



# Models for Survival Analysis with Covariates

Janet Raboud

CHL 5225: Advanced Statistical Methods for Clinical Trials



## Topics

- Survival terminology
- Proportional hazards models
  - Partial likelihood
  - Checking assumptions
  - Residuals
  - Time dependent covariates
  - Multiple failures




# Measuring Survival Time

- Time is measured from
  - Start of the risk period or study period
- Clinical trials
  - Time of randomization
  - Time of intervention
- Cohort Studies
  - Enrollment into cohort?
  - Age?
  - Time the exposure started?
  - Calendar year?



# Censoring

- Censoring is the defining feature of survival analysis, making it distinct from other kinds of analysis.
- Some failures are not observed
- **Right Censoring**
  - Most common kind
  - Individuals are known to not to have experienced the event of interest before a certain time  $t$  but it is not known if they have the event later or at what time the event occurs
- Reasons for censoring
  - Loss to follow-up
  - End of study



## Censoring (continued)

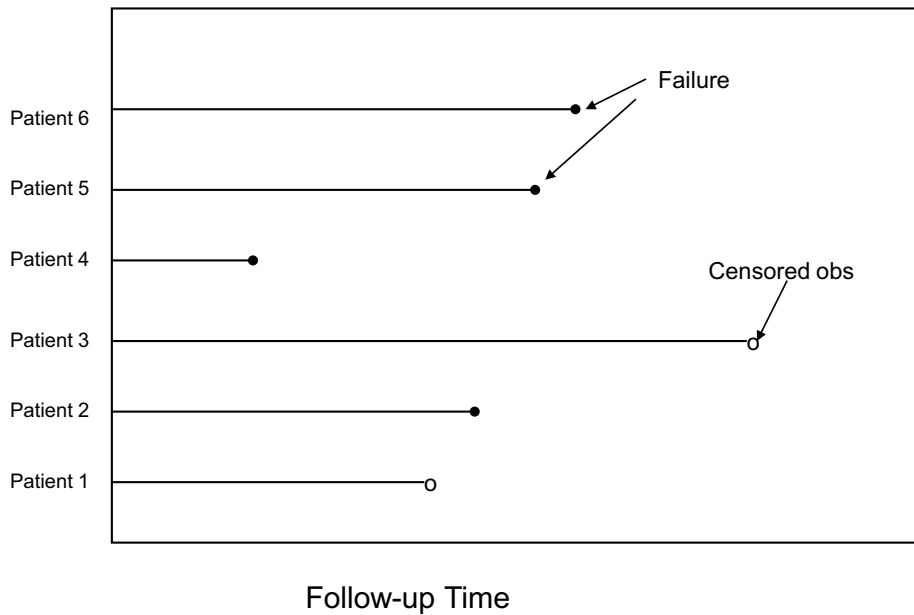
- **Left Censoring** – some failures/events occurred before observation started
- **Interval censoring** – time is only known to fail between two dates
  - Diagnosis of diabetes
  - Infection with HIV (between last negative and first positive test)
  - Suppression of HIV virus
- Assumption: Censoring occurs at random and is unrelated to failure process



## Notation

- Outcome of interest is failure time  $T$ 
  - If all failures were observed, we could model  $f(T)$  directly
- $C$  = censoring time
- $X = \min(T, C)$  = observed “end” time
- $\delta = I\{T < C\}$  indicates that  $X$  is a failure rather than a censored observation.
- The  $k$  distinct failure times from  $N$  individuals can be labeled as

$$t_{(1)}, t_{(2)}, \dots, t_{(k)}$$



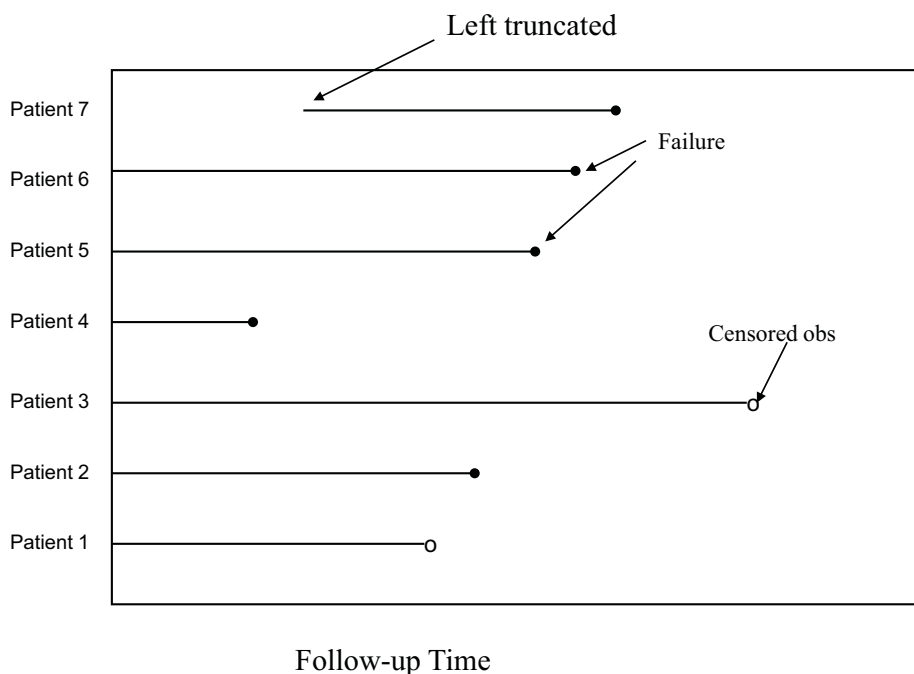
## Data for survival analysis

- Time
- Censoring indicator
- Covariate(s)

ID	Time	Failure	x
1	12	1	25
2	7	0	30
3	21	1	31
4	15	0	27
5	12	1	28
6	18	0	22
7	28	1	32

# Left Truncation

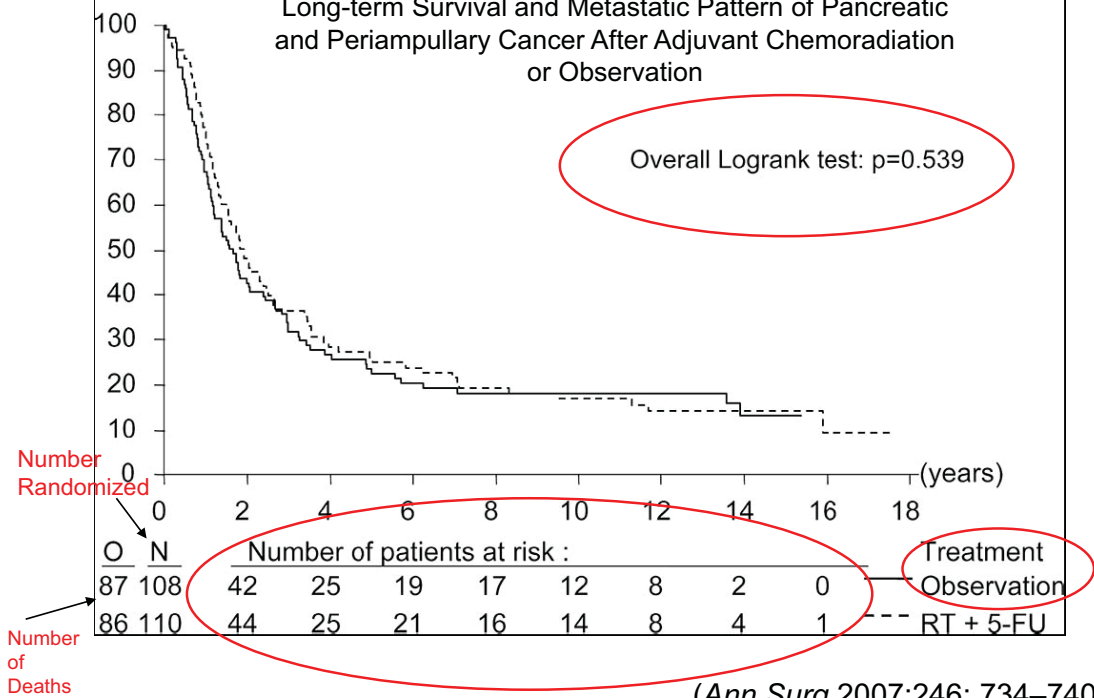
- Left truncation occurs when an individual comes into observation some time after the natural starting point of the phenomenon.
- Eg. Measuring time from HIV infection to AIDS – some individuals are not followed from the time of infection but come into observation some time later.
- Want to make sure that these observations are excluded from the risk sets of failures which occur before they come under observation.





