Survival Analysis of Breast Cancer Data from the TCGA Dataset

Ramaa Nathan

6/25/2019

Contents

1 Introduction

Survival Analysis is a branch of statistics to study the expected duration of time until one or more events occur, such as death in biological systems, failure in meachanical systems, loan performance in economic systems, time to retirement, time to finding a job in etc. In case of cancer studies, one of the primary objectives is to assess the time to an event of interest like relapse of cancer, death, etc.

Survival Analysis is especially helpful in analyzing these studies when one or more of the cohorts do not experience the event and are considered censored for various reasons like death due to a different cause,

loss-to-follow-up, end of study, etc. The basic quantity used to describe time-to-event data is the survival function which is the probability of surviving beyond time x.

The survival function can be modeled using parametric methods (Exponential, Weibull, etc), semi-parametric methods (Cox proportional hazards model), and non-parametric methods (Kaplan Meier model). The difference in survival times due to different treatment groups can also be compared using Logrank tests. The Cox proportional hazards model is useful in modeling the survival function in the presence of covariates.

The Cancer Genome Atlas (TCGA) Program, a joint effort between the National Cancer Institute and the Human Genome Research Institute, provides publicly-available clinical and high-throughput genomic data for thirty-three different types of cancers. This rich data source is widely used by researchers and has led to vast improvements in diagnosing, treating, and preventing cancer. The TCGA dataset will be used in this study to do survival analysis of breast cancer data.

2 Research Questions

Here are a few questions that we could answer with this study.

1. What is the probability of survival for breast cancer?
2. How do different cancers (breast, ovarian, lung) affect survival rates?
3. What are the important factors that influence estimation of survival rate for Breast Cancer?
4. What are the effects of each factor on survival?

5. What is the probability of survival for breast cancer when other clinical covariates are considered?

3 Survival Analysis Basics

Before we start analyzing the data, lets first try to understand the basic terminologies related to survival analysis.

3.1 Terminologies

3.1.1 Event Times

Event times often are useful endpoints in clinical trials. Examples include survival time from onset of diagnosis, time until progression from one stage of disease to another, and time from surgery until hospital discharge. In each case, time is measured from study entry until the event occurs.

3.1.2 Censoring

With an endpoint that is based on an event time, there always is the chance of censoring. An event time is censored if there is some amount of follow-up on a subject and the event is not observed during the study period. There are different types of censoring (Lawrence M. Friedman 2015)

3.1.3 Survival Function

The survival function is the probability of surviving beyond time t or the probability of experiencing the event beyond time t. The survival function takes value 1 at the origin and 0 at infinity.

$$S(t) = P(T > t)$$

If f(t) is the probability density function (pdf) that describes the time-to-event, and F(t) is the corresponding cumulative distributive function, then

$$F(t) = \int_0^t f(x)\ dx \tag{1}$$

$$S(t) = 1 - F(t) = \int_t^\infty f(x)\ dx \tag{2}$$

3.1.4 Hazard Rate or Hazard Function

The hazard function, h(x) is defined as the instantaneous risk of the event or the probability that if a person survives to t, they will experience the event in the next instant. The hazard function or hazard rate can be considered to be the slope of the survival function.

$$h(t) = \lim_{\Delta t -> 0} \frac{P[t \leq T < t + \Delta\, t | T \geq t]}{\Delta t}$$

The hazard function can also be expressed as the ratio of the probability density function to the survival function

$$h(t) = \frac{f(t)}{S(t)}$$

3.1.5 Cumulative hazard

Cumulative hazard is the accumulation of hazard rate over time. It is not a probability and is a measure of the risk. The greater the value of H(t), the greater the risk of failure by time t. When the distribution is continuous, the cumulative hazard is the integral of the hazard rate.

$$H(t) = \int_0^t h(x)dx = -log(S(t))$$

3.1.6 Hazard Ratio

Hazard ratio is defined as the ratio of of two hazard functions, corresponding to two treatment groups

$$\Lambda = \frac{h_1(t)}{h_2(t)}$$

3.1.7 Proportional Hazard

The hazard ratio can be used to compare two treatment groups. When the hazard ratio is constant and independent of time, the hazards for the two treatment groups are said to be proportional. Basically, the relative risk of the event is constant over time.

3.1.8 Non-parametric Kaplan Meier Model

The Kaplan Meier survival curve is a non-parametric teachnique for estimating the probability of survival, even in the presence of censoring. In this model, there is the notion of a risk set, which is the set of all individuals who are at risk to have an event at time t. This includes individuals who are known to be alive at time t and those who have the actual event at time t.

