SESSION 2:
WEIGHTED LOG RANK TESTS

Module 9: Survival Analysis for Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Elizabeth R. Brown, ScD
Member, Fred Hutchinson Cancer Research Center
and
Research Professor
Department of Biostatistics
University of Washington

OVERVIEW

- Session 1
  - Review basics
  - Cox model for adjustment and interaction
  - Estimating baseline hazards and survival
- Session 2
  - Weighted logrank tests
- Session 3
  - Other two-sample tests
- Session 4
  - Choice of outcome variable
  - Power and sample size
  - Information accrual under sequential monitoring

KEY IN CLINICAL TRIALS

- Group comparisons
  - Two groups
  - k groups
  - Test for (linear) trend

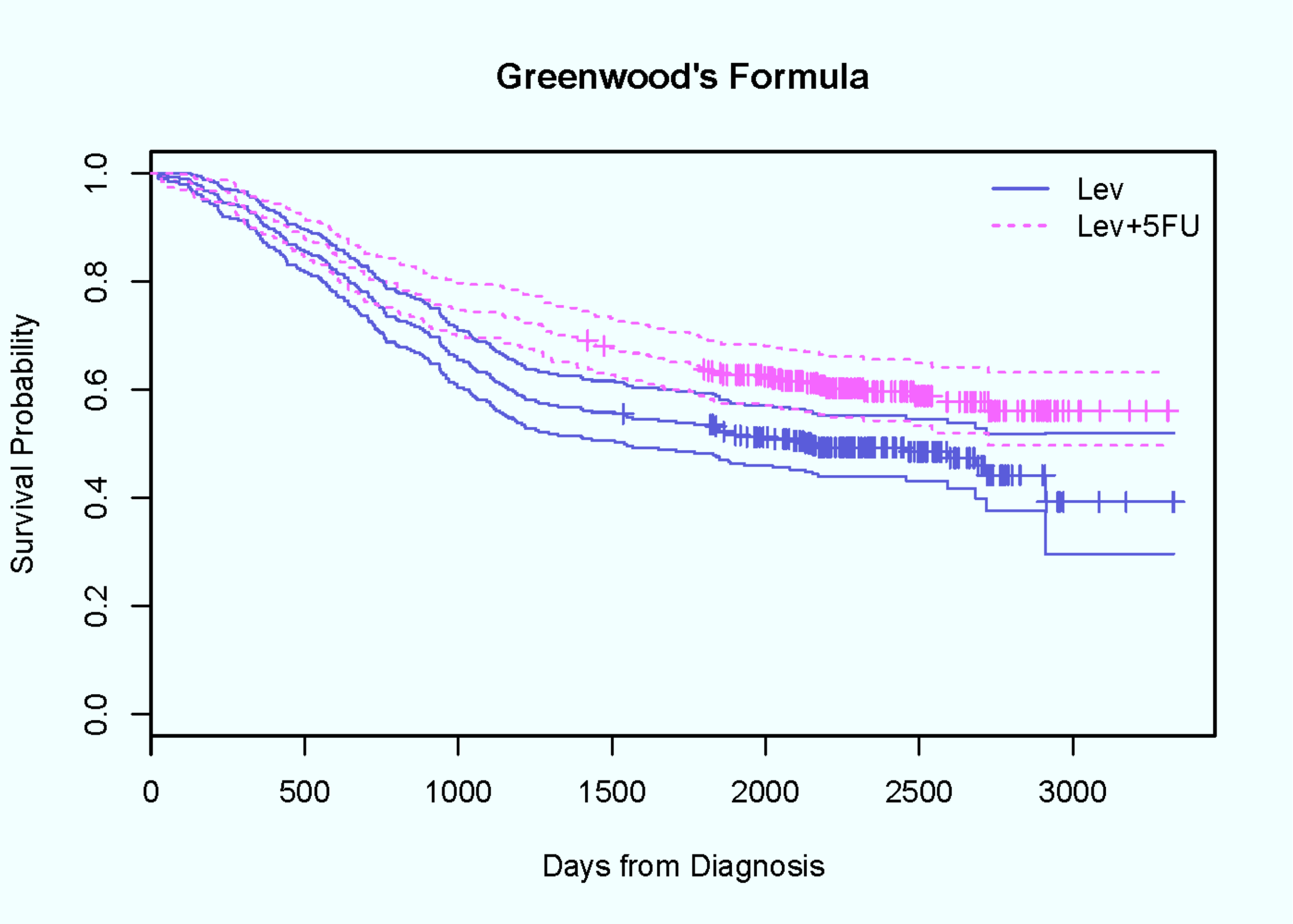
- Assume, H_0 : no differences between groups

EXAMPLE

- Levamisole and Fluorouracil for adjuvant therapy of resected colon carcinoma
  Moertel et al, 1990, 1995
- 1296 patients
- Stage B_2 or C
- 3 unblinded treatment groups
  - Observation only
  - Levamisole (oral, 1yr)
  - Levamisole (oral, 1yr) + fluorouracil (intravenous 1yr)

COLON DATA EXAMPLE

- Kaplan-Meier plots and pointwise CIs



THE P-VALUE QUESTION

- Statistical significance?

TWO-GROUP COMPARISONS

- A number of statistical tests available
- The calculation of each test is based on a contingency table of group by status at each observed survival (event) time t_j, $j=1,\ldots m$, as shown in the Table below.

Event/Group	**1**	**2**	**Total**
Die	$d_{1(j)}$	$d_{2(j)}$	$D_{(j)}$
Do Not Die	$n_{1(j)}-d_{1(j)}= s_{1(j)}$	$n_{2(j)}-d_{2(j)} = s_{2(j)}$	$N_{(j)}-D_{(j)} = S_{(j)}$
At Risk	$n_{1(j)}$	$n_{2(j)}$	$N_{(j)}$

TWO-GROUP COMPARISONS

- The contribution to the test statistic at each event time is obtained by calculating the expected number of deaths in group 1(or 0), assuming that the survival function is the same in each of the two groups.
- This yields the usual "*row total times column total divided by grand total*" estimator. For example, using group 1, the estimator is

$$\hat{E}_{1(j)} = \frac{n_{1(j)} D_{(j)}}{N_{(j)}}$$

- Most software packages base their estimator of the variance on the hypergeometric distribution, defined as follows:

$$\hat{V}_{(j)} = \frac{n_{1(j)} n_{2(j)} D_{(j)} \left(N_{(j)} - D_{(j)}\right)}{N_{(j)}^{2} \left(N_{(j)} - 1\right)}$$

TWO-GROUP COMPARISONS

- Each test may be expressed in the form of a ratio of weighted sums over the observed survival times as follows

$$Q = \frac{\left[\sum_{j=1}^{m} W_{(j)}\left(d_{1(j)} - \hat{E}_{1(j)}\right)\right]^2}{\sum_{j=1}^{m} W_{(j)}^2 \hat{V}_{(j)}}$$

