Singapore Med J 2004 Vol 45(6) : 249

Biostatistics 203. Survival analysis

Y H Chan

Table I. Summary of the common univariate/multivariate biostatistical techniques to analyse quantitative and qualitative data types.

Quantitative data[1]		Qualitative data[2]	
Normality/homogeneity of variance assumptions satisfied?		Independent sample	Matched case-control
YES Parametric tests	**NO** Non-parametric tests		
1 Sample T Paired T	Sign test Wilcoxon Signed Rank	Chi Square/ Fisher Exact	McNemar test
2 Sample T	Wilcoxon Rank Sum/ Mann Whitney U		
ANOVA	Kruskal Wallis		
Multivariate tests			
Multiple linear regression[3]		Logistic regression[4]	Conditional logistic regression

In this article, we shall discuss the use of survival analysis on a quantitative type of data corresponding to the time from a well-defined time origin until the occurrence of some particular event of interest or end-point.

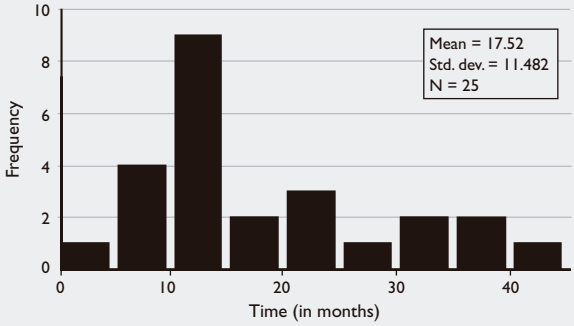
Medical examples are:
- Duration – time from randomisation to relapse
- Pressure sore – time to development
- Survival – time from randomisation until death

Non-medical examples are:
- Banking – time from making a loan to full-repayment
- Economy – time from graduation to get 1st job
- Social – time from being single to getting married

Since survival time is a quantitative variable, why can't we just use the usual techniques from Table I? Before we explain the main reason why we use survival analysis, let' us consider a simple example on the survival times (in months) for 25 lung cancer patients who all died; the timings are : 1, 5, 6, 6, 9, 10, 10, 10, 12, 12, 12, 12, 12, 13, 15, 16, 20, 24, 24, 27, 32, 34, 36, 36, 44 months.

Performing a simple descriptive, we have n = 25, mean (sd) = 17.52 (11.48) months and median = 12 months.

Fig. 1 The distribution of the survival times.



It is obvious that the distribution is not normal (Fig. 1) as expected from survival-time data.

Kaplan Meier is the usual technique performed to analyse survival-time data. Table II shows the Kaplan Meier analysis for the above 25 subjects (all died of lung cancer):

Table II. Kaplan Meier analysis (no censoring).

Kaplan Meier technique (All subjects died)			
	Survival time	Standard error	95% CI
Mean	17.52	2.30	13.02, 22.02
Median	12.00	1.25	9.55, 14.45

What do we observe? The Kaplan Meier results of Table II is exactly the same to that of the descriptive results above. So why do we need to do a survival analysis? To quote a Chinese saying, we have used "a bull knife to kill a chicken": an "overkill in analysis"! The reason here is: since all the subjects died (presumably of lung cancer), we have no extra information to require us to perform a survival analysis – **no censored data**.

Clinical Trials and Epidemiology Research Unit
226 Outram Road
Blk B #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

Correspondence to:
Dr Y H Chan
Tel: (65) 6325 7070
Fax: (65) 6324 2700
Email: chanyh@cteru.com.sg

What are censored observations? Censored observations arise in cases for which

- the critical event has not yet occurred
- lost to follow-up
- other interventions offered
- event occurred but unrelated cause

Let us consider the situation where we have more information (censored cases) for our 25 lung cancer patients : 1#, 5#, 6, 6, 9#, 10, 10, 10#, 12, 12, 12, 12, 12#, 13#, 15#, 16#, 20#, 24, 24#, 27#, 32, 34#, 36#, 36#, 44# months (where # denotes censored observations).

