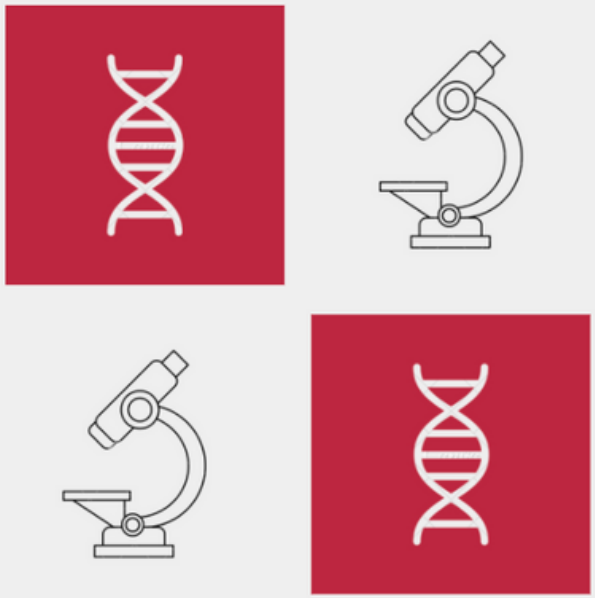
CAN WE PREDICT 7-YEAR SURVIVAL?