In the modeling process, the actual failure times (event times) are first ordered in an increasing order. At each event time t_k, the number of subjects still at risk (n_k), the number of events (d_k), number of censored (lost to follow-up) subjects since the last event time ($n_k - d_k$) are recorded. The risk set does not include the

subjects lost to follow-up. The Kaplan Meier Survival probability at time t_k utilizes conditional probability - the probability of surviving at time t, given that the person has survived upto time t.

$$S(t) = \begin{cases} 1 & t_0 \leq t \leq t_1 \\ \prod_{i=1}^{i=k}(\frac{n_i - d_i}{n_i}) & t_k \leq t \leq t_{k+1}, k = 1, 2, ..., K \end{cases}$$

Tha basic assumptions in a Kaplan Meier model are:

```
1. The censoring is independent of prognosis
2. The survival probabilities are the same for all subjects recruited at any time in the study
3. Events happened at the time specified.
```

3.1.9 Parametric Exponential Distribution

$$pdf : f(t) = \lambda e^{-\lambda t} \tag{3}$$

$$cdf : F(t) = 1 - e^{-\lambda t} \tag{4}$$

$$Survival\ Function : S(t) = 1 - F(t) = e^{-\lambda t} \tag{5}$$

$$Hazard\ Function : h(t) = \frac{f(t)}{S(t)} = \lambda \tag{6}$$

$$Hazard\ Ratio = \frac{\lambda_1}{\lambda_2} \tag{7}$$

$$Cumulative\ Hazard : H(t) = -logS(t) = \lambda t \tag{8}$$

In the case of an exponential distribution, the hazard function or the hazard rate is a constant λ and the hazard ratio is proportional as it is independent of time.

3.1.10 Semi-Parametric Cox Proportional Hazard Regression Model

The Cox Regression Model is used to model the hazard at time t in the presence of multiple covariates, each of which could be categorical or quantitative. The Cox model is similar to the exponential model where the survival time is given by $S(t) = e^{\lambda t)}$. But the hazard rate λ is now considered to be a linear combination of several covariates $Z = Z_1, Z_2, .., Z_p$ and so

$$\lambda(Z_1, Z_2, .., Z_p) = \beta_1 Z_1 + \beta_2 Z_2 + .. + \beta_p Z_p$$

The Cox regression model can then be expressed as

$$h(t|Z) = h_0(t) exp(\Sigma_{k=1}^{p} \beta_k Z_k)$$

where $h(t|Z)$ is the hazard at time t for an individual with covariates Z and $h_0(t) = h(t|Z = 0)$ is the baseline hazard rate.

The Cox model is semi-parametric, containing both parametric and non-parametric components.

1. The $h_0(t)$ is the non-parametric component and can take any form as along as $h_0(t) \geq 0$
2. $exp(\Sigma_{k=1}^{p} \beta_k Z_k)$ is the parametric component.

For example, suppose we have three predictors - therapy, age and race. Then

$$h(t) = h_0(t) exp(\beta_1 therapy + \beta_2 age + \beta_3 race)$$

So, the coefficient β_1 is the log of the hazard ratio and $exp(\beta_1)$ is the hazard ratio for the individual on treatment B compared to treatment A, when the race and age covariates are the same for both individuals. The values of $exp(\beta 1)$ provide the following interpretations,

- $exp(\beta_1) > 1$ indicates higher hazard or lower survival rate compared to the base hazard function.
- $exp(\beta_1) < 1$ indicates lower hazard or higher survival rate
- $exp(\beta_1) = 1$ indicates no association

In R, we use the coxph function to model the Cox PH model. It provides the following output values:

1. coef = the estimate of β_i
2. exp(coef) - the estimate of $exp(\beta_i)$
3. se(coef) - the standard error of the estimate of β_i
4. z = $\frac{z}{se(coef)}$ = the Wald statistic for testing the null hypothesis that $\beta_i = 0$ assuming that z follows a standard normal distribution
5. p = two sided p-value

One of the basic assumptions of the Cox Proportional Hazards Model is that the hazards are proportional. This can be tested by checking that the Schoenfeld residuals exhibit a random pattern against time. In R, cox.zph function in the survival package and ggcoxzph from the survminer package can be used to check for the plot of these residuals.