- Where *j = 1,…,m* are the ordered unique event times
- Under the null hypothesis and assuming that the censoring experience is independent of group, and that the total number of observed events and the sum of the expected number of events is large, then the *p*-value for *Q* may be obtained using the chi-square distribution with one degree-of-freedom,

$$p = \Pr\left(\chi^2(1) \geq Q\right)$$

WEIGHTING

- Weights used by different tests

- Log Rank: $W_j = 1$

- Wilcoxon: $W_j = N_j$

- Tarone-Ware: $W_j = \sqrt{N_j}$

Most frequently used test weights later times relatively more heavily, while Wilcoxon weights early times more heavily

- Peto-Prentice: $W_j = S\left(t_{(j)}\right)$ where $S(t) = \prod_{t_{(i)} \le t} \left(\frac{N_i + 1 - D_i}{N_i + 1} \right)$

- Fleming-Harrington: $W_j = \left[\hat{S}\left(t_{(j-1)}\right) \right]^p \times \left[1 - \hat{S}\left(t_{(j-1)}\right) \right]^q$

  $p = q = 0 \quad \Rightarrow \quad W_j = 1$

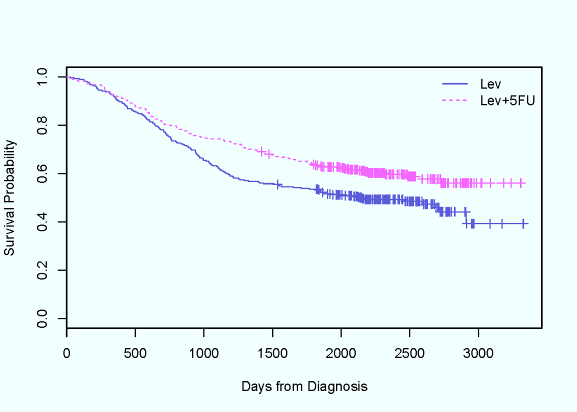
  $p = 1, q = 0 \quad \Rightarrow \quad W_j =$ Kaplan-Meier estimate at previous survival time

- and $\hat{S}\left(t_{(j-1)}\right)$ is the Kaplan-Meier estimator at time t_{j-1}

COLON CANCER EXAMPLE

- Comparing Lev vs Lev+5FU

Group	N	Obs	Exp
Lev	310	161	136.9
Lev+5FU	304	123	147.1
Total	614	284	284.0



- Log-rank test: $\chi^2(1) = 8.2$, p-value = 0.0042
- Peto-Prentice: $\chi^2(1) = 7.6$, p-value = 0.0058
- Wilcoxon: $\chi^2(1) = 7.3$, p-value = 0.0069
- Tarone-Ware: $\chi^2(1) = 7.7$, p-value = 0.0055
- Flem-Harr(1,.0): $\chi^2(1) = 7.6$, p-value = 0.0056
- Flem-Harr(1,.3): $\chi^2(1) = 9.5$, p-value = 0.0020

- Example where choice of weights makes a difference

EXAMPLE: LOW BIRTH WEIGHT INFANTS

- Data from UMass
- Goal: determine factors that predict the length of time low birth weight infants (<1500 grams) with bronchopulmonary dysplasia (BPD) were treated with oxygen
- Note: observational study, not clinical trial
- 78 infants total, 35 (43 not) receiving surfactant replacement therapy
- Outcome variable: total number of days the baby required supplemental oxygen therapy

SUMMARY STATISTICS - LBWI

- The estimated median number of days of therapy
    - for those babies who did not have surfactant replacement therapy
        - 107 {95% CI: (71, 217)},
    - for those who had the therapy is
        - 71 {95% CI: (56, 110)}

    - The median number of days of therapy for the babies not on surfactant is about 1.5 times longer than those using the therapy.

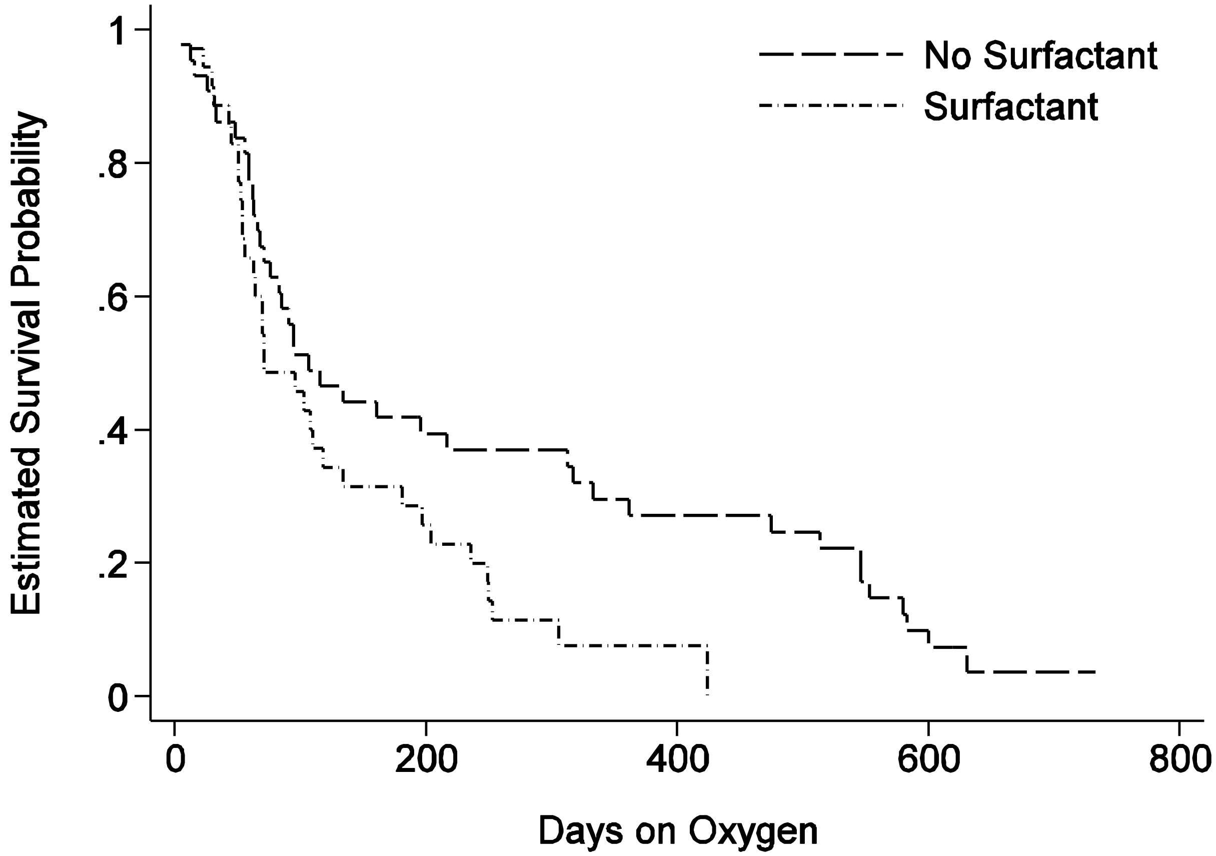
TWO-GROUP COMPARISONS LBWI

- Different weighting approaches

Test	Statistic	p – value
Log-rank	5.62	0.018
Wilcoxon	2.49	0.115
Tarone-Ware	3.70	0.055
Peto-Prentice	2.53	0.111
Flem-Harr(1,0)	2.66	0.103
Flem-Harr(0,1)	9.07	0.0026

EXAMPLE: LBWI

- Kaplan-Meier plot



WEIGHTS

- How should weights be chosen?
  - Must be determined during design phase. It is not reasonable to look at the survival curves first, then choose weights
  - Is there a reason to believe we will have non-proportional hazards?
    - If not, go with the logrank test
    - If so, consider what survival differences are most meaningful (early vs late)
- Ordinarily: No weights = log rank test

TRIALS WHERE WEIGHTS ARE IMPORTANT ?

