

# Survival Data

Su-Chun Cheng

March 31, 2009

[scheng@biostat.ucsf.edu](mailto:scheng@biostat.ucsf.edu)

# Data Project

- You need to complete a data analysis project
- A 2,000 word write-up is due on 5/22
- Presentation sessions on 5/26, 5/28 or 6/2, or 6/4
- E-mail me by 4/7 (next week)
  - a one-paragraph description of your data and primary aim
  - working with an biostatistician? (name if yes)
  - unavailability for any presentation sessions?

# Data Project

- I will assign you to a faculty advisor
- Your advisor will meet with you to help guide your analysis
- Students in groups of 6 for one presentation session with their advisors
  - 25 minute presentation, 3 hour session
- Sessions to be scheduled with your advisor
- Details in “[project guidelines 2009.pdf](#)”

# Pediatric Kidney Transplant

- 9750 kids under 18 yo with kidney tx
- Outcome: time to death post transplant
- UNOS database covering 1990-2002  
*38,000 patient years, 429 deaths*
- What are predictors of post tx mortality?  
*e.g., donor source: cadaveric v. living*

Example of **Survival** Data

**Survival Data has special  
features**

# Features of UNOS Data

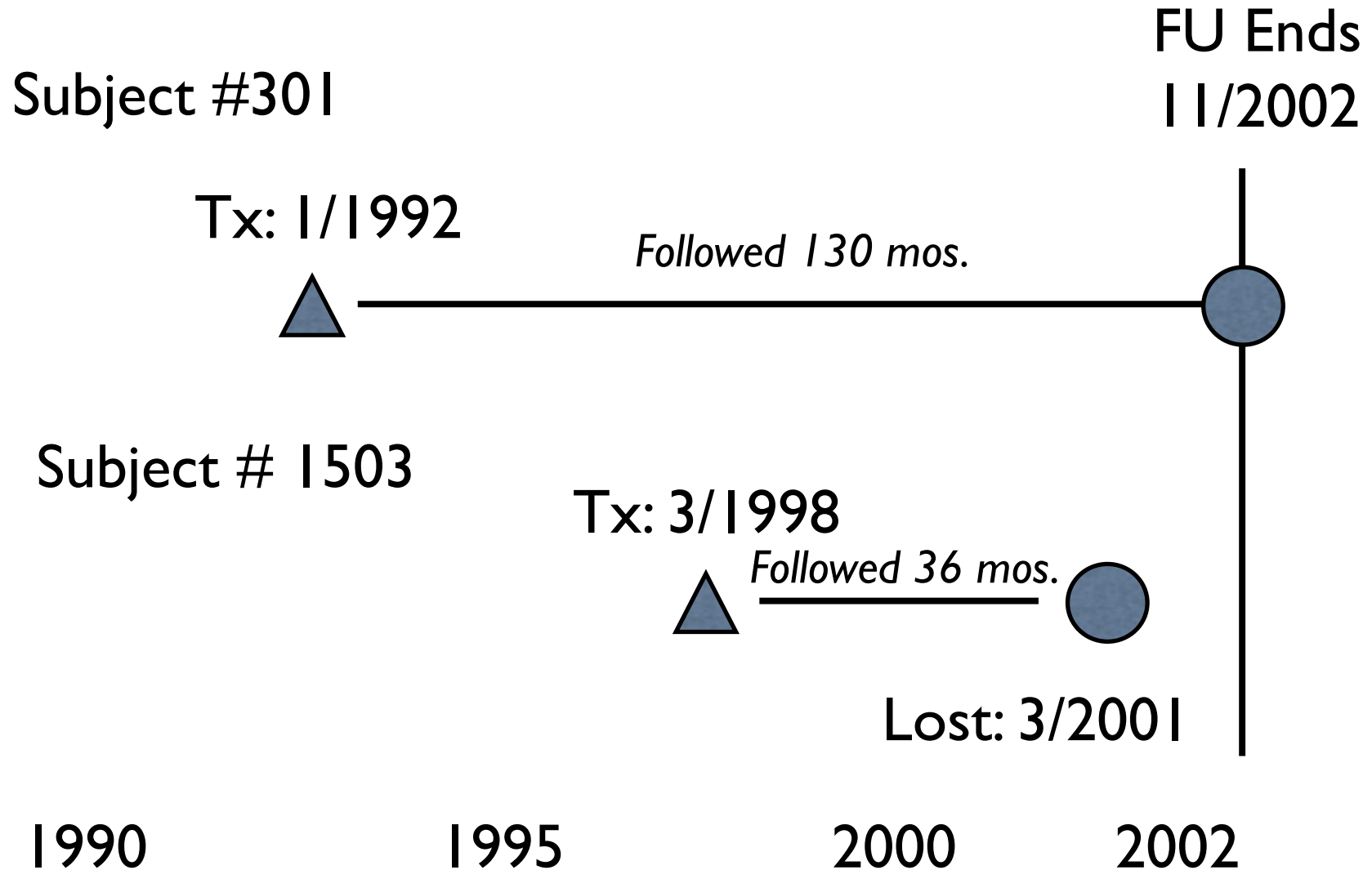
- Risk of death depends on length of follow-up
- Follow-up ranges from 1 days to 12 years
- Most kids are alive  
*429 of 9750 subjects have events (4%)*
- Thus, 96% of subjects are **right censored**