4 R Functions for Survival Analysis

The survival analysis in this study has been done in R and here is a list of the important functions used in the analysis.

```
## Warning in readLines(knitr::current_input()): incomplete final line found
## on 'Survival_Analysis_mini.Rmd'
```

R Functions for Survival Analysis

Purpose	**Package::Function**	**Package::Graphical Wrapper**
Extract Survival Data from TCGA	rtcga::survivalTCGA	
Create Survival Object	survival::Surv()	
Fit a Kaplan Meier Curve	survival::survfit	survminer::ggsurvplot()
Compare Kaplan Meier Curves using logrank	survival::survdiff()	survminer::ggsurvplot()
Fit a parametric model	survival::survreg	
Fit Cox Proportional Hazards Model	survival::coxph()	survminer::ggforest()
Test for Proportional Hazards	survival::cox.zph()	survminer::ggcoxzph()
Display Adjusted Survival Curves for Cox Proportional Hazards Model for a Factor		survminer::ggcoxadjustedcurves
Split the survival data set at specific cut times to accommodate the time-dependent covaraites for Cox	survival::survSplit	
Convert Weibull results to an easy interpretable form	SurvRegCensCov::ConvertWeibull	

5 Data Source and Description

The Cancer Genome Atlas (TCGA) Program [1] provides publicly-available clinical and high-throughput genomic data for thirty-three different types of cancers. For this study of survival analysis of Breast Cancer, we use the Breast Cancer (BRCA) clinical data that is readily available as BRCA.clinical. This dataset has 3703 columns from which we pick the following columns containing demographic and cancer stage information as important predictors of survival analysis.

```
## Warning in readLines(knitr::current_input()): incomplete final line found
## on 'Survival_Analysis_mini.Rmd'
```

ColumnName	DataType	Description
Gender	categorical	Gender
Race	categorical	Race
Ethnicity	categorical	Ethnicity
Age	integer	Age at first diagnosis
Vital Status	binary	Vital Status (1 - dead (event), 0 - alive/censored)
Days to Death	integer	Number of days to death from first diagnosis
Days to Followup	integer	Number of days to last follow-up from first diagnosis
Therapy type	categorical	Therapy Type (Chemo, Harmone, Immuno, etc.
Pathologic Stage	categorical	Cancer stage - based on T,M, and N labeling
Pathology T	categorical	Tumor (T) stage describing size and location of tumor
Pathology N	categorical	Lymph (N) nodes status describing if cancer has spread into nearby lymph nodes
Pathology M	categorical	Metastasis (M) status describing if cancer has spread to other parts of the body

6 Exploratory Data Analysis

The clinical data set from the The Cancer Genome Atlas (TCGA) Program is a snapshot of the data from 2015-11-01 and is used here for studying survival analysis.

6.1 Data Extraction

The RTCGA package in R is used for extracting the clinical data for the Breast Invasive Carcinoma Clinical Data (BRCA). In addition, the survival and survminer packages in R are used for the analysis.

The survivalTCGA function in the RTCGA package is used to extract the relevant columns. This function also uses the vital status variable that indicates if the observation was an event or a censor and combines the number of days to death from first diagnosis and number of days to last follow-up from first diagnosis into a new "times" variable.

6.2 Checks for Missing and Invalid Data

The extracted data is first checked for any missing data or invalid data. There are two observations that have negative values for the time to event that are filtered out during the data cleanup and transformation step.

The data is generally clean with only some of the demographic information like race and ethnicity missing for a few of the observations. We also find that more than 40% of the rows have NAs in one or more columns. So, we will filter out the missing data, as needed during the analysis.

6.3 Data Cleaning and Transformation

Next, the following cleaning and transformations are applied to the TCCGA BRCA.clinical data to get a clean and compact dataset:

1. Rename the long variable names to short names.
2. Filter out the 12 observations corresponding to males diagnosed with breast cancer.
3. Filter out the 2 observations with negative "times" value.
4. Create a "age" variable that contains the number of days at first diagnosis to age in years at first diagnosis.
5. Create a "years_to_event" variable which is the "times"" variable converted from days to years.
6. Data in the pathology columns contain information on both stage and sub-stage. Transform the data to only contain the high level stage information.
7. Modify the "therapy_type"" to contain three types - chemotherapy, hormone therapy and Other (lump all the other infrequent types into Other)