The subject with 44# definitely is a surviving person at the point of analysis (we cannot "ask" the patient to die – not ethical!). The 1# could be one who just enrolled into the study recently and still surviving. Perhaps, the 5# could be one who (after five months) decided to seek other help and did not return to the study; his survival status is unknown. Lastly, the 13# could be one who died but not because of lung cancer. In all, 10 of the 25 subjects died from lung cancer.

How do we present this data in SPSS? Table III shows the 1st six cases, as an example.

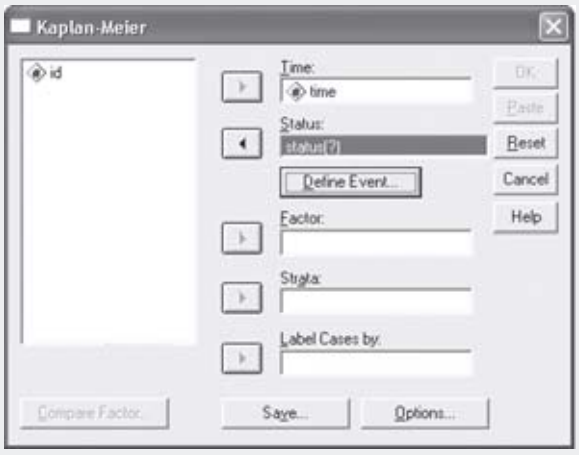
Table III. Survival analysis dataset in SPSS.

Subject number	Survival time	Status
1	1	0
2	5	0
3	6	1
4	6	1
5	9	0
6	10	1
	etc	

The last variable "Status" tells SPSS which case is censored (denoted by 0) and which case is an event (dying of lung-cancer, denoted by 1).

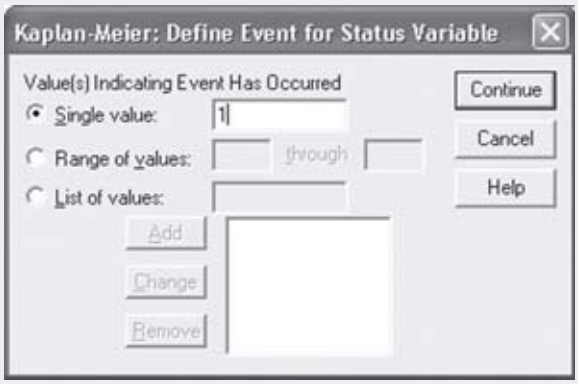
To perform a Kaplan Meier analysis in SPSS, go to Analyze, Survival, Kaplan Meier to get Template I.

Template I. Kaplan Meier analysis.



Put the variables "time" and "status" at their appropriate options, click on 'Define Event' button to get Template II.

Template II. Defining the event.



Put a 1 as an event as defined accordingly. Click "Continue". In Template I, click on the "Options" folder and checked the boxes as shown in Template III.

Template III. Kaplan Meier options.

Ticking on the "Mean and median survival" option gives Table IV.

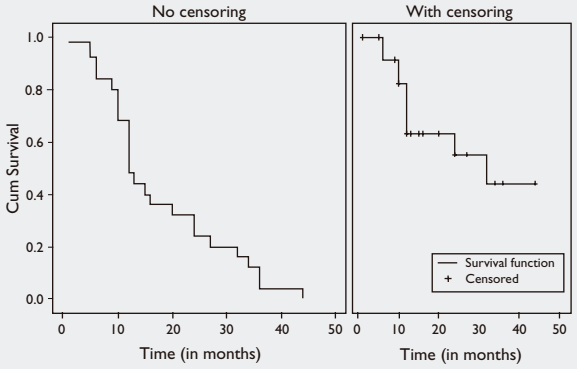
Table IV. Kaplan Meier analysis (with censoring).

	Kaplan Meier technique		
	Survival time	Standard error	95% CI
Mean	28.51	3.54	21.58, 35.44
Median	32.00	14.43	3.71, 60.29

Table IV shows the Kaplan Meier analysis with censored data information taken into account. We observe that the median survival time has increased from 12 months (without censoring) to 32 months.

