Survival Analysis: An Introduction

Jaine Blayney
Bioinformatics, CCRCB

j.blayney@qub.ac.uk

Definition Of Survival Analysis

Survival analysis examines and
models the time it takes for events to occur.

Events?
Could be: recurrence of disease, death, re-offending, return to drug use...

SURVIVAL TOPICS

What this workshop covers:

- Introduction to terms commonly used in survival analysis
- Examples of most popular analysis techniques
- Suggested ways to evaluate and present results

What this workshop doesn't cover:

- In-depth mathematical formulae, big sums!
- Power analysis for survival
- Interactions

BOTH THEORY AND PRACTICE...

“He who loves practice without theory is like the sailor
who boards ship without a rudder and compass
and never knows where he may be cast."

Leonardo da Vinci, 1452-1519

SURVIVAL TOPICS

Specific areas covered:

- Overall and recurrence-free survival definitions
- Censoring
- Calculating/coding recurrence-free/overall survival from data
- Kaplan-Meier analysis
- Sub-group analysis
- Cox Proportional Hazards
- Putting it all together

SURVIVAL ANALYSIS - SOFTWARE

Survival analysis software tools:

- GraphPad Prism

- SPSS (copy/licensing available from School/IT)

- **R (open source: http://cran.r-project.org/). Relevant R packages: survival, survcomp, HMISC, Design, MASS**

Overall and Recurrence-Free Survival Definitions

Overall Survival (OS)

Overall survival refers to the time a patient survives after a particular event, e.g. date of first treatment or date of surgery. If the patient is still alive, OS is taken until the last follow-up time. If a patient has died, the end date is the date of death. This includes death from any cause.

Recurrence-Free Survival (RFS)

Recurrence-free survival refers to the time a patient survives without evidence of the disease. **End-points will vary and definitions will depend on context.** One example:

- if a patient recurs, RFS is taken from date of surgery/first treatment to date of recurrence.
- if a patient is non-recurring/alive the end date is the last follow-up.
- if a patient dies (non-recurring), the date of death is the end date.

SURVIVAL/CENSORING

Patients may have **censored survival times** if death or recurrence has not yet occurred (or there is no evidence to show that either has occurred). This could happen when:

- they drop-out of the study, e.g. they stop attending clinics for follow-up
- the study has a fixed time-line and recurrence or death occurs after the cut-off

Overall Survival/Censoring

Background:

We are interested in **overall survival** for patients (Group 1) after surgery for non-small-cell lung cancer (NSCLC):

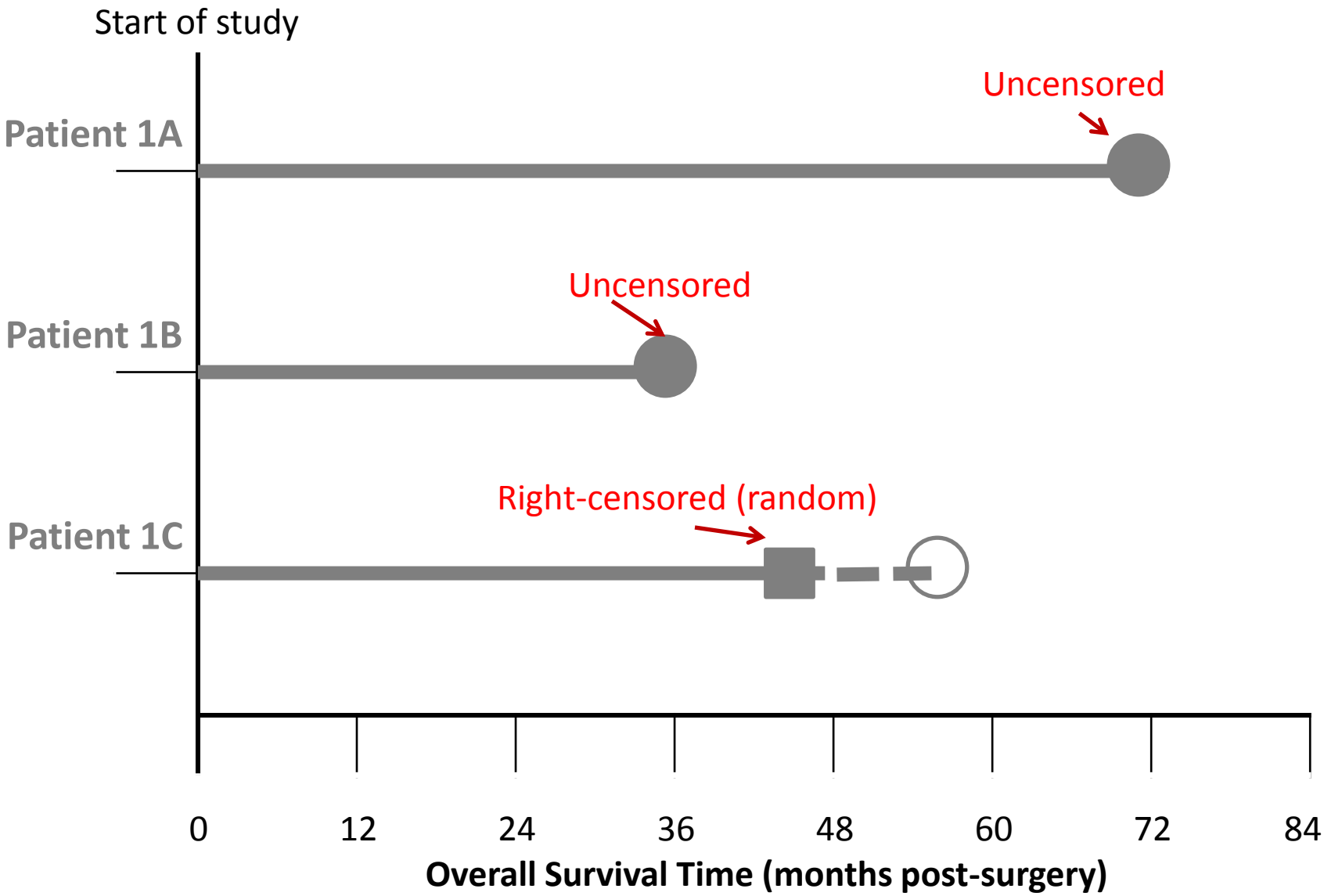
- **Patient 1A:**
  Patient 1A dies 72 months after surgery.
  Uncensored, as event has occurred.

- **Patient 1B:**
  Patient 1B dies 36 months after surgery.
  Uncensored, as event has occurred.

- **Patient 1C:**
  Patient 1C is followed for 47 months after surgery, then stops attending clinic. They die at 58 months, though this is unknown to the study.
  Right-censoring (random), as they were still alive at last check-up.

Adapted from: J Fox, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Overall Survival/Censoring



Adapted from: J Fox, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Recurrence-Free Survival/Censoring

Background:

We are interested in **recurrence-free survival (recurrence or death)** for patients after surgery for non-small-cell lung cancer (NSCLC).

- **Patient 1A:**
  Patient 1A recurs at 49 months post surgery (dies 72 months after surgery).
  Uncensored, as event (recurrence) has occurred.

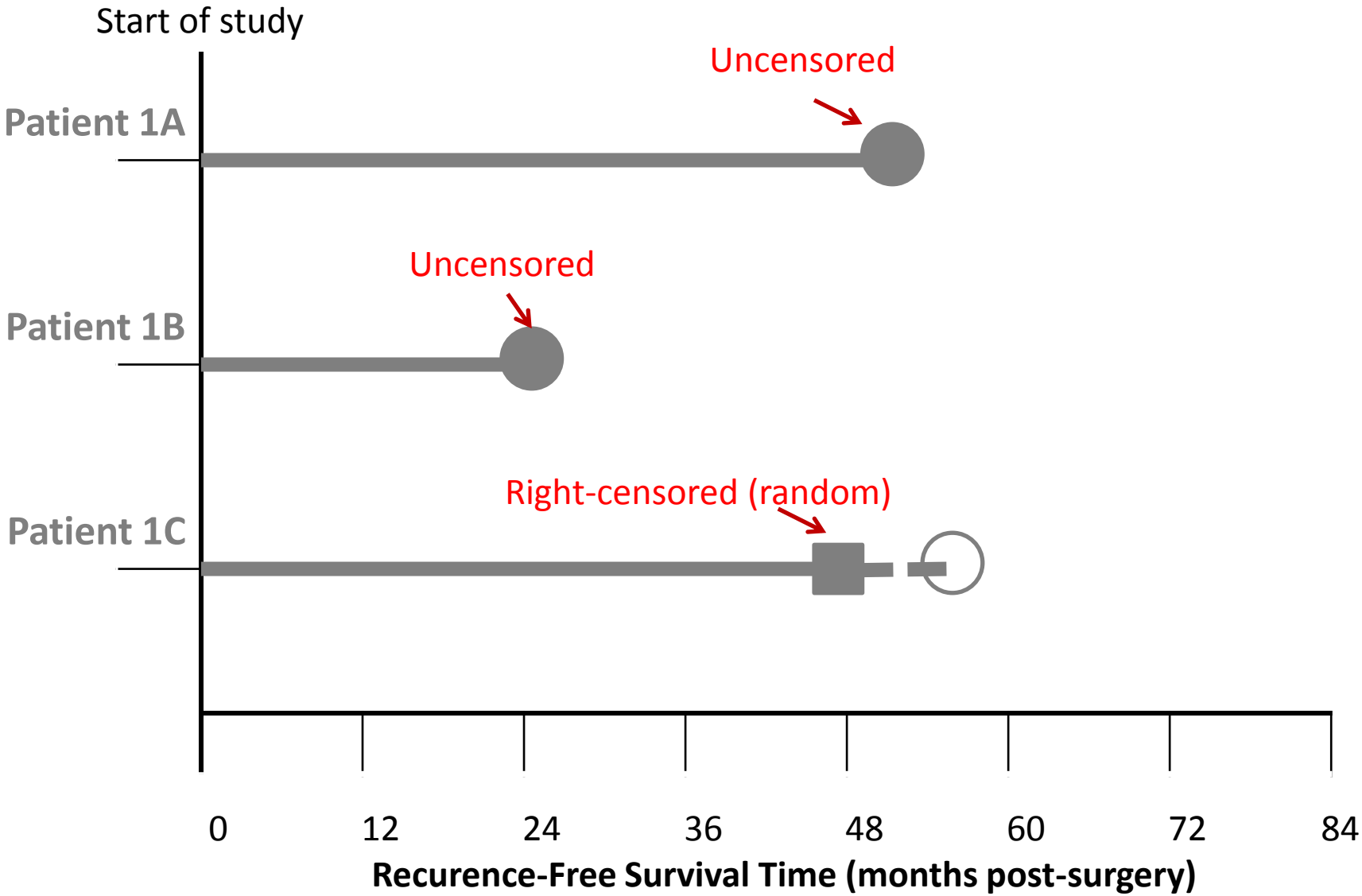
Note: 2 events, choose first

- **Patient 1B:**
  Patient 1B recurs at 25 months post surgery (dies 36 months after surgery).
  Uncensored, as event (recurrence) has occurred.

- **Patient 1C:**
  Patient 1C is followed for 47 months after surgery, then stops attending clinic. There was no evidence of recurrence up to this point (they later die at 58 months, unknown to the study).
  Right-censoring (random), as they were still alive and non-recurring at last check-up.

Adapted from: J Fox, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Recurrence-Free Survival/Censoring



Adapted from: J Fox, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Overall Survival (60-months)/Censoring

Background:

We are interested in **60-months overall survival** for patients after surgery for non-small-cell lung cancer (NSCLC).

