

Lesson 8

Survival Analysis

BIS 505b

Yale University
Department of Biostatistics

Pagano Chapter 21

Date Modified: 04/16/2021

Goals for this Lesson

Addressing a Research Question

- ① How to recognize **time-to-event data**
- ② How to summarize survival data and estimate **survival probabilities**
- ③ How to **compare survival experiences between groups**
- ④ How to investigate the **association of risk factors** with a **time-to-event outcome**

Contents

1 Survival Data

- Motivation for Survival Analysis
- Components of Survival Data
- Terminology and Notation

2 Estimating Survivor Function

- Data Structure
- Kaplan-Meier Method

3 Group Comparisons and Modeling

- Log-Rank Test
- Cox Proportional Hazards Model
- Cox Adjusted Survival Curves

Progress this Unit

- 1 Survival Data
 - Motivation for Survival Analysis
 - Components of Survival Data
 - Terminology and Notation
- 2 Estimating Survivor Function
 - Data Structure
 - Kaplan-Meier Method
- 3 Group Comparisons and Modeling
 - Log-Rank Test
 - Cox Proportional Hazards Model
 - Cox Adjusted Survival Curves

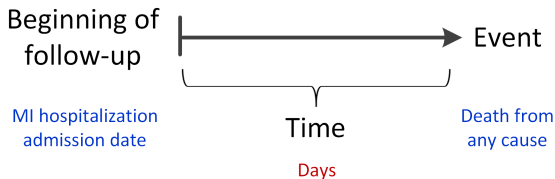
Traditional Methods

- The type of **outcome variable** analyzed (Y) drives the method of analysis

Outcome	Technique	Mathematical Model	Yields
Continuous	Linear Regression	$\mu = \alpha + \beta x$	Mean Difference
Binary	Logistic Regression	$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$	Odds Ratio
Count	Poisson Regression	$\log(\mu) = \alpha + \beta x$	Mean Ratio
Rate		$\log\left(\frac{\mu}{t}\right) = \alpha + \beta x$	Rate Ratio

Survival Analysis

- In **survival analysis**, the outcome variable of interest is **time** until a target **event** of interest occurs



- Worcester Heart Attack Study (WHAS)
 - Study factors associated with long-term survival following acute MI
 - Effect of sex, age, BMI at time of hospitalization for MI on length of survival

What are Survival Data?

- Survival data consist of two pieces:

1. **Time:** Time (days, weeks, months, years) from a distinct start point until the event of interest occurs
2. Occurrence of the **event/endpoint**

Often called **survival time** because it gives the length of time that an individual has “survived” or been event-free over follow-up period

Often called a **failure** because the event of interest is usually death, disease incidence, negative experience

Determining Survival Time

- To determine the **survival time** T , must precisely define:
 1. Time origin or starting point
 2. Ending event of interest
 3. Measurement scale for passage of time
- For example, life span T from birth (**time origin**) to death (**ending event**) in years (**measurement scale**)

Time Origin

Date of birth
Start date of new treatment (rand date)
Hospital admission date
Nursing home admission date

Ending Event

Death
Relapse of disease (e.g., leukemia)
Discharge to the community from hospital or nursing home

- **Survival time** = Distance on the time scale between these two points

Motivation for Survival Analysis

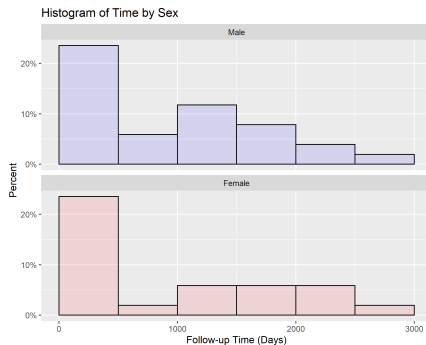
- If the objective is to compare the **proportion** experiencing the event between two or more groups, then why not compute an odds ratio, perform a **chi-square test** or use **logistic regression**?

	Died	Alive	Total
Female	23 (66%)	12	35
Male	28 (43%)	37	65

- $\widehat{RR} = \frac{23}{35} / \frac{28}{65} = \frac{0.66}{0.43} = 1.5$
- $\widehat{OR} = (23 \times 37) / (12 \times 28) = 2.5$
- Ignores time**; does not account for varying follow-up time

Motivation for Survival Analysis

- If the objective is to compare survival times (quantitative) between two or more groups, then why not compare **mean** survival time to event using a ***t*-test** or **linear regression**?



- Survival times tend to be positively (right) skewed (non-Normal)
- Looking at the time measurement alone does not properly account for those **not experiencing event**

Survival Data Notation

Two Components of Survival Data

1. A dichotomous variable indicating whether the event was observed, δ
2. A quantitative time component, or the survival time/right censoring time, T

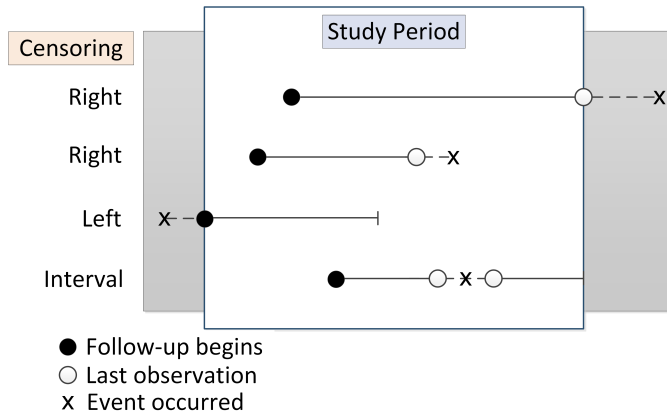
- δ_i : Event indicator = $\begin{cases} 1 & \text{if event (failure, disease, death)} \\ 0 & \text{if subject } i \text{ is censored} \end{cases}$
- $t_i = \begin{cases} \text{Time from origin to event} & \text{if } \delta_i = 1 \\ \text{Time from origin to censoring} & \text{if } \delta_i = 0 \end{cases}$
- Available information on the outcome: the pair (t_i, δ_i)



Censoring

- The main feature of survival data that renders standard methods inappropriate is that survival times are frequently incomplete, or **censored**
- An individual is said to be **censored** ($\delta_i = 0$) when the endpoint of interest **has not been observed for that individual**. We do not know the survival time exactly.
- **Censoring time** - Last time subject known to be at risk without having experienced the event

Censoring

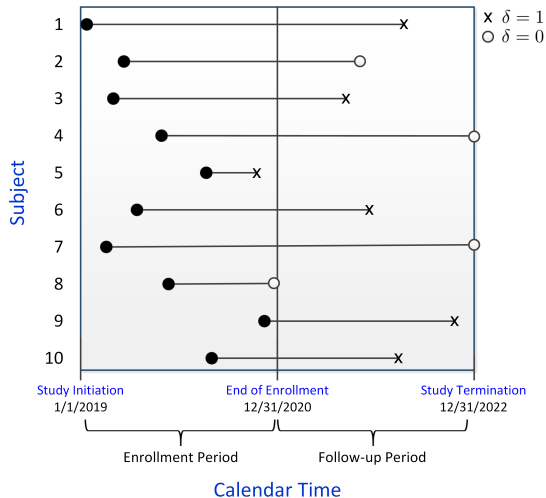


- **Right censoring** - Actual survival time is greater than that observed, or to the *right* of the censoring time
- **Interval censoring** - Individuals are known to have experienced event within an interval of time
- **Left censoring** - Actual survival time of an individual is less than that observed, or to the *left* of the censoring time

Right Censoring

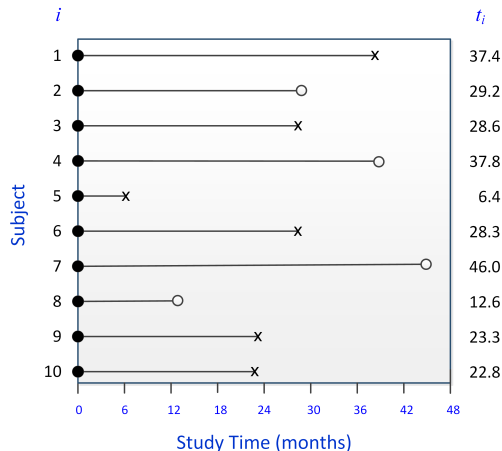
- **Right censoring time** - Last time subject known to be at risk without having experienced the event
- There are generally three reasons why right censoring occurs:
 1. A person does not experience the event before the study ends (administrative censoring)
 2. A person is lost to follow-up during the study period
 3. A person withdraws from (drops out of) the study
- Knowing that they left the study without having experienced the event contains information. Want to account for censoring in the analysis.

