2025

Time to event cancer recurrence project

NIKOL MARTINOVA STAYKOVA

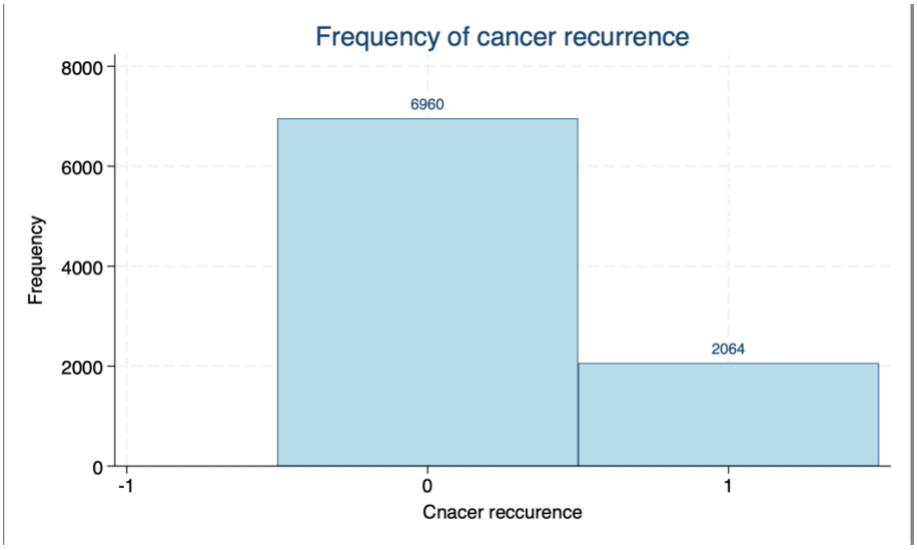
Getting to know the data

Before I started comparing the survival curves of the people on therapy vs the not on therapy, I produced a few graphs of the whole data to get a better understanding of what are the relationships in it. The first thing that I did was to declare that the data is a survival-time data, the time variable I set to be **rectime** (representing the duration until cancer recurrence) and the failure variable is **censrec** (1 means there is a recurrence 0 means there is no recurrence).

After declaring the data to be a survival data we can see that there are 9,024 observations in the data and 2064 failures (there is a recurrence for 2064 people). We can't say for sure that the rest of the people didn't get their cancer back as they could have simply left the study or died before the event happened.

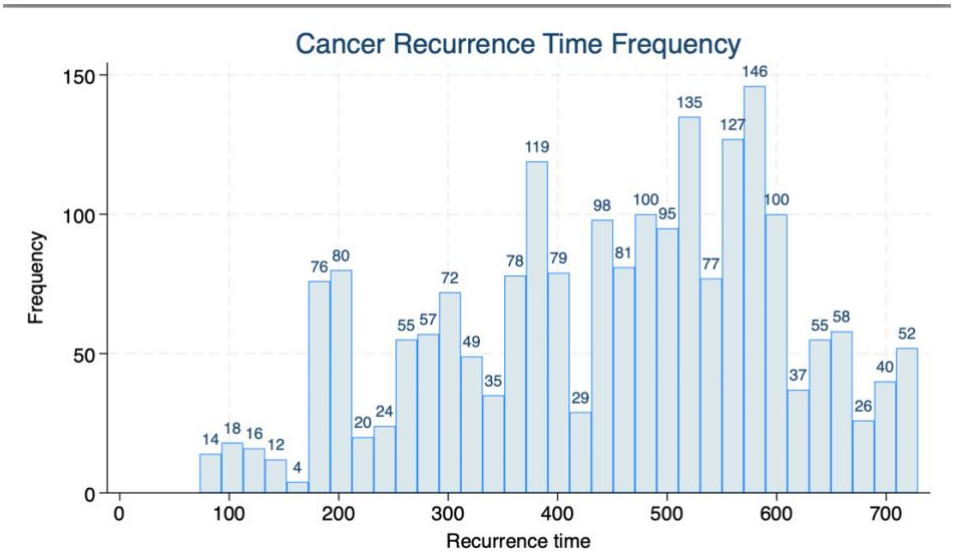
total observations	9,024
exclusions	0
failures in single-record data	2,064

I decided to also produce a histogram to see the cancer recurrence in a graphical way (Graph 1). From the histogram we can see the same result as when we declared the data to be survival data. The Freuquency of failure (the number of times when censrec was 1) is 2064 and the number of people who didn't experience recurrence is 6960.