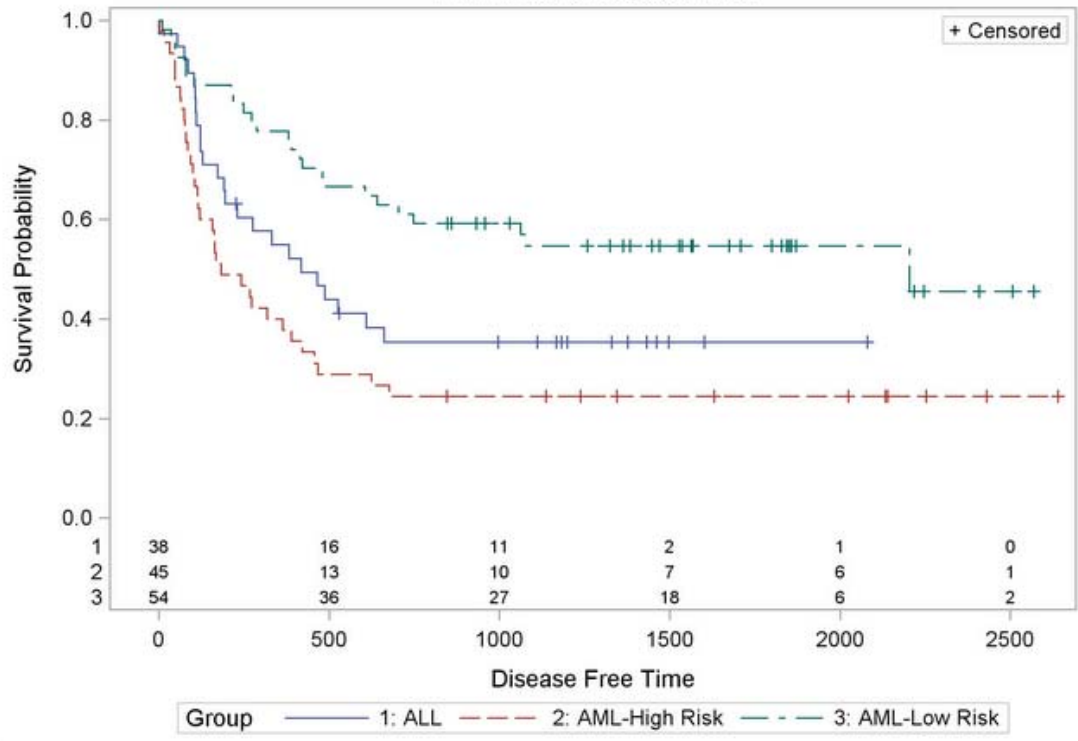
# Survival


Long-term Survival and Metastatic Pattern of Pancreatic and Periampullary Cancer After Adjuvant Chemoradiation or Observation



## Product-Limit Survival Estimates

With Number of Subjects at Risk





```
ODS html;  
ODS graphics on;  
PROC LIFETEST data=BMT plots=survival  
  (atrisk=0 to 2500 by 500) maxtime=2500;  
  TIME T * Status(0);  
  STRATA Group / test=logrank;  
  RUN;  
ODS graphics off;  
ODS html close;
```



## Median Survival Time

- The median survival time can be estimated as the time at which the survival curve reaches 50%, ie. where  $F(t) = .50$
- Can't estimate median survival time if  $F(t)$  never reaches .50.
- The median survival time is *\*not\** the median of the survival times of individuals who failed.

# Survival Functions

## ■ Survival Function

$$F(t) = \Pr(T \geq t) = \exp\left[-\int_0^t \lambda(s) ds\right]$$

## ■ Probability density function

$$f(t) = \lim_{\Delta \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta)}{\Delta} = \lambda(t) \exp\left[-\int_0^t \lambda(s) ds\right]$$

## ■ Hazard function

$$\lambda(t) = \frac{f(t)}{F(t)}$$

# Full Likelihood

## ■ Full likelihood for survival analysis

- $T_i \sim f(t; \theta)$  survivor function  $F(t; \theta)$
- $C_i \sim g(t; \eta)$  censoring function  $G(t; \eta)$

## ■ Assuming censoring is independent of failure

$$L(\theta, \eta) = \prod_{i=1}^n \{f(t_i; \theta)G(t_i; \eta)\}^{\delta_i} \{F(t_i; \theta)g(t_i; \eta)\}^{1-\delta_i}$$

## ■ If G contains no information about parameters

$$\begin{aligned} L(\theta, \eta) &\propto \prod_{i=1}^n f(t_i; \theta)^{\delta_i} F(t_i; \theta)^{1-\delta_i} \\ &= \prod_{i=1}^n [\lambda(t_i) \exp[-\int_0^{t_i} \lambda(s) ds]]^{\delta_i} [\exp[-\int_0^{t_i} \lambda(s) ds]]^{1-\delta_i} \\ &= \prod_{i=1}^n [\lambda(t_i)^{\delta_i} \exp[-\int_0^{t_i} \lambda(s) ds]] \end{aligned}$$



# Proportional Hazards Models

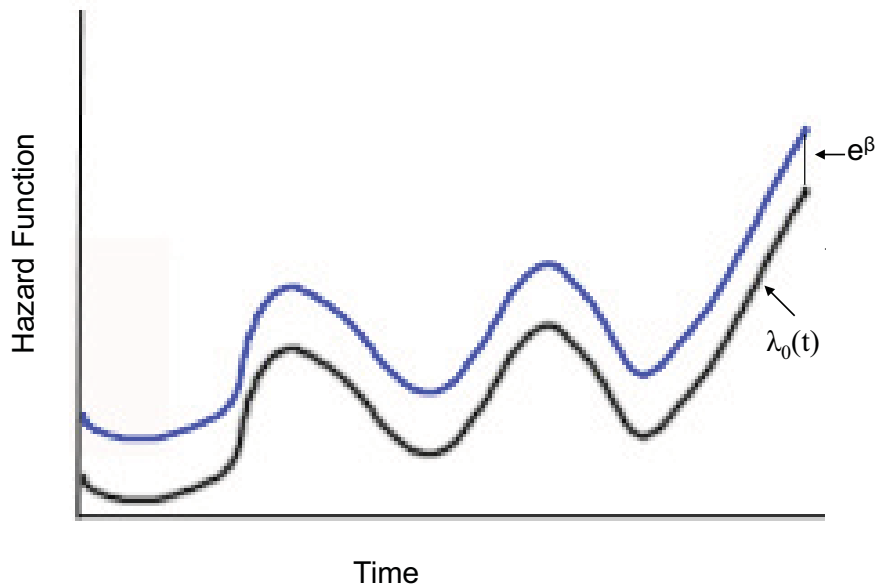
Cox 1972, JRSS(B)

- Hazard function

$$\lambda(t) = \lambda_0(t)e^{\beta X}$$

- where  $\lambda_0(t)$  is an arbitrary and unspecified baseline hazard function that does not depend on  $\beta$
- $X$  is a vector of explanatory variables
- $\beta$  is a vector of regression coefficients associated with  $X$

Example of Baseline Hazard Function






## Ratio of Hazard Functions

$$\frac{\lambda(t | z_1)}{\lambda(t | z_0)} = e^{(z_1 - z_0)\beta}$$

is constant over time.

$$\frac{\lambda(t | z_k + 1)}{\lambda(t | z_k)} = e^{\beta_k} \quad \text{For all } t > 0.$$

$e^{\beta_k}$  is the HR associated with a 1 unit increase in  $z_k$ .