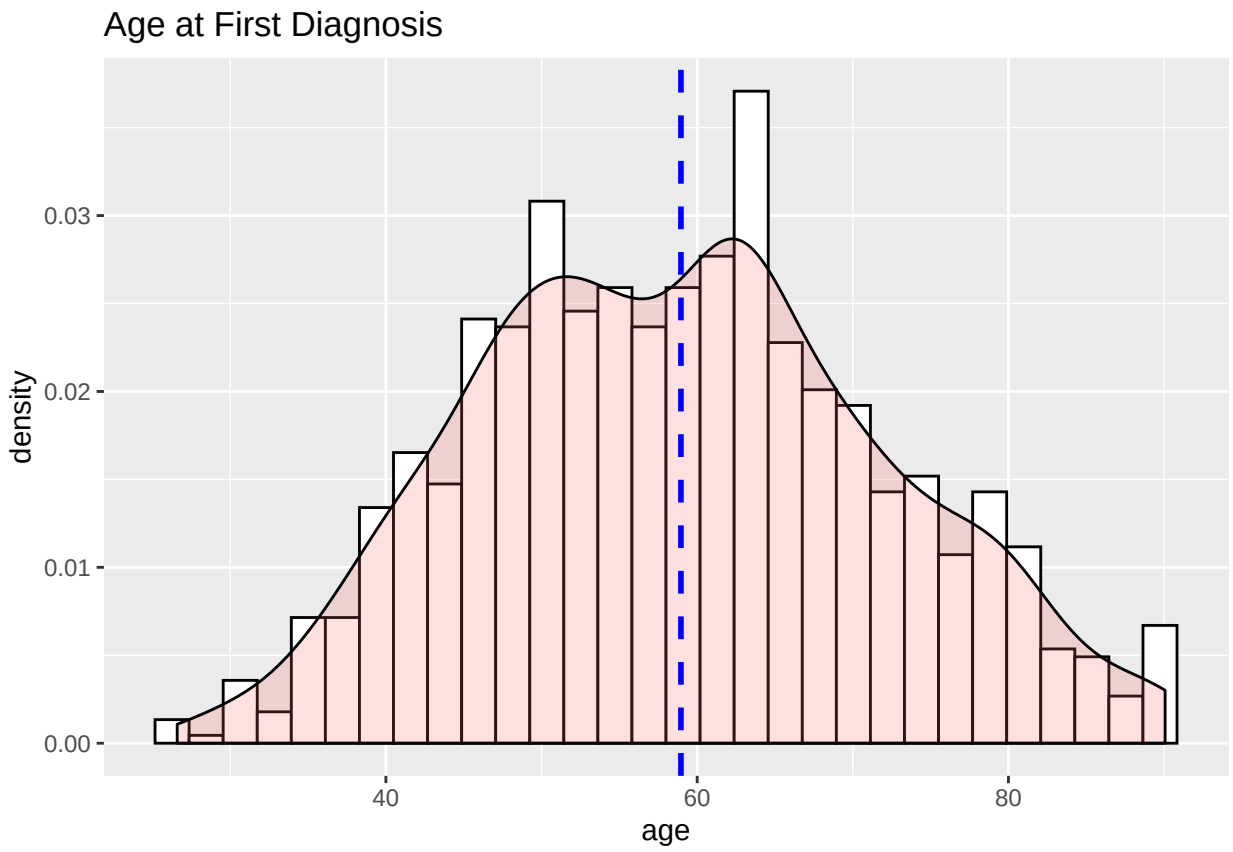
8. Modify the “race”" to contain three types - black or african american and white (lump the other two types into Other)