- Question: Examples of settings where log rank and Cox model
  - Might be inappropriate?
  - Have low power?

K-GROUPS

- K-Group Comparisons

Group	1	2	…	*k*	…	*K*	Total
Die	$d_{1(j)}$	$d_{2(j)}$	…	$d_{k(j)}$	…	$d_{K(j)}$	$D_{(j)}$
Not Die	$s_{1(j)}$	$s_{2(j)}$	…	$s_{k(j)}$	…	$s_{K(j)}$	$S_{(j)}$
At Risk	$n_{1(j)}$	$n_{2(j)}$	…	$n_{k(j)}$	…	$n_{K(j)}$	$N_{(j)}$

- In a manner similar to the two-group case, we estimate the expected number of events for each group under an assumption of equal survival functions as

$$\hat{E}_{k(j)} = \frac{D_{(j)} n_{k(j)}}{N_{(j)}}, \; k = 1,2, \quad ,K$$

K-GROUP COMPARISON

- Again, compare observed vs expected
- Quadratic form Q
- Under the null hypothesis and
  if the summed estimated expected number of events is large
- Test statistic $p = \Pr\left(\chi^2(K-1) \geq Q\right)$

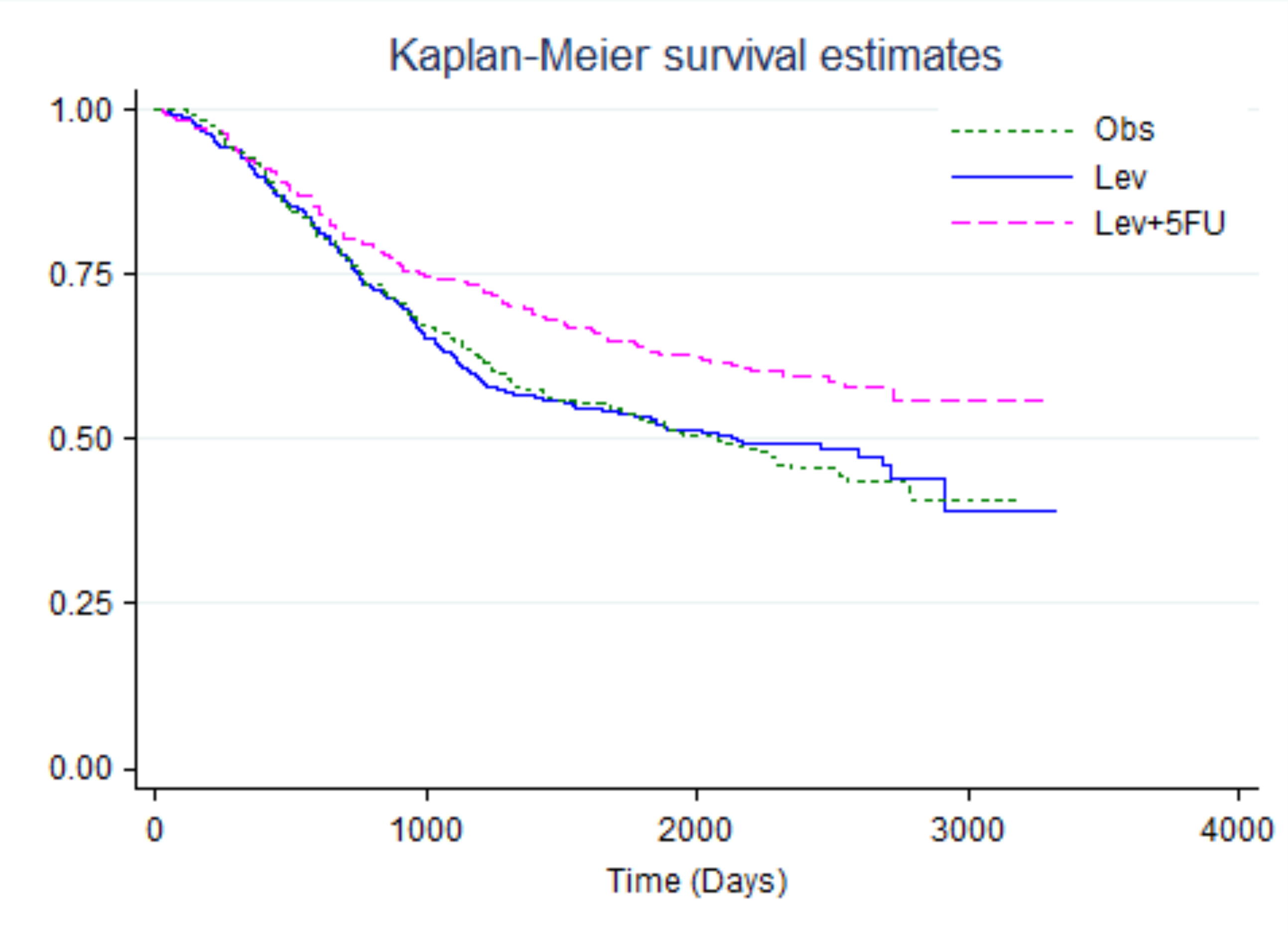
COLON CANCER EXAMPLE

- Obs vs Lev vs Lev+5FU

- Log-rank test: $\chi^2(2)$ = 11.7, p-value = 0.0029
- Wilcoxon: $\chi^2(2)$ = 9.7, p-value = 0.0078
- Peto-Prentice: $\chi^2(2)$ = 10.3, p-value = 0.0059
- Tarone-Ware: $\chi^2(2)$ = 10.6, p-value = 0.0049
- Flem-Harr(1,0): $\chi^2(2)$ = 10.4, p-value = 0.0056
- Flem-Harr(1,.3): $\chi^2(2)$ = 13.7, p-value = 0.0011

COLON CANCER EXAMPLE

- Obs vs Lev vs Lev+5FU



TREND TEST – EXAMPLE 1 (COLON)

- Obs vs Lev vs Lev+5FU
- Coding ?