WITH SURVIVAL ANALYSIS

Prostate Cancer Case

Anisha Joshi | Healthcare Analytics

Our Agenda for Today

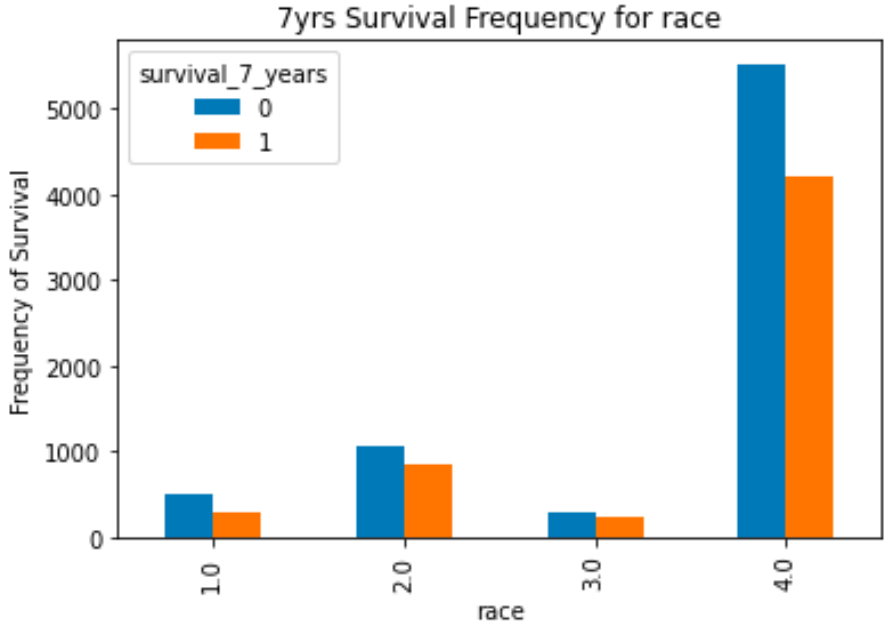


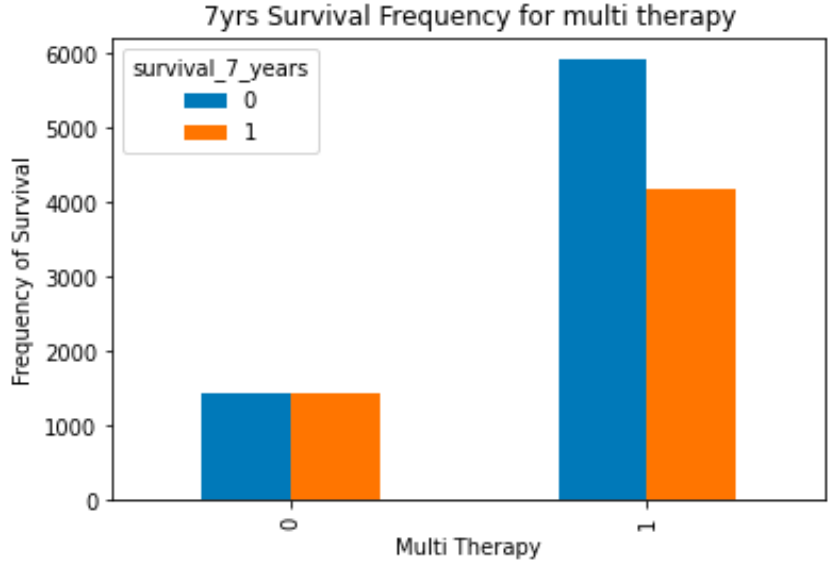
LIST OF KEY CONCEPTS

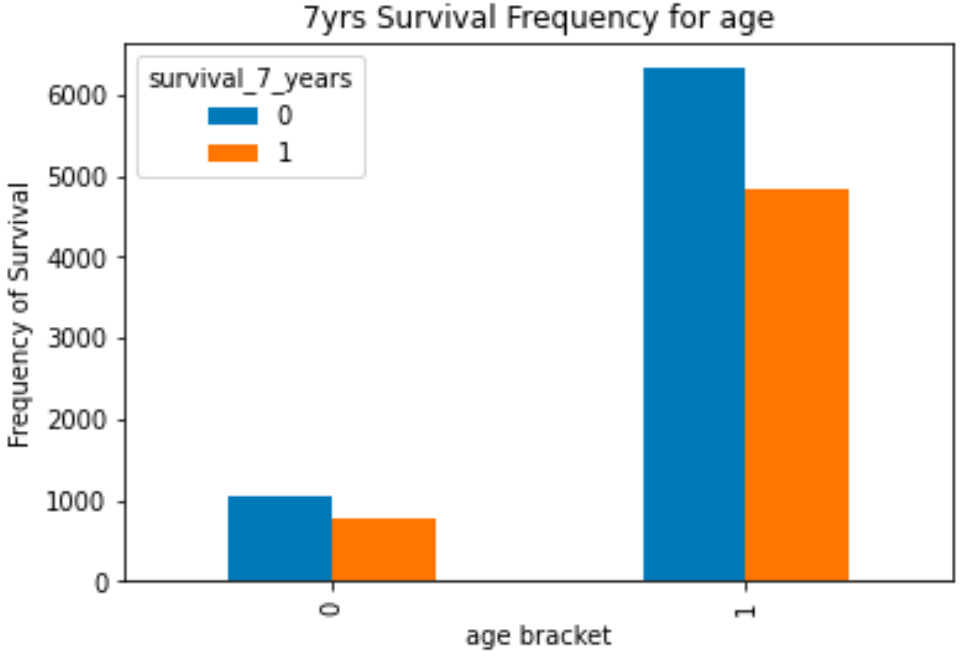
