Survival Analysis Report

Data Analysis for the Pan-Cancer Data

Prepared By

Ashutosh Anand

[ashutoshanand3007@gmail.com]

August 4, 2022

Survival Analysis

A branch of statistics focused on analysing time to an event. Survival data relates to the time taken for an individual to reach a certain event. There is a concept of censoring in which if a person survived more than the observed duration, then it will be a right-censored. Whereas, If the survival duration is less than observed duration then it will be a left-censored data. In other words, Censoring means that an individual has not experienced the event by the end of the study e.g., they withdrew from the study or died from an unrelated event.

Data Description:

The data consist of 10 variables/columns. The variable description is presented as the following:

bcr_patient_barcode: Individual Patient Barcode

type: Type (Categorised) of Cancer.

age: Age in years

gender: Male, Female

race: Category of Humankind.

tumor_stage: The stage of tumour, such as Stage I,II,III,IV,X,IS,etc.

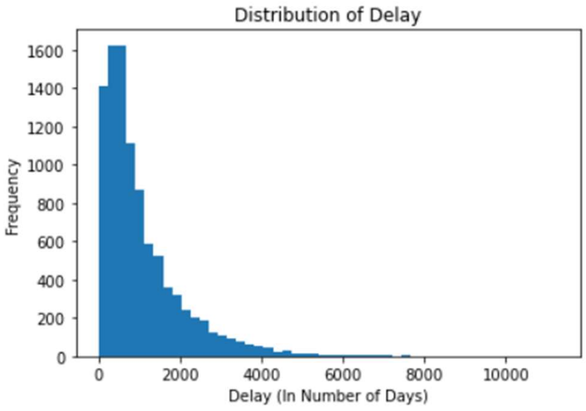
vital_status: Dead, Alive

treatment_outcome: What is the effect of treatment on patients, such as Complete or Partial Remission/Response, Stable Disease.

Event: censoring status 0 = censored, 1 = dead

Delay: Survival time in days.

Data Distribution:



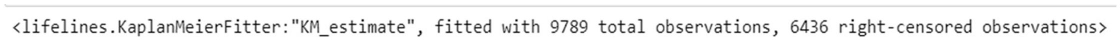
This Histogram Plot shows the Frequency Distribution of Delay Column as per the plot we can say that the data is left skewed.

Kaplan-Meier Estimator

The standard non-parametric technique to estimate the survival function of time-to-event data is proposed by Kaplan and Meier, is called the Product-limit estimator.

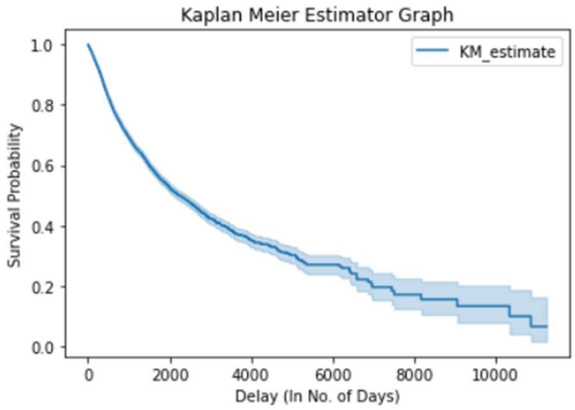
The Kaplan Meier Fitter is in Lifelines library and can be imported from there.

We can fit the model for the data using .fit() function and it gives us below output that shows how much data is right-censored observations



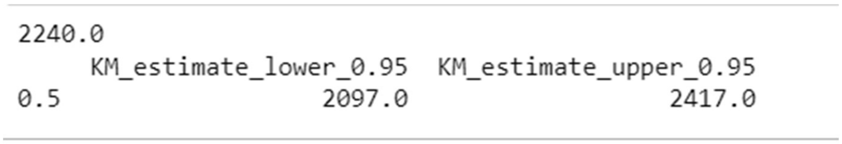
```
<lifelines.KaplanMeierFitter:"KM_estimate", fitted with 9789 total observations, 6436 right-censored observations>
```

Findings 1:



This graph is plotted using Kaplan Meier Fitter’s plot_survival_function_ and its y-axis represents the probability of experiencing the event after surviving up to time *t*, represented on the x-axis. Each drop in the survival function is caused by the event of interest happening for at least one observation.