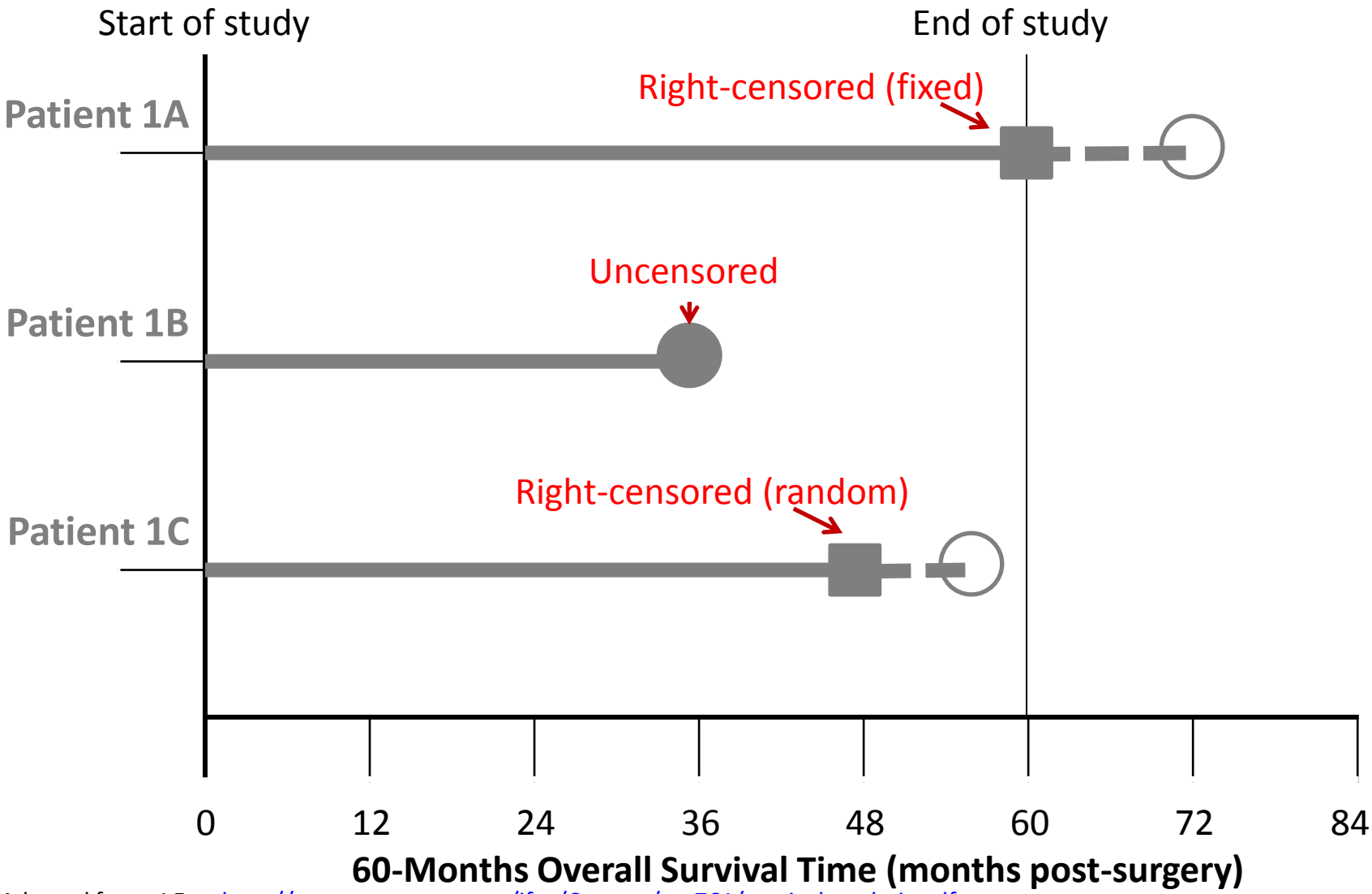
- **Patient 1A:**
  Patient 1A is followed for 60+ months years after surgery. They are still alive at the end of the 60 months, but die in month 72.
  Right-censoring (fixed), as they are still alive by the allotted time period.

- **Patient 1B:**
  Patient 1B dies 36 months after surgery.
  Uncensored, as event has occurred.

- **Patient 1C:**
  Patient 1C is followed for 47 months after surgery, then stops attending clinic. They die at 58 months, though this is unknown to the study.
  Right-censoring (random), as they were still alive at last check-up.

Adapted from: J Fox, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Overall Survival (60-months)/Censoring



Adapted from: J Fox, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Example Survival Data – Coding for Events

	Recurring	Dead/ Alive	OS (mths)	OS Status	RFS (mths)	RFS Status	OS _60mnths (mths)	OS_60mnths Status
Patient 1A	Yes	Dead	72	1	49	1	60	0
Patient 1B	Yes	Dead	36	1	25	1	36	1
Patient 1C	No	Alive	47	0	47	0	47	0

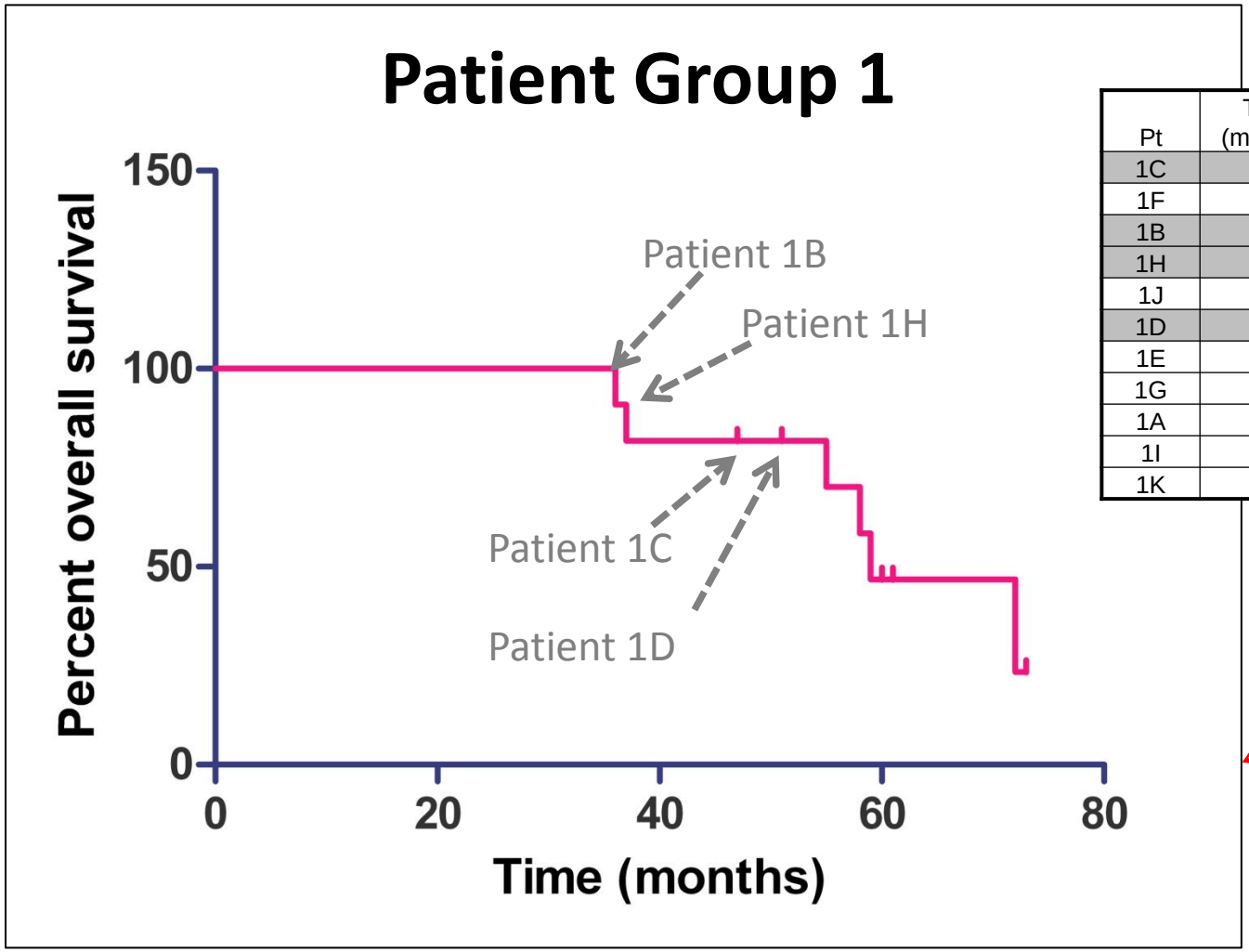
Example Survival Data – Coding for OS Events

	Date of Surgery	Date of Recurrence	Date of Death	Date of Last Follow-Up	Recurring	Dead/ Alive	OS (mths)	OS Status	RFS (mths)	RFS Status
Patient 1C	05/06/1998	NA	NA	01/06/2002	No	Alive				
Patient 1F	06/11/1999	NA	NA	04/12/2004	No	Alive				
Patient 1B	03/04/2000	03/05/2002	10/04/2003	10/04/2003	No	Dead				
Patient 1H	23/07/2002	NA	01/09/2005	01/09/2005	No	Dead				
Patient 1J	06/07/1999	NA	12/06/2004	12/06/2004	No	Dead				
Patient 1D	09/08/2001	04/11/2003	NA	06/12/2005	Yes	Alive	51	0	26	1
Patient 1E	12/04/2000	17/05/2003	NA	12/05/2005	Yes	Alive				
Patient 1G	09/04/1999	10/05/2001	NA	02/06/2005	Yes	Alive				
Patient 1A	**18/05/1999**	20/06/2003	**18/05/2005**	18/05/2005	Yes	**Dead**	**72**	**1**	49	1
Patient 1I	27/02/2000	12/05/2003	13/01/2005	13/01/2005	Yes	Dead				
Patient 1K	18/03/2000	12/04/2001	17/11/2004	17/11/2004	Yes	Dead				

Example Survival Data – Coding for RFS Events

	Date of Surgery	Date of Recurrence	Date of Death	Date of Last Follow-Up	Recurring	Dead/ Alive	OS (mths)	OS Status	RFS (mths)	RFS Status
Patient 1C	05/06/1998	NA	NA	01/06/2002	NA	NA				
Patient 1F	06/11/1999	NA	NA	04/12/2004	No	Alive				
Patient 1B	03/04/2000	03/05/2002	10/04/2003	10/04/2003	No	Dead				
Patient 1H	23/07/2002	NA	01/09/2005	01/09/2005	No	Dead				
Patient 1J	06/07/1999	NA	12/06/2004	12/06/2004	No	Dead				
Patient 1D	**09/08/2001**	**04/11/2003**	**NA**	**06/12/2005**	**Yes**	**Alive**	**51**	**0**	**26**	**1**
Patient 1E	12/04/2000	17/05/2003	NA	12/05/2005	Yes	Alive				
Patient 1G	09/04/1999	10/05/2001	NA	02/06/2005	Yes	Alive				
Patient 1A	**18/05/1999**	20/06/2003	**18/05/2005**	18/05/2005	Yes	**Dead**	**72**	**1**	49	1
Patient 1I	27/02/2000	12/05/2003	13/01/2005	13/01/2005	Yes	Dead				
Patient 1K	18/03/2000	12/04/2001	17/11/2004	17/11/2004	Yes	Dead				

SURVIVAL CURVES



Pt	Time (months)	Censor
1C	47.	0
1F	60.	0
1B	36.	1
1H	37.	1
1J	59.	1
1D	51.	0
1E	61.	0
1G	73.	0
1A	72.	1
1I	58.	1
1K	55.	1

Input

Output

Created using: GraphPad Prism, http://www.graphpad.com/prism

SURVIVAL CURVES – KAPLAN-MEIER

Previous example of survival curve using Kaplan-Meier estimate of the survivor function $\hat{S}(t)$ = the probability that a subject survives longer than time t.

Assumptions in calculating $\hat{S}(t)$:

- Let n_j denote the number of individuals alive (at risk) **just before** time *t(j)*, i.e. at the beginning of the day/week/month (including those who will die at time *t(j)* and those who are censored at time *t(j)*)
- Both deaths and censoring are taken as occurring **immediately after** time *t(j)*, i.e. *at the end of the day/week/month*
- Let d_j denote the number of failures (deaths) at time *t(j),* i.e. *at the end of the day/week/month*

- Survivor function, $\hat{S}(t)$, is calculated using:

Deaths at end of day etc

Alive at start of day etc

$$\frac{n_1-d_1}{n_1} * \frac{n_2-d_2}{n_2} * \frac{n_3-d_3}{n_3} \ \ldots * \frac{n_k-d_k}{n_k}$$

Adapted from: Ventre and Fine, http://www.lexjansen.com/pharmasug/2011/cc/pharmasug-2011-cc16.pdf

Survival Curves – Kaplan-Meier

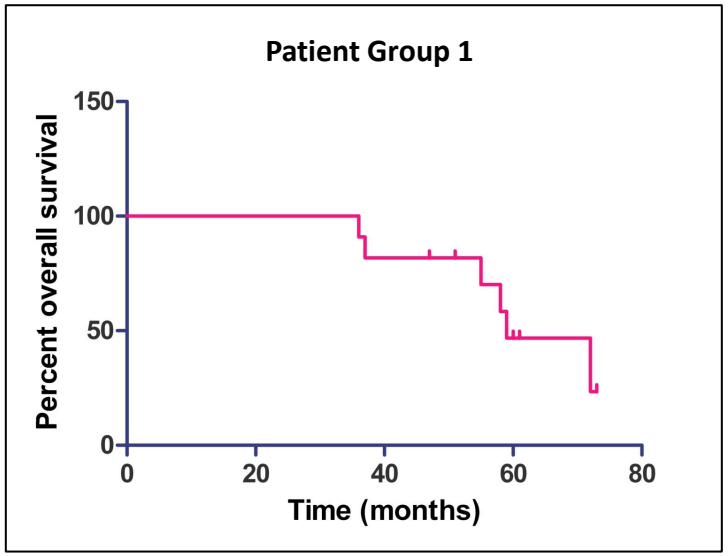
	t(j)	event
Patient 1B	36	1
Patient 1H	37	1
Patient 1C	47	0
Patient 1D	51	0
Patient 1K	55	1
Patient 1I	58	1
Patient 1J	59	1
Patient 1F	60	0
Patient 1E	61	0
Patient 1A	72	1
Patient 1G	73	0

Note: The censored cases at t(j) = 47 and 51 are omitted from n_j at t(j) = 55 Likewise, the censored cases at t(j) = 60 and 61 are omitted from nj at t(j) = 72

Start of day

End of day

j	t(j)	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0	0	11	0	1	1
1	36	11	1	0.9091	0.9091
2	37	10	1	0.9000	0.8182
3	55	7	1	0.8571	0.7013
4	58				
5	59				
6	72				



Comparison of Survival Curves – Kaplan-Meier

And if we want to compare two or more survival curves?