This means that with the factoring in of the "extra" information, we are being "realistic" about the survival time of, in this case, lung cancer or being "fair" to the treatment under study with the intent of extending the survival time of these subjects. Fig. 2 shows the survival plots for both censored and no-censored scenarios.

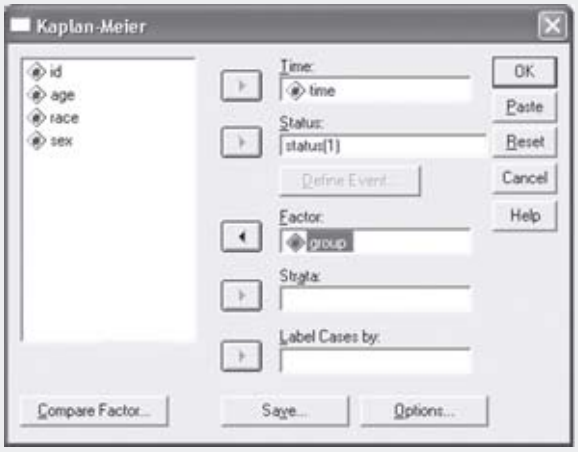
Fig. 2 Survival plots – lung cancer example.



COMPARING TWO SURVIVAL CURVES

Kaplan Meier can be used to compare two treatment groups on their survival times. Put the variable "group" in the "Factor" option, see Template IV.

Template IV. Defining the factor for comparison.



Click on "Compare Factor" on the left-hand corner of Template IV to invoke the log-rank test to compare the two groups (Template V).

Template V. The log-rank test

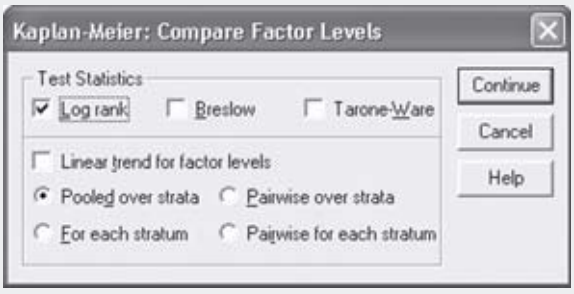


Table V shows the mean/median survival times for the control and active groups with log-rank test p = 0.1835 – no differences between the active and control on having a shorter time to event, with the survival plot given in Fig. 3. One common misconception of survival analysis is that some researchers interpret the result as one group being more likely to have deaths (this should be given by logistic regression!). It is the time to event which is the primary response here.

Table V. Kaplan Meier analysis for comparison between two groups.

Survival analysis for time

Factor group = control

	Survival time	Standard error	95% confidence interval
Mean (Limited to 36)	21	5	(12, 30)
Median	12	2	(7, 17)

Factor group = active

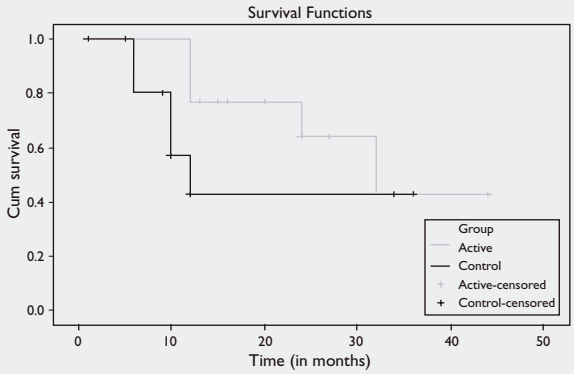
	Survival time	Standard error	95% confidence interval
Mean (Limited to 44)	31	4	(23, 39)
Median	32	8	(17, 47)

	Total	Number of events	Number censored	Percent censored
Group control	12	5	7	58.33
Group active	13	5	8	61.54
Overall	25	10	15	60.00

Test statistics for equality of survival distributions for group

	Statistic	df	Significance
Log rank	1.77	1	.1835

Fig. 3 Survival plot for comparison of two groups.