Calculating Time, Hypothetical RCT



- Subjects enter study at different times (**staggered entry**)
- Subjects entering at different times will have variable lengths of maximum follow-up time

Calculating Time, Hypothetical RCT



- Regardless of calendar time, each subject's time of enrollment defines the $t = 0$ point
- In practice, value of time is found by calculating the number of days (or months, years, etc.) between two calendar dates
- Collect data in **calendar time** and convert to **analysis (study) time**

Calculating Time, Hypothetical RCT

- t_i : 37.4, 29.2+, 28.6, 37.8+, 6.4, ... Censored Observations

Subject i	t_i	δ_i
1	37.4	1
2	29.2	0
3	28.6	1
4	37.8	0
5	6.4	1

Censoring times Event times

Calculating Time: Example

R Code, Working with Dates

```
> class(whas$ADDATE)  # Dates are imported as character
[1] "character"
> whas$ADDATE[1:5]    # Date of entry into study
[1] "3/13/1995" "1/14/1995" "2/17/1995" "4/7/1995" "2/9/1995"
> whas$FOLDATE[1:5]   # Date of death or censoring
[1] "3/19/1995" "1/23/1996" "10/4/2001" "7/14/1995" "5/29/1998"

> library(lubridate)  # Load required package
> whas$ADDATE <- mdy(whas$ADDATE)
> whas$FOLDATE <- mdy(whas$FOLDATE)
> class(whas$ADDATE)
[1] "Date"

# Duration of the interval between ADDATE and FOLDATE, converted to days using (ddays(1))
> whas$TIME <- as.duration(whas$ADDATE %--% whas$FOLDATE)/ddays(1)
> whas$TIME[1:5]
[1] 6 374 2421 98 1205
```

Terminology

- Basic **mathematical terminology and notation** for survival analysis:
 - T : The random variable for a person's survival time ($T \geq 0$)
 - t : Any specific value of interest for the random variable T
- For example, if we are interested in evaluating the likelihood that a person survives for more than $t = 5$ years after hospitalization for acute MI, are interested in estimating $P(T > 5)$
- Survival data are generally described and modeled in terms of two related functions:
 - **Survival/survivor function** $S(t)$
 - **Hazard function** $h(t)$ or $\lambda(t)$

Survivor Function

Survivor Function

The survivor function $S(t)$ reports the probability of surviving longer than some specified time t , $S(t) = P(T > t)$

- Survival probabilities are indexed by time, $S(t)$
- Phrase “at a particular time” is important in the interpretation of survival probabilities
- Probability of survival is expected to change over the time period
- Report 90-day survival, 6-month survival, 1-year survival

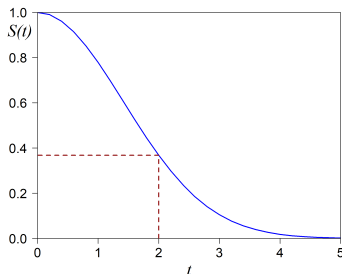
t	$S(t)$
1	$S(1) = P(T > 1)$
2	$S(2) = P(T > 2)$
3	$S(3) = P(T > 3)$
\vdots	\vdots



Survivor Function

- The **survivor function**, $S(t)$, is fundamental to a survival analysis, because obtaining survival probabilities for different values of t allows us to summarize survival data

Figure: Survivor function



- All survivor functions have the following characteristics:
 - Nonincreasing
 - $S(0) = P(T > 0) = 1$
 - $S(2) = P(T > 2) = 0.37$
 - $S(\infty) = 0$
- In practice, **estimate** $S(t)$ using the data, $\hat{S}(t)$

Hazard Function

- When the outcome is time-to-event, the **event rate** is likely to change over time

Hazard Function

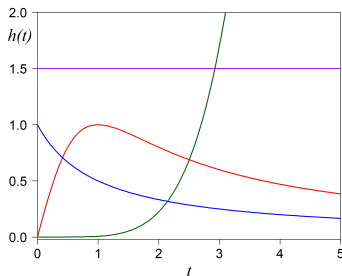
The hazard function $h(t)$ gives the instantaneous rate of failure at time t , given that the individual has survived up to time t (still at risk at time t)

- $$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$
- The **conditional probability** gives the probability that a person's survival time, T , will lie in the time interval between t and $t + \Delta t$, **given that** the survival time is greater than or equal to t
- The hazard is a **rate** because it is divided by the the small time interval, Δt

Hazard Function

- A **hazard function**, $h(t)$, is a description of the event rate at different time points; can be graphed over t

Figure: Examples of hazard functions



- Unlike the event rates discussed in Poisson regression, hazard rates are not assumed to be constant across all time
- Hazard functions are:
 - Nonnegative
 - Have no upper bound
- Constant hazard: $h(t) = \lambda$, Instantaneous potential for becoming ill at any time does not change throughout the study period

Survival vs. Hazard

Figure: Examples of hazard functions

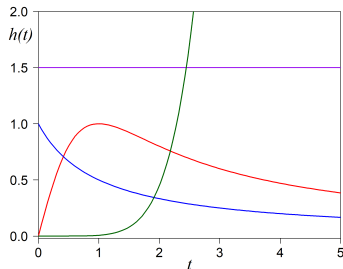
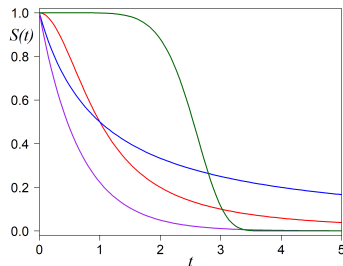


Figure: Examples of survival functions



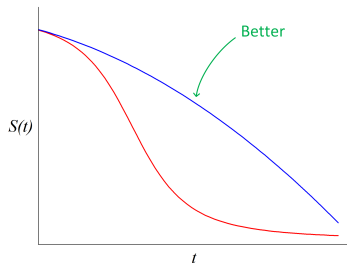
- Although the survival functions have same basic shape, hazard functions are dramatically different
- In contrast to the survivor function, which focuses on *not having an event*, the hazard function focuses on the event *occurring*; hazard function usually more informative about the underlying mechanism of failure than the survival function

Goals of Survival Analysis

- **Goal 1:** Estimate and interpret survivor function from survival data
 - Kaplan-Meier estimate of the survivor function
- **Goal 2:** Compare survivor functions
 - Log-rank test
- **Goal 3:** Assess the relationship between explanatory variables and survival time/hazard of event
 - Cox proportional hazards regression

Comparing Survival Experiences

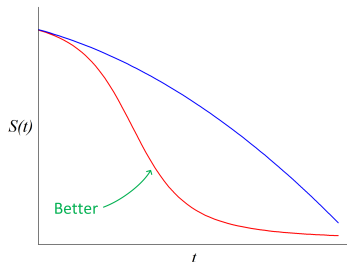
- Two survivor functions below describe two different survival experiences over time
- Suppose the event is a **negative** one (e.g., death or relapse)



- **Red curve:** Quick drop in survival probabilities early in follow-up
- **Blue curve:** Shows a steady decrease in survival probabilities throughout follow-up (better survival experience)
- **Blue** group tends to “survive” longer (longer time until death or relapse in patients in the **blue** group as compared to the **red** group)

Comparing Survival Experiences

- Suppose the event is a **positive** one (e.g., return to work after surgery)



- Estimated survival curve for those in the **blue** group is above those in the **red** group
 - **Blue** group tends to “survive” longer (longer time until return to work in the **blue** group as compared to the **red** group)
 - **Red** group returns to work quicker than the **blue** group
- If it is a positive endpoint, want time to positive endpoint to be shorter

Progress this Unit

- 1 Survival Data
 - Motivation for Survival Analysis
 - Components of Survival Data
 - Terminology and Notation
- 2 Estimating Survivor Function
 - Data Structure
 - Kaplan-Meier Method
- 3 Group Comparisons and Modeling
 - Log-Rank Test
 - Cox Proportional Hazards Model
 - Cox Adjusted Survival Curves

Typical Data

- Survival data consist of a survival time (t_i) for each subject, some of which may be censored, and an event indicator (δ_i)
- In addition, there might be some *covariates* (x) on each patient (e.g. treatment group, age, weight, sex, smoking status)

Subject	t	δ	x
1	t_1	δ_1	x_1
2	t_2	δ_2	x_2
\vdots			
n	t_n	δ_n	x_n
$\sum \delta_i = \text{Total number of failures}$			

Typical Data: Example

Subject	t	δ	x (group)
1	6	1	1
2	6	1	1
3	6	1	1
4	6	0	1
5	7	1	1
6	9	0	1
7	10	1	1
8	10	0	1
9	11	0	1
10	13	1	1
11	16	1	1
...			
22	1	1	0
23	1	1	0
...			
41	22	1	0
42	23	1	0

- **Example:** Clinical trial in leukemia patients in remission. 21 subjects randomized to chemotherapy treatment (6-MP); 21 subjects randomized to placebo.
- Event (failure) of interest: Relapse. Time measured in weeks.
- A person is censored if:
 - He/she remains in remission until the end of the study,
 - Is lost to follow-up,
 - Or withdraws before the end of the study

Censored Observations

- Even though censored observations are incomplete, in that a subject's survival time is not known exactly, can still make use of the information we have on a censored individual up to the time they are censored
- They are at risk of experiencing the event up until the time they are censored

Kaplan-Meier Estimator

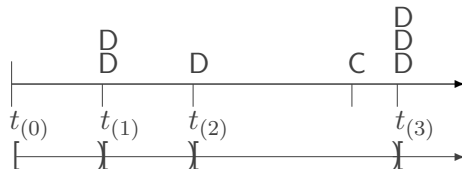
- **Kaplan-Meier/Product-Limit Method** - Most commonly used method for estimating $S(t)$
- Can **non-parametrically** (i.e., without a probability distribution assumption) estimate the survival probabilities/survivor function from the observed survival times, both censored and uncensored
- At each time point, a subject is either:
 1. Still waiting for the event,
 2. Censored, or
 3. Has experienced the event.

Kaplan-Meier Method

- Suppose there are k distinct **event times** in the period of follow-up

1. Construct a series of **time intervals** based on unique event times

- One event time in each interval
- Event is assumed to occur at the *start* of the interval (events define the intervals)



- **Ordered** failure times from earliest to latest: $t_{(1)}, \dots, t_{(k)}$ such that $t_{(1)} < t_{(2)} < \dots < t_{(k)}$
- $t_{(0)} \equiv \text{time } 0$



Kaplan-Meier Method

2. Summarize the number of failures (d_j), number censored (c_j), and number of subjects at risk (n_j) in each time interval

j	Ordered Failure Times	Interval	d_j	c_j	n_j
0	$t_{(0)} = 0$	$I_0 = [t_{(0)}, t_{(1)})$	$d_0 = 0$	c_0	$n_0 = N$
1	$t_{(1)}$	$I_1 = [t_{(1)}, t_{(2)})$	d_1	c_1	n_1
2	$t_{(2)}$	$I_2 = [t_{(2)}, t_{(3)})$	d_2	c_2	n_2
\vdots					
k	$t_{(k)}$	$I_k = [t_{(k)}, t_{max}]$	d_k	c_k	n_k

- d_j Number of failures occurring at $t_{(j)}$
- c_j Number censored in interval j , $[t_{(j)}, t_{(j+1)})$
- $n_j = n_{j-1} - d_{j-1} - c_{j-1}$, Number alive instant before $t_{(j)}$ (risk set for interval j)
 - Number entering the interval alive; includes those censored and failing at $t_{(j)}$
 - Censored observations are taken into account when determining the number at risk at $t_{(j)}$

