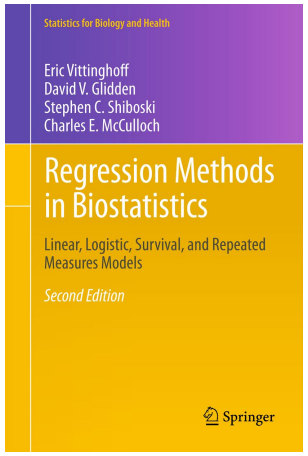
Biostat 209
Survival Data I

John Kornak
March 31, 2016
John.kornak@ucsf.edu

Reading VGSM 3.5 (review) and 6.1 - 6.2.4



http://www.epibiostat.ucsf.edu/biostat/vgsm/

Logistics

- Lectures in Mission Hall Room 1400 every Tuesday 10.30-12.00 ***EXCEPT*** additional lecture Thursday 4/21/16 (also in 1400)
- Labs every Thursday in room 1400 10.30-12.30 (except 4/21/16) + 2 homework sessions after lectures
- 3.5 lectures survival analysis, 1.5 common biostatistical problems (John Kornak), 3 repeated measures analysis (Dr. McCulloch)
- Lectures are to be recorded and posted online
- Note that homework given out on 4/19/16 will be due at start of lecture on 4/21/16
- Grades based on 5 hw assignments (70%) + data analysis project (30%, due 5/19/16)
- Check materials online (and Biostat 209 discussion forum?)

Data Analysis Project

- You need to complete a data analysis project
- DOES NOT need to be on survival analysis, but would preferably use regression methods from Biostat 208 or 209 (see project guidelines on web site syllabus page)
- A 2,000 word write-up is due on 5/19/16
- Presentation sessions on 5/24/16, 5/26/16, 5/31/16, and 6/2/16
- Email me by 4/5/16 (next Tuesday)

1. **whether unavailable for any presentation sessions?**
2. one **brief** paragraph description of your data **(within the email – not in a separate file)**, primary aim, possible/ probable analysis plan
3. if working with a biostatistician? (name if yes)

Data Analysis Project

- I will assign you to a faculty advisor
- Your advisor will meet with you (at your request) to help guide your analysis
- Students in groups of 6 for each presentation session
  - 25 minute presentation, 3 hour session
- A faculty advisor (not necessarily your own) will preside over your presentation session
- Please be considerate and allow sufficient time to schedule meetings
- Example projects and presentations on Syllabus page

TICR Professional Conduct Statement

- I will maintain the highest standards of academic honesty
- I will neither give nor receive aid in examinations or assignments unless such cooperation is expressly permitted by the instructor (ok for mutual collaboration on Survival Biostat 209 assignments)
- I will conduct research in an unbiased manner, report results truthfully, and credit ideas developed and work done by others
- I will not use answer keys from prior years
- I will write answers in my own words, and, when collaboration is permitted, acknowledge collaborators when answers are jointly formulated (here primarily for projects)

Survival Analysis Overview

- Kaplan-Meir survival curves, logrank test & Wilcoxon
- Cox proportional hazards modeling
- Cox proportional hazard model checking
- Extensions to the Cox proportional hazards model – time-varying covariates, competing risks
- Overview of other methods: clustered data, interval censoring…
- All methods illustrated with STATA
- Ask questions

In this lecture…

- Survival data and censoring
- Functions for describing survival
- Kaplan-Meier (KM) curves
- Log-rank & Wilcoxon test
- Hazard function
- Proportional hazards model
- Cox model
- The Cox model in STATA (`stcox`)

Survival data examples

Measure time to an event on a set of subjects

- Years to death post surgery
- Months to tumor recurrence post-resection
- Weeks to relapse post rehab
- Days to hospital discharge post-surgery
- Hours until a battery runs out

Different types of outcome with different time-scales

Survival data
have special features

Example – Lymphoma data

Survival (in days) of patients with lymphoma

6, 19, 32, 42, 42, 43*, 94, 126, 169*, 207, 211*,
227, 253, 255*, 270*, 310*, 316*, 335*, 346*

… so what makes survival data special/different?

Example – Lymphoma data

Survival (in days) of patients with lymphoma

6, 19, 32, 42, 42, 43*, 94, 126, 169*, 207, 211*,
227, 253, 255*, 270*, 310*, 316*, 335*, 346*

… so what makes survival data special/different?

Answer: right censoring, * = still alive at end of follow up

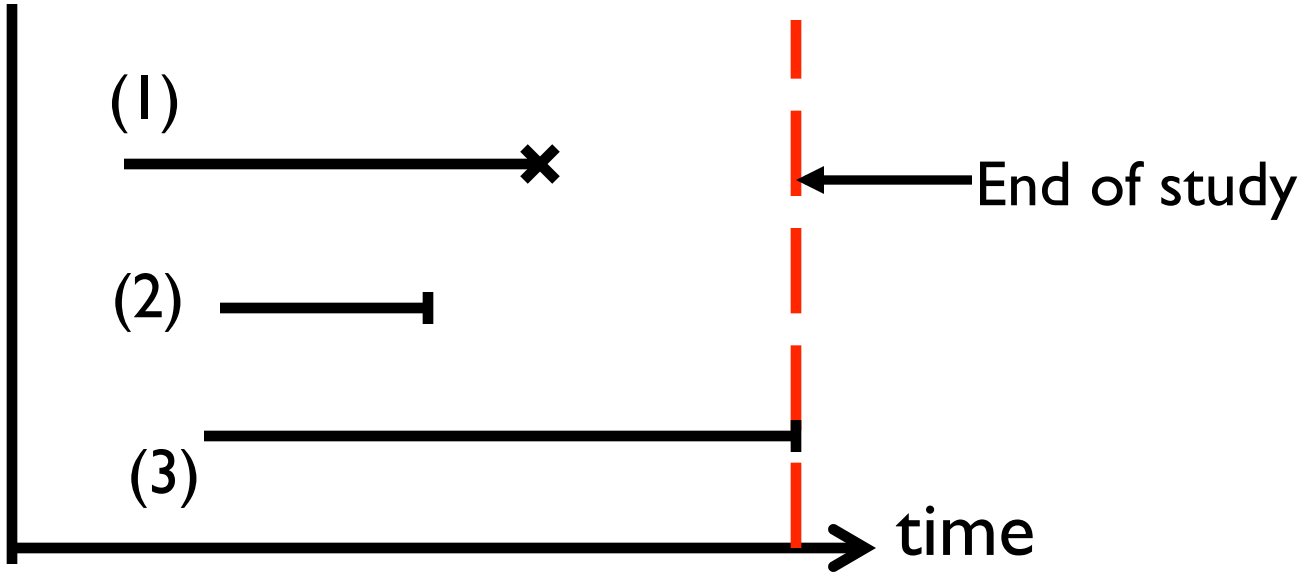
Right Censoring

Definition:

A survival time is "right censored at time t" if we only know know that it is greater than t

Example: A subject is followed for 18 months. Follow up ends and the subject is alive. The subject is then *right censored* at 18 months

Right censoring – not everyone is followed to death



(1) Subject followed to death (not censored)
(2) Subject right censored by loss to follow up
(3) Subject administratively censored: end of study

We will assume censoring times are *independent* of survival, i.e. censoring times are *non-informative*.

Pediatric Kidney Transplant

Example of Survival Data

- United Network for Organ Sharing (UNOS) database
- 9750 children under 18 yrs with kidney tx,1990-2002
- Outcome: time to death post transplant
- *38,000 patient years, 429 deaths*
- What are predictors of post tx mortality?
  e.g., donor source: cadaveric v. living

Features of UNOS Data

- Risk of death depends on length of follow-up (an individual has more chance of dying during the study if followed for 10 vs. 5 yrs)
- Follow-up ranges from 1 day to 12.5 yrs
- Most children are alive at end of study
  438 of 9750 subjects have events (4.5%)
- Thus, 95.5% of subjects are *right censored*

UNOS Data

$Y_i = \texttt{time}$ (years since transplant)

$$\delta_i = \texttt{indic} \begin{cases} 1: & \text{Death at } Y_i \\ 0: & \text{Alive as of } Y_i \end{cases}$$

$$X_i = \texttt{txtype} \begin{cases} 1: & \text{Cadaveric} \\ 0: & \text{Living} \end{cases}$$

Survival Data in STATA

- Declare the data to be survival data
- Use stset command
- stset Y, failure(δ) [can shorten to f(δ)]
  UNOS data:
  stset time, failure(indic)
- In STATA, code δ carefully
  1 should be event, 0 a censored observation
  read stset output summary to check

How can we analyze survival data?

What happens if we try regression methods from Biostat 208?

Treat as Continuous?

Question: What is the mean survival time?

- Outcome: Time to Death
- Problem:

Treat as Continuous?

Question: What is the mean survival time?

- Outcome: Time to Death
- Problem: Most subjects still alive at study end
- Mean time of death can be highly misleading
- Example: 100 subjects censored at 500 months and only 2 deaths at 1 and 3 months
- Mean time to death: 2 months....misleading!

Mean time to death is not generally a useful summary for survival data!

Treat as Binary?

Question: what proportion of subjects survive?

- Outcome: Whether “death” occurs
- Problem:

Treat as Binary?