6.4 Data Visualizations

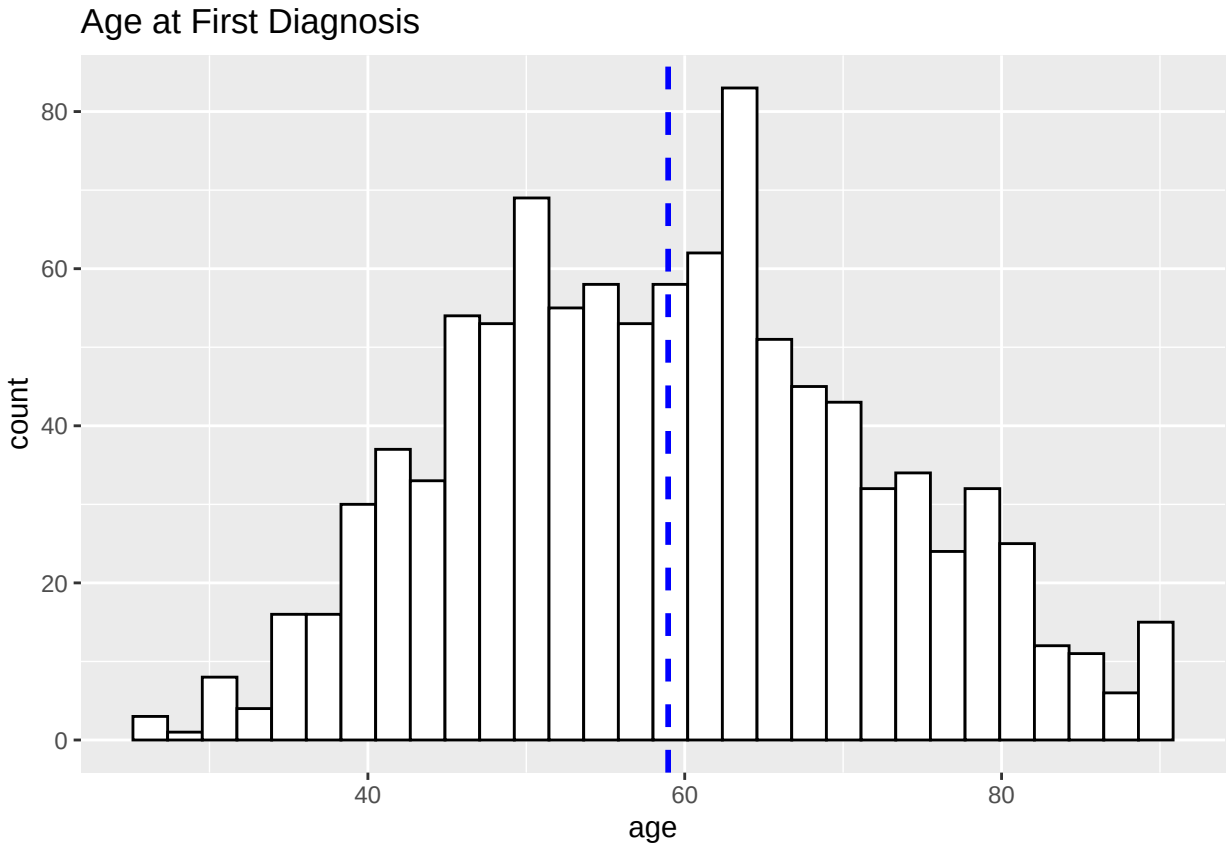
6.4.1 Histograms

Age at First Diagnosis

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



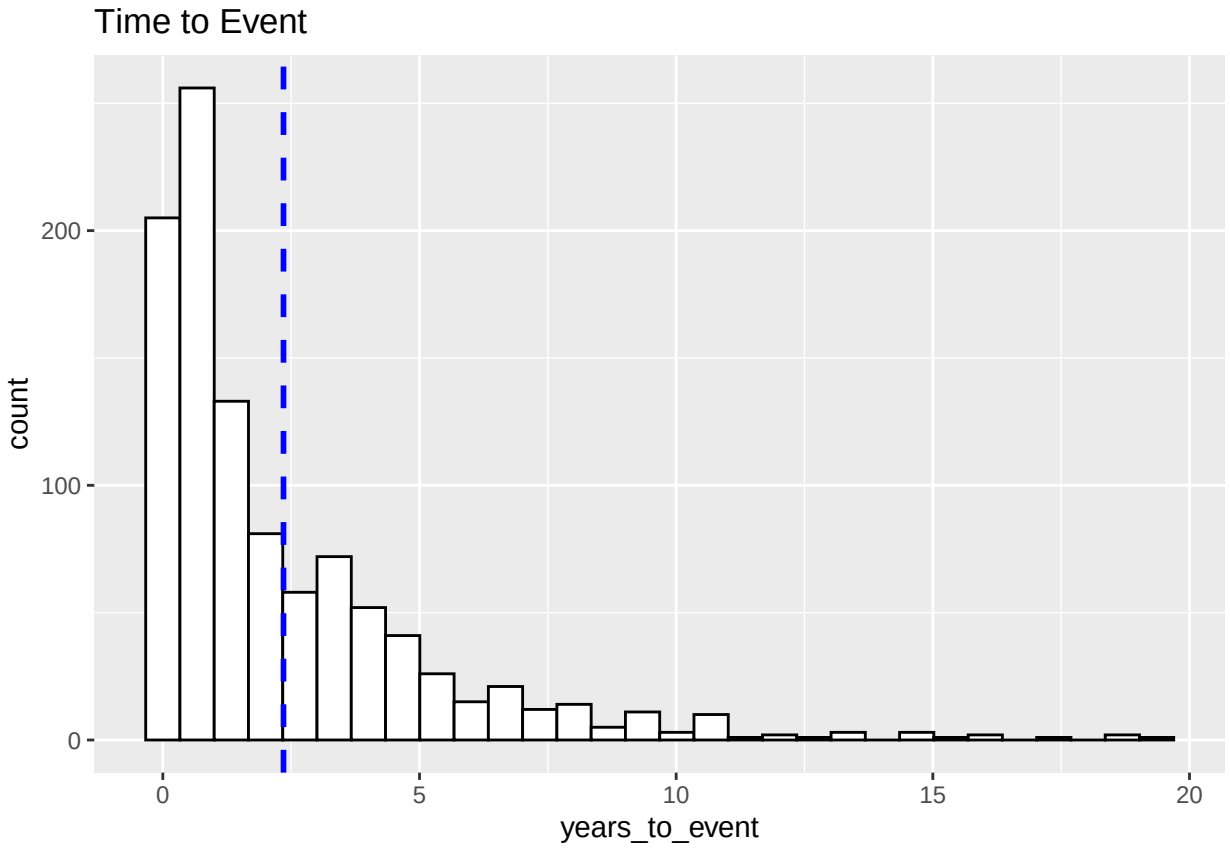

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The average age at first diagnosis seems to 59 years and the distribution of age is mostly symmetrical with a small bimodal effect.

Time to Event

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

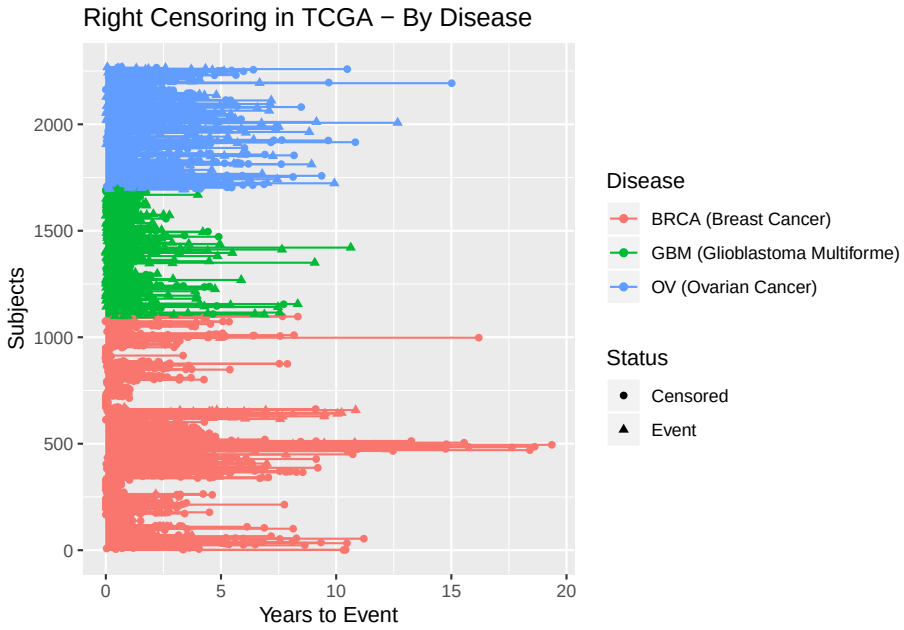



```
##       age
##  Min.   :26.59
##  1st Qu.:49.28
##  Median :58.94
##  Mean   :58.95
##  3rd Qu.:68.00
##  Max.   :90.06
##  NA's   :9
```

Without considering censoring the mean survival time seems to be less than 2.5 years. This is not helpful information as censoring information is an important component of survival data. The TCGA clinical data set is a survival dataset containing right censor information. We will first visualize the distribution of the right censor data by different categories. The Censoring Event plots here were inspired by a workshop on Survival Analysis.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
therapy_type	3	22770.15	7590.051	50.06829	0
Residuals	1019	154474.27	151.594	NA	NA