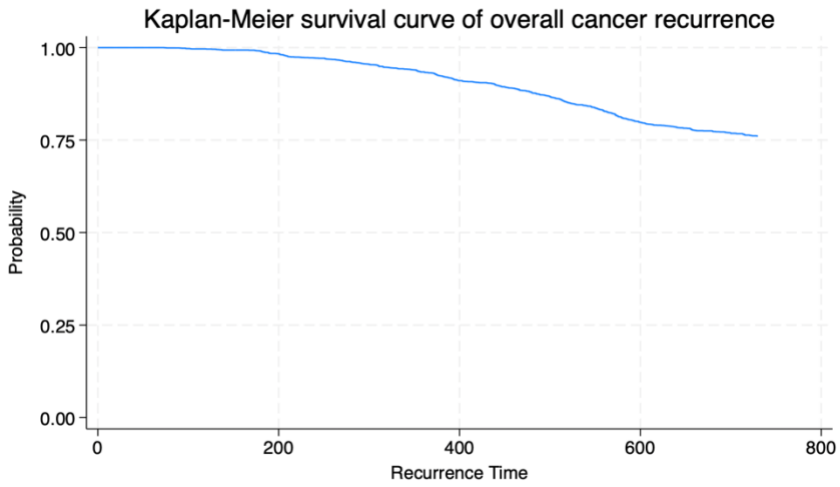
Graph 1

I also decided to produce a histogram of the recurrence time when the cancer has returned to see when it is most likely for the event to happen (graph 2). From this histogram we can make the insight that the cancer is most likely to return in the period of 500-600 days as the amount of recurrences is the biggest reaching its peak of 146 events at around 580 days.



Graph 2

The last thing I did before I started comparing the people on hormone therapy vs the people not on it was to produce a Kaplan-Meier survival curve for the overall cancer recurrence (Graph 3). From it we can see that the probability of not experiencing a recurrence drops to 75% by the time we reach the 720th day.