- 
- Since the baseline hazard is completely unspecified, we can't use the ordinary likelihood to estimate  $\beta$
  - Cox proposed the idea of a partial likelihood to remove the nuisance parameter  $\lambda_0(t)$  from the estimating equation.

## Development of the Partial Likelihood

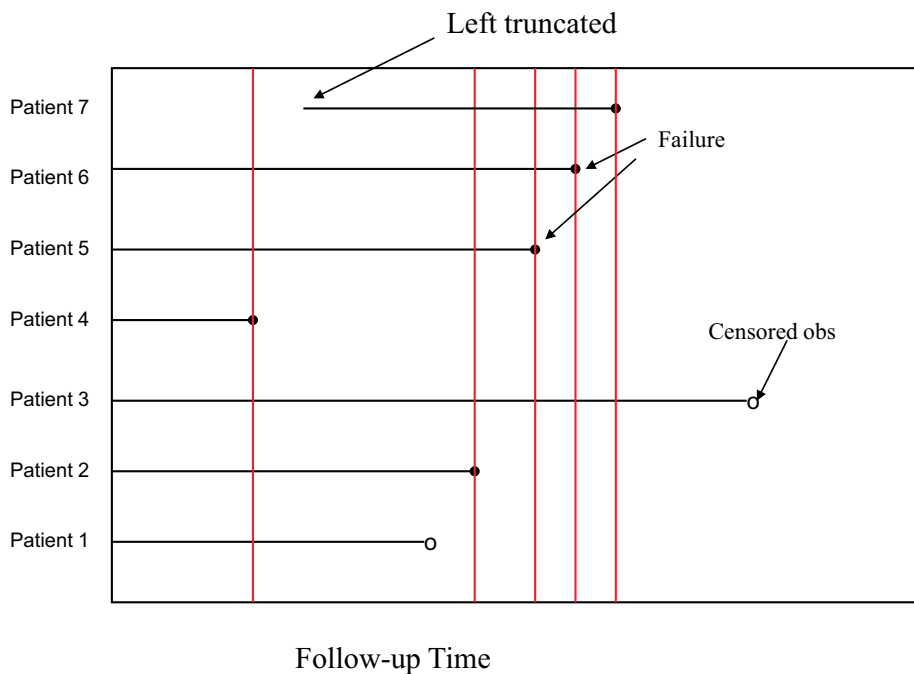
We consider the probability that individual  $i$ , with covariate vector  $x_i$ , is the one to experience the event at time  $t_i$ , given that there is an event at time  $t_i$ .

$$\begin{aligned} & \Pr(\text{Individual } i \text{ has event at time } t_i \mid \text{one event at } t_i) \\ &= \frac{\Pr(\text{Individual } i \text{ has event at time } t_i \mid \text{survival to } t_i)}{\Pr(\text{one event at } t_i \mid \text{survival to } t_i)} \\ &= \frac{h(t_i \mid x_i)}{\sum_{j \in R_i} h(t_i \mid x_j)} \\ &= \frac{h_0(t_i) \exp \beta^T x_i}{\sum_{j \in R_i} h_0(t_i) \exp \beta^T x_j} \\ &= \frac{\exp \beta^T x_i}{\sum_{j \in R_i} \exp \beta^T x_j} \text{ where } R_i \text{ is the set of individuals at risk at } t_i \end{aligned}$$

The product of all of these conditional probabilities is called the **partial likelihood**.

$$PL = \prod_i \frac{\exp \beta^T x_i}{\sum_{j \in R_i} \exp \beta^T x_j}$$

We discard information about the actual times at which events happen.





# Ties in Event Times

- Death times are assumed to be unique when constructing the partial likelihood
- Ties can occur when a “truly continuous” time variable is not measured accurately enough
- When deaths and censoring times coincide, the censoring is assumed to occur immediately after all the deaths
- When death times coincide, the true arrangement of the ranks is unknown – many possible permutations
- Fortunately, permutations within different death times can be treated independently



## Partial Likelihood when there are Ties in Event Times

(Kalbfleisch and Prentice)

- Assume there are  $d_i$  events at time  $t_i$
- Let  $Q_i$  be the set of all subsets of size  $d_i$ , which can be selected from  $R_i$  individuals.

Pr(subset  $q_i$  have event at time  $t_i$  |  $d_i$  events at time  $t_i$ )

$$\begin{aligned}
 &= \frac{\prod_{k \in q_i} h_0(t_i) \exp \beta^T x_k}{\sum_{q_j \in Q_i} \prod_{j \in q_j} h_0(t_i) \exp \beta^T x_j} \\
 &= \frac{\exp \beta^T \sum_{k \in q_i} x_k}{\sum_{q_j \in Q_i} \exp \beta^T \sum_{j \in q_j} x_j}
 \end{aligned}$$

$$\text{Then, exact PL} = \prod_i \frac{\exp \beta^T \sum_{k \in q_i} x_k}{\sum_{q_j \in Q_i} \exp \beta^T \sum_{j \in q_j} x_j}$$

Can be very computationally intensive.

## Approximations for Ties

■ Breslow: 
$$PL = \prod_i \frac{\exp \beta^T \sum_{k \in q_i} x_k}{[\sum_{j \in R_i} \exp \beta^T x_j]^{d_i}}$$

Breslow's method is the default in SAS and is a good approximation when ties are not extensive.

■ Efron: 
$$PL = \prod_i \frac{\exp \beta^T \sum_{k \in q_i} x_k}{\prod_{l=1}^{d_i} [\sum_{j \in R_i} \exp \beta^T x_j - \frac{l-1}{d} \sum_{m \in q_i} \exp \beta^T x_m]}$$

Efron's method is a closer approximation of the exact PL.