Question: what proportion of subjects survive?

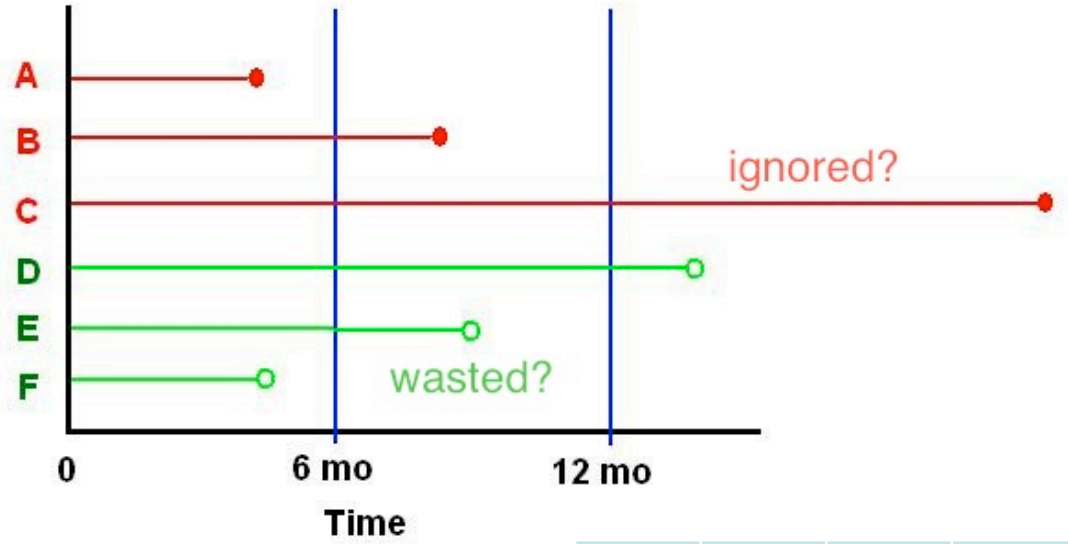
- Outcome: Whether "death" occurs
- Problem: Different follow up times
- Requires an arbitrary cut-off time (e.g., 1 yr)
- What if a subject is censored at 360 days?
  their data is wasted: not followed for a year
- Deaths after 1 year are ignored, *but most deaths may occur after 1 year!*

Treat as Binary?

e.g., cut-off at 6 months or 1 year?

observed events

censored data



	A	B	C	D	E	F
6 mo	✓	✗	✗	✓	✓	✗
12 mo	✓	✓	✗	✓	✗	✗

Right *and* left censoring with no mechanism to deal with it!

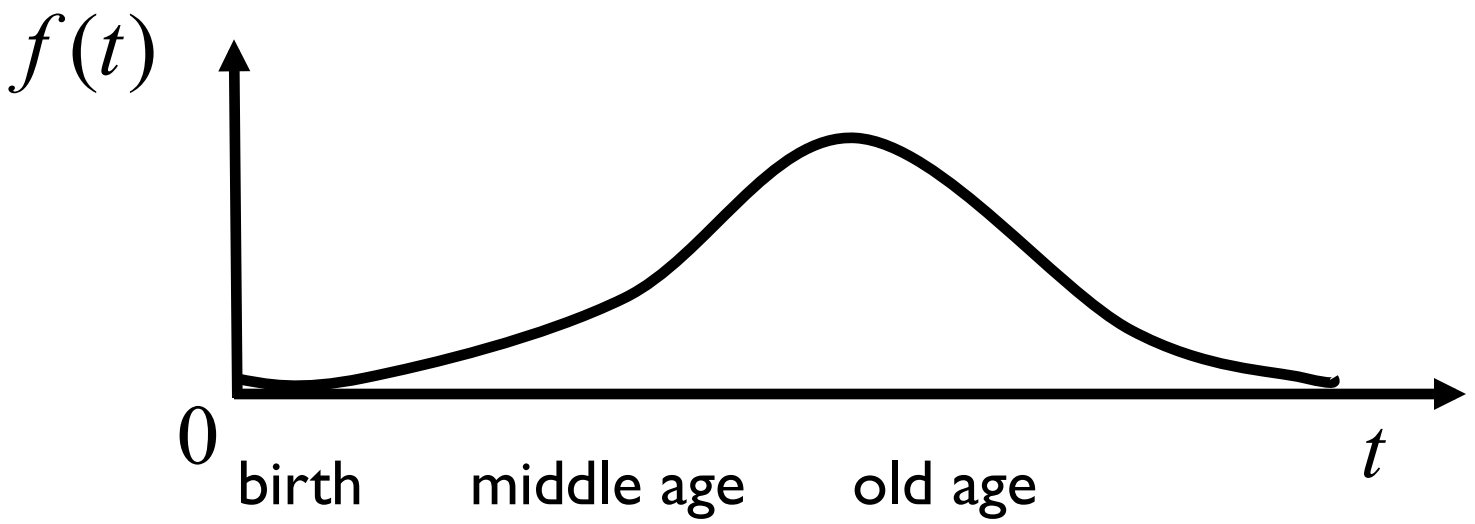
Aims of a Survival Analysis

- Summarize the distribution of survival times
  Tool: Kaplan-Meier estimates (of survival distribution)
- Compare survival distributions between groups
  Tool: logrank test / Wilcoxon test
- Investigate predictors of survival
  Tool: Cox regression model

Kaplan-Meier and logrank covered previously but will review

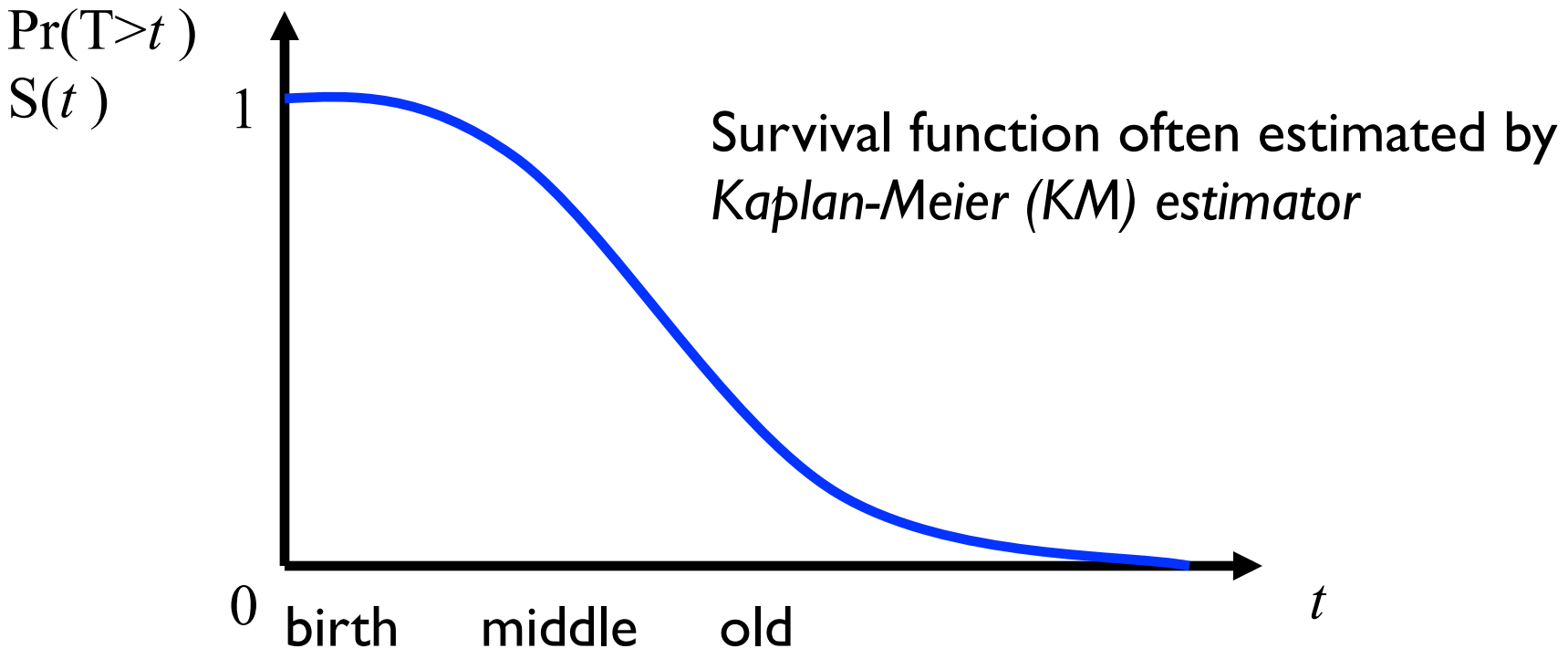
Functions for Describing Survival

Probability density (distribution) - of death



Functions for Describing Survival

Survival function: the probability an individual will survive longer than a particular time t



Survival function often estimated by *Kaplan-Meier (KM) estimator*

Estimating the Survival function with the Kaplan-Meier (KM) estimator

The probability of death in any particular time interval can be estimated by:

$$\frac{\text{\# observed deaths}}{\text{\# at risk}}$$

E.g. in a study of 2000 genetically bred mice, 230 died of heart failure between 0 and 3 weeks; estimate probability of death between 0 and 3 weeks to be 230/2000

Estimating the Survival function with the Kaplan-Meier (KM) estimator

- The survival curve = 1 at t=0 (everyone is alive early on)
- KM approach assumes that at the times we observe deaths survival curve should drop *proportionally*.
- Note that the KM estimator must account for the censored data! Once a subject is censored their death cannot be observed and observable "at risk" set drops.
- The subsequent heights of drops for each death in the KM survival curve increases with each individual lost to follow up.