Kaplan-Meier Method

3. Estimate the *conditional* probability of **event** at $t_{(j)}$ given they are at risk at time $t_{(j)}$

- $\frac{d_j}{n_j}$: Proportion of individuals who experience the event at time $t_{(j)}$ given survival to $t_{(j)}$
- d_j : Number of deaths occurring at $t_{(j)}$
- n_j : Number at risk at $t_{(j)}$

4. Estimate the *conditional* probability of **surviving** beyond $t_{(j)}$ given they are at risk at time $t_{(j)}$

- $1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}$

Kaplan-Meier Method

5. As events are assumed to occur **independently** of one another, **multiply** the conditional probabilities for the current interval and all preceding intervals to find the *unconditional* probability of surviving beyond $t_{(j)}$, $\hat{S}(t)$

Kaplan-Meier (Product-Limit) Estimator

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

- \prod symbol: “**multiply** all of the individual conditional probabilities together” for the current interval and all intervals before

Kaplan-Meier Method

- The Kaplan-Meier estimator may also be written as:

Kaplan-Meier (Product-Limit) Estimator

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \times \left(1 - \frac{d_j}{n_j}\right)$$

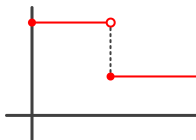
- Example:

$$\begin{aligned}\hat{S}(t_{(4)}) &= \underbrace{\left(1 - \frac{d_1}{n_1}\right) \times \left(1 - \frac{d_2}{n_2}\right) \times \left(1 - \frac{d_3}{n_3}\right)}_{\hat{S}(t_{(3)})} \times \left(1 - \frac{d_4}{n_4}\right) \\ &= \hat{S}(t_{(3)}) \times \left(1 - \frac{d_4}{n_4}\right)\end{aligned}$$

Kaplan-Meier Survival Curve

- The results of the Kaplan-Meier analysis are often graphed; graphs are known as **Kaplan-Meier survival curves**
- Right-continuous stepwise function
 - Jumps only at the observed failure times, constant between failure times

Figure: Right-continuous function



Kaplan-Meier Survival Curve

- **Beginning** of function:
 - $\hat{S}(0) = 1$, remains at 1 until the first event time $t_{(1)}$
- **End** of the function:
 - If all remaining subjects **fail** at largest observed time t_{max} , $\hat{S}(t_{max}) = 0$
 - If **censoring** occurs at t_{max} , then $\hat{S}(t_{max}) > 0$
- Comparing the survival curves of two different populations can yield insightful information about the timing of deaths in response to different environmental conditions
- Often in the literature, you will see the survival curves for two different subgroups on the same graph so that you can compare the two easily

Kaplan-Meier Survival Probabilities: Example

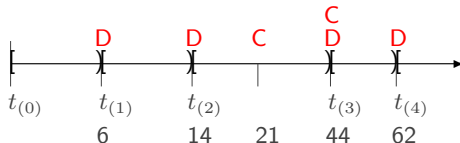
- Simple example to illustrate the calculation of the Kaplan-Meier survival function:

Subject i	t_i	δ_i
1	6	1
2	14	1
3	21	0
4	44	1
5	44	0
6	62	1

- Each interval begins at an observed event time ($\delta = 1$) and ends just before the next ordered event time
- $I_0 = \{t : 0 \leq t < 6\} = [0, 6)$

Kaplan-Meier Survival Probabilities: Example

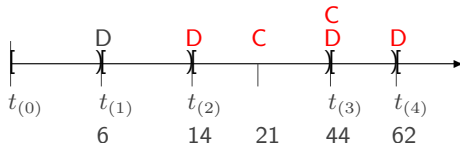
j	$t_{(j)}$	Interval	d_j	c_j	n_j	d_j/n_j	$1 - (d_j/n_j)$
0	0	$I_0 = [0, 6)$	0	0	6	0/6	6/6
1	6	$I_1 = [6, 14)$	1	0	6	1/6	5/6
2	14	$I_2 = [14, 44)$	1	1			
3	44	$I_3 = [44, 62)$	1	1			
4	62	$I_4 = [62, 62]$	1	0			



- All 6 subjects are alive at time 0 and remain alive until 1 subject dies at 6 days. The fraction of subjects surviving past time 0 is $6/6$.
- The risk set at 6 weeks also consists of 6 persons, because all 6 survived at least 6 weeks. Of the subjects who survived to week 6, $5/6$ survived past week 6.

Kaplan-Meier Survival Probabilities: Example

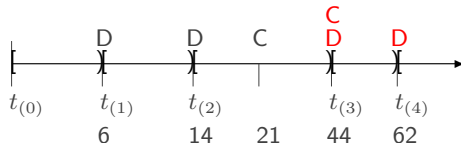
j	$t_{(j)}$	Interval	d_j	c_j	n_j	d_j/n_j	$1 - (d_j/n_j)$
0	0	$I_0 = [0, 6)$	0	0	6	0/6	6/6
1	6	$I_1 = [6, 14)$	1	0	6	1/6	5/6
2	14	$I_2 = [14, 44)$	1	1	5	1/5	4/5
3	44	$I_3 = [44, 62)$	1	1			
4	62	$I_4 = [62, 62]$	1	0			



- The risk set at 14 weeks consists of the **5** persons who survived at least 14 weeks. Of the subjects who survived to week 14, **4/5** survived past week 14
- n_j does not include the subject who experienced the event at 6 weeks (did not survive at least 14 weeks)

Kaplan-Meier Survival Probabilities: Example

j	$t_{(j)}$	Interval	d_j	c_j	n_j	d_j/n_j	$1 - (d_j/n_j)$
0	0	$I_0 = [0, 6)$	0	0	6	0/6	6/6
1	6	$I_1 = [6, 14)$	1	0	6	1/6	5/6
2	14	$I_2 = [14, 44)$	1	1	5	1/5	4/5
3	44	$I_3 = [44, 62)$	1	1	3	1/3	2/3
4	62	$I_4 = [62, 62]$	1	0			

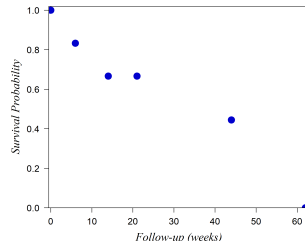


- The risk set at 44 weeks consists of the **3** persons who survived at least 44 weeks. Of the subjects who survived to week 44, **2/3** survived past week 44
- Assume the status of the patient censored at 44 weeks known not to have experienced the event at week 44. He survived past week 44.

Kaplan-Meier Survival Probabilities: Example

$t_{(j)}$	Interval	d_j	c_j	n_j	$1 - (d_j/n_j)$	$\hat{S}(t_{(j)})$
0	[0, 6)	0	0	6	$6/6 = 1$	1
6	[6, 14)	1	0	6	$5/6 = 0.83$	$1 \times 0.83 = 0.83$
14	[14, 44)	1	1	5	$4/5 = 0.8$	$1 \times 0.83 \times 0.8 = 0.67$
44	[44, 62)	1	1	3	$2/3 = 0.67$	$1 \times 0.83 \times 0.8 \times 0.67 = 0.44$
62	[62, 62]	1	0	1	$0/1 = 0$	$1 \times 0.83 \times 0.8 \times 0.67 \times 0 = 0$

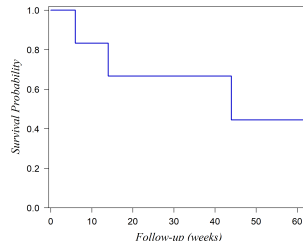
- Estimated survival probability changes at the observed failure times and is constant between event times
- Estimated probability of surviving past 6 weeks = $\hat{S}(6) = 0.83$
- Estimated probability of surviving past 44 weeks = $\hat{S}(44) = 0.44$
- In absence of censoring, $\hat{S}(t)$ reduces to ratio of number of individuals event free at time t divided by the number of people who entered the study



Kaplan-Meier Survival Probabilities: Example

$t_{(j)}$	Interval	d_j	c_j	n_j	$1 - (d_j/n_j)$	$\hat{S}(t_{(j)})$
0	[0, 6)	0	0	6	$6/6 = 1$	1
6	[6, 14)	1	0	6	$5/6 = 0.83$	$1 \times 0.83 = 0.83$
14	[14, 44)	1	1	5	$4/5 = 0.8$	$1 \times 0.83 \times 0.8 = 0.67$
44	[44, 62)	1	1	3	$2/3 = 0.67$	$1 \times 0.83 \times 0.8 \times 0.67 = 0.44$
62	[62, 62]	1	0	1	$0/1 = 0$	$1 \times 0.83 \times 0.8 \times 0.67 \times 0 = 0$

- Estimated survival probability changes at the observed failure times and is constant between event times
- Estimated probability of surviving past 6 weeks = $\hat{S}(6) = 0.83$
- Estimated probability of surviving past 44 weeks = $\hat{S}(44) = 0.44$
- In absence of censoring, $\hat{S}(t)$ reduces to ratio of number of individuals event free at time t divided by the number of people who entered the study



Kaplan-Meier Survival Probabilities: Example

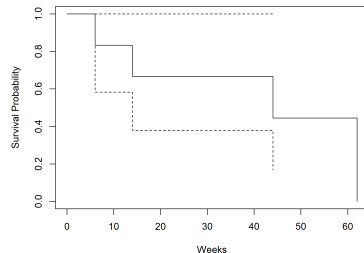
R Code, Kaplan-Meier Method

```
# Load required package
> library(survival)

# Estimate KM probabilities (full sample)
> kmsurv <- survfit(Surv(time, censor) ~ 1, data = example)
> summary(kmsurv)

Call: survfit(formula = Surv(time, censor) ~ 1, data = example)
   time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6      6      1    0.833   0.152    0.583      1
   14      5      1    0.667   0.192    0.379      1
   44      3      1    0.444   0.222    0.167      1
   62      1      1    0.000    NaN      NA      NA

> plot(kmsurv, xlab = "Weeks", ylab = "Survival Probability",
       mark.time = F)
```