## Estimation of Coefficients

- The Newton-Raphson method is used to find an estimate of  $\hat{\beta}$
- $\hat{\beta}$  is maximized to find a solution to the likelihood equations

$$\frac{\delta \log L(\beta)}{\delta \beta} = 0$$

With  $\hat{\beta}^0 = 0$  as an initial solution, estimates of the coefficients are found through an iterative process.

$$\hat{\beta}^{j+1} = \hat{\beta}^j - \left[ \frac{\delta^2 l(\hat{\beta}^j)}{\delta \beta^2} \right]^{-1} \frac{\delta l(\hat{\beta}^j)}{\delta \beta}$$



## Testing the Global Null Hypothesis

- Likelihood Ratio Test

$$\chi_{LR}^2 = 2[l(\hat{\beta}) - l(0)]$$

- Wald's Test

$$\chi_w^2 = \hat{\beta}' [\widehat{V}(\hat{\beta})]^{-1} \hat{\beta} = \hat{\beta} / se(\hat{\beta}) \text{ for one covariate}$$

- Score Test

$$\chi_s^2 = \left[ \frac{\delta l(0)}{\delta \beta} \right]' I^{-1}(0) \left[ \frac{\delta l(0)}{\delta \beta} \right] \text{ where } I(0) = \frac{\delta^2 l(0)}{\delta \beta^2}$$



## Comparison of Test Statistics

- Each statistic has a chi square distribution with  $p$  degrees of freedom, where  $p$  is the number of covariates.
- All three tests are equivalent asymptotically.
- The likelihood ratio test is considered the most reliable, the Wald test the least reliable.
- The efficient score statistic is based on one iteration of the Newton-Raphson algorithm.



## Example of Comparing Survival with PH Models in a Clinical Trial

- Randomized clinical trial (ACTG 320) of a new treatment for HIV disease compared to placebo
  - 577 patients on placebo
  - 574 patients on indinavir
- Outcome = new AIDS defining illness or death
- Randomization was stratified by CD4 count




## SAS code for PH models

```
PROC PHREG data = data;  
  MODEL time * censor (0) = covariate;  
  BY covariate;  
  STRATA subgroup;  
  RUN;
```



# SAS code for ACTG 320

```
PROC PHREG data = actg320;  
  MODEL time * censor(0) = tx /risklimits;  
RUN;
```



Model Information		
Data Set	WORK.ACTG320	
Dependent Variable	time	time
Censoring Variable	sensor	sensor
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	1151
Number of Observations Used	1151

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
1151	96	1055	91.66





Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1316.931	1306.236
AIC	1316.931	1308.236
SBC	1316.931	1310.800

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.6952	1	0.0011
Score	10.5399	1	0.0012
Wald	10.1365	1	0.0015



Analysis of Maximum Likelihood Estimates					
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
tx	1	-0.68425	0.21492	10.1365	0.0015

Hazard Ratio	95% Hazard Ratio Confidence Limits		Variable Label
0.504	0.331	0.769	tx



## Proportional Hazard Models

Covariates	Unadjusted		Adjusted	
	Hazards Ratio (95% CI)	p value	Hazards Ratio (95% CI)	p value
Treatment	0.50 (.33,.769)	0.0015	0.49 (.32,.75)	0.001
CD4 > 50 cells/mm <sup>3</sup>	0.26 (.16, .40)	<0.0001	0.25 (.16, .39)	<0.0001
Age (per year)	1.02 (.999, 1.04)	0.06		
Age (per 10 years)	1.23 (0.99, 1.52)	0.06		
Age > 50 years	1.86 (1.13, 3.07)	0.015	2.11 (1.27, 3.48)	0.004
Hemophiliac	1.02 (0.32, 3.22)	0.97		
Years of prior ZDV	0.97 (0.89, 1.06)	0.51		



## Interpretation of Model Output

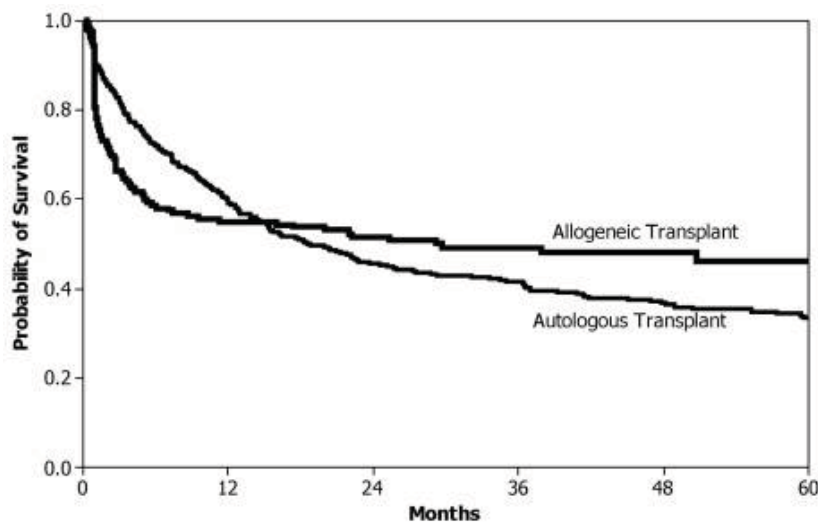
- The hazard ratio associated with treatment is 0.49. Patients receiving indinavir are 0.49 times as likely to progress to an AIDS defining event or death as patients receiving placebo.
- Patients with CD4 counts > 50 are 0.25 times as likely to progress to an AIDS defining event or death as patients with CD4 counts < 50.
- Patients more than 50 years old were 2.11 times as likely to progress to an AIDS defining event or death.

# Checking Assumptions of the Proportional Hazards Model

- The key assumption of this model is proportionality
- I.e. the ratio of hazards for any two subjects  $i$  and  $j$  is independent of time

$$\frac{\lambda_0(t)e^{X_i\beta}}{\lambda_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}}$$

## KM curves should not cross



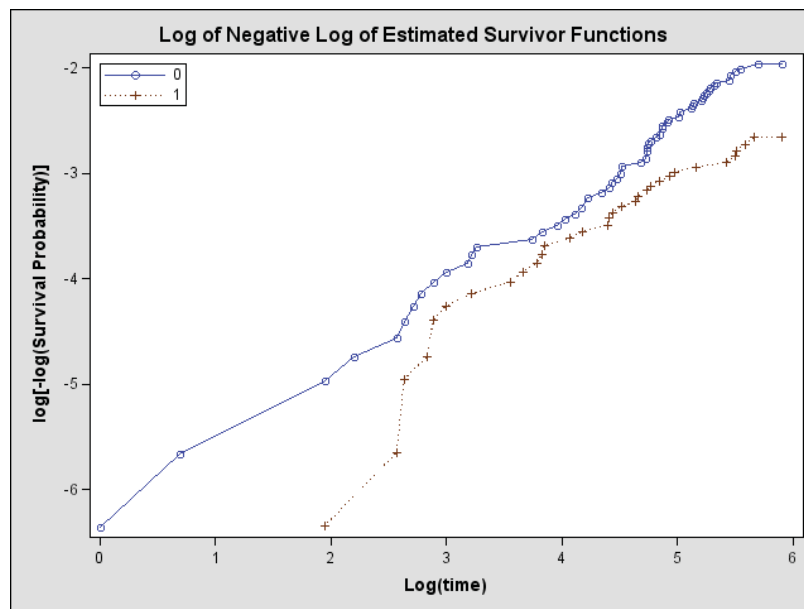
# Checking PH Assumptions

- For covariates with a small number of levels, plots of  $\log[-\log(F_i(t))]$  vs  $\log t$  are useful.
- The plots should be roughly parallel if the proportional hazards model is appropriate

Recall  $F(t) = \Pr(T \geq t) = \exp\left[-\int_0^t \lambda(s) ds\right] = \exp\left[-\int_0^t \lambda_0(s) e^{\beta x} ds\right]$

$$-\log F(t) = \int_0^t \lambda_0(s) e^{\beta x} ds$$

$$\log[-\log F(t)] = \log \int_0^t \lambda_0(s) ds + \beta x$$





## Checking PH assumptions with residuals

- If the covariate has many levels or is continuous, KM plots and  $\log(-\log(F(t)))$  plots are not as useful.
- Plots of the Schoenfeld residuals can be used to assess departure from the PH assumption.
  - cumulative sum of Schoenfeld residuals
  - weighted Schoenfeld residuals
- **References**
  - Schoenfeld D. Partial residuals for the proportional hazards model. *Biometrika* 1982;69:239-41
  - Therneau & Grambsch. *Modeling Survival Data*. 2000. Ch 6.