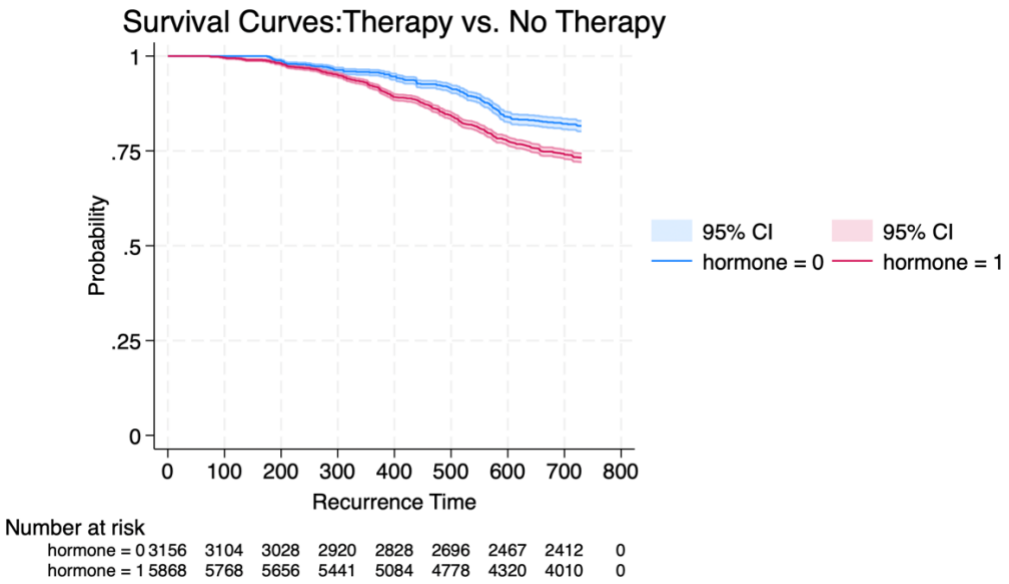


Graph 3

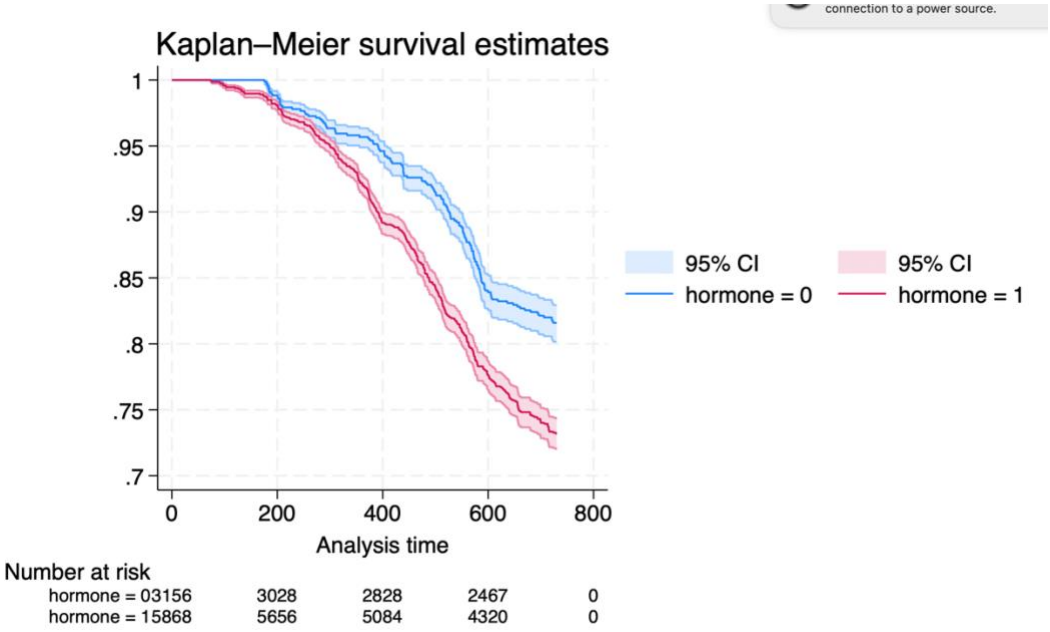
Comparing the people on/not-on therapies

In order to compare if there is a difference between the recurrence time of people who are on the therapies vs the people who are not, I produced a Kaplan-Meier survival curves to see the result graphically (Graph 4). I also included the confidence intervals.

From the graph we can see that the probability of not experiencing recurrence for people who do hormone therapies drops significantly by the 720th day. I included a more zoomed in graph (Graph 5) so we can see this in more details as well. From the more detailed graph we can see that for the people who do hormone therapies the probability of not experiencing recurrence drops below 75% by the end of the study (there is around 73% to not get your cancer back after the 720th day), while for those who don't do hormone therapy the probability of not experiencing occurrence stays around 83% (there is around 83% your cancer not to come back after the 720th day). This indicates that doing the hormone therapy may not be beneficial.



Graph 4



Graph 5

To test the difference we run a log rank test and we get the result:

Hormone	Observed events	Expected events

0	548	741.00
1	1516	1323.00
Total	2064	2064.00
		chi2(1) = 78.51 Pr>chi2 = 0.000

H0: there is no difference between the people using hormone therapy vs the people not using hormone therapy

Ha: there is difference.

The p-value of **7.977e-19** is vanishingly small (essentially zero) <**0.05**, which provides very strong statistical evidence against the null hypothesis. Combining the visual evidence from the survival curves with this extremely significant p-value, the data strongly suggests that hormone therapy in this dataset is associated with a higher risk of cancer recurrence and worse survival outcomes.

Incidence e Rate

By reporting incidence-rate comparison I get that that the number of failures for the people exposed the hormone therapy vs the people who were not exposed to it we found:

- Among individuals exposed to hormone therapy, there were **1,516** failures. The total time at risk for this group was **3,725,627** person-days.
- Among individuals not exposed to hormone therapy, there were **548** failures. The total time at risk for this group was **2,058,491** person-days.