Kaplan-Meier

To generate a KM estimated survival curve we consider intervals of "zero" length in time:

$$\Pr(\text{death at time } t) = \frac{\#\text{ observed deaths at time } t}{\#\text{ at risk at time } t}$$

This leads to an instant drop in the survival curve of size "Pr(death at time *t*)" every time there is a death (or an "event").

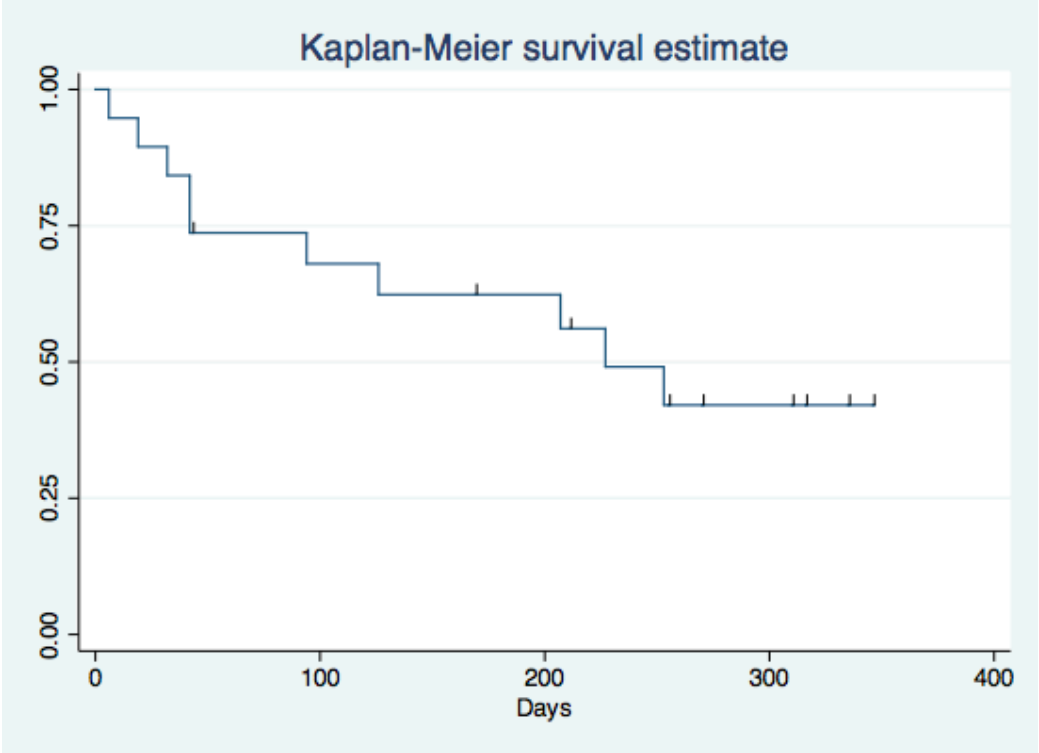
Note that the "number at risk" in the denominator accounts for the censored data! Once a subject is censored they are no longer in the "at risk" group.

Kaplan-Meier

Lymphoma data

```
load lymph.dta      // Stata commands
stset Days, failure(relapse)
sts list
sts graph, censored(single) xtitle(Days)
```

Days	Beg. Total	Fail	Net Lost	Survivor Function
6	19	1	0	0.9474
19	18	1	0	0.8947
32	17	1	0	0.8421
42	16	2	0	0.7368
43	14	0	1	0.7368
94	13	1	0	0.6802
126	12	1	0	0.6235
169	11	0	1	0.6235
207	10	1	0	0.5611
211	9	0	1	0.5611
227	8	1	0	0.4910
253	7	1	0	0.4209
255	6	0	1	0.4209
270	5	0	1	0.4209
310	4	0	1	0.4209
316	3	0	1	0.4209
335	2	0	1	0.4209
346	1	0	1	0.4209

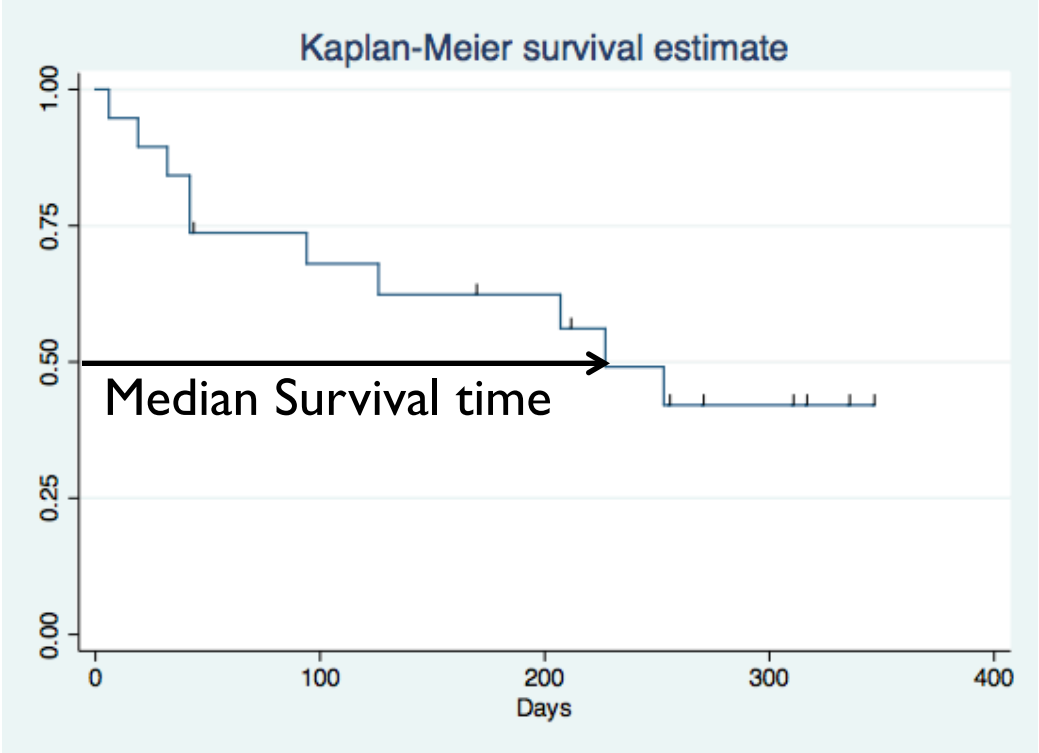


Kaplan-Meier

Lymphoma data

```
load lymph.dta      // Stata commands
stset Days, failure(relapse)
sts list
sts graph, censored(single) xtitle(Days)
```

Days	Beg. Total	Fail	Net Lost	Survivor Function
6	19	1	0	0.9474
19	18	1	0	0.8947
32	17	1	0	0.8421
42	16	2	0	0.7368
43	14	0	1	0.7368
94	13	1	0	0.6802
126	12	1	0	0.6235
169	11	0	1	0.6235
207	10	1	0	0.5611
211	9	0	1	0.5611
227	8	1	0	0.4910
253	7	1	0	0.4209
255	6	0	1	0.4209
270	5	0	1	0.4209
310	4	0	1	0.4209
316	3	0	1	0.4209
335	2	0	1	0.4209
346	1	0	1	0.4209



Example – ALL Data

Comparing survival curves between 2 groups:

6-mercaptopurine (6-MP) maintenance thpy for children in remission from acute lymphoblastic leukemia (ALL)
=> outcome = time to relapse (weeks)

Placebo: `1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23`

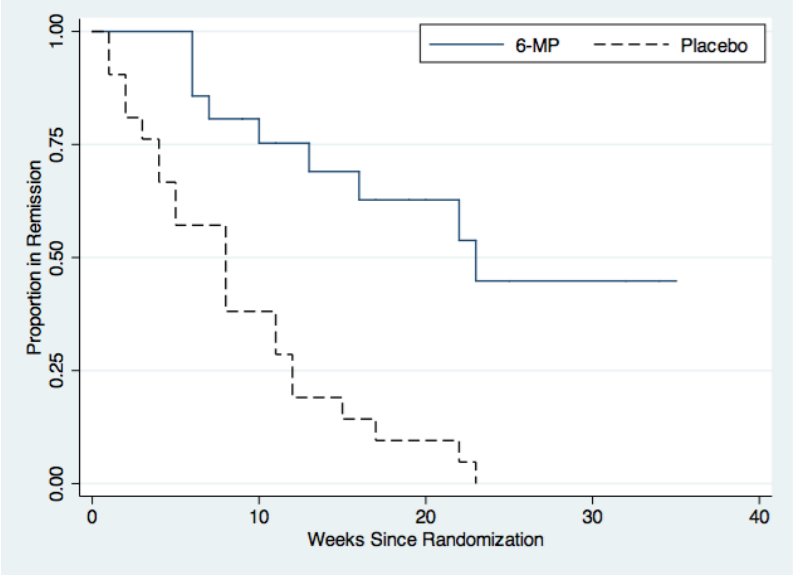
6-MP: `6,6,6,6*,7,9*,10,10*,11*,13,16,17*,19*,20*,22,23,25*,32*,34*,35*`

… is there a difference in time to relapse of children on 6-MP vs. Placebo?