## Cumulative Sum of Schoenfeld Residuals

- One residual for each covariate for each subject
  - The residual is the difference between the observed value of  $x$  and its conditional expectation
- Only defined at observed event times.
- Plot of cumulative sum of residuals against time or log time should be a random walk starting and ending at zero and centered around zero.
- Plots can be hard to interpret.



# Weighted Schoenfeld Residuals

- If PH doesn't hold, an alternative is

$$\lambda(t) = \lambda_0(t) \exp[x\beta(t)]$$

- If PH holds, then  $\beta_j(t)$  vs  $t$  will be a horizontal line
- It has been shown that

$$E(s_{kj}^*) + \beta_j = \beta_j(t_k)$$

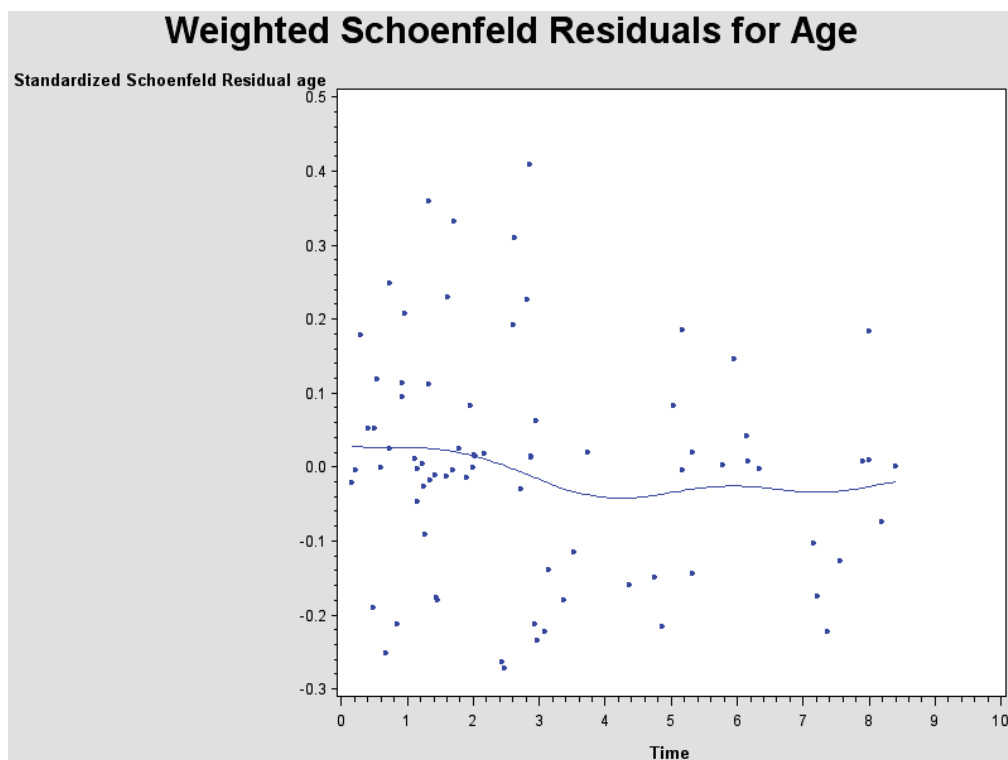
- where  $s_{kj}^*$  = scaled Schoenfeld residual
- So, a plot of the scaled Schoenfeld residuals vs time should have a zero slope if the PH assumption holds.



# Outputting Residuals in SAS

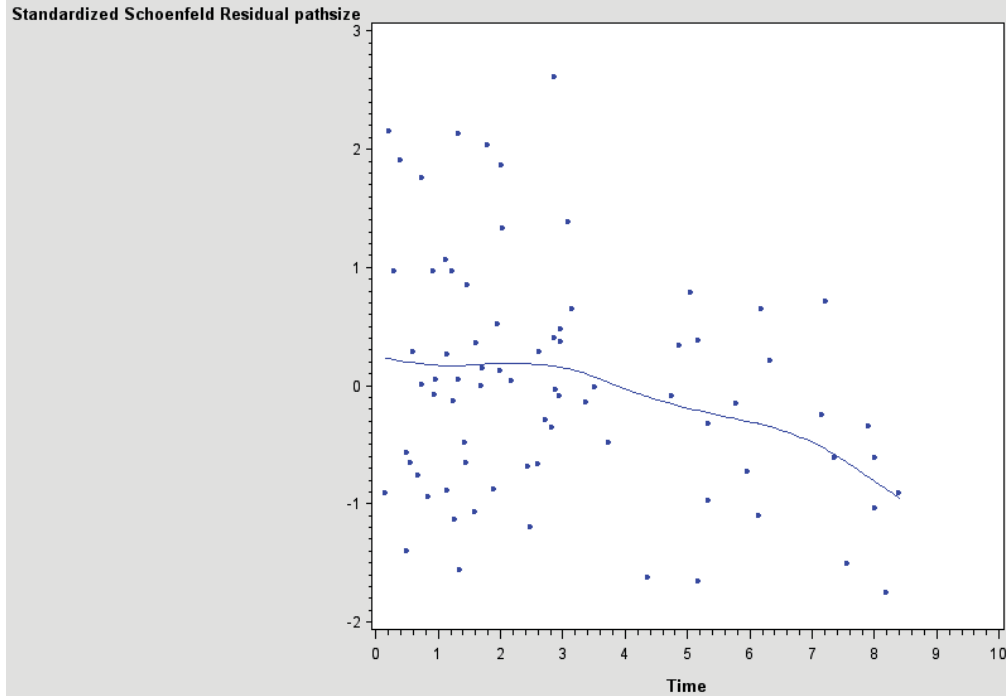
```
ODS html;  
ODS graphics on;  
PROC PHREG;  
MODEL TIME * CENSOR = age tumorsize;  
OUTPUT OUT = residuals  
  RESDEV = deviance  
  RESMART = martingale  
  RESSCH = schoenfeld  
  RESSCO = score  
  WTRESSCH = wtschoenfeld_age wtschoenfeld_tumor;  
RUN;  
ODS graphics off;  
ODS html close;
```

```
ods html;
ods graphics on;
proc gplot data=residuals;
  title 'Weighted Schoenfeld Residuals for Age';
  plot wtschoenfeld_age*timemin;
  label timemin = 'Time';
run;
ods graphics off;
ods html close;
```





## Weighted Schoenfeld Residuals for Tumor Size



- Plot is indicative of non proportional hazards
- Next step
  - Fit an interaction of time and tumor size
  - In this case, the interaction is significant  $\beta$
- How to proceed?
  - It may not matter
  - Should consider the size of the variation in residuals relative to the estimated coefficient





## **When PH assumptions don't hold:**

- Can include an interaction term with time and the covariate
- Can use a parametric model
- Use a time-dependent coefficient



## **Check functional form of covariates**

- For continuous variables, a linear relationship may or may not be the best fit
- It may be better to:
  - create categories for the variable
  - transform with logarithms, square root, etc
  - model as a quadratic
- Plot martingale residuals vs continuous covariates to check functional form of covariates



## Residual Plots

- Plot martingale residuals vs continuous covariates
  - To check functional form of covariates
- Plot deviance residuals vs observation number, to check for outliers.