- **Log-rank test** is the most popular method as a test significance of significance

- No assumption regarding the distribution of survival times.

- Log-rank tests the null hypothesis that there is no difference between the populations in the probability of an event, e.g. death or recurrence, at any time point

Comparison of Survival Curves – Log-Rank

Group 2 Week	Death (=1)	Group 3 Week	Death (=1)
6	1	10	1
13	1	10	1
21	1	12	1
30	1	13	1
31	0	14	1
37	1	15	1
38	1	16	1
47	0	17	1
49	1	18	1
50	1	20	1
63	1	24	1
79	1	24	1
80	0	25	1
82	0	28	1
82	0	30	1
86	1	33	1
98	1	34	0
149	0	35	1
202	1	37	1
219	1	40	1
		40	1
		40	0
		46	1
		48	1
		70	0
		76	1
		81	1
		82	1
		91	1
		112	1
		181	1

Week	Overall Observed Deaths	Expected Deaths – Group 2	Expected Deaths – Group 3	Observed Remainder – Group 2	Observed Remainder – Group 3
6	1/51	0.392157	0.607843	19	31
10	2/50	0.76	1.24	19	29
12					
13					
14					
15					
…					
Total (Expected)		Sum	Sum		
Total (Observed)		14	28		

X^2 (Chi-square) statistic to test null hypothesis =

$$\frac{(\text{Observed Group 2} - \text{Expected Group 2})^2}{\text{Expected Group 2}} + \frac{(\text{Observed Group 3} - \text{Expected Group 3})^2}{\text{Expected Group 3}}$$

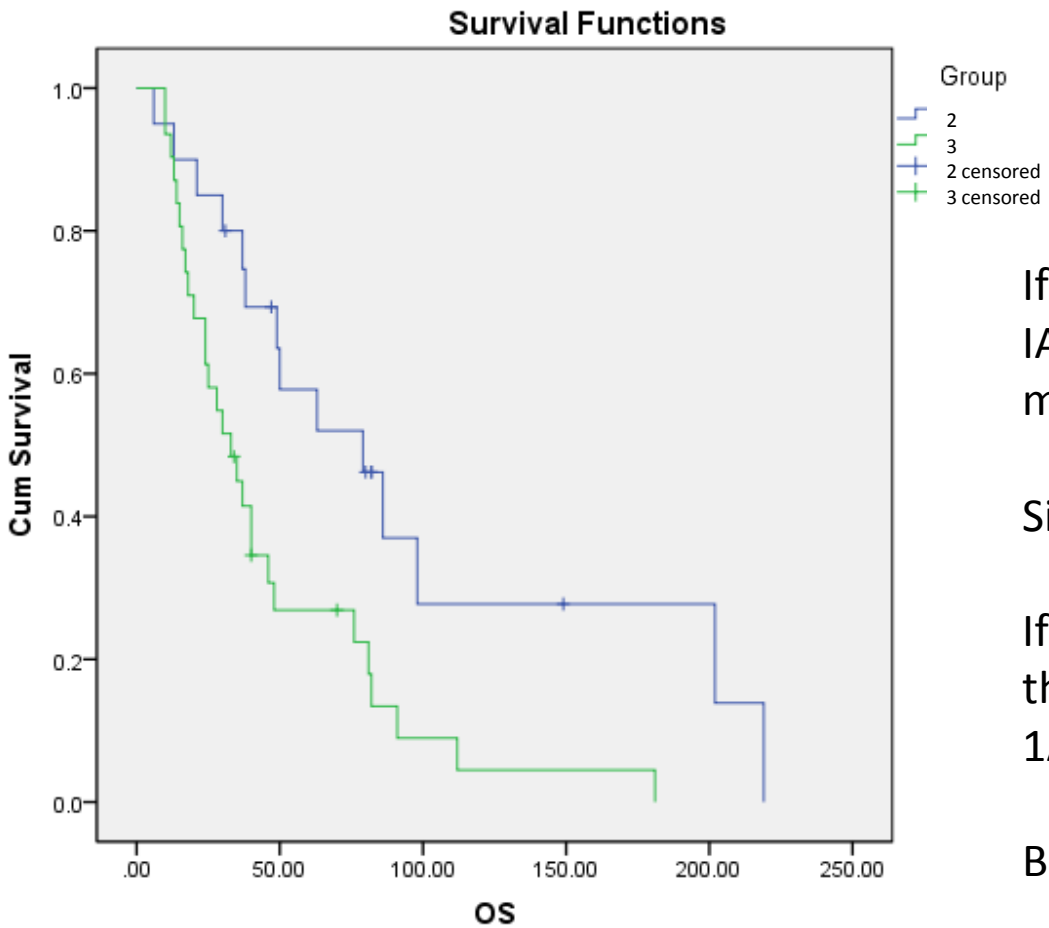
Degrees of freedom = Number of groups – 1 = 2- 1 = 1

From table of X^2 distributions, find p-value

Probability that such an (extreme) chi-square value could be obtained by chance =0.05/0.01

Taken from: Bland and Altman, BMJ, 328, 1073, 2004

Comparison of Survival Curves – Log-Rank



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	7.497	1	.006

Test of equality of survival distributions for the different levels of Group.

Groups 2 and 3 be drawn from factors such as pathological staging, e.g. NSCLC Stage IA vs IB

If comparing multiple groups e.g. Stage IA, IB, IIA, IIB, then need to correct for multiple comparisons

Simplest method: Bonferroni correction

If comparing 3 curves, Stage IA, IB, IIA, then 3 possible comparisons: IA/IB, 1A/IIA and IB/IIA

Bonferroni correction

$$= \frac{\text{significance cut-off}}{\text{number of comparisons}}$$

$$= 0.05/3 = 0.0167$$

Calculated using SPSS 19 using data from Bland and Altman, BMJ, 328, 1073, 2004

SUB-GROUP ANALYSIS

And if we want to consider the effect of other variables?

- E.g. How does survival differ for different age groups/pathological staging?
- Could separate groups eg <60yrs/Stage IA; ≥60yrs/Stage IA; <60yrs/Stage IB; ≥60yrs/Stage IB and analyse by Kaplan-Meier?
- Sub-group analysis, proceed with caution! (see Sleight, 2000)

"In retrospect, perhaps one of the most important results in the ISIS trials was the analysis of the results by astrological star sign .... We were ... able to divide our population into 12 subgroups by astrological star sign. Even in a highly positive trial such as ISIS-2 (International Study of Infarct Survival), in which the overall statistical benefit for aspirin over placebo was extreme (P <0.00001), division into only 12 subgroups threw up two (Gemini and Libra) for which aspirin had a nonsignificantly adverse effect"

Sleight, Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them!, Curr Control Trials Cardiovasc Med, 1, 25–27, 2000

Cox Proportional Hazards Regression Model

If we want to consider the effect of multiple variables in relation to RFS and/or OS, consider Cox <u>proportional hazards</u> regression model

Hazard Function and Hazard Ratio:

Hazard function: h(t) is a function of the probability of an event in the time interval [t, t+i], given that the individual has survived up to time t

In the **single variable** case eg age, given the baseline hazard function, i.e. no explanatory variable : $h_0(t)$

$$h(t) = h_0(t)*\exp(X_1\beta_1)$$

Not dependent on time

Where X_1 is an explanatory continuous or categorical variable that is modelled to predict an individual's hazard and where β_1 is a regression coefficient of the predictor variable.

Eg: model of age in ovarian cancer (estimated β_1) :

$$h(t|age) = h_0(t)*\exp(age * 0.02)$$

Unit increase

Taken from: http://www.medcalc.org/manual/cox_proportional_hazards.php

Cox Proportional Hazards Regression Model

Hazard ratio:

Consider two patients, patient 1 aged x_1 and patient 2 aged x_2. We are interested in the relationship of this covariate with RFS.

$$\text{Hazard Ratio} = \frac{h_0(t)\exp(x_2\hat{\beta})}{h_0(t)\exp(x_1\hat{\beta})}$$

$$= \exp((x_2 - x_1)\hat{\beta})$$

$\hat{\beta}$ = Estimate of regression coefficient

If, in this case $\hat{\beta}$ = 0.02, and patient 1 is aged 60 and patient 2 is aged 70 then patient 2 is 1.22 (=exp(10*0.02)) times more likely to experience recurrence or death than patient 1

Cox PH Regression Model assumes proportional hazards and linearity within each group of covariates

Adapted from: http://userwww.service.emory.edu/~poldd/survival3.pdf

Cox Proportional Hazards Regression Model

We are interested in the effect of five covariates (only variables for which we have complete data): Histology, Performance Status, Age, Grade and Stage within three cohorts of cancer patients in relation to RFS and OS.

```
#In R – input:

uni_ps <- read.table("file.txt", header= TRUE)

names(uni_ps)

attach(uni_ps)

fitps <- coxph(Surv(OS, OS_event)~(PS)+strata(study), method = "breslow")

summary(fitps)
```

Input file

Returns variables and attaches them to R workspace

Calls Cox PH function, considering effect of covariate Performance Status (PS) on overall survival (OS and OS_event). Method can be Breslow or Efron, this refers to approach for dealing with event time ties. Efron is more accurate.

Histology
Histology1
Histology2
Performance Status
Age (yrs)
<60
=>60
Grade
I
II
Stage
I, II, III
IV

Cox Proportional Hazards Regression Model

Hazard Ratio for unit increase in PS, values: 0, 1, 2

```
#In R – output from summary(fitps):

Output:
Call:
coxph(formula = Surv(OS, Os_stat) ~ (PS), data = strata(study), method = "breslow")
 n= 161

                coef        exp(coef)    se(coef)     z Pr(>|z|)
PS              0.3894      1.4762       0.1489       2.615        0.00893 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef)   exp(-coef)   lower .95 upper .95
PS              1.476       0.6774       1.102     1.977

....
```

$\hat{\beta}$

P-value

95% Confidence Interval

Cox Proportional Hazards Regression Model

We now need to check if the covariate meets the proportional hazards assumption.

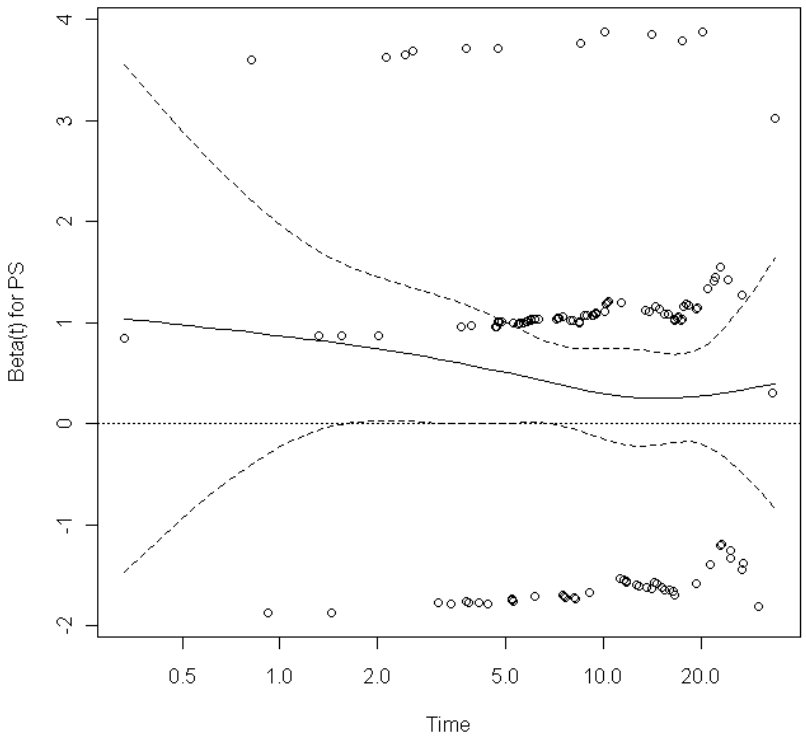
Fit from previous stage

#In R – input:

```
ph_fit_ps <-cox.zph(fitps, transform ='log')
ph_fit_ps
plot(ph_fit_ps)
abline(h=0, lty=3)
```

Returns summary

Graphical plot of the Schoenfeld residuals versus log(time)



#In R – output:

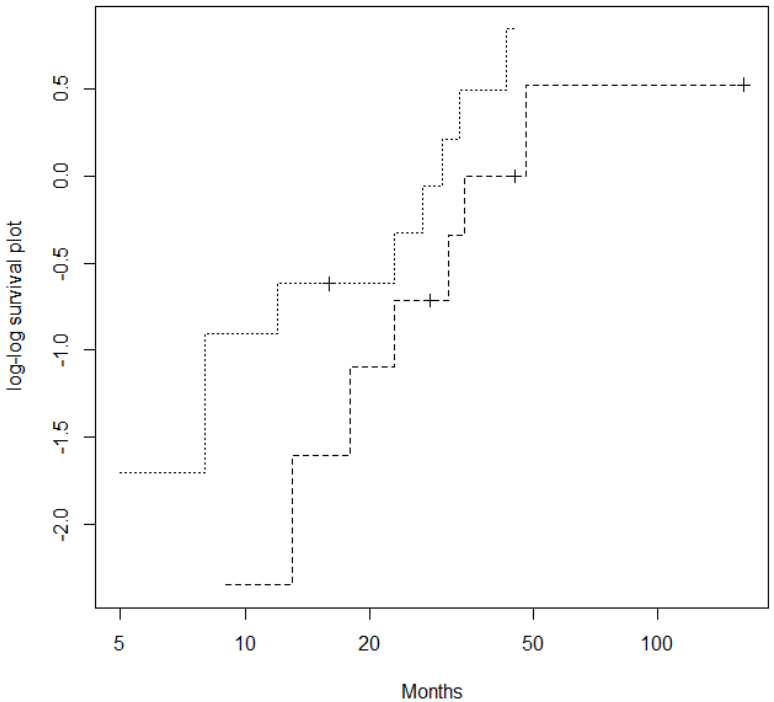
	rho	chisq	p
PS	-0.0896	0.994	0.319

P is >0.05!
Check both plots and output. PS appears to fit the PH assumption

Cox Proportional Hazards Regression Model

NOTE: You can check the PH assumption in a number of ways including:

Log-log plot
By this method, a plot of the logarithm of time against the logarithm of the negative logarithm of the estimated survivor function. If curves are not crossing each other then PH assumption is satisfied.




```
>data(leukemia)
>leukemia.surv <- survfit(Surv(time, status) ~ x, data = leukemia)
>plot(leukemia.surv, lty = 2:3)
>lsurv2 <- survfit(Surv(time, status) ~ x, data= leukemia)
>plot(lsurv2, lty=2:3, fun="cloglog", xlab="Months", ylab="log-log survival plot")
```

Cox Proportional Hazards Regression Model

And if the covariate does not meet the proportional hazards assumption?