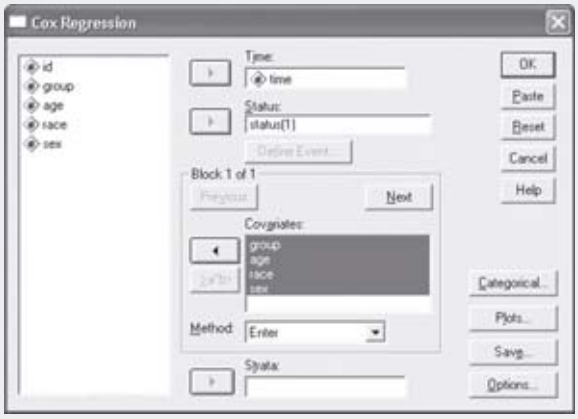


The Kaplan Meier technique is the univariate version of survival analysis. To take into account confounders into the analysis, we have to use cox regression.

COX REGRESSION

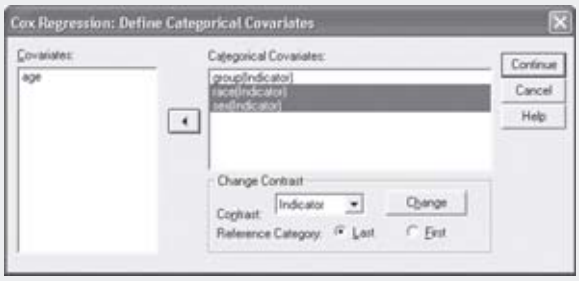
For the above lung cancer example, we have collected information on race, age and gender, and want to look at a confounder model to determine whether the two groups differ after adjusting for demographics. To perform a cox regression, go to Analyse, Survival, Cox regression to get Template VI.

Template VI. Cox regression: lung cancer example.



The declaration for the categorical variables is similar to that discussed in the logistic regression article[4] by clicking on the "Categorical" folder and put group, race and sex as the categorical covariates (Template VII)

Template VII. Declaration of categorical variables.

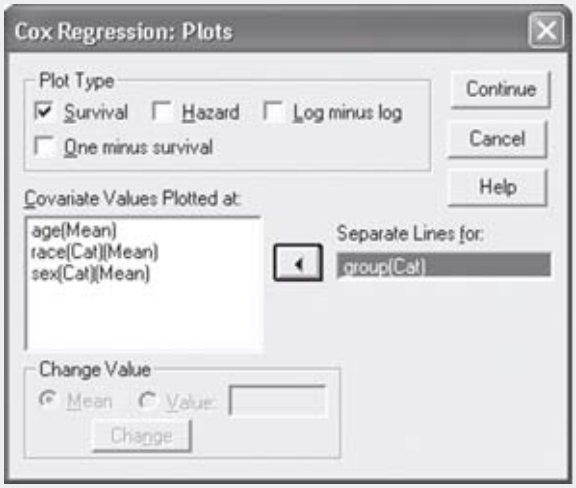


In Template VI, click on "Options" to invoke the 95% CI for the hazard ratio (HR), given by the expression exp(B) – which is also the same expression for odds ratios in logistic regression. This is another common mistake – researchers at times refer to odds ratio in survival analysis (mistaken by the same symbol). The interpretation for the hazard ratio is similar to that of the odds ratio. A value of one means there is no differences between two groups in having a "shorter time to event". A HR >1 means that the group of interest comparing to the reference group (to be observed from the categorical declaration) likely have a shorter time to event. A HR <1 means that the group of interest less likely to have a shorter time to event.

Template VIII. Invoking the 95% CI for the hazard ratio.

From Template VI, ask for plots to get Template IX – click on "Survival" and Separate Lines for "group".

Template IX. Survival plot for Cox regression.



The following Tables VIa – e show the results for the Cox regression.

Table VIa. Categorical definition.

