Multivariable survival analysis

S9

Michael Hauptmann
Netherlands Cancer Institute
Amsterdam, The Netherlands
m.hauptmann@nki.nl

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

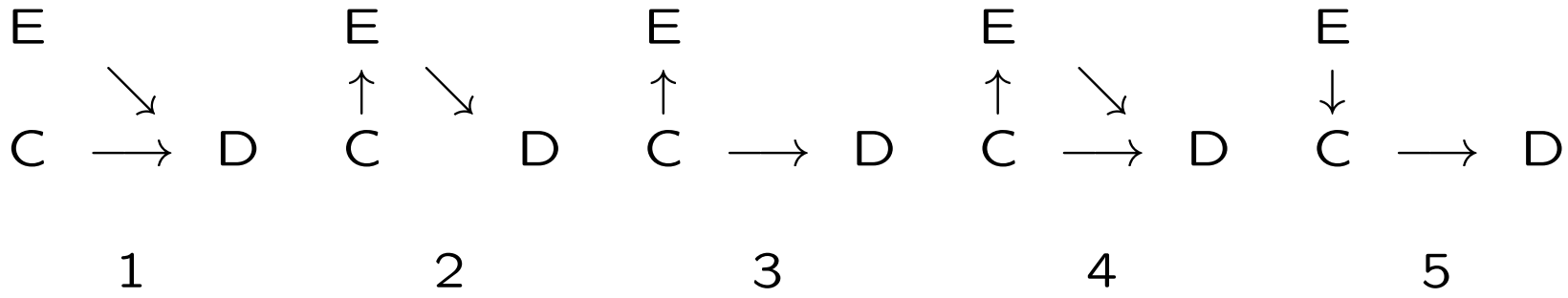
Confounding

A potential confounder is

- Correlated with the variable of interest (e.g., treatment or exposure)
- Correlated with the outcome
- Not in the causal pathway between the variable of interest and the outcome

Confounding: Causal diagrams

E=exposure, D=disease, C=potential confounder



1. C has independent effect on D (C is not a confounder)
2. Effect of C on D is completely contained in E (C is not a confounder)
3. Apparent association between E and D is completely explained by C (C is a confounder)
4. Association between E and D partly due to C (C is a confounder)
5. C is in the causal pathway between E and D (C is not a confounder)

How to prevent/control confounding

- Prevention by design
  - Restriction to one stratum (do study among smokers only if important variables are correlated with ever/never smoking and outcome, limits generalizability)
  - Matching

- Control by analysis
  - Collect data on potential confounders
  - Stratified analysis
  - Multivariable analysis

Identification of confounders

- Based on mechanistic understanding

- Based on statistical significance of association with disease

  - Confounder has to be correlated with outcome, but correlation coefficient or test may not be significant

  - Risk of residual confounding due to limited power

  - Magnitude of confounder-outcome association more important

- Comparison between crude and adjusted effect estimates

Important message

Confounding is about bias (the point estimate), not variance (the CI)

- As more variables are added to a regression model, confounding bias, if any, will decrease

- But uncertainty around estimated coefficient for variable of interest will increase

Need for covariate adjustment

Age at DX, T stage (size and/or extent of primary tumor), N stage, tumor site & treatment may be confounders

Table 3. Results of the univariate and multivariate analysis for local control and overall survival.

				Univariate analysis		Multivariate analysis	
	Variable		No. of events	HR (95% CI)	*p* value	HR (95% CI)	*p* value
Local control*	Tumor volume, cm^3	≤20	24	1.0 (ref)	<.001	1.0 (ref)	<.001
		21–40	34	1.4 (0.9–2.4)		1.2 (0.7–2.1)	
		41–60	18	1.7 (0.9–3.2)		1.4 (0.7–2.6)	
		>60	28	3.0 (1.7–5.2)		2.8 (1.5–5.2)	
	T status	T2	3	0.6 (0.2–1.9)	.0268	0.7 (0.2–2.3)	.292
		T3	23	0.6 (0.4–0.9)		0.8 (0.5–1.3)	
		T4	78	1.0 (ref)		1.0 (ref)	
	N status	N0	20	0.9 (0.6–1.5)	.481	–	–
		N1	13	0.9 (0.5–1.7)			
		N2	61	1.0 (ref)			
		N3	10	1.3 (0.6–2.5)			
Overall survival†	Tumor volume, cm^3	≤ 20	44	1.0 (ref)	<.001	1.0 (ref)	<.001
		21–40	59	1.3 (0.9–1.9)		1.2 (0.8–1.8)	
		41–60	36	1.8 (1.1–2.7)		1.5 (0.9–2.4)	
		>60	47	3.0 (2.0–4.5)		2.8 (1.8–4.3)	
	T status	T2	6	0.8 (0.4–1.9)	.0352	1.1 (0.5–2.4)	>.5
		T3	43	0.7 (0.5–0.9)		0.9 (0.6–1.3)	
		T4	139	1.0 (ref)		1.0 (ref)	
	N status	N0	28	0.7 (0.4–1.0)	.0017	0.7 (0.5–1.1)	.0133
		N1	28	1.2 (0.8–1.8)		1.3 (0.8–1.9)	
		N2	106	1.0 (ref)		1.0 (ref)	
		N3	25	2.2 (1.4–3.4)		2.5 (1.6–3.9)	

Abbreviations: HR, hazard ratio; CI, confidence interval; ref, reference.
**Multivariate model included tumor volume, T status, age at diagnosis, tumor site, and treatment.*
†Multivariate model included tumor volume, T status, N status, age at diagnosis, and tumor site.