Options:

- Omit the covariate
- Check for interactions of the covariate with time (this has occurred for tumour size in both breast and lung cancer) see Fox: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf
- Stratify by the covariate

#In R – input:

```
fitps <- coxph(Surv(OS, OS_event)~(PS)+strata(study), method = "breslow")
```

Strata = stratified Cox model; separate baseline hazard functions are allowed for each stratum. The stratum-specific analyses are pooled to get an overall estimate. In multi-centre studies, you normally stratify by 'centre'. Can stratify by more than one variable.

Cox Proportional Hazards Regression Model

Next, we need to check if the covariate (continuous) is linear:

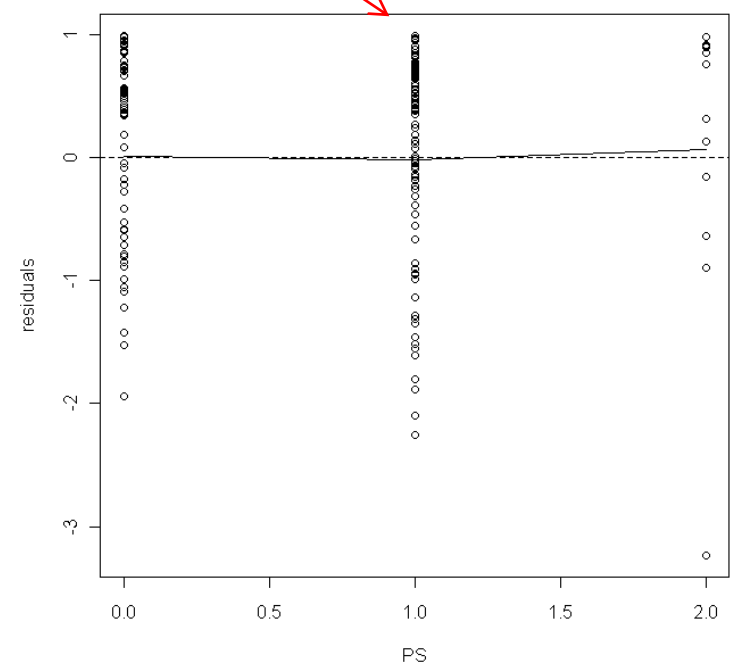
- Use Martingale residuals:

Graphical plot of the Martingale residuals

#In R – input:

```
res <- residuals(fitps, type='martingale')
X <- as.matrix(uni_ps[,("PS")]) # matrix of covariates

plot(X[,1], res, xlab=c("PS")[1], ylab="residuals")
abline(h=0, lty=2) + lines(lowess(X[,1], res, iter=0))
```



In this case, PS was modelled as continuous and appears to be linear

If non-linear, what do you?

- Consider discretizing into ranges, though be careful as these need to be meaningful cut-offs. Even then can result in serious loss of information.
- OR fit a spline model eg restricted cubic spline (see Harrell, http://lib.stat.cmu.edu/S/Harrell/help/Hmisc/html/rcspline.plot.html)

Cox Proportional Hazards Regression Model

Consider univariate analysis, then multivariate analysis using Cox Proportional Hazards model.

Definitions:

- **Univariate**

  One covariate – similar to the log-rank test

- **Multivariate**

  More than one covariate. In medical science a multivariate model refers to multi-explanatory variables in relation to RFS and OS etc.

  However, in statistics, the term of multivariate model is used in the sense of multivariate responses (outcomes). Clearly define your terms!

Cox Proportional Hazards Regression Model

Clinical Factor		Univariate – Disease-Free Survival		
	N(n)	**HR**	**CI**	**p-value**
Histology				
Histology1	50 (24)	1		
Histology2	45 (30)	3.245	2.600-4.235	<0.001
Performance Status	95 (54)	1.305	1.100-1.600	0.304
Age				
<60	46 (21)	1		
=>60	44 (23)	1.304	1.091-1.821	0.204
Grade				
I	35 (20)	1		
II	60 (34)	0.803	0.503-0.916	0.612
Stage				
I, II, III	50 (26)	1		
IV	45 (28)	1.200	1.023-1.603	0.402

Cox Proportional Hazards Regression Model

How to select variables to take forward into multivariate model?

- Not by univariate p-value, more benefit from clinical knowledge and relevance
- Can use Akaike's information criterion (AIC) to select variables (backward selection)

Then assess quality of the fitted model using:

- **AIC**
- BIC (Bayesian information criterion)
- Deviance (-2 log likelihood)

And check for discriminatory power of model using:

- Concordance index (c-index)

Cox Proportional Hazards Regression Model

#In R – using stepAIC (Mass)

```
fitall <- coxph(Surv(OS, Os_stat)~agec+histc+stagec+grade+PS+strata(study), method = "breslow")
stepAIC(fitall, direction=c("backward"))
```

#Output

```
Start:  AIC=1080.18
Surv(OS, Os_stat) ~ agec + histc + stagec + grade + PS

          Df   AIC
- grade    1 1078.2
- agec     1 1078.3
- histc    1 1079.8
<none>       1080.2
- stagec   1 1081.5
-PS        1 1085.1
-....
Step:  AIC=1076.35
Surv(OS, Os_stat) ~ histc + stagec + PS

          Df   AIC
- histc    1 1076.1
<none>       1076.3
- stagec   1 1077.8
- PS       1 1082.3
```

#Output contd...

Want this value as low as possible

```
Step:  AIC=1076.13
Surv(OS, Os_stat) ~ stagec + PS

          Df   AIC
<none>        1076.1
- stagec   1  1078.0
-PS        1  1081.8

Call:
coxph(formula = Surv(OS, Os_stat) ~ stagec + PS, data = strata(study),
   method = "breslow")

          coef     exp(coef) se(coef)   z     p
stagec    0.119    1.13      0.0605 1.97 0.0490
PS        0.420    1.52      0.1498 2.80 0.0051

Likelihood ratio test=10.6  on 2 df, p=0.00495  n= 161
```

Does removal of a covariate push the AIC up?

Cox Proportional Hazards Regression Model

What if two (or more) models, e.g. Stage/PS and Stage/PS/Histology, produce similar values?

- Check AIC, BIC and deviance values, then c-index

```
fit3 <- coxph(Surv(OS, Os_stat)~histc+stagec+PS+strata(study), method = "breslow")
fit2 <- coxph(Surv(OS, Os_stat)~stagec+PS+strata(study), method = "breslow")

stepAIC(fit3, direction=c("backward"))      #AIC
stepAIC(fit3, k=log(161))                   #BIC
anova(fit3)                                 #Deviance
```

#Output: deviance

-> anova(fit3)

-Analysis of Deviance Table - Cox model: response is Surv(OS, Os_stat)

--Terms added sequentially (first to last)

-	loglik	Chisq	Df	Pr(>|Chi|)
-NULL	-541.37			
-histc	-540.50	1.7471	1	0.186240
-stagec	-539.17	2.6625	1	0.102737
-PS	-535.17	7.9858	1	0.004714 **

-Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Look for loglik and p-value columns

Cox Proportional Hazards Regression Model

-Output: AIC

```
-> stepAIC(fit3, direction=c("backward"))
-Start:  AIC=1076.35
-Surv(OS, Os_stat) ~ histc + stagec + PS

-                Df    AIC
-- histc          1 1076.1
-<none>             1076.3
-- stagec         1 1077.8
-- PS             1 1082.3

-Step:  AIC=1076.13
-Surv(OS, Os_stat) ~ stagec + PS

-                Df    AIC
-<none>               1076.1
-- stagec           1 1078.0
-- PS               1 1081.8
-Call:
-coxph(formula = Surv(OS, Os_stat) ~ stagec + PS, data =
strata(study), method = "breslow")

-                coef     exp(coef) se(coef)    z      p
-stagec          0.119      1.13   0.0605 1.97 0.0490
-PS              0.420      1.52   0.1498 2.80 0.0051

-Likelihood ratio test=10.6  on 2 df, p=0.00495  n= 161
```

Reference values

Does removal of a covariate push the AIC up?

-Output: BIC

```
-> stepAIC(fit3, k=log(161))
-Start:  AIC=1085.59
-Surv(OS, Os_stat) ~ histc + stagec + PS

-                Df    AIC
-- histc          1 1082.3
-- stagec         1 1084.0
-<none>             1085.6
-- PS             1 1088.5

-Step:  AIC=1082.29
-Surv(OS, Os_stat) ~ stagec + PS

-                Df    AIC
-- stagec         1 1081.1
-<none>             1082.3
-- PS             1 1084.9

-Step:  AIC=1081.1
-Surv(OS, Os_stat) ~ PS

-                Df    AIC
-<none>               1081.1
-- PS             1     1082.7
-Call: coxph(formula = Surv(OS, Os_stat) ~ PS, data =
strata(study),  method = "breslow")

-                coef  exp(coef) se(coef)    z      p
-PS              0.389 1.48       0.149      2.61 0.0089

-Likelihood ratio test=6.73  on 1 df, p=0.0095  n= 161
```

Cox Proportional Hazards Regression Model

Deviance and BIC scores suggest a one-factor model (PS). AIC suggests a two-factor model: PS and stage

Check for discriminatory power using c-index (score of 0.5 is no better than random, usually 0.6-0.7 for survival models)

Also parsimony principle

```
#R input

fitPS <- coxph(Surv(OS, Os_stat)~PS,strata(study), method = "breslow")
fitPSStage <- coxph(Surv(OS, Os_stat)~stagec+PS,strata(study), method = "breslow")

survConcordance(Surv(OS, Os_stat) ~predict(fitPS), data=uni_ps)
survConcordance(Surv(OS, Os_stat) ~predict(fitPSStage), data=uni_ps)
```

Value for stage and PS model is higher: 0.5852062. So go with stage and PS

```
#R input  (PS only)

[1] 0.5612245
$stats
     agree    disagree      tied.x   tied.time incomparable
      3966       2544        5103           6        1261
$call
survConcordance(formula = Surv(OS, Os_stat) ~ predict(fitPS), data = uni_ps)
```

Cox Proportional Hazards Regression Model

Clinical Factors	Composition
Grade	I (50), II (45)
Age	Median: 63 yrs
Stage	I (10), II (20), III (20), IV (45)
Histology	Epithelial (60), Mixed (25), Sarcomatous (20)
Performance status	0 (30), 1 (40), 2 (25)

95 patients used in univariate and multivariate analysis. Explain why 17 were excluded.

Results:

Of a total of 112 patients, those with incomplete information for the following clinical factors were excluded: histology (n = 3), stage (n = 4) and histology/staging (n = 10). The final multivariate data-set thus comprised 95 patients (Table 1). The median RFS for this group was 6.5 (CI 5.4-7.2) months; for OS the median was 12.1 (CI 10.1-15.1) months.

Cox Proportional Hazards Regression Model

Table 1. Prognostic value of clinical factors for progression-free survival.