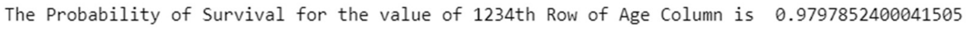
The length of the vertical line represents the fraction of observations at risk that experienced the event at time *t*. The height of the drop can also tell us about the number of observations at risk.



```
2240.0
     KM_estimate_lower_0.95  KM_estimate_upper_0.95
0.5                  2097.0                  2417.0
```

Here, the median survival time is 310 days, which indicates that 50% of the sample live 2240 days and 50% dies within this time. The 95% Confidence Interval lower limit is 2097 days, while the upper limit is 2417 days.

Using the .median_survival_time_ we can get the median survival time and with .confidence_interval_ we get Confidence Interval Limits.



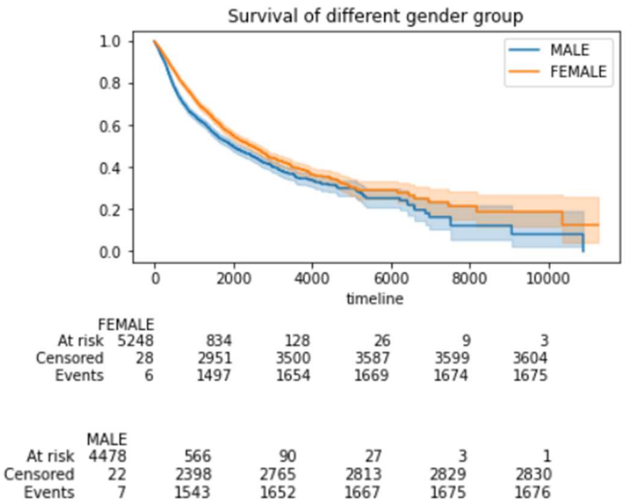
```
The Probability of Survival for the value of 1234th Row of Age Column is  0.9797852400041505
```

Here, 1234th row of Age Column is taken for the prediction of the probability of survival using Kaplan Meier Fitter’s .predict() function.

```
The Probability of Survival for the value of 11089th Row of Delay Column is  0.13478085327940661
```

Similarly, 11089th row of Delay Column is taken for the prediction of the probability of survival using Kaplan Meier Fitter’s .predict() function.

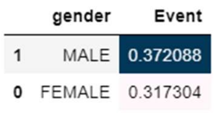
Findings 2:



```
        FEMALE
At risk  5248    834    128     26      9      3
Censored   28   2951   3500   3587   3599   3604
Events      6   1497   1654   1669   1674   1675


          MALE
At risk   4478    566     90     27      3      1
Censored    22   2398   2765   2813   2829   2830
Events       7   1543   1652   1667   1675   1676
```

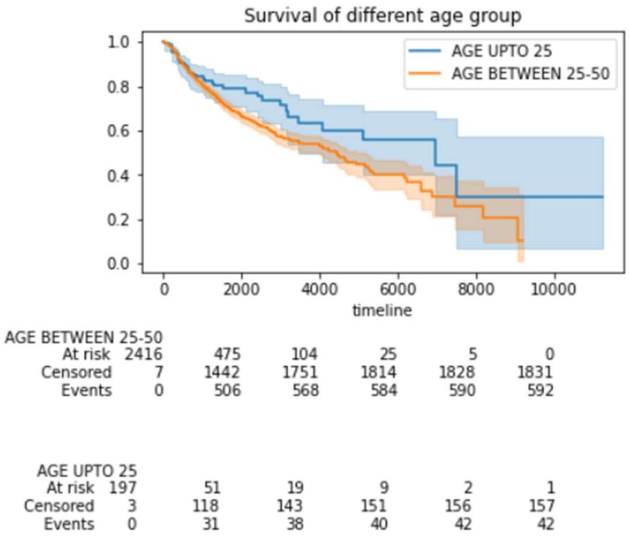
This figure is the survivor curve of each stratification of sex fitted with KM estimator, where the orange line means the survival probability of female for the patients and the blue line means survival probability of male for the individuals. Form this figure, we could interpret that the lifetime of female for the patients is longer than males. On the other hand, there has larger death risk for males than females.



	gender	Event
1	MALE	0.372088
0	FEMALE	0.317304

This table shows the Mean value of Event for each gender level, the Male have higher value, that means they are more possibility of event happening if a patient is a male than if he/she were a female.