Cox (semi-parametric) proportional hazards model

$$h(t) = h_0(t) * \exp\left[\beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_p\right]$$

where

- Hazard function $h(t)$ depends on p covariates $x_1, \ldots, x_p$ whose impact is measured by regression coefficients $\beta_1, \ldots, \beta_p$

- h_0 is baseline hazard, i.e., hazard if all x_i are equal to zero ($e^0 = 1$), estimated nonparametrically $\rightarrow$ no distributional assumption about survival times necessary

- Hazards may vary over time t

Interpretation of Cox model

- Essentially, Cox model is multivariable linear regression of logarithm of the hazard on variables x_i with baseline hazard as intercept that varies over time

- Covariates act multiplicatively on hazard at any point in time

- Key assumption: hazard of event in any group is a constant multiple of hazard in any other $\rightarrow$ hazard curves for groups should be proportional & do not cross

Interpretation of model coefficients

- $\exp(\beta_i)$ is hazard ratio (HR) for covariate x_i
- $\beta_i > 0 \rightarrow \text{HR} > 1 \rightarrow$ as value of x_i increases, event hazard increases & length of survival decreases
- HR describes risk change per one unit change in covariate x_i, i.e., difference in risk between two subjects with identical covariate values except for covariate x_i which differs by one unit (constant across range of continuous x_i)
- Assumption: change in risk is the same no matter what the other covariate values are, i.e., risk difference is average across risk differences for all possible covariate value combinations
- Departures from assumptions can be evaluated with interaction terms

Parameter estimation

- Partial likelihood function (behaves like a proper likelihood for most practical purposes)

$$\mathsf{L}(\beta) = \prod_{j=1}^{J} \frac{\exp(\beta' x_j)}{\sum_{i \text{ in } \mathsf{R}_j} \exp(\beta' x_i)}$$

- Product consists of one factor for each event time

- At each event time j, all subjects still at risk are in the risk set R_j & contribute

- Contribution of the whole risk set is probability that the subject that experienced the event (with its x_i's) actually experiences the event

Getting ready for some analyses

- Create 4-category age at DX variable:
  ageg4=1 if age≤50, 2 if 50<age≤60, 3 if 60<age≤65, 4 if age>65

```
COMPUTE ageg4=1+(age>50)+(age>60)+(age>65).  [click Transform - Compute Variable]
EXECUTE.
```

- Patientnr=9803282 excluded because of 393.83 ml tumor volume

- Patientnr=20000926 excluded due to negative values (-1.05) for variables `Time_meta_first` & `Time_any_meta` [started treatment 11/04/2000, suspicious lung lesions seen on pretreatment chest CT-scan of 10/03/2000 (confirmed after treatment); nevertheless, pt was treated in RADPLAT protocol; suspicion: metastases already present at start of treatment; pt was later (19/03/2003) diagnosed with liver metastases]

- Patientnr=323452 excluded because of missing `survival_status` [pt was in hospice on the last date of follow-up, date of death unknown, partial response after treatment (persistent disease), never without disease, no (known) distant metastases]

Exclusions in SPSS[1]