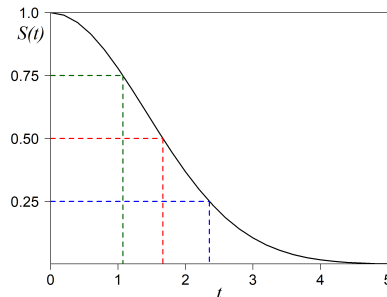


- Estimated survival function and confidence intervals provide a useful descriptive measure of overall survival experience

Survival Percentiles

- Because the distribution of survival times tends to be positively skewed, the **median survival time** is the preferred summary measure
- **Median survival time:** Time beyond which 50% of the population is expected to survive. Also the time by which 50% of the population is expected to die; t_{50} such that $S(t_{50}) = 0.50$
- Commonly reported **survival percentiles**:
 - Lower quartile (25th percentile): $S(t_{25}) = 0.75$
 - Median (50th percentile): $S(t_{50}) = 0.50$
 - Upper quartile (75th percentile): $S(t_{75}) = 0.25$

Figure: Common survival percentiles



Estimating Median Survival Time

- Median survival time and other percentiles of T may be estimated graphically using the estimated Kaplan-Meier survival function
- $\hat{S}(t)$ is a step-function; unlikely there will be a t such that $\hat{S}(t) \equiv 0.5$
 - Estimated median survival time is the **smallest** observed survival time for which $\hat{S}(t)$ is less than or equal to 0.5
 - $\hat{t}_{50} = \min \{t \mid \hat{S}(t) \leq 0.5\}$
- If $\hat{S}(t) \equiv 0.5$, average event times at the endpoints of the segment
- \hat{t}_{50} is not defined if $\hat{S}(t)$ is never ≤ 0.5

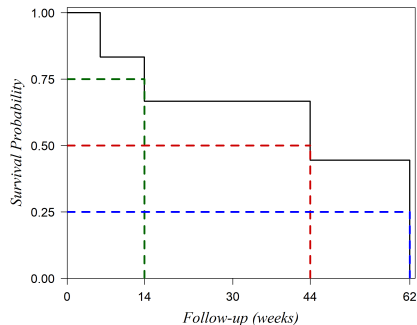
Median Survival Time: Example

- \hat{t}_{50} : Smallest observed survival time where $\hat{S}(t) \leq 0.5$

t	$\hat{S}(t)$
$0 \leq t < 6$	1
$6 \leq t < 14$	0.8
$14 \leq t < 44$	0.6
$44 \leq t < 62$	0.3
$t = 62$	0

- $\hat{t}_{50} = 44$ weeks

Figure: Kaplan-Meier survival curve and quartiles



Percentiles of Survival Time: Example

R Code, Quantiles of Survival Time

```
> quantile(kmsurv)
$quantile
25 50 75
14 44 62

$lower
25 50 75
6 14 44

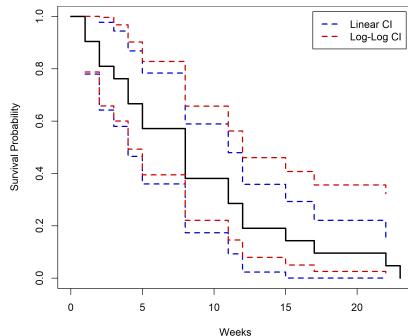
$upper
25 50 75
NA NA NA
```

- 25th percentile of survival time $\hat{t}_{25} = 14$ weeks [95% CI (6, .)]
 - Estimate that 75% will survive at least 14 weeks
- Median survival time (50th percentile)
 $\hat{t}_{50} = 44$ weeks [95% CI (14, .)]
 - Estimate that 50% will survive at least 44 weeks
- 75th percentile of survival time $\hat{t}_{75} = 62$ weeks [95% CI (44, .)]
 - Estimate that 25% will survive at least 62 weeks

Properties of KM Estimator

- The Kaplan-Meier estimator $\hat{S}(t)$ provides **point estimate** of $S(t)$
- Like all statistics estimate is subject to random variation
- Use estimated **standard error** to create pointwise confidence intervals for $S(t)$
- Two types of confidence intervals for $S(t)$:
 1. **Linear** (“plain”) CI
 2. **Log-log** CI
- Both valid; log-log more commonly used

Figure: Confidence intervals for $S(t)$



Confidence Interval for $S(t)$: Linear CI

- $\hat{S}(t)$ is approximately normally distributed
- Linear confidence interval for $S(t)$ at time t has form:

100(1 - α)% Linear Confidence Interval for $S(t)$ at t

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \widehat{SE} \left[\hat{S}(t_{(j)}) \right]$$

- For $t_{(j)} \leq t < t_{(j+1)}$
- Yields symmetric confidence intervals
- Problem with this method: Can give confidence limits outside range of $[0, 1]$
- **Solution 1:** Replace any computed limit out of this range by 0 or 1

Confidence Interval for $S(t)$: Log-log CI

- **Solution 2:** Compute confidence interval for log-log survivor function, $\log(-\log S(t))$ and transform to give CI for $S(t)$

100(1 - α)% Log-log Confidence Interval for $S(t)$ at t

- Use: $c_L = \log(-\log \hat{S}(t)) - z_{1-\frac{\alpha}{2}} \widehat{SE}[\log(-\log \hat{S}(t_{(j)}))]$
 $c_U = \log(-\log \hat{S}(t)) + z_{1-\frac{\alpha}{2}} \widehat{SE}[\log(-\log \hat{S}(t_{(j)}))]$
 - Transform to get CI for $S(t)$: $(\exp(-e^{c_U}), \exp(-e^{c_L}))$
- Endpoints always between $[0, 1]$
 - Non-symmetric confidence intervals
 - Default R confidence intervals

Confidence Intervals for $S(t)$: Example

R Code, Confidence Intervals

```
# Linear CI
```

```
> kmsurv.linear <- survfit(Surv(time, censor) ~ 1, data = example, conf.type = "plain")
```

```
> summary(kmsurv.linear)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	6	1	0.833	0.152	0.5351	1.00
14	5	1	0.667	0.192	0.2895	1.00
44	3	1	0.444	0.222	0.0089	0.88
62	1	1	0.000	NaN	NaN	NaN

```
# Log-log CI
```

```
> kmsurv.loglog <- survfit(Surv(time, censor) ~ 1, data = example, conf.type = "log")
```

```
> summary(kmsurv.loglog)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	6	1	0.833	0.152	0.583	1
14	5	1	0.667	0.192	0.379	1
44	3	1	0.444	0.222	0.167	1
62	1	1	0.000	NaN	NA	NA

```
# Plot KM estimate and CI (conf.int=FALSE option to suppress CI)
```

```
> plot(kmsurv.linear)
```

Group Comparisons: Example

- A comparison of survival functions is often necessary to address a research question
- Begin with a visual inspection of the Kaplan-Meier curves for both groups
- **Example:** Clinical trial in leukemia patients in remission
 - 21 subjects randomized to chemotherapy treatment (6-MP)
 - 21 subjects randomized to placebo
- Outcome = Time to relapse (**negative** event)

Group Comparisons: Example

R Code, KM Probabilities by Group

```
> kmsurv2 <- survfit(Surv(time, censor) ~ group_factor, data = leuk)
> summary(kmsurv2)
Call: survfit(formula = Surv(time, censor) ~ group, data = leuk)
```

```
      group=Placebo
time  n.risk  n.event  survival  std.err  lower 95% CI  upper 95% CI
  1      21       2    0.9048   0.0641    0.78754    1.000
  2      19       2    0.8095   0.0857    0.65785    0.996
  3      17       1    0.7619   0.0929    0.59988    0.968
  4      16       2    0.6667   0.1029    0.49268    0.902
  5      14       2    0.5714   0.1080    0.39455    0.828
  8      12       4    0.3810   0.1060    0.22085    0.657
 11       8       2    0.2857   0.0986    0.14529    0.562
 12       6       2    0.1905   0.0857    0.07887    0.460
 15       4       1    0.1429   0.0764    0.05011    0.407
 17       3       1    0.0952   0.0641    0.02549    0.356
 22       2       1    0.0476   0.0465    0.00703    0.322
 23       1       1    0.0000      NaN          NA          NA
```

- $\hat{t}_{50} = 8$ weeks in the Placebo arm

Group Comparisons: Example

R Code, KM Probabilities by Group (con't), Quantiles

```
group=6-MP
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  6      21       3   0.857  0.0764   0.720   1.000
  7      17       1   0.807  0.0869   0.653   0.996
 10      15       1   0.753  0.0963   0.586   0.968
 13      12       1   0.690  0.1068   0.510   0.935
 16      11       1   0.627  0.1141   0.439   0.896
 22       7       1   0.538  0.1282   0.337   0.858
 23       6       1   0.448  0.1346   0.249   0.807

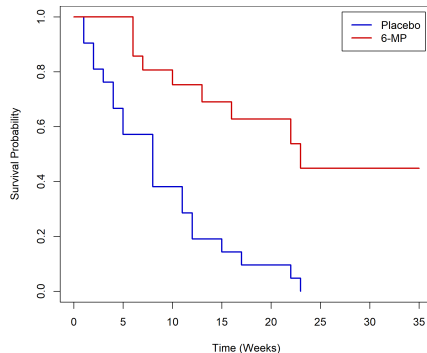
> quantile(kmsurv2)$quantile
      25 50 75
group_factor=Placebo  4  8 12
group_factor=6-MP    13 23 NA

> plot(kmsurv2, xlab = "Time (weeks)", ylab = "Survival Probability",
      mark.time = F, col = c("blue3", "red3"), lwd = 2)
> legend("topright", levels(leuk$group_factor),
      col = c("blue3", "red3"), lwd = 2)
```

- $\hat{t}_{50} = 23$ weeks in the 6-MP arm

Group Comparisons: Example

Figure: Kaplan-Meier survival curves. Time to relapse in leukemia patients by treatment



- Survival experience in **6-MP chemotherapy group** is better than survival experience in **placebo group**