## Martingale Residuals

- $M_i(t)$  = Martingale residual at time  $\tau_i$  is the difference over  $[0, \tau_i]$  between the observed and expected number of events.
- The Martingale residual for the  $i^{th}$  subject is the sum of the residuals over all time periods;  $M_i = M_i(\infty)$

$$M_i = \delta_i - \Delta_0(t_i) \exp(\hat{\beta}'z_i)$$

- Range  $-\infty$  to 1



## Deviance Residuals

- A transformation of the martingale residuals to achieve a more symmetric distribution.
- Symmetrically distributed about zero.
- Negative for observations with longer survival times than expected and positive for observations with shorter survival than expected
- Extreme values may indicate an outlier
- Unusual patterns may suggest that the model is not a good fit to the data.



## Model Assessment

- “ASSESS” statement in PROC PHREG
- Uses cumulative sums of martingale residuals over follow-up times or covariate values
- Can check
  - Functional form of a covariate
  - PH assumption for each covariate

# Model Assessment in SAS

ODS html;

ODS graphics on;

PROC PHREG;

MODEL TIME \* CENSOR = AGE;

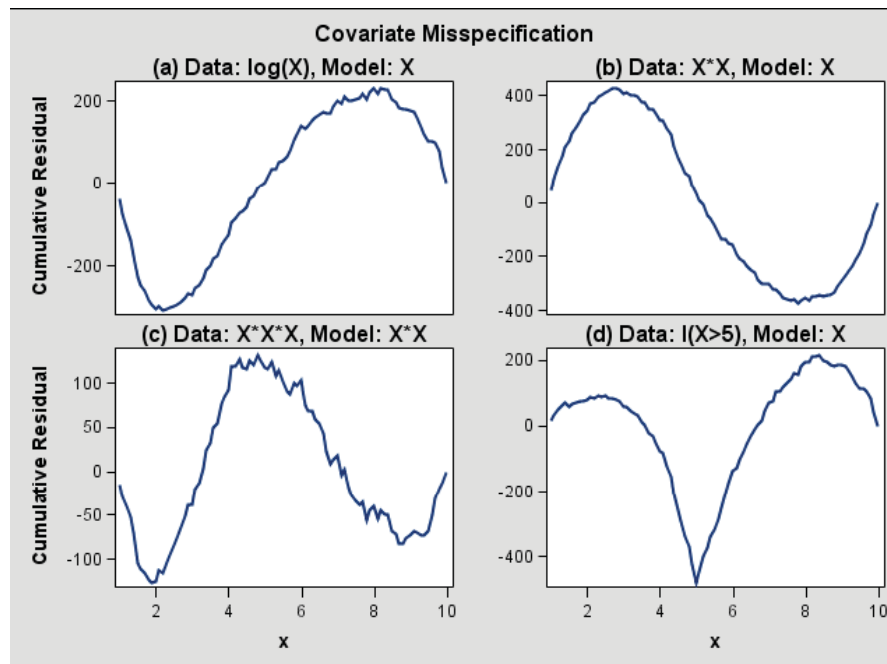
ASSESS VAR = (age) PH / crpanel resample seed=2011;

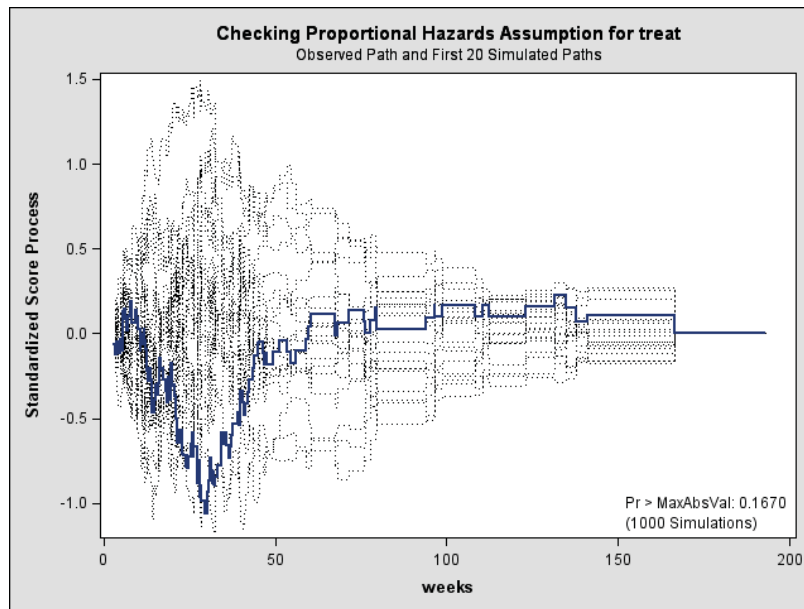
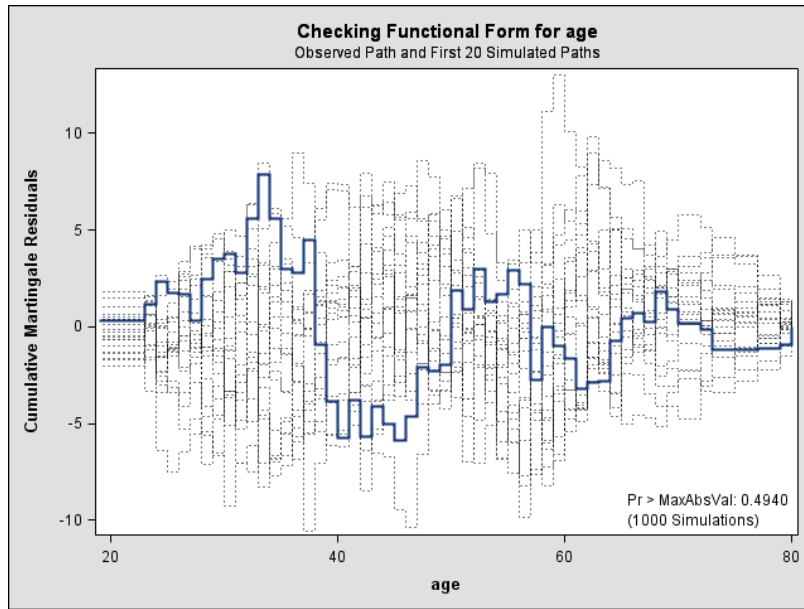
RUN;

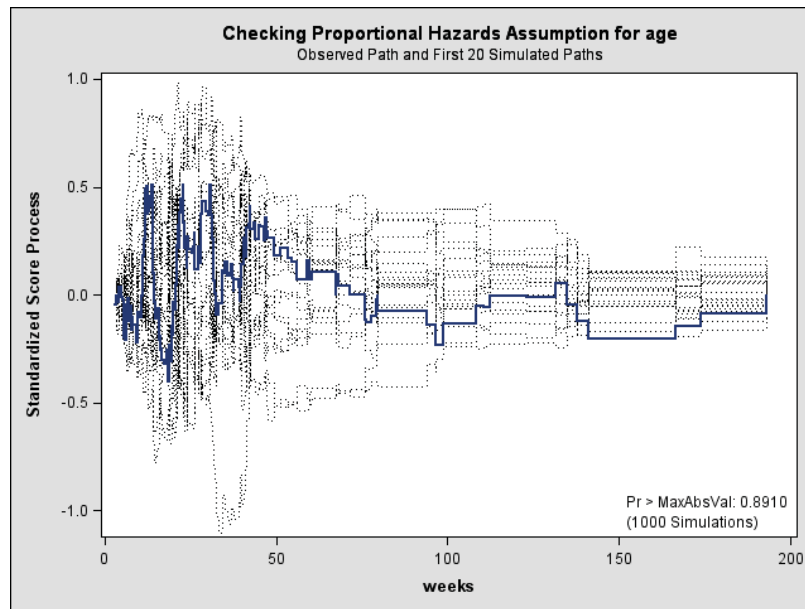
ODS graphics off;

ODS html close;

“Resample” gives the p value. Setting the seed makes sure you get the same p value each time.