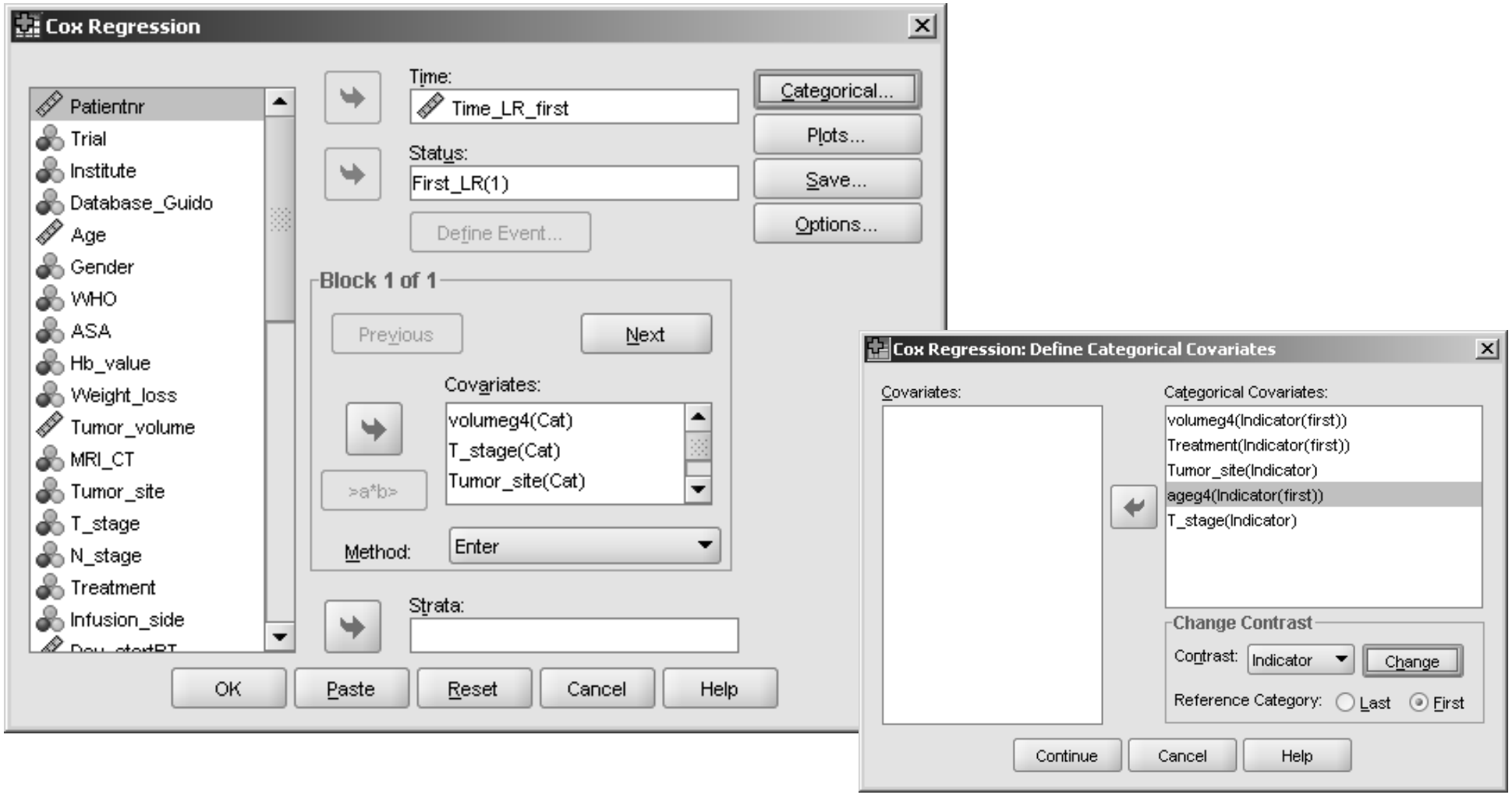
- Click: Data – Select Cases – If

- Enter:

```
(Patientnr~=9803282 AND Patientnr~=20000926 AND Patientnr~=323452)
```

SPSS code for multivariable Cox regression

Click: Analysis – Survival – Cox Regression



NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Determine reference category

(applies to any regression model)

- Don't use category with small number of events.
- For each of the k categories, except for the reference category, SPSS creates a dummy variable with value 1 for subjects in that category and 0 otherwise.
- All $k-1$ dummy variables are included in the regression model.
- The parameter estimated for a particular dummy variable is the effect of that category compared with the reference category.

Example: T stage

Number of events per category[2]

Report

Sum

T_stage	First_LR
2	3
3	23
4	77
Total	103

Categorical Variable Codings[b,c,d,e,f]

		Frequency	(1)	(2)	(3)
Tumor_site[a]	1	80	1	0	
	2	225	0	1	
	3	55	0	0	
T_stage[a]	2	14	1	0	
	3	111	0	1	
	4	235	0	0	
Treatment[a]	1	172	0	0	0
	2	88	1	0	0
	3	44	0	1	0
	4	56	0	0	1
volumeg4[a]	1.00	116	0	0	0
	2.00	125	1	0	0
	3.00	59	0	1	0
	4.00	60	0	0	1
ageg4[a]	1.00	91	0	0	0
	2.00	148	1	0	0
	3.00	55	0	1	0
	4.00	66	0	0	1

HRs and 95% CI from multivariable Cox regression[3]

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
volumeg4			12.609	3	.006			
volumeg4(1)	.174	.284	.377	1	.539	1.190	.682	2.077
volumeg4(2)	.269	.333	.655	1	.418	1.309	.682	2.514
volumeg4(3)	.980	.312	9.847	1	.002	2.665	1.445	4.915
T_stage			1.252	2	.535			
T_stage(1)	-.386	.611	.399	1	.528	.680	.205	2.252
T_stage(2)	-.266	.261	1.041	1	.307	.766	.459	1.278
Tumor_site			5.957	2	.051			
Tumor_site(1)	.645	.333	3.749	1	.053	1.905	.992	3.659
Tumor_site(2)	.152	.317	.229	1	.633	1.164	.625	2.166
Treatment			13.891	3	.003			
Treatment(1)	.078	.261	.088	1	.766	1.081	.648	1.803
Treatment(2)	.104	.353	.087	1	.768	1.110	.556	2.215
Treatment(3)	.997	.280	12.704	1	.000	2.709	1.566	4.686
ageg4			9.297	3	.026			
ageg4(1)	-.389	.275	2.008	1	.157	.678	.396	1.161
ageg4(2)	.361	.299	1.456	1	.228	1.434	.798	2.577
ageg4(3)	.301	.295	1.039	1	.308	1.351	.757	2.411

Beyond the HR

Survival proportion for a given risk group, i.e., with certain values for $x_1, \ldots, x_p$

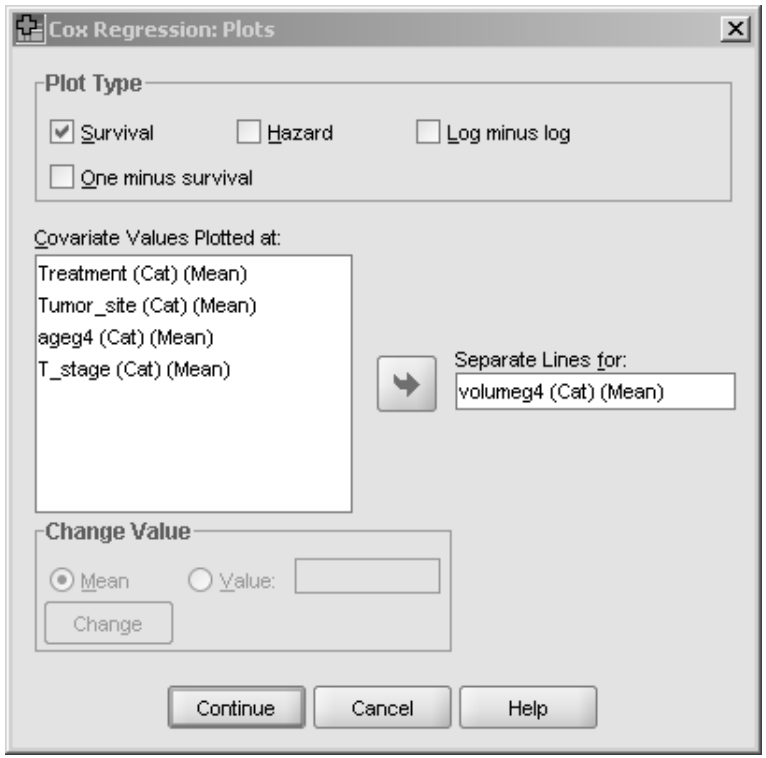
$$\mathrm{S}(t) = \mathrm{S}_0(t)^{\exp(\gamma)}$$

