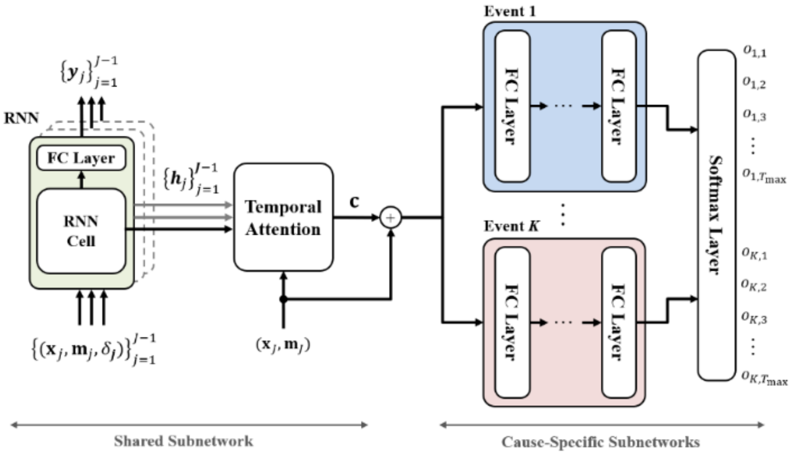
$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

The Kaplan-Meier estimator is one of the most frequently used methods of survival analysis. The estimate may be useful to examine recovery rates, the probability of death, and the effectiveness of treatment. It is limited in its ability to estimate survival adjusted for covariates; parametric survival models and the Cox proportional hazards model may be useful to estimate covariate-adjusted survival.



## 6.2 Dynamic-DeepHit

Dynamic-DeepHit is a multi-task network, which consists of two types of subnetworks: a shared subnetwork that handles the history of longitudinal measurements and predicts the next measurements of time-varying covariates, and a set of cause-specific subnetworks which estimates the joint distribution of the first hitting time and competing events. As the multitask learning has been successful across different applications [33]-[36], we jointly optimize the two subnetworks to help the overall network capture associations between the time-to-event under competing risks and i) the static covariates and ii) the progression of underlying process that governs the time-varying covariates.



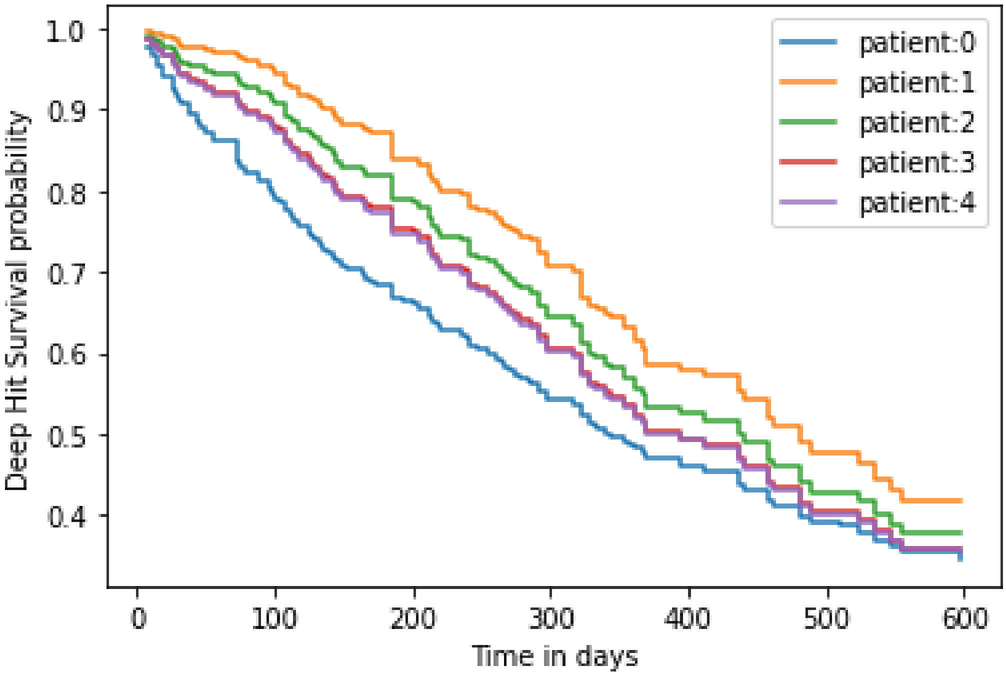
This figure illustrates the overall architecture of Dynamic-DeepHit which comprises of:

1. Shared Subnetwork: The shared subnetwork consists of two components: i) an RNN structure to flexibly handle the longitudinal data with each subject having different numbers of measurements, that are captured at irregular time intervals and are partially missing and ii) an attention mechanism to unravel the temporal importance of the history of measurements in making risk predictions.