- Pretend you did not see any results yet …

TREND TEST

- H_0: survival functions are equal
- H_A: survival functions are rank-ordered and follow the trend specified by a vector of coefficients

- Examples
  - Drug dosing
  - Age

TREND

- Null hypothesis: $\lambda_1(t) \equiv \lambda_2(t) \equiv \ldots \equiv \lambda_k(t)$
- Specific alternative hypothesis:

$$c^{s_1}\lambda_1(t) \equiv c^{s_2}\lambda_2(t) \equiv \ldots \equiv c^{s_k}\lambda_k(t), c \neq 1$$

- The test statistic for trend uses "scores": s_1, s_2,…, s_k

$$\frac{\left(\sum_{i=1}^{k} s_i \sum_{j=1}^{J_k} \left(d_{ij} - E_{ij}\right)\right)^2}{s'Vs}$$

- Good power when average difference between observed and expected events grows or diminishes with increasing s_i

TREND ANALYSIS

- Trend test

Groups				
Obs	0			
Lev	1			
Lev+5FU	2			
	p – value			
Log-rank				
Wilcoxon				
Tarone-Ware				
Peto-Prentice				

TREND ANALYSIS

- Trend test

Groups				
Obs	0			
Lev	1			
Lev+5FU	2			
	p – value			
Log-rank	0.002			
Wilcoxon	0.007			
Tarone-Ware	0.004			
Peto-Prentice	0.005			

TREND ANALYSIS

- Trend test

Groups				
Obs	0	0		
Lev	1	0.25		
Lev+5FU	2	1		
	p – value			
Log-rank	0.002	0.0007		
Wilcoxon	0.007	0.002		
Tarone-Ware	0.004	0.001		
Peto-Prentice	0.005	0.002		

TREND ANALYSIS

- Trend test

Groups				
Obs	0	0	0	
Lev	1	0.25	0.75	
Lev+5FU	2	1	1	
	p – value			
Log-rank	0.002	0.0007	0.01	
Wilcoxon	0.007	0.002	0.008	
Tarone-Ware	0.004	0.001	0.02	
Peto-Prentice	0.005	0.002	0.02	

TREND ANALYSIS

- Trend test

Groups				
Obs	0	0	0	0
Lev	1	0.25	0.75	?
Lev+5FU	2	1	1	1
	p – value			
Log-rank	0.002	0.0007	0.01	0.79
Wilcoxon	0.007	0.002	0.008	0.96
Tarone-Ware	0.004	0.001	0.02	0.87
Peto-Prentice	0.005	0.002	0.02	0.93
Flem-Harr(1,.3)	0.0007	0.0002	0.004	0.69

- Another example regarding trend

TREND – EXAMPLE 2

- Thomas et al. (1977)
- Also Marubini and Valsecchi (1995, p 126)
- 29 Animals
- 3 level of carcinogenic agent (0, 1.5, 2.0)
- Outcome: time to tumor formation

Group	**Dose**	**N**	**Times to event *(t)* or censoring *(t+)***
0	0	9	73+,74+,75+,76,76,76+,99,166,246+
1	1.5	10	43+,44+,45+,67,68+,136,136,150,150,150
2	2.0	10	41+,41+,47,47+,47+,58,58,58,100+,117

TREND TEST

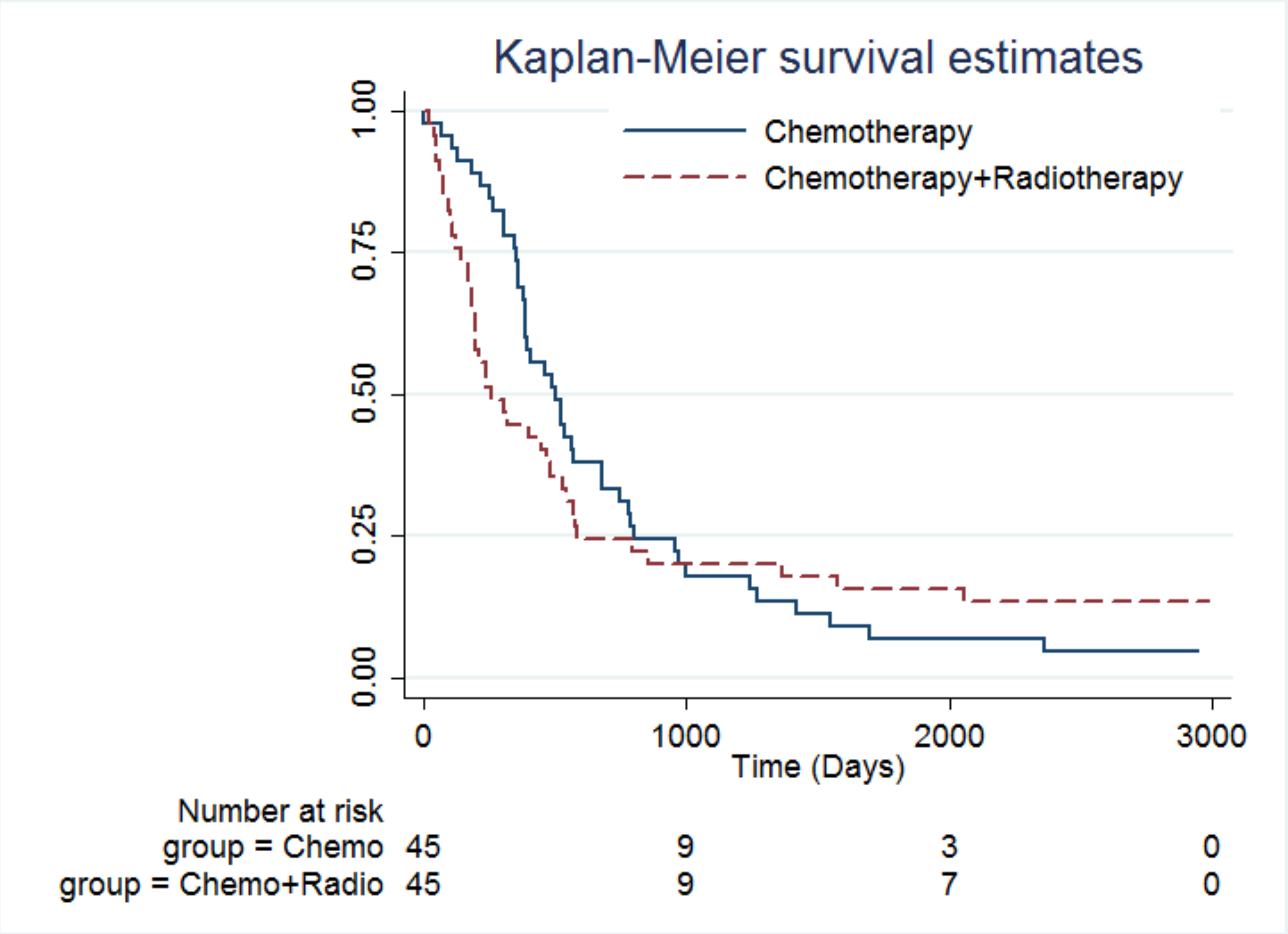
- Dose example, 29 animals

Test (Group differences)	df	Chi2	P-value
Log-rank	2	8.05	0.018
Wilcoxon	2	9.04	0.011
Trend test			
Log-rank (1,2,3)	1	5.87	0.015
Wilcoxon (1,2,3)	1	6.26	0.012
Log-rank (0,1.5,2)	1	3.66	0.056
Wilcoxon (0,1.5,2)	1	3.81	0.051

EXAMPLE 3

- Stablein and Koutrouvelis (1985)
- Gastrointestinal Tumor Study Group (1982)
- Chemotherapy vs.
  Chemotherapy and Radiotherapy
- 90 patients (45 per group)

KAPLAN-MEIER SURVIVAL CURVES



TEST STATISTICS – EXAMPLE 3

Test	Statistic	p – value
Log-rank		?
Wilcoxon		?
Peto-Prentice		?
Tarone-Ware		?
Fl-Ha(1,0)		?
Fl-Ha(0,1)		?