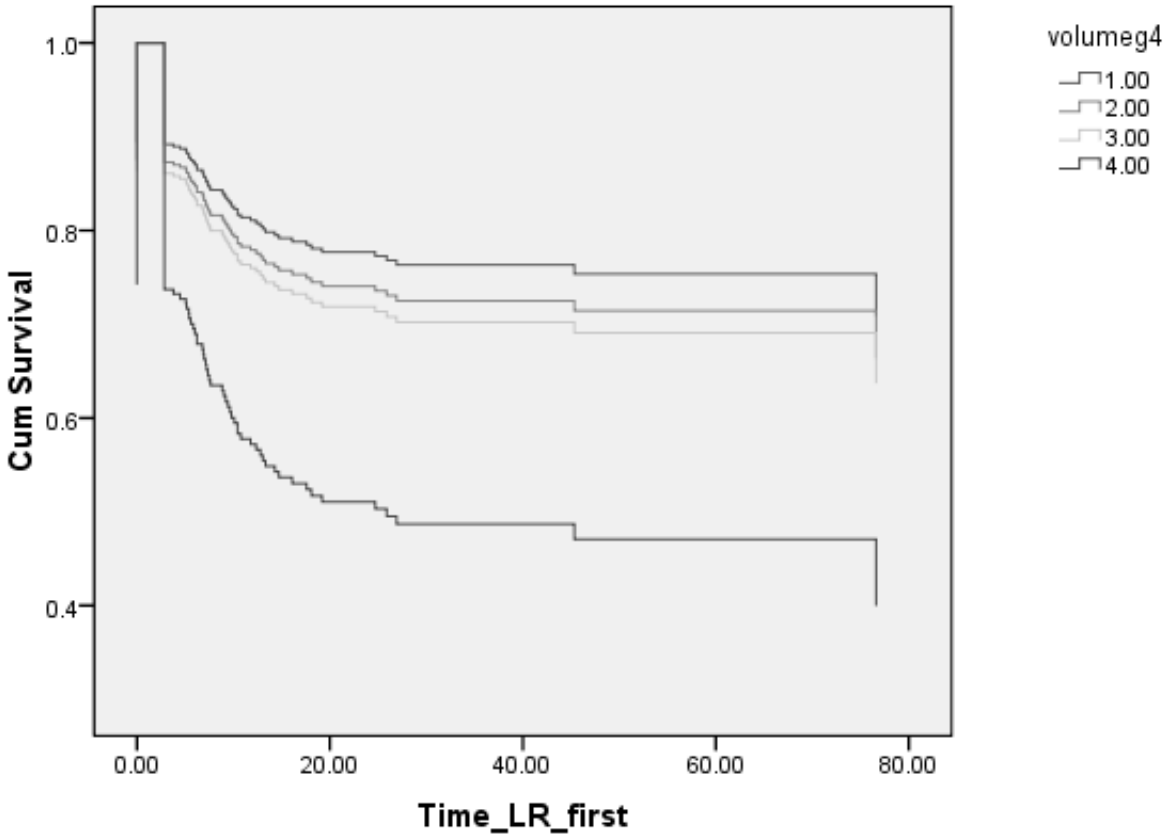
where $\mathrm{S}_0(t)$ is baseline survival (survival proportion when all covariates are equal to zero) and

$$\gamma = \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_p * x_p$$

Predicted recurrence-free survival by time since DX and tumor volume

In Analyze - Survival - Cox Regression, click on Plots





NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Use of continuous covariate as categorical or continuous?

- Continuous
  - Uses all available data
  - Requires only 1 DF
  - Assumes linear relationship with outcome, e.g., log HR in Cox regression
  - Can be extended by more flexible semi- and nonparametric methods (e.g., polynomials)

Categorical vs. continuous

- Categorical
  - Does not use all available data
  - Assumes homogeneity of effect within categories
  - Requires choice of #categories and cutpoints
  - Requires several DF (overall tests have low power)
  - Avoids strong assumptions about shape of exposure-response relationship (estimate for one part of the exposure range should not affect that at another)

Categorical vs. continuous

- Distinguish confounders from risk factors: adequate control of confounding in most cases by 4–5 categories or linear function, but shape of exposure-response important for main risk factor of interest

- How to categorize
  - Predetermined cutpoints (quartiles, quintiles), preferably meaningful
  - Don't choose cutpoints which minimize p-values (bias)
  - >2 categories to reduce loss of information & illustrate trend
  - Sufficient # subjects & events/category (percentiles among cases)

- Trend test: use continuous variable alongside with categorical version to provide best linear approximation