The incidence rate for the group exposed to hormone therapy is:

1516/ 3725627= 0.00040 events per day

For the group not exposed to hormone therapy, the incidence rate is:

548/ 2058491= 0.00027 events per day

Thus, the overall combined incidence rate for both groups is:

2064/ 5784118= .00036 events per day

The **incidence rate ratio** is calculated by dividing the incidence rate of the exposed group by the incidence rate of the non-exposed group. In this case, I obtained an incidence rate of **1.53**. In simple words compared to those not using hormone therapy, individuals using hormone therapy see the event happening about 1.5 times more often. A comparison of incidence rates is limited because it assumes a constant rate of cancer occurrence over time, ignoring potential variations in risk at different time points.

Cox proportional hazards model

I computed a Cox proportional hazard model with hormone as the only risk factor and I got the following result:

Variable	**Hazard Ratio (HR)**	**Std. Error**	**z-value**	**p-value**	**95% Confidence Interval (CI)**
Hormone Therapy	**1.55**	0.077	8.79	<0.001	(1.41 – 1.71)

A hazard ratio of **1.55** means that the hazard (the likelihood of cancer recurrence at a particular moment) is **55%** higher in the hormone therapy group than in the non-exposed group at any given time. The p-value less than **0.001** indicates that the result is highly statistically significant

I tested the **proportional hazards assumption**. I used the test to see whether the effect of hormone therapy changes over time and got the following result:

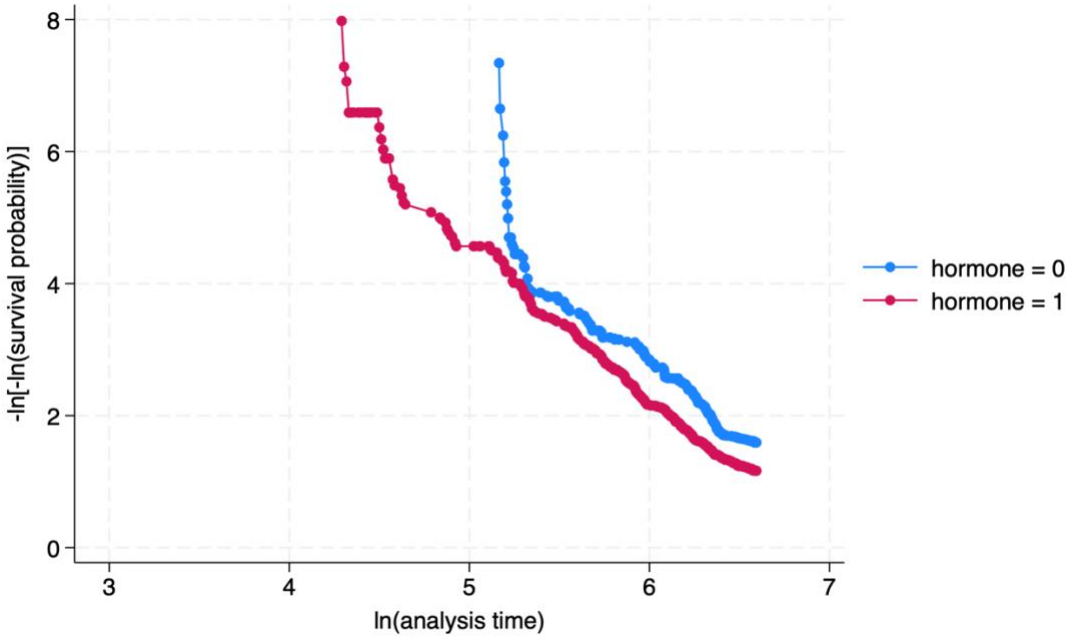
Test of proportional-hazards assumption

	Chi2	df	Prob>chi2
Global test	4.94	1	0.0262

H0: the hazard ratio remains constant over time

Ha: the hazard ratio changes over time

My p-value <**0.05** which means we reject the null hipothesis However, my log-minus-log survival plot(Graph 6) has a slight curvature which indicated that the hazard ratio may change slightly over time. This means that the relationship between the hazard rate and hormone is not constant over time, and the assumption of proportional hazards does not hold for this covariate.