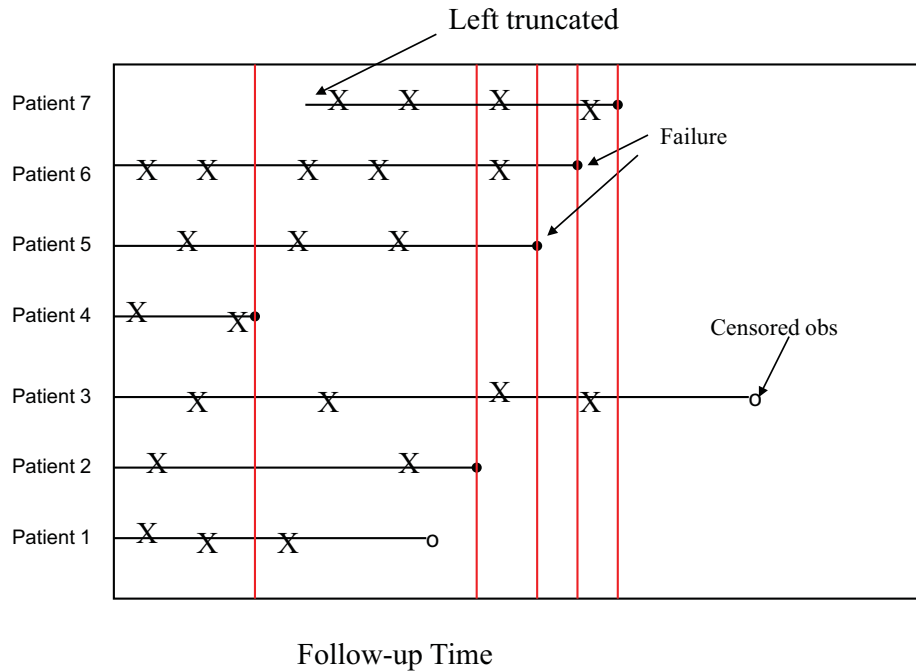






## Time dependent Covariates

- External covariates: not directly related to failure mechanism
  - Time of year
  - Air pollutants
  - Age
- Internal covariates: generated by the individual
  - Blood pressure
  - CD4 cell count
  - Blood pressure
- Fundamental assumption for time-dependent covariates is that the change in exposure occurs randomly.



## Modeling External time dependent Covariates in SAS

Can create time dependent covariates directly in the PROC statement for simple situations:

```
PROC PHREG data= Heart;  
  MODEL Time*Status(0)= XStatus Acc_Age;  
  IF (WaitTime = . or Time < WaitTime) then  
    XStatus=0.;  
  ELSE XStatus= 1.0;  
RUN;
```



## Data set up for time dependent covariates (counting process style)

- Each individual has one row of data for each different value of their time-dependent covariate
- For each period, enter the start and stop date of that period, the value of the covariate during that period and an indicator of whether or not the person experienced the event at the end of the period.



## Data Setup for Time dependent Covariates

ID	Start	Stop	Event	CD4	gender
1	0	10	0	250	M
1	10	15	0	187	M
1	15	20	1	123	M
2	0	6	0	300	F
2	6	9	0	281	F
2	9	13	0	260	F
2	13	14	1	200	F
4	0	8	1	105	M
3	0	12	0	86	F





# Smoking and Survival with Heart Disease

- Problem:
  - Individuals stop smoking just before death
  - Smoking appears to improve survival!
  - In fact – the converse is likely true, but measures of severe illness (hospitalization or heart failure, for example), lead to smoking cessation
  - Smoking status is correlated with severity of illness
- Covariate should not be on causal pathway
- Solutions
  - Time lagged covariates
  - % of follow-up where individual smoked

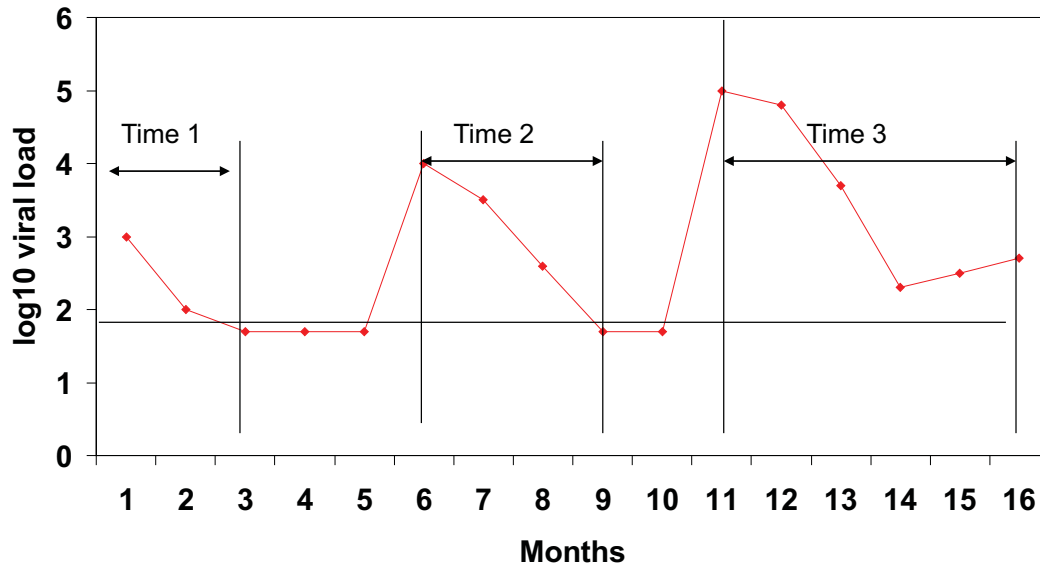


# Multivariate Failure Time

- Multivariate failure time data arise when
  - Individuals under study can experience multiple events of the same type such as heart attacks or recurrence of disease during the study period (recurrent events) or
  - Individuals under study can experience multiple events of different types (competing risks)
  - Individuals under study are in natural clusters, such as members of the same family, etc. (clustered failure times).



## Determining Times to Virologic Suppression



## Strategies to Handle Multiple failure times per person

- Analyze only the time to the first failure
  - Loss of information
- Model time between different events separately
- Model correlation between events within subjects by introducing a random effect or frailty
- Marginal model – no assumptions about the nature of within subject dependence



## Kinds of models

- Marginal model (Wei, Lin, Weissfeld (1989))
- Intensity model (Andersen and Gill, 1982)
- Gap time model (Prentice, Williams and Peterson, 1981)
- Proportional rates/means model (Lin, Wei, Yang and Ying (2000), Lawless and Nadeau (1995), Pepe and Cai (1993))

Kelly & Lim Statistics in Medicine, 2000; 19:13-33.