Continuous tumor volume

Variables in the Equation

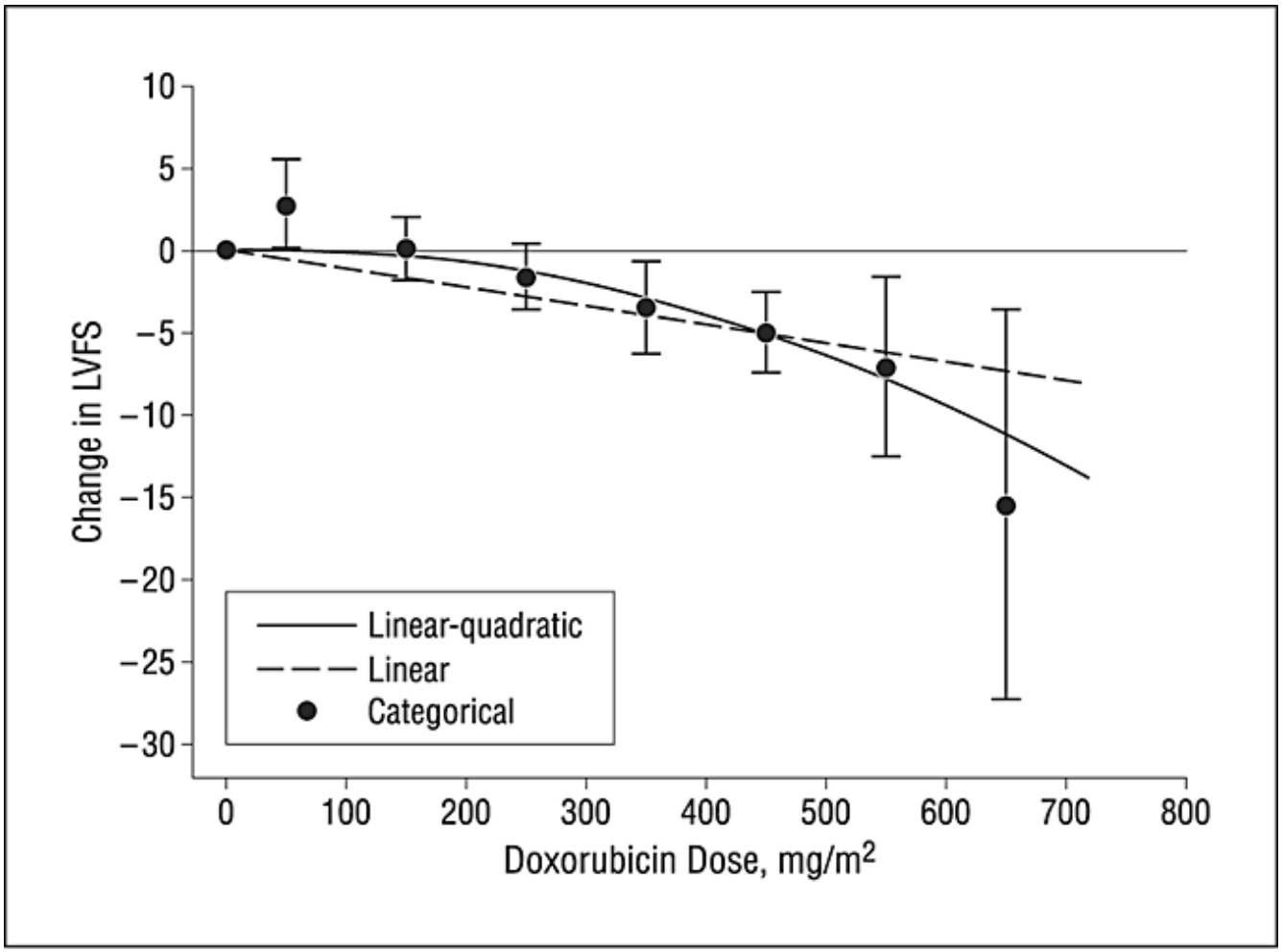
	B	SE	Wald	df	Sig.	Exp(B)
ageg4			8.915	3	.030	
ageg4(1)	-.403	.274	2.160	1	.142	.668
ageg4(2)	.311	.299	1.082	1	.298	1.365
ageg4(3)	.298	.294	1.024	1	.312	1.347
Treatment			13.460	3	.004	
Treatment(1)	.010	.262	.001	1	.970	1.010
Treatment(2)	.120	.353	.117	1	.733	1.128
Treatment(3)	.963	.279	11.892	1	.001	2.620
Tumor_site			6.826	2	.033	
Tumor_site(1)	.632	.330	3.683	1	.055	1.882
Tumor_site(2)	.084	.319	.070	1	.791	1.088
T_stage			.731	2	.694	
T_stage(1)	-.314	.606	.268	1	.604	.731
T_stage(2)	-.195	.258	.570	1	.450	.823
Tumor_volume	.013	.003	18.740	1	.000	1.013

Categorical vs. continuous

- In most cases we expect monotonic relationships and use models that respect such restrictions.

- Departure from (log-)linearity assumption with continuous variable can be evaluated by adding terms which allow for curvature, e.g., quadratic term, fractional polynomials, more flexible approaches (splines, generalized additive models).