# Right Censoring

Definition: *A survival time is not known exactly but only known to be greater than some value*

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



# Right Censoring

Right censoring can occur for many reasons

- Study occurs over a finite period of time
- Staggered entry
- Loss to Follow-Up

⇒ All lead to right censoring

Assume censored subjects “not different” in risk  
(non-informative, independent censoring)

# UNOS Data

$Y_i =$  time years since transplant

$\delta_i =$  indic  $\left\{ \begin{array}{l} 1 : \text{Death at } Y_i \\ 0 : \text{Alive as of } Y_i \end{array} \right.$

$X_i =$  txttype  $\left\{ \begin{array}{l} 1 : \text{Cadaveric} \\ 0 : \text{Living} \end{array} \right.$

$i = 1, \dots, 9750$

# Survival Data in STATA

- Declare the data to be survival data
- Use `stset` command
- `stset Y, failure( $\delta$ )`  
UNOS data:  
`stset time, failure(indic)`
- In STATA, code  $\delta$  carefully  
1 should be event, 0 a censoring

# Mean for Survival Data?

- Outcome: Time to Death
- Problem: Most subjects still alive
- Average death time is greatly misleading
- Example: 100 subjects censored at 500 months and deaths at 1 and 3 months
- Average death time: 2 months....misleading!

**Mean is not a useful summary for survival data!**

# Treat as Binary

- Proportion of subjects “alive”
- Requires an arbitrary time (e.g., 1 year)
- What if a subject is censored 360 days?  
their data is wasted: not followed for year
- Deaths after 1 year are ignored  
most deaths are after 1 year

Again....can't handle censoring!

# Aims of a Survival Analysis

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

*Kaplan Meier and logrank covered previously*

# KM and logrank in STATA

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

# Regression by Outcome

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

# Hazard Function

- Rate of failure per (small) unit time
- Hazard is like a daily death rate  
 $h(t) = \# \text{ die at day } t / \# \text{ followed to } t$
- Rate of death among those alive (at risk)
- Easily estimated for censored data
- A measure of “risk”  
*higher hazard => greater risk of death*
- Doesn't have to be death