Progress this Unit

- 1 Survival Data
 - Motivation for Survival Analysis
 - Components of Survival Data
 - Terminology and Notation
- 2 Estimating Survivor Function
 - Data Structure
 - Kaplan-Meier Method
- 3 Group Comparisons and Modeling
 - Log-Rank Test
 - Cox Proportional Hazards Model
 - Cox Adjusted Survival Curves

Log-Rank Test

- While the Kaplan-Meier survival curves suggest a difference, a formal **hypothesis test** is needed to determine if the difference in the sample data is large enough to conclude that the survival curves for the populations are different
- **Log-rank test** is a formal statistical inference procedure used to determine if the population survival curves are significantly different over the range of time t

Log-Rank Test Hypotheses

$$H_0 : S_1(t) = S_2(t) \text{ for all times } t \text{ vs.}$$

$$H_1 : S_1(t) \neq S_2(t) \text{ for some value of time } t$$

- Under the null hypothesis, distribution of survival times is identical in the two groups

Log-Rank Test

- Assuming we have two independent populations, the log-rank test compares:
 - The total number of **observed events** in group 1 to
 - The total number of **expected events** in group 1 under H_0 (assuming the survival experiences for group 1 and group 2 are identical)
- Using sample data, a test statistic is computed
 - Involves looking at each distinct event time (stratum)
 - Contributions over all event times are accumulated using a Mantel-Haenszel test statistic

Log-Rank Test

Log-Rank Test Statistic: Comparing 2 Groups

$$\chi^2 = \frac{(\text{total observed events in group 1} - \text{total expected events in group 1})^2}{\text{total variance for the number of event occurrences for group 1}}$$

- Observed counts that disagree substantially with the expected counts result in a **large** test statistic, leading to **rejection** of H_0 that the survival experiences are the same
- Test statistic is compared to a χ^2 distribution with $df = \text{number of groups} - 1$
 - When two groups are being compared, χ^2 is compared to χ^2_1
 - When three groups are being compared, χ^2 is compared to χ^2_2



Log-Rank Test: Example

- **Example:** Is the survival experience in 6-MP significantly different from that in placebo? $H_0 : S_1(t) = S_2(t)$ for all t vs. $H_1 : S_1(t) \neq S_2(t)$ for some t

R Code, Log-Rank Test

```
# Log-rank test
> survdiff(Surv(time, censor) ~ group_factor, data = leuk)
Call:
survdiff(formula = Surv(time, censor) ~ group_factor, data = leuk)

           N Observed Expected (O-E)^2/E (O-E)^2/V
group_factor=Placebo 21         21   10.7      9.77    16.8
group_factor=6-MP    21         9   19.3      5.46    16.8

Chisq= 16.8 on 1 degrees of freedom, p= 0.00004
```

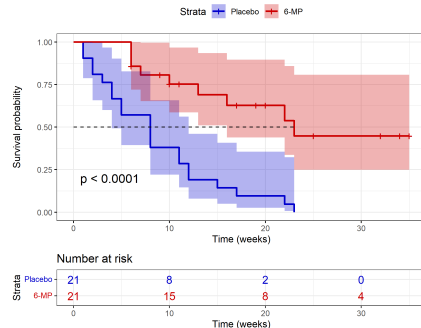
- Log-rank test statistic $X^2 = 16.79$ compared to χ_1^2 `pval = 1-pchisq(16.8, df = 1)`
 - Reject H_0 if $X^2 \geq \chi_{1-\alpha}^2(1) = \chi_{.95}^2(1) = 3.84$
 - $X^2 = 16.8 > 3.84 \rightarrow$ Reject H_0
 - $p = P(\chi^2 \geq 16.8) < .0001$
 - $p < 0.05 \rightarrow$ Reject H_0
- **Conclusion:** There is evidence to reject the null hypothesis and conclude the survival curves for those receiving 6-MP and those receiving placebo are significantly different ($p < .0001$)

Summarizing Group Differences

- There is a statistically significant difference in the survival functions between the two treatment groups ($p < .0001$). The drug effectively delays the time to relapse in this patient population.

Table: Median survival time and log rank test comparing groups

6-MP Chemotherapy	Placebo	Log-rank p -value
23 weeks	8 weeks	$< .0001$



Wilcoxon Test

- Another test for comparing survival curves used in practice is the [Wilcoxon test](#)
- Slight variant of the log-rank test statistic, which places more weight on differences in the survival curves at *earlier* times (when the number at risk is larger)
- Can be better at detecting differences in the survival curves that exist at earlier time periods
- Same H_0 , H_1 , tests comparing 2 groups also compared to χ^2_1

R Code, Wilcoxon Test

```
> survdiff(Surv(time, censor) ~ group_factor, data = leuk, rho = 1)
Call:
survdiff(formula = Surv(time, censor) ~ group, data = leuk, rho = 1)

      N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 21    14.55     7.68     6.16     14.5
group=1 21     5.12    12.00     3.94     14.5

Chisq= 14.5 on 1 degrees of freedom, p= 0.0001
```

Comment About Multiple Group Comparisons

- Can compare **more than 2 groups** using the log-rank and Wilcoxon tests, χ^2_{g-1}
- Kaplan-Meier curves become difficult to examine if there are many levels of a categorical variable or if more than one categorical variable is of interest
- Furthermore, unless a continuous variable is categorized, log-rank/Wilcoxon tests and Kaplan-Meier curves cannot be used to investigate the relationship

Motivation for Modeling

- The log-rank test allowed us to perform unadjusted comparisons of the survival experiences between groups
- To adjust for the effect of confounding variables and to investigate multiple risk factors simultaneously, **regression** can be used
- One of the advantages of regression modeling is the ability to examine the effect of multiple predictor variables (dichotomous, nominal, ordinal, and/or continuous) on the outcome of interest

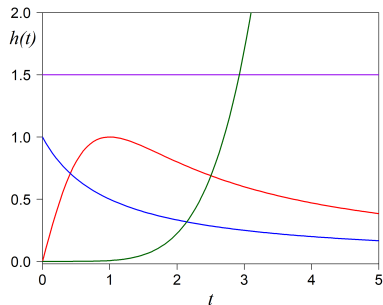
Regression Models

Y Variable	Outcome Measured in Terms of	Regression Model	Yields
Continuous	μ	Linear Regression	Mean Difference
Dichotomous	$\log\left(\frac{p}{1-p}\right)$	Logistic Regression	Odds Ratio
Count	$\log(\mu)$ or $\log(\lambda)$	Poisson Regression	Mean or Rate Ratio
Survival	$\log(h(t))$	Cox Regression	Hazard Ratio

- The most commonly used regression model for censored time-to-event data is the **Cox proportional hazards model** (a.k.a., the Cox regression model)

Hazard, Revisited

Figure: Examples of hazard functions



- The **hazard function** or **hazard rate** at time t , $h(t)$, is the instantaneous failure rate at time t
- Hazard rate: Probability that if you survive to t , you will experience to the event in the next instant, Δt , divided by the length of the interval, Δt , giving a rate

Cox Proportional Hazards Model

Hazard Function: Cox PH Model

$$h(t; x) = h_0(t) \exp(\beta x)$$

Simple Cox PH Regression Model

$$\log(h(t; x)) = \log(h_0(t)) + \beta x$$

- Because the hazard function is a **rate**, it must be strictly positive; exponential part of the model ensures the fitted model will always give an estimated hazard that is non-negative
- $h_0(t)$ is the hazard function for an individual who has all covariates = 0 (“**baseline hazard**”)
- In Cox regression, the baseline hazard is left unspecified (i.e., no intercept term is estimated in a Cox model)

Time-to-event response variable → Cox PH regression

Cox Proportional Hazards Model

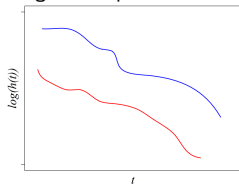
$$h(t; x) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$$

1. $h_0(t)$: Baseline hazard function characterizes how the hazard function changes as a function of survival time
 - $h_0(t)$ is a function of t , not a function of X
2. $\exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$: Characterizes how the hazard changes as a function of subject covariates
 - Function of X , not a function of t (time-independent)

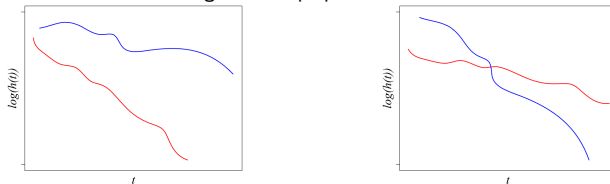
Assumptions of the Cox Model

- The primary assumption of the Cox **proportional hazards** model is that the hazards are **proportional** over the study period
- For any two different subjects, the log of the hazard functions, $\log(h(t, x = x_1))$ vs. $\log(h(t, x = x_0))$, are **parallel**

Figure: Proportional hazards




Figures: Nonproportional hazards



- **Proportional hazards**: Ratio of hazards will be the same over time
- **Nonproportional hazards**: Ratio of hazards will depend on time

Interpretation of β : Change in Log Hazard

- Can compare the log hazard ($\log(h(t; x))$) under two conditions, e.g., $x = x_1$ vs. x_0
 - $\log(h(t; x)) = \log(h_0(t)) + \beta x$


$$\begin{aligned}\log(h(t; x = x_1)) - \log(h(t; x = x_0)) &= [\log(h_0(t)) + \beta x_1] - [\log(h_0(t)) + \beta x_0] \\ &= \log(h_0(t)) + \beta x_1 - \log(h_0(t)) - \beta x_0 \\ &= (x_1 - x_0) \beta\end{aligned}$$

- For example, a 1-unit increase in x ($x_1 - x_0 = 1$) gives an expected difference of β in the log hazard

$$\log(h(t; x = 1)) - \log(h(t; x = 0)) = \beta$$



Interpretation of β and e^β

$$\beta = \log(h(t; x = 1)) - \log(h(t; x = 0)) = \log\left(\frac{h(t; x = 1)}{h(t; x = 0)}\right) = \log(\text{HR})$$