Curves for ALL data

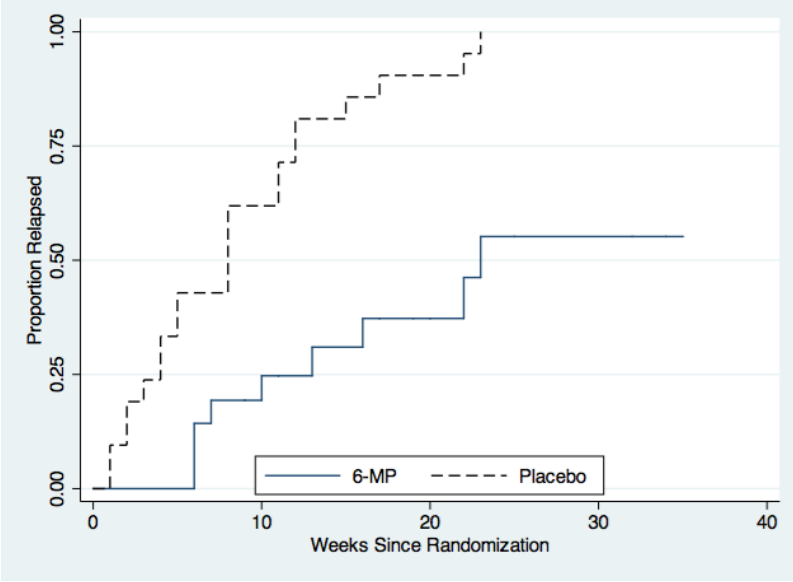
K-M Survival function

```
use leuk.dta,
stset time, f(cens)
sts graph, by(group)
```



Cumulative relapse incidence

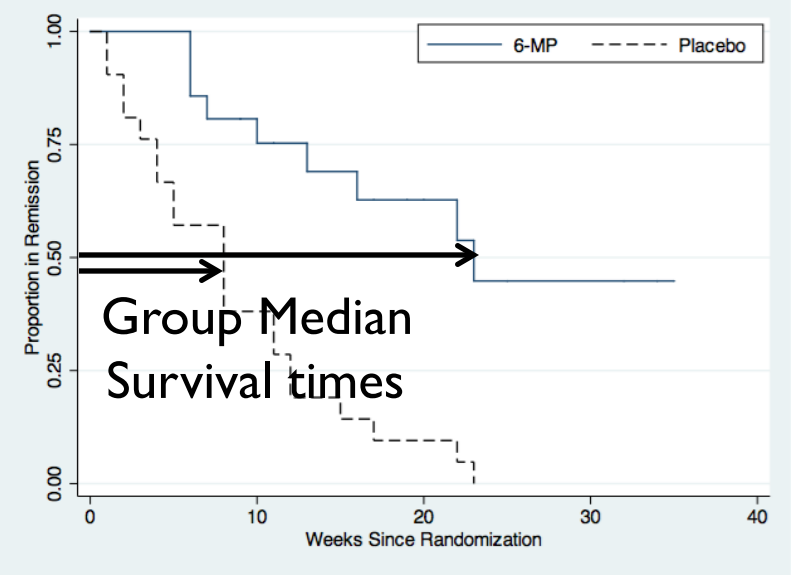
```
use leuk.dta,
stset time, f(cens)
sts graph, by(group) failure
```



Curves for ALL data

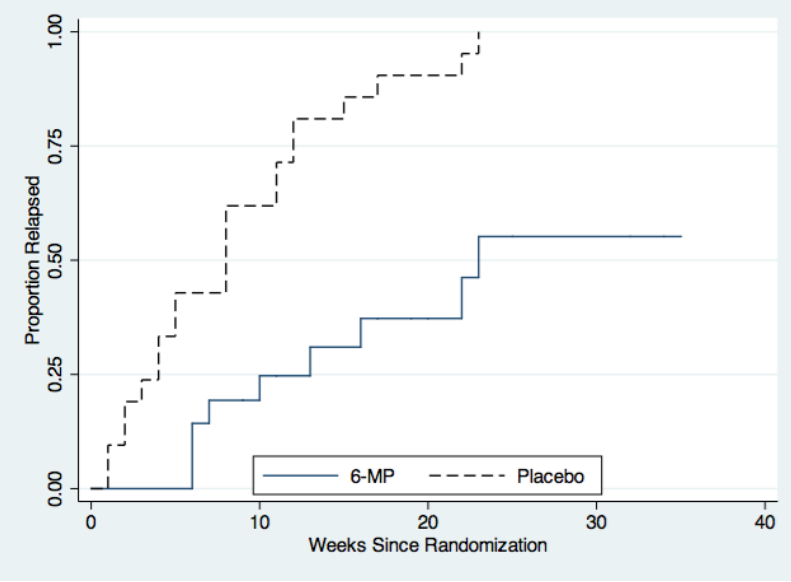
K-M Survival function

```
use leuk.dta,
stset time, f(cens)
sts graph, by(group)
```



Cumulative relapse incidence

```
use leuk.dta,
stset time, f(cens)
sts graph, by(group) failure
```



Non-parametric tests for comparing Two Survival Curves

- Logrank test compares different KM estimated survival functions (e.g. between groups)
- Does not give estimated size of effect – use Kaplan-Meier curve (e.g. via median survival times) or Cox Model for that purpose
- The idea of the test is that (under the null hypothesis of "no group difference") the proportion of deaths within each group at any time should not be too different than for the combined data from all groups because *all groups have the same survival distribution.*

Logrank test for comparing Two Survival Curves

- Basically a large sample chi-square test:
  "observed - expected" across all time points (based on ranks)
- Main alternate non-parametric test is the Wilcoxon test for censored data: used if either a) strong non-proportional hazards (see later), or b) want to give more weight to earlier events
- Logrank test is recommended as a default

Logrank test for ALL data

```
use leuk.dta
stset time, f(cens)
sts test group

        failure _d:  cens
  analysis time _t:  time


Log-rank test for equality of survivor functions

        |   Events         Events
group   |  observed       expected
--------+-------------------------
6 MP    |         9           19.25
Placebo |        21           10.75
--------+-------------------------
Total   |        30           30.00

             chi2(1) =      16.79
             Pr>chi2 =     0.0000
```

Wilcoxon test for ALL data

```
use leuk.dta
stset time, f(cens)
sts test group, wilcoxon
 failure _d:  cens
   analysis time _t:  time


Wilcoxon (Breslow) test for equality of survivor
functions

         |   Events          Events         Sum of
group    |  observed        expected         ranks
---------+-------------------------------------------
6 MP     |         9           19.25          -271
Placebo  |        21           10.75           271
---------+-------------------------------------------
Total    |        30           30.00             0

                chi2(1) =      13.46
                Pr>chi2 =     0.0002
```

Summary of KM and logrank in STATA

- `stset` command declares outcome as survival
  doesn't need to be specified again
- `sts list, by(txtype)`
  print Kaplan-Meier by variable txtype
- `sts graph, by(txtype)`
  graph Kaplan-Meier by variable txtype
- `sts test txtype`
  calculates logrank test for variable txtype

Aims of a Survival Analysis

- Summarize the distribution of survival times
  Tool: Kaplan-Meier estimates (of survival distribution)
- Compare survival distributions between groups
  Tool: logrank test
- Investigate predictors of survival
  Tool: Cox regression model

Regression by Outcome

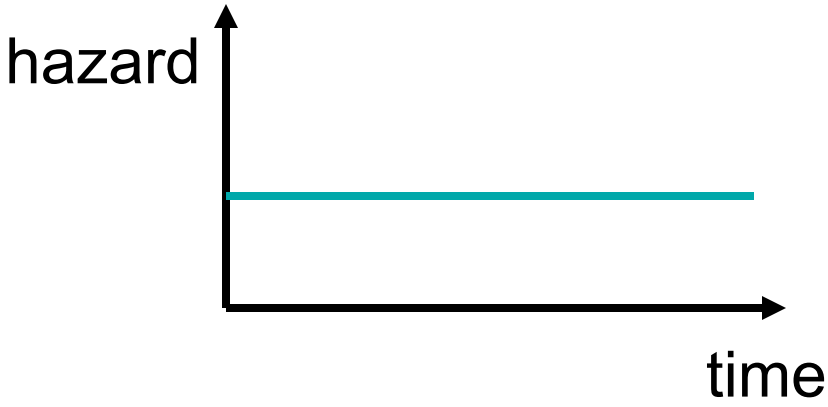
Outcome Data	Regression	Data Summary
Continuous	Linear	Mean
Binary	Logistic	Odds
Survival	Cox	Hazard