Clinical factors	N	Progression or death, n	Univariate HR	Univariate [95% CI]	Univariate p-Value	Multivariate HR	Multivariate [95% CI]	Multivariate p-Value
Age (years)					0.040			0.045
⩾60	232219		1			1		
<60	291286		**1.21**	**[1.01–1.46]**		**1.26**	**[1.01–1.58]**	
Histological type					0.007			0.004
Epithelial	310299		1			1		
Mixed	10299		**1.39**	**[1.09–1.77]**		**1.43**	**[1.12–1.82]**	
Sarcomatous	38 36		**1.44**	**[1.01–2.06]**				
Stage of disease					0.027			0.044
Stage I or II	118 112		1			1		
Stage III	190184		1.06	[0.83–1.35]				
Stage IV	154151		**1.37**	**[1.05–1.8]**		**1.28**	**[1.01–1.62]**	
Performance status					<0.001			0.001
0	136128		1			1		
1	313305		**1.44**	**[1.24–1.68]**		**1.36**	**[1.13–1.63]**	
2	74 72		**2.08**	**(Log linear trend)**		**1.84**	**(Log linear trend)**	
Haemoglobin concentration					<0.001			0.044
⩾12 g/dl	379364		1			1		
<12 g/dl	141139		**1.47**	**[1.2–1.81]**		**1.29**	**[1.01–1.64]**	
Histological diagnosis					0.343			
Probable	54 52		1					
Definite	447431		1.16	[0.86–1.56]				
WBC count (10^9/l)					0.006			
–, 7.4[	166159		1					
[7.4–9.5[	180174		**1.37**	**[1.09–1.71]**				
[9.5, –	175170		**1.38**	**[1.1–1.72]**				
Interaction: histological diagnosis × WBC count								0.031
When histological diagnosis is probable and WBC count is	NA							
<7.4 × 10^9/l						1		
⩾7.4 × 10^9/l						0.48	[0.19–1.22]	
When histological diagnosis is definite and WBC count is	NA							
<7.4 × 10^9/l						1		
⩾7.4 × 10^9/l						**1.39**	**[1.09–1.78]**	
Platelets count (10^9/l)					0.041			
–, 315[	173162		1					
[315–435[	172169		1.22	[0.97–1.52]				
[435, –	176172		**1.33**	**[1.06–1.67]**				0.451

At least 10 events per predictor variable

Results of AIC backward selection. Present final model separately

Taken from: Francart J, Vaes E, Henrard S, et al: A prognostic index for progression-free survival in malignant mesothelioma with application to the design of phase II trials: A combined analysis of 10 EORTC trials. Eur J Cancer 45: 2304-2311, 2009

Cox Proportional Hazards Regression Model

More results (summary of patients)...

From the 598 patients registered, 75 were excluded (for ineligibility (*n* = 41), for incoherent or missing data (*n* = 9), histological diagnosis not definite or probable (*n* = 25)). The remaining 523 patients were predominantly male (83%) with a performance status of 0 or 1 (86%). The median age was 58 years (range: 19–80 years). Mesothelioma diagnosis was definite in 89% and probable in 11%. Histological type was epithelial in 69%, sarcomatous in 8% and mixed in 23%. Forty-one percent of patients had a stage III disease and 33% had a stage IV disease. The median WBC count, platelet count and haemoglobin concentration were 8.4×10^9/l (range: 3.2–18.3×10^9/l), 374×10^9/l (range: 153–968×10^9/l) and 13.2 g/dl (range: 6.4–19.6 g/dl), respectively. LDH level was abnormal in 18% (58/322) and alkaline phosphatase level was abnormal in 31% (109/355).

Median follow-up time was 9.9 months (IQR: 4.5–22.8 months). Of the 523 patients, 485 (93%) progressed during follow-up and 445 (85%) died, leading to 3% (18/523) of progression-free survivors after the follow-up. Median survival and median PFS were 9.1 months (95% confidence interval (CI): 8.3–10.2 months) and 3.9 months (95% CI: 3.4–4.3 months), respectively.

Taken from: Francart J, Vaes E, Henrard S, et al: A prognostic index for progression-free survival in malignant mesothelioma with application to the design of phase II trials: A combined analysis of 10 EORTC trials. Eur J Cancer 45: 2304-2311, 2009

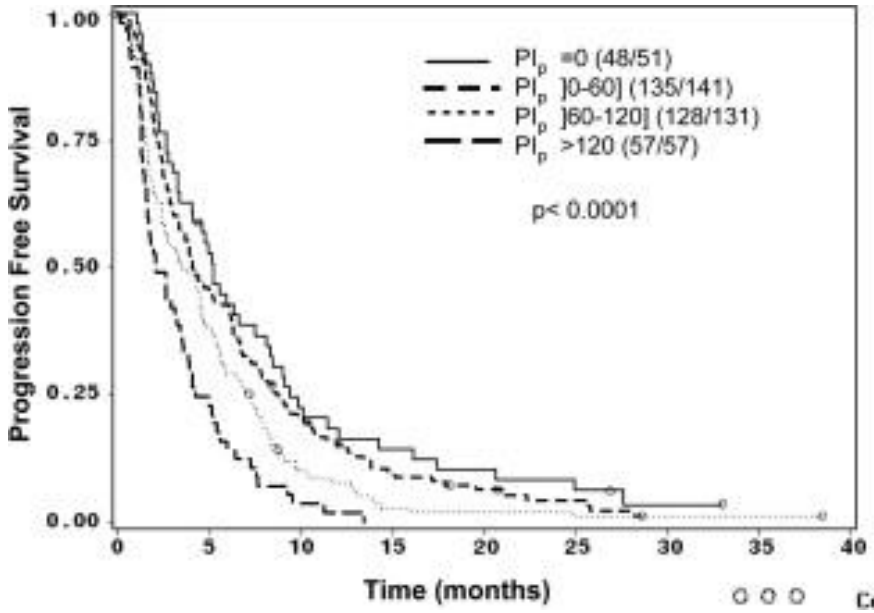
Cox Proportional Hazards Regression Model

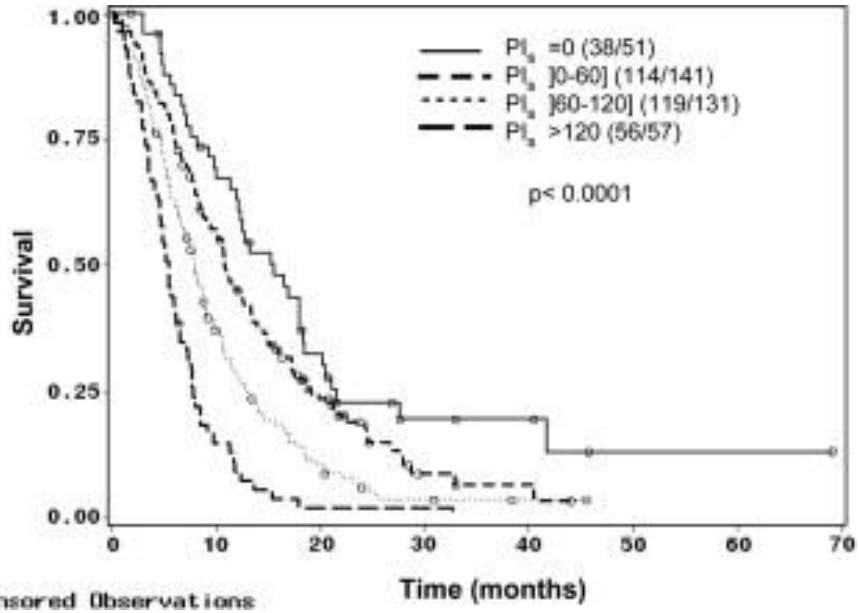
More results (reporting univariate)...

In univariate analysis of PFS (Table 1), poor prognosis was associated with age < 60 years, high WBC count (7.4×10^9/l), haemoglobin concentration < 12 g/dl, performance status greater than 0 and an abnormal LDH level. A moderate increase in platelet count was associated with **an increased risk** but without reaching statistical significance; a high level was clearly significantly associated with an increased risk. Stage IV disease was **associated with a poor prognosis** compared with stage I and II disease. **A better prognosis** was associated with epithelial histological type as compared with mixed or sarcomatous type. Time interval since diagnosis ($p = 0.450$), gender ($p = 0.875$), alkaline phosphatase level ($p = 0.121$) and certainty of histological diagnosis ($p = 0.343$) did not reach significance for predicting PFS. **For the variables of histological type, stage of disease and WBC count, different categories had similar HRs, and thus they were regrouped before their inclusion in the multivariate analysis.**

Taken from: Francart J, Vaes E, Henrard S, et al: A prognostic index for progression-free survival in malignant mesothelioma with application to the design of phase II trials: A combined analysis of 10 EORTC trials. Eur J Cancer 45: 2304-2311, 2009

Cox Proportional Hazards Regression Model



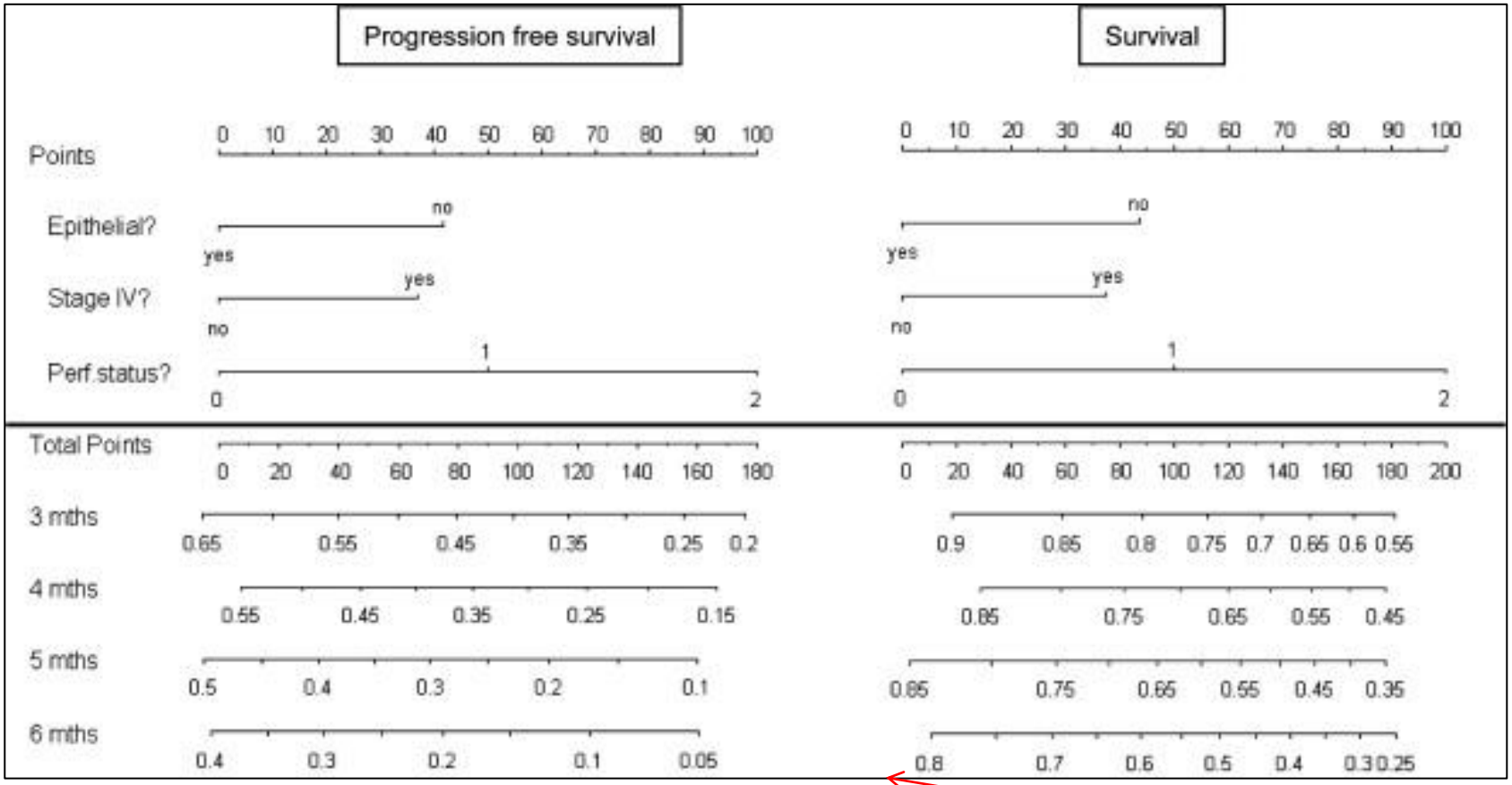


Censored Observations

Risk categories were identified from Cox model based on scores for PS, histology and staging.

Survival curves taken from: Francart J, Vaes E, Henrard S, et al: A prognostic index for progression-free survival in malignant mesothelioma with application to the design of phase II trials: A combined analysis of 10 EORTC trials. Eur J Cancer 45: 2304-2311, 2009

Cox Proportional Hazards Regression Model



Nomograms –visual interpretations of Cox models.Taken from: Francart J, Vaes E, Henrard S, et al: A prognostic index for progression-free survival in malignant mesothelioma with application to the design of phase II trials: A combined analysis of 10 EORTC trials. Eur J Cancer 45: 2304-2311, 2009

AND FINALLY...