# Daily Death Rate

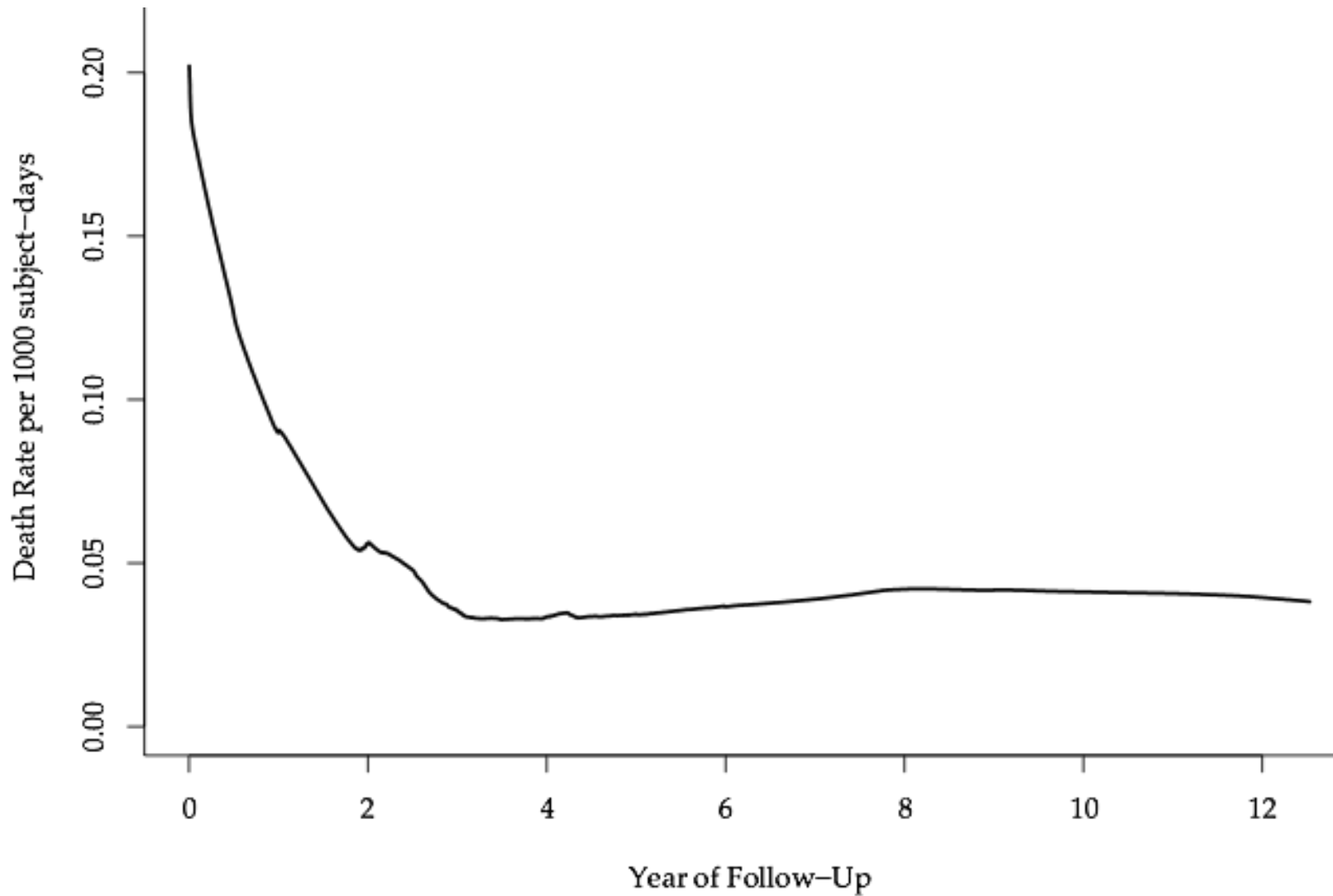
---

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

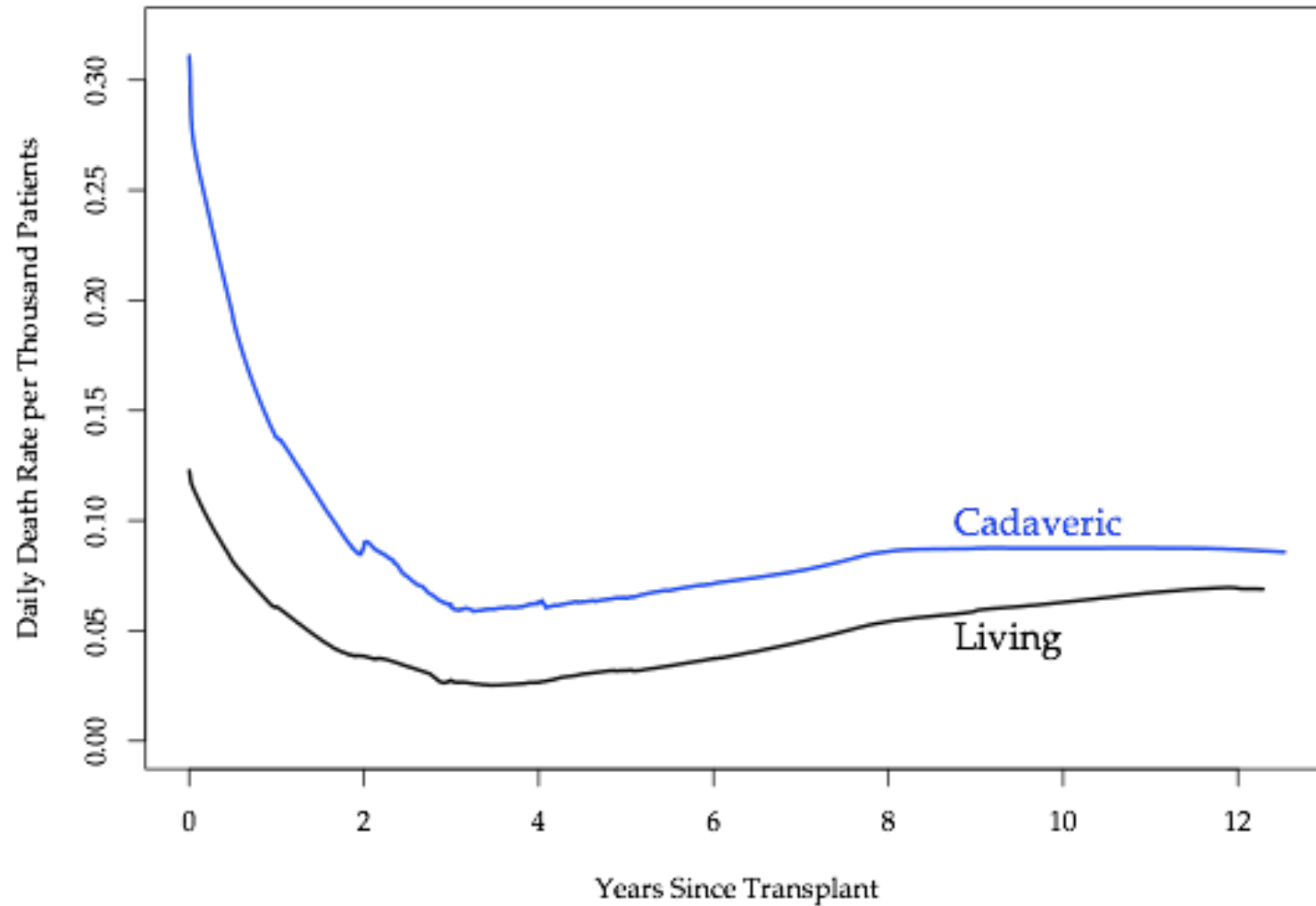
# $h(t)$ for Kidney Tx Data



# Death Rate for UNOS data

- Peaks in weeks after transplant
- Maximum rate = 0.2 deaths/1000 pt days
- Steadily decrease for 3 years
- Risk of death fall with time
- Rate is not a simple function of time