-Introduction to data
-Exploratory Data Analysis
-Feature Engineering
-Feature Selection
-Define Time & Event
-Survival Probability v/s Timeline
-Cox's Propotional Hazard

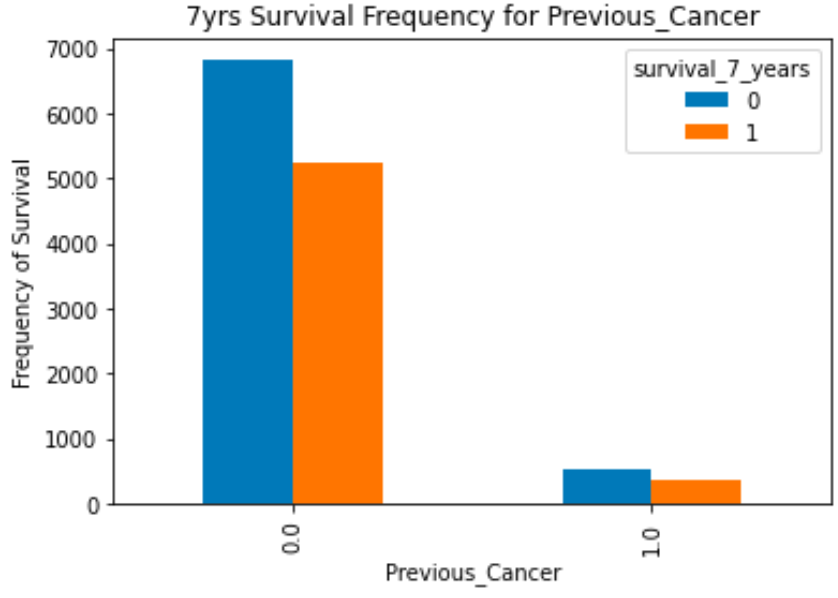
Assuming the provided data is of 2021 & 2022