2. Cause-specific Subnetworks: Each cause-specific subnetwork utilizes a feed-forward network composed of fully connected layers to capture relations between the cause specific risk and the history of measurements. The inputs to these subnetworks is the output from the shared subnetwork. Each cause-specific subnetwork captures the latent patterns that are distinct to each competing event.

3. Output Layer: Dynamic-DeepHit employs a soft-max layer in order to summarize the outcomes of each cause specific subnetwork, $f_{c_1}(\cdot), \cdots, f_{c_K}(\cdot)$, and to map into a proper probability measure. Overall, the network

produces an estimated joint distribution of the first hitting time and competing events. In particular, given a subject with $\mathcal{X}^*$, each output node represents the probability of having event $k$ at time $\tau$.



## 6.3 Random Survival Forests

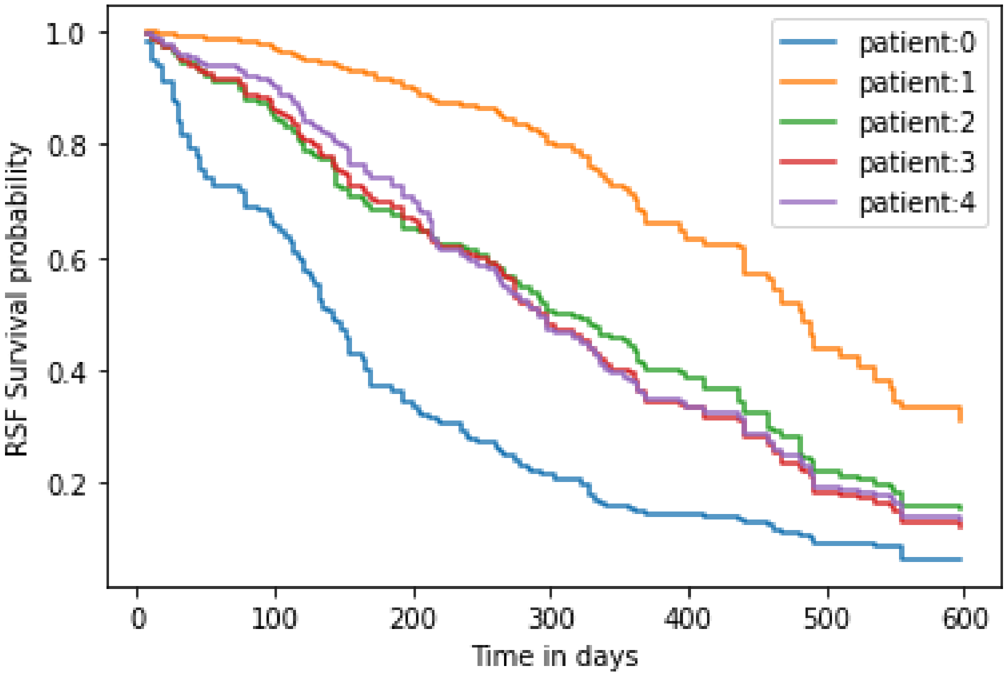
A random survival forest (RSF) is an ensemble of trees method for analysis of right censored time-to-event data and an extension of Brieman's random forest method. A Random Survival Forest ensures that individual trees are de-correlated by building each tree on a different bootstrap sample of the original training data, and then at each node, only evaluate the split criterion for a randomly selected subset of features and thresholds. Predictions are formed by aggregating predictions of individual trees in the ensemble.

- A survival tree is built with the idea of partitioning the covariate space recursively to form groups of subjects who are similar according to the time-to-event outcome.