Findings 3:

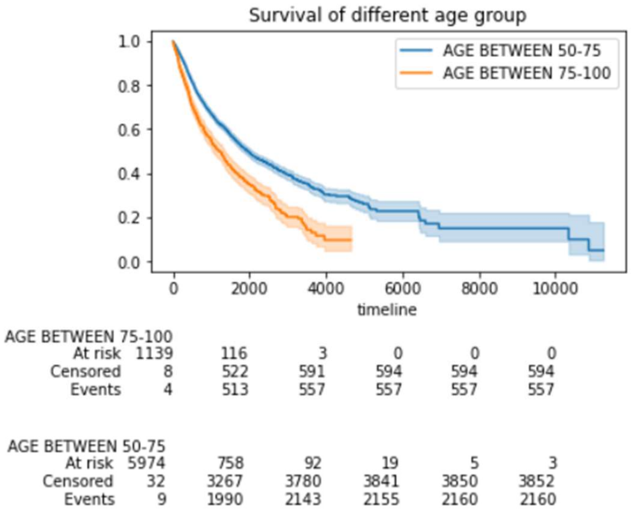


AGE BETWEEN 25-50						
At risk	2416	475	104	25	5	0
Censored	7	1442	1751	1814	1828	1831
Events	0	506	568	584	590	592

AGE UPTO 25						
At risk	197	51	19	9	2	1
Censored	3	118	143	151	156	157
Events	0	31	38	40	42	42

This figure plots the estimators of survival function for two categories of age with the Kaplan-Meier estimator. It shows that the patients of Age up to 25 have more larger survival probability than patients between Age of 25-50.

The table below the graph shows how many are At Risk, Censored, Events occurred at every section of timeline (first column shows data of 0-2000 in timeline and so on.).

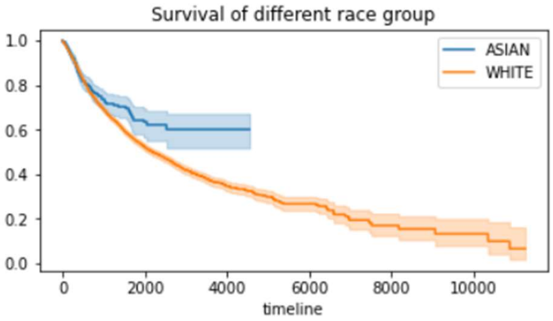


AGE BETWEEN 75-100						
At risk	1139	116	3	0	0	0
Censored	8	522	591	594	594	594
Events	4	513	557	557	557	557

AGE BETWEEN 50-75						
At risk	5974	758	92	19	5	3
Censored	32	3267	3780	3841	3850	3852
Events	9	1990	2143	2155	2160	2160

The graph shows that the patients of Age between 50-75 have somewhat more larger survival probability than patients between Age of 75-100.

Findings 4:

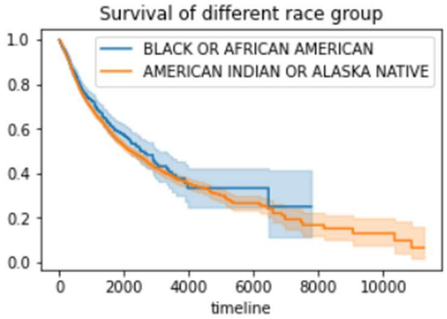


```
                WHITE
       At risk  9061     1344      217       53       12        4
     Censored    43     4874     5737     5871     5899     5905
       Events    13     2899     3163     3193     3206     3208


                ASIAN
       At risk  665          152          56          11          1
     Censored     7          392         475         518        528
       Events     0          128         141         143        143
```

Patients of Asian race seems to be survived more than White race patients.

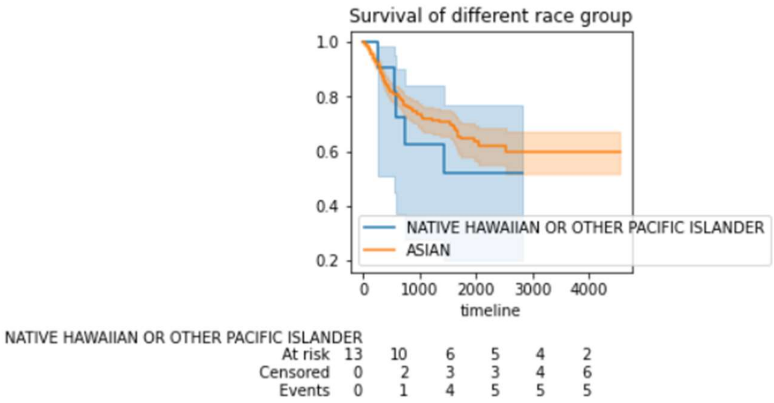
At any given duration, a higher proportion of Asian patients lived more than white patients.