For further reading

Anything by:

- Frank Harrell (both statistical theory and R implementations)
- Terry Therneau (statistics theory and R implementations) and
- Doug Altman (from a medical statistics point of view)

Survival Analysis:
An Introduction

R Tutorial

Jaine Blayney

Bioinformatics, CCRCB

j.blayney@qub.ac.uk

WORKING WITH SURVIVAL DATA

1 KAPLAN-MEIER ANALYSIS **2**

1.1 Data Source 2

1.2 Reading in Data 3

1.3 FormattingData 5

1.4 Load Survival Package 6

1.5 Create a Survival Object 6

1.6 Plot Survival Curves 7

1.7 Compare Survival Curves 8

PRACTICE EXAMPLES **9**

Question 1: Plotting/Comparing Survival Curves Part I 9

Question 2: Plotting/Comparing Survival Curves Part II 11

2 COX PROPORTIONAL HAZARDS **13**

2.1 Testing PH Assumption 13

2.2 Testing Linearity 15

2.3 Univariate Analysis 17

2.3 Multivariate Analysis 19

PRACTICE EXAMPLES **21**

Question 3: Developing A Cox PH Model 21

Question 4: Variable Selection 22

SUGGESTED SOLUTIONS **23**

The material in this tutorial has been adapted in part from the UCLA Statistical Computing Seminars – Survival Analysis Series (http://www.ats.ucla.edu/stat/)

1 KAPLAN-MEIER ANALYSIS

1.1 Data Source

The hmohiv data-set is drawn from a study of HIV positive patients. The study examined whether there was a difference in survival times of HIV positive patients between those who had used intravenous drugs and those who had not. This data-set has been taken from [http://www.ats.ucla.edu/stat/r/examples/asa/asa_ch2_r.htm].

The hmohiv data-set contains the variables: **patient ID, overall survival time, study entry date, date last seen/study end date, age, drug use** and an **event/censored** variable.

ID	time	age	drug	censor	entdate	enddate
1	5	46	0	1	15/05/1990	14/10/1990
2	6	35	1	0	19/09/1989	20/03/1990
3	8	30	1	1	21/04/1991	20/12/1991
4	3	30	1	1	03/01/1991	04/04/1991
5	22	36	0	1	18/09/1989	19/07/1991
6	1	32	1	0	18/03/1991	17/04/1991
7	7	36	1	1	11/11/1989	11/06/1990
8	9	31	1	1	25/11/1989	25/08/1990
…	…	…	…	…	…	…

1.2 Reading in Data

First, read in the table. The variables in the table are separated by commas and there is a header row.

```
>hmohiv<-read.table("http://www.ats.ucla.edu/stat/R/examples/asa/hmohiv.csv",
sep=",", header = TRUE)
```

To check to see what sort of object (format) that you have created, use:

```
>class(hmohiv)
```

To confirm the data that is contained within hmohiv:

```
>hmohiv
```

In order to access variables:

```
>hmohiv$ID
>hmohiv$time
```

NOTE: The attach() function in R can be used to make objects within dataframes (dataframes=tables, where rows = patients and columns = variables) accessible in R with fewer keystrokes. As an example instead of hmohiv$ID just type:

```
> attach(hmohiv)
>ID
>time
```

Though be careful if you have a number of data-sets open with the same variable names! To detach:

```
>detach(hmohiv)
>ID
>time
```

Or if you were interested in viewing patient ID and age together (no arguments before “,” = all rows, c(1,3) = column 1 and column 3, where c() is a function to combine arguments to form a vector), you would use:

```
> hmohiv[,c(1,3)]
```

You could create a smaller data-frame with only patients 1-5:

```
> attach(hmohiv)
> mini<-hmohiv[ID<=5,]
```

To check the content and format of the object that you have created:

```
>mini
>class(mini)
```

To check all objects that you have in memory:

```
>ls()
```

To remove all objects in memory:

```
>rm(list =ls())
```

1.3 Formatting Data

In some cases, you may only have dates in your data. In order to calculate the actual survival times, you could first import your data into Excel and read the revised table into R. Or you could achieve this within R itself.

First, create a new column:

```
>hmohiv["TIME_CHECK"]<-NA
```

Then, calculate the difference, in days between end date (column 7) and start date (column 6) Use as.numeric(), otherwise you will be given a wordy answer. Use as.Date() as the original data is in a data.frame format.

```
>hmohiv$TIME_CHECK<-as.numeric(as.Date(hmohiv[,7], "%m/%d/%Y")-
as.Date(hmohiv[,6], "%m/%d/%Y"))
```

Check on output:

```
>hmohiv$TIME_CHECK
```

To convert from days to months (rough workaround):

```
>hmohiv$TIME_CHECK<-round(hmohiv$TIME_CHECK/30.5)
```

To double-check your answer against the original time in the table:

```
>hmohiv[,c(2,8)]
```

1.4 Load Survival Package

First, load the survival package:

```
>install.packages("survival")
>library(survival)
```

You can get more information on the package from: http://cran.r-project.org/web/packages/survival/survival.pdf

1.5 Create a Survival Object

The initial step is to create a survival object:

```
>s_obj<-Surv(time, censor)
```

What does the survival object do? Look at it in the context of time/censoring. Create a new column and copy the results of survival object into it:

```
>hmohiv["s_obj"]<-NA
>hmohiv$s_obj<-Surv(time, censor)
```

Then compare it to the time and censor columns:

```
>hmohiv[,c(2,5,9)]
```

In the survival object, patients with censored times are represented by 4+, 1+ etc. Patients who experienced an event have unaltered times, eg 7, 9 etc.

Now, try:

```
>detach(hmohiv)
> my.surv<-Surv(time, censor)
```

What happens and why?

An alternative, if you don't want to use 'attach' is:

```
>with(hmohiv, Surv(time, censor))
```

1.6 Plot Survival Curves

Use the survival object within survfit() to create survival curves.

For one single curve with a basic plot:

```
> my.KMest1 <- survfit(Surv(time, censor)~ 1, conf.type="none")
> plot(my.Kmest1)
```

With confidence intervals:

```
>my.surv<-Surv(time, censor)
>my.KMest2 <- survfit(my.surv~1, conf.int=0.95)
>plot(my.KMest2)
```

Or, all in one go:

```
>plot(survfit(Surv(time, censor)~1, conf.int=0.95))
```

For two curves, using two separate groups based on drug use, again with a basic plot:

```
> my.KMest3 <- survfit(my.surv~drug,data=hmohiv)
> plot(my.KMest3)
```

You can add a title later to the plot:

```
> title(main="TEST KAPLAN-MEIER CURVE", col.main="black", xlab="Time (Months)",
ylab="Overall Survival Proportion",  col.lab="blue", cex.lab=0.9)
```

Or at the same time:

```
> plot(my.KMest3, main="TEST KAPLAN-MEIER CURVE", col.main="black",
xlab="Time (Months)", ylab="Overall Survival Proportion",col.lab="blue",
cex.lab=0.9)
```

Change curve colours, mark="+" = censored, lty provides different line formats:

```
>plot(my.KMest3, main="TEST KAPLAN-MEIER CURVE", col.main="black", xlab="Time
(Months)", ylab="Overall Survival Proportion",col.lab="blue", cex.lab=0.9, mark="+",
col=c(2,4), lty = 2:3)
```

Then add a legend:

```
>legend(50, .9, c("No Drug Use", "Drug Use"), lty = 2:3,col=c(2,4))
```

NOTE: You can use lty =1, in plot and legend for non-broken lines in both curves. You can assume that as ‘no drug use’ =0 and ‘drug use’ =1, then the first colour (red =2) corresponds to ‘no drug use’ and blue (=4) to ‘drug use’.

1.7 Compare Survival Curves

To determine if the survival curves are different, use survdiff():

```
>survdiff(Surv(time, censor) ~ drug, data = hmohiv)
```

This will give you a p-value for the difference using the log-rank test (default). It is possible to determine the hazard ratio using the log-rank approach, but this will be covered in the Cox Proportional Hazards section.

PRACTICE EXAMPLES

Question 1: Plotting/Comparing Survival Curves Part I

Load in the ovarian data-set:

```
>data(ovarian)
>ovarian
```

This data is from a randomised trial comparing two treatments for ovarian cancer. Look in http://cran.r-project.org/web/packages/survival/survival.pdf, the variables are as follows:

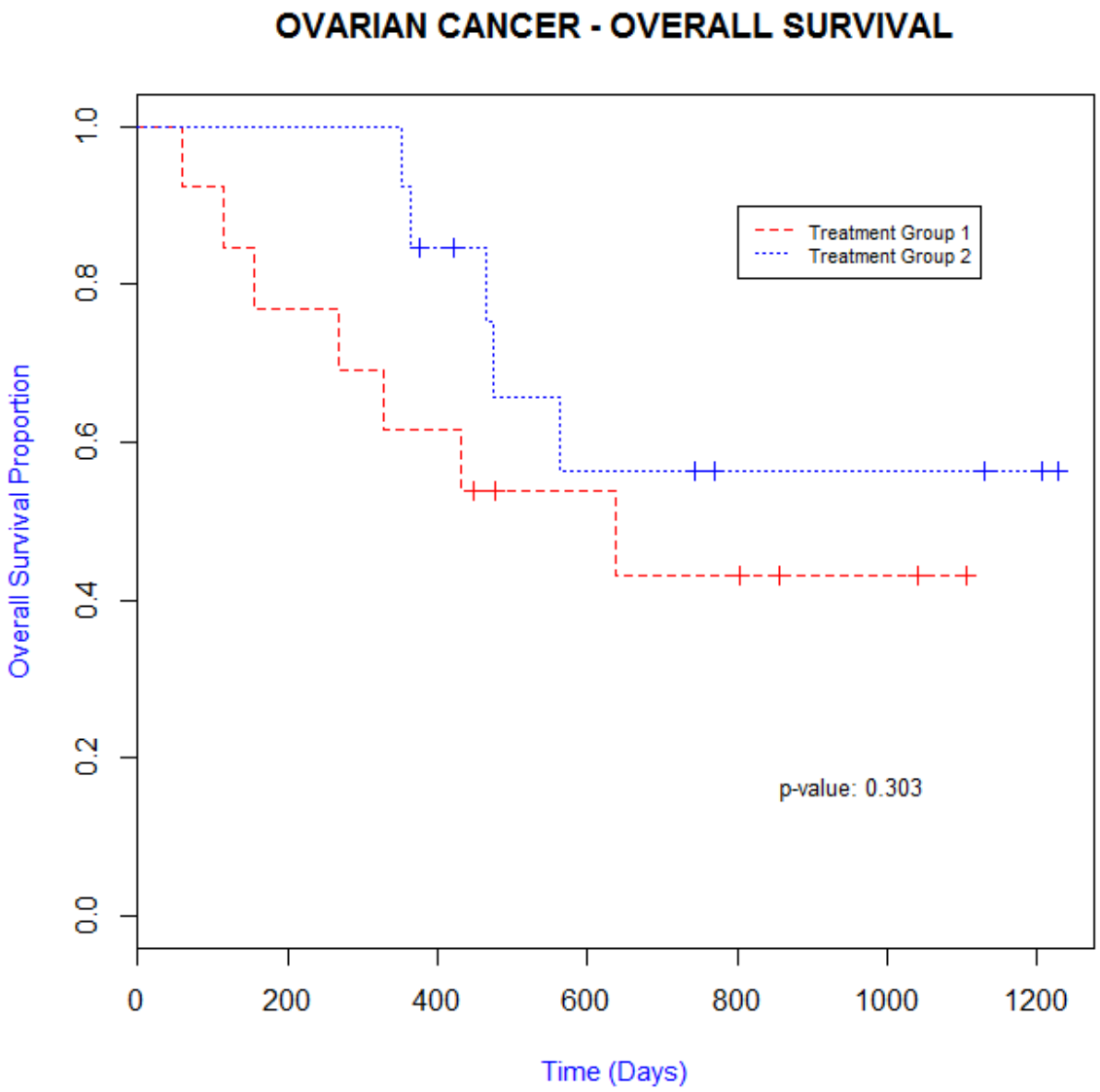
futime: survival or censoring time
fustat: censoring status
age: in years
resid.ds: residual disease present (1=no, 2=yes)
rx: treatment group
ecog.ps: ECOG performance status (1 is better)

Compare the two treatment groups (**treatment 1** versus **treatment 2**). Plot the respective survival curves, indicating censored subjects. You can distinguish between the two groups using different colours or different line formats or both. Label both x- and y-axes. Add a suitable title. Also, add a legend indicating which line corresponds to which line format.

Finally compare the two survival curves (log-rank) and add a p-value to the bottom-right of the plot.

HINT: You add two or more legends to the plot; to add a legend without a border use box.col="white".