Functions for Describing Survival: Hazard Function

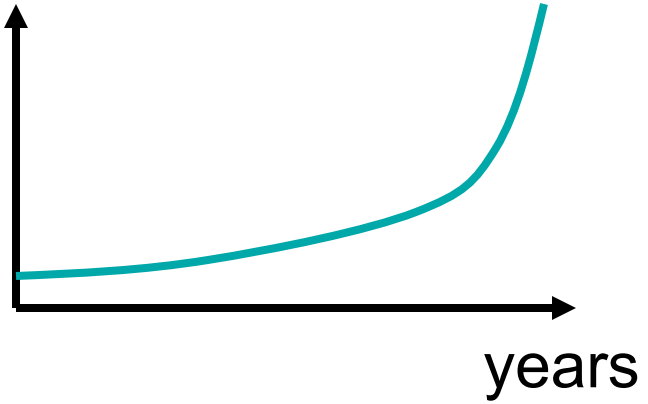
- Rate of failure per (small) unit time
- Hazard is like an instantaneous (daily?) death rate:
  h(t) = # die at day t / # followed to t
- Rate of death among those still alive/followed
- Easily estimated for censored data
- A measure of "risk"
  higher hazard => greater risk of death
- Of course, outcome does not have to be death

Hazard function examples

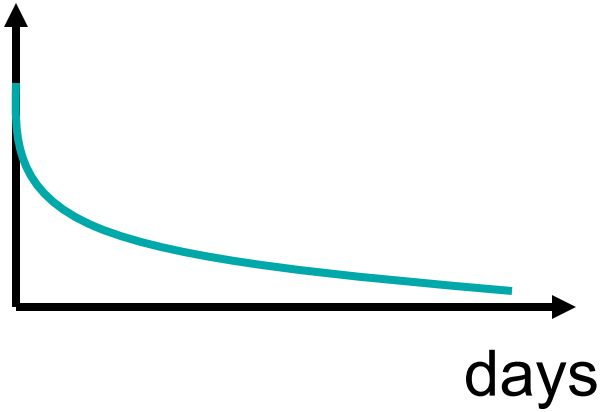
Exponential survival



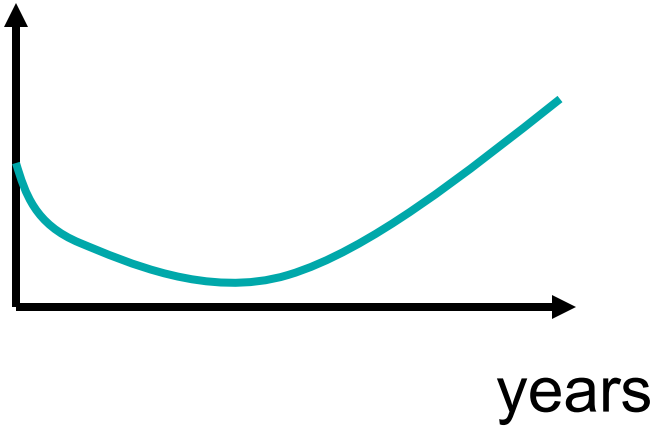
Normal aging



Post heart attack risk



Post heart bypass surgery

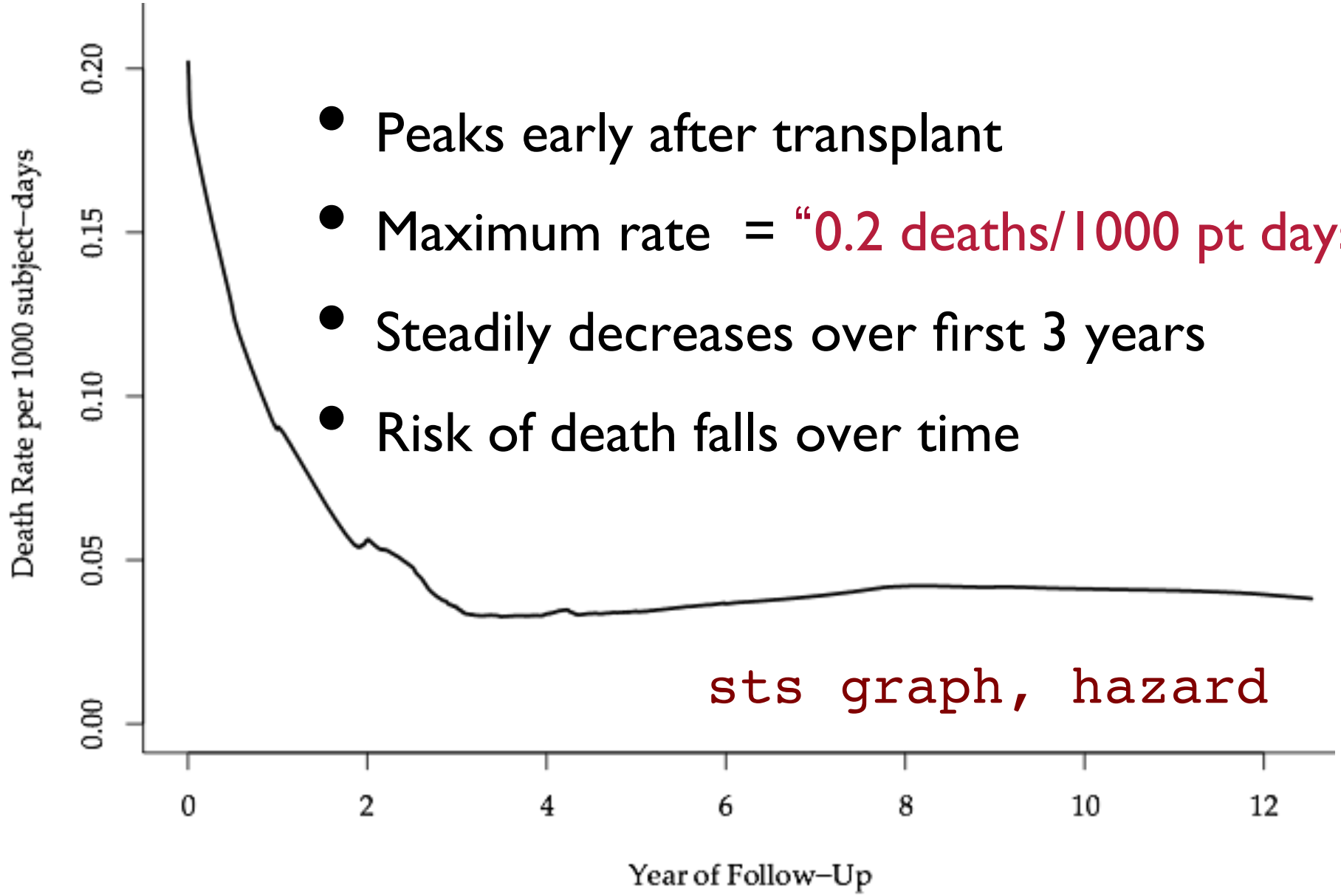


UNOS Data: Daily Death Rate

Day of FU	No. Fol.	No. Died	No. Cens.	Death Rate	Death Rate per 1000 Subj-Days
1	9752	7	14	7/9752	0.72
2	9731	5	8	5/9731	0.51
3	9718	5	12	5/9718	0.51
4	9701	7	41	7/9701	0.72
5	9653	3	54	3/9653	0.31
6	9596	2	57	2/9596	0.21
7	9537	0	50	0/9537	0.00
8	9487	4	49	4/9487	0.42
9	9434	1	49	1/9434	0.11
10	9384	3	28	3/9384	0.32