EXPLORATORY ANALYSIS

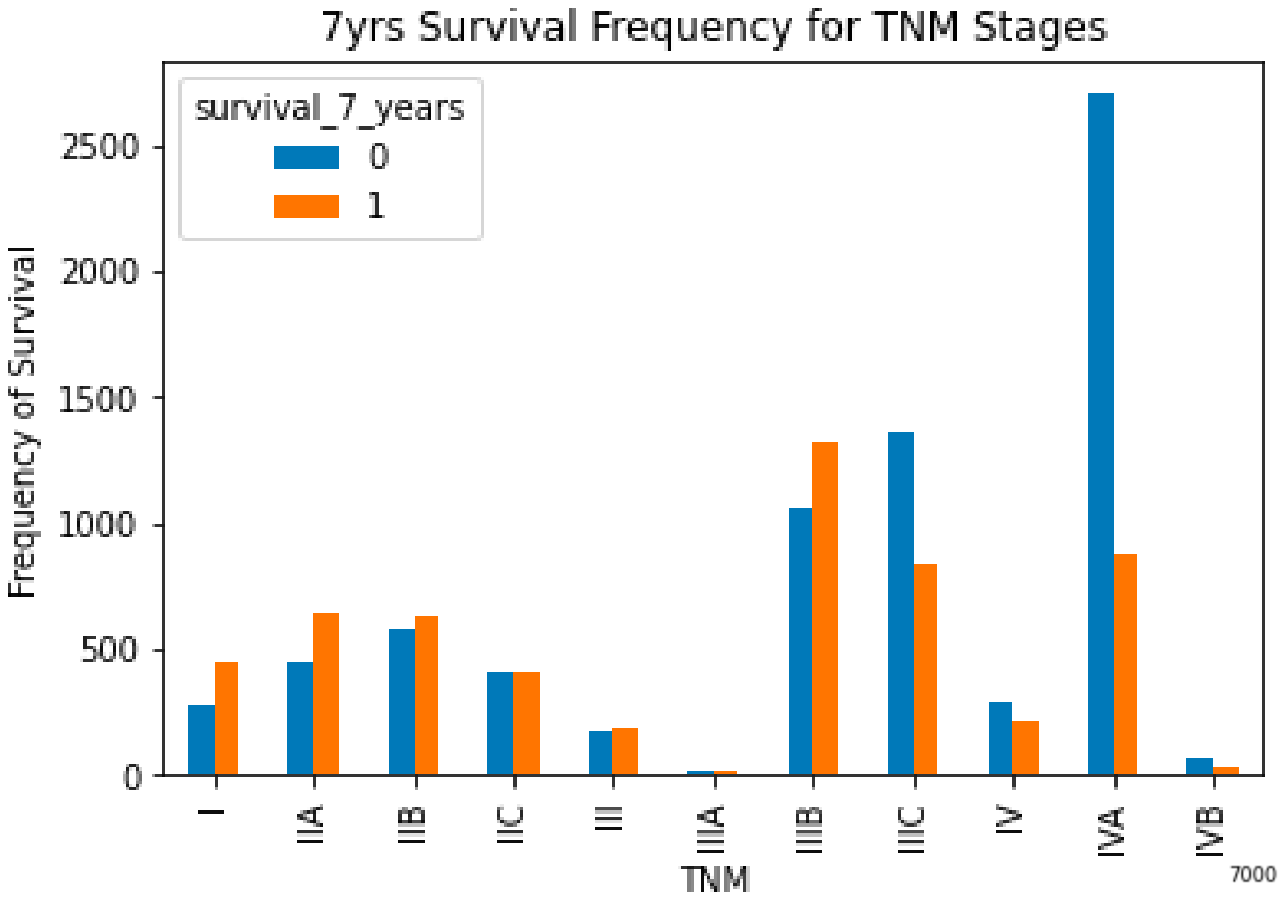


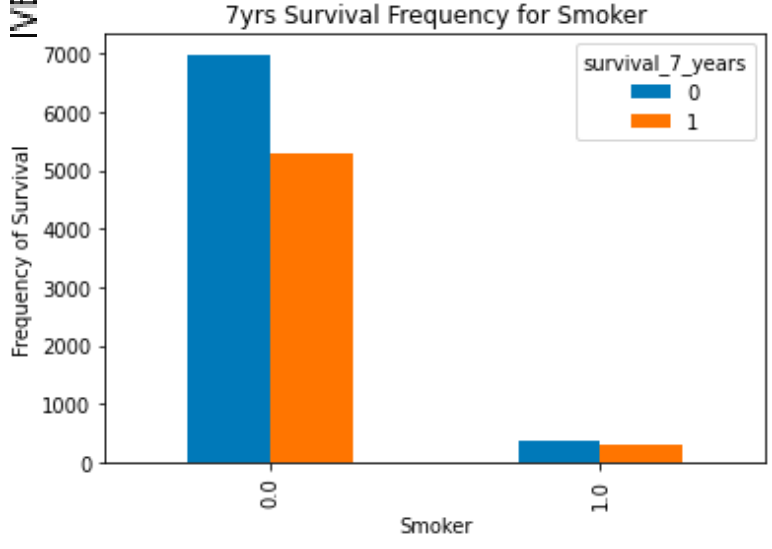






EXPLORATORY ANALYSIS





Subject Matter

The TNM system for prostate cancer is based on 5 key pieces of information:

- The extent of the main (primary) **tumor** (**T** category)
- Whether cancer has spread to nearby lymph **nodes** (**N** category)
- Whether cancer has spread (**metastasized**) to other parts of the body (**M** category)
- The **PSA level** at the time of diagnosis
- The **Grade Group** (based on the **Gleason score**), is a measure of how likely the cancer is to grow and spread quickly. This is determined by the results of the prostate biopsy (or surgery).

Instead of these 5 categories have TNM score

Ref - https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/staging.html

Feature Engineering

Delete Date, ID and combined symptoms

SYMPTOMS

Dummy One Hot Key Encoding, Add later, Sym > 3 = 1 as Sym

AGE

In different Brackets Initially, later 65+

DIFF. IN TUMOUR SIZE

From date of diagnosis to 6 months/year

HEIGHT & WEIGHT

Replaced with BMI

Define Time/Event

RFE with Estimator as Logistic Regression

If Survival_7_years =1, then Event(dead) = 0, else dead =1

if Survival_7_year =1 >> time = 2555 days
If Survival_7_years = 0 and Survival_1_year =0 >> time = (random num(366,2555 days))
If Survival_7_years =0 & Survival_1_year =0

- Check PSA_6_mon >> If not NA then time = (random num(183,364 days))
- If PSA_6_mon is NA then time = (random num(1,182 days))

#	Column	Non-Null Count	Dtype
0	rd_thrpy	15385 non-null	int64
1	rad_rem	15385 non-null	int64
2	survival_1_year	15385 non-null	int64
3	age	14637 non-null	float64
4	stage	15385 non-null	object
5	TNM_sc	15385 non-null	object
6	survival_7_years	15385 non-null	int64
7	race4	15385 non-null	int64
8	time	15385 non-null	int64
9	dead	15385 non-null	int64