- The basic approach for building a survival tree is by using a binary split on a single predictor. In tree building, a binary split is such that the two daughter nodes obtained from the parent node are dissimilar and several split-rules (different impurity measure) for time-to-event data have been suggested over the years. For a categorical covariant $X$, a split is defined as $X \leq c$ where $c$ is some constant. For a categorical covariate $X$ with many split-points, the potential split is $X \in \{c_1, \ldots, ck\}$ where $c\ 1, \ldots, ck$ are potential split values of a predictor variable $X$

- The goal in survival tree building is to identify prognostic factors that are predictive of the time-to-event outcome.

- RSF offers great flexibility and can automatically detect certain types of interactions without the need to specify them beforehand.



## 6.4 Cox-Time

Time-varying covariance occurs when a covariate changes over time during the follow-up period. Such variables can be analyzed with the Cox regression model to estimate its effect on survival time. The Cox-proportional hazards model is a popular choice for analysis of right censored time-to-event data. When it comes to predicting the survival function for a specific unit, the Cox Proportional Hazard Model is usually the go-to model. The model is convenient for its flexibility and simplicity, however, it has been criticised for its restrictive proportional hazards (PH) assumption which is often violated.

One approach for using time-varying covariate data is to extend the Cox proportional hazard model to allow time-varying covariates.

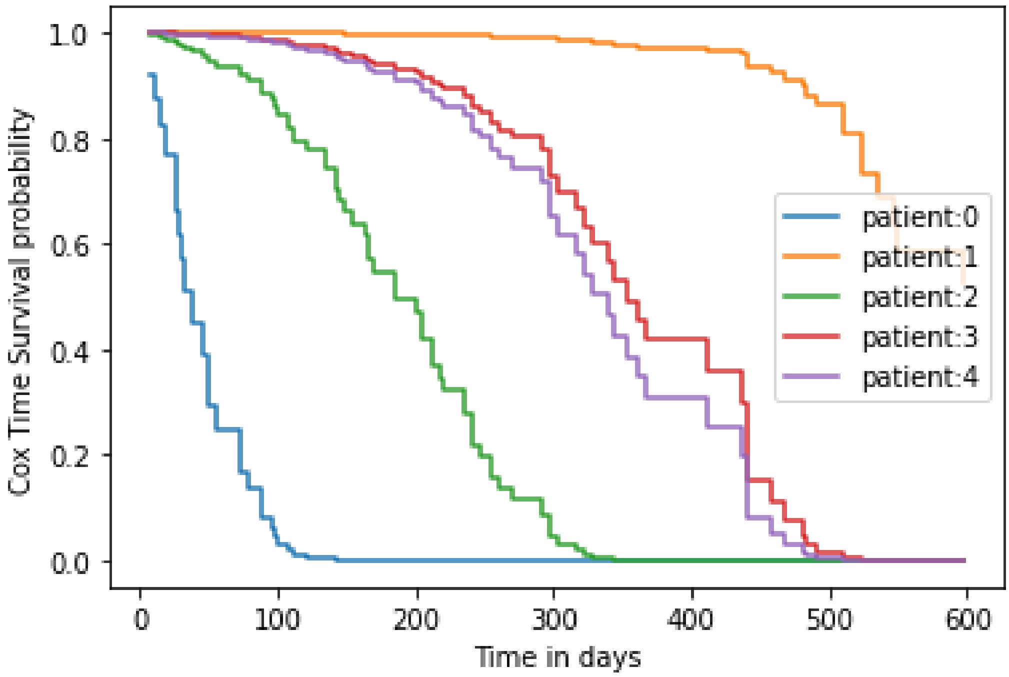
$$\lambda(\mathrm{t} \mid \mathrm{Z(t)}) = \lambda\mathrm{o}(t) \exp\left(\beta' x + \gamma' X g(\mathrm{t})\right)$$

where $\beta'$ and $\gamma'$ are coefficients of time-fixed and time-varying covariate respectively. Suppose we let $Z(t)$ represent the covariate, then:

$$\mathrm{Z(t)} = [x1, x2 \ldots xp, X1g(\mathrm{t}), X2g(\mathrm{t}) \ldots, Xq\ \mathrm{g(t)}]$$

and the hazard ratio which is non-constant is given by:

$$\widehat{HR} = \left(\frac{\lambda(t; Z(t)}{\lambda\left(t; Z(t)^*\right.}\right) = \exp\left(\beta' x^* + \gamma' X g(t)^*\right)$$