Let's smooth these and extend

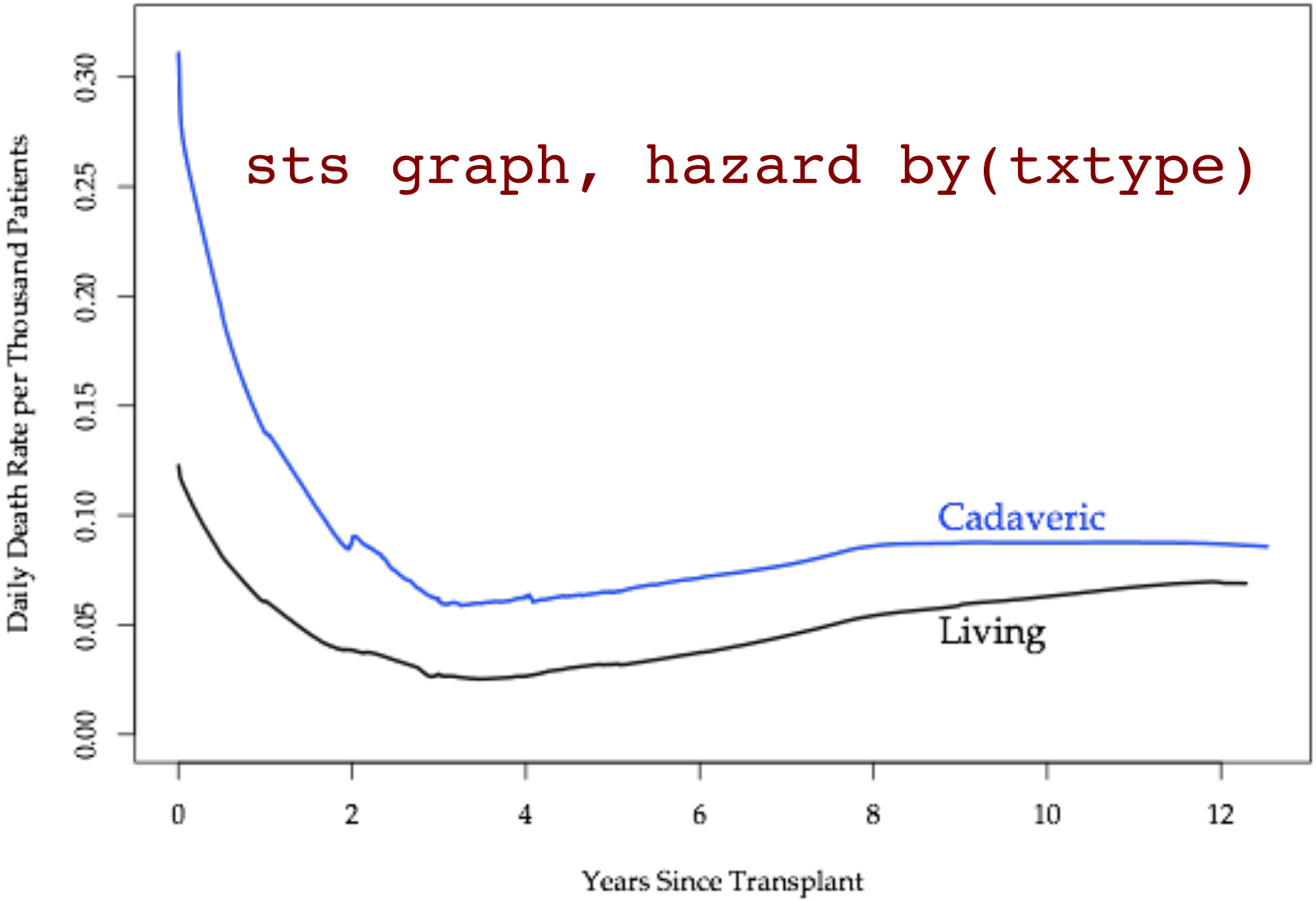
h(t) for UNOS data



- Peaks early after transplant
- Maximum rate = "0.2 deaths/1000 pt days"?
- Steadily decreases over first 3 years
- Risk of death falls over time

```
sts graph, hazard
```

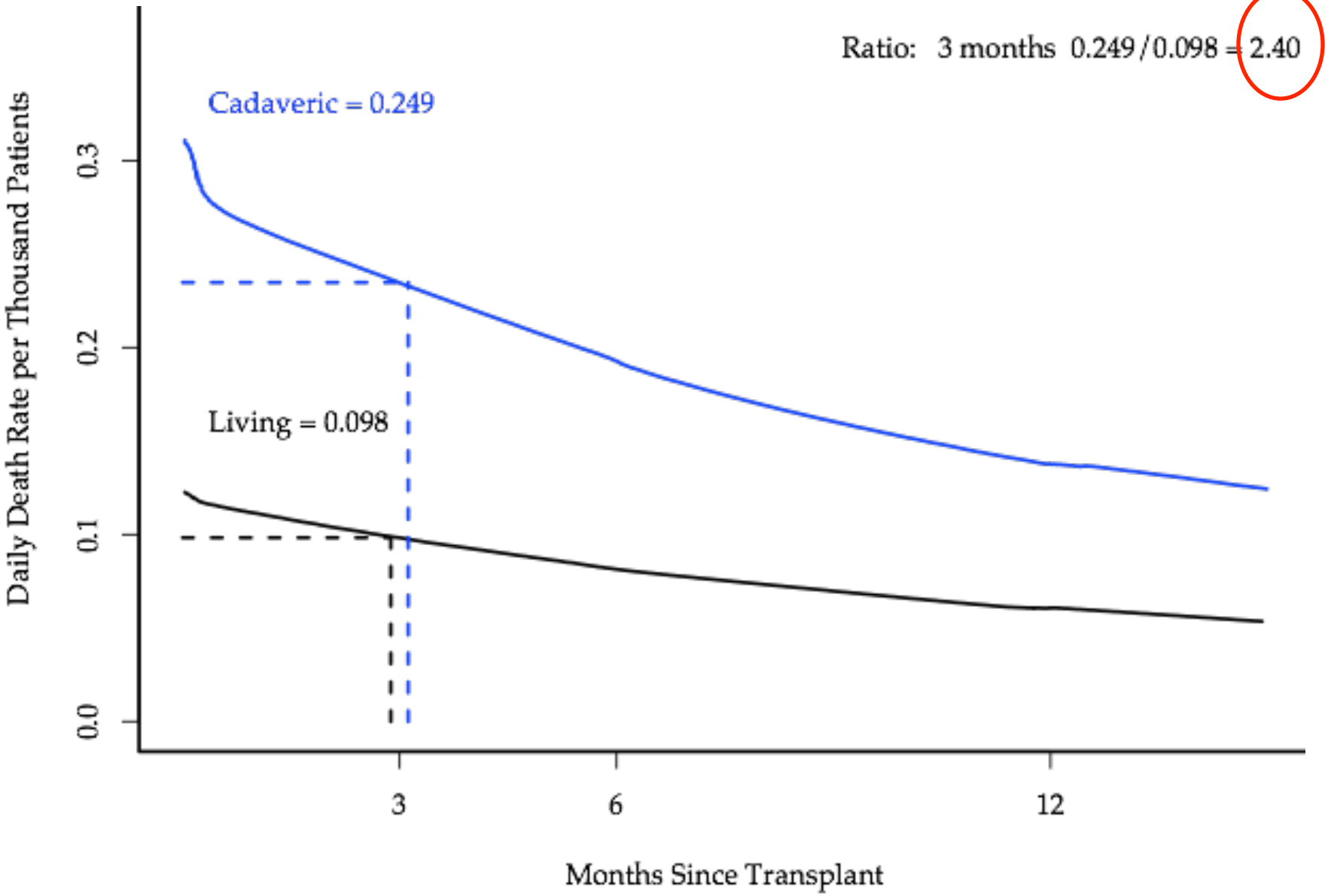
Hazard by Type of Tx



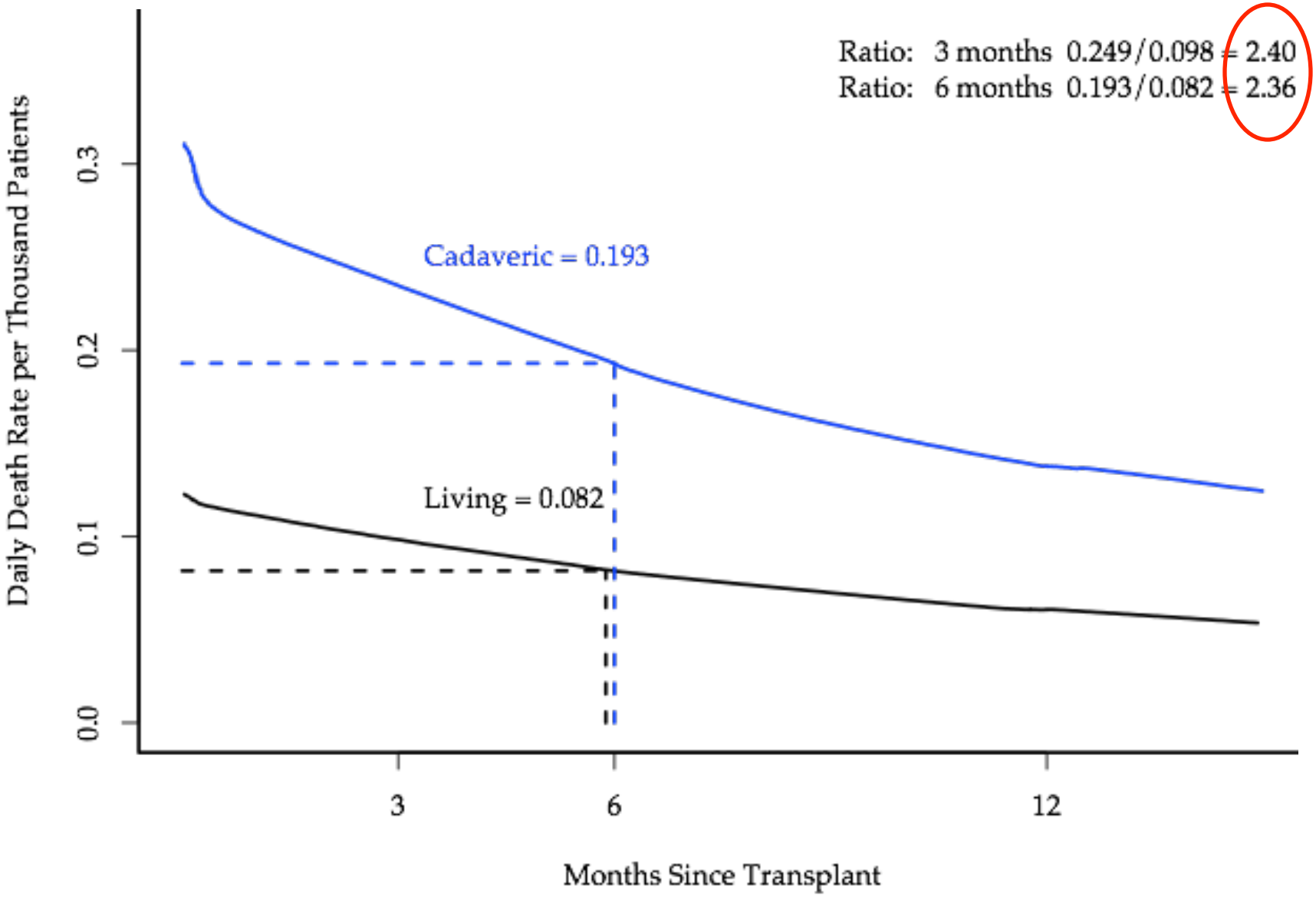
Comparing Hazards

- Hazards are between 0 and infinity (c.f. odds)
- Reasonable to divide when comparing (c.f. odds)
- Leads to “hazard ratio”
- Consider the hazard ratio at different times…

