MODULE 13: SURVIVAL ANALYSIS FOR CLINICAL TRIALS

Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Susanne May, Ph.D.
Barbara McKnight, Ph.D.
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 based on functionals and metrics
- Session 4
  - Choice of outcome variable
  - Surrogate endpoints
  - Power and sample size
  - Information accrual under sequential monitoring

SESSION 1:
REVIEW, COX MODEL FOR ADJUSTMENT AND INTERACTION, AND ESTIMATION OF BASELINE HAZARDS AND SURVIVAL

Module 13: Survival Analysis in Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Barbara McKnight, Ph.D.
Professor
Department of Biostatistics
University of Washington

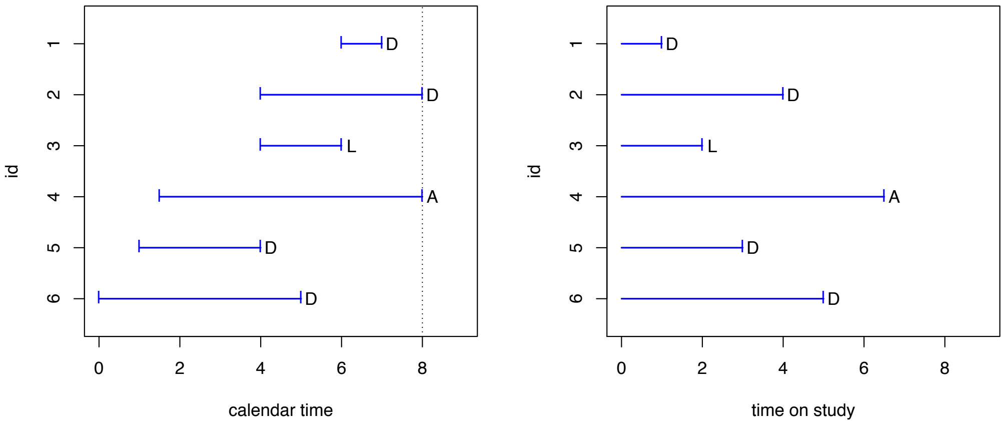
OUTLINE

- Review of censored data, KM estimation, logrank test and Cox model basics
- Covariate adjustment in Cox model
- Precision in Cox model
- Interaction (Effect Modification) in Cox Model
- Stratification adjustment in Cox model
- Estimation of baseline hazards and survival based on Cox model fit

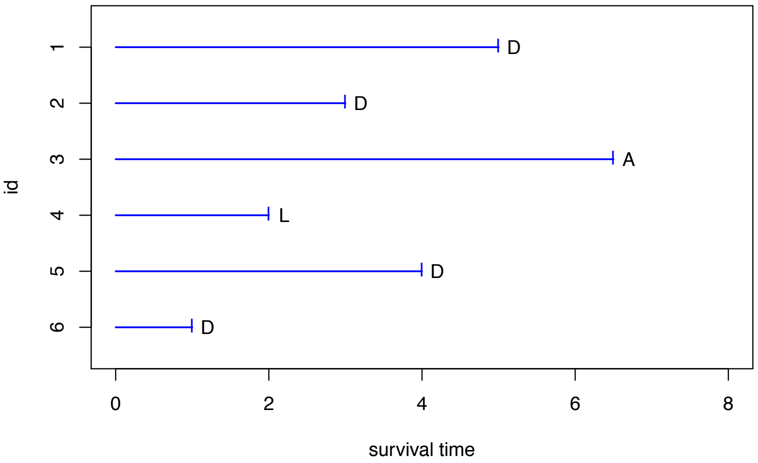
OUTLINE

- **Review of censored data, KM estimation, logrank test and Cox model basics**
- Covariate adjustment in Cox model
- Precision in Cox model
- Interaction (Effect Modification) in Cox Model
- Stratification adjustment in Cox model
- Estimation of baseline hazards and survival based on Cox model fit

TIME IN A CLINICAL TRIAL



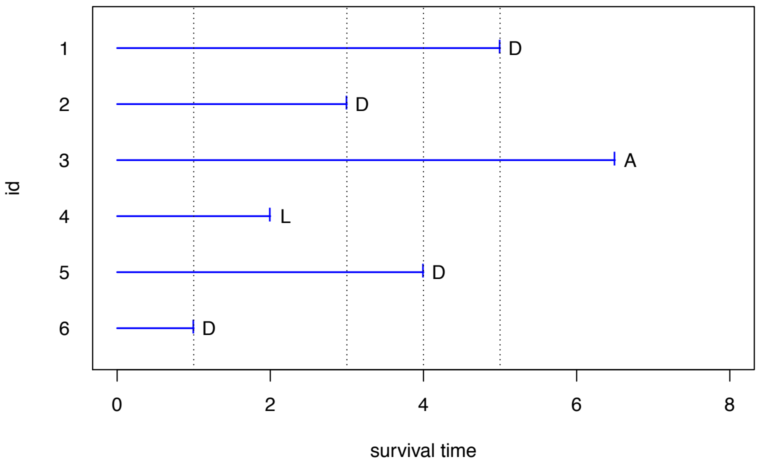
CENSORED DATA



id	Y	δ
1	5	1
2	3	1
3	6.5	0
4	2	0
5	4	1
6	1	1

“Censored” observations give some information about their survival time.

RISK SETS

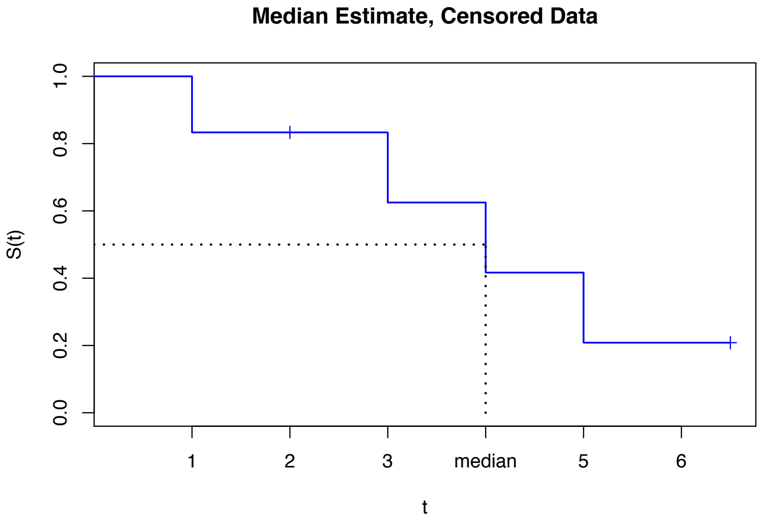


R_1	R_2	R_3	R_4
{1,2,3,4,5,6}	{1,2,3,5}	{1,3,5}	{1,3}

CENSORED DATA ASSUMPTION

- Important assumption: subjects who are censored at time t are at the same risk of dying at t as those at risk but not censored at time t.

MEDIAN & SURVIVAL CENSORED DATA



EQUIVALENT CHARACTERIZATIONS

- Any one of the density function(f(t)), the survival function(S(t)) or the hazard function(λ(t)) is enough to determine the survival distribution.
- They are each functions of each other:

  - $S(t) = \int_t^\infty f(s)ds = e^{-\int_0^t \lambda(s)ds}$
  - $f(t) = -\frac{d}{dt}S(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}$
  - $\lambda(t) = \frac{f(t)}{S(t)}$

LOGRANK TEST

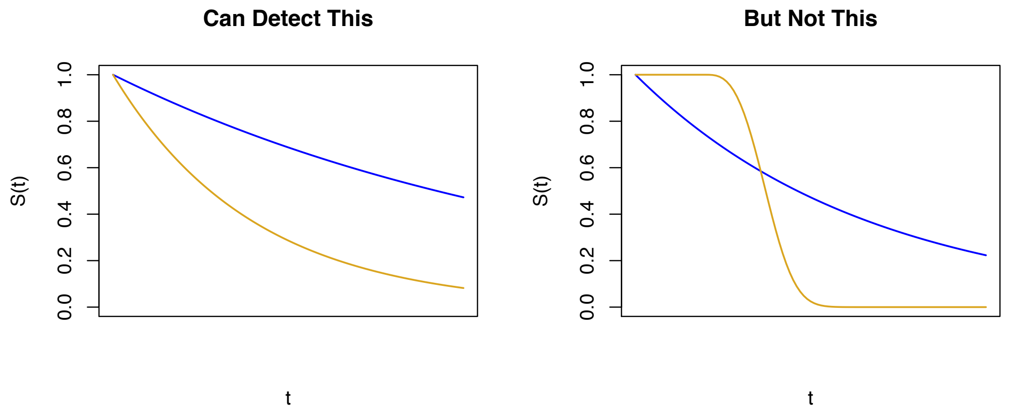
- The test is based on a 2x2 table of group by current status at each observed failure time (ie for each risk set)
- $T_{(j)}$, *j=1,…m*, as shown in the Table below.

Event/Group	1	2	Total
Die	$d_{1(j)}$	$d_{2(j)}$	$D_{(j)}$
Survive	$n_{1(j)}$-$d_{1(j)}$= $s_{1(j)}$	$n_{2(j)}$-$d_{2(j)}$ = $s_{2(j)}$	$N_{(j)}$-$D_{(j)}$ = $S_{(j)}$
At Risk	$n_{1(j)}$	$n_{2(j)}$	$N_{(j)}$

LOGRANK TEST

- Detects consistent differences between survival curves over time.
- Best power when:
  - H_0: $S_1(t) = S_2(t)$ for all t vs H_A: $S_1(t) = [S_2(t)]^c$, or
  - H_0: $\lambda_1(t) = \lambda_2(t)$ for all t vs H_A: $\lambda_1(t) = c\,\lambda_2(t)$
- Good power whenever hazard function ratio is on consistent side of one.

LOGRANK TEST



Other tests (generalized Wilcoxon and others) can give more weight to early or late differences.

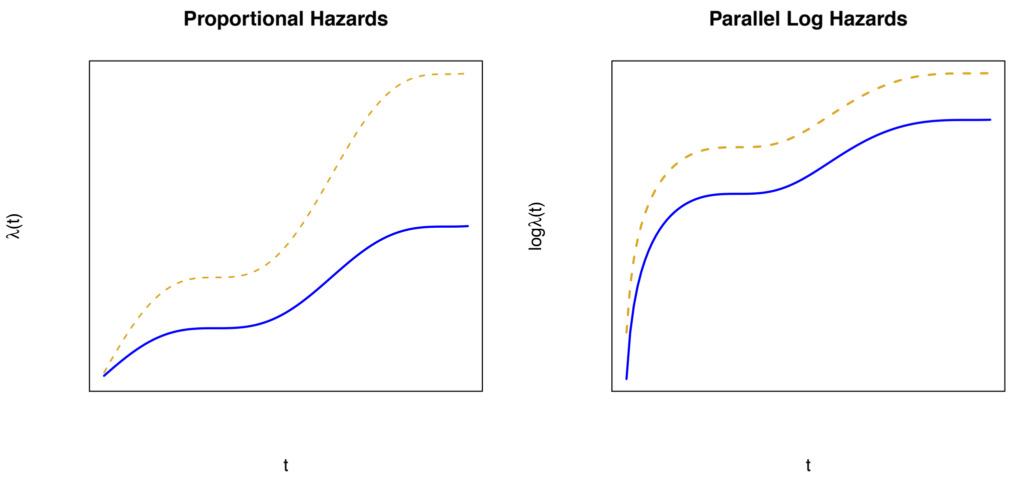
COX REGRESSION MODEL

- Usually written in terms of the hazard function
- As a function of independent variables $x_1, x_2, \ldots x_k$,

$$\lambda(t) = \lambda_0(t)\underset{\underset{\text{relative risk / hazard ratio}}{\uparrow}}{e^{\beta_1 x_1 + \cdots + \beta_k x_k}}$$

$$\log\lambda(t) = \underset{\underset{\text{intercept}}{\uparrow}}{\log\lambda_0(t)} + \beta_1 x_1 + \cdots + \beta_k x_k$$

EXAMPLE



RELATIONSHIP TO SURVIVAL FUNCTION

Single binary x:

$$x = \begin{cases} 1 & \text{Test treatment} \\ 0 & \text{Standard treatment} \end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta x} \implies S(t) = [S_0(t)]^{e^{\beta x}}$$

In terms of $S_0(t)$:

$S(t)$ for $x = 1$: $[S_0(t)]^{e^{\beta \cdot 1}} = [S_0(t)]^{e^{\beta}}$

$S(t)$ for $x = 0$: $[S_0(t)]^{e^{\beta \cdot 0}} = [S_0(t)]^{1} = S_0(t)$

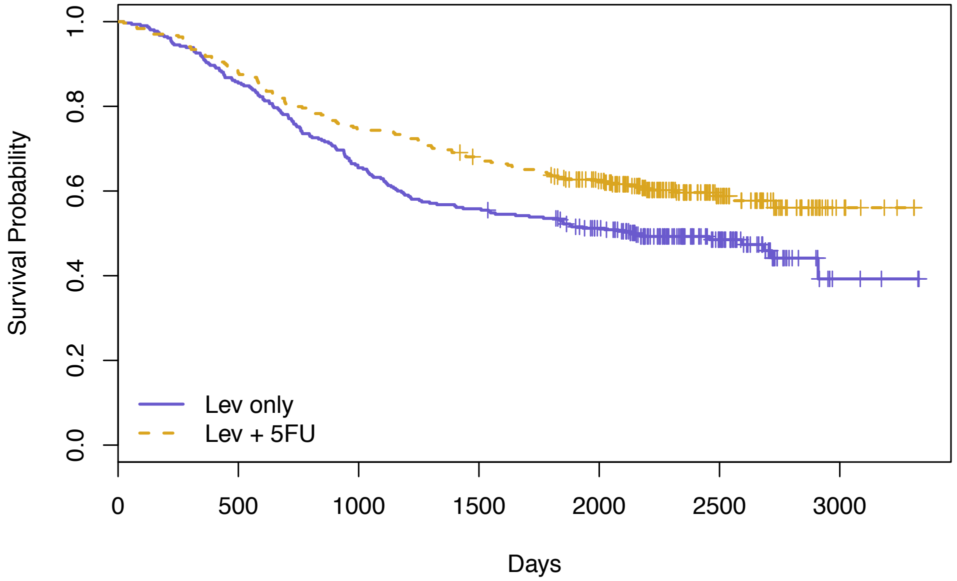
CONFOUNDING/PRECISION

- Because of randomization not truly a problem, but imbalance may be an issue , especially in small trials.
- As in linear regression, regression models for censored survival data allow group comparisons among subjects with similar values of adjustment or “precision” variables (more later).
- Fairer and more powerful comparison as long as adjustment variables are not the result of treatment.

COLON CANCER EXAMPLE

- Levamisole and Fluorouracil for adjuvant therapy of resected colon carcinoma
  - Moertel et al. *New England Journal of Medicine*. 1990;322(6):352–358.
  - Moertel et al. *Annals of internal medicine*. 1995;122(5):321–326.
- 1296 patients
- Stage B_2 or C
- 3 unblinded treatment groups
  - Observation only
  - Levamisole (oral, 1yr)
  - Levamisole (oral, 1yr) + 5 fluorouracil (intravenous 1yr)
- Will examine two treatment arms in Stage C patients only

COLON CANCER EXAMPLE



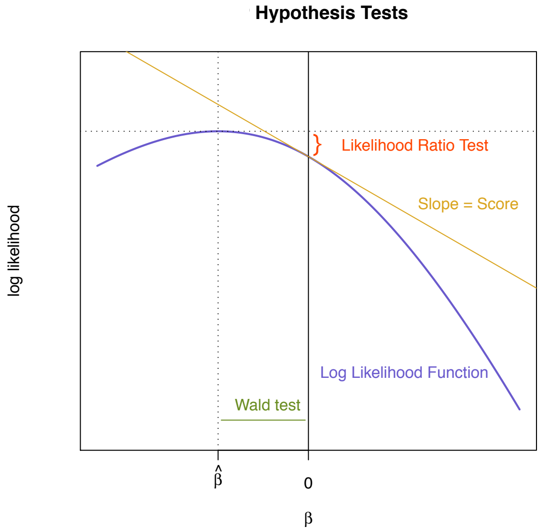
COLON CANCER EXAMPLE

Variable	n	Deaths	Hazard ratio	CI	P-value
Levamisole Only	310	161	1.0 (reference)	--	--
Levamisole + 5FU	304	123	0.71	(0.56, 0.90)	.004

Q: Which group has better survival?

A:

LIKELIHOODS AND TESTS



TEST COMPARISON

Test	Statistic	P-value
Wald's	8.13	.004
Score	8.21	.004
Likelihood Ratio	8.21	.004

Two-sided tests

OUTLINE

- Review of censored data, KM estimation, logrank test and Cox model basics
- **Covariate adjustment in Cox model**
- **Precision in Cox model**
- Interaction (Effect Modification) in Cox Model
- Stratification adjustment in Cox model
- Estimation of baseline hazards and survival based on Cox model fit

STRATIFIED RANDOMIZATION

- For strong predictors: concern about possible randomization imbalance
  - Clinic or center
  - Stage of disease
  - Sex
  - Age
- Adjust for stratification variables in analysis
  - More powerful if predictors are strong
  - Same conditioning as the sampling

ADJUSTMENT AND PRECISION

- In Cox regression, addition of variables to a model that are associated only with the outcome can improve power.
- There is little effect on the coefficient estimate for other variables (eg treatment) or their standard errors, except when the association between outcome and the added variable is very strong.
- When there is an effect of adding a predictive variable, this is what happens to inference for the treatment variable or other variable of interest:
  - The standard error of its coefficient increases
  - The estimate of the coefficient moves farther from zero
  - The test of whether the coefficient is zero has more power.

ANALYSES

- Primary analysis: If randomization was blocked on prognostic variables, adjust for them.
  - Depth of invasion (extent)
  - Interval since surgery
  - Number of positive nodes (≥ 4)

- Secondary analysis: Adjust for additional prognostic variables: Observed at time of randomization and therefore not affected by treatment
  - Obstruction
  - Histologic differentiation

PROGNOSTIC VARIABLE ADJUSTMENT

$$x_1 = \begin{cases} 1 & \text{moderate differentiation} \\ 0 & \text{otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{poor differentiation} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{tumor obstructed bowel} \\ 0 & \text{otherwise} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{4+ nodes positive} \\ 0 & \text{otherwise} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{extent to muscle} \\ 0 & \text{otherwise} \end{cases} \qquad x_6 = \begin{cases} 1 & \text{extent to serosa} \\ 0 & \text{otherwise} \end{cases}$$

$$x_7 = \begin{cases} 1 & \text{extent to contiguous structures} \\ 0 & \text{otherwise} \end{cases} \qquad x_8 = \begin{cases} 1 & \text{Levamisole only} \\ 0 & \text{otherwise} \end{cases}$$

$$x_9 = \begin{cases} 1 & \text{Levamisole + 5FU} \\ 0 & \text{otherwise} \end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9}$$

PROGNOSTIC VARIABLE ADJUSTMENT

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\beta_7 x_7+\beta_8 x_8+\beta_9 x_9}$$

Interpretation of e^{β_8}:

"Relative risk (or hazard ratio) comparing Levamisole Only to Observation among those with the same values of prognostic variables".

Interpretation of e^{β_9}:

"Relative risk (or hazard ratio) comparing Levamisole + 5FU to Observation among those with the same values of prognostic variables".