## Risk Intervals

- Gap time
  - resets the clock to zero after each event
- Total time
  - counts time from baseline for each event
- Counting process
  - Measures both start and stop times from baseline

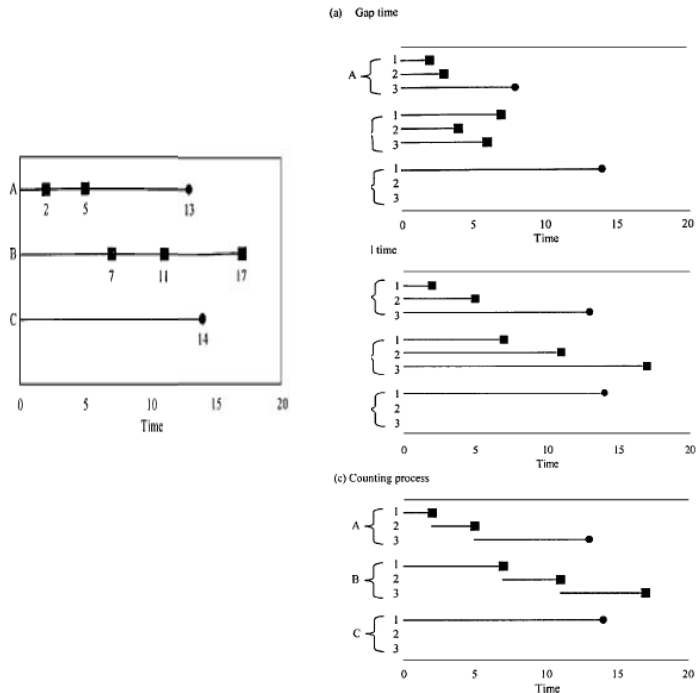


Figure 2. Illustrations of the risk interval formulations: (a) gap time; (b) total time; (c) counting process, using the hypothetical data from Figure 1, where ■ is an event and ● is censoring. Each time to an event or censoring is a separate risk interval, hence subjects A and B have three separate intervals

## Baseline Hazard

- Common baseline hazard
- Event-specific baseline hazard



## Risk Set

- Unrestricted
  - All subjects' risk intervals contribute to the risk set for each event, regardless of how many events they have had.
- Restricted
  - Only the  $k^{th}$  events are included in the  $k^{th}$  risk set
- Semi-restricted



## Within-Subject correlation

- Conditional
  - Assumes the current event is unaffected by earlier events in that subject
- Marginal
  - Assumes that events within a subject are independent
- Random effects (aka frailty model)



# Marginal Cox Models for Multiple Events Data

- Each event is considered as a separate process.
- $N$  subjects, each subject can experience up to  $K$  events.
- $Z_{ki}(\cdot)$  is the covariate associated with the  $k^{\text{th}}$  event for the  $i^{\text{th}}$  subject.

$$\lambda_k(t; Z_{ki}) = \lambda_{k0}(t) \exp[\beta'_k Z_{ki}(t)], \quad k = 1, \dots, K; \quad i = 1, \dots, N$$

- Where  $\lambda_{k0}(t)$  is the event specific baseline hazard function for the  $k^{\text{th}}$  event and  $\beta_k$  is the vector of regression coefficients for the  $k^{\text{th}}$  event.
- Marginal models do not condition on the time since the study start or the previous inter-event times.



The input data set should contain

- an ID variable for identifying the subject so that all observations of the same subject have the same ID value
- an event number variable to index the multiple events. For example, Event = 1 for the first event, Event = 2 for the second event, and so on.
- a Time variable to represent the observed time from some time origin for the event. For recurrence events data, it is the time from the study entry to each recurrence.
- a Status variable to indicate whether the Time value is a censored or uncensored time. For example, Status=1 indicates an uncensored time and Status=0 indicates a censored time.
- covariates:  $X_1, X_2, \dots$



# SAS Code to fit marginal model

```
proc phreg covs(aggregate);
  model Time*Status(0)=Z11 Z12 Z13 Z21 Z22 Z23;
  strata Enum;
  id ID;
  Z11= Z1 * (Enum=1);
  Z12= Z1 * (Enum=2);
  Z13= Z1 * (Enum=3);
  Z21= Z2 * (Enum=1);
  Z22= Z2 * (Enum=2);
  Z23= Z2 * (Enum=3);
run;
```



## Intensity Model (Andersen and Gill, 1982)

- Each subject can experience multiple events of the same type.
- $N(t)$  is the number of events a subject experiences over the interval  $[0,t]$
- The intensity model is

$$\lambda_z(t)dt = E\{dN(t) | F_{t-}\} = \lambda_0(t) \exp[\beta'Z(t)]dt$$

- Data for each subject needs to be entered in the counting process style, with a start time, stop time and censoring indicator for each event



## Data for Andersen Gill Model

ID	Start	Stop	Censoring	Age
1	0	52	0	41
2	0	16	1	43
2	16	35	1	43
2	35	49	0	43
3	0	13	1	49
4	0	26	1	35
4	26	39	1	45



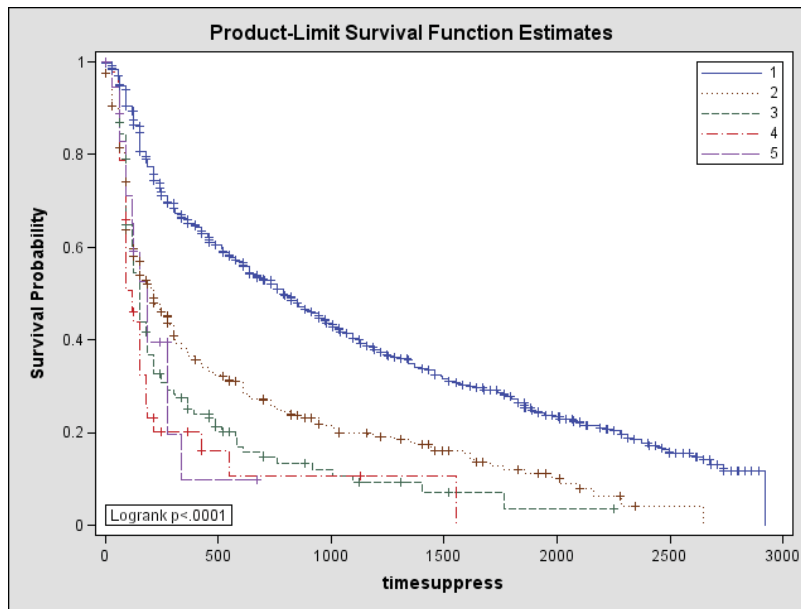
## SAS code for Andersen Gill model

```
PROC PHREG covs(aggregate);  
MODEL (Tstart,Tstop)*Status(0)=Trt Age;  
id ID;
```





## Probability of Unsuppressed Viral Load by Number of Viral Load Breakthrough



### Summary of the Number of Censored and Uncensored Values

Stratum	numsuppress	Total	Failed	Censored	Percent Censored
1	1	989	640	349	35.29
2	2	357	262	95	26.61
3	3	142	110	32	22.54
4	4	56	39	17	30.36
5	5	23	13	10	43.48
<b>Total</b>		1567	1064	503	32.10



## Recurrent Event Proportional Hazard Models

Model	Event #	Covariates	Hazards Ratio (95% CI)	P-value
Andersen-Gill	-	Cal Year > 2003	0.44 (.38, .50)	<0.0001
Marginal	1	Cal Year > 2003	0.23 (.19, .28)	<0.0001
	2	Cal Year > 2003	0.39 (.31, .50)	<.0001
	3	Cal Year > 2003	0.78 (.55, 1.10)	.16
	4	Cal Year > 2003	0.53 (.31, .90)	.02
	5	Cal Year > 2003	0.46 (.20, 1.05)	0.07



## Assignment

- Data and assignment on course website
- Email to me by Oct 27<sup>th</sup>
- No more than 20 pages
- 5 of 25 marks are for “writing and conciseness”.