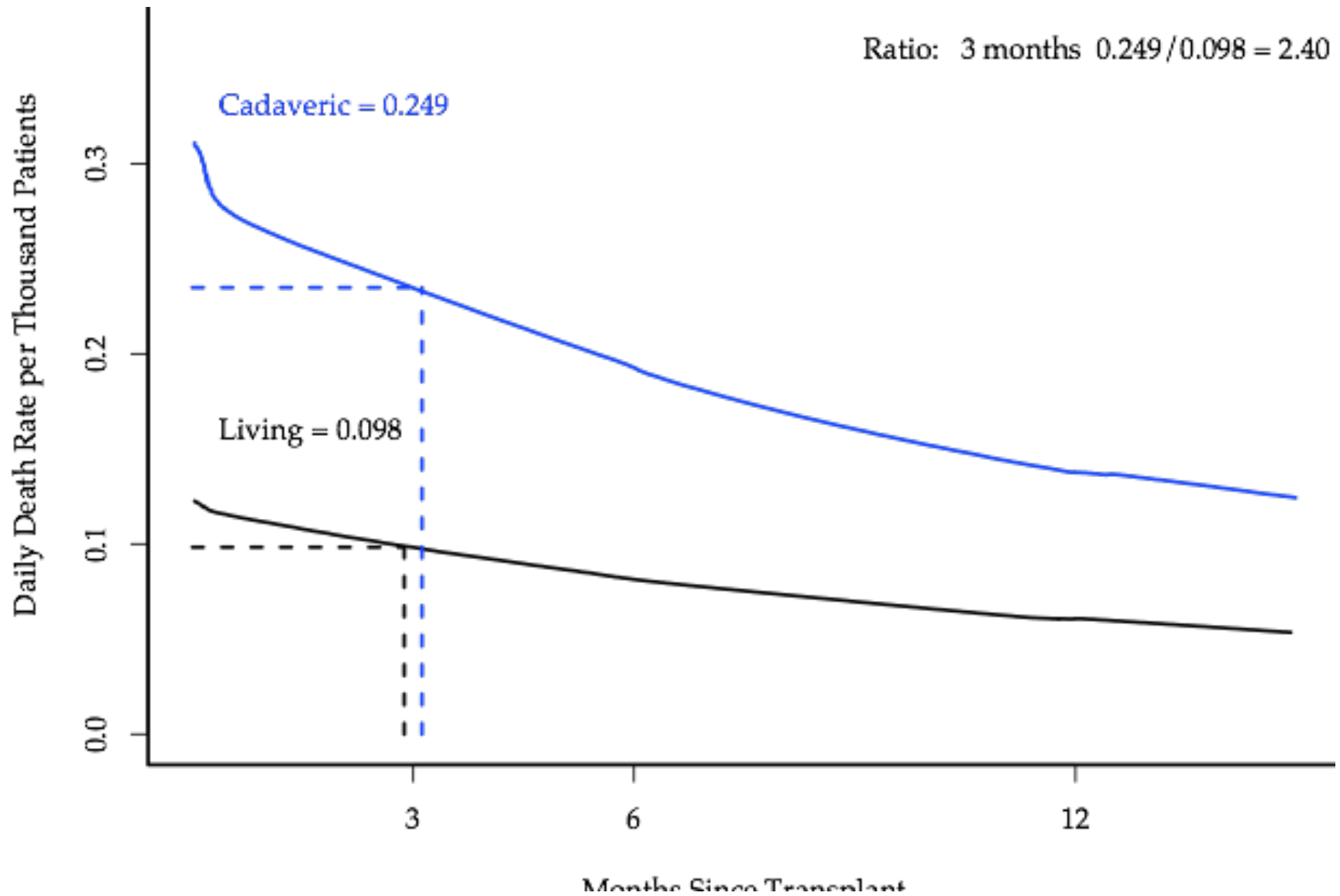
## Hazard by Transplant Type



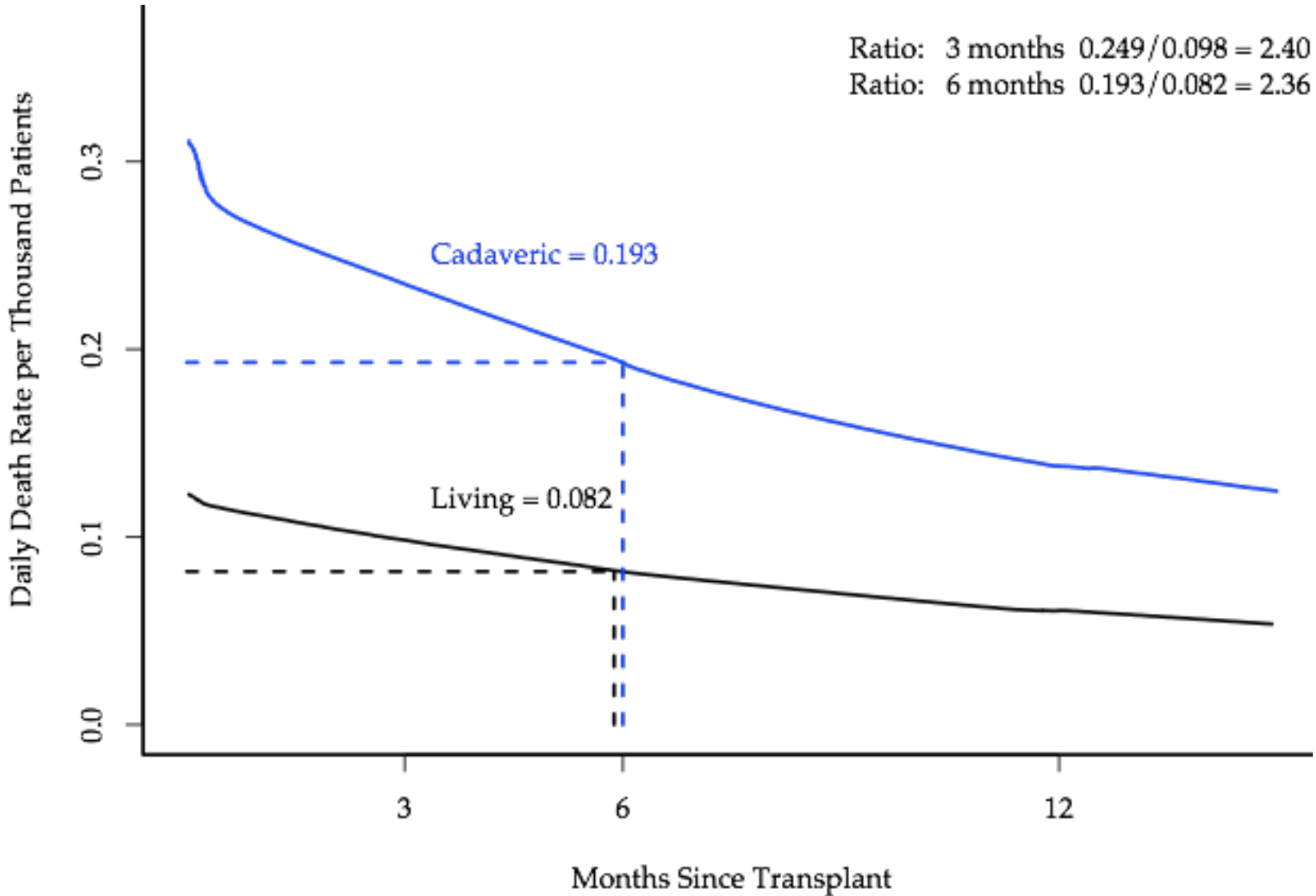
# Comparing Hazards

- Hazards are between 0 and infinity like the odds
- Makes sense to divide when comparing like the odds
- Gives a hazard ratio
- Let's calculate some by time