Final data

15,385

Rows

33

Features

14,637

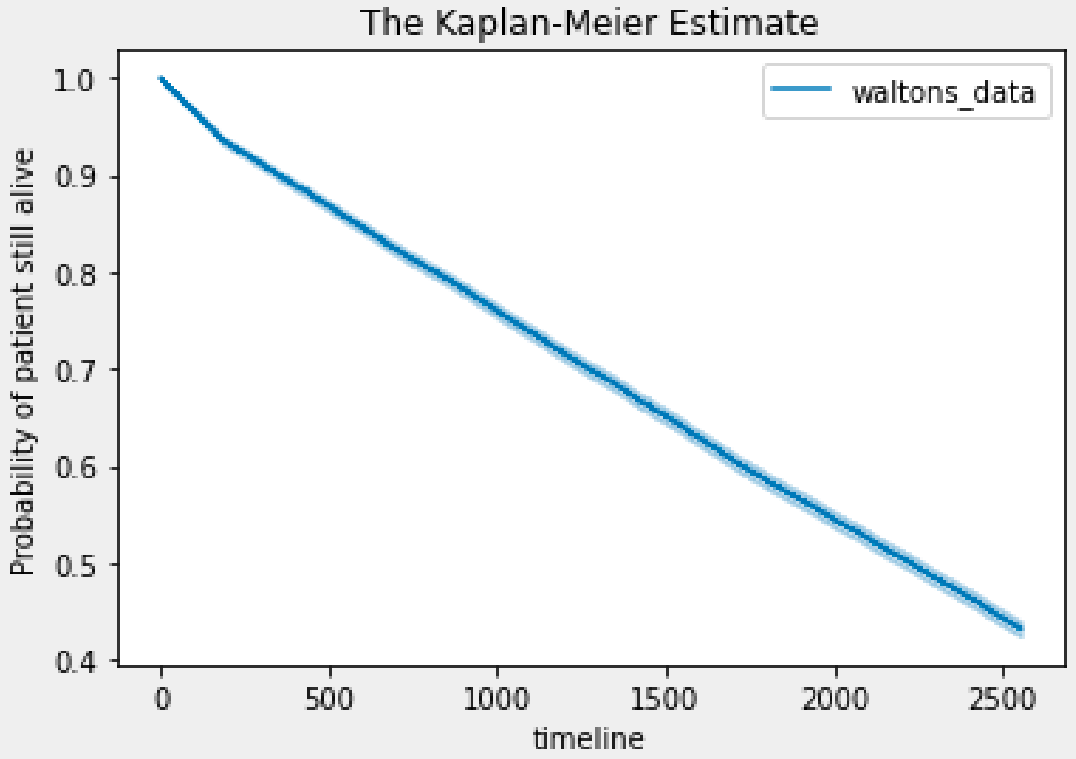
Final Rows

10

Final Features

After Dummy Variables - 22 final features

Probability of survival versus timeline

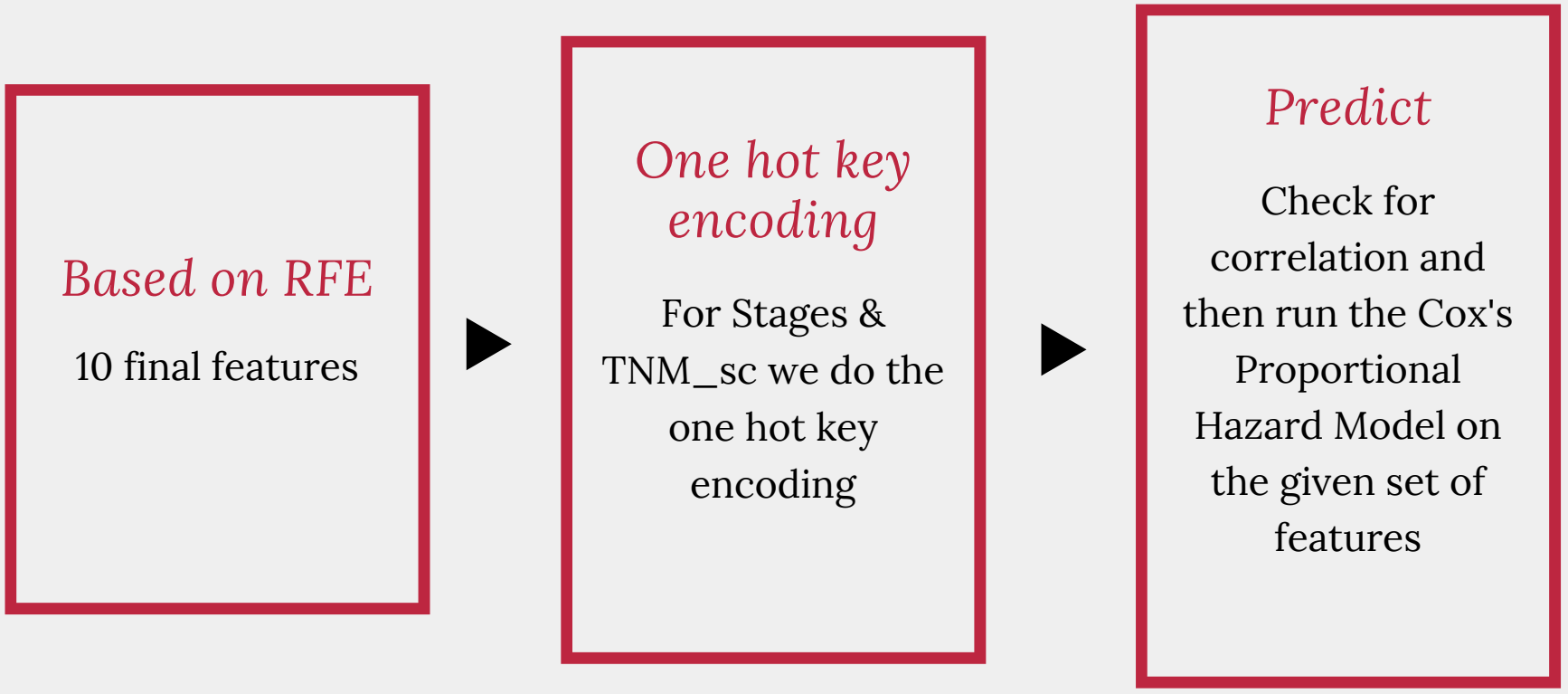


Based on the Given Dataset 2228 is number of days where on average 50% of patients died

The survival probability for a patient at timeline 0 is 1. The probability that a person dies on the 1st day of diagnosis is nearly equal to 0. So we can say that the survival probability is as high as possible. As the timeline increases, the probability of survival decreases for a patient.

Cox's Proportional Hazard

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. It allows us to examine how specified factors influence the rate of a particular event happening (death) at a particular point in time.



Cox's Proportional Hazard

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
rd_thrpy	0.23	1.25	0.02	0.19	0.27	1.20	1.31	10.80	<0.005	87.95
rad_rem	0.14	1.15	0.03	0.08	0.19	1.08	1.21	4.56	<0.005	17.57