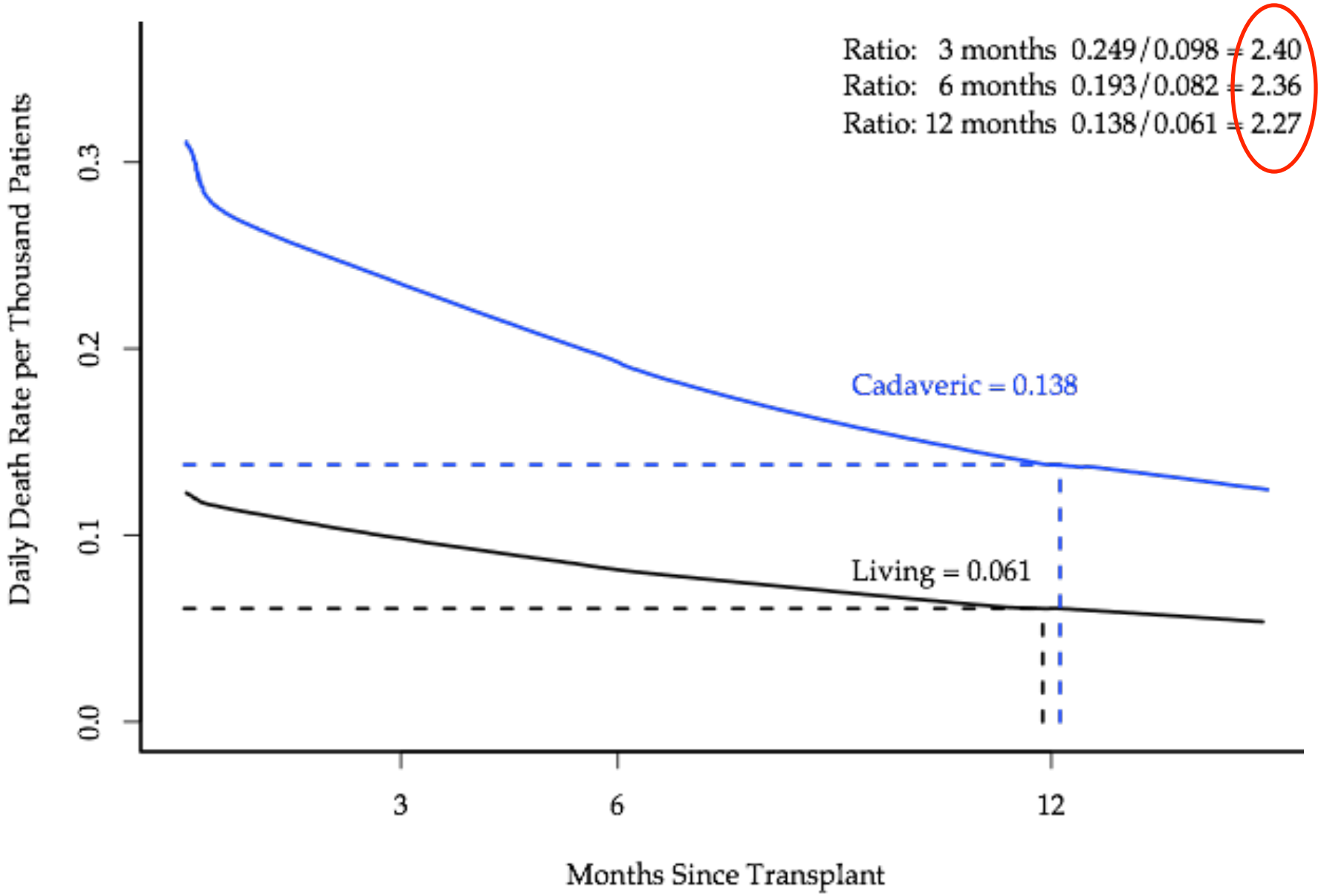
Hazard at 3 months



Hazard at 6 months



Hazard at 12 months



Hazard rates over time

	1000 × Haz Rate		*Relative*
Time	*Cadaveric*	*Living*	*Rate*
3mo	0.235	0.098	2.40
6mo	0.193	0.082	2.36
1yr	0.138	0.061	2.27
2yr	0.088	0.038	2.30
3yr	0.061	0.027	2.25
4yr	0.063	0.026	2.37
5yr	0.065	0.032	2.03

- $h_0(t)$: hazard for living recipients at t (col. 3)
- $h_1(t)$: hazard for cadaveric recipients at t (col. 2)
- $h_1(t)/h_0(t)$: relative short-term risk (col. 4)
  "hazard ratio" at time t

Estimated hazard ratio differs greatly over time?

Proportional hazard ratio

- If $h_1(t) = r h_0(t)$ for all t, hazards are proportional, i.e., *hazards have the same shape* (proportional curves)
- *r* is the relative hazard (constant hazard ratio)
- *r* is a useful quantity for discussing predictor effects

Hazard Ratio in UNOS Data

- Hazards changed with time
  cadaveric: 0.24 at 3 mos, 0.07 at 5 years
  living: 0.10 at 3 mos, 0.03 at 5 years
- But, their ratio (estimated hazard ratio) didn't vary much: between 2.03 and 2.40
- "Cadaveric recipients have a little over twice the death rate of living recipients"
- Proportional hazards model: gives predictor effects based on hazards, i.e., regression estimates in terms of hazard ratios.

The Proportional Hazards Model for Survival Data

- $x = (x_1, \ldots, x_p)$: the set of predictors
- $h(t|x)$: hazard of someone with predictors x
- $h(t|x) = h_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$
- $h(t|x)/h_0(t) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$
- $\log(h(t|x)) = \log(h_0(t)) + \beta_1 x_1 + \ldots + \beta_p x_p$
  because $\log(a/b) = \log(a) - \log(b)$
- Formulation is much like logistic regression but change *odds* to *hazards*

Cox Model

- The "baseline" hazard $h_0(t)$ is unspecified; *plays the role of intercept*
- Predictor effects in terms of hazard ratios; *relative rates of failure*
- Key point: don't need to know $h_0(t)$ to understand the effects of predictors
- The effect of one unit increase in predictor x_p is to multiply hazard by $\exp(\beta_p)$ (holding all other predictors constant)….

+1 unit change in x_p

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p V) \qquad x_p = V$$

versus

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p (V+1)) \qquad x_p = V+1$$

$$\text{Ratio} = \frac{h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p (V+1))}{h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

+1 unit change in x_p

$$\text{Ratio} = \frac{\cancel{h_0(t)}\exp(\beta_1 x_1 + \ldots + \beta_p(V+1))}{\cancel{h_0(t)}\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

$$= \frac{\exp(\beta_1 x_1 + \ldots + \beta_p (V+1))}{\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

Ratio does not depend on t !

+1 unit change in x_p

$$\text{Ratio} = \frac{\exp(\beta_1 x_1 + \ldots + \beta_p (V+1))}{\exp(\beta_1 x_1 + \ldots + \beta_p V)}$$

$$= \exp(\beta_1 x_1 + \ldots + \beta_p (V+1) - (\beta_1 x_1 + \ldots + \beta_p V))$$

because exp(a)/exp(b) = exp(a-b)

$$= \exp(\beta_p (V+1) - \beta_p V)$$ b/c same other predictors

$$= \exp(\beta_p)$$ b/c $\beta_p V$ terms cancel

Hazard Ratio (HR)

- β is the regression coefficient
- If a predictor variable has no effect then $\beta=0$
- HR for a unit increase in a predictor is $\exp(\beta)$ -- if predictor has no effect then $\exp(\beta) = \exp(0) = 1$
- Can readily switch from summarizing models in terms of coefficients, β, or in terms of hazard ratios, $\exp(\beta)$ – HR's work better for interpretation, but math simpler based on coefficients
- HR provides useful way to discuss predictor effects (i.e., increases or decreases Hazard by a factor)

`Stata` Command

- First use the command `stset` to declare the data as survival
- Then fit Cox model with

`stcox` *predictorlist*

Stata Output (UNOS)

```
. stcox txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          438
Time at risk    =  38236.09865

                                               LR chi2(1)       =     49.23
Log likelihood  =   -3762.6976                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   1.974683    .195328     6.88   0.000     1.626672    2.397149
------------------------------------------------------------------------------
```