Cardiac function in 5-year survivors of childhood cancer



Van der Pal et al. *Arch Intern Med* 2010

Assessing adequacy of model

- Residuals: essentially, difference between observed & model-predicted survival

- Analysis of residuals not trivial: interpretation complicated by censoring, skewed so smoothing needed, many different residuals suggested based on theoretic grounds

- Inclusion/exclusion of covariates: contribution to goodness of fit (magnitude of HRs & likelihood ratio test)

- Functional form of covariate: continuous vs. categorical, quadratic term to model non-linearity

- PH assumption: hazards are proportional at all points in time

Graphical evaluation of PH assumption

- Plotting hazards is of limited use: empirical hazards poorly estimated & difficult to assess visually

- Instead, plot cum. hazard vs. survival time (lines should not cross)

- Better

$$-\log \text{cumulative hazard} = -\log[-\log(\text{survival})]$$

  plotted against log(time) should be parallel
  (continuous variables need to be categorized into groups)

- Non-parallel lines due to non-PH or omission of important covariate

Alternatives

- Schoenfeld residuals

- Time-dependent covariate test

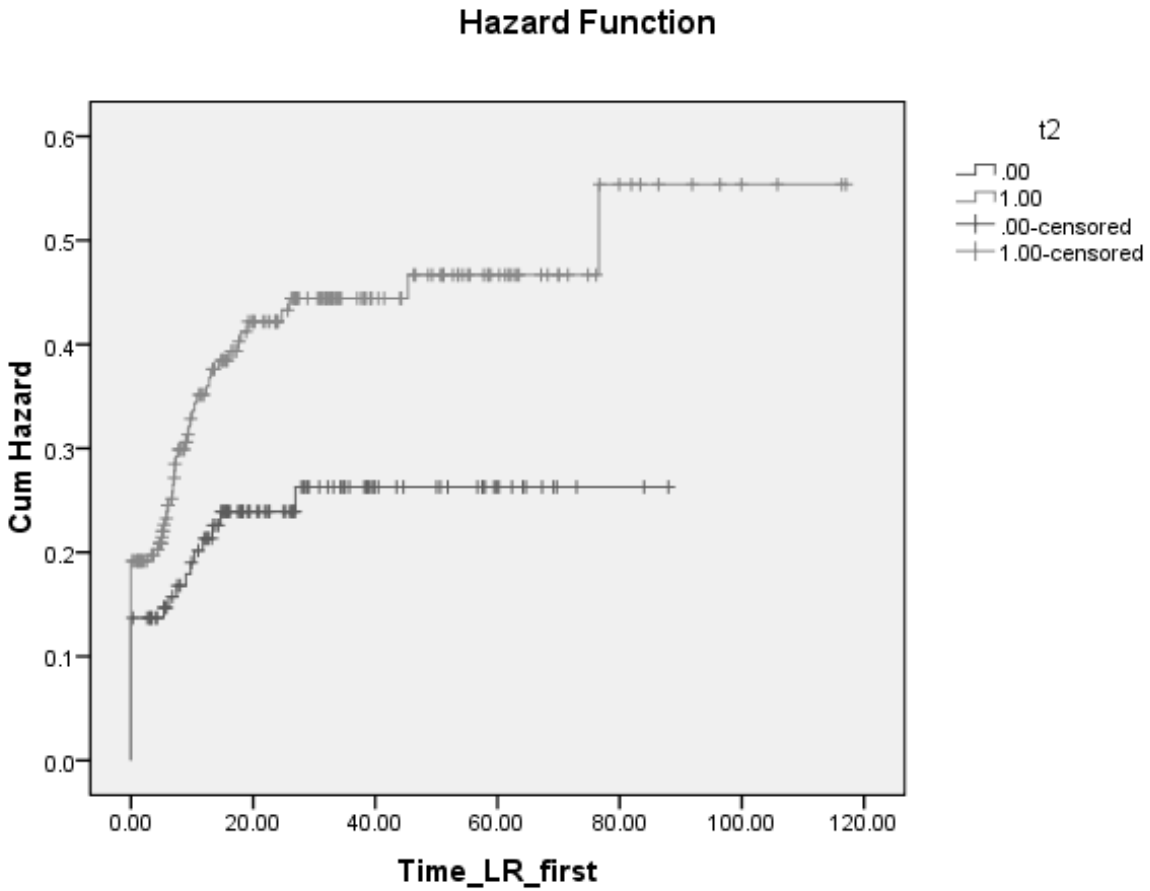
If PH not fulfilled

- Stratification
  - Stratify Cox regression on variable with non-PH
  - Assumption is relaxed to piecewise (stratum-wise) PH
  - No effect size estimated for stratification variable
- Include interaction between non-PH variable and log(time+1)