## 6.5 MTLR

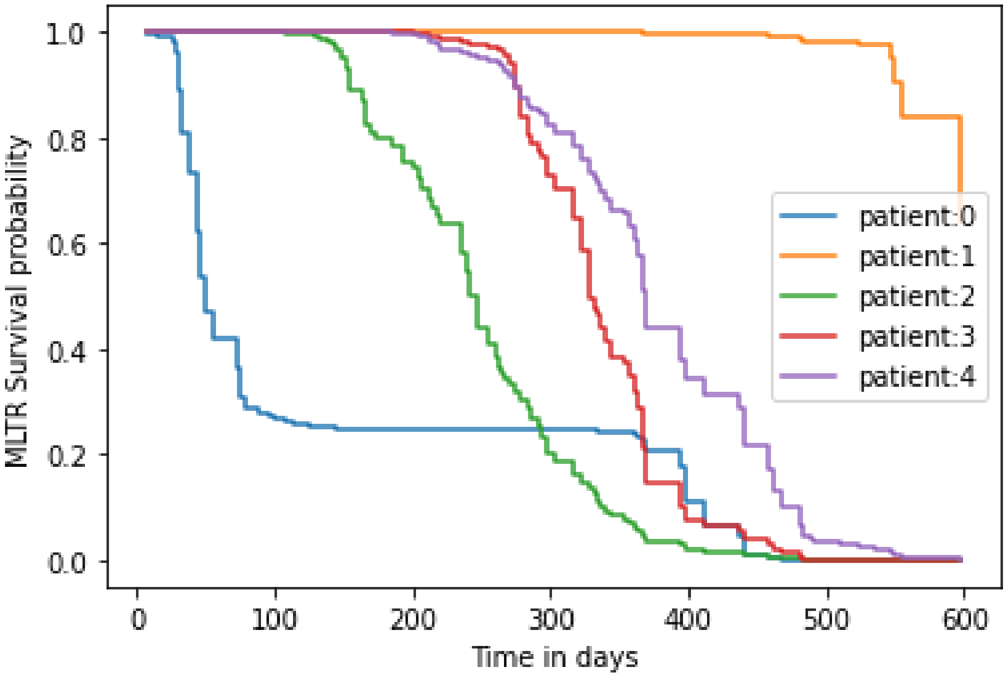
The Multi-Task Logistic Regression (MTLR) model is an alternative to Cox's proportional hazard model. It can be seen as a series of logistic regression models built on different time intervals so as to estimate the probability that the event of interest happened within each interval.

But in the presence of nonlinear elements in the data, it will stop yielding satisfactory performances as it is powered by a linear transformation. In order to introduce more modeling flexibility, the Neural Multi-Task Logistic Regression model (N-MTLR) which used the neural networks within the original MTLR design.

In the case of Neural Multi-Task Logistic Regression, the density and survival functions become:

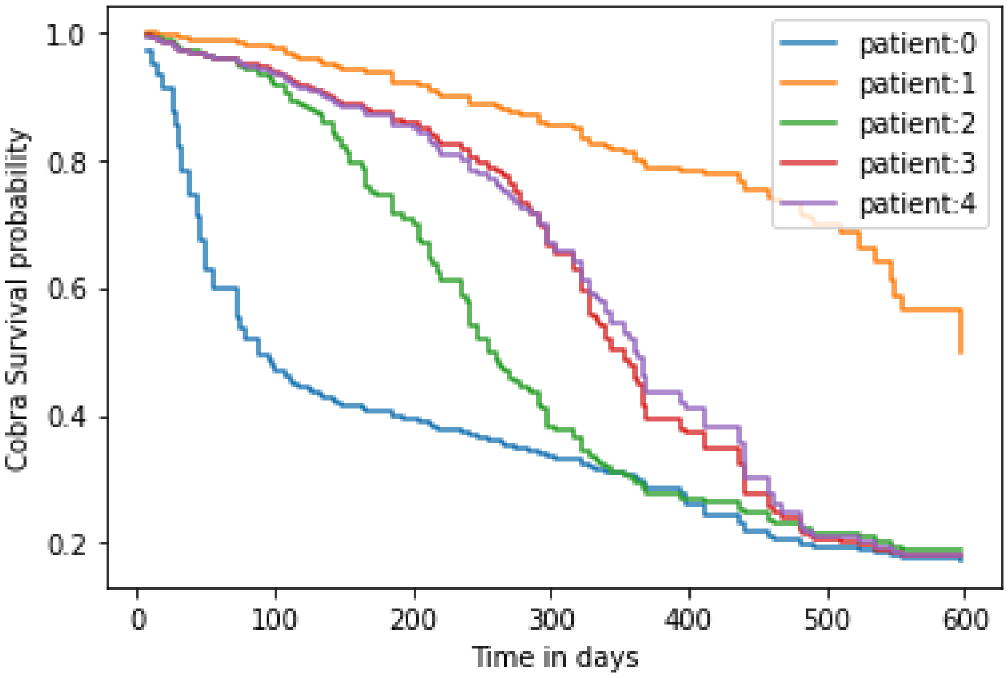
$$f\left(a_s, \vec{x}\right) = P\left[T \in [\tau_{s-1}, \tau_s) \mid \vec{x}\right] = \frac{\exp(\psi(\vec{x}) \cdot \Delta) \circ \vec{Y}}{Z(\psi(\vec{x}))}$$

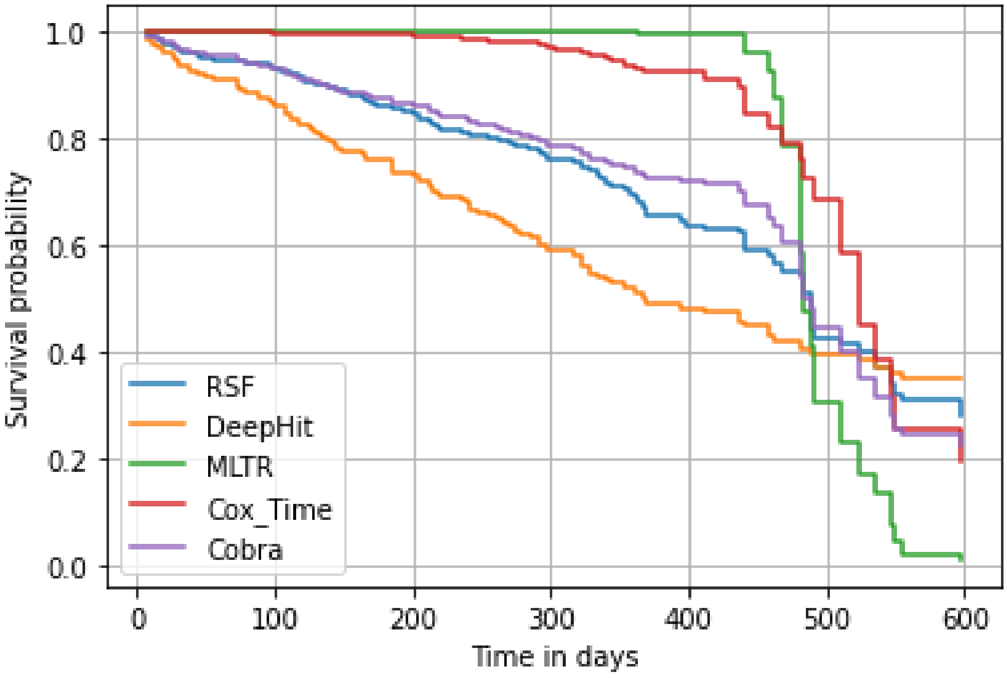
$$S\left(\tau_{s-1}, \vec{x}\right) = \sum_{k=s}^{J} \frac{\exp(\psi(\vec{x}) \cdot \Delta) \circ \vec{Y}}{Z(\psi(\vec{x}))}$$



# 7 COBRA Aggregation

Using the above machines we obtain survival functions which measures probability of patient surviving after a given time $\tau$. It is a cumulative probability measure used in most of the survival analysis methods. We create a pickle dictionary to store values of survival function at pre-defined event_times for each of 312 patients which can be used to make prediction. Thus we get an array of 312 x 113 prediction i.e Patients x Times.