6.4.2 Censoring and Event Plots by Disease

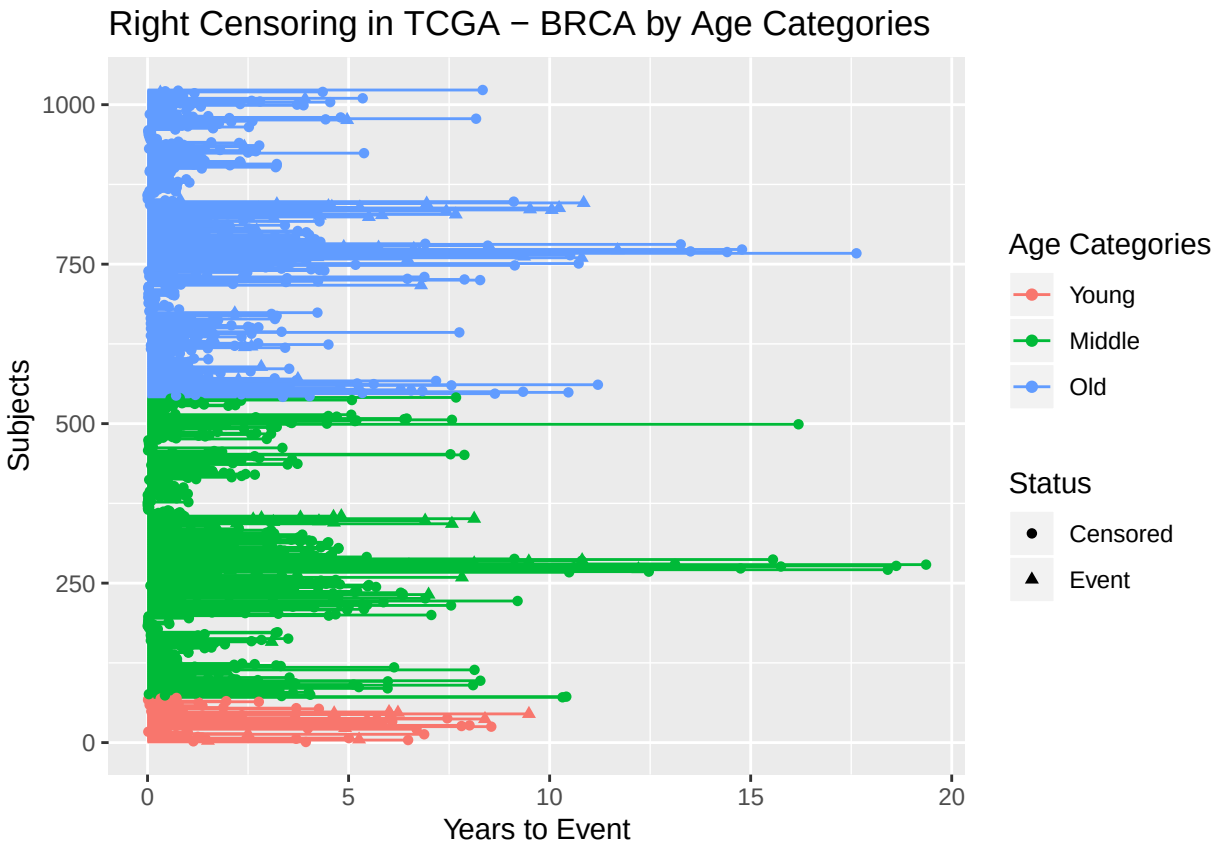


	0	1	Sum
brca	994	104	1098
gbm	149	446	595
ov	279	297	576
Sum	1422	847	2269

From the contingency table and the plots comparing the diseases of BRCA (breast cancer), GBM (Glioblastoma Multiforme) and OV (Ovarian Cancer), it can be observed that less that almost 50% of the cases are for Breast Cancer and the rest are almost equally split between ovarian and GBM. In these, 10% of the subjects with Breast cancer, 75% of the subjects with GBM and 50% of the subjects with ovarian cancer did not survive (had events).