Graph 6

Commands used in stata to produce the result

stset rectime, failure(censrec==1) scale(1)

Survival-time data settings

Failure event: censrec==1

Observed time interval: (0, rectime]

Exit on or before: failure

9,024 total observations

0 exclusions

9,024 observations remaining, representing

2,064 failures in single-record/single-failure data

5,784,118 total analysis time at risk and under observation

At risk from t = 0

Earliest observed entry t = 0

Last observed exit t = 730

histogram censrec, discrete frequency fcolor(ltblue) lcolor(navy) addlabel addlabopts(mlabcolor(navy))

> ytitle("Frequency") xtitle("Cnacer reccurence") xlabel(minmax) title("Cancer Recurrence Time Frequency"

> , color(navy))

(start=0, width=1)

. histogram rectime if censrec==1, frequency fcolor(ltblue) lcolor(navy) addlabel ytitle("Frequency") xt

> itle("Recurrence time") title("Cancer recurence time frequency")

(bin=33, start=73, width=19.878788)

sts graph, by(hormone) ci risktable risktable(, size(small)) ytitle("Probability") xtitle("Recurrence > Time") title("Survival Curves:Therapy vs. No Therapy")

sts graph, by(hormone) ci risktable risktable(, size(small)) ylabel(#9)

sts test hormone, logrank

```
      Failure _d: censrec==1
Analysis time _t: rectime

Equality of survivor functions
Log-rank test

       | Observed       Expected
hormone |    events         events
--------+-------------------------
      0 |       548         741.00
      1 |      1516        1323.00
--------+-------------------------
  Total |      2064        2064.00

                chi2(1) =  78.51
                Pr>chi2 = 0.0000
```

stir hormone

```
      Failure _d: censrec==1
```

Analysis time _t: rectime

Incidence-rate comparison

Exposed: hormone = 1

Unexposed: hormone = 0

```
                |       hormone        |
                |  Exposed   Unexposed |      Total
----------------+----------------------+-----------
       Failures |      1516        548 |       2064
           Time |   3725627    2058491 |    5784118
----------------+----------------------+-----------
                |                      |
 Incidence rate |  .0004069    .0002662 |   .0003568
                |                      |
                |     Point estimate   |    [95% conf. interval]
                |----------------------+-----------------------
 Inc. rate diff. |        .0001407      |    .0001104     .000171
 Inc. rate ratio |         1.52851      |    1.385325    1.688453 (exact)
 Attr. frac. ex. |         .345768      |    .2781475    .4077419 (exact)
 Attr. frac. pop |         .2539653     |
                +-------------------------------------------------
```

Mid-p-values for tests of incidence-rate difference:

Adj Pr(Exposed failures <= 1516) = 1.0000 (lower one-sided)

Adj Pr(Exposed failures >= 1516) = 0.0000 (upper one-sided)

Two-sided p-value = 0.0000

stcox hormone

```
      Failure _d: censrec==1
 Analysis time _t: rectime

Iteration 0:  Log likelihood = -18441.795
Iteration 1:  Log likelihood = -18400.894
Iteration 2:  Log likelihood = -18400.692
Iteration 3:  Log likelihood = -18400.692
Refining estimates:
Iteration 0:  Log likelihood = -18400.692

Cox regression with Breslow method for ties

No. of subjects =     9,024                       Number of obs =  9,024
No. of failures =     2,064
Time at risk    = 5,784,118
                                                  LR chi2(1)    =  82.21
Log likelihood = -18400.692                       Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     hormone |   1.549601    .077247     8.79   0.000     1.405361    1.708646
------------------------------------------------------------------------------
```

estat phtest

```
Test of proportional-hazards assumption

Time function: Analysis time
```

```
-----------------------------------------------
            |    chi2       df       Prob>chi2
------------+----------------------------------
 Global test |     4.94        1          0.0262
-----------------------------------------------
```

stphplot, by(hormone)