You should achieve something like this:


OVARIAN CANCER - OVERALL SURVIVAL
Treatment Group 1
Treatment Group 2
p-value: 0.303
Overall Survival Proportion
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
1200
Time (Days)

Question 2: Plotting/Comparing Survival Curves Part II

Load in the Leukaemia-free survival/transplant data-set:

```
>install.packages("KMsurv")
>library(KMsurv)
>data(alloauto)
>alloauto
```

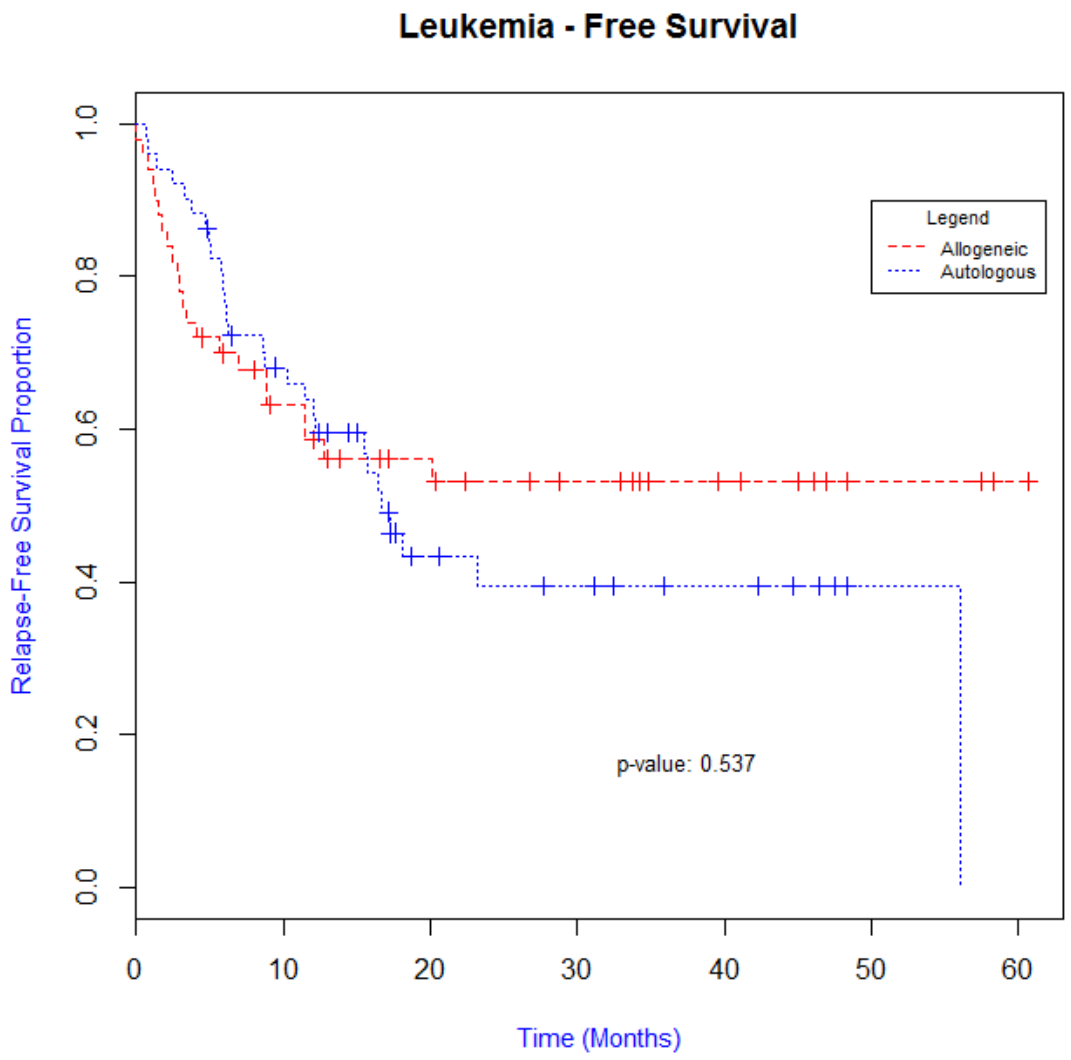
This data considers two transplant types in relation to leukaemia-free survival. Look in http://cran.r-project.org/web/packages/KMsurv/KMsurv.pdf, the variables are as follows:

Time: Time to death or relapse, months
Type: Type of transplant (1=allogeneic, 2=autologous)
Delta: Leukemia-free survival indicator (0=alive without relapse, 1=dead or relapse)

Compare the two transplant types (**allogeneic** vs **autologous**). Plot the respective relapse-free survival curves, indicating censored subjects. You can distinguish between the two groups using different colours or different line formats or both. Label both x- and y-axes. Add a suitable title. Also, add a legend indicating which line corresponds to which line format.

Finally compare the two survival curves (log-rank) and add a p-value to the bottom-right of the plot.

You should achieve something like this:


Leukemia - Free Survival
Relapse-Free Survival Proportion
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Time (Months)
Legend
Allogeneic
Autologous
p-value: 0.537

2 COX PROPORTIONAL HAZARDS

We'll return to the hmohiv data-set and look at it from a Cox Proportional Hazards (PH) perspective. Previously, we only looked at drug treatment, now we can look at the effect of age.

The first thing to do is to check is both variables meet the proportional hazards (PH) assumptions.

Start off by calling the coxph() function. The method parameter can also use the term "breslow".

```
> drug.coxph <- coxph(Surv(time,censor)~drug, method="efron", data=hmohiv)
```

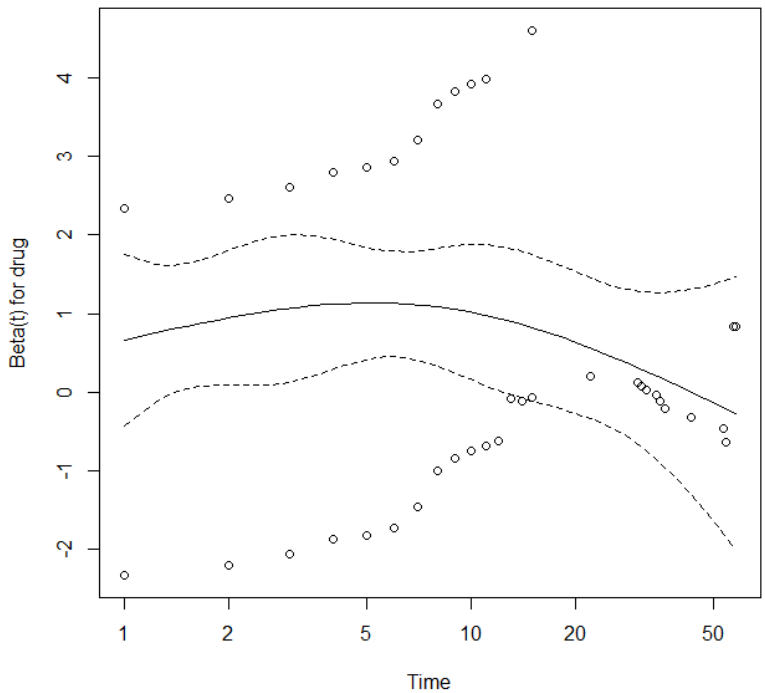
2.1 Testing PH assumption

We'll look at the drug variable first. We use both a plot and p-value to examine whether the variable meets the PH assumption.

```
> drug_ph <-cox.zph(drug.coxph, transform ='log')
> plot(drug_ph[1,])
> drug_ph
```

This is the plot and output that you should get:

	rho	chisq	p
drug	-0.0746	0.417	0.518

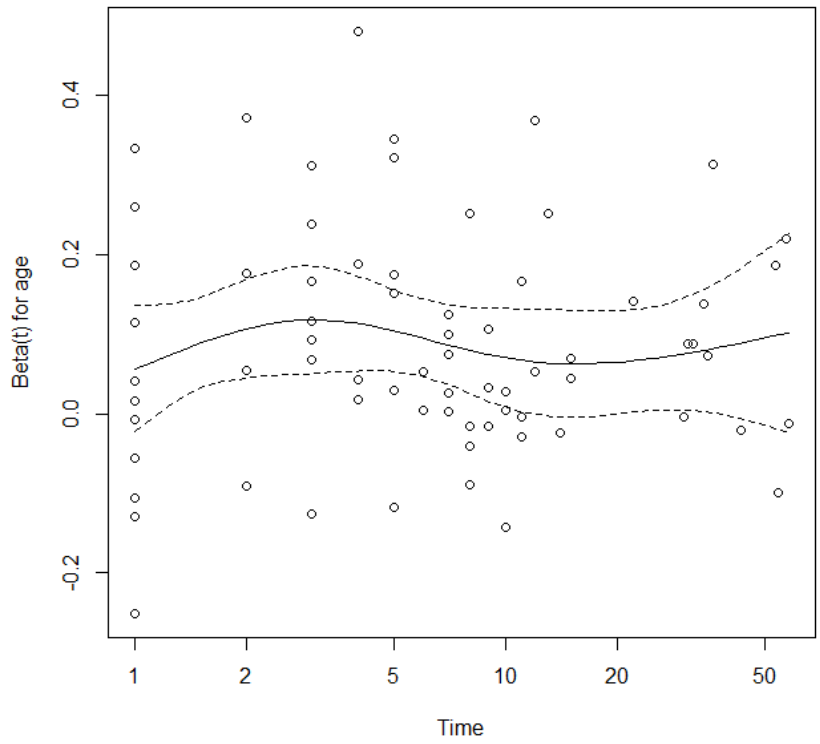


The plot is curved (we are looking for a straight line), especially towards the last third of a graph, though the range is small, and the p-value is not significant. We shall continue on, but sometimes it can be a very tough call as to whether a variable is PH or not.

We'll look at the age variable now:

```
>age.coxph <- coxph(Surv(time,censor)~age, method="efron", data=hmohiv)
>age_ph <-cox.zph(age.coxph, transform ='log')
>plot(age_ph[1,])
>age_ph
```

	rho	chisq	p
age	0.0164	0.0179	0.894



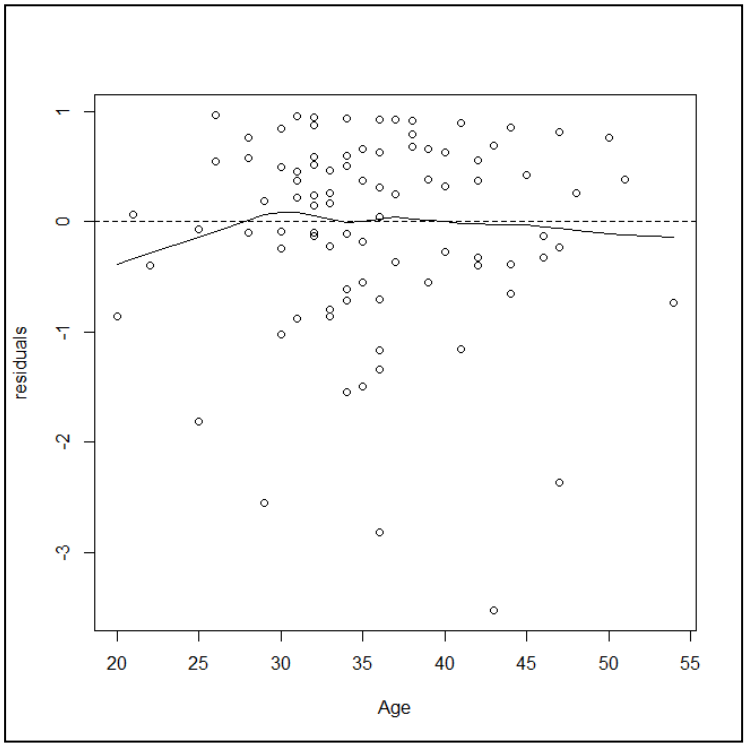
The p-value for age is not significant, so according to the test, age conforms to the PH assumption. Looking at the plot, it is variable about a straight line, so we'll go with the PH assumption.

2.2 Testing Linearity

As age is a continuous variable, we also to check for linearity.

```
>res <- residuals(age.coxph, type='martingale')
>X <- as.matrix(hmohiv[,"age"]) # matrix of covariates
>plot(X[,1], res, xlab=c("Age")[1], ylab="residuals")
>abline(h=0, lty=2) + lines(lowess(X[,1], res, iter=0))
```

This is the output:



At the start of the plot, there appears to some doubt over age's linearity. One option is to transform the variable using a cubic spline or similar, but this beyond the scope of this tutorial. **Alternatively we could split the variable into categories, but there needs to a very good reason for doing so!** For example, if there were key age categories in relation to the subject already defined in the literature.