PROGNOSTIC VARIABLE ADJUSTMENT

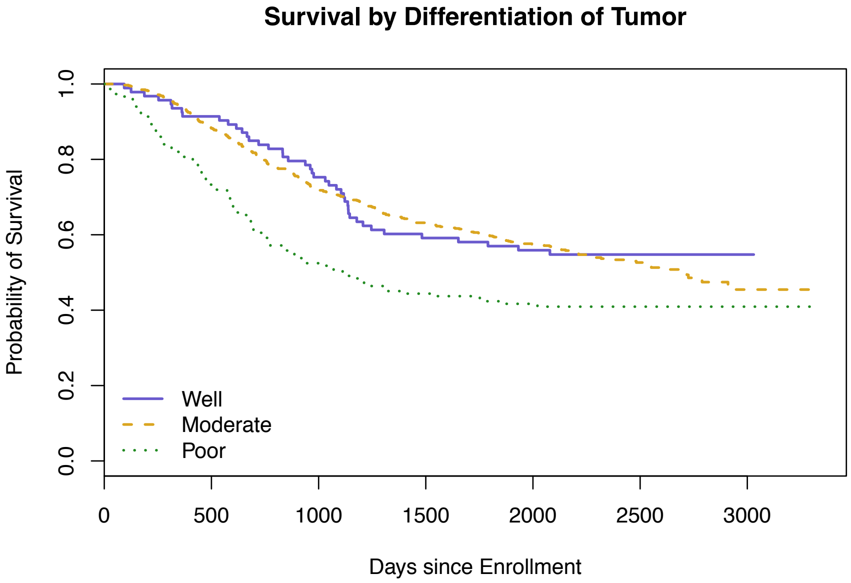
$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\beta_7 x_7+\beta_8 x_8+\beta_9 x_9}$$

Interpretation of $e^{\beta_9-\beta_8}$:

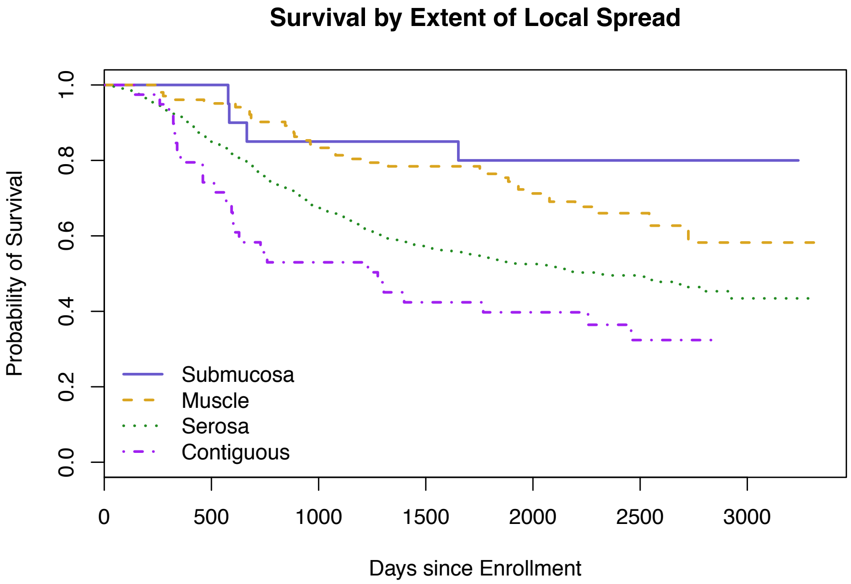
"Relative risk (or hazard ratio) comparing Levamisole + 5FU to Levamisole Only among those with the same values of prognostic variables".

$$\begin{aligned}
\lambda(t) \text{ for } x_1,\ldots,x_7 \text{ and } x_8 = 0 \text{ and } x_9 = 1: &\quad \lambda_0(t)e^{\beta_1 x_1+\cdots+\beta_7 x_7+\beta_8\cdot 0+\beta_9\cdot 1}\\
\lambda(t) \text{ for } x_1,\ldots,x_7 \text{ and } x_8 = 1 \text{ and } x_9 = 0: &\quad \lambda_0(t)e^{\beta_1 x_1+\cdots+\beta_7 x_7+\beta_8\cdot 1+\beta_9\cdot 0}\\
\text{ratio:} &\quad e^{\beta_8(0-1)+\beta_9(1-0)} = e^{\beta_9-\beta_8}
\end{aligned}$$

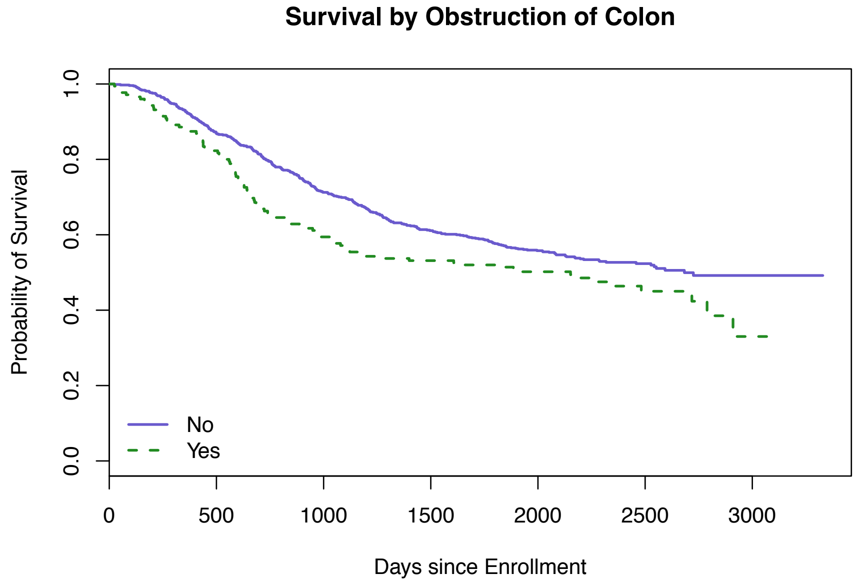
PROGNOSTIC VARIABLES



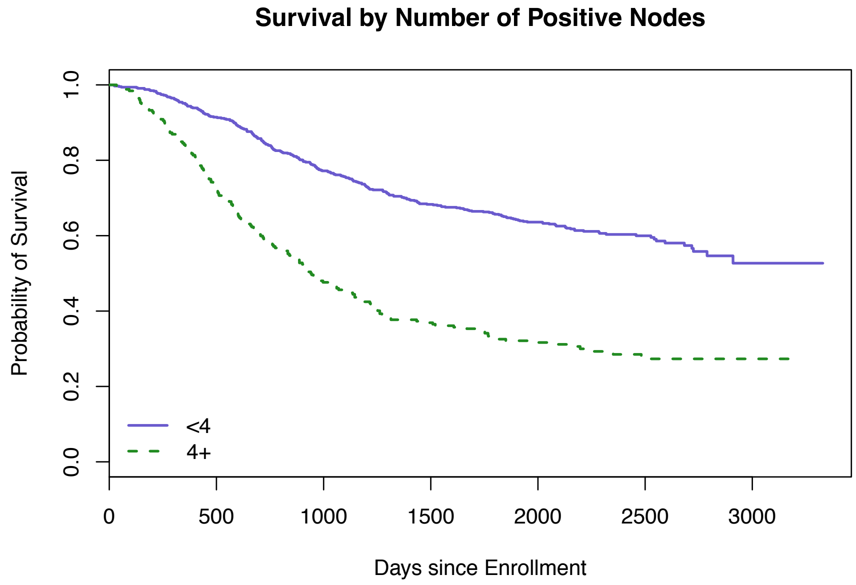
PROGNOSTIC VARIABLES



PROGNOSTIC VARIABLES



PROGNOSTIC VARIABLES



ADJUSTED

Group	Hazard Ratio	95% CI	P-value
Observation Only	1.0 (reference)	--	--
Levamisole Only	0.97	(0.78, 1.21)	0.79
Levamisole + 5FU	0.69	(0.54, 0.87)	0.002

Adjusted for tumor differentiation (well, moderate, poor), colon obstruction (yes, no), < 4 nodes positive, extent (submucosa, muscle, serosa, contiguous tissues)

ADJUSTMENT VARIABLES

Variable	Hazard Ratio	95% CI
Moderate Differentiation	0.94	(0.67, 1.29)
Poor Differentiation	1.38	(0.95, 2.00)
Obstructed bowel	1.30	(1.03, 1.63)
4+ nodes positive	2.45	(2.03, 2.98)
Extent: muscle	1.41	(0.50, 3.99)
Extent: serosa	2.29	(0.85, 6.16)
Extent: contiguous	3.34	(1.15, 9.65)

Usually not presented.

ANOTHER SIMPLER EXAMPLE

Two binary variables, x_1 and x_2 and 2 treatment groups:

$$x_1 = \begin{cases} 1 & \text{Levamisole + 5FU} \\ 0 & \text{Levamisole Only} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{4+ Nodes Positive} \\ 0 & \text{<4 Nodes Positive} \end{cases}$$

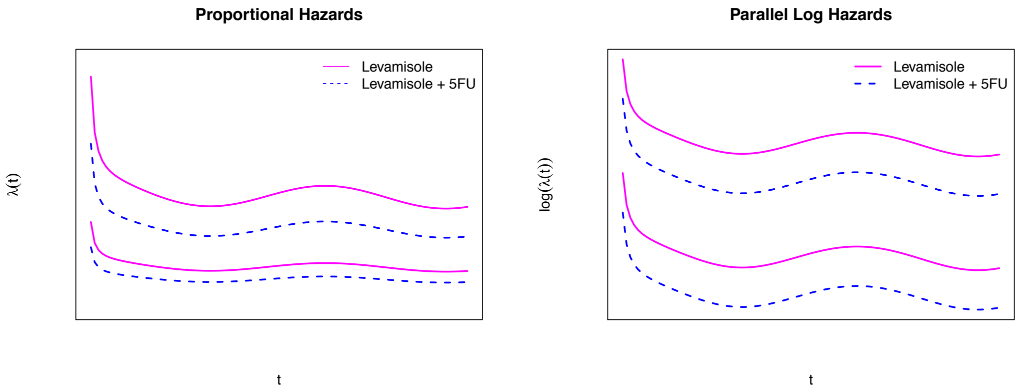
$$\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2}$$

Interpretation of e^{β_1}:

"Relative risk (or hazard ratio) comparing Levamisole + 5FU to Levamisole Only among those with similar numbers of positive nodes".

$$\begin{aligned} \lambda(t) \text{ for } x_1 = 1 \text{ and } x_2: &\quad \lambda_0(t) e^{\beta_1 \cdot 1 + \beta_2 x_2} \\ \lambda(t) \text{ for } x_1 = 0 \text{ and } x_2: &\quad \lambda_0(t) e^{\beta_1 \cdot 0 + \beta_2 x_2} \\ \text{ratio:} &\quad e^{\beta_1(1-0)+\beta_2(x_2-x_2)} = e^{\beta_1} \end{aligned}$$

HEURISTIC HAZARDS

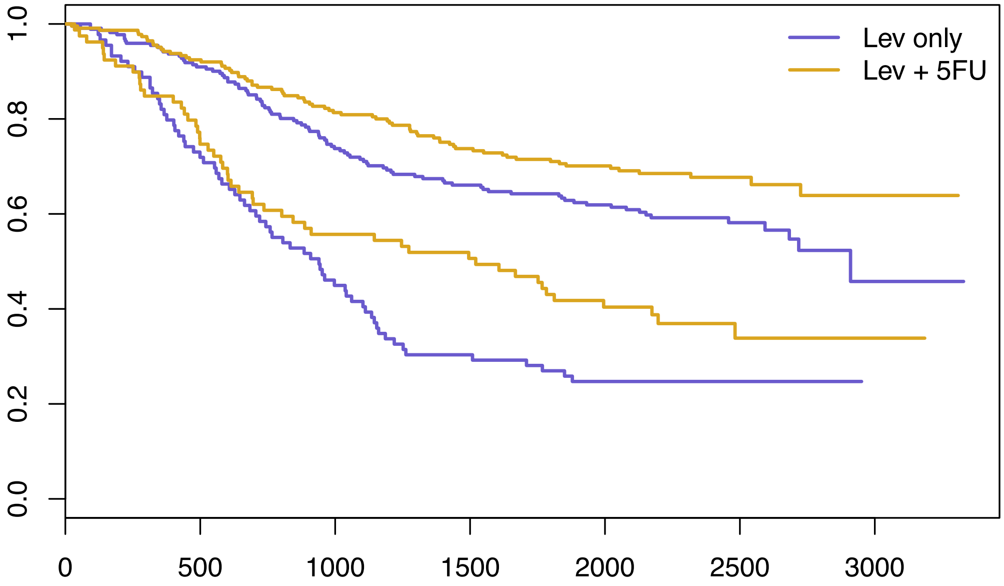


SIMPLER MODEL

Variable	Hazard ratio	95% CI	P-value
Levamisole + FU	0.71	(0.56, 0.90)	0.005
4+ nodes positive	2.67	(2.10, 3.38)	< .0001

Often, second row would not be given, and group sample sizes and numbers of deaths would be presented

COLON CANCER TRIAL DATA



RESULTS

"There was strong evidence that adjuvant treatment with 5FU + Levamisole improves survival in stage C colon cancer patients compared to Levamisole alone. After adjustment for number of positive nodes (<4, 4+) the hazard ratio comparing 5FU + Levamisole to Levamisole was 0.71, (95% CI 0.56 - 0.90, P = .004)."

OUTLINE

- Review of censored data, KM estimation, logrank test and Cox model basics
- Covariate adjustment in Cox model
- Precision in Cox model
- **Interaction (Effect Modification) in Cox Model**
- Stratification adjustment in Cox model
- Estimation of baseline hazards and survival based on Cox model fit

MORE SECONDARY ANALYSES

- Often interested in examining a small number of subgroups to determine subjects especially benefitted by treatment.
- Should be specified in advance!
- Should be few in number.
- Test results are usually corrected for multiple comparisons.
- Should test for interaction, not just notice that the estimated hazard ratios look different.

INTERACTION

Two binary variables, x_1 and x_2 with interaction:

$$x_1 = \begin{cases} 1 & \text{5FU + Levamisole} \\ 0 & \text{Levamisole alone} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{4+ nodes positive} \\ 0 & \text{<4 nodes positive} \end{cases}$$

$$\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$$

Interpretation of e^{β_1}:

HR comparing 5FU + Levamisole to Levamisole only among those with fewer than 4 positive nodes.

Interpretation of $e^{\beta_1 + \beta_3}$:

HR comparing 5FU + Levamisole to Levamisole only among those with at least 4 positive nodes.

WITH INTERACTION

Two binary variables, x_1 and x_2 with interaction:

$$x_1 = \begin{cases} 1 & \text{5FU + Levamisole} \\ 0 & \text{Levamisole alone} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{4+ nodes positive} \\ 0 & \text{<4 nodes positive} \end{cases}$$

$$\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$$

$\lambda(t)$ for $x_1 = 1$ and $x_2 = 0$: $\lambda_0(t)e^{\beta_1 \cdot 1}$ $\quad$ $\lambda(t)$ for $x_1 = 1$ and $x_2 = 1$: $\lambda_0(t)e^{\beta_1 \cdot 1 + \beta_2 \cdot 1 + \beta_3 \cdot 1}$

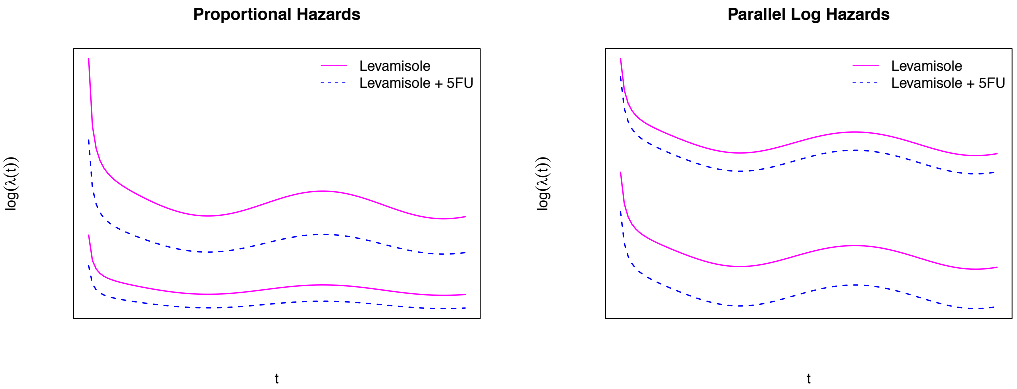
$\lambda(t)$ for $x_1 = 0$ and $x_2 = 0$: $\lambda_0(t)e^{\beta_1 \cdot 0}$ $\quad$ $\lambda(t)$ for $x_1 = 0$ and $x_2 = 1$: $\lambda_0(t)e^{\beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 0}$

ratio: $e^{\beta_1(1-0)} = e^{\beta_1}$ $\quad$ ratio: $e^{\beta_1(1-0) + \beta_3(1-0)} = e^{\beta_1 + \beta_3}$

PRESENTATION

- Usually we present hazard ratios at different values of the interacting/effect modifying variable with CIs and results of a test for interaction.
- Interaction term coefficient β or e^{β} usually not of primary interest.
- In previous example:
  - Treatment HR when <4 nodes positive: $e^{\beta 1}$
  - Treatment HR when 4+ nodes positive: $e^{\beta 1 + \beta 3}$

HEURISTIC HAZARDS



RESULTS

	HR (5FU + Lev/Lev)	95% CI	P-value
< 4 nodes positive	0.72	(0.53, 0.97)	0.03221
4+ notes positive	0.71	(0.49, 1.02)	0.06368
Test for interaction			0.95726

RESULTS

- “We did not find evidence that the hazard ratio associated with treatment differed depending on whether the patient had four or more positive nodes. (P = .96).”

OUTLINE

- Review of censored data, KM estimation, logrank test and Cox model basics
- Covariate adjustment in Cox model
- Precision in Cox model
- Interaction (Effect Modification) in Cox Model
- **Stratification adjustment in Cox model**
- Estimation of baseline hazards and survival based on Cox model fit

RISK SET STRATIFICATION

There are two ways to adjust for a binary (or other categorical) variable:

$$x_1 = \begin{cases} 1 & \text{Levamisole + 5FU} \\ 0 & \text{Levamisole Only} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{4+ Positive Nodes} \\ 0 & \text{<4 Positive Nodes} \end{cases}$$

Dummy variable stratification:

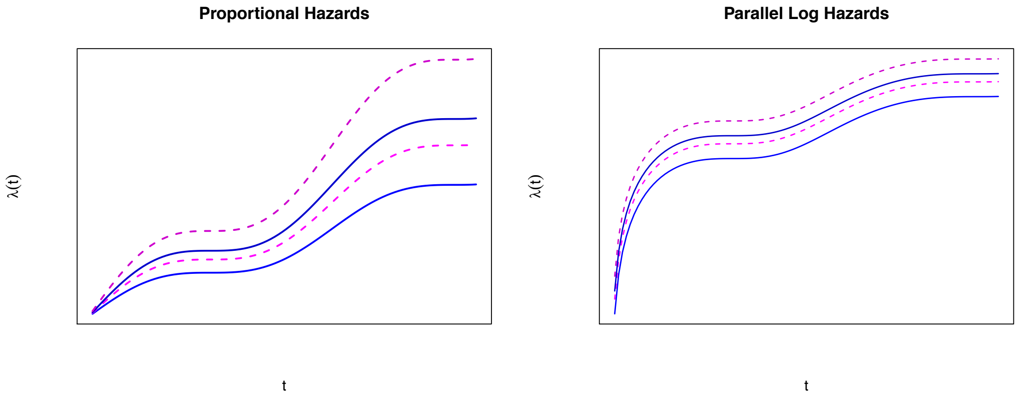
$$\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2}$$

True stratification:

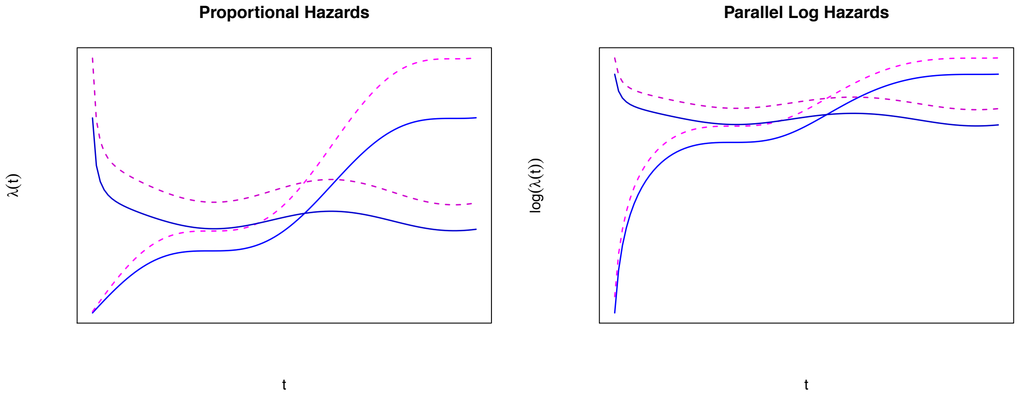
$$\lambda(t) = \lambda_{0x_2}(t) e^{\beta_1 x_1}$$

Stratified logrank test ≈ score test of $H_0 : \beta_1 = 0$ in true stratification model.

DUMMY VARIABLE STRATIFICATION



TRUE STRATIFICATION



RESULTS

"There was strong evidence that adjuvant treatment with 5FU + Levamisole improves survival in stage C colon cancer patients compared to Levamisole alone. After adjustment for number of positive nodes (<4, 4+) the hazard ratio comparing 5FU + Levamisole to Levamisole was 0.72, (95% CI: 0.57 - 0.91) P=0.005."

Very similar to covariate adjustment.

ADDING INTERACTION

Can include interaction for variable with true stratification:

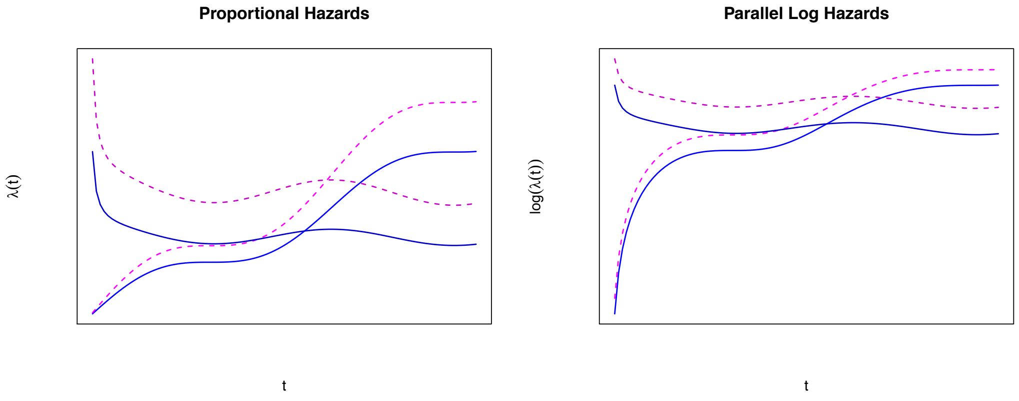
$$x_1 = \begin{cases} 1 & \text{Test treatment} \\ 0 & \text{Standard treatment} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{Failed prior treatment} \\ 0 & \text{No prior treatment} \end{cases}$$

True stratification with interaction:

$$\lambda(t) = \lambda_{0x_2}(t)e^{\beta_1 x_1 + \beta_2 x_1 x_2}$$

HEURISTIC HAZARDS



INTERACTION AND STRATIFICATION

- The interaction model does not violate rules about including main effects for terms that are part of interactions in a regression model.
- The “main effect” of x_2 is included in the $\lambda_{0x2}(t)$ term.

RESULTS

	HR (5FU + Lev/Lev)	95% CI	P-value
< 4 nodes positive	0.71	(0.53, 0.97)	0.03076
4+ notes positive	0.72	(0.5, 1.04)	0.07969
Test for interaction			0.97371

Very similar to covariate node4 model.

OUTLINE

- Review of censored data, KM estimation, logrank test and Cox model basics
- Covariate adjustment in Cox model
- Precision in Cox model
- Interaction (Effect Modification) in Cox Model
- Stratification adjustment in Cox model
- **Estimation of baseline hazards and survival based on Cox model fit**

ESTIMATING THE FUNCTIONS

- After fitting the Cox model,

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

  we may be interested in estimating

  - hazard: $\lambda(t)$
  - cumulative hazard: $\Lambda(t)$ and
  - survival function: $S(t)$

  at values of x, consistent with the model.
- Can be done by estimating baseline versions of these:

  $\lambda_0(t), \Lambda_0(t)$, and $S_0(t)$,

  and multiplying by $e^{\hat{\beta}x}$.

BASELINE CUMULATIVE HAZARD

$$\hat{\Lambda}_0(t) \quad = \underset{\substack{\uparrow \\ \text{observed} \\ \text{failure times}}}{\sum_{j:t_{(j)}\leq t}} \frac{D_j}{\underset{\substack{\uparrow \\ \text{risk set}}}{\sum_{i\in R_j}} e^{\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_K x_{Ki}}}$$

- Estimate depends on $\hat{\beta}_1, \ldots, \hat{\beta}_K$.
- Actually makes sense. Consider special cases.