TEST STATISTICS – EXAMPLE 3

Test	Statistic	p – value
Log-rank	0.23	0.64
Wilcoxon		
Peto-Prentice		
Tarone-Ware		
Fl-Ha(1,0)		
Fl-Ha(0,1)		

TEST STATISTICS – EXAMPLE 3

Test	Statistic	p – value
Log-rank	0.23	0.64
Wilcoxon	3.96	0.047
Peto-Prentice		
Tarone-Ware		
Fl-Ha(1,0)		
Fl-Ha(0,1)		

TEST STATISTICS – EXAMPLE 3

Test	Statistic	p – value
Log-rank	0.23	0.64
Wilcoxon	3.96	0.047
Peto-Prentice	4.00	0.046
Tarone-Ware	1.90	0.17
Fl-Ha(1,0)		
Fl-Ha(0,1)		

TEST STATISTICS – EXAMPLE 3

Test	Statistic	p – value
Log-rank	0.23	0.64
Wilcoxon	3.96	0.047
Peto-Prentice	4.00	0.046
Tarone-Ware	1.90	0.17
Fl-Ha(1,0)	2.59	0.11
Fl-Ha(0,1)	4.72	0.03

TEST STATISTICS – EXAMPLE 3

Test	Statistic	p – value
Log-rank	0.23	0.64
Wilcoxon	3.96	0.047
Peto-Prentice	4.00	0.046
Tarone-Ware	1.90	0.17
Fl-Ha(1,0)	3.96	0.047
Fl-Ha(0,1)	2.06	0.15

- Why the difference?

GROUP COMPARISONS

- H_0: $S_1(t) = S_2(t)$ $\lambda_1(t) = \lambda_2(t)$

- Possible alternative
  - Survival function: $S_2(t) = S_1(t)^C, C \neq 1$
  - Hazard function: $\lambda_2(t) = C\lambda_1(t), C \neq 1$

$$\ln\left(\lambda_2(t)\right) = \ln\left(\lambda_1(t)\right) + C, \quad C \neq 1$$

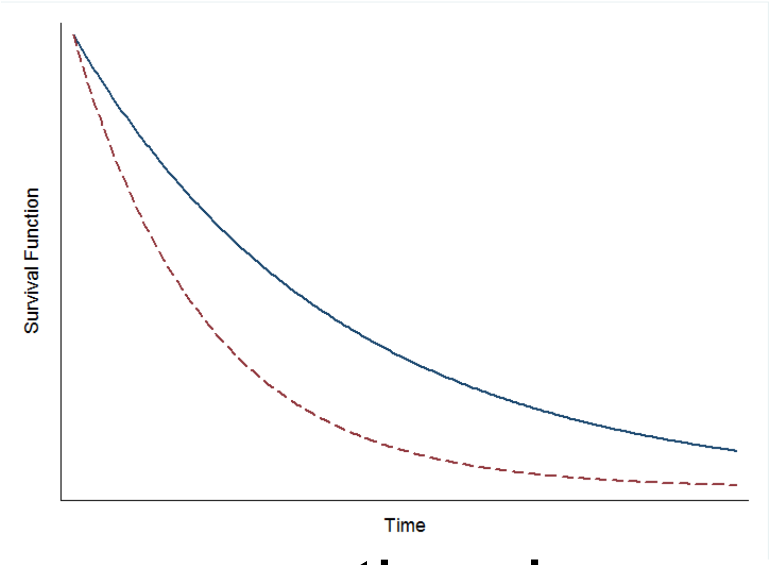
- Log-rank test most powerful
  if hazards are proportional

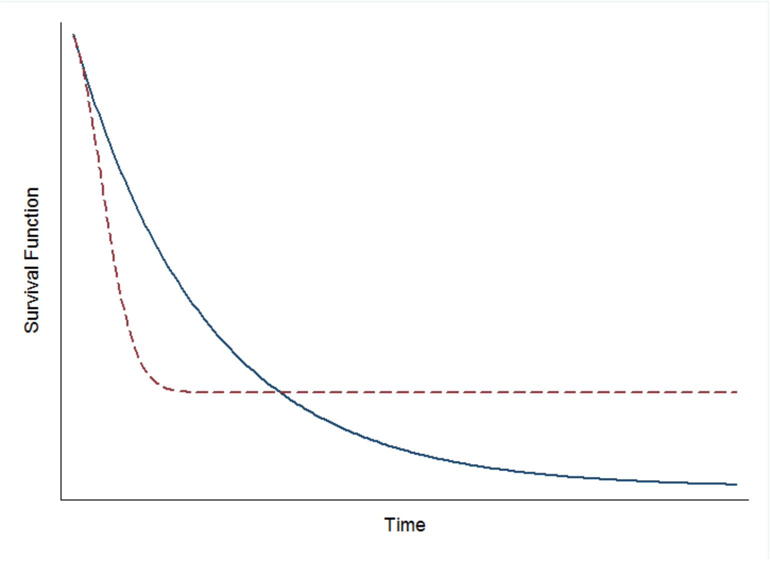
SURVIVAL FUNCTIONS

- We can detect

this

but ordinarily not this





proportional

not proportional

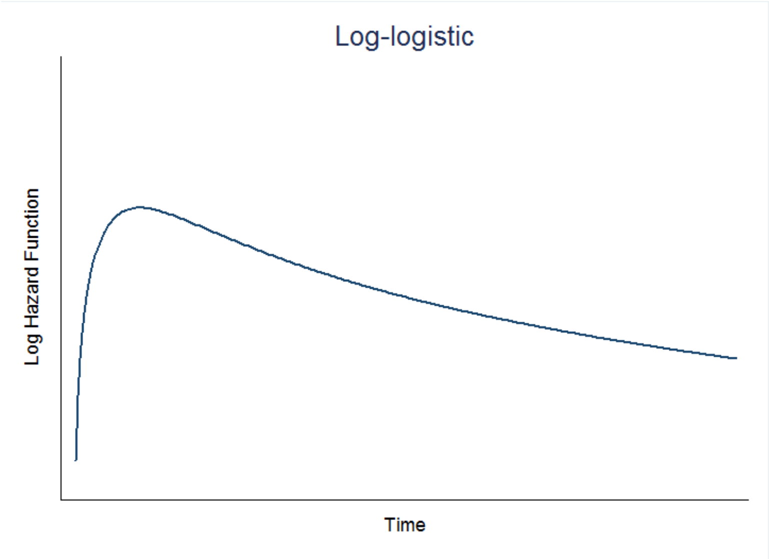
(generated as 2 exponential distributions)