survival_1_year	-3.75	0.02	0.05	-3.85	-3.66	0.02	0.03	-76.00	<0.005	inf
age	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	1.28	0.20	2.31
race4	-0.01	0.99	0.02	-0.05	0.04	0.95	1.04	-0.27	0.78	0.35
stage_I	-0.25	0.78	0.09	-0.42	-0.08	0.66	0.92	-2.95	<0.005	8.29
stage_IIA	-0.12	0.88	0.05	-0.22	-0.03	0.80	0.97	-2.50	0.01	6.35
stage_IIB	0.02	1.03	0.04	-0.05	0.10	0.95	1.11	0.63	0.53	0.91
stage_III	-0.09	0.92	0.04	-0.17	-0.00	0.85	1.00	-2.07	0.04	4.70
stage_IV	0.12	1.13	0.04	0.05	0.20	1.05	1.22	3.28	<0.005	9.92
TNM_I	-0.17	0.84	0.07	-0.30	-0.04	0.74	0.96	-2.53	0.01	6.44
TNM_IIA	-0.16	0.85	0.05	-0.27	-0.06	0.76	0.95	-3.01	<0.005	8.56
TNM_IIB	-0.06	0.94	0.05	-0.15	0.03	0.86	1.03	-1.26	0.21	2.26
TNM_IIC	-0.09	0.92	0.05	-0.19	0.02	0.83	1.02	-1.66	0.10	3.38
TNM_IIIA	-0.19	0.82	0.26	-0.70	0.32	0.50	1.37	-0.74	0.46	1.12
TNM_IIIB	-0.21	0.81	0.04	-0.27	-0.14	0.76	0.87	-5.79	<0.005	27.08
TNM_IIIC	0.13	1.14	0.03	0.06	0.20	1.06	1.22	3.75	<0.005	12.47
TNM_IVA	0.36	1.43	0.03	0.30	0.42	1.35	1.53	11.09	<0.005	92.51
TNM_IVB	0.37	1.45	0.11	0.16	0.59	1.17	1.80	3.40	<0.005	10.53