BASELINE CUMULATIVE HAZARD

$$\hat{\Lambda}_0(t) \quad = \sum_{j:t_{(j)}\leq t} \frac{D_j}{\sum_{i\in R_j} e^{\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_K x_{Ki}}}$$

1. One group, no covariates ($\hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_K x_{Ki} = 0$):

$$\underset{\substack{\uparrow \\ \text{For the single} \\ \text{homogeneous group}}}{\hat{\Lambda}_0(t)} \quad = \sum_{j:t_{(j)}\leq t} \frac{D_j}{\sum_{i\in R_j} 1} \quad = \underset{\substack{\uparrow \\ \text{Standard} \\ \text{Estimator}}}{\sum_{j:t_{(j)}\leq t} \frac{D_j}{N_j}}$$

BASELINE CUMULATIVE HAZARD

$$\hat{\Lambda}_0(t) \quad = \quad \sum_{j:t_{(j)}\le t} \frac{D_j}{\sum_{i\in R_j} e^{\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_K x_{Ki}}}$$

2. Two groups, one binary covariate:

$$x = \begin{cases} 1 & \text{group 2} \\ 0 & \text{group 1} \end{cases}$$

$$\underset{\substack{\uparrow \\ \text{For Group 1}}}{\hat{\Lambda}_0(t)} \quad = \quad \sum\nolimits_{j:t_{(j)}\le t} \frac{D_j}{\sum_{i\in R_j} e^{\hat{\beta}x_i}} \quad = \quad \sum\nolimits_{j:t_{(j)}\le t} \frac{D_j}{\sum_{\substack{i\in R_j \\ \text{Group 1}}} e^{\hat{\beta}x_i}+\sum_{\substack{i\in R_j \\ \text{Group 2}}} e^{\hat{\beta}x_i}}$$

$$= \quad \sum\nolimits_{j:t_{(j)}\le t} \frac{D_j}{\underbrace{n_{1j}+e^{\hat{\beta}}n_{2j}}_{\substack{\text{Effective risk set size} \\ \text{in group 1}}}}$$

BASELINE CUMULATIVE HAZARD

$$\hat{\Lambda}_0(t) \quad = \quad \sum_{j:t_{(j)}\le t} \frac{D_j}{\sum_{i\in R_j} e^{\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_K x_{Ki}}}$$

In general:

The denominator $\sum_{i\in R_j} e^{\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_K x_{Ki}}$ is

- Bigger than N_j when the average risk for a subject in R_j is bigger than the risk for a subject in R_j with $x_{1i} = x_{2i} = \cdots = x_{Ki} = 0$
- Smaller than N_j when the average risk for a subject in R_j is smaller than the risk for a subject in R_j with $x_{1i} = x_{2i} = \cdots = x_{Ki} = 0$

BASELINE CUMULATIVE HAZARD

$$\hat{\Lambda}_0(t) \quad = \sum_{j:t_{(j)}\leq t} \frac{D_j}{n_{1j} + e^{\hat{\beta}} n_{2j}}$$

↑

Group 1

D_j counts deaths in both groups.

$\hat{\beta} > 0 \quad \Longrightarrow$ More deaths in group 2
Effective risk set size must be increased to estimate risk in group 1.

$\hat{\beta} < 0 \quad \Longrightarrow$ More deaths in group 1
Effective risk set size must be decreased to estimate risk in group 1.

COLON CANCER TRIAL DATA

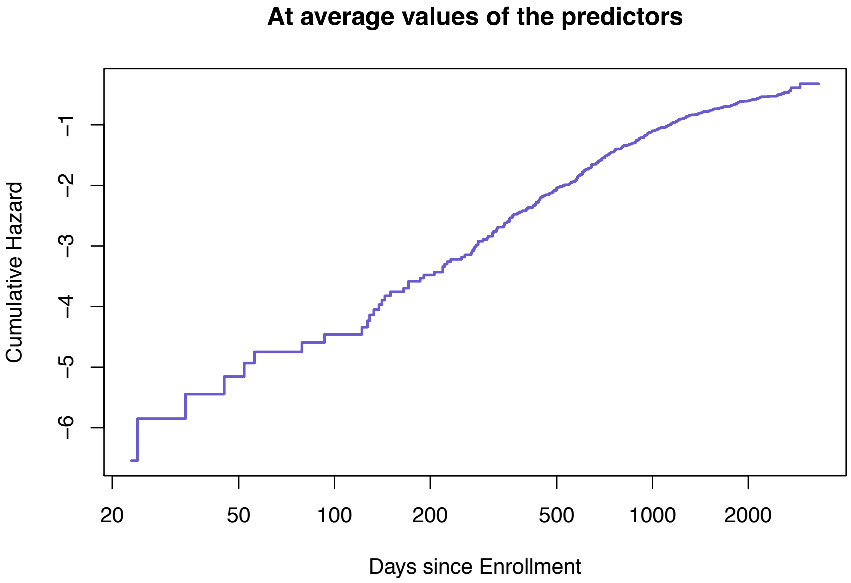
Observation Arm Omitted

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	z	Pr(>|z|)
5FU + Lev	-0.34	0.71	0.12	-2.83	0.0064
4+ Nodes Pos	0.98	2.67	0.12	8.08	<0.0001

$e^{\beta_{Rx}}$ CI: (0.5629, 0.9008)

LRT: 8.098 on 1 df, P = 0.0044

COLON CANCER TRIAL DATA

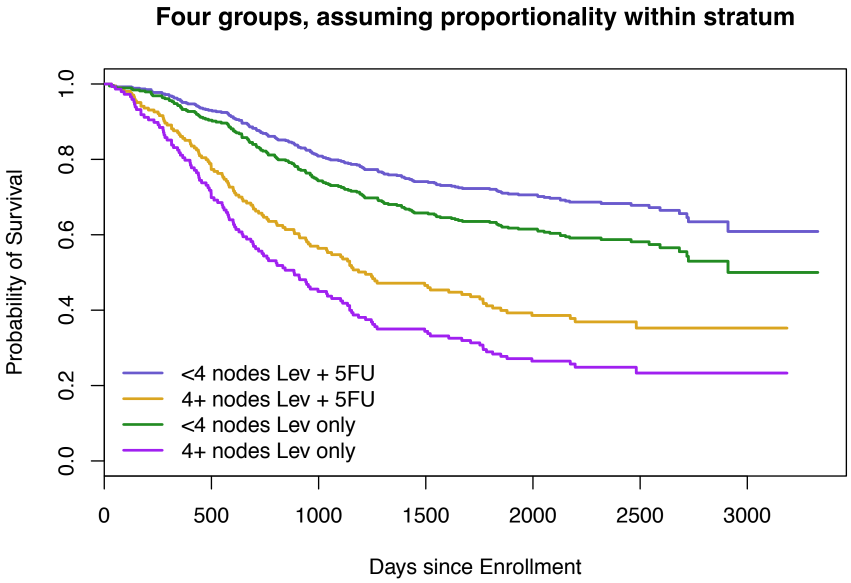

At average values of the predictors

ESTIMATING Λ AND AT COVARIATE VALUES

- Baseline survival function: $\hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)}$ (Since $S(t) = e^{-\Lambda(t)}$).
- At other values:

$$
\begin{aligned}
\hat{\Lambda}(t|x_{1i}, x_{2i}, \ldots, x_{ki}) &= \hat{\Lambda}_0(t) e^{\hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}} \\
\hat{S}(t|x_{1i}, x_{2i}, \ldots, x_{ki}) &= [\hat{S}_0(t)]^{e^{\hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}}}
\end{aligned}
$$

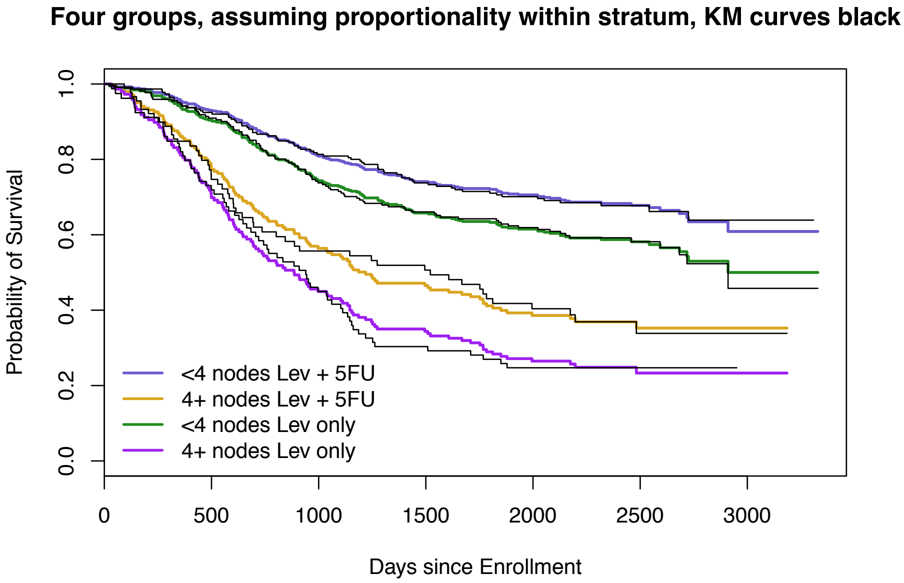
COLON CANCER TRIAL DATA



USES FOR BASELINE AND SPECIFIC-X FUNCTIONS

- To estimate survival for different covariate combinations, according to the model.
- To check the fit of the model, by comparing $\hat{\Lambda}_x(t)$ or $\hat{S}_x(t)$ to $\hat{\Lambda}(t)$ or $\hat{S}(t)$ for groups with like values of $\hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_K x_{Ki}$.
- To check whether hazards in different risk set strata are proportional.

COLON CANCER TRIAL DATA



TO WATCH OUT FOR:

- Coefficients in Cox regression are positively associated with risk, not survival.
  - Positive β means large values of x are associated with shorter survival.
- Without certain types of time-dependent covariates (more later), Cox regression does not depend on the actual times, just their order.
  - Can add a constant to all times to remove zeros (which are removed by some software) without changing inference
- For LRT, nested models must be compared based on same subjects.
  - If some values of variables in larger model are missing, these subjects must be removed from fit of smaller model.
- Coefficient interpretation depends on what other variables are in the model and how they are coded (ie. interaction terms, 0/1 vs 1/-1 etc.)
- Hazards may not be proportional

SESSION 2:
WEIGHTED LOG RANK TESTS

Module 13: Survival Analysis for Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Susanne May, Ph.D.
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

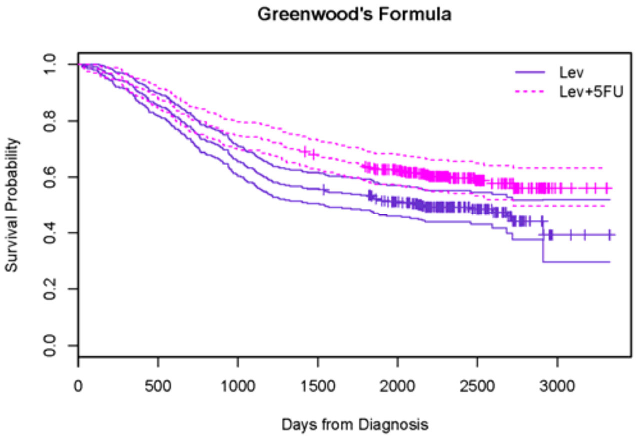
- Group comparisons
  - Two groups
  - k groups
  - Test for (linear) trend
- Assume, H_0 : no differences between groups

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	$n_{1(j)}$	$n_{2(j)}$	$N_{(j)}$

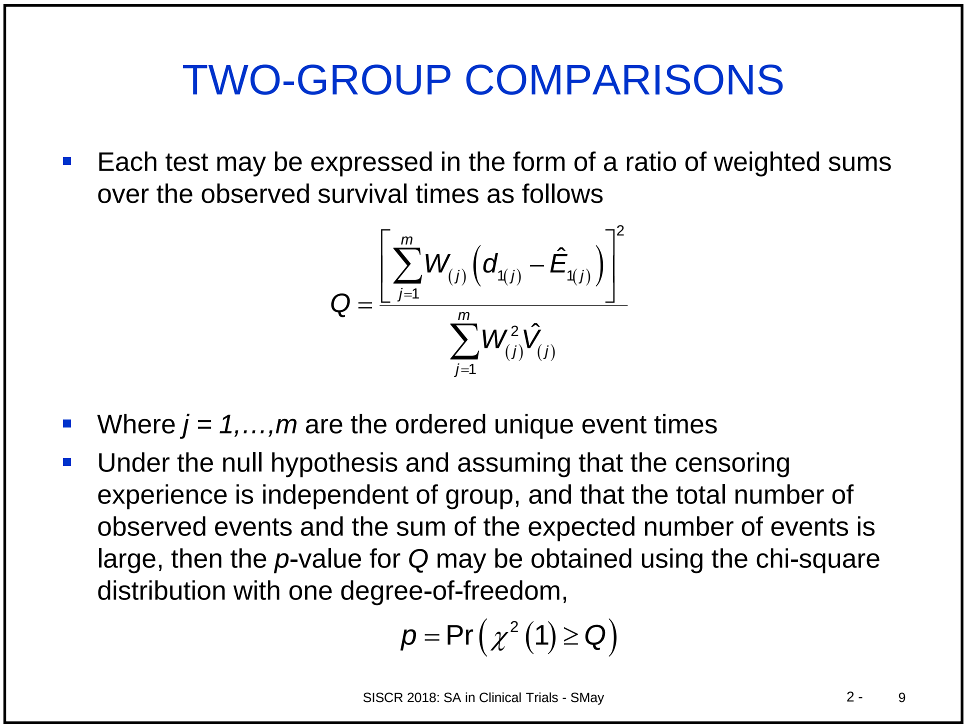
TWO-GROUP COMPARISONS

- 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:

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



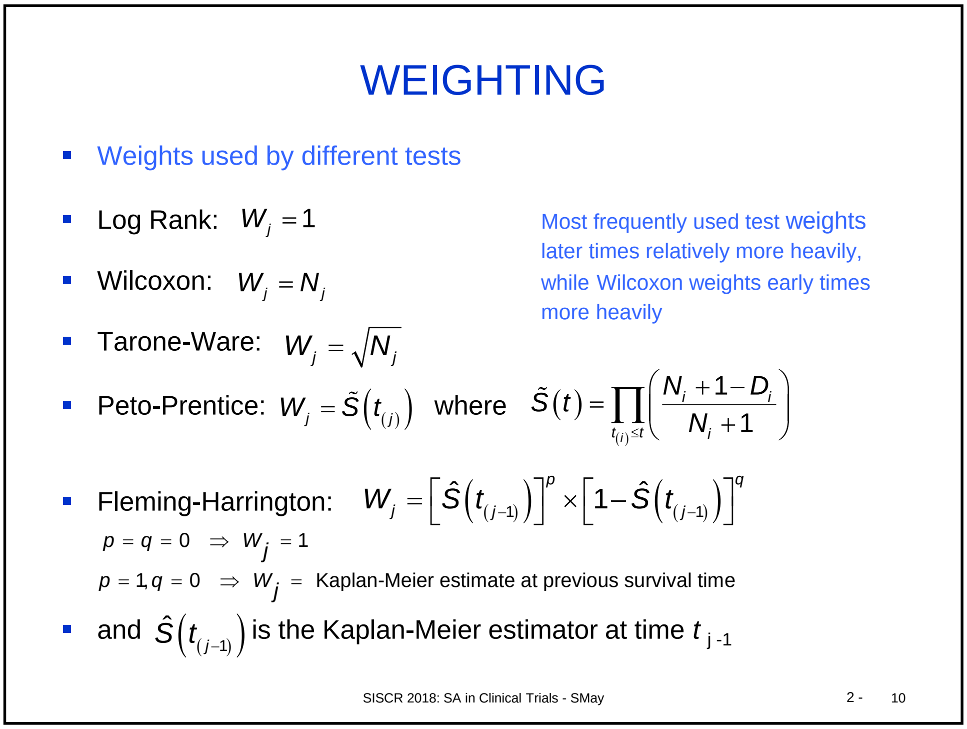
TWO-GROUP COMPARISONS

- 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: $W_j = \sqrt{N_j}$
- Peto-Prentice: $W_j = \tilde{S}\left(t_{(j)}\right)$ where $\tilde{S}(t) = \prod_{t_{(i)} \leq t}\left(\frac{N_i + 1 - D_i}{N_i + 1}\right)$
- Fleming-Harrington: $W_j = \left[\hat{S}\left(t_{(j-1)}\right)\right]^p \times \left[1 - \hat{S}\left(t_{(j-1)}\right)\right]^q$

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

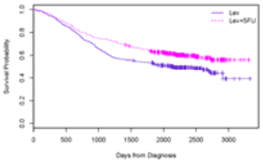
  $p = 1, q = 0 \Rightarrow W_j =$ Kaplan-Meier estimate at previous survival time
- and $\hat{S}\left(t_{(j-1)}\right)$ is the Kaplan-Meier estimator at time $t_{j\text{-}1}$

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

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.

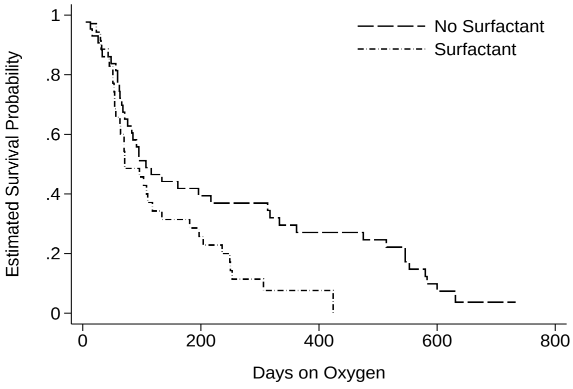
TWO-GROUP COMPARISONS LBWI

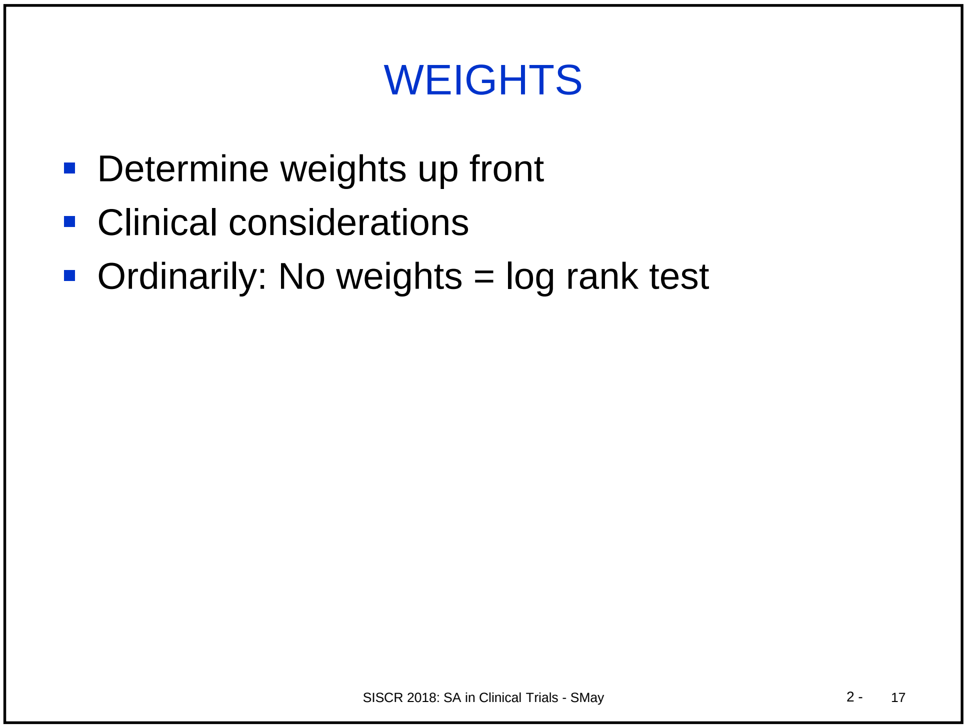
- Different weighting approaches

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

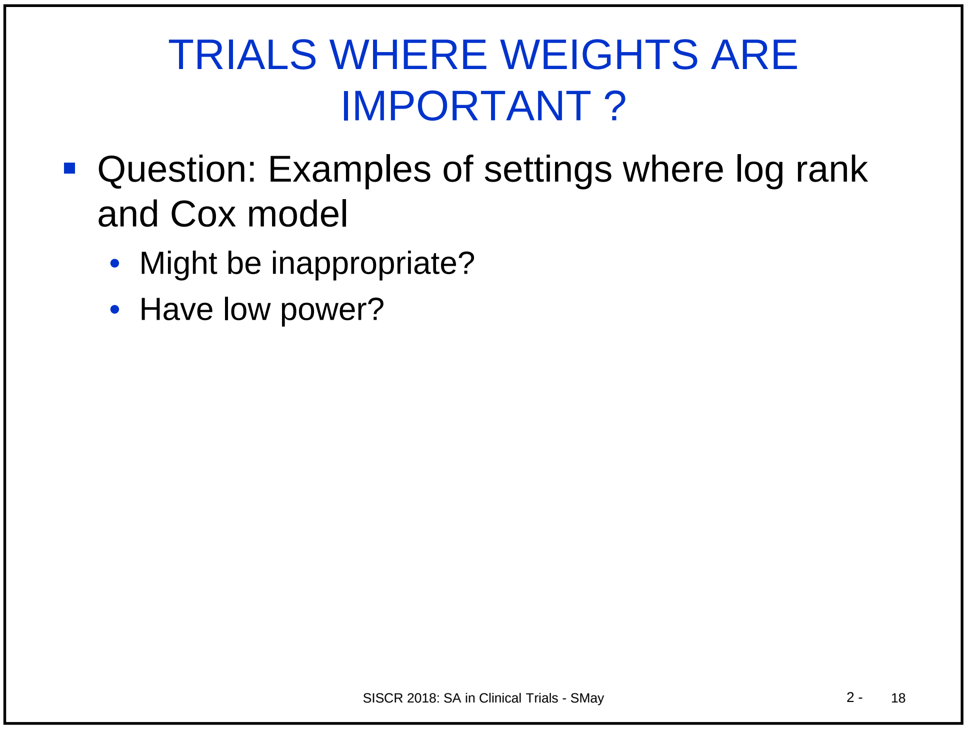
- Kaplan-Meier plot





WEIGHTS

- Determine weights up front
- Clinical considerations
- 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-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