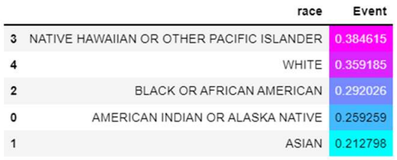
```
BLACK OR AFRICAN AMERICAN
                 At risk  923   310   131    43    15     7     6     2     0
               Censored     5   421   553   622   643   651   652   655   657
                 Events     0   197   244   263   270   270   270   271   271


    AMERICAN INDIAN OR ALASKA NATIVE
                         At risk  8803   1269    203     47     12      4
                       Censored     45   4796   5622   5748   5771   5777
                         Events     13   2796   3036   3066   3078   3080
```

Patients of Black or African American race seems to be survived more than American Indian or Alaska Native race patients. At some point, the confidence intervals overlap, that means it is less likely that there is real difference between the curves.



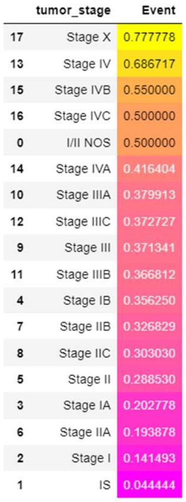
Patients of Asian race seems to survive more than Native Hawaiian race patient. At some point, the confidence intervals overlap, that means it is less likely that there is real difference between the curves.



	race	Event
3	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	0.384615
4	WHITE	0.359185
2	BLACK OR AFRICAN AMERICAN	0.292026
0	AMERICAN INDIAN OR ALASKA NATIVE	0.259259
1	ASIAN	0.212798

The tabular representation of each level of Race and its Event’s Mean value of each particular level, value closer to 1 means, those patients are more not likely to survive.

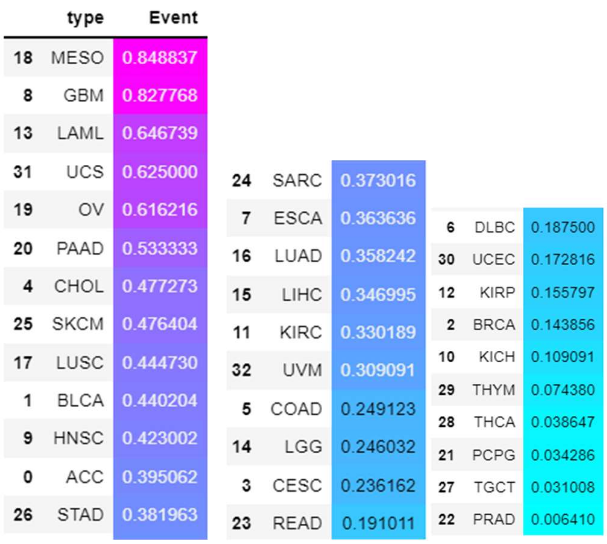
Findings 5:



	tumor_stage	Event
17	Stage X	0.777778
13	Stage IV	0.686717
15	Stage IVB	0.550000
16	Stage IVC	0.500000
0	I/II NOS	0.500000
14	Stage IVA	0.416404
10	Stage IIIA	0.379913
12	Stage IIIC	0.372727
9	Stage III	0.371341
11	Stage IIIB	0.366812
4	Stage IB	0.356250
7	Stage IIB	0.326829
8	Stage IIC	0.303030
5	Stage II	0.288530
3	Stage IA	0.202778
6	Stage IIA	0.193878
2	Stage I	0.141493
1	IS	0.044444

The tabular representation of each level of Tumor_Stage and Mean value of Event of each particular Tumour Stage, value closer to 1 means, those people are more not likely to survive. As per table, Patient at Stage X is more proximate to death than the patient at Stage IVs.