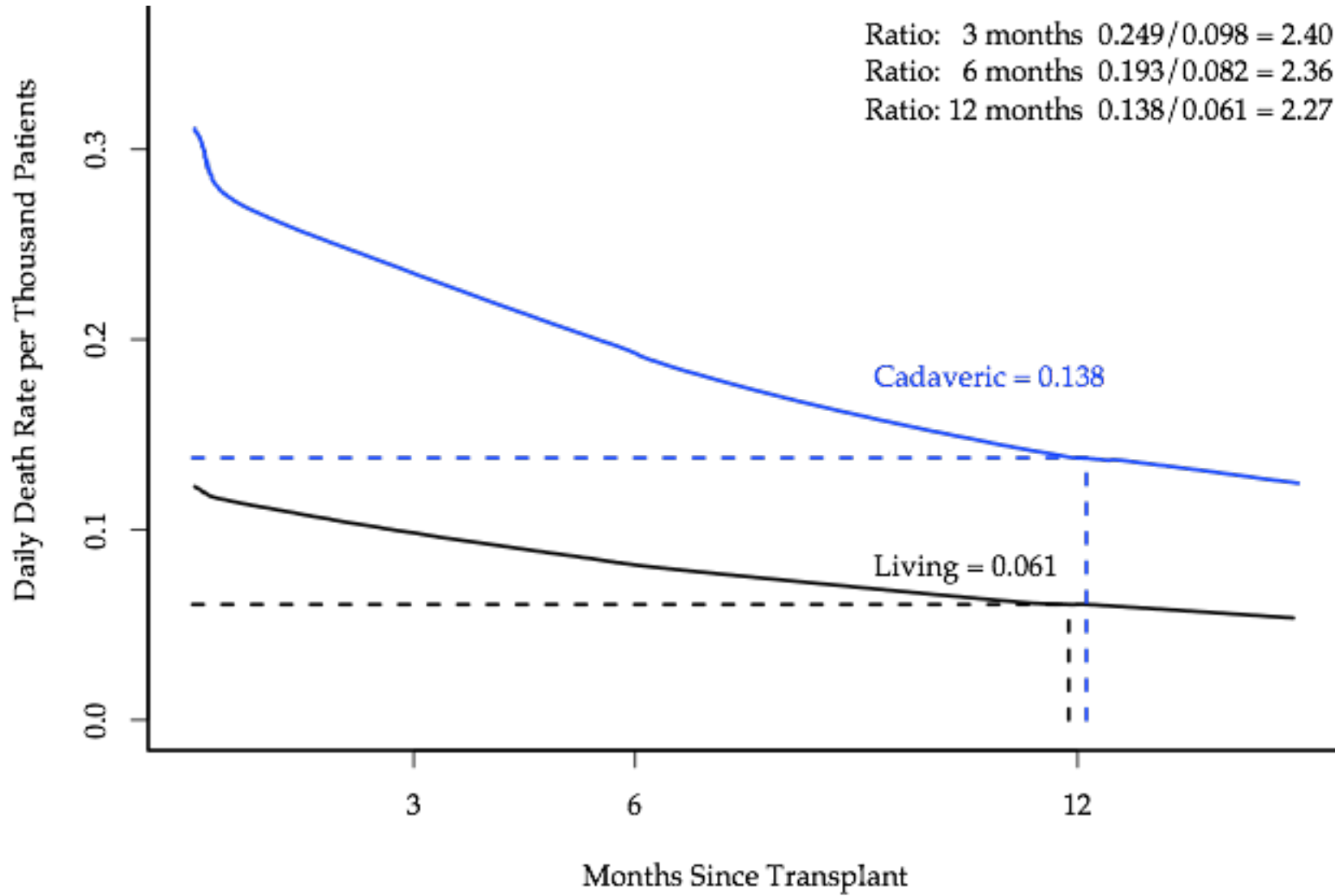
# Hazard at 3 months



# Hazard at 6 months



# Hazard at 12 months



# Smoothed Rates

<i>Time</i>	<i>1000 × Haz Rate</i>		<i>Relative Rate</i>
	<i>Cadaveric</i>	<i>Living</i>	
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

Relative Rates differ greatly over time?

# Hazard Ratio

- $h_0(t)$ : hzd for living recpt at  $t$  (col 3)
- $h_1(t)$ : hzd for cadaveric recpt at  $t$  (col 2)
- $h_1(t)/h_0(t)$ : relative short-term risk (col 4)  
“**hazard ratio**” at time  $t$
- If  $h_1(t) = r h_0(t)$ , hazards proportional  
*hazards have the same shape* (parallel curves)
- $r$  is the relative hazard
- 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
- Their ratio (hazard ratio) didn't vary greatly  
2.03 to 2.40
- “Cadaveric recipients have about twice the death rate as living recipients”
- Cox model: regression based on hazards  
predictor effects in terms of hazard ratios

# The Cox Model for Survival Data

- Let  $\mathbf{x} = (x_1, \dots, x_p)$
- $h(t|\mathbf{x})$ : hazard of someone w/ predictors  $\mathbf{x}$
- $h(t|\mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p)$
- $\log( h(t|\mathbf{x}) ) = \log( h_0(t) ) + \beta_1 x_1 + \dots + \beta_p x_p$
- Much like logistic regression  
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*
- Don't need to know  $h_0(t)$   
*to understand predictor effects*
- 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 $x_p$

$$\begin{aligned} h(t|x) &= h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p V) & x_p = V \\ &= h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p (V+1)) & x_p = V+1 \end{aligned}$$

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

# + 1 unit change in $x_p$

$$\begin{aligned}h(t|x) &= h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p (V+1)) & x_p = V+1 \\ &= h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p V) & x_p = V\end{aligned}$$

$$\begin{aligned}\text{Ratio} &= \frac{\cancel{h_0(t)} \exp(\beta_1 x_1 + \dots + \beta_p (V+1))}{\cancel{h_0(t)} \exp(\beta_1 x_1 + \dots + \beta_p V)} \\ &= \frac{\exp(\beta_1 x_1 + \dots + \beta_p (V+1))}{\exp(\beta_1 x_1 + \dots + \beta_p V)}\end{aligned}$$

# + 1 unit change in $x_p$

$$\begin{aligned} \text{Ratio} &= \frac{\exp(\beta_1 x_1 + \dots + \beta_p (V+1))}{\exp(\beta_1 x_1 + \dots + \beta_p V)} \\ &= \exp(\beta_1 x_1 + \dots + \beta_p (V+1) - (\beta_1 x_1 + \dots + \beta_p V)) \\ &\quad \text{because } \exp(x) / \exp(y) = \exp(x-y) \\ &= \exp(\beta_p (V+1) - \beta_p V) \quad \text{b/c same other predictors} \\ &= \exp(\beta_p) \quad \text{b/c } \beta_p V \text{ term cancel} \end{aligned}$$

# Hazard Ratio

- $\beta$  is the regression coefficient  
If no effect of a predictor variable  $\rightarrow \beta=0$
- HR for a variable is  $\exp(\beta)$   
If no effect of variable  $\rightarrow \exp(\beta)=1$
- HR is the effect of a unit change on hazard
- A useful way for discussing predictor effects

# Stata Command

- First use the command `stset` to declare the data as survival
- `stcox` *predictorlist*

# Stata Output

stcox txtype

```
No. of subjects =          9752          Number of obs =          9752
No. of failures =           383
Time at risk   = 29809.96986
Log likelihood = -3380.3913          LR chi2(1) =          43.12
                                          Prob > chi2 =          0.0000
```