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

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

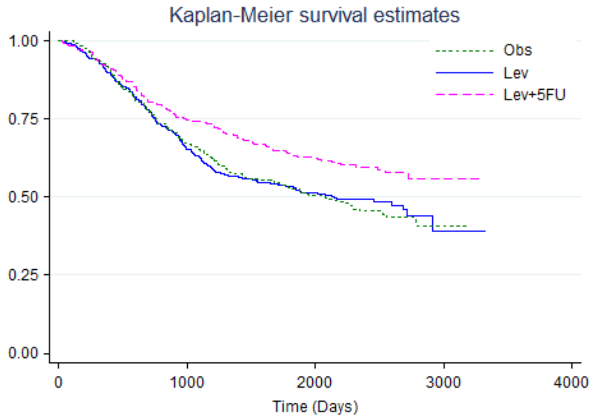
COLON CANCER EXAMPLE

- 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

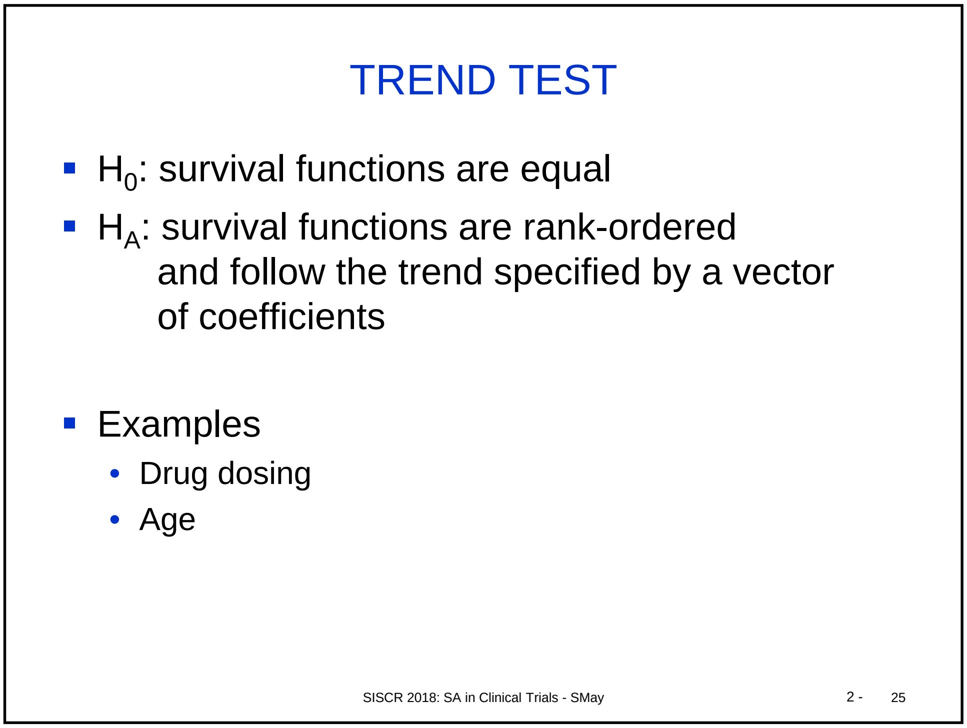
COLON CANCER EXAMPLE

- Obs vs Lev vs Lev+5FU



TREND TEST – EXAMPLE 1 (COLON)

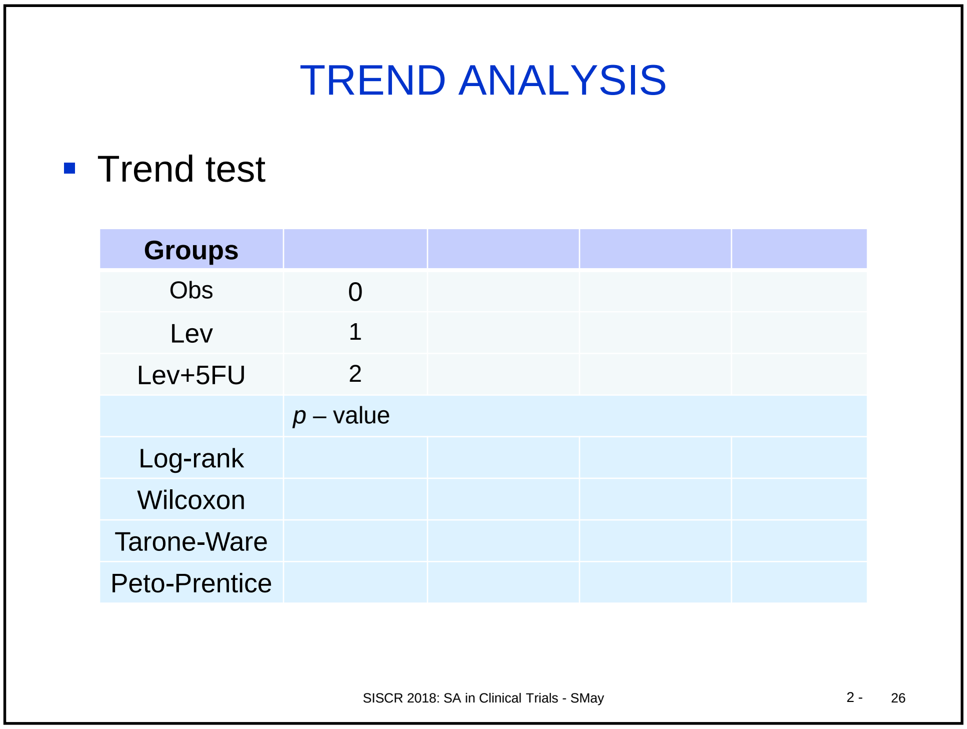
- Obs vs Lev vs Lev+5FU
- Coding ?
- Pretend you did not see any results yet …



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 ANALYSIS

- Trend test

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

TREND ANALYSIS

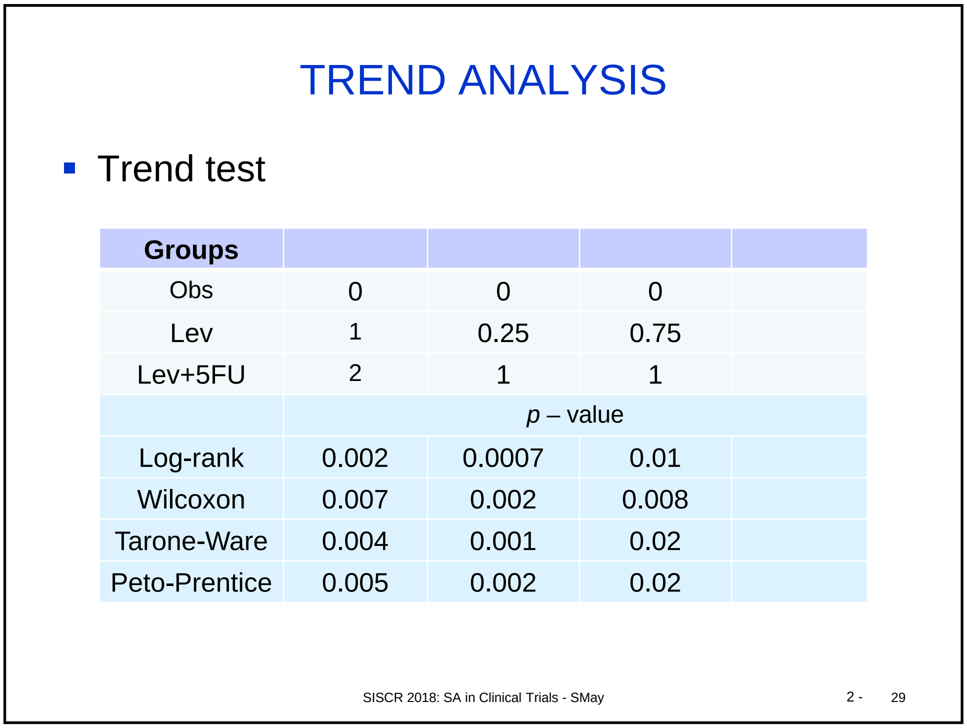
- Trend test

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

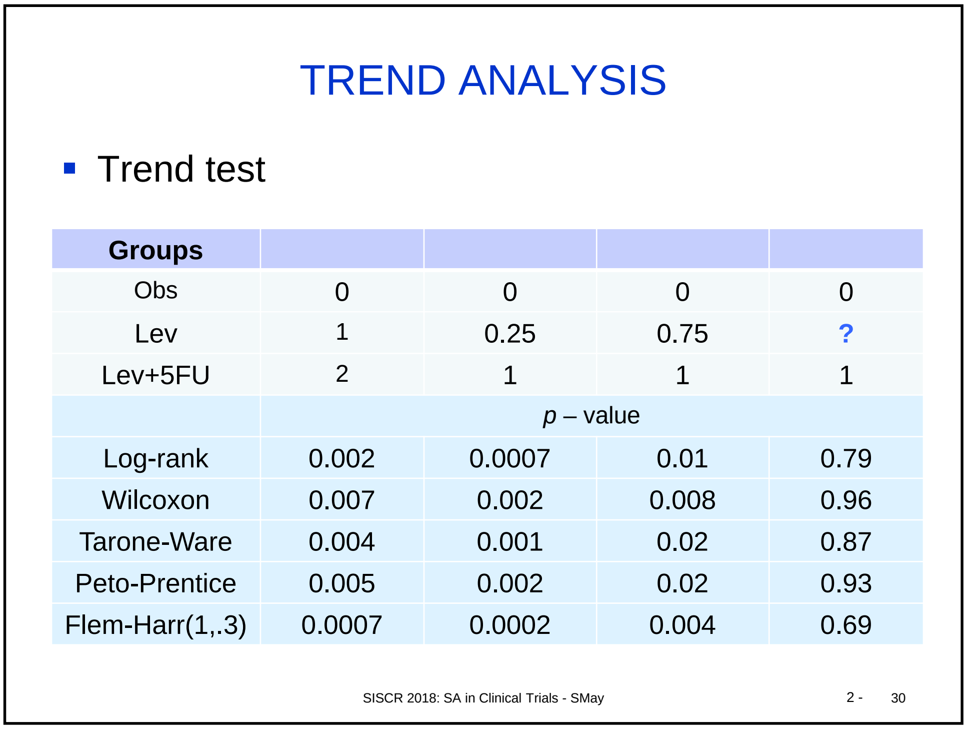
Groups				
Obs	0	0		
Lev	1	0.25		
Lev+5FU	2	1		
	p – value			
Log-rank	0.002	0.0007		
Wilcoxon	0.007	0.002		
Tarone-Ware	0.004	0.001		
Peto-Prentice	0.005	0.002		



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

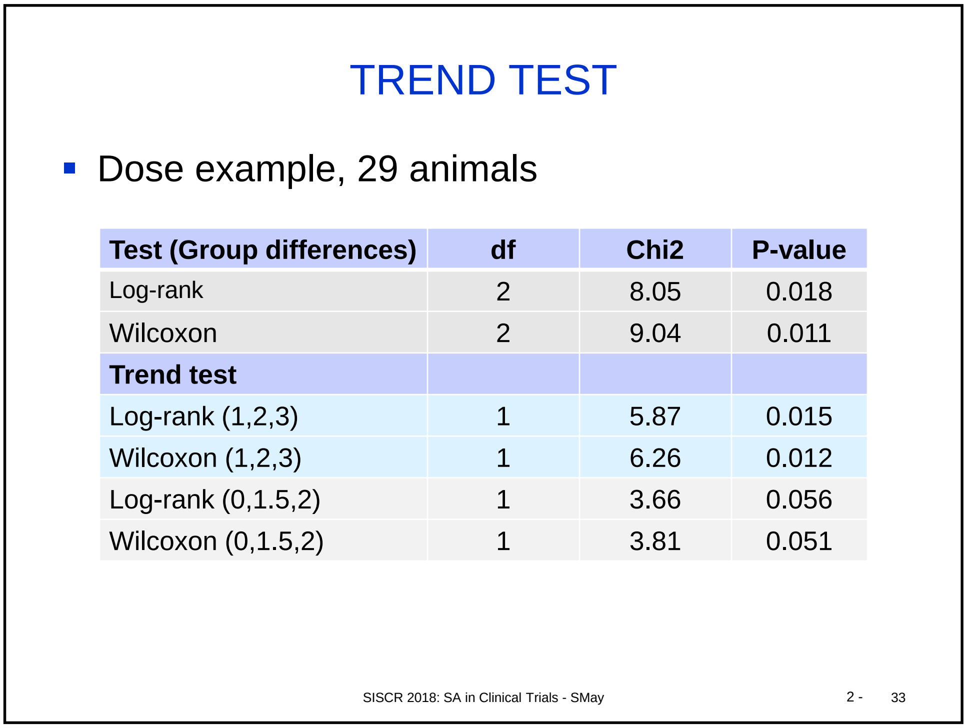
Groups				
Obs	0	0	0	0
Lev	1	0.25	0.75	?
Lev+5FU	2	1	1	1
	p – value			
Log-rank	0.002	0.0007	0.01	0.79
Wilcoxon	0.007	0.002	0.008	0.96
Tarone-Ware	0.004	0.001	0.02	0.87
Peto-Prentice	0.005	0.002	0.02	0.93
Flem-Harr(1,.3)	0.0007	0.0002	0.004	0.69

- 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

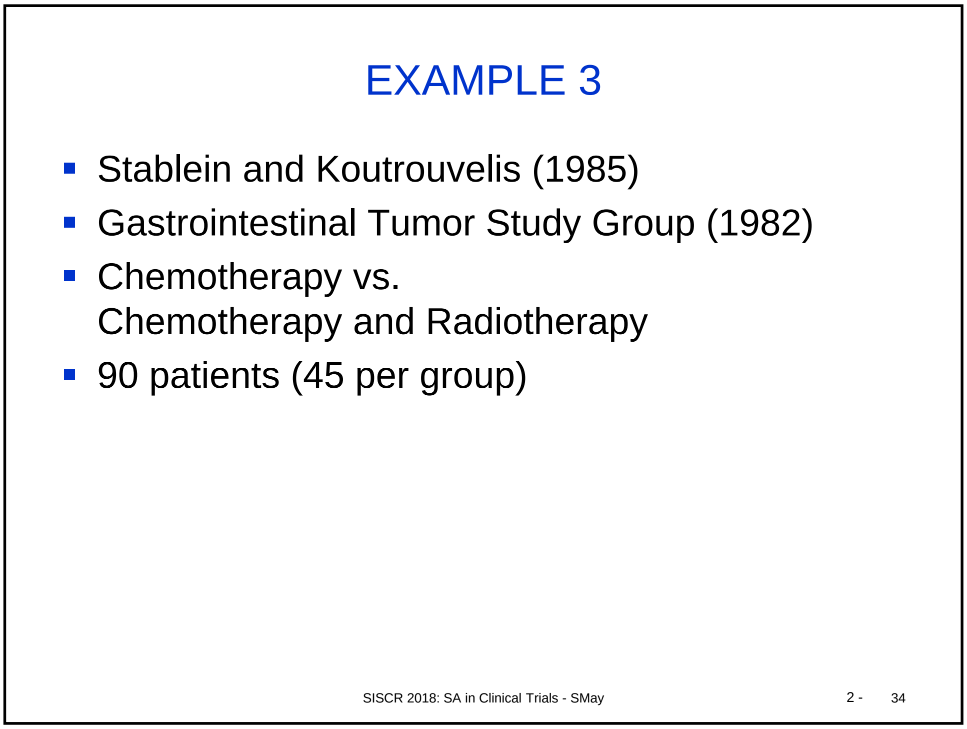
Group	**Dose**	**N**	**Times to event *(t)* or censoring *(t+)***
0	0	9	73+,74+,75+,76,76,76+,99,166,246+
1	1.5	10	43+,44+,45+,67,68+,136,136,150,150,150
2	2.0	10	41+,41+,47,47+,47+,58,58,58,100+,117



TREND TEST

- Dose example, 29 animals

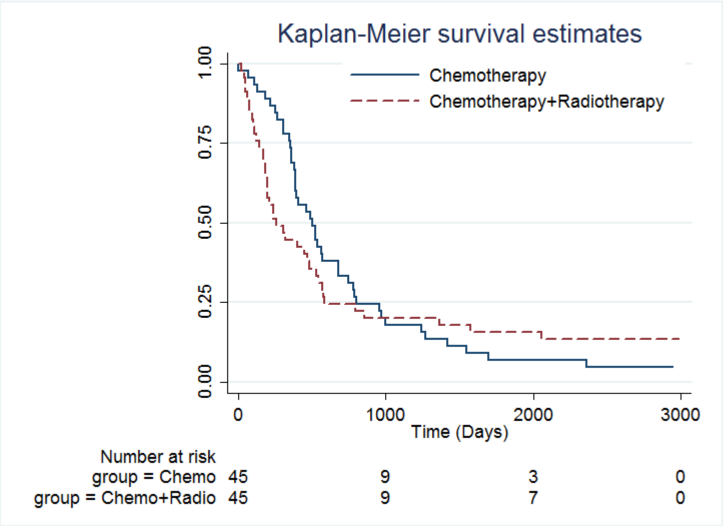
Test (Group differences)	df	Chi2	P-value
Log-rank	2	8.05	0.018
Wilcoxon	2	9.04	0.011
Trend test			
Log-rank (1,2,3)	1	5.87	0.015
Wilcoxon (1,2,3)	1	6.26	0.012
Log-rank (0,1.5,2)	1	3.66	0.056
Wilcoxon (0,1.5,2)	1	3.81	0.051



EXAMPLE 3

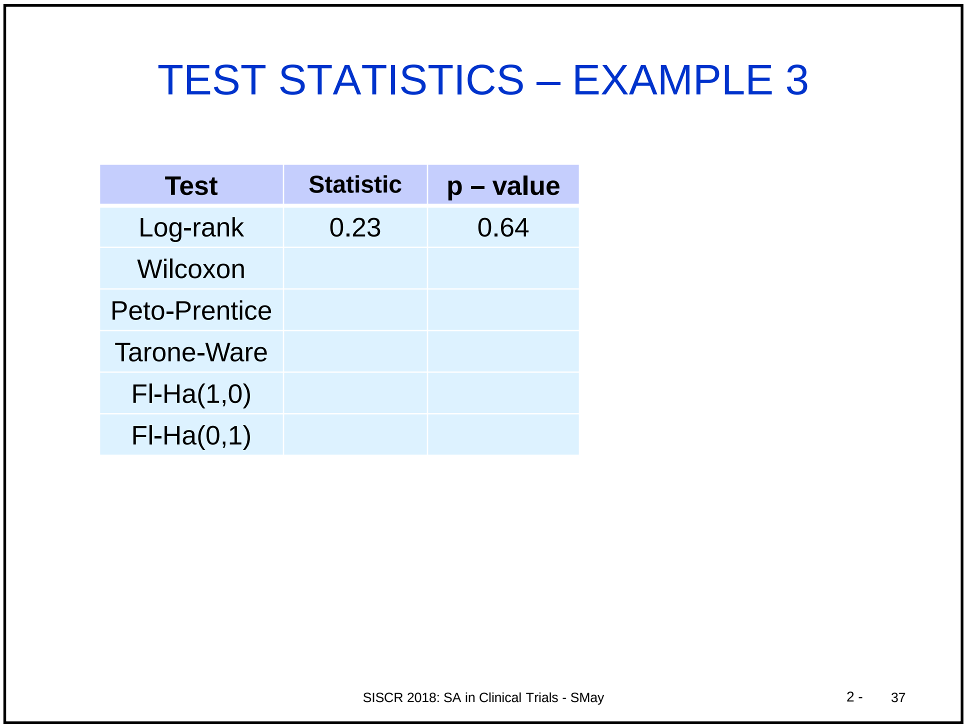
- Stablein and Koutrouvelis (1985)
- Gastrointestinal Tumor Study Group (1982)
- Chemotherapy vs.
  Chemotherapy and Radiotherapy
- 90 patients (45 per group)