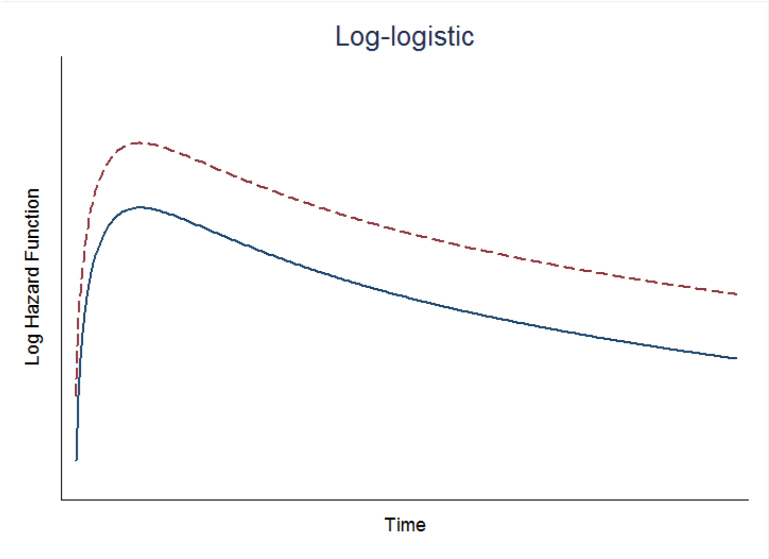
PROPORTIONAL HAZARDS

- Easier to visualize on log hazard scale

GROUP COMPARISONS

- Proportional hazards – use log hazards scale
- Example: log-logistic survival times
- Hazards plotted on log scale





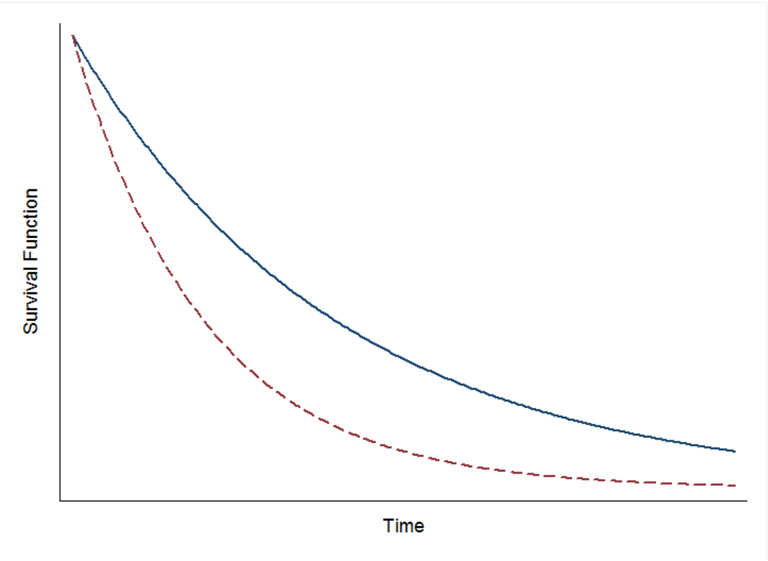
SO FAR

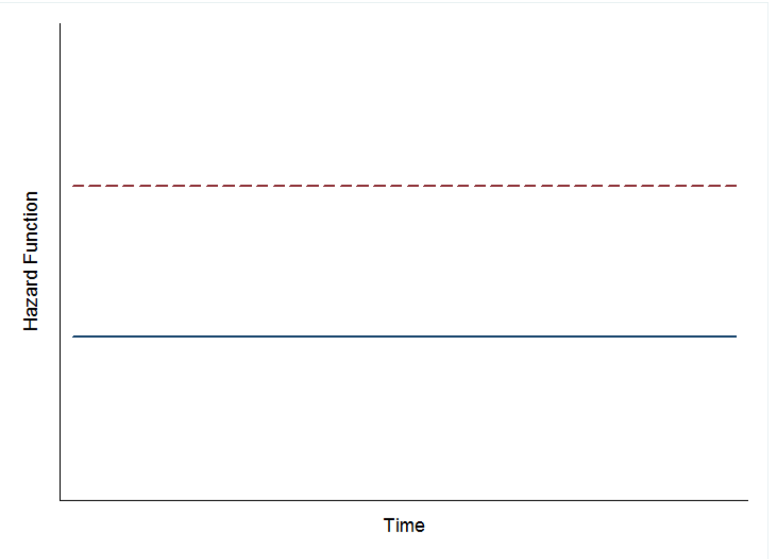
- Two and K – group comparisons
- Trend tests

- Non-parametric
- Did not make use of actual values of time

PARAMETRIC MODELS

- Control group: Exponential(0.5)
- Example
- Survival functions   Hazard functions