6.4.3 Censoring and Event Plots by Age Category

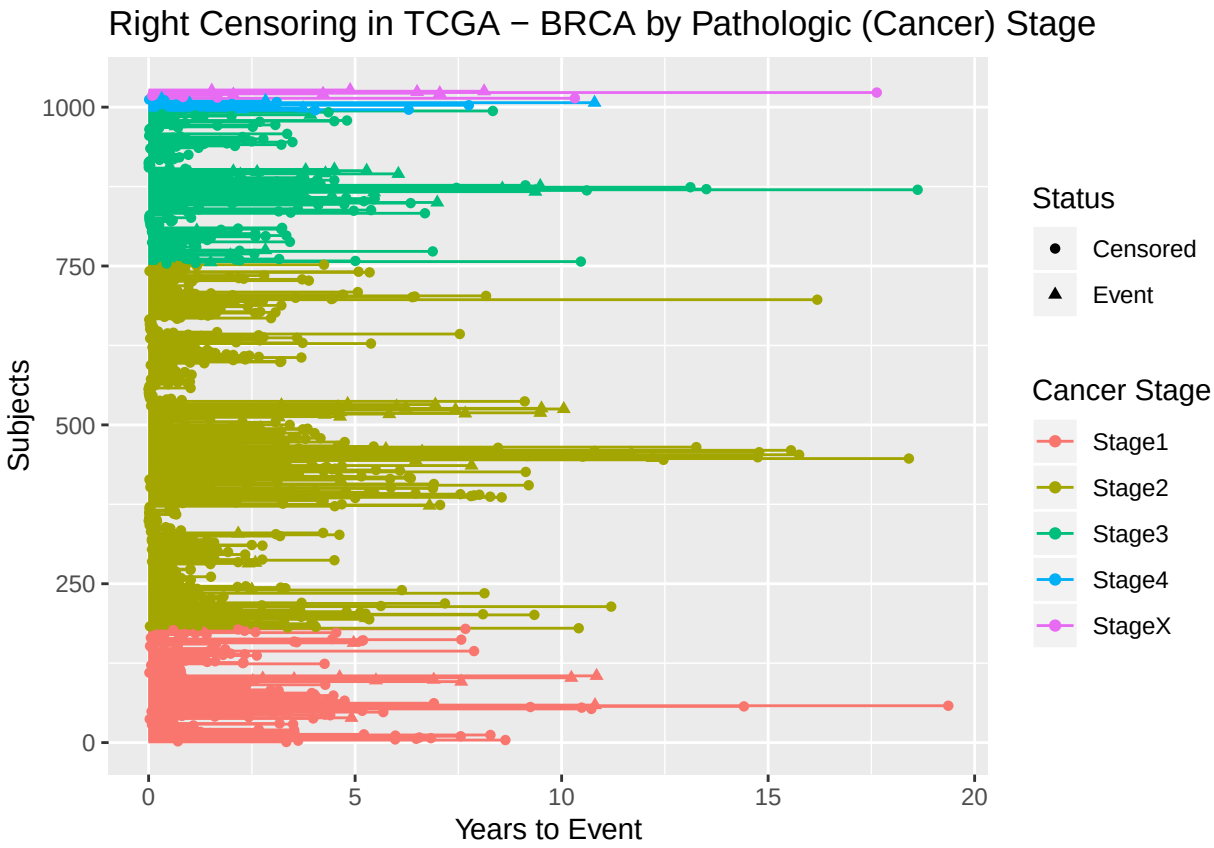
```
##          vital_status
## agecat       0    1  Sum
##   young     59   11   70
##   middle   436   35  471
##   old      424   58  482
##   Sum      919  104 1023
```



From the contingency table and the plots, it can be observed that less than 10% of the subjects are less than 40 years old. The number of subjects in the other two groups are almost equally distributed of which more number of events were observed in the older age group.

6.4.4 Censoring and Event Plots by Pathological Stage

```
##                  vital_status
## pathologic_stage    0    1  Sum
##           stage1  166   13  179
##           stage2  530   44  574
##           stage3  211   30  241
##           stage4   10    9   19
##           stageX    7    7   14
##           Sum     924  103 1027
```



From the contingency table and the censor plots, it can be observed that almost 50% of the subjects are diagnosed with stage 2 cancer, with less than 2% of the subjects are either in stage 4 or in stage X. The percentage of events is the highest (almost 50%) for the stage 4 and stage X groups and the lowest for stage 1.

7 Answers to Research Questions:

1. What is the median survival time for breast cancer?
   - In the absence of any predictors, the median survival time is 9.5 years.
2. How do different cancers (breast, ovarian, lung) affect survival rates?
   - Of the three types of cancers BRCA (breast cancer), GBM (Glioblastoma Multiforme) and OV (Ovarian Cancer) that were compared, BRCA has the highest survival rate.
3. What are the important factors that influence risk of death for Breast Cancer?
   - When considered individually or together age at first diagnosis, no therapy, cancer stage 3 and cancer stage 4, and any metastasis into the lymph nodes significantly reduce survival time.
4. What are the effects of each factor on survival?
   - The number of lymph nodes that have been metasized into seems to be the strongest predictor of survival.
5. What is the probability of survival for breast cancer when other clinical covariates are considered?
   - The results do not change when the predictors are all accounted for indicating an additive model with no interactions.

References

Campus, PSU World. n.d. “STAT 509: Design and Analysis of Clinical Trials.” https://newonlinecourses.science.psu.edu/stat509/.

Lawrence M. Friedman, et al. 2015. *Fundamentals of Clinical Trials*. Springer.

M J Bradburn, S B Love, T G Clark. n.d. "Survival Analysis Part Ii: Multivariate Data Analysis – an Introduction to Concepts and Methods." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394368/.

National Cancer Institute. n.d. "The Cancer Genome Atlas Program." https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

STHDA. n.d. "Guide to Survminer 0.3.0." http://www.sthda.com/english/wiki/survminer-0-3-0.

Sweeney, Elizabeth. n.d. "New York R Survival Analysis Workshop." https://github.com/emsweene/New_York_R_Survival_Analysis_Workshop.

Zhang, Zhongheng. n.d. "Parametric Regression Model for Survival Data: Weibull Regression Model as an Example." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5233524/.