---

<u>_t</u>							
<u>_d</u>		Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
txtype		1.978032	2099168	6.43	0.000	1.606572	2.435378

---

Estimated hazard ratio: *hazard of death about double for cadaveric recipients*  
(living  $x=0$ ; cadaveric  $x=1$ )

# Stata Output

stcox txtype

```
No. of subjects =          9752          Number of obs   =          9752
No. of failures =           383
Time at risk    = 29809.96986
Log likelihood  = -3380.3913          LR chi2(1)        =          43.12
                                          Prob > chi2       =          0.0000
```

---

<u>_t</u>						
<u>_d</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
txtype	1.978032	.2099168	6.43	0.000	1.606572	2.435378

---

SE of hazard ratio

# Stata Output

stcox txttype

```
No. of subjects =          9752          Number of obs   =          9752
No. of failures =           383
Time at risk    = 29809.96986
Log likelihood   = -3380.3913          LR chi2(1)        =          43.12
                                          Prob > chi2       =          0.0000
```

---

<u>_t</u>						
<u>_d</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
txttype	1.978032	.2099168	6.43	0.000	1.606572	2.435378

---

Wald test p-value < 0.05: *hazards are significantly different in the two groups*

# Stata Output

stcox txttype

```
No. of subjects =          9752          Number of obs   =          9752
No. of failures =           383
Time at risk    = 29809.96986
Log likelihood   = -3380.3913          LR chi2(1)        =          43.12
                                          Prob > chi2       =          0.0000
```

---

<u>_t</u>						
<u>_d</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
txttype	1.978032	.2099168	6.43	0.000	1.606572	2.435378

---

**Likelihood ratio test p-value < 0.05: hazards are significantly different in the two groups**

# Stata Output

stcox txtype

```
No. of subjects =          9752          Number of obs   =          9752
No. of failures =           383
Time at risk    = 29809.96986
Log likelihood  = -3380.3913          LR chi2(1)       =          43.12
                                          Prob > chi2     =          0.0000
```

---

<u>_t</u>						
<u>_d</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
txtype	1.978032	.2099168	6.43	0.000	1.606572	2.435378

---

**95% CI for Hazard ratio: *the hazard ratio is likely between 1.61 and 2.44.***

# Interpretation

“The hazard ratio of mortality for the recipient of a cadaveric kidney is 2.0 compared to living organ ( $p < 0.001$ ).

The 95% CI for the hazard ratio is 1.6 to 2.4”

# Why the Cox Model?

- Model for effect of a continuous predictor
- Method for adjusting for confounders
- Framework for interaction, mediation
- Obtain prediction of personal prognosis

# Proportional Hazard Model

- Allows multiple predictors  
*continuous, binary, categorical*
- Confounding  
*adjust by adding confounders to the model*
- Interaction  
*create and add product terms*
- Predictor selection  
*same issues as in linear and logistic regression*
- What different:  
*interpretation, assumptions, model checking*

# Survival by HLA loci

Matching HLA loci range from 0 to 6

xi: 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

<u>t</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Ihlamat_1	.9543573	.1697015	-0.26	0.793	.673523 1.352289
_Ihlamat_2	.697097	.1288643	-1.95	0.051	.485223 1.001486
_Ihlamat_3	.4644883	.0788633	-4.52	0.000	.333007 .6478825
_Ihlamat_4	.4249741	.090394	-4.02	0.000	.2800966 .6447883
_Ihlamat_5	.5870101	.1700124	-1.84	0.066	.3327492 1.035558
_Ihlamat_6	.3834296	.1396405	-2.63	0.008	.1877968 .7828581

# Interpretation

HLA is a significant predictor of mortality  
Chi<sup>2</sup>=47,  $p < 0.001$  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
Log likelihood   =  -3766.4598          LR chi2(1)       =          23.62
                                          Prob > chi2     =          0.0000
```

---

<u>_t</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.9581318	.0083224	-4.92	0.000	.9419582	.9745831

---

# Interpretation

Age is a significant predictor of mortality  
 $Z = -4.92$ ,  $p < 0.001$  by the Wald test.

Each one-year increase in age (at transplant)  
reduces the hazard of mortality by 4%,  
95% CI (2% to 6% reduction)

# Main Points

- Survival Data
  - characterized by censoring
  - requires new methods
- Cox Model
  - based on hazard functions
  - use hazard ratio for predictor effects
  - assumes proportional hazards
  - important similarities with other regressions