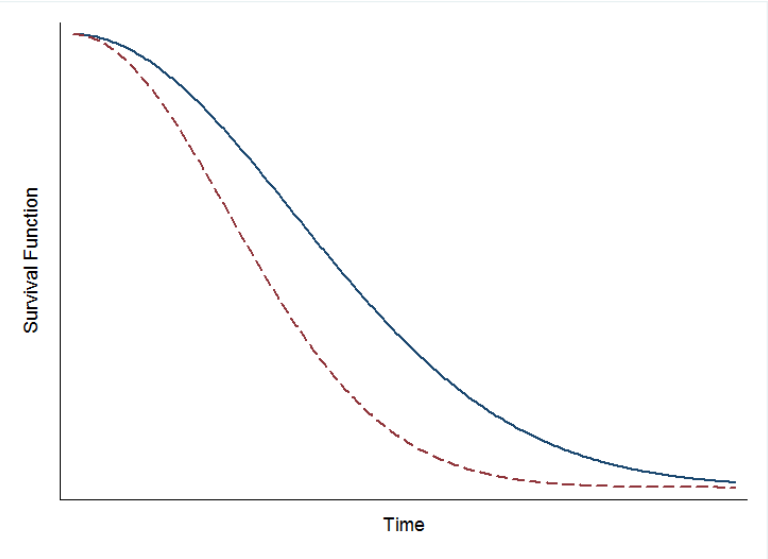


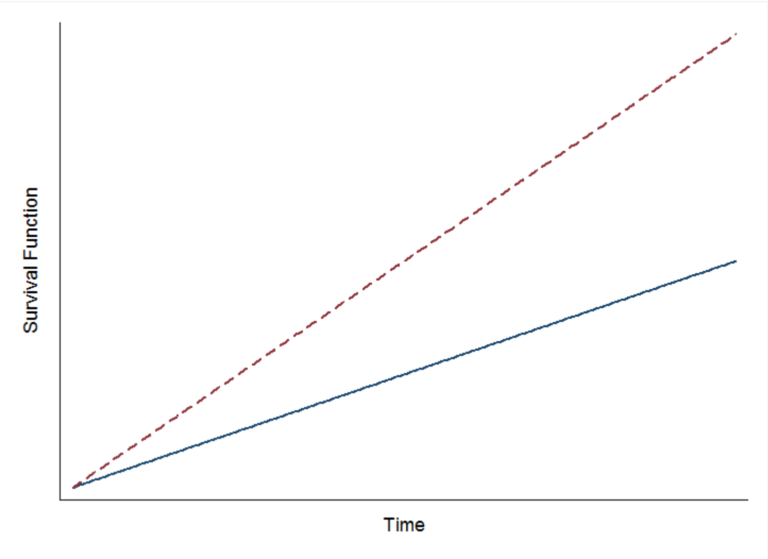


PARAMETRIC MODELS

- Control group: Weibull(0.5,2)
- Example
- Survival Functions

Hazard Functions

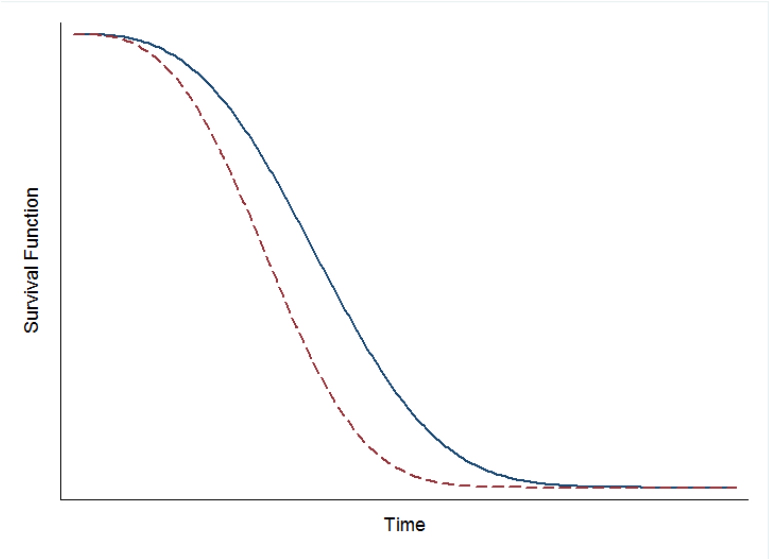


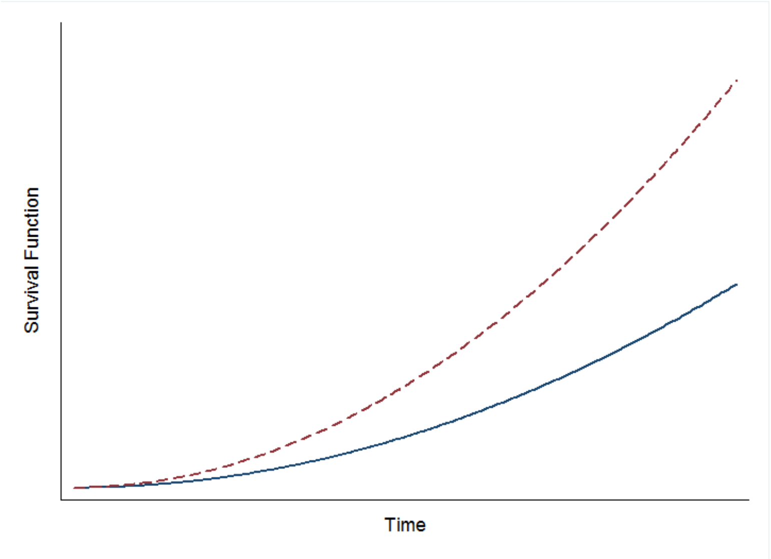


PARAMETRIC MODELS

- Control group: Weibull(0.5,3)
- Example
- Survival Functions

Hazard Functions



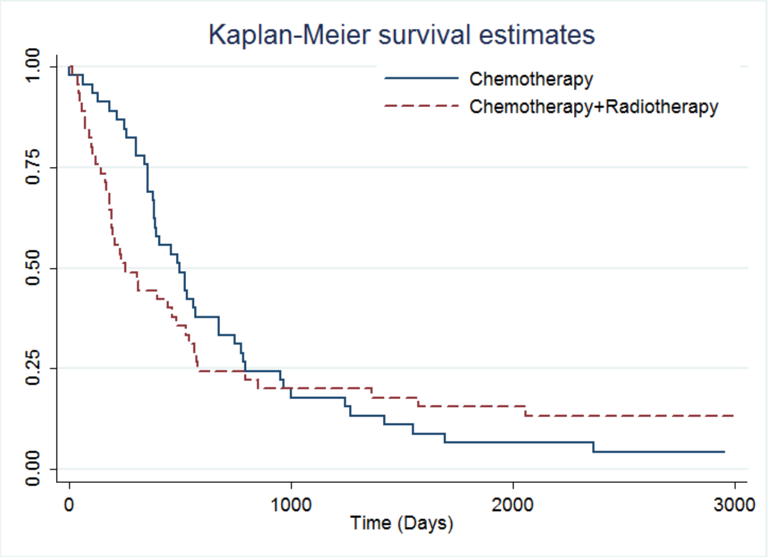


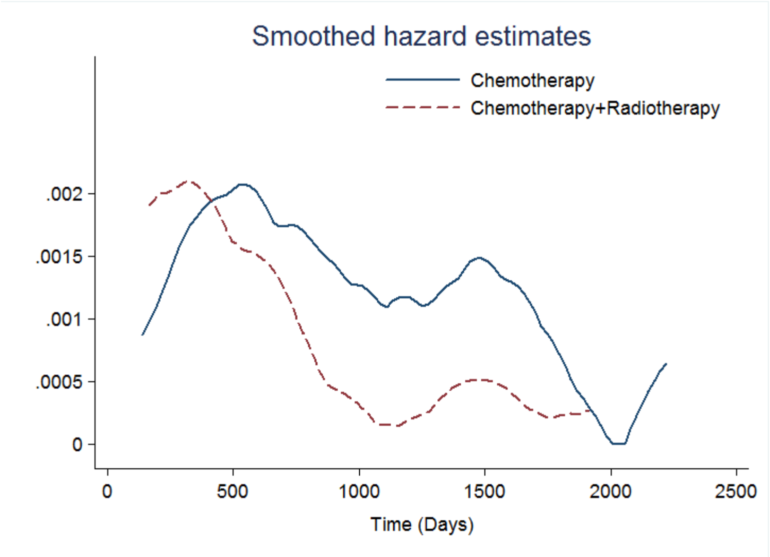
PARAMETRIC APPROACHES

- Weibull and exponential
  - Proportional hazards assumption
  - Distributional assumptions
- Be careful with interpretation of parameter estimates when working with these models.

BACK TO EXAMPLE 3

- Gastrointestinal Tumor Study
- Survival Functions Hazard Functions





- Other covariates

EXAMPLE 1: COLON CANCER – REVISITED

- Tumor differentiation and survival

Group	Observed Events	Expected Events
Well	42	47.5
Moderate	311	334.9
Poor	88	58.6
	441	441

- $\chi(2) = 17.2$,
- p – value = 0.0002

EXAMPLE 1 REVISITED

- Tumor differentiation by treatment group

Groups	Obs	Lev	Lev+5FU	Total
Well	27	37	29	93
Moderate	229	219	215	663
Poor	52	44	54	150
Total	308	300	298	906

STRATIFIED LOG-RANK TEST

- Assume *R* strata (*r* = 1,…,*R*)
- Recall (non-stratified) log-rank test statistic

$$Q = \frac{\left[\sum_{j=1}^{m}\left(d_{1(j)} - \hat{E}_{1(j)}\right)\right]^2}{\sum_{j=1}^{m}\hat{V}_{(j)}}$$