For predicting the indicator outcome for an input from the test data, we do the following:

- For each machine, we find its prediction on the considered test point and we iterate through the prediction table to find $\epsilon$ close predictions. We mark these predictions for those machines

- We pick up the reference-training data points which have all machines giving $\epsilon$ close predictions (The total number of machines/models is given as $\alpha$ )

- Taking the mean of prediction probabilities corresponding to these data-points, which we already have since this is the training data, we get the outcome corresponding to our required test data point.

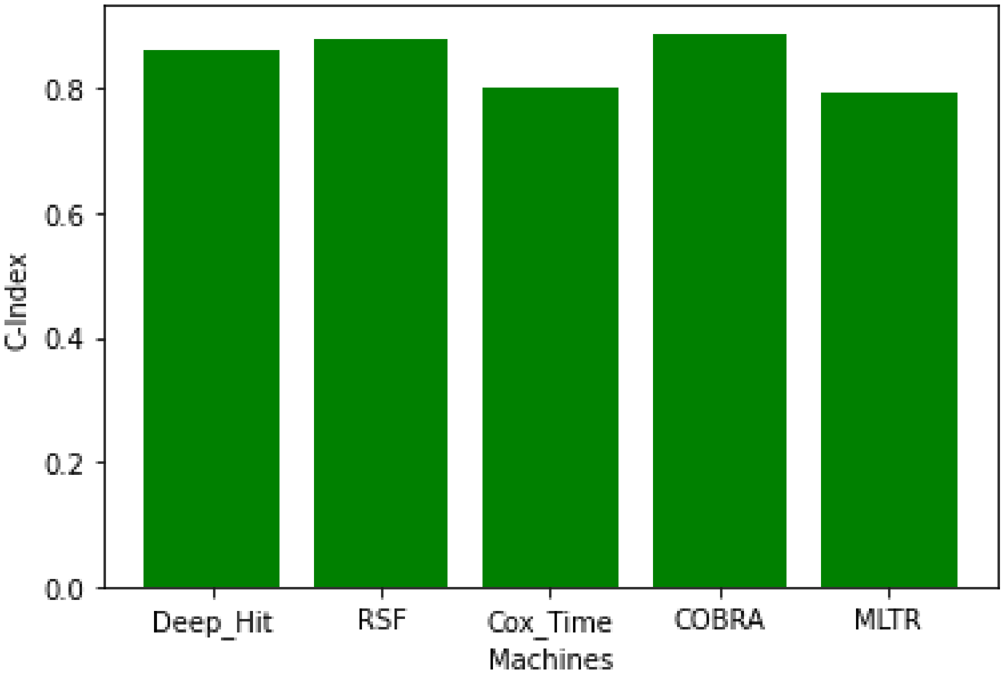
Since our time dimension is constant we iterate on our event_times to perform Cobra Aggregation learning method which takes average of common points within $\epsilon$ range. Where $\epsilon$ is chosen to be most optimal at $10^{-3}$. We perform this for each time step to get an cobra_prediction dictionary with Patients x Time dimension.

# 8 Evaluation

We use the two most common performance metrics in survival analysis to compare our models. These are calculated at a specific horizon time to give comparable results across all patients.

## 8.1 C-Index

The most frequently used evaluation metric of survival models is the concordance index (c index, c statistic). It is the measure of rank correlation between predicted risk scores $\hat{f}$ and observed time points $y$ that is closely related to Kendall's $\tau$. It is defined as the ratio of correctly ordered (concordant) pairs to comparable pairs. Two samples $i$ and $j$ are comparable if the sample with lower observed time $y$ experienced an event, i.e., if $y_j > y_i$ and $\delta_i = 1$, where $\delta_i$ is a binary event indicator. A comparable pair $(i, j)$ is concordant if the estimated risk $\hat{f}$ by a survival model is higher for subjects with lower survival time, i.e., $\hat{f}_i > \hat{f}_j \wedge y_j > y_i$, otherwise the pair is discordant.



The concordance index or C-index is a generalization of the area under the ROC curve ($A$ that can take into account censored data. It represents the global assessment of the model discrimination power: this is the model's ability to correctly provide a reliable ranking of the survival times based on the individual risk scores. It can be computed with the following form

$$\mathrm{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

with: - $\eta_i$, the risk score of a unit $i$ - $1_{T_j<T_i} = 1$ if $T_j < T_i$ else 0 - $1_{\eta_j>\eta_i} = 1$

if $\eta_j > \eta_i$ else 0 Similarly to the AUC, C-index = 1 corresponds to the best model prediction, and C-index = 0.5 represents a random prediction.