Slope Parameter from Cox Regression Model

Slope β from a Cox regression model is the **log hazard ratio** associated with a 1-unit increase in risk factor x

$$e^\beta = \frac{h(t; x = 1)}{h(t; x = 0)} = \text{HR}$$

e^β from Cox Regression Model

e^β from a Cox regression model is the **hazard ratio** associated with a 1-unit increase in risk factor x

Proportional Hazards

- Because the baseline hazard cancels out, the hazard ratio does not depend on time

$$\text{HR} = \frac{h(t; x = 1)}{h(t; x = 0)} = \frac{h_0(t) \exp(\beta \times 1)}{h_0(t) \exp(\beta \times 0)} = e^\beta$$

- Constant hazard ratio at all time points: **proportional hazards**

- If $\beta = 0$, then $e^\beta = \frac{h(t; x = 1)}{h(t; x = 0)} = 1$

- If $\beta > 0$, then $e^\beta = \frac{h(t; x = 1)}{h(t; x = 0)} > 1$ $h(t; x = 1)$ is $100 \times (e^\beta - 1)\%$ **larger** than $h(t; x = 0)$

- If $\beta < 0$, then $e^\beta = \frac{h(t; x = 1)}{h(t; x = 0)} < 1$ $h(t; x = 1)$ is $100 \times (1 - e^\beta)\%$ **smaller** than $h(t; x = 0)$

Simple Cox PH Regression Model: Example

- **Example:** Estimate the hazard ratio associated with treatment received to determine if treatment is associated with time to leukemia relapse

- x : Group, dichotomous = $\begin{cases} 1 & \text{6-MP Chemotherapy} \\ 0 & \text{Placebo} \end{cases}$

$$\log(\hat{h}(t; x)) = \log(\hat{h}_0(t)) + b \text{ Group}$$

- $\log(\widehat{\text{HR}})$ and $\widehat{\text{HR}}$ in the drug group vs. the placebo group:

Interpretation of Estimated Slope

$$\log(\widehat{\text{HR}}) = b \qquad \widehat{\text{HR}} = e^b$$

Simple Cox PH Regression Model: Example

$$\log(h(t; x)) = \log(h_0(t)) + \beta x$$

R Code, Cox PH Regression

```
> cox1 <- coxph(Surv(time, censor) ~ group_factor, data = leuk)
> summary(cox1)
```

Call:
coxph(formula = Surv(time, censor) ~ group_factor, data = leuk)

n= 42, number of events= 30

	coef	exp(coef)	se(coef)	z	Pr(> z)
group_factor6-MP	-1.5721	0.2076	0.4124	-3.812	0.000138

	exp(coef)	exp(-coef)	lower .95	upper .95
group_factor6-MP	0.2076	4.817	0.09251	0.4659

$$\log(\hat{h}(t, x)) = \log(\hat{h}_0(t)) - 1.572 \text{ Group}$$

- Baseline hazard term is not estimated (no intercept is reported)

Simple Cox PH Regression Model: Example

- The **log hazard** of relapse is:

$$\log(\hat{h}(t, x)) = \log(\hat{h}_0(t)) - 1.572 \text{ Group}$$

R Code, HR

```
# "b"  
> coef(cox1)  
group_factor6-MP  
-1.572125  
  
# exp(b)  
> exp(coef(cox1))  
group_factor6-MP  
0.2076035
```

- Log hazard ratio between the 6-MP group and control group = -1.572

$$\log \left(\frac{\hat{h}(t; x = 1)}{\hat{h}(t; x = 0)} \right) = -1.572$$

- Estimated hazard ratio: $\widehat{HR} = \frac{\hat{h}(t; x = 1)}{\hat{h}(t; x = 0)} = e^{-1.572} = 0.208$
 - 6-MP treatment reduces the rate of relapse by 77.9% over placebo; consistent with KM curves



Hypothesis Test for β

$$\log(h(t; x)) = \log(h_0(t)) + \beta x$$

- The next step is to determine if there is a significant relationship between x and the **hazard of failure**
- **Hypothesis test** for the slope parameter β (Wald Test)
 - $H_0: \beta = 0$, equivalent to HR = 1
 - $H_1: \beta \neq 0$, equivalent to HR $\neq 1$

Wald Test Statistic for Slope

$$Z = \frac{b}{s_b} \sim N(0, 1)$$

Hypothesis Test for β : Example

- **Example:** Is there a difference in the **hazard** of relapse in the 6-MP chemotherapy group vs. the placebo group?

R Code, Cox PH Regression

```
> summary(cox1)
...
              coef exp(coef) se(coef)      z Pr(>|z|)
group_factor6-MP -1.5721    0.2076  0.4124 -3.812 0.000138
              exp(coef) exp(-coef) lower .95 upper .95
group_factor6-MP    0.2076      4.817  0.09251   0.4659
```

- $z = \frac{b}{s_b} = \frac{-1.5721}{0.4124} = -3.812$ compared to $N(0, 1)$ `pval = 2*(1-pnorm(3.812))`
- Reject H_0 if $|z| \geq z_{1-\frac{\alpha}{2}} = z_{.975} = z^* = 1.96$
- $|z| = 3.812 \geq z^* \rightarrow$ Reject H_0
- $p = 2 \times P(Z \geq 3.812) = 0.00014$
- $p < 0.05 \rightarrow$ Reject H_0
- **Conclusion:** There is a significant difference in hazard of relapse between the 6-MP and the placebo group ($p = 0.0001$)

Confidence Interval for $\log(\text{HR})$ and HR

- **Confidence interval** for the slope β , $\log(\text{HR})$, has the form:

100(1 - α)% Confidence Interval for $\log(\text{HR}), \beta$

$$b \pm z_{1-\frac{\alpha}{2}} s_b = (c_L, c_U)$$

- To find the confidence interval for the **hazard ratio**, exponentiate the lower and upper bounds of the CI for $\log(\text{HR})$

100(1 - α)% Confidence Interval for HR, e^β

$$(e^{c_L}, e^{c_U})$$

- **Note:** The confidence interval for the HR will *exclude* 1 if we rejected H_0 . The CI will *include* 1 if we failed to reject H_0 .

Confidence Interval for $\log(\text{HR})$ and HR: Example

- **Example:** Confidence interval for the HR of remission in 6-MP group vs. placebo group

R Code, Confidence Intervals

```
> cbind(logHR = coef(cox1), confint.default(cox1)) # logHR (b1) and CI for beta1
              logHR      2.5 %      97.5 %
group_factor6-MP -1.572125 -2.380408 -0.7638424
> exp(cbind(HR = coef(cox1), confint.default(cox1))) # HR and CI for HR
              HR      2.5 %      97.5 %
group_factor6-MP 0.2076035 0.09251284 0.4658729
```

- 95% CI for $\log(\text{HR})$, β :

$$b \pm z_{1-\frac{\alpha}{2}} s_b = -1.5721 \pm 1.96 \times 0.4124 = (c_L = -2.380, c_U = -0.764)$$

- 95% CI for HR: $(e^{-2.380}, e^{-0.764}) = (0.0825, 0.466)$, which excludes 1

Likelihood Ratio Test

- Just as in logistic and Poisson regression, a likelihood ratio test can be used to simultaneously test multiple β
- **Full (F) model:** $\log(h(t; x)) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2$
- **Reduced (R) model:** $\log(h(t; x)) = \log(h_0(t)) + \beta_1 x_1$
 - $H_0: \beta_2 = \beta_3 = 0$
 - $H_1: \beta_2$ and β_3 are not both 0

Likelihood Ratio Test Statistic

$$\begin{aligned} G &= -2 \log\text{-likelihood}(R) - (-2 \log\text{-likelihood}(F)) \\ &= -2 [\log\text{-likelihood}(R) - \log\text{-likelihood}(F)] \end{aligned}$$

- $G \sim \chi^2_{df}$, where df = Number of parameters tested under H_0

Multiple Cox PH Regression Model: Example

- **Question:** Is treatment associated with time to leukemia relapse after adjusting for the possible confounding variable, white blood cell count (WBC)

R Code, Cox PH Regression

```
> cox2 <- coxph(Surv(time, censor) ~ group_factor + logWBC, data = leuk)
> exp(cbind(HR = coef(cox2), confint.default
```

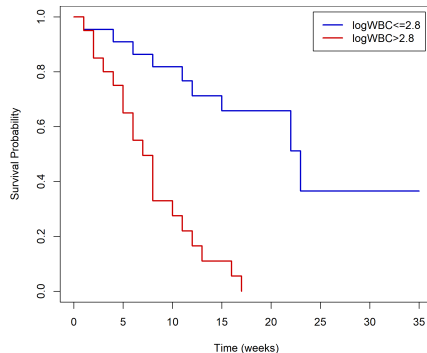
$$\log(\hat{h}(t, x)) = \log(\hat{h}_0(t)) - 1.386 \text{ Group} + 1.691 \log \text{WBC}$$

	Estimate	SE	HR	95% CI	p-value
Group (6-MP vs. Pbo)	-1.386	0.425	0.250	(0.109, 0.575)	0.001
logWBC	1.691	0.336	5.424	(2.808, 10.478)	<.0001

- The adjusted hazard ratio of relapse in the 6-MP group is training group is 0.25 [95% CI (0.109, 0.575)]; 6-MP treatment is associated with a 75% reduction in the relapse rate, adjusting for all other variables in the model ($p = 0.001$)

Categorizing a Continuous Variable: Example

Figure: Kaplan-Meier survival curves. Time to relapse in leukemia patients by logWBC



- Survival in those with **lower WBC** better than survival experience in those with **higher WBC**
- The adjusted hazard ratio associated with a 1-unit increase in logWBC is **5.424** [95% CI (2.808, 10.478)] ($p < .001$)

Checking the Proportional Hazards Assumption

- If the survival curves **cross** over levels of a covariate, the proportional hazards assumption is **violated** for that variable
 - The hazard functions in the groups of interest (i.e., exposed/unexposed) are not proportional over time
- However, if the survival curves **do not cross**, does **not** necessarily mean that the PH assumption is **satisfied**
- A commonly-used diagnostic plot for assessing the PH assumption is to plot the log-cumulative hazard $\log(\hat{H}(t))$ vs. $\log(t)$; curves will be parallel if the proportional hazards model is valid