		Categorical variable codings			
		Frequency	(1)	(2)	(3)
Group	1.00=control	12	1		
	2.00=active	13	0		
Race	1=chinese	15	1	0	0
	2=indian	5	0	1	0
	3=malay	2	0	0	1
	4=other	3	0	0	0
Sex	1=male	17	1		
	2=female	8	0		

The reference category for group is active, race is "other race" and sex is female.

Table VIb gives the p-values (Sig) and the hazard ratios (Exp(B)) of the variables. Firstly, we have to check for multicolinearity by observing whether the SE of all the variables are small (see logistic regression[4] for a detailed discussion on this checking).

Table VIb. Estimates of variables in Cox regression.

Variables in the equation								
							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Group	1.841	.911	4.086	1	.043	6.302	1.058	37.550
Sex	3.670	1.435	6.542	1	.011	39.263	2.358	653.769
Age	.115	.043	7.137	1	.008	1.122	1.031	1.220
Race			2.066	3	.559			
Race(1)	-.307	1.181	.068	1	.795	.735	.073	7.448
Race(2)	.983	1.299	.573	1	.449	2.672	.210	34.060
Race(3)	.907	1.469	.381	1	.537	2.476	.139	44.085

Since this is an adjusting for confounder model, our interest is only in the variable group. 'Thankfully' the p-value is 0.043 (statistically significant!) compared to the Kaplan Meier analysis (well, we do not always get this happy ending). The HR is 6.302 (95% CI 1.058 - 37.55), comparing the control with the active (obtained from the categorical definition table IVa), the control likely to have a shorter time to event and in this example, the event is death.

What is going on here? Why now a statistical difference? Table VIb also showed that there are statistical differences for gender and also age – the men and older people were doing worst. Performing a cross-tabulation shows that there are more men and less women in the control group (p = 0.673) and mean age is higher in the active group. See Tables VIc and VId.

Table VIc. Cross-tabulation between group and gender.

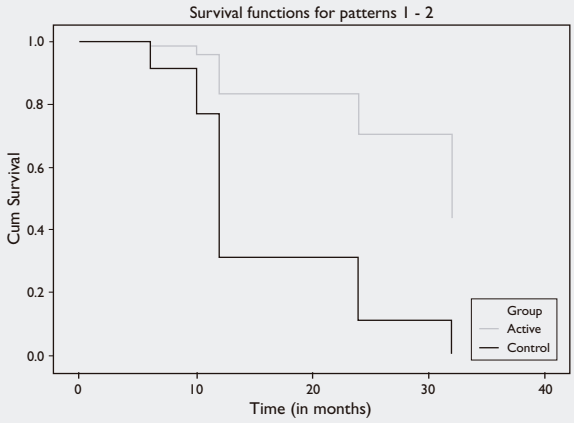
The sex of the patient * group cross-tabulation					
			Group		
			Control	Active	Total
Sex of patient	Male	Count	9	8	17
		% within group	75.0%	61.5%	68.0%
	Female	Count	3	5	8
		% within group	25.0%	38.5%	32.0%
Total		Count	12	13	25
		% within group	100.0%	100.0%	100.0%

Table VId. Age differences between group (p=0.737).

Group statistics					
	Group	N	Mean	Std. deviation	Std. error mean
Age	active	13	31.6923	16.16263	4.48271
	control	12	29.5833	14.73683	4.25416

Thus taking into account these information, a treatment difference is found, as observed from the survival plot in Fig. 4.

Fig. 4 Survival plot for the lung cancer example.



The above exercise showed that it is not relevant to stop at the univariate analysis but to always perform a multivariate analysis to present the realistic situation!