- Stratified log-rank test

$$Q = \frac{\left[\sum_{j_1=1}^{m_1}\left(d_{1,1(j)} - \hat{E}_{1,1(j)}\right) + \ldots + \sum_{j_r=1}^{m_r}\left(d_{1r(j)} - \hat{E}_{1r(j)}\right) + \ldots + \sum_{j_R=1}^{m_R}\left(d_{1R(j)} - \hat{E}_{1R(j)}\right)\right]^2}{\sum_{j_1=1}^{m_1}\hat{V}_{1(j)} + \ldots + \sum_{j_r=1}^{m_r}\hat{V}_{r(j)} + \ldots + \sum_{j_R=1}^{m_R}\hat{V}_{R(j)}}$$

STRATIFIED LOG-RANK TEST

- H_0: $\lambda_{1r}(t) = \lambda_{2r}(t)$ for all $r = 1,\ldots,R$
- H_A: $\lambda_{1r}(t) = c\lambda_{2r}(t), c \neq 1$ for all $r = 1,\ldots,R$
- Under H_0 test statistic ~ $\chi^2(K-1)$
- The $d_{1r(j)}, \hat{E}_{1r(j)}$ and $\hat{V}_{r(j)}$ are solely based on subjects from the *r*-th strata

STRATIFIED LOG-RANK TEST

Well differentiated	**Observed Events**	**Expected Events**
Obs	18	16.7
Lev	16	10.6
Lev+5FU	8	14.7
	42	42

Moderately differentiated	**Observed Events**	**Expected Events**
Obs	109	98.7
Lev	115	105.4
Lev+5FU	87	106.9
	311	311.0

STRATIFIED LOG-RANK TEST

Poorly differentiated	Observed Events	Expected Events
Obs	27	24.8
Lev	34	30.5
Lev+5FU	27	32.7
	88	88.0

Combined over differentiation strata	Observed Events	Expected Events
Obs	154	140.1
Lev	165	146.5
Lev+5FU	122	154.4
	441	441.0

- $\chi(2) = 10.5$
- P-value: 0.005

COMPARISON STRATA VS NO STRATA

- $\chi(2) = 10.5$
- P-value: 0.005

Combined over differentiation strata	Observed Events	Expected Events
Obs	154	140.1
Lev	165	146.5
Lev+5FU	122	154.4
	441	441.0

- $\chi(2) = 11.7$
- P-value: 0.003

Without strata	Observed Events	Expected Events
Obs	161	146.1
Lev	168	148.4
Lev+5FU	123	157.5
	452	452

COMPARISON STRATA VS NO STRATA

- Why are the observed and expected different?

COMPARISON STRATA VS NO STRATA

- Why are the observed and expected different?

- Answer: There are 23 individuals with missing differentiation level (11 of whom experienced the event)
  - Not a “fair” comparison

(FAIR) COMPARISON STRATA VS NO STRATA

- $\chi(2) = 10.5$
- P-value: 0.005

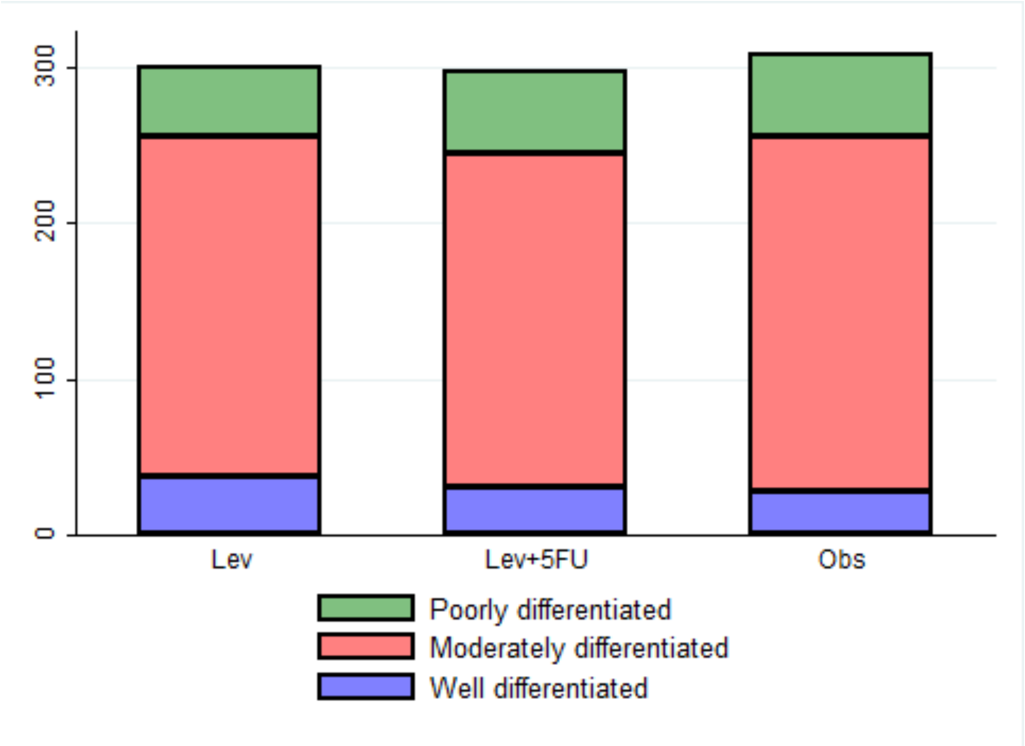
Combined over differentiation strata	Observed Events	Expected Events
Obs	154	140.1
Lev	165	146.5
Lev+5FU	122	154.4
	441	441.0

- $\chi(2) = 10.6$
- P-value: 0.005

Without strata	Observed Events	Expected Events
Obs	154	141.4
Lev	165	145.3
Lev+5FU	122	154.3
	441	441.0

DIFFERENTIATION BY TREATMENT GROUP

- Randomization worked



- Example with more strata

MORE STRATA - EXAMPLE 5

- Van Belle et al (Biostatistics, 2nd Edition)
- Based on Passamani et al (1982)
- Patients with chest pain
- Studied for possible coronary artery disease
  - Definitely angina
  - Probably angina
  - Probably not angina
  - Definitely not angina
- Physician diagnosis
- Outcome: Survival

30 STRATA

	# of prox. vessels			
# vessels	0	1	2	3
0	5-11			
0	12-16			
0	17-30			
1	5-11	5-11		
1	12-16	12-16		
1	17-30	17-30		
2	5-11	5-11	5-11	
2	12-16	12-16	12-16	
2	17-30	17-30	17-30	
3	5-11	5-11	5-11	5-11
3	12-16	12-16	12-16	12-16
3	17-30	17-30	17-30	17-30

Left
Ventricular
Score

30 STRATA

- Chi^2 (3) = 1.47
- P – value = 0.69

- Comparing 4 groups across 30 strata
- Adjusting for these strata showed initial findings of differences between groups may have been due to confounding.

SUMMARY

- Two sample tests
- Different flavors (weighted) two sample tests
- K – sample test
- Trend test
- Stratified test

TO WATCH OUT FOR:

- Only ranks are used for “standard” tests
- Observations with time = 0
- Crossing survival functions
- Independent censoring
- Clinical relevance
  - Log rank test and Cox
  - A difference between 3 and 6 days is judged the same as a difference between 3 years and 6 years

- Questions ?