$\log(-\log S(t))$ Plot: Optional Derivation

- The hazard of death at any time t is: $h(t; x) = h_0(t) \exp(\beta x)$
- Integrating both sides of this equation between 0 and t gives the cumulative hazard:

$$\int_0^t h(u; x) du = \exp(\beta x) \int_0^t h_0(u) du$$

$$H(t; x) = H_0(t) \exp(\beta x)$$

$$\log H(t; x) = \log H_0(t) + \beta x$$

- Since $H(t) = -\log S(t)$, the plot of $\log H(t)$ vs. $\log(t)$ is generally referred to as the $\log(-\log S(t))$ plot
- Use the estimated Kaplan-Meier survival probabilities $\hat{S}(t)$ to create this plot

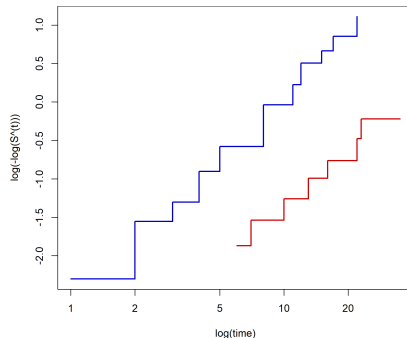
Checking the Proportional Hazards Assumption: Example

R Code, Assessing PH Assumption

```
# Kaplan-Meier survival probabilities by group
> kmsurv2 <- survfit(Surv(time, censor) ~
                    group_factor, data = leuk)

# log-log S(t) vs. log(t) for group
> plot(kmsurv2, col = c("blue3", "red3"),
      fun = "cloglog", xlab="log(time)",
      ylab="log(-log(S^(t)))")
```

Figure: Log cumulative hazard plot for group



- cloglog stands for complementary log-log (term used for $\log(-\log(\cdot))$)
- PH assumption does **not** appear to be violated for the treatment group variable
- Must categorize a **quantitative variable** in order to assess PH assumption

Violation of Proportional Hazards Assumption: Example

- **Example:** Examine the effect of **sex** on relapse in the leukemia data

R Code, KM Curves, Log-rank Test

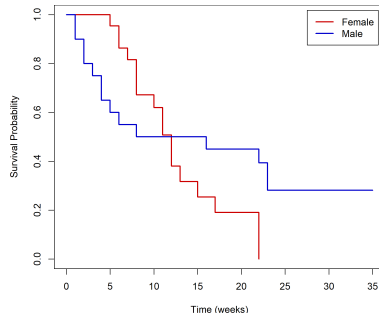
```
# Kaplan-Meier survival curves by sex
> kmsurv3 <- survfit(Surv(time, censor) ~ sex_factor,
  data = leuk)
> plot(kmsurv3, xlab = "Time (weeks)", ylab = "Survival
  Probability", mark.time = F, col = c("red3", "blue3"))

# Log-rank test
> survdiff(Surv(time, censor) ~ sex_factor, data = leuk)

      N Observed Expected (O-E)^2/E (O-E)^2/V
sex_factor=Female 22      16      14.2    0.240    0.557
sex_factor=Male   20      14      15.8    0.214    0.557

Chisq= 0.6  on 1 degrees of freedom, p= 0.5
```

Figure: Kaplan-Meier survival curves by sex



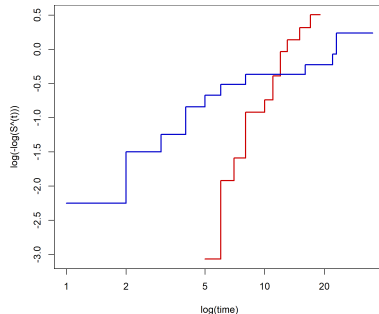
Violation of Proportional Hazards Assumption: Example

R Code, Assessing PH Assumption

```
# log-log S(t) vs. log(t) for sex  
> plot(kmsurv3, col = c("red3", "blue3"), fun = "cloglog",  
       xlab = "log(time)", ylab = "log(-log(S^(t)))")
```

- The two log-log survival plots clearly intersect and are, therefore, nonparallel
- Thus, sex, when considered by itself, appears to violate the PH assumption

Figure: Log cumulative hazard plot for sex



Extended Cox Model

- Can introduce **time-dependent variable** (function of time) into the model to estimate the hazard ratio for different points in time: **extended Cox model**

Extended Cox Regression Model

$$\log(h(t; x)) = \log(h_0(t)) + \beta x + \alpha(xt)$$

- **Note:** This is **not** a simple interaction term, but first requires data to be converted into **counting process format**; then, interaction term is created
- Can also assess the PH assumption by testing for the significance of the product term ($H_0: \alpha = 0$) using a Wald or likelihood ratio test
 - Under H_0 , extended Cox model reduces to traditional Cox PH model

Extended Cox Model

$$\log(h(t; x)) = \log(h_0(t)) + \beta x + \alpha(xt)$$

- By including an interaction with time, the hazard ratio is now a function of time:

$$\text{HR} = \frac{h_0(t) \exp(\beta(1) + \alpha(1)t)}{h_0(t) \exp(\beta(0) + \alpha(0)t)} = \exp(\beta + \alpha t)$$

- If $\alpha < 0$, the relative hazard *decreases* with time
- If $\alpha > 0$, the relative hazard *increases* with time

Extended Cox Model: Example

$$\log(h(t; x)) = \log(h_0(t)) + \beta \text{Sex} + \alpha(\text{Sex} \times t)$$

- Extend the Cox model to explore the effect of sex in the leukemia data
- In order to properly create the interaction, the total time at risk for an individual is sub-divided into smaller time intervals, providing a way for values of variables to change from time interval-to-interval for the same individual ([counting process data layout](#))

R Code, Creating Data Frame for Extended Cox Model

```
# Convert data into counting process style, defined by vector of unique event times
> cut.points <- unique(leuk$time[leuk$censor == 1])
> cut.points    # Unique event times in data
[1] 1  2  3  4  5  8 11 12 15 17 22 23  6  7 10 13 16
> leuk2 <- survSplit(data = leuk, cut = cut.points, end = "time", start = "time0",
                    event = "censor")
# Creating time-varying time*sex (using numerical version of sex, not sex_factor)
> leuk2$sextime <- leuk2$time*leuk2$sex
```

Extended Cox Model: Example

R Code, Creating Data Frame for Extended Cox Model

```
> subset(leuk[,c("ID", "sex", "time", "censor")], leuk$ID == 36)
  ID sex time censor
36 36  1  22      1
> subset(leuk2[,c("ID", "sex", "time0", "time", "censor", "severtime")], leuk2$ID == 36)
  ID sex time0 time censor severtime
304 36  1    0   1     0          1
305 36  1    1   2     0          2
306 36  1    2   3     0          3
307 36  1    3   4     0          4
308 36  1    4   5     0          5
309 36  1    5   6     0          6
310 36  1    6   7     0          7
311 36  1    7   8     0          8
312 36  1    8  10     0         10
313 36  1   10  11     0         11
314 36  1   11  12     0         12
315 36  1   12  13     0         13
316 36  1   13  15     0         15
317 36  1   15  16     0         16
318 36  1   16  17     0         17
319 36  1   17  22     1         22
```

Extended Cox Model: Example

R Code, Extended Cox Model

```
# Extended Cox model
> cox_extended <- coxph(Surv(time0, time, censor) ~ sex + sextime + cluster(ID), data = leuk2)
> summary(cox_extended)
Call:
coxph(formula = Surv(time0, time, censor) ~ sex + sextime, data = leuk2,
      cluster = ID)

n= 426, number of events= 30

              coef exp(coef) se(coef) robust se      z Pr(>|z|)
sex           2.07622   7.97430  0.81392   0.83396  2.490  0.01279
sextime      -0.27909   0.75647  0.08445   0.10067 -2.772  0.00556

      exp(coef) exp(-coef) lower .95 upper .95
sex           7.9743      0.1254      1.555   40.8849
sextime       0.7565      1.3219      0.621    0.9215
```

$$\log(\hat{h}(t; x)) = \log(\hat{h}_0(t)) + 2.076 \text{ Sex} - 0.279(\text{Sex} \times t) \quad \text{where Sex} = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

Extended Cox Model: Example

- Estimated HR for sex (male vs. female) varies over time (not constant over time)

$$\hat{HR} = \exp(b + at) = \exp(2.076 - 0.279t)$$

- HR of relapse in males vs. females at $t = 5$ weeks:

- $\hat{HR} = e^{2.076 - 0.279(5)} = 1.976$

- At $t = 10$ weeks:

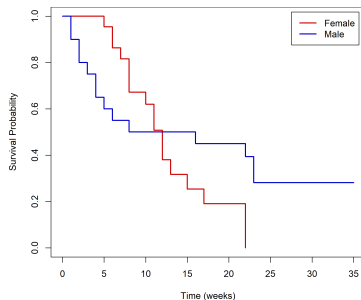
- $\hat{HR} = e^{2.076 - 0.279(10)} = 0.490$

- At $t = 15$ weeks:

- $\hat{HR} = e^{2.076 - 0.279(15)} = 0.121$

- Since $a = -0.279 < 0$, \hat{HR} decreases as t increases

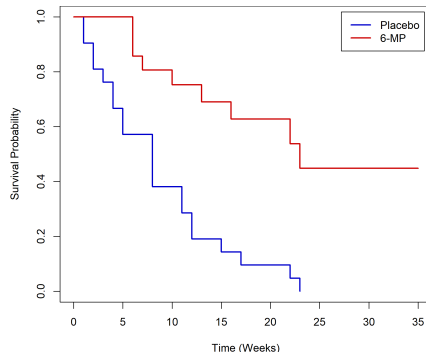
Figure: Kaplan-Meier curves by sex



Adjusted Survival Curves

- When a Cox model is used to fit survival data, we can plot **adjusted survival curves** (adjusting for explanatory variables used as predictors)
- A survival curve can be estimated using the Kaplan-Meier method; however, the Kaplan-Meier curve **does not adjust or control for any covariates**