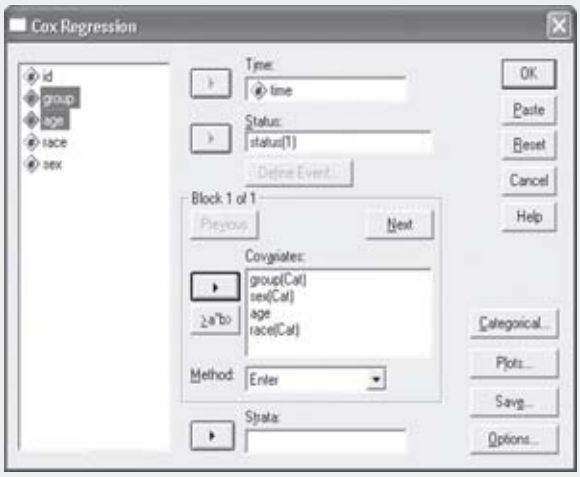
Since we found a difference between treatment groups, do you want to stop here? How about interaction between gender and group, or age and group? Question of interest would be: is there a particular group (female on active, for example) performing better? Note that we will start to ask these questions only when the "main effects" model showed significant differences in the variables of interest.

How to put in the interaction term? In Template VI, highlight group 1st, hold the ctrl key and highlight age – observe the button >a*b> becomes "visible" – click on this button – see Template X.

Table VIe. Result with interaction terms.

Variables in the equation								
							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Group	-5.524	4.891	1.276	1	.259	.004	.000	58.121
Sex	1.687	1.716	.966	1	.326	5.401	.187	156.115
Age	.082	.055	2.186	1	.139	1.085	.974	1.200
Race			3.171	3	.366			
Race(1)	-.869	1.341	.420	1	.517	.419	.303	5.804
Race(2)	1.112	1.261	.777	1	.378	3.041	.257	36.039
Race(3)	1.018	1.570	.421	1	.517	2.769	.128	60.107
Age*group	.121	.089	1.823	1	.177	1.128	.947	1.344
Group*sex	5.584	3.261	2.933	1	.087	266.224	.447	158709.101

Template X. Preparing to put an interaction term group*age.



Click on >a*b> button to activate age*group(Cat) – see Template XI. Likewise do the same for gender*group.

Template XI. Activating an interaction term.

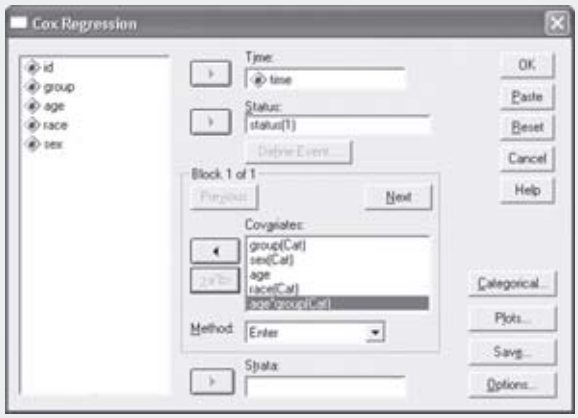


Table VIe shows that none of the interaction terms are significant. This implies that regardless of age or gender, the active group is performing better (from Table VIb).

Let us discuss another example on the use of interaction term – using the breast cancer survival dataset from SPSS. Variables collected were age and the categorical histology grade, oestrogen receptor status, progesterone receptor status, pathological tumour size and lymph node status. The interest is to determine the predictors for a shorter survival time to death.

Table VIIa. Categorical definition – breast cancer example.

Categorical variable codings				
		Frequency	(1)	(2)
histgrad	1=1	56	0	0
	2=2	352	1	0
	3=3	252	0	1
cr	0=negative	262	0	
	1=positive	398	1	
pr	0=negative	299	0	
	1=positive	361	1	
pathscat	1=<=2cm	457	0	0
	2=2-5cm	196	1	0
	3=>5cm	7	0	1
ln_yesno	0=no	485	0	
	1=yes	175	1	

Reference group for histology grade is grade 1, for er, pr and lymph node is negative and tumour size is ≤2cm.

Table VIIb. Main effects model – breast cancer example.