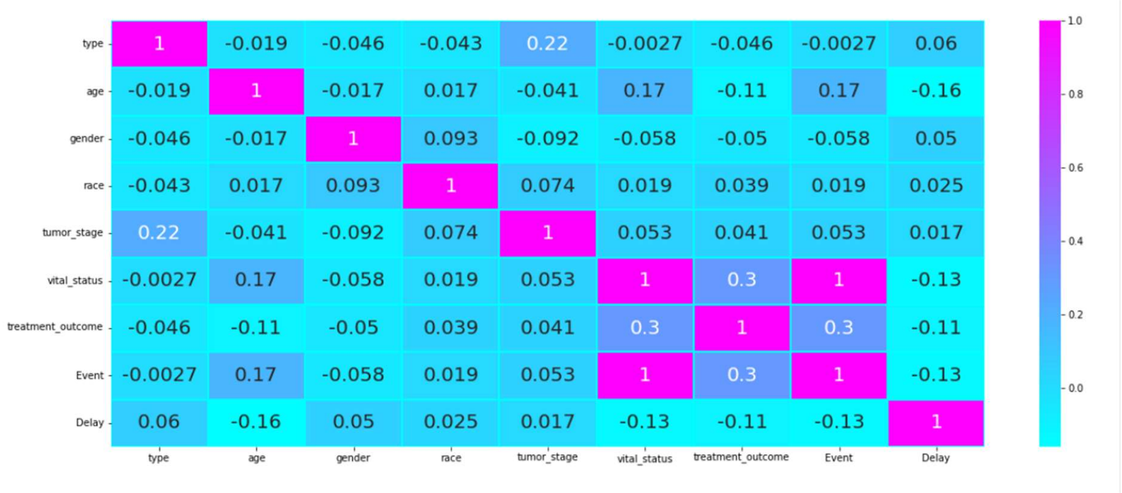
Findings 6:



	type	Event
18	MESO	0.848837
8	GBM	0.827768
13	LAML	0.646739
31	UCS	0.625000
19	OV	0.616216
20	PAAD	0.533333
4	CHOL	0.477273
25	SKCM	0.476404
17	LUSC	0.444730
1	BLCA	0.440204
9	HNSC	0.423002
0	ACC	0.395062
26	STAD	0.381963
24	SARC	0.373016
7	ESCA	0.363636
16	LUAD	0.358242
15	LIHC	0.346995
11	KIRC	0.330189
32	UVM	0.309091
5	COAD	0.249123
14	LGG	0.246032
3	CESC	0.236162
23	READ	0.191011
6	DLBC	0.187500
30	UCEC	0.172816
12	KIRP	0.155797
2	BRCA	0.143856
10	KICH	0.109091
29	THYM	0.074380
28	THCA	0.038647
21	PCPG	0.034286
27	TGCT	0.031008
22	PRAD	0.006410

The tabular representation of each level of Type and Mean value of Event of each particular Type. Showing the malignant type as MESO and benign as PRAD.

Findings 7:



	type	age	gender	race	tumor_stage	vital_status	treatment_outcome	Event	Delay
type	1	-0.019	-0.046	-0.043	0.22	-0.0027	-0.046	-0.0027	0.06
age	-0.019	1	-0.017	0.017	-0.041	0.17	-0.11	0.17	-0.16
gender	-0.046	-0.017	1	0.093	-0.092	-0.058	-0.05	-0.058	0.05
race	-0.043	0.017	0.093	1	0.074	0.019	0.039	0.019	0.025
tumor_stage	0.22	-0.041	-0.092	0.074	1	0.053	0.041	0.053	0.017
vital_status	-0.0027	0.17	-0.058	0.019	0.053	1	0.3	1	-0.13
treatment_outcome	-0.046	-0.11	-0.05	0.039	0.041	0.3	1	0.3	-0.11
Event	-0.0027	0.17	-0.058	0.019	0.053	1	0.3	1	-0.13
Delay	0.06	-0.16	0.05	0.025	0.017	-0.13	-0.11	-0.13	1

Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The value of treatment_outcome and Event have highest correlation and second highest is of tumor_stage and type, close to the 1 and thus if one increases so does the other and the closer to 1 the stronger this relationship is.

A correlation closer to -1 is similar as in the case of Age and Delay, but instead of both increasing one variable will decrease as the other increases. The diagonals are all 1/ pink because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and darker the color the higher the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares such as Event and Vital_status.