This can be achieved by:

```
>install.packages("car")
>library(car)
>agecat<-recode(age, "20:29='A'; 30:34='B'; 35:39='C';40:54='D'", as.factor=T)
>agecat.coxph <- coxph( Surv(time, censor)~agecat, method="efron")
>summary(agecat.ph)
```

You can check the age range using:

```
>range(age)
```

You could also carry out the coding in reverse:

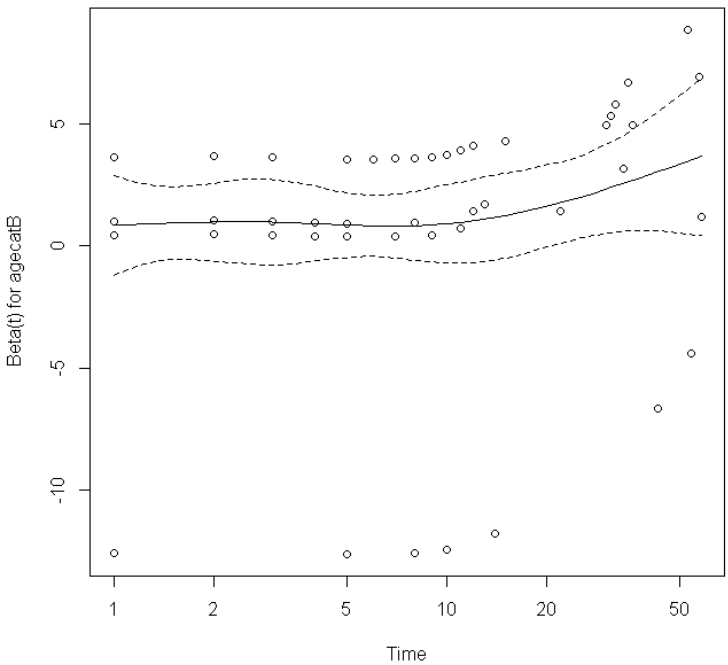
```
>agecat<-recode(age, "20:29='D'; 30:34='B'; 35:39='C';40:54='A'", as.factor=T)
```

Double-check that the age categorical variable is still PH.

```
>agecat_ph <-cox.zph(agecat.coxph, transform ='log')
>plot(agecat_ph[1,])
>plot(agecat_ph[2,])
>plot(agecat_ph[3,])
>agecat_ph
```

The new age category is still PH (first plot only shown)!

	rho	chisq	p
agecatB	0.12269	1.136567	0.286
agecatC	0.00131	0.000138	0.991
agecatD	-0.05443	0.237351	0.626
GLOBAL	NA	2.220486	0.528



2.3 Univariate Analysis

Let's look at the results of the individual variables, drug and age (categorical):

```
>drug.coxph
>agecat.coxph
```

```
> drug.coxph

Call:

coxph(formula = Surv(time, censor) ~ drug, data = hmohiv, method = "efron")

              coef    exp(coef)    se(coef)    z      p

drug          0.831   2.3          0.242       3.44   0.00059

Likelihood ratio test=11.6  on 1 df, p=0.000659  n= 100, number of events= 80

> agecat.coxph

Call:

coxph(formula = Surv(time, censor) ~ agecat, method = "efron")

              coef    exp(coef)    se(coef)    z      p

 agecatB      1.20    3.33         0.450       2.67   7.5e-03
agecatC       1.33    3.80         0.458       2.91   3.6e-03
agecatD        1.91   6.78         0.468       4.09   4.3e-05
```

The age categories have a reference category A; categories B, C and D have all been compared to this category. It is not clear which direction the drug category is being compared. Usually, if numeric, the lowest category, in this case drug =0, ie 'no drug use' is used as the reference. You can specify within the model that a variable is a factor. Although not required as drug is an either/or variable, you can do it like this:

```
>drug.coxph <- coxph(Surv(time,censor)~factor(drug), method="efron",
data=hmohiv)
```

Calling the summary function for both of the models will be more informative:

```
>summary(drug.coxph)
>summary(agecat.coxph)
```

```
>summary(drug.coxph)
Call:

coxph(formula = Surv(time, censor) ~ factor(drug), data = hmohiv, method =
"efron")

 n= 100, number of events= 80

                  coef     exp(coef)     se(coef)     z      Pr(>|z|)

factor(drug)1     0.8309   2.2953        0.2418       3.436  0.00059 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef)     exp(-coef)     lower .95     upper .95
factor(drug)1     2.295         0.4357         1.429         3.687

Concordance= 0.611  (se = 0.036 )
Rsquare= 0.11   (max possible= 0.997 )
Likelihood ratio test= 11.6  on 1 df,   p=0.0006593
Wald test          = 11.81  on 1 df,   p=0.0005903
Score (logrank) test = 12.33  on 1 df,   p=0.0004464

>summary(age.coxph)
Call:

coxph(formula = Surv(time, censor) ~ agecat, method = "efron")

n= 100, number of events= 80

                  coef     exp(coef)     se(coef)     z      Pr(>|z|)
agecatB           1.2030   3.3301        0.4503       2.672  0.00755 **
agecatC           1.3337   3.7951        0.4580       2.912  0.00359 **
agecatD           1.9144   6.7831        0.4679       4.091  4.29e-05 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef)     exp(-coef)     lower .95     upper .95

agecatB           3.330         0.3003         1.378         8.049
agecatC           3.795         0.2635         1.547         9.313
agecatD           6.783         0.1474         2.711         16.971

Concordance= 0.642  (se = 0.04 )
Rsquare= 0.189   (max possible= 0.997 )
Likelihood ratio test= 20.92  on 3 df,   p=0.0001091
Wald test          = 17.85  on 3 df,   p=0.0004724
Score (logrank) test = 19.83  on 3 df,   p=0.0001843
```

Looking at each variable individually is termed univariate analysis. **Are both variables significant? Would do the concordance index values suggest? Why are the the confidence intervals for the age categories so wide?**

2.3 Multivariate Analysis

What happens if both variables are considered together (multivariate)? Will they remain significant?

```
>agecat_drug.coxph <- coxph( Surv(time, censor)~agecat+drug, method="efron")
```

```
> summary(agecat.coxph)

Call:

coxph(formula = Surv(time, censor) ~ agecat + drug, method = "efron")

 n= 100, number of events= 80

              coef     exp(coef)    se(coef)     z       Pr(>|z|)
agecatB       1.2569   3.5146       0.4596       2.735   0.00624 **
agecatC       1.3194   3.7411       0.4671       2.825   0.00473 **
agecatD       2.0152   7.5019       0.4804       4.195   2.73e-05 ***
drug          0.8926   2.4415       0.2530       3.527   0.00042 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef)    exp(-coef)    lower .95    upper .95
agecatB       3.515        0.2845        1.428        8.651
agecatC       3.741        0.2673        1.498        9.345
agecatD       7.502        0.1333        2.926        19.234
drug          2.442        0.4096        1.487        4.009

Concordance= 0.681  (se = 0.042 )
Rsquare= 0.284   (max possible= 0.997 )
Likelihood ratio test= 33.35  on 4 df,   p=1.015e-06
Wald test            = 28.7  on 4 df,   p=9.012e-06
Score (logrank) test = 31.57  on 4 df,   p=2.339e-06
```

Does the concordance index improve with both variables in the model?

You could then present the results something like this:

			UNIVARIATE			MULTIVARIATE		
		N (n)	**HR**	**%95 CI**	**p-value**	**HR**	**%95 CI**	**p-value**
Age (yrs)	*20-29*	12 (8)	1			1		
	30-34	34 (29)	3.330	1.378-8.049	0.0076	3.515	1.428-8.651	0.0062
	35-39	25 (20)	3.795	1.547-9.313	0.0036	3.741	1.498-9.345	0.0047
	40-54	29 (23)	6.783	2.711-16.971	<0.0001	7.502	2.926-19.234	<0.0001
Drug Use	*No*	51 (42)	1			1		
	Yes	49 (38)	2.295	1.429-3.687	0.0006	2.442	1.487-4.009	0.0004

PRACTICE EXAMPLES

Question 3: Developing a Cox Model

Look at the hmoiv data-set again.

Recode the age variables (see below) and repeat the univariate and multivariate analyses.

```
>agecat<-recode(age, "20:29='D'; 30:34='B'; 35:39='C';40:54='A'", as.factor=T)
```

Does using a larger reference group for the age help in any way? What do you notice about the hazard ratios and confidence intervals? Are the p-values or the concordance index affected?

Question 4: Variable Selection

For this example, we'll look at data involving drug treatment programs. The UIS data-set compares the time for return to drug use for patients enrolled in two different residential treatment programs that differed in length (**treat**=0 is the short program and **treat**=1 is the long program). The patients were randomly assigned to two different sites (**site**=0 is site A and **site**=1 is site B). The variable **age** indicates age at enrolment, **hercoc** indicates heroin or cocaine use in the past three months (hercoc=1 indicates heroin and cocaine use, hercoc=2 indicates either heroin or cocaine use and hercoc=3 indicates neither heroin nor cocaine use) and **ndrugtx** indicates the number of previous drug treatments. The variable **time** is the time until return to drug use and the **event** variable indicates whether the subject returned to drug use (event=1 indicates return to drug use and event=0 otherwise). **NOTE:** you can assume PH for all variables and linearity for age and ndrugtx (continuous). **This is not strictly true**, but this example problem is directed more towards gaining experience in variable selection.

First, read in the table:

```
>uis<-read.table("http://www.ats.ucla.edu/stat/R/examples/asa/uis.csv", sep=",",
header = TRUE)
>attach(uis)
>head(uis)
```

For simplicity, we'll only work with the variables mentioned above:

```
>uis_small<-uis[,c(1,2,4,6,8,9,11,12)]
```

For further simplicity, we'll remove the patients with missing values:

```
>tiny_uis<-uis_small[apply(uis_small,1,function(x)!any(is.na(x))),]
```

We'll assume that the two different sites are different centres, so we'll want to stratify by this variable. You can do this via:

```
>age.coxph <- coxph(Surv(time,censor)~age+strata(site), method="efron",
data=tiny_uis)
```

Now, repeat the univariate analyses for the variables: treat, age, ndrugtx and hercoc. Make a note of the concordance index for each. Create a multivariate model using all four variables. What is the concordance index?

Now run the following (AIC) to determine what variables should be retained:

```
>install.packages("MASS")
>library(MASS)
>stepAIC(fit_four_cox_model)
```

What variables are retained? What is the concordance index of the final model?

SUGGESTED SOLUTIONS

PRACTICE EXAMPLES

QUESTION 1

```
>data(ovarian)
>ovarian

>my.KMest4 <- survfit(Surv(futime, fustat)~rx,data=ovarian)

>plot(my.KMest4, main="OVARIAN CANCER - OVERALL SURVIVAL", col.main="black", xlab="Time
(Days)", ylab="Overall Survival Proportion",col.lab="blue",
cex.lab=0.9,col=c("red","blue"),mark.time=TRUE, lty = 2:3)

>legend(800, .9, title="Legend",c("Treatment Group 1", "Treatment Group 2"), lty =
2:3,col=c("red","blue"),cex=0.7)

>survdiff(Surv(futime, fustat) ~ rx, data = ovarian)

>legend(800, .2, c("p-value: 0.303"), cex=0.8,box.col="white")
```

QUESTION 2

```
>install.packages("KMsurv")
>library(KMsurv)
>library(help=KMsurv)
>data(alloauto)
>alloauto

>my.KMest4 <- survfit(Surv(time, delta)~type,data=alloauto,conf.type="none")


>plot(my.KMest4, main="Leukemia - Free Survival", col.main="black", xlab="Time (Months)",
ylab="Relapse-Free Survival Proportion",col.lab="blue",
cex.lab=0.9,col=c("red","blue"),mark.time=TRUE, lty = 2:3)


>survdiff(Surv(time, delta) ~ type, data = alloauto)
#1=allogeneic, 2=autologous
#p= 0.537


>legend(50, .9, title="Legend",c("Allogeneic", "Autologous"), lty = 2:3,col=c("red","blue"),cex=0.7)


>legend(30, .2, c("p-value: 0.537"), cex=0.8,box.col="white")
```

QUESTION 3

Univariate:

Drug variable: as before

coxph(formula = Surv(time, censor) ~ agecat, method = "efron")

n= 100, number of events= 80

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
agecatB	-0.5807	0.5595	0.3122	-1.860	0.0629 .
agecatC	-0.7114	0.4909	0.2870	-2.479	0.0132 *
agecatD	-1.9144	0.1474	0.4679	-4.091	4.29e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
agecatB	0.5595	1.787	0.30344	1.0316
agecatC	0.4909	2.037	0.27971	0.8617
agecatD	0.1474	6.783	0.05892	0.3689

Concordance= 0.642 (se = 0.04)
Rsquare= 0.189 (max possible= 0.997)
Likelihood ratio test= 20.92 on 3 df, p=0.0001091
Wald test = 17.85 on 3 df, p=0.0004724
Score (logrank) test = 19.83 on 3 df, p=0.0001843

Multivariate:

coxph(formula = Surv(time, censor) ~ agecat + drug, method = "efron")

n= 100, number of events= 80

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
agecatB	-0.6958	0.4987	0.3151	-2.208	0.02722 *
agecatC	-0.7582	0.4685	0.2899	-2.615	0.00891 **
agecatD	-2.0152	0.1333	0.4804	-4.195	2.73e-05 ***
drug	0.8926	2.4415	0.2530	3.527	0.00042 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
agecatB	0.4987	2.0053	0.26892	0.9247
agecatC	0.4685	2.1345	0.26541	0.8269
agecatD	0.1333	7.5019	0.05199	0.3418
drug	2.4415	0.4096	1.48685	4.0092

Concordance= 0.681 (se = 0.042)
Rsquare= 0.284 (max possible= 0.997)
Likelihood ratio test= 33.35 on 4 df, p=1.015e-06
Wald test = 28.7 on 4 df, p=9.012e-06
Score (logrank) test = 31.57 on 4 df, p=2.339e-06

QUESTION 4

Hercoc: Concordance= 0.536

Ndrugtx: Concordance= 0.549

Age: Concordance= 0.532

Treat: Concordance= 0.543

Hercoc,Ndugtx,Age,Treat: Concordance=0.585

AIC selects: Ndugtx,Age,Treat

Ndugtx,Age,Treat :Concordance= 0.587