Variables in the equation								
							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Age	-.021	.014	2.200	1	.138	.980	.953	1.007
histgrad			.872	2	.647			
histgrad(1)	.778	1.036	.564	1	.453	2.177	.286	16.587
histgrad(2)	.942	1.056	.796	1	.972	2.564	.324	20.300
cr	-.022	.432	.003	1	.959	.978	.419	2.281
pr	-.455	.422	1.159	1	.282	.635	.277	1.452
pathscat			6.005	2	.050			
pathscat(1)	.638	.336	3.614	1	.057	1.893	.980	3.657
pathscat(2)	1.484	.776	3.658	1	.056	4.412	.964	20.200
ln_yesno	.724	.337	4.605	1	.032	2.063	1.065	3.997

Table VIIc. Interaction terms – breast cancer example.

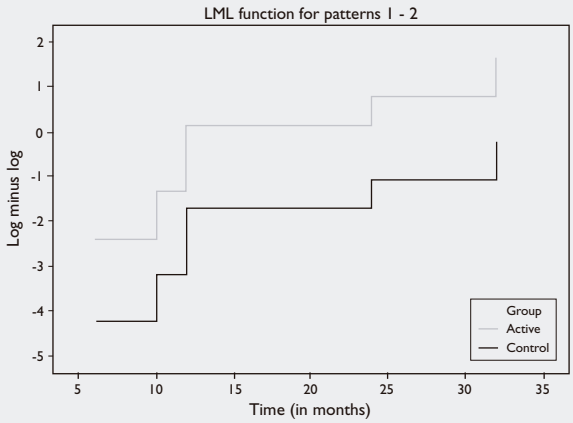
Variables in the equation								
							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Age	-.023	.014	2.845	1	.092	.977	.951	1.004
histgrad			1.165	2	.559			
histgrad(1)	1.047	1.067	.962	1	.327	2.848	.352	23.068
histgrad(2)	1.161	1.081	1.153	1	.283	3.192	.384	26.563
cr	-.063	.424	.022	1	.881	.939	.409	2.156
pr	-.516	.413	1.556	1	.212	.597	.266	1.342
pathscat			8.520	2	.014			
pathscat(1)	-.179	.501	.128	1	.721	.836	.313	2.233
pathscat(2)	3.100	1.102	7.904	1	.005	22.189	2.557	192.566
ln_yesno	.006	.505	.000	1	.990	1.006	.374	2.706
ln_yesno*pathscat			8.564	2	.014			
ln_yesno*pathscat(1)	1.670	.707	5.574	1	.018	5.312	1.328	21.248
ln_yesno*pathscat(2)	-1.847	1.547	1.425	1	.233	.158	.008	3.274

Those with a positive lymph node more likely to have a shorter time to death (HR = 2.06, 95% CI 1.07 - 4.0, p = 0.032). Tumour size is "just off statistical significance". Should we conclude that only women with a positive lymph node are at a higher risk? *Chotto matte (wait a minute)* – what happens if we include a lymph node * tumor size interaction (see Table VIIc).

Here we can see that lymph node status is no more statistically significant but tumour size and their interaction are! The results are telling us that regardless of the lymph node status, subjects with tumour size >5cm are at risk (HR=22.19, 95% CI 2.56 - 192.57, p=0.005) and for subjects with tumour size 2 - 5cm, they are at a higher risk if they have a positive lymph node (HR=5.31, 95% CI 1.33 - 21.25, p=0.018).

One last assumption to check: **proportional hazard model**. From the lung cancer example, in Template IX, click on the "log-minus-log" plot option to get Fig. 5, we do not want the lines to cross each other. When the proportional hazard assumption is not satisfied, we will have to use Cox regression with time-dependent covariate to analyse the data.

Fig. 5 Log-minus-log plot for proportional hazard checking.



Our next article will be "Biostatistics 301. Repeated measurement analysis".

REFERENCES

1. Chan YH. Biostatistics 102. Quantitative data – parametric and non-parametric tests. Singapore Med J 2003; 44:391-6.
2. Chan YH. Biostatistics 103: Qualitative data – tests of independence. Singapore Med J 2003; 44:498-503.
3. Chan YH. Biostatistics 201. Linear regression analysis. Singapore Med J 2004; 45:55-61.
4. Chan YH. Biostatistics 202. Logistic regression analysis. Singapore Med J 2004; 45:149-53.