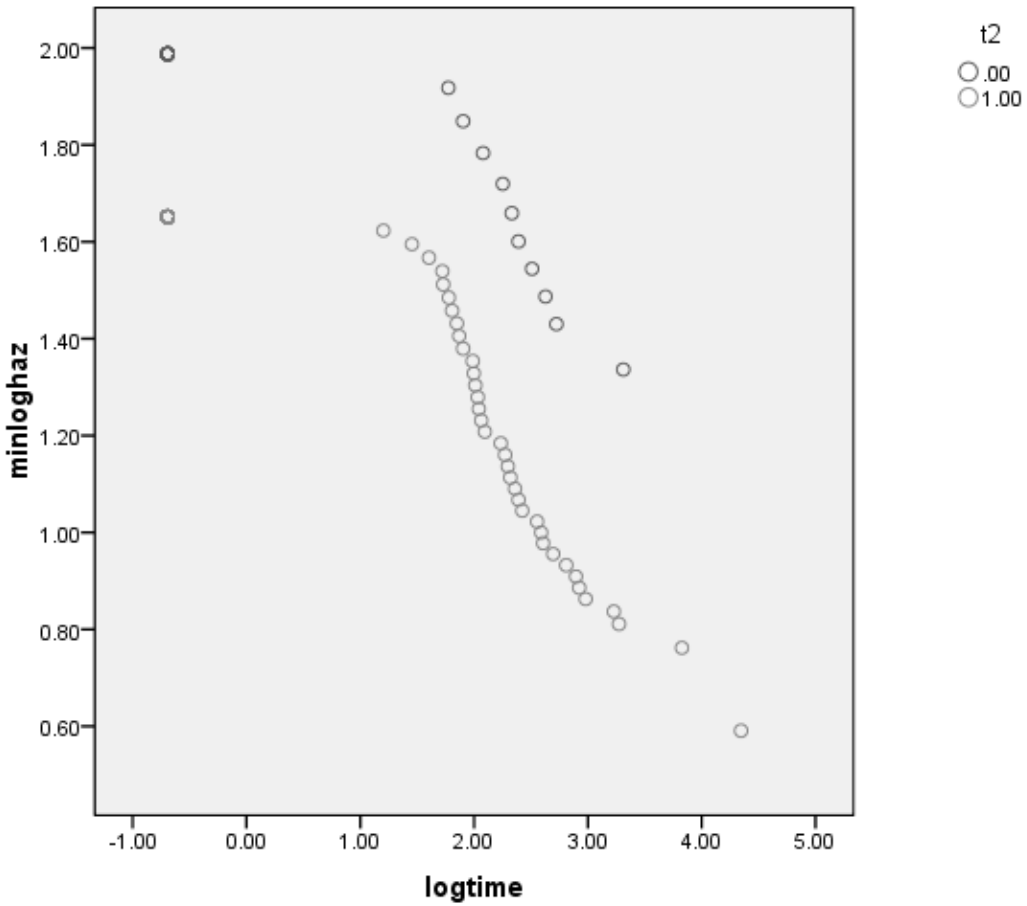
Example: T stage

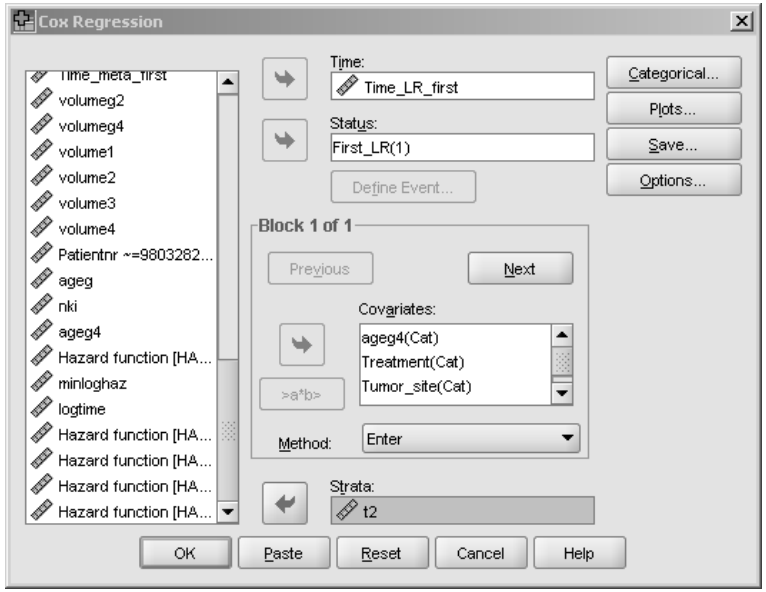
- Only few pts with T=2 $\longrightarrow$ combine with T=3
  `T2`=1 if `T_stage`=4, zero otherwise[4]

- Plot cumulative hazard by T stage[5]



Negative log cumulative hazard vs. log survival time[6]





No evidence of non-PH – Stratification for T2 not needed

Interaction

- Hazard ratios

	M	
E	0	1
0	1.0 (ref)	HR(M)
1	HR(E)	HR(M)∗HR(E)∗IHR

- Interaction

$$\begin{aligned} \text{HR(M,E)} &= e^{\beta_1(\text{M=1})+\beta_2(\text{E=1})+\beta_3(\text{M=E=1})} \\ &= e^{\beta_1(\text{M=1})} * e^{\beta_2(\text{E=1})} * e^{\beta_3(\text{M=E=1})} \\ &= \text{HR(M)} * \text{HR(E)} * \text{IHR} \end{aligned}$$