Figure: Kaplan-Meier survival curves



Adjusted Survival Curves

- There is no closed form for the adjusted survival curve
- Cox estimated survival function:

Estimated Survival Function from Cox PH Model

$$\hat{S}(t; \mathbf{x}) = \left[\hat{S}_0(t) \right]^{\exp(b_1 x_1 + b_2 x_2 + \dots + b_k x_k)}$$

- **Example:** $\hat{S}(t, \mathbf{x}) = \left[\hat{S}_0(t) \right]^{\exp(-1.386 \text{ Group} + 1.691 \log \text{WBC})}$
 - Can specify fixed values for continuous covariates; sample mean of continuous covariate is a common choice (e.g., assume $\log \text{WBC} = 2.93$)

Adjusted Survival Curves: Example

- The estimated adjusted survival function is:

$$\hat{S}(t; \mathbf{x}) = \left[\hat{S}_0(t) \right]^{\exp(-1.386 \text{ Group} + 1.691 \log \text{WBC})}$$

- If $\text{Group} = 0$ (Placebo, reference), $\log \text{WBC} = 2.93$ ($\overline{\log \text{WBC}}$):

$$\hat{S}(t; \mathbf{x}) = \left[\hat{S}_0(t) \right]^{\exp(-1.386 \times 0 + 1.691 \times 2.93)} = \left[\hat{S}_0(t) \right]^{\exp(4.95)} = \left[\hat{S}_0(t) \right]^{141.83}$$

- If $\text{Group} = 1$ (6-MP group), $\log \text{WBC} = 2.93$:

$$\hat{S}(t; \mathbf{x}) = \left[\hat{S}_0(t) \right]^{\exp(-1.386 \times 1 + 1.691 \times 2.93)} = \left[\hat{S}_0(t) \right]^{\exp(3.57)} = \left[\hat{S}_0(t) \right]^{35.47}$$

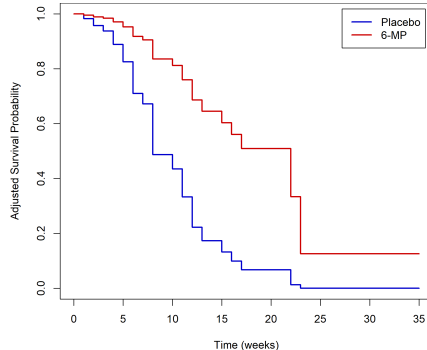
- The adjusted survival curves are estimated at particular covariate values
- Plot using software

Adjusted Survival Curves: Example

R Code, Adjusted Survival Curves

```
> cox2 <- coxph(Surv(time, censor) ~ group_factor + logWBC,  
  data = leuk)  
  
# Fitted S(t) by group assuming logWBC = its mean  
> pred.x <- data.frame(group_factor=levels(leuk$group_factor),  
  logWBC = mean(leuk$logWBC, na.rm = TRUE))  
  
# Adjusted survival probabilities  
> Shat <- survfit(cox2, newdata = pred.x, data = leuk)  
  
# Generate plot of adjusted survival curves  
> plot(Shat, xlab = "Time (weeks)", ylab = "Adjusted Survival  
  Probability", mark.time = F, col = c("blue3", "red3"))  
> legend("topright", levels(leuk$group_factor),  
  col = c("blue3", "red3"))
```

Figure: Cox Adjusted Survival Curves by Treatment at Mean logWBC=2.93



Adjusted Survival Probabilities: Example

R Code, Adjusted Survival Probabilities

```
> levels(leuk$group_factor)
[1] "Placebo" "6-MP"
> cbind(Shat$time, Shat$urv)
      1      2
[1,] 1 0.9826000933 0.9956204
[2,] 2 0.9574636904 0.9891896
[3,] 3 0.9372903338 0.9839363
[4,] 4 0.8887605634 0.9709422
[5,] 5 0.8249756494 0.9530281
[6,] 6 0.7099164641 0.9178964
[7,] 7 0.6717802071 0.9053101
[8,] 8 0.4873318107 0.8354859
[9,] 9 0.4873318107 0.8354859
...<omitted for space>
[18,] 20 0.0673460794 0.5093473
[19,] 22 0.0123416313 0.3332259
[20,] 23 0.0002519488 0.1259306
[21,] 25 0.0002519488 0.1259306
[22,] 32 0.0002519488 0.1259306
[23,] 34 0.0002519488 0.1259306
[24,] 35 0.0002519488 0.1259306
```

- Adjusted median survival time in the Placebo group: $\hat{t}_{50} = 8$ weeks
- Adjusted median survival time in the 6-MP group: $\hat{t}_{50} = 22$ weeks

Public Health Application: KM, Log-Rank, Cox PH Regression

Public Health Application

Body Mass and Smoking Are Modifiable Risk Factors for Recurrent Bladder Cancer

Asaf Wyszynski, PhD^{1,2}; Sam A. Tanyos, BS¹; Judy R. Rees, BM, BCh, PhD¹; Carmen J. Marsit, PhD^{1,2}; Karl T. Kelsey, MD³; Alan R. Schned, MD⁴; Eben M. Pendleton, BS¹; Maria O. Celaya, MPH¹; Michael S. Zens, PhD¹; Margaret R. Karagas, PhD¹; and Angeline S. Andrew, PhD¹

BACKGROUND: In the Western world, bladder cancer is the fourth most common cancer in men and the eighth most common in women. Recurrences frequently occur, and continued surveillance is necessary to identify and treat recurrent tumors. Efforts to identify risk factors that are potentially modifiable to reduce the rate of recurrence are needed. **METHODS:** Cigarette smoking behavior and body mass index were investigated at diagnosis for associations with bladder cancer recurrence in a population-based study of 726 patients with bladder cancer in New Hampshire, United States. Patients diagnosed with non-muscle invasive urothelial cell carcinoma were followed to ascertain long-term prognosis. Analysis of time to recurrence was performed using multivariate Cox regression models. **RESULTS:** Smokers experienced shorter time to recurrence (continuing smoker hazard ratio [HR] = 1.51, 95% confidence interval [CI] = 1.08-2.13). Although being overweight (body mass index > 24.9 kg/m²) at diagnosis was not a strong independent factor (HR = 1.33, 95% CI = 0.94-1.89), among continuing smokers, being overweight more than doubled the risk of recurrence compared to smokers of normal weight (HR = 2.67, 95% CI = 1.14-6.28). **CONCLUSIONS:** These observational results suggest that adiposity is a risk factor for bladder cancer recurrence, particularly among tobacco users. Future intervention studies are warranted to evaluate whether both smoking cessation and weight reduction strategies reduce bladder tumor recurrences. *Cancer* 2013;000:000-000.

- Multiple Cox proportional hazards regression model used to examine relationship between smoking and time to recurrence of bladder cancer, adjusted for age, gender, stage/grade, treatment, tumor size, and multiplicity

Public Health Application: KM, Log-Rank, Cox PH Regression

The main goals of the statistical analysis were to assess the relationship between cigarette smoking status, being overweight, and bladder cancer recurrence among non-muscle invasive urothelial cell carcinoma cases ($n = 726$). Time to first recurrence analysis was performed using Kaplan-Meier plots and differences were assessed using the log-rank test. To adjust for additional factors related to patient survival, Cox proportional hazards regression analysis was performed in non-muscle invasive tumors (stages 0 and I) with adjustment for age at diagnosis, sex, smoking status (never, former, continuing), as well as grade (low, high), presence of carcinoma in situ (Tis), first-course treatment (transurethral resection [TURB], immunotherapy, chemotherapy, radiotherapy, cystectomy), tumor size (< 3 cm, ≥ 3 cm), and multiplicity (single, > 1) in the model. P values represent 2-sided statistical tests with statistical significance at $P < .05$. Statistical significances of the interactions were assessed using likelihood ratio tests comparing the models with and without interaction terms.

Journal Example: KM, Log-Rank, Cox PH Regression

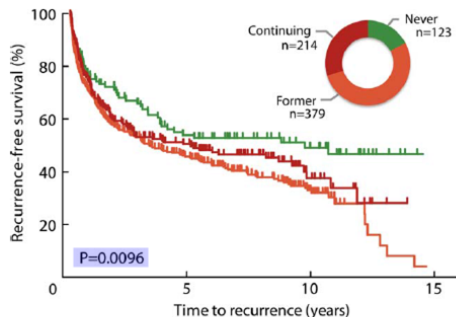


Figure 1. Bladder cancer recurrence is shown by cigarette smoking status after diagnosis. The Kaplan-Meier plot indicates shorter time to recurrence among patients who reported being either former or continued smokers when compared to never-smokers (log-rank $P=.0096$). Inset graphic shows relative group size of never (green; 17.2%), former (orange; 52.9%), and current (red; 29.9%) smokers.

TABLE 2. Bladder Cancer Recurrence and Smoking After Diagnosis

Smoking status at follow-up	Time to recurrence				
	N	(%)	Median years	HR ^a (95%CI)	P-value
Never	123	(17)	9.76	1.00 (ref)	
Former	379	(53)	3.48	1.61 (1.17–2.20)	0.0031
Continuing	214	(30)	5.14	1.51 (1.08–2.13)	0.018

^a adjusted for age, gender, stage/grade, treatment, smoking status, tumor size, and multiplicity.

Cigarette smoking was associated with shorter time to first recurrence (log-rank $P=.0096$, Fig. 1). After adjustment for age, sex, stage/grade, tumor size, multiplicity, and treatment, patients were at a higher risk of recurrence if they had a history of smoking when compared to nonsmokers, whether they continued smoking after diagnosis (hazard ratio [HR] = 1.51, 95% confidence interval [CI] = 1.08–2.13) or quit at or before diagnosis (HR = 1.61, 95% CI = 1.17–2.20; Table 2).

Lesson Summary

- When analyzing a time-to-event endpoint, (t_i, δ_i) , interested in describing the survival experience and investigating factors associated with time-to-event/hazard of the event

Describing Data

Survivor Function

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

Median Survival Time

$$S(t_{50}) = 0.5$$

Regression Modeling

$$\log(h(t; x)) = \log(h_0(t)) + \beta x$$