KAPLAN-MEIER SURVIVAL CURVES



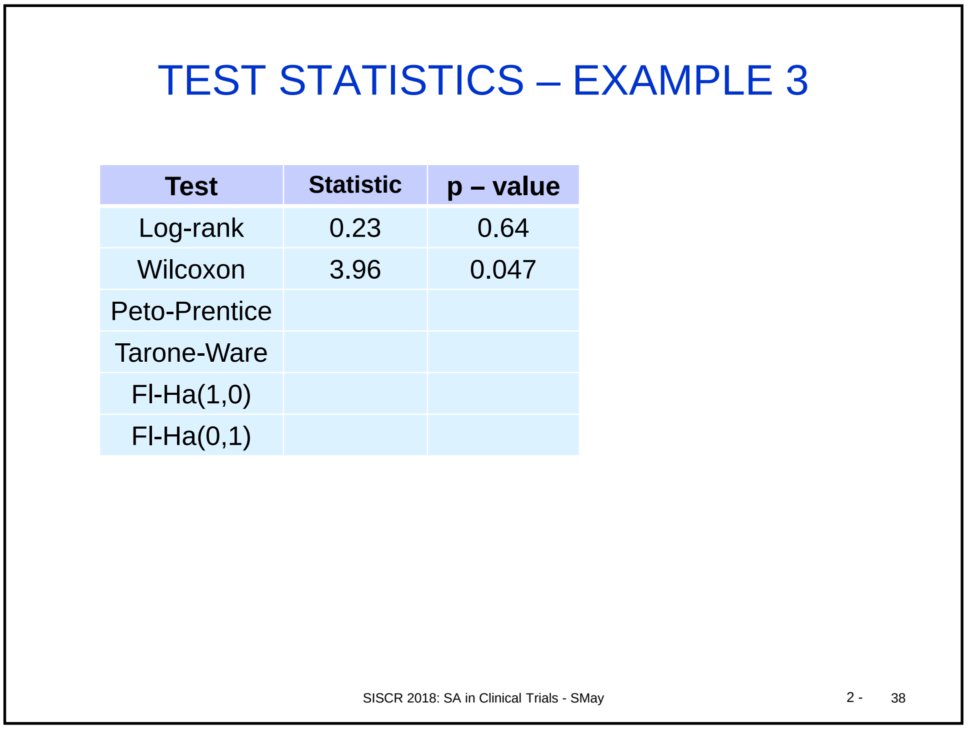
TEST STATISTICS – EXAMPLE 3

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



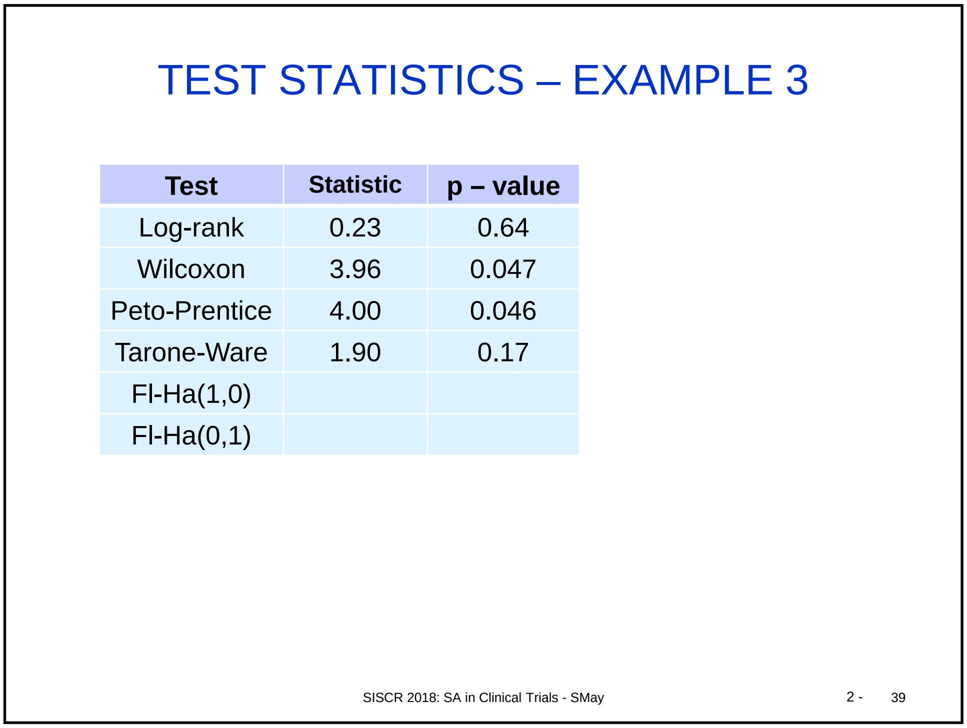
TEST STATISTICS – EXAMPLE 3

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



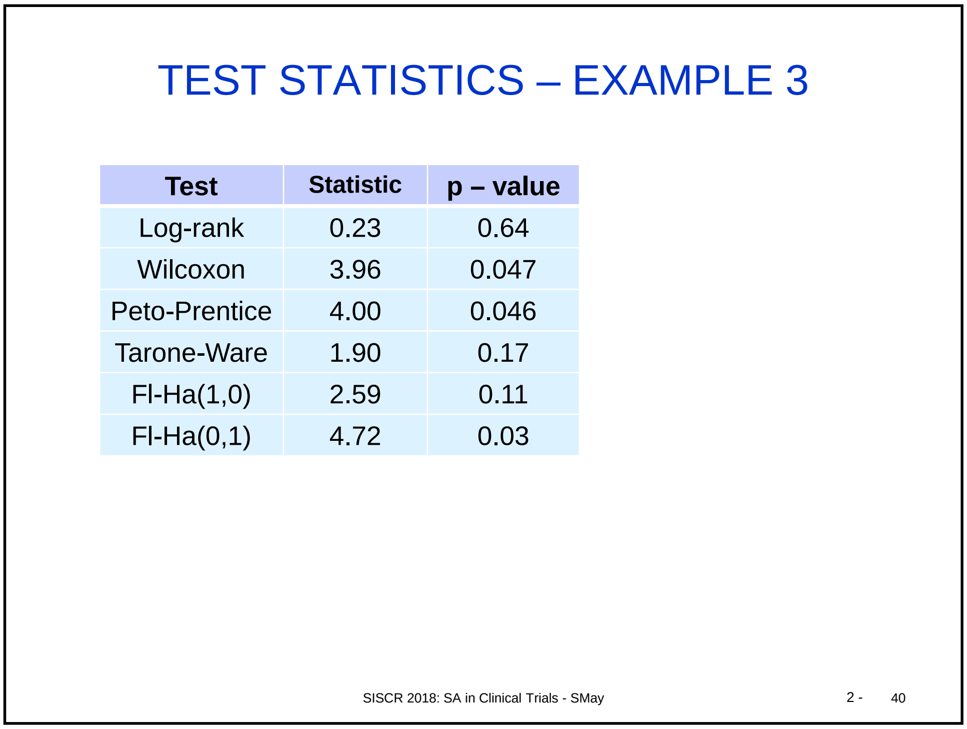
TEST STATISTICS – EXAMPLE 3

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)		



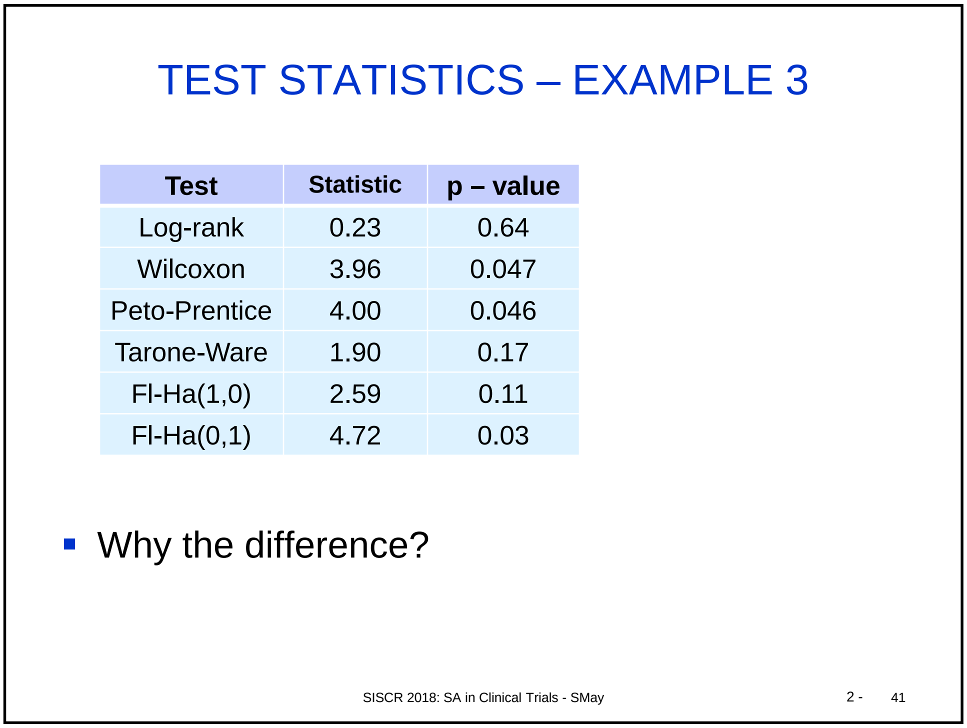
TEST STATISTICS – EXAMPLE 3

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

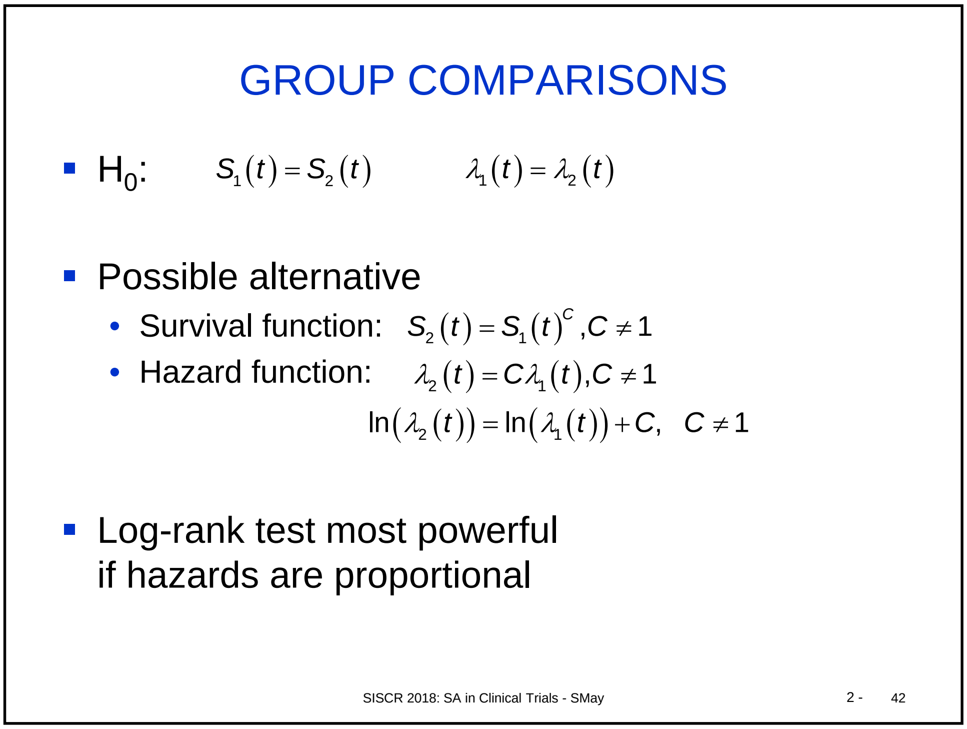
Test	Statistic	p – value
Log-rank	0.23	0.64
Wilcoxon	3.96	0.047
Peto-Prentice	4.00	0.046
Tarone-Ware	1.90	0.17
Fl-Ha(1,0)	2.59	0.11
Fl-Ha(0,1)	4.72	0.03



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)	2.59	0.11
Fl-Ha(0,1)	4.72	0.03

- 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: $\lambda_2(t) = C\lambda_1(t), C \neq 1$

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

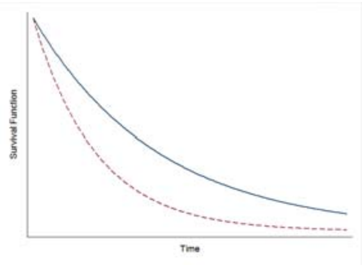
- Log-rank test most powerful if hazards are proportional

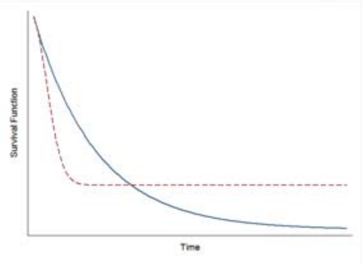
SURVIVAL FUNCTIONS

- We can detect

this

but ordinarily not this





proportional

not proportional

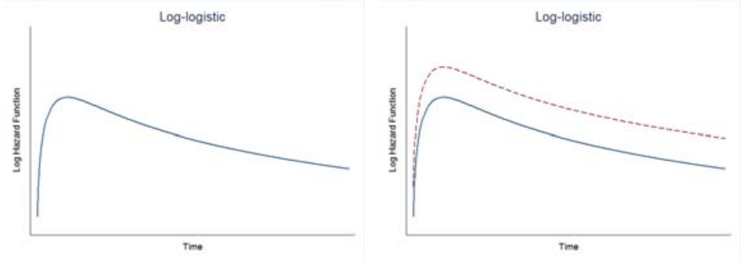
(generated as 2 exponential distributions)

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



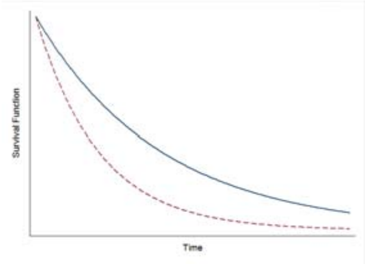
SO FAR

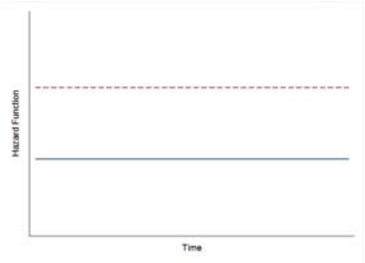
- Two and K – group comparisons
- Trend tests

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

PARAMETRIC MODELS

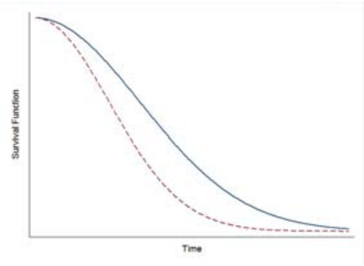
- Control group: Exponential(0.5)
- Example
- Survival functions Hazard functions

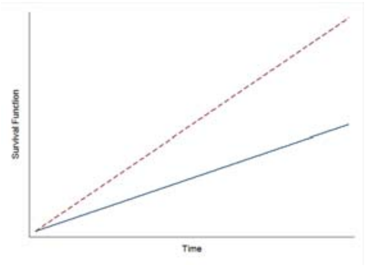




PARAMETRIC MODELS

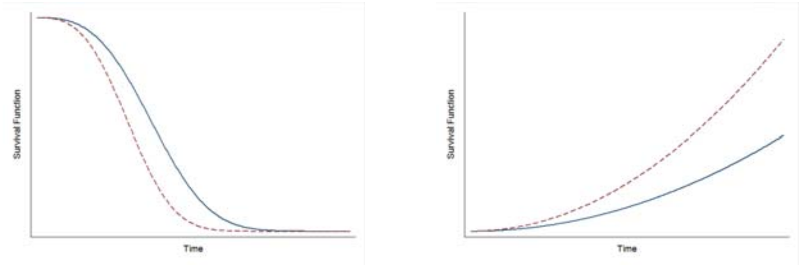
- Control group: Weibull(0.5,2)
- Example
- Survival Functions Hazard Functions





PARAMETRIC MODELS

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

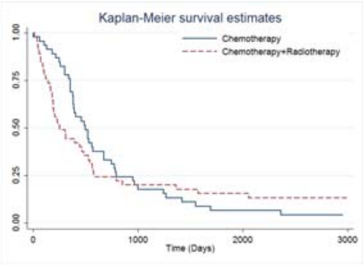


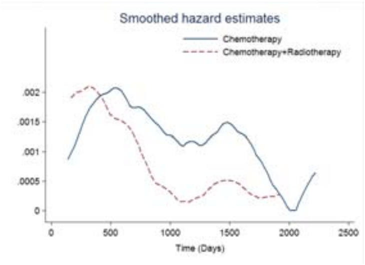
PARAMETRIC APPROACHES

- Weibull and exponential
  - Proportional hazards assumption
  - Distributional assumptions

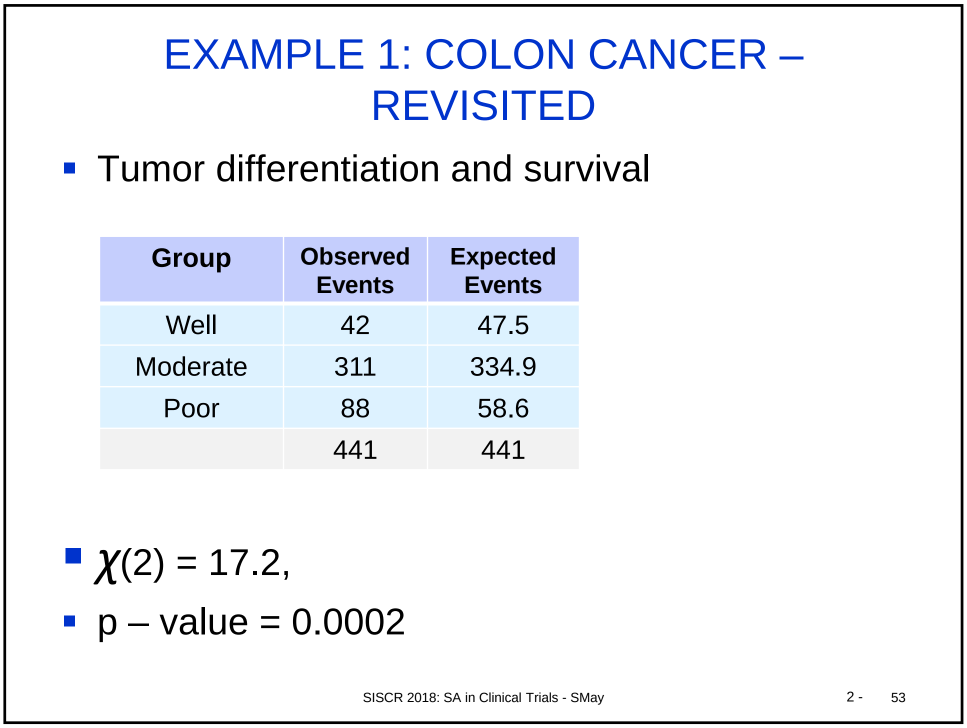
BACK TO EXAMPLE 3

- Gastrointestinal Tumor Study
- Survival Functions Hazard Functions





- Other covariates

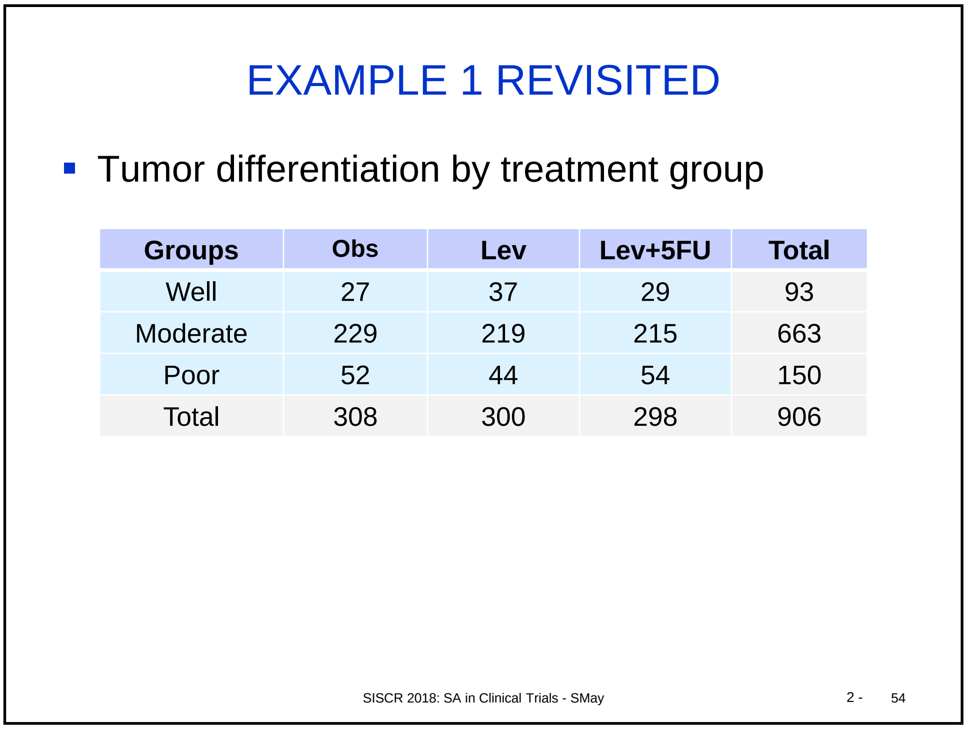


EXAMPLE 1: COLON CANCER – REVISITED

- 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

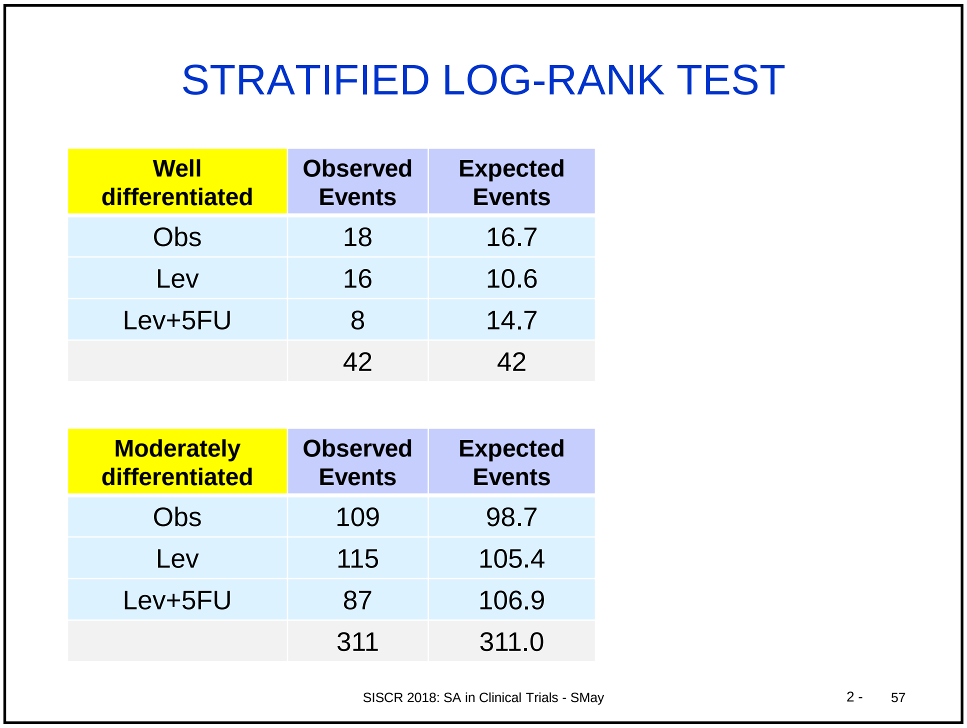
$$Q=\frac{\left[\sum_{j=1}^{m}\left(d_{1(j)}-\hat{E}_{1(j)}\right)\right]^2}{\sum_{j=1}^{m}\hat{V}_{(j)}}$$