- Hazard ratios

	M	
E	0	1
0	1.0 (ref)	e^{β_1}
1	e^{β_2}	$e^{\beta_1+\beta_2+\beta_3}$

Notation: (M = 1) = 1 if M = 1, 0 otherwise, dito for E

Example

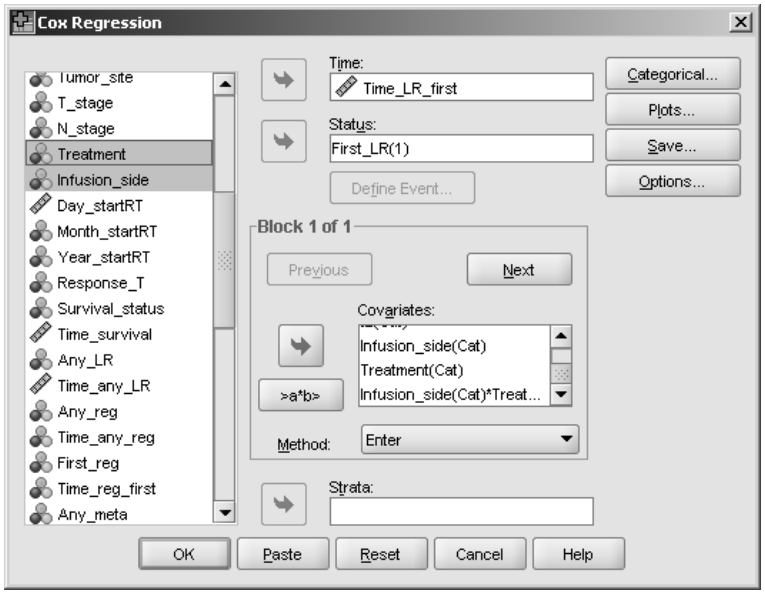
- Focus on pts with intra-arterial (IA) or intravenous (IV) chemoradiation
- IV: 3*100 mg/m2 cisplatin plus 70 Gy in 36 fractions
- IA: 4*150 mg/m2 cisplatin in tumor-feeding artery immediately followed by systemic rescue, RT dito
- IA preferably single sided but switched to double sided with equal distribution of cisplatin if tumor invasion was >1 cm across anatomical midline
- Interaction between IV/IA and infusion side?

Analysis

- Only keep pts w/ `treatment`=1 (IA) or `treatment`=2 (IV)
  (Data - Select Cases - If)

- Fit Cox model as before and add `infusion_side` (1=single, 2=double)

- Also add interaction between `infusion_side` & `treatment`
  (select both variables and click on ">a*b>")



Result

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
ageg4			8.369	3	.039			
ageg4(1)	-.287	.333	.744	1	.388	.750	.391	1.441
ageg4(2)	.686	.359	3.659	1	.056	1.986	.983	4.013
ageg4(3)	.304	.355	.731	1	.393	1.355	.676	2.716
Tumor_site			6.454	2	.040			
Tumor_site(1)	.838	.426	3.876	1	.049	2.312	1.004	5.325
Tumor_site(2)	.180	.395	.207	1	.649	1.197	.552	2.596
volumeg2	.703	.274	6.572	1	.010	2.019	1.180	3.456
t2	.195	.334	.340	1	.560	1.215	.631	2.339
Infusion_side	.354	.316	1.258	1	.262	1.425	.767	2.648
Treatment	.624	.412	2.297	1	.130	1.867	.833	4.183
Infusion_side*Treatment	-.978	.533	3.372	1	.066	.376	.132	1.068

HR table

Infusion side	Treatment	
	IA	IV
Single	1.0 (ref)	1.867
Double	1.425	1.867*1.425*.376=1.000#

- # expected without interaction: 1.867*1.425=2.660
- HR(IV vs. IA) among single-sided = 1.867
- HR(IV vs. IA) among double-sided = 1.000/1.425=.702
- Test of heterogeneity of IV vs. IA effect, p=.066
- Full story: Rasch et al., *Cancer* 2010

SPSS code (syntax and clicking)

1. Exclusions in SPSS
   Click Data – Select Cases – If and fill in
   `(Patientnr~=9803282 AND Patientnr~=20000926 AND Patientnr~=323452)`

```
USE ALL.
COMPUTE filter_$=(Patientnr~=9803282 AND Patientnr~=20000926 AND Patientnr~=323452).
VARIABLE LABEL filter_$ 'Patientnr ~=9803282 AND Patientnr~=20000926 AND Patientnr~=323452 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

2. Number of events per T stage category
   Click Analyze – Compare Means – Means and select variables `First_LR` and `T_stage`. Under Options, select Sum.

```
MEANS TABLES=First_LR BY T_stage
   /CELLS SUM.
```

3. HRs and 95% CI from multivariable Cox regression
   Click Analyze – Survival – Cox Regression, select the Time variable and the Status variable, provide the value indicating an event, and select covariates. Click Categorical, select each covariate to be treated as a categorical variable, and select the reference category (Last or First). Do not forget to click Change after selection of the reference category.

```
COXREG Time_LR_first
  /STATUS=First_LR(1)
  /CONTRAST (Treatment)=Indicator(1)
  /CONTRAST (volumeg4)=Indicator(1)
  /CONTRAST (T_stage)=Indicator
  /CONTRAST (Tumor_site)=Indicator
  /CONTRAST (ageg4)=Indicator(1)
  /METHOD=ENTER volumeg4 T_stage Tumor_site Treatment ageg4
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

4. Combine pts with T=2 and T=3 in one category, i.e., create new binary variable `T2`=1 if `T_stage`=4, zero otherwise
   Transform – Compute
   Target variable: `T2`
   Numeric expression: `T_stage`=4

5. Plot cumulative hazard: Analyze – Survival – Kaplan-Meier
   Time: `Time_LR_first`
   Status: `First_LR`
   Click "Define event" and write 1 for single value.
   Factor: `T2`
   Click Options – Plots – Hazard

6. Calculate negative log cumulative hazard and plot vs. log survival time

- Save hazard
  In Analyze – Survival – Kaplan-Meier as above, click Save – Hazard
  Creates new variable `HAZ_1`

- Calculate –log(hazard)
  Transform – Compute
  Target variable: `minloghaz`
  Numeric expression: –LN(`HAZ_1`)

- Calculate log(survival time)
  Transform – Compute
  Target variable: `logtime`
  Numeric expression: LN(`Time_LR_first`+.5 —)

- Plot of `minloghaz` against `logtime`
  Graphs – Scatter – Simple
  Y-axis: `minloghaz`
  X-axis: `logtime`
  Put `T2` in the "Set Markers By" box