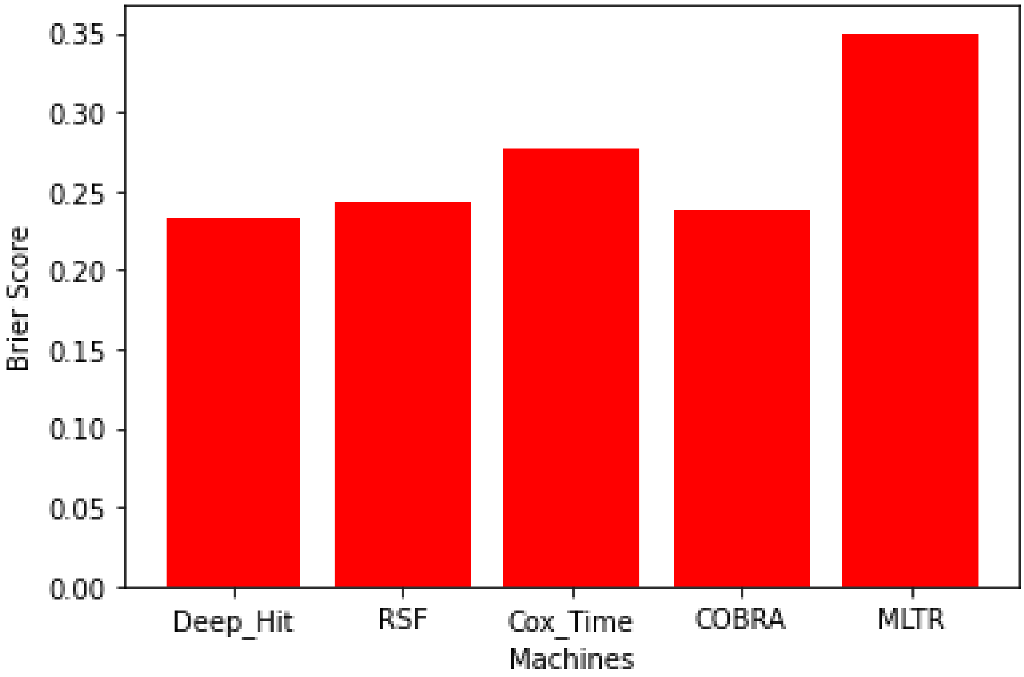
## 8.2 Brier Score

The Brier score is used to evaluate the accuracy of a predicted survival function at a given time $t$; it represents the average squared distances between the observed survival status and the predicted survival probability and is always a number between 0 and 1 , with 0 being the best possible value.

Given a dataset of $N$ samples, $\forall i \in 1, N, (\vec{x}_i, \delta_i, T_i)$ is the format of a datapoint, and the predicted survival function is $\hat{S}\left(t, \vec{x}_i\right), \forall t \in \mathbb{R}^{+}$.

In the absence of right censoring, the Brier score can be calculated as:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(1_{T_i > t} - \hat{S}\left(t, \vec{x}_i\right)\right)^2$$



# 9 Conclusion

We analysed the Cystic fibrosis dataset which is a longitudinal data collected by the UK Cystic Fibrosis Registry and applied various regression strategies, namely i) Dynamic-DeepHit, ii) Random Survival Forest, iii) Cox-Time and iv) MTLR, on it to predict the survival function. We then used COBRA (COmBined Regression Alternative) to combine all these methods and predict the survival function. From our results, we saw that the predictions given by COBRA were more accurate than any of the other four methods. Thus we can conclude that COBRA is a better regression strategy than any of the individual regression strategies it takes into account.

# References

[1] Gérard Biau, Aurélie Fischer, Benjamin Guedj, and James D. Malley. Cobra: A combined regression strategy. *Journal of Multivariate Analysis*, 146:18–28, Apr 2016.

[2] Frank Emmert-Streib and Matthias Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019.

[3] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression, 2019.

[4] Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020.

[5] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.