- 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

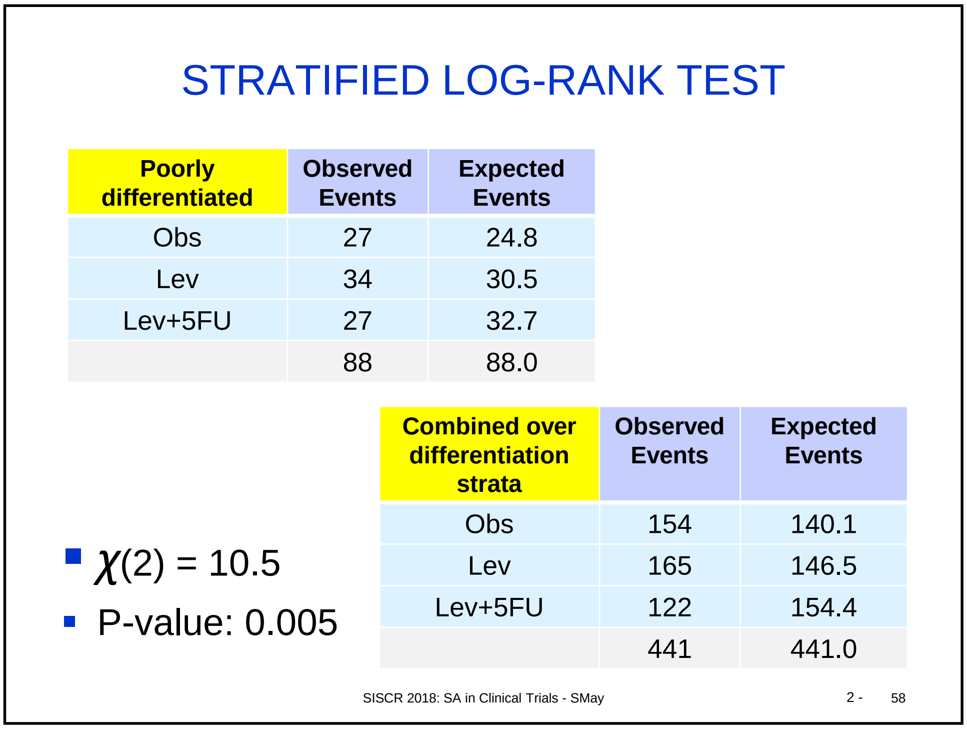
- H_0: $\lambda_{1r}(t)=\lambda_{2r}(t)$ for all *r* = 1,…,*R*
- H_A: $\lambda_{1r}(t)=c\lambda_{2r}(t), c\neq 1$ for all *r* = 1,…,*R*
- Under H_0 test statistic ~ $\chi^2(K-1)$
- The $d_{1r(j)}, \hat{E}_{1r(j)}$ and $\hat{V}_{r(j)}$ are solely based on subjects from the *r*-th strata



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

(FAIR) COMPARISON STRATA VS NO STRATA

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

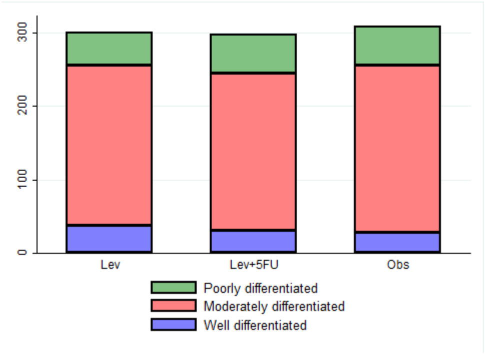
Combined over differentiation strata	Observed Events	Expected Events
Obs	154	140.1
Lev	165	146.5
Lev+5FU	122	154.4
	441	441.0

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

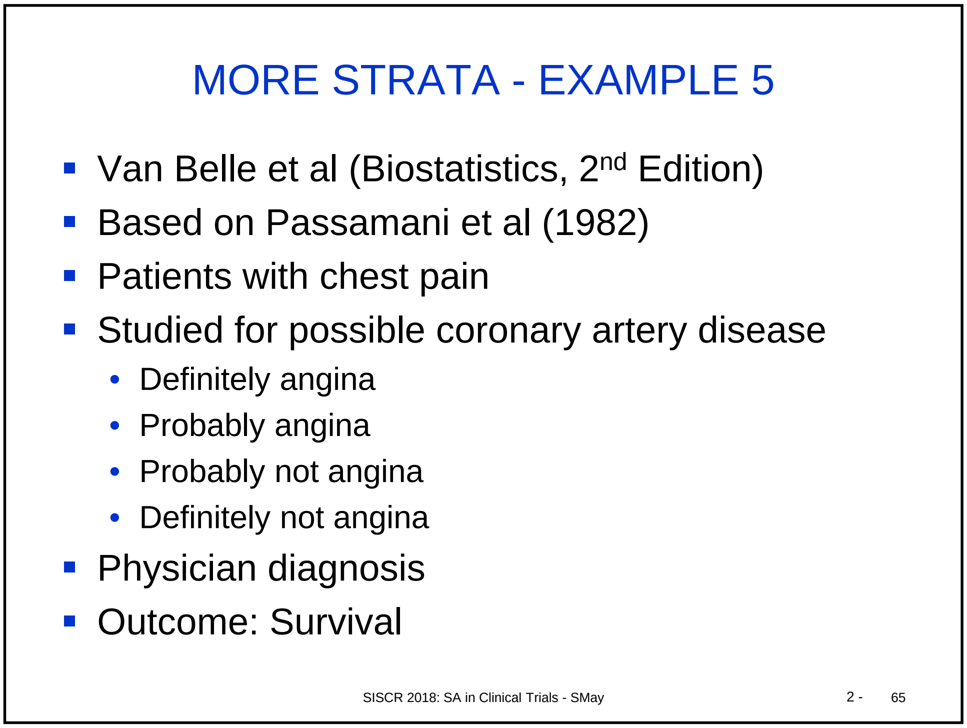
Without strata	Observed Events	Expected Events
Obs	154	141.4
Lev	165	145.3
Lev+5FU	122	154.3
	441	441.0

DIFFERENTIATION BY TREATMENT GROUP

- Randomization worked

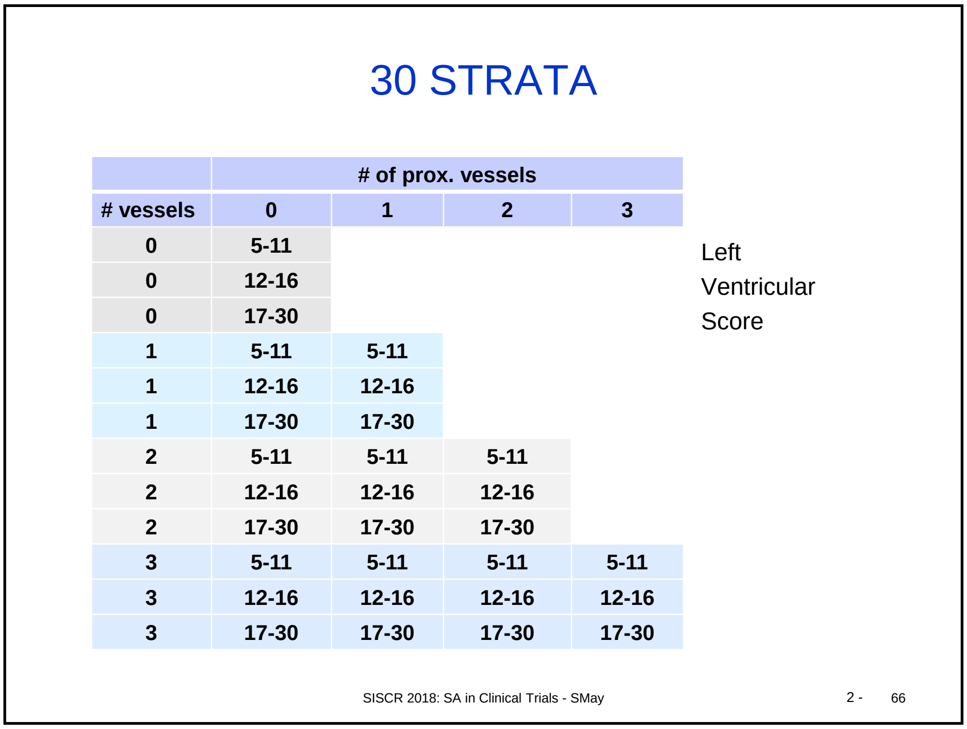


- Example with more strata



MORE STRATA - EXAMPLE 5

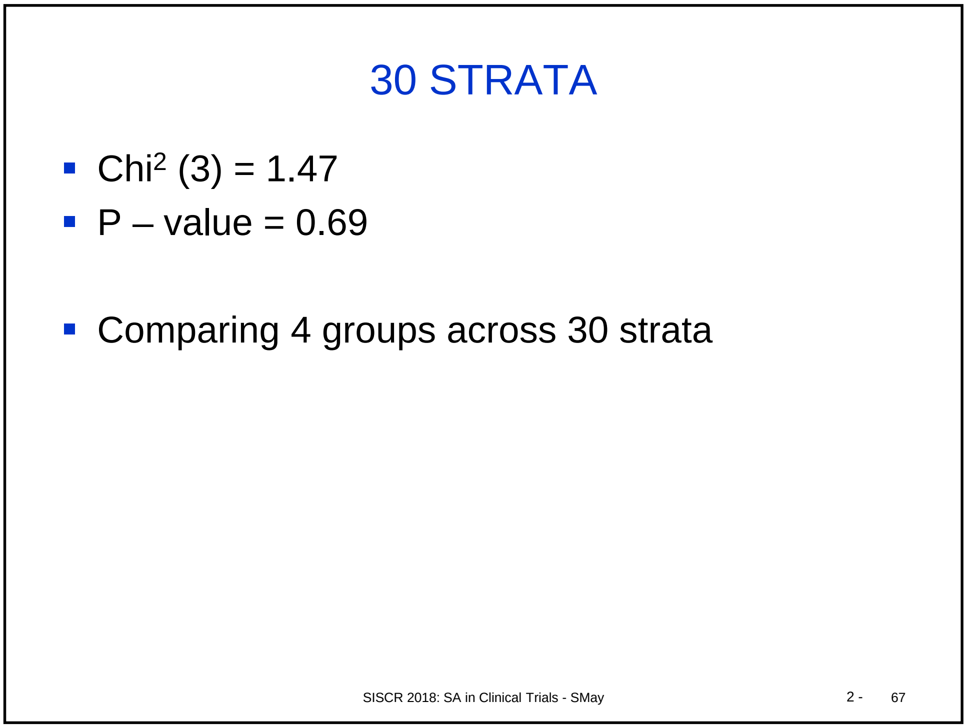
- Van Belle et al (Biostatistics, 2nd Edition)
- Based on Passamani et al (1982)
- Patients with chest pain
- Studied for possible coronary artery disease
  - Definitely angina
  - Probably angina
  - Probably not angina
  - Definitely not angina
- Physician diagnosis
- Outcome: Survival



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

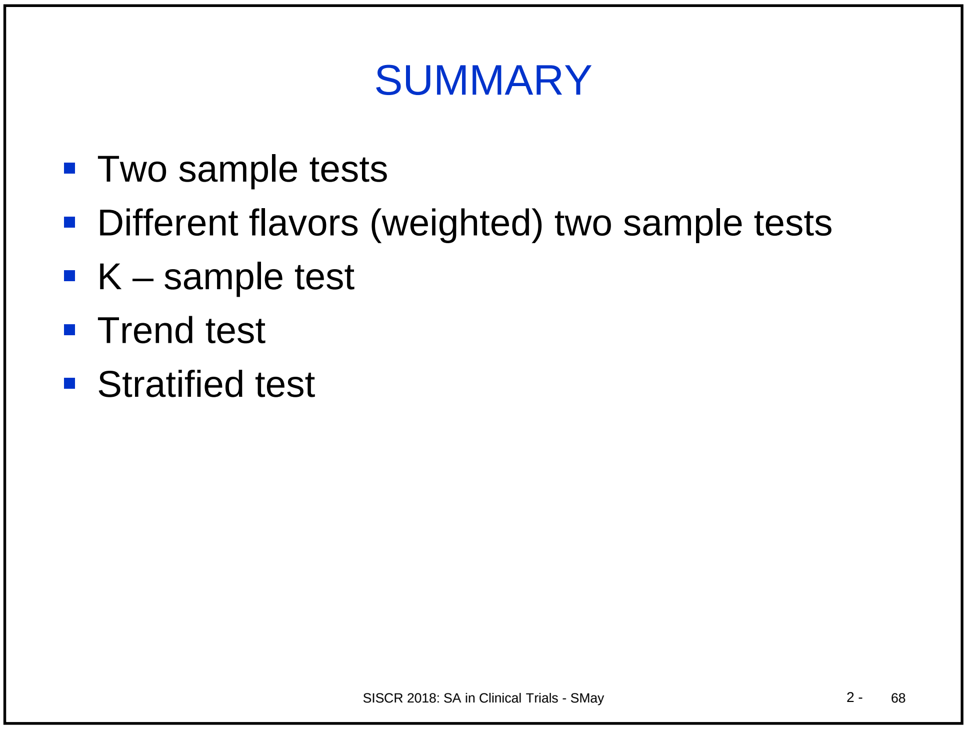
Left
Ventricular
Score



30 STRATA

- Chi^2 (3) = 1.47
- P – value = 0.69

- Comparing 4 groups across 30 strata



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 ?

SESSION 3:
ADDITIONAL TWO-SAMPLE TESTS

Module 13: Survival Analysis in Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2017

Barbara McKnight, Ph.D.
Professor
Department of Biostatistics
University of Washington

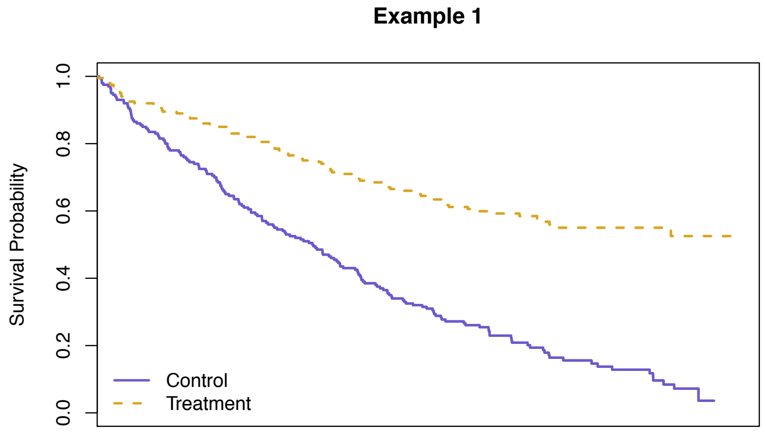
OUTLINE

- Limitations of proportional hazards
- Other contrasts based on functionals of S(t)
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- Other metrics to describe the distance between survival curves
  - Weighted difference in S(t)
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

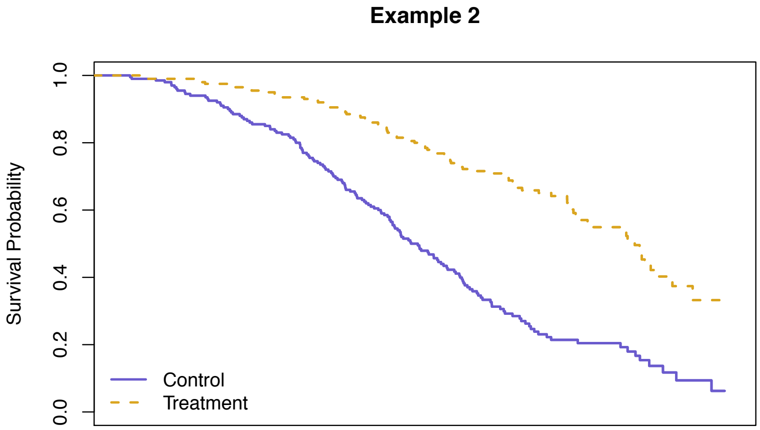
OUTLINE

- **Limitations of proportional hazards**
- Other contrasts based on functionals of S(t)
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- Other metrics to describe the distance between survival curves
  - Weighted difference in S(t)
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

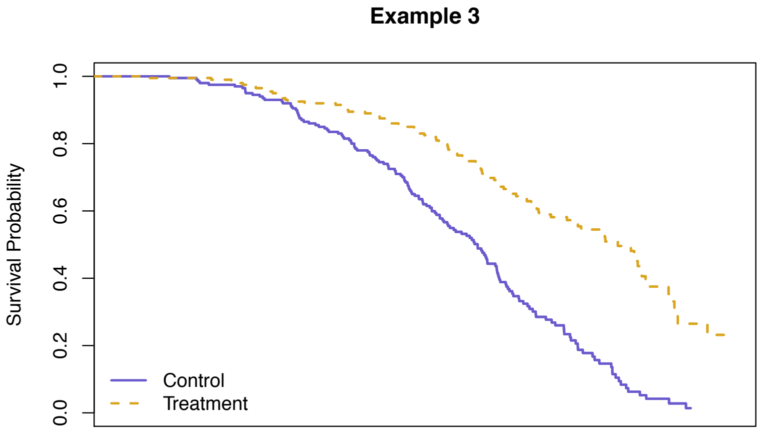
PROPORTIONAL HAZARDS EXAMPLES



PROPORTIONAL HAZARDS EXAMPLES



PROPORTIONAL HAZARDS EXAMPLES

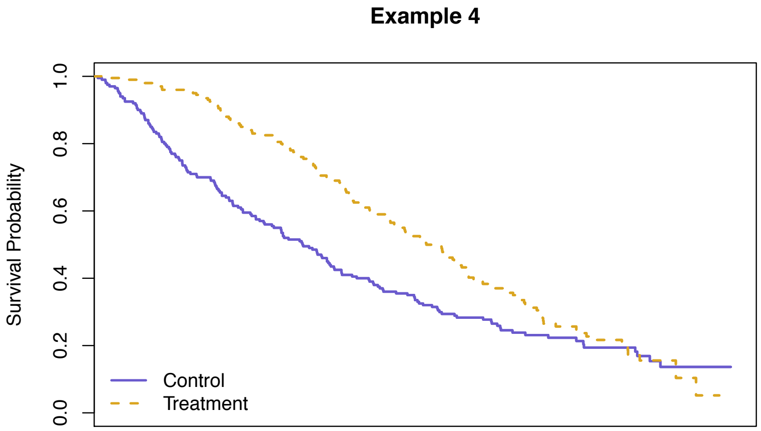


PROPORTIONAL HAZARDS EXAMPLES

Q: Which group has better survival in these examples?

A:

NON-PROPORTIONAL HAZARDS EXAMPLES



NON-PROPORTIONAL HAZARDS EXAMPLES

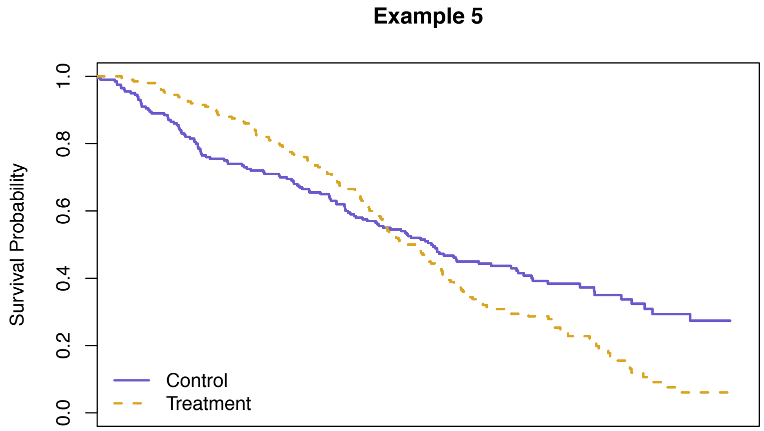
Q: Why does it appear the hazards are not proportional?

A:

Q: Which group has better survival?

A:

NON-PROPORTIONAL HAZARDS EXAMPLES

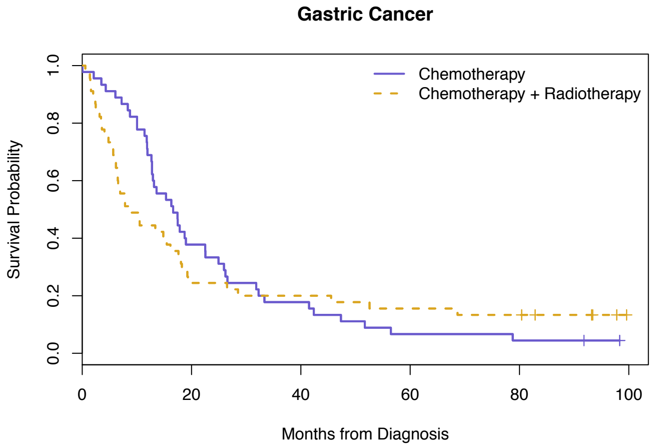


YOUR CHOICE

- Which group has better survival?

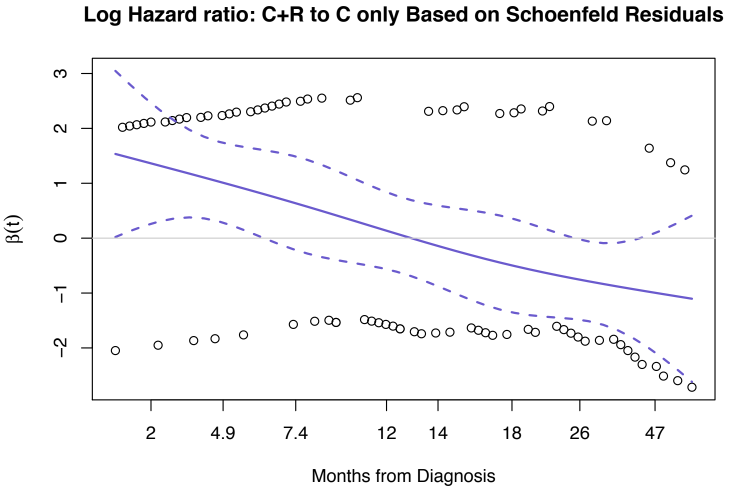
- You are a newly diagnosed patient. What would you want to know before choosing whether to take treatment?

REAL DATA



Schein PS, Gastrointestinal Tumor Study Group. A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. Cancer. 1982 May 1;49(9):1771–1777.

HAZARD RATIO



Test of proportional hazards based on Schoenfeld residuals: P = 0.0003

HAZARD RATIO

	Hazard Ratio	95% CI	P-value
Chemotherapy	1.0 (reference)	--	--
Chemotherapy + Radiotherapy	1.1	(0.72, 1.7)	.63

Assuming hazard ratio is constant…

CROSSING HAZARDS

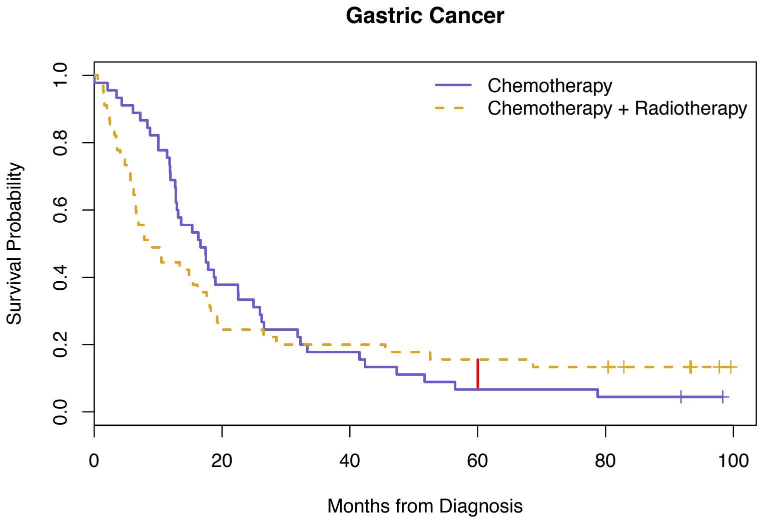
When the proportional hazards assumption doesn't hold:

- Cox model will give weighted-average of time-specific hazard ratios (weights depend on censoring distribution)
- log rank test will test whether a weighted-average difference of hazards is zero
  - statistic numerator $= \sum_j \frac{n_{1j}n_{2j}}{(n_{1j}+n_{2j})}(\frac{d_{1j}}{n_{1j}} - \frac{d_{2j}}{n_{2j}})$
  - More weight at earlier times when number at risk is larger
- May not be the quantity on which you want to base inference (estimation and testing)

OUTLINE

- Limitations of proportional hazards
- **Other contrasts based on functionals of S(t)**
  - **S(t) at fixed time point**
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- Other metrics to describe the distance between survival curves
  - Weighted difference in S(t)
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

FIVE-YEAR SURVIVAL



FIVE-YEAR SURVIVAL

- Compares only at a single point in time
- Ignores earlier survival differences, which may be important to some patients, given that in this example survival to 5 years in either group is low

S(t) AT A CHOSEN TIME t

- Choose time t for comparison at design stage.
- Compare $\hat{S}_1(t)$ to $\hat{S}_2(t)$ using

$$\frac{\hat{S}_1(t) - \hat{S}_2(t)}{\sqrt{\widehat{\mathrm{var}}(\hat{S}_1(t)) + \widehat{\mathrm{var}}(\hat{S}_2(t))}}$$

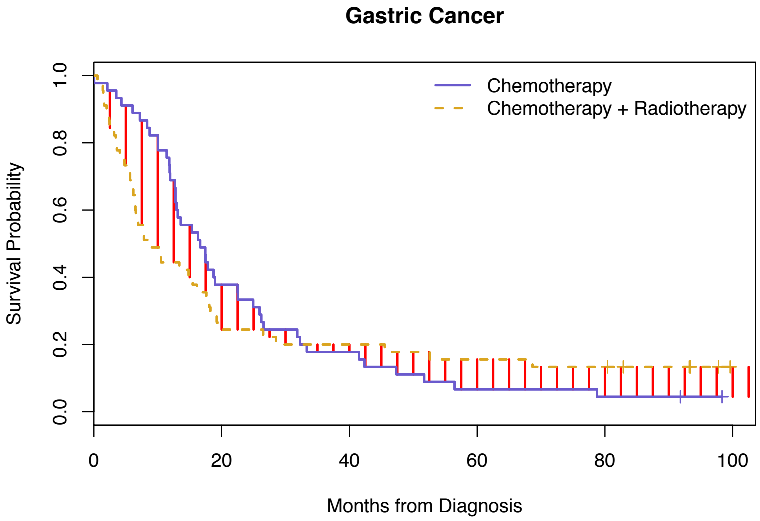
where $\widehat{\mathrm{var}}(\hat{S}_2(t))$ is computed using Greenwood's formula or another large-sample formula such as the one based on the complementary log-log of $\hat{S}(t)$.

FIVE-YEAR SURVIVAL DIFFERENCE

Gastric Cancer

Difference	se(Difference)	Z Statistic	P-value
.0889	.0656	1.36	.1753

COMPARISON AT MORE THAN ONE TIME



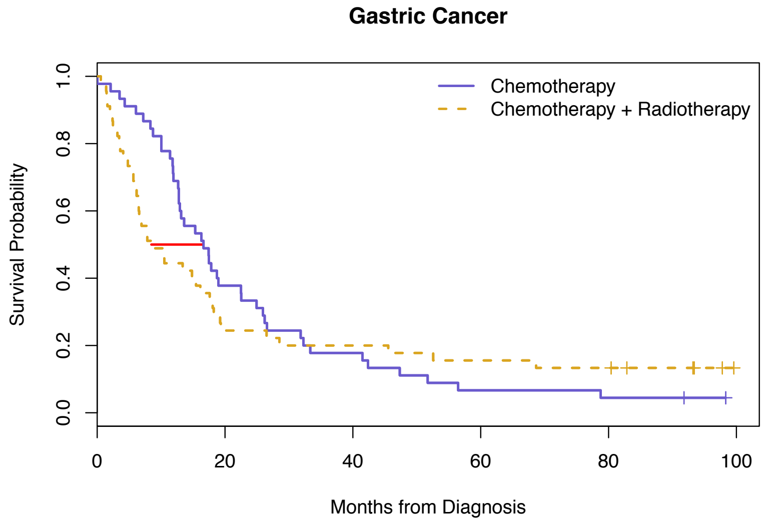
AVERAGE DIFFERENCES

- Average difference between survival curves over time might be of interest
- In gastric cancer example, differences are of different signs at different times, so there would be cancellation
- Allows poorer survival after survival curves cross to detract from better survival before
- Interpretation?
- Also related to average quantile difference

OUTLINE

- Limitations of proportional hazards
- **Other contrasts based on functionals of S(t)**
  - S(t) at fixed time point
  - **Quantiles (eg. median)**
  - Mean survival time
  - Restricted mean survival time
- Other metrics to describe the distance between survival curves
  - Weighted difference in S(t)
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

MEDIAN SURVIVAL



MEDIAN SURVIVAL

- Compares only a single quantile
- Hard for some patients to interpret the difference in medians

MEDIAN TEST

Idea: Define $\hat{M}_1$ and $\hat{M}_2$ to be the median survival times in the two samples.

Then let the overall median survival time be defined by the weighted average.

$$\hat{M} = \frac{N_1}{N}\hat{M}_1 + \frac{N_2}{N}\hat{M}_2$$

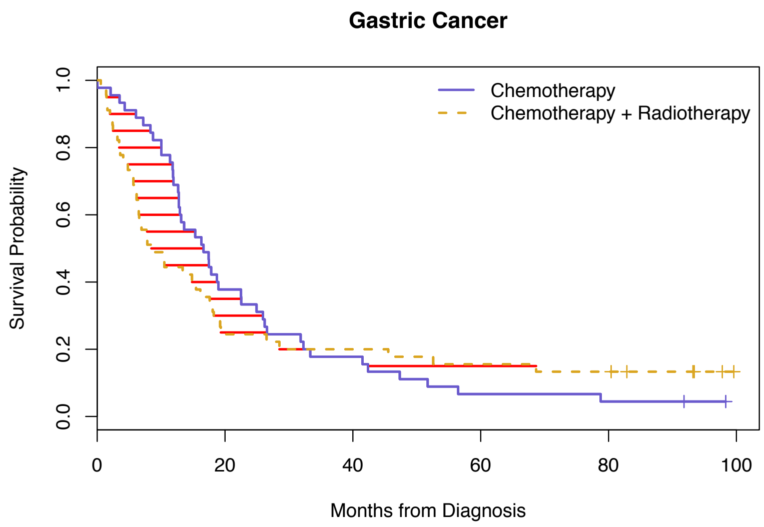
A test of $H_0 : M_1 = M_2$ can be performed by testing

$$H_0 : S_1(\hat{M}) = S_2(\hat{M})$$

Reference distribution based on joint asymptotic distribution of $(S_1(\hat{M}), S_2(\hat{M}))$.

Brookmeyer R, Crowley J. JASA 1982;77(378):433–440.

MORE THAN ONE QUANTILE



OUTLINE

- Limitations of proportional hazards
- **Other contrasts based on functionals of S(t)**
  – S(t) at fixed time point
  – Quantiles (eg. median)
  – **Mean survival time**
  – Restricted mean survival time
- Other metrics to describe the distance between survival curves
  – Weighted difference in S(t)
  – Maximum difference (Kolmogorov – Smirnov)
  – Integrated squared difference (Cramér von Mises)

MEAN SURVIVAL TIME

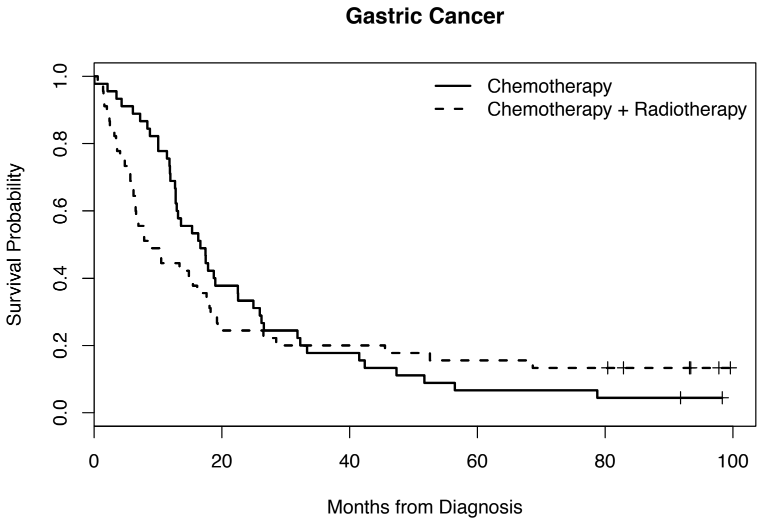
Useful Fact: $\int_0^\infty S(t)dt = E(T) = \int_0^\infty tf(t)dt$

Proof: $\int_0^\infty S(t)dt = S(t)t|_0^\infty - \int_0^\infty t(-f(t))dt = \int_0^\infty tf(t)dt$

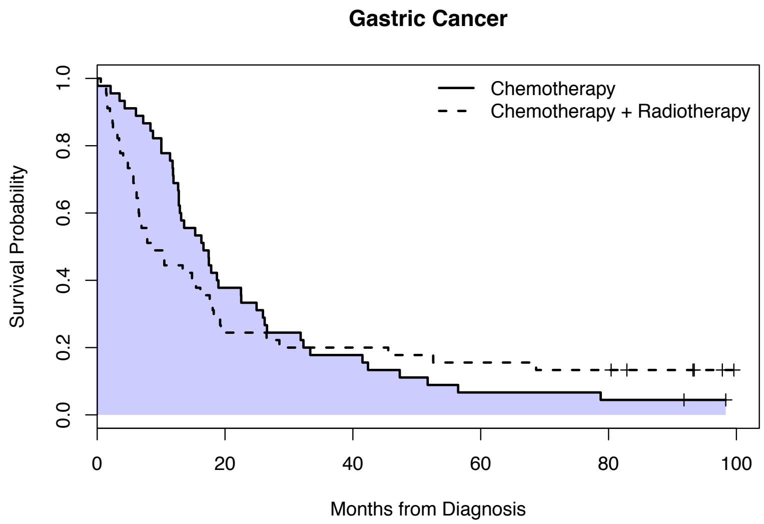
by integration by parts and

the fact that $E(T) < \infty \Rightarrow tS(t) \overset{t\to\infty}{\to} 0$.

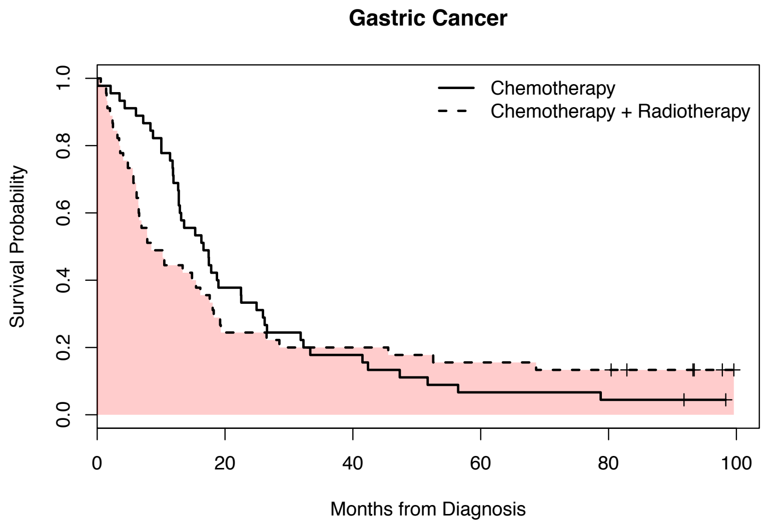
MEAN SURVIVAL TIME



MEAN SURVIVAL TIME



MEAN SURVIVAL TIME



MEAN SURVIVAL TIME

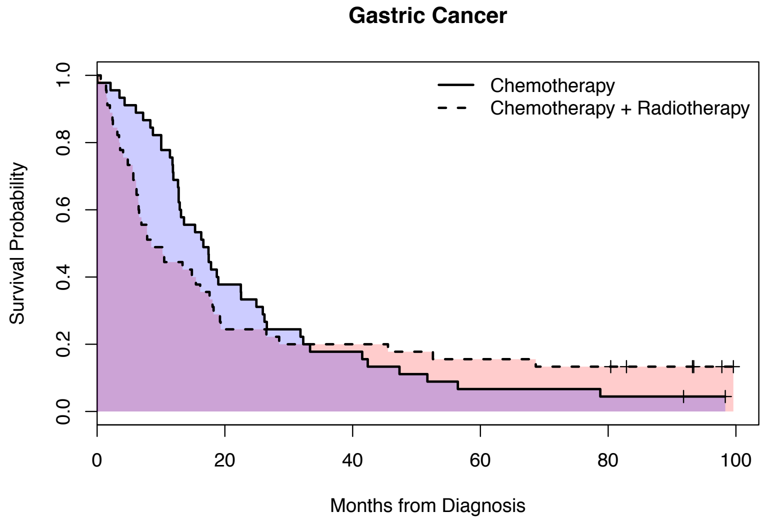
- Mean survival time $\mu = \int_0^\infty S(t)dt$
- Large sample (asymptotic) distribution proved by Gill in The Annals of Statistics. 1983;11(1):49–58.
- In finite samples, can be infinite if last time is a censoring
  - Integrate to last failure time only
  - Integrate to last observed time only

MEAN SURVIVAL TIME

	Mean Survival*	SE
Chemotherapy	24.1 months	3.3 months
Chemotherapy + Radiotherapy	24.3 months	4.8 months

* Up to 99.6 months (last observed time in either group)

MEAN SURVIVAL TIME



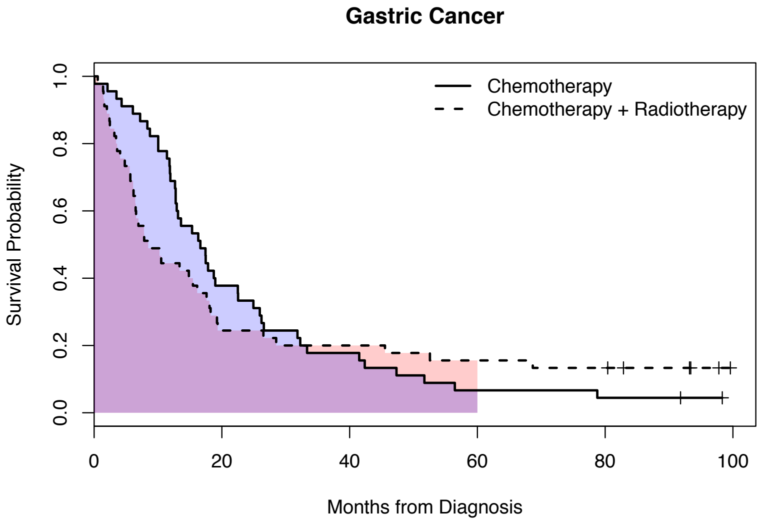
MEAN SURVIVAL TIME DIFFERENCE

- Average of survival function differences over time
- Average of survival quantile differences over quantiles
- Allows cancellation
- Not much information at late times where few are at risk.
- Infinite estimate if KM curve doesn't descend to zero
- May want to truncate to a shorter interval, restricting to times where *S(t)* estimates are precise

OUTLINE

- Limitations of proportional hazards
- **Other contrasts based on functionals of S(t)**
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - **Restricted mean survival time**
- Other metrics to describe the distance between survival curves
  - Weighted difference in S(t)
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

RESTRICTED MEAN SURVIVAL TIME



MOTIVATION

- Clinically Interpretable ("over the next five years, patients like you live, on average, 13 months longer")
- Power/precision depends on length of observation time as well as number of events. Can achieve enough power/precision for meaningful comparisons with smaller studies.
- May be better measure for non-inferiority safety studies where events are rare. (Uno H et al. Ann Intern Med 2015; 21;163(2):127–134.) "Average number of days out of n event free."
- Excellent motivation when survival curves do not cross.

RESTRICTED MEAN SURVIVAL TIME

- Interpretation: average time lived in the interval $[0, \tau]$.
- Interpretation for differences: on average, the amount more time lived in $[0, \tau]$ on treatment A than on treatment B.
- Some asymptotically equivalent ways to estimate it:
  - $\hat{\mu} = \int_0^\tau \hat{S}(t)dt$
  - $\frac{1}{n}\sum_{i=1}^n \frac{d_i y_i}{\hat{S}_{c(y_i)}}$ where $\hat{S}_{c(y_i)}$ is the KM estimated survival function of the censoring distribution
  - Using pseudo-observations based on the jackknife.

$$\hat{\mu} = \sum_{i=1}^n \hat{\mu}_i,$$

RESTRICTED MEAN SURVIVAL DIFFERENCE

- Standard estimation and testing:
  - $\hat{\mu}_k = \int_0^\tau \hat{S}_k(t)dt$
  - $\widehat{\text{var}}(\hat{\mu}_k) = \sum_{j=1}^{J} [\int_{t_j}^{\tau} \hat{S}_K(t)dt]^2 \frac{D_{jk}}{N_{jk}(N_{jk}-D_{jk})}$
  - Compare test statistic:

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\widehat{\text{var}}(\hat{\mu}_1) + \widehat{\text{var}}(\hat{\mu}_2)}}$$

    to standard normal distribution (asymptotic).

RESTRICTED MEAN SURVIVAL TIME

$$E[\min(T, \tau)] = \widehat{E[Y]} = \int_0^\tau \hat{S}(t)dt$$

Several approaches to variance estimation:

- Asymptotic
- Random perturbation resampling method (Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostat. 2014 Apr 1;15(2):222–233.)
- Variance of pseudo observations

PSEUDO OBSERVATIONS

- There are a number of other less direct ways to estimate $\mu_k = \int_0^\tau \hat{S}_k(t)dt$ that make generalizing to regression models easier.
- One appealing method uses pseudo-observations based on the jackknife.
  - Group means computed in the usual way from pseudo-observations
  - Standard errors computed from pseudo-observations in the usual way.
  - Test statistic based on two-sample t-test (unequal variances) with pseudo-observations.

PSEUDO OBSERVATIONS

Estimation of μ using pseudo-observations based on the jackknife.

$$\hat{\mu} = \sum_{i=1}^{n} \hat{\mu}_i,$$

where $\hat{\mu}_i = n\hat{\mu} - (n-1)\hat{\mu}_{-i}$.

- $\hat{\mu}$ is computed by the first method from the pooled sample, and
- $\hat{\mu}_{-i}$ is computed the same way but leaving out the i^{th} observation.
- Andersen et al. Lifetime Data Anal. 2004;10(4):335–350.
- Functions available in Stata, R and SAS.

RESTRICTED MEAN SURVIVAL TIME

	Restricted Mean Survival (2000 days)	SE
Chemotherapy	673	77.8
Chemotherapy + Radiotherapy	599	101.1

Comparison Method	P-value
Asymptotic	.560
Pseudo observations	.566

DESIGN AND INFERENCE ISSUES

- Not much information / precision available at late times when few subjects are at risk
  - If a restricted mean over an interval [0, τ] is of interest, important to follow subjects enough longer than τ to have an adequate number still at risk at time τ.

EXAMPLE

- Schermerhorn et al. (2015) compared survival in a matched cohort of 39,966 pairs of Medicare patients who received either endovascular or open repair of an abdominal aortic aneurism.
  - Perioperative mortality and complication rates were higher in those given open repair: 5.2% vs 1.6% for mortality and 12.9% vs 3.8%
  - The estimated hazard ratio for death comparing endovascular to open repair varied over time:
    - HR = .32 (95% CI: .29 - .35) over the first 30 days
    - HR = .64(95% CI: .58 -.71) for 30 – 90 days
    - HR = 1.17(95% C: I 1.13 – 1.21) for 90 days – 4 years
    - HR = 1.05 (95% CI: 1.00 - 1.09) after 4 year.

Schermerhorn ML, Buck DB, O'Malley AJ et al. NEJM 2015 Jul 23;373(4):328–338.

EXAMPLE

- Because of non-proportional hazards they estimated differences in restricted mean survival using the pseudo observation approach of Andersen et al with the matched-pair data.
  - Over the first 4 years, the endovascular group lived an average of 12.4 days longer (95% CI 9.0 – 15.6)
  - Over the first 7 years, the endovascular group lived an average of 8.2 days longer (95% CI: 1.5-14.4)
  - The authors concluded that the advantage of endovascular repair persisted to 7 years.
- The pseudo-observation approach makes it easy to accommodate the matched design.

SCREENING TRIAL

- 202,546 women 50-72 years of age, England, Wales, Northern Ireland
- Randomized to one of three arms in 1:1:2 ratio between June 1, 2001 and Oct 21, 2005.
  - Annual multimodal screening (serun CA 125 + algorithm)
  - Annual transvaginal ultrasound
  - No screening
- Screening ended Dec 31, 2011.
- Not blinded
- Primary outcome: death from ovarian cancer (by end of 2014)

Jacobs IJ, Menon U, Ryan A, et al. (2016) The Lancet. 387(10022):945–956.

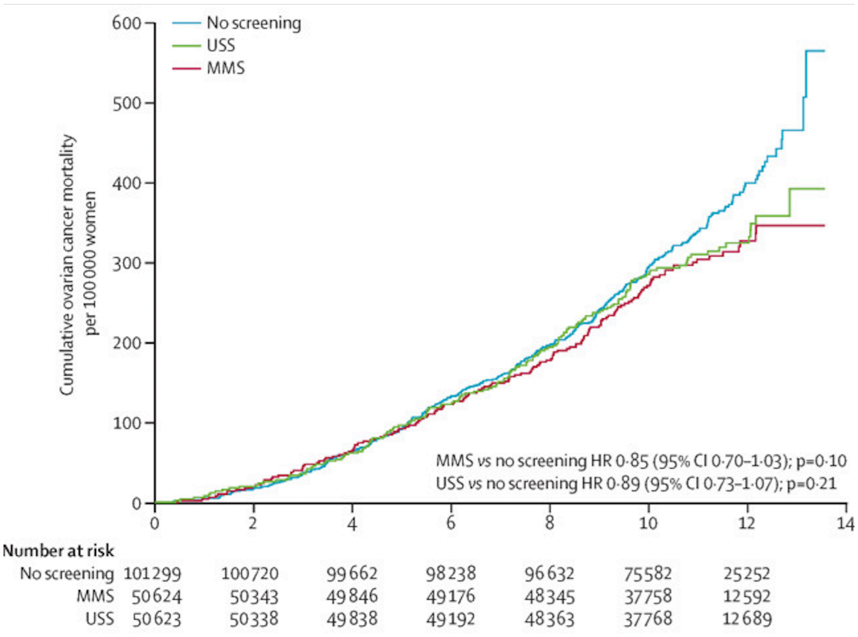
OVARIAN CANCER SCREENING TRIAL

- Primary analysis: Cox regression (proportional hazards)
  - MMS vs. no screening: Mortality reduction =

    (1 – HR)100 = 15% (95% CI: -1% – 33%) P = .10
  - USS vs. no screening: Mortality reduction =

    (1 – HR) 100 = 11% (95% CI: -7% - 27%) P = .21

OVARIAN CANCER SCREENING TRIAL



MMS *vs* no screening HR 0·85 (95% CI 0·70–1·03); p=0·10
USS *vs* no screening HR 0·89 (95% CI 0·73–1·07); p=0·21

Number at risk

No screening	101 299	100 720	99 662	98 238	96 632	75 582	25 252
MMS	50 624	50 343	49 846	49 176	48 345	37 758	12 592
USS	50 623	50 338	49 838	49 192	48 363	37 768	12 689

OVARIAN CANCER SCREENING TRIAL

- Secondary analyses, excluding prevalent cases:
- Post-hoc Weighted* logrank test:
  - MMS mortality reduction = 22% (3-38%) P = .023
  - USS mortality reduction = 20% (0 – 35%) P = .049

* by pooled cumulative mortality

SURVEY

- Trinquart et al. (JCO. 2016; 20; 34(15):1813–1819) surveyed oncology RCTs reported in five journals during the last six months of 2014.
  - 54 trials, 33,212 patients
  - Reconstructed data
  - 13 (24%) had evidence of non-proportional hazards
  - Compared tests based on HR treatment effect with tests based on ratio and difference of RMST.
  - Statistical significance in agreement between HR-based and RMST-based tests for 53 out of 54 trials.

OUTLINE

- Limitations of proportional hazards
- Other contrasts based on functionals of S(t)
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- **Other metrics to describe the distance between survival curves**
  - **Weighted difference in S(t)**
  - Maximum difference (Kolmogorov – Smirnov)
  - Integrated squared difference (Cramér von Mises)

ANOTHER OPTION: METRICS

- Tests based on detecting consistent differences between survival curves or hazard across time lose power when the hazards or survival curves cross.
- Weighting can focus on a time period when direction of differences is consistent.
- Other metrics can measure distance between survival functions or hazard functions in a way that does not require the direction of differences to be consistent
- Tests based on them can have more power to detect a difference when survival functions or hazards cross. (Need to think about whether the difference detected is of interest.)

METRICS

- Weighted difference between Kaplan-Meier estimates (Pepe MS, Fleming TR. Biometrics. 1989;497–507).

  Choose weights based on toxicity profile, for example.

$$\sqrt{\frac{n_1 n_2}{n}}\int_0^\infty \hat{w}(t)[\hat{S}_2(t)-\hat{S}_1(t)]dt$$

- Weighted difference between Kaplan-Meier estimates with adaptively chosen weights (Uno et al. Statistics in Medicine, 2015; 34(28):3680–3695).

  Hard to know what parameter is being compared.

OUTLINE

- Limitations of proportional hazards
- Other contrasts based on functionals of S(t)
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- **Other metrics to describe the distance between survival curves**
  - Weighted difference in S(t)
  - **Maximum difference (Kolmogorov – Smirnov)**
  - Integrated squared difference (Cramér von Mises)

METRICS

- Supremum: Tests based on the supremum of a difference of cumulative weighted hazard functions over $[0, t_m]$:

$$\sup_{t\in[0,t_m]} \sum_{i:t_i<t} W_i \frac{n_{1i}n_{2i}}{n_{1i}+n_{2i}} \left(\frac{d_{1i}}{n_{1i}} - \frac{d_{1i}}{n_{1i}}\right)$$

  - Gill, R.D. (1980). Censoring and stochastic integrals. Math. Centre Tracts 124, Mathematisch Centrum Amsterdam.
  - Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Biometrics. 1980;36(4):607–625.
  - Fleming TR, Harrington DP, O'Sullivan M. JASA. 1987;82(397):312–320.

OUTLINE

- Limitations of proportional hazards
- Other contrasts based on functionals of S(t)
  - S(t) at fixed time point
  - Quantiles (eg. median)
  - Mean survival time
  - Restricted mean survival time
- Other metrics to describe the distance between survival curves
  - Weighted difference in S(t)
  - Maximum difference (Kolmogorov – Smirnov)
  - **Integrated squared difference (Cramér von Mises)**

METRICS

- l^2: Tests based on the integrated squared difference of survival or cumulative hazard functions over $[0, t_m]$:

$$\sum_{t_i : t_i \leq t_m, \delta_i = 1} (\hat{S}_2(t_i) - \hat{S}_1(t_i))^2 d(-\hat{S}(t_i))$$

or

$$\sum_{t_i : t_i \leq t_m, \delta_i = 1} ((\hat{S}_2(t_i) - \hat{S}_1(t_i))W_i)^2 d(\hat{H}(t_i))$$

where the weight function W_i and H are functions of the asymptotic covariance of the cumulative hazard estimator at different times.

  - Koziol Biom. J. 1978;20(6):603–608.
  - Koziol, Yuh . Biom. J. 1982;24(8):743–750.
  - Schumacher. International Statistical Review 1984;52(3):263–281.

ISSUE

- Hard to think of a good scientific hypothesis that specifies which of these metrics and associated tests is consistent with the hypothesis.
- Large temptation to choose the type of test after looking at the data and noticing crossing hazards or crossing survival functions in the search for a powerful test.
- Scientific hypotheses more likely to be consistent with a difference between functionals of the survival function S(t).

OTHER POSSIBILITIES

- Test based on Cox model with time-dependent interaction terms (time-dependent coefficients). Some on this tomorrow.
- Test based on specific richer model for how hazard ratio depends on time (Yang S, Prentice R. Biometrika. 2005;92(1):1–17).

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\theta_0\theta_\infty}{\theta_0 + (\theta_\infty - \theta_0)S_1(t)}$$

parameterized by θ_0, the limiting hazard ratio as $t \to 0$ and θ_∞, the limiting hazard ratio as $t \to \infty$

TO WATCH OUT FOR

- Base quantity to be compared (weighted sum for logrank, time, quantile or restricted mean) on what would be meaningful in the context of the trial.
- Important to choose it before looking at the data.

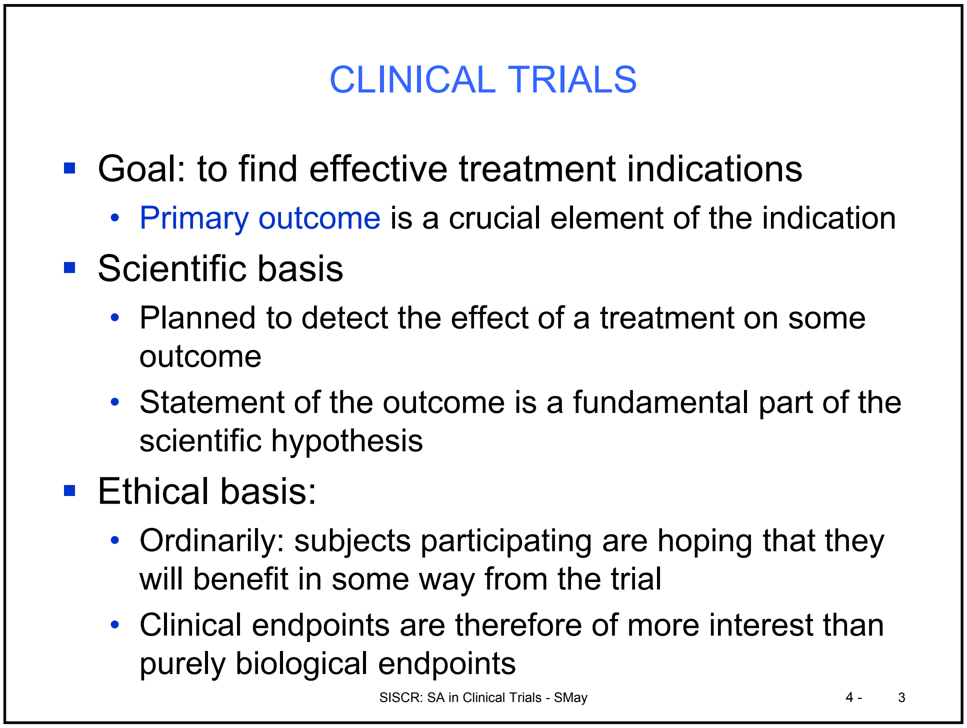
SESSION 4:
SELECTED TOPICS

Module 13: Survival Analysis for Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Susanne May, Ph.D.
Professor
Department of Biostatistics
University of Washington

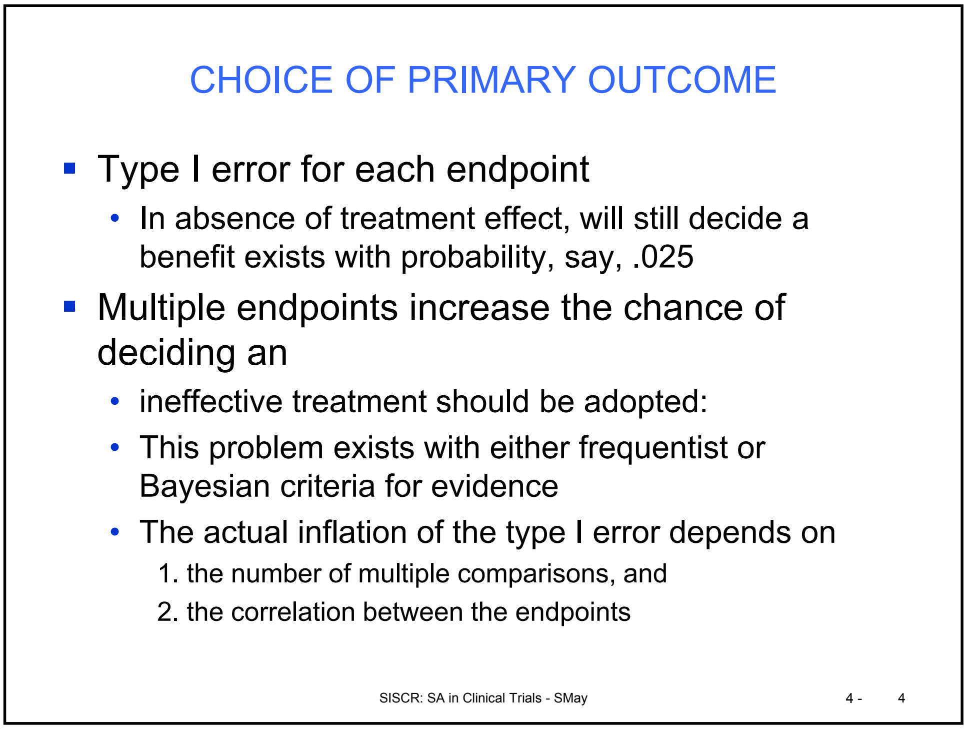
OVERVIEW

- 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



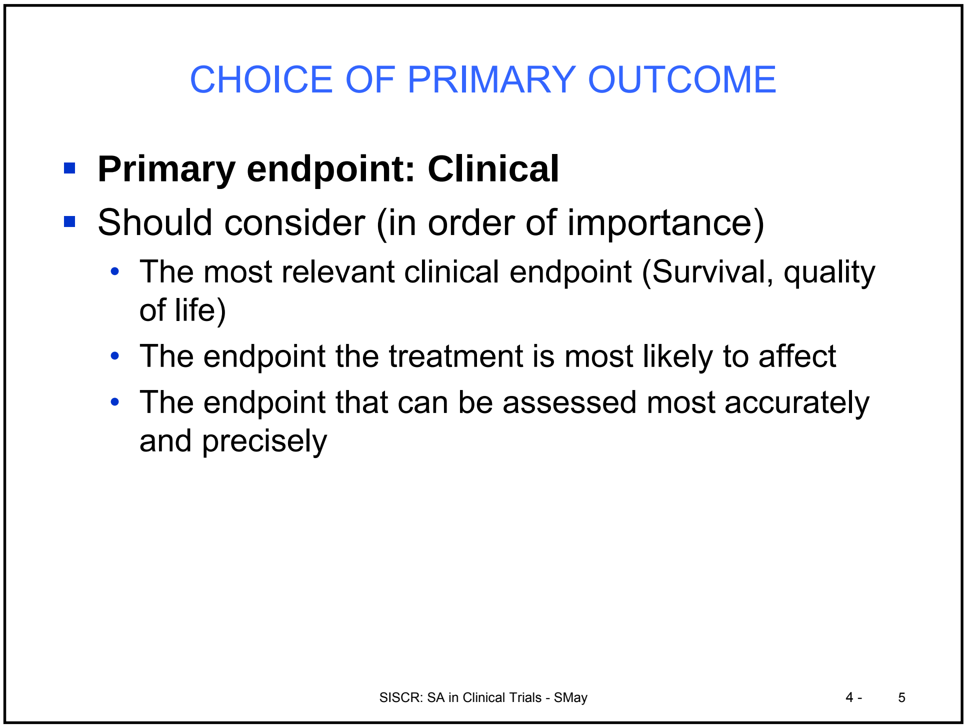
CLINICAL TRIALS

- Goal: to find effective treatment indications
  - Primary outcome is a crucial element of the indication
- Scientific basis
  - Planned to detect the effect of a treatment on some outcome
  - Statement of the outcome is a fundamental part of the scientific hypothesis
- Ethical basis:
  - Ordinarily: subjects participating are hoping that they will benefit in some way from the trial
  - Clinical endpoints are therefore of more interest than purely biological endpoints



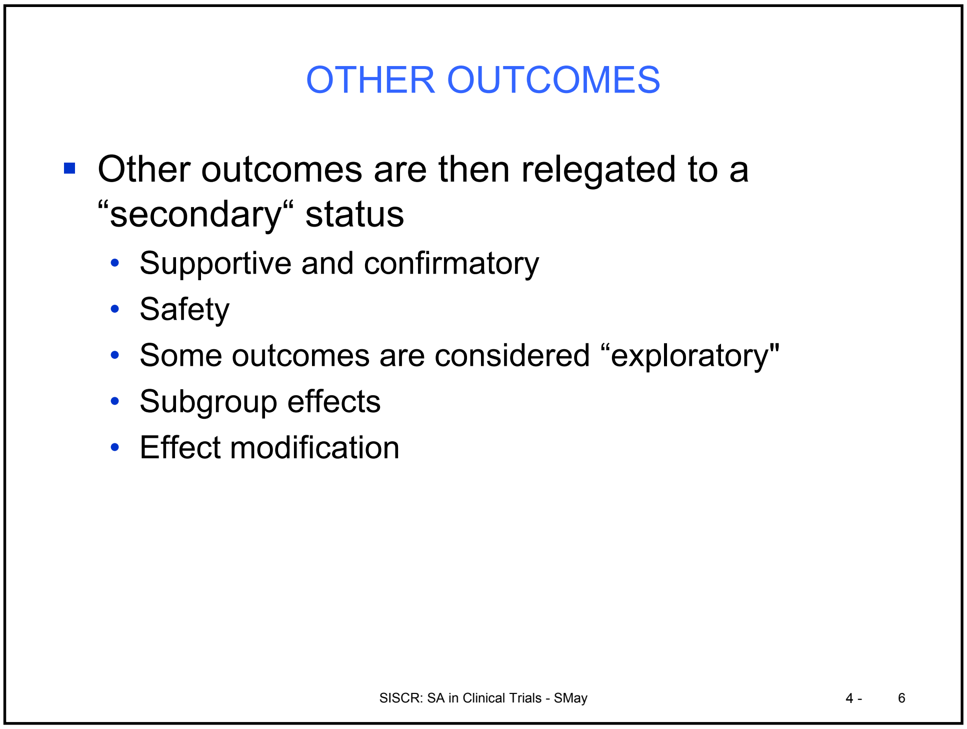
CHOICE OF PRIMARY OUTCOME

- Type I error for each endpoint
  - In absence of treatment effect, will still decide a benefit exists with probability, say, .025
- Multiple endpoints increase the chance of deciding an
  - ineffective treatment should be adopted:
  - This problem exists with either frequentist or Bayesian criteria for evidence
  - The actual inflation of the type I error depends on
    1. the number of multiple comparisons, and
    2. the correlation between the endpoints



CHOICE OF PRIMARY OUTCOME

- **Primary endpoint: Clinical**
- Should consider (in order of importance)
  - The most relevant clinical endpoint (Survival, quality of life)
  - The endpoint the treatment is most likely to affect
  - The endpoint that can be assessed most accurately and precisely



OTHER OUTCOMES

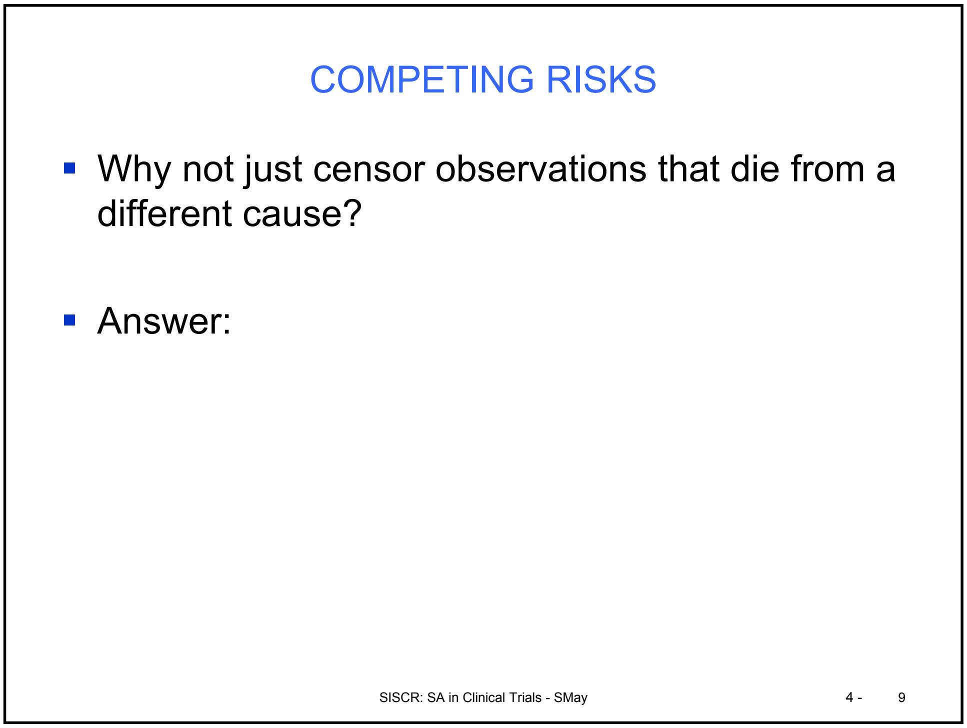
- Other outcomes are then relegated to a “secondary“ status
  - Supportive and confirmatory
  - Safety
  - Some outcomes are considered “exploratory"
  - Subgroup effects
  - Effect modification

CHOICE OF PRIMARY OUTCOME

- Should consider (in order of importance)
  - The phase of study: What is current burden of proof?
  - The most relevant clinical endpoint (Survival, quality of life)
    - Proven surrogates for relevant clinical endpoint (???)
  - The endpoint the treatment is most likely to affect
    - Therapies directed toward improving survival
    - Therapies directed toward decreasing AEs
  - The endpoint that can be assessed most accurately and precisely
    - Avoid unnecessarily highly invasive measurements
    - Avoid poorly reproducible endpoints

COMPETING RISKS

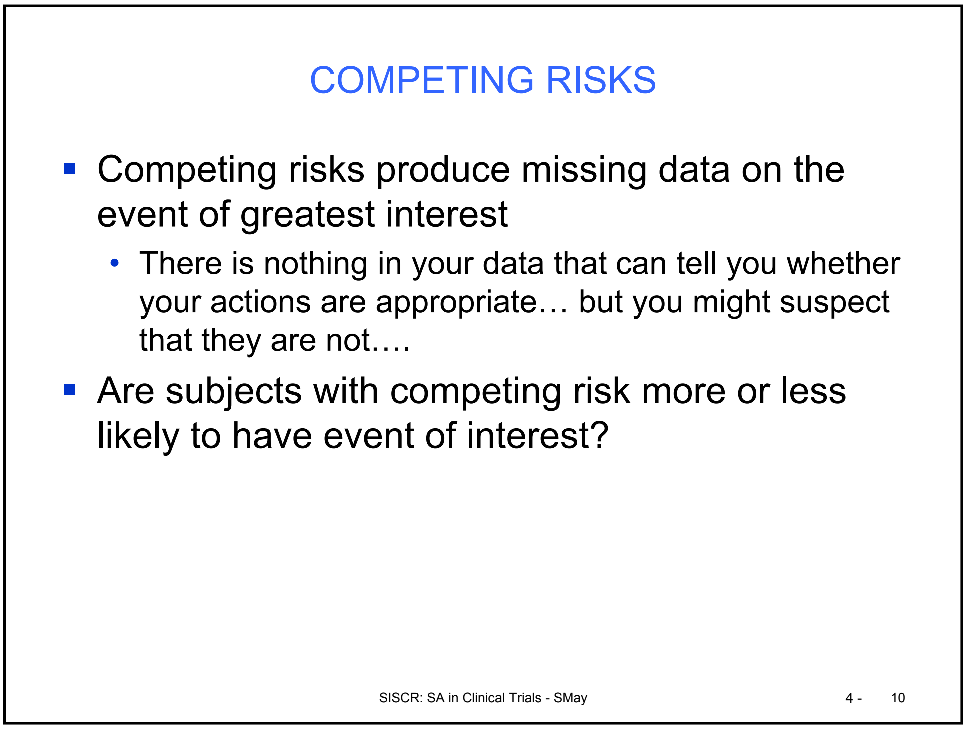
- Occurrence of some other event precludes observation of the event of greatest interest, because
  - Further observation impossible
    