(living *x=0; cadaveric x=1*)

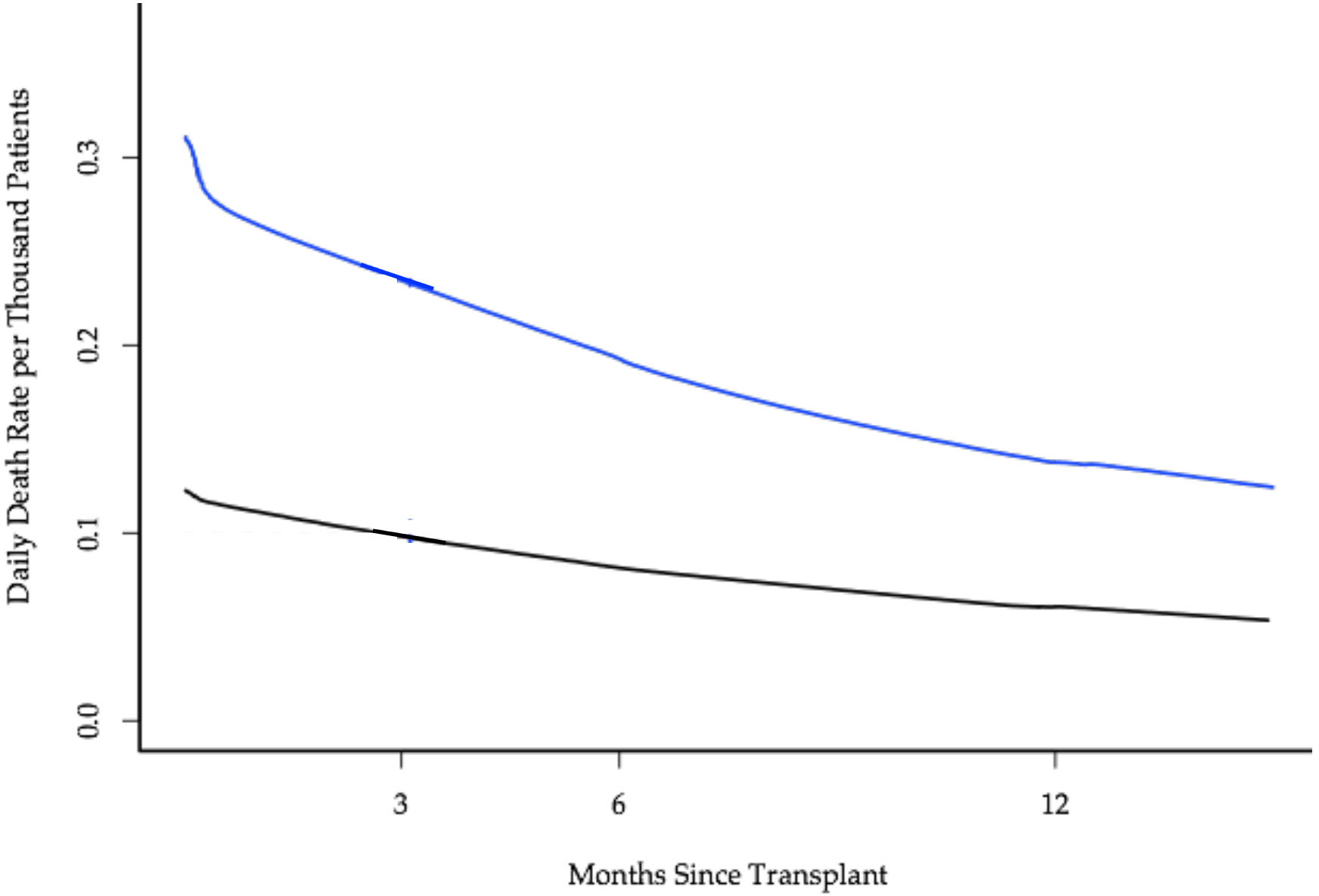
Stata Output

```
. stcox txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          438
Time at risk    =   38236.09865

                                               LR chi2(1)       =     49.23
Log likelihood  =   -3762.6976                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   1.974683   .195328     6.88   0.000     1.626672    2.397149
------------------------------------------------------------------------------
```

Estimated hazard ratio = 1.97: *hazard of death about double for cadaveric recipients*

(living *x=0; cadaveric x=1*)

Hazard over 12 months



Stata Output

```
. stcox txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          438
Time at risk    =  38236.09865
                                               LR chi2(1)       =     49.23
Log likelihood  =   -3762.6976                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   1.974683   .195328     6.88   0.000     1.626672    2.397149
------------------------------------------------------------------------------
```

SE of hazard ratio

Stata Output

```
. stcox txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          438
Time at risk    =  38236.09865
                                               LR chi2(1)       =     49.23
Log likelihood  =   -3762.6976                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   1.974683    .195328     6.88   0.000     1.626672    2.397149
------------------------------------------------------------------------------
```

Wald test p-value < 0.05: *group hazards are statistically significantly different*

Stata Output

```
. stcox txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          438
Time at risk    =  38236.09865
                                               LR chi2(1)       =     49.23
Log likelihood  =   -3762.6976                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   1.974683    .195328     6.88   0.000     1.626672    2.397149
------------------------------------------------------------------------------
```

Likelihood ratio test p-value < 0.05: *group hazards are statistically significantly different*

Stata Output

```
. stcox txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          438
Time at risk    =  38236.09865
                                               LR chi2(1)       =     49.23
Log likelihood  =   -3762.6976                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   1.974683    .195328     6.88   0.000     1.626672    2.397149
------------------------------------------------------------------------------
```

95% CI for Hazard ratio: *A plausible range for the hazard ratio is between 1.63 and 2.40.*

Interpretation

"The hazard ratio of mortality for the recipient of a cadaveric kidney is about 1.97 relative to a living organ ($p < 0.001$).
The 95% CI for the hazard ratio is 1.63 to 2.40"

Survival by HLA loci

Number of matching HLA loci range from 0 to 6

```
. stcox i.hla
No. of subjects =         9517                 Number of obs    =      9517
No. of failures =          424
Time at risk    =  37439.62467                 LR chi2(6)       =     47.20
Log likelihood  =   -3629.6294                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hlamat |
           1 |   .9543573   .1697015    -0.26   0.793      .673523    1.352289
           2 |    .697097   .1288643    -1.95   0.051      .485223    1.001486
           3 |   .4644883   .0788633    -4.52   0.000      .333007    .6478825
           4 |   .4249741    .090394    -4.02   0.000     .2800966    .6447883
           5 |   .5870101   .1700124    -1.84   0.066     .3327492    1.035558
           6 |   .3834296   .1396405    -2.63   0.008     .1877968    .7828581
------------------------------------------------------------------------------
```

Interpretation

HLA is a statistically significant predictor of mortality
Chi2=47, $p < 0.0001$ by the likelihood ratio test.

# Loci	HR compared to 0 matching loci	% Change in Hazard of Death
1	0.95	-5%
2	0.70	-30%
3	0.46	-54%
4	0.42	-58%
5	0.58	-42%
6	0.38	-62%

Effect of Age

```
. stcox age
No. of subjects =         9742                     Number of obs   =      9742
No. of failures =          437
Time at risk    =  38217.71783
                                                   LR chi2(1)      =     23.62
Log likelihood  =   -3766.4598                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9581318   .0083224    -4.92   0.000     .9419582    .9745831
------------------------------------------------------------------------------
```

Interpretation

Age is a statistically significant predictor of mortality
$p < 0.001$, by the Wald test or likelihood ratio test.
Each one-year increase in age (at transplant)
reduces the hazard of mortality by 4.2%,
95% CI (2.5% to 5.8% reduction)

Why the Cox Model?

- Can be fitted without an explicit model for the hazard
- Can model the effect of a continuous predictor
- Can model multiple predictors: *continuous, binary, categorical*
- Can adjust for confounders: *adjust by adding confounders to the model*
- Can incorporate interaction, mediation: *create and add product terms (simpler in Stata 11 on)*
- Can test for and estimate predictors for patient-level prognosis

Comparison with other forms of regression

- *Same issues as in linear and logistic regression:* predictor selection

- *differences:* interpretation, assumptions, model checking

Main take home points

- **Survival Data**
  - characterized by censoring
  - requires special methods: Kaplan-Meier, logrank, Cox
- **Cox Model**
  - based on hazard functions
  - assumes proportional hazards
  - uses hazard ratio for predictor effects
  - does not require model for baseline hazard
  - important similarities with other regressions

Next Lecture

- Binary, categorical and continuous predictors in the Cox model
- Review of interaction effects
- Confounding, mediation and interactions in the Cox model
- Adjusted survival curves

Don't forget…

- Lab this Thursday Rm 1400 Mission Hall (with Julie Morisset, julie.morisset@ucsf.edu or Anjana Sharma, anjana.sharma@ucsf.edu) - lab is posted - recommend looking in advance
- Project idea and presentation date restrictions to me by next Tuesday 4/7/16 (+Look at posted project guidelines)
- First homework (posted) due on 4/12/16 (will need knowledge from next lecture to complete)
- Reading for next lecture VGSM 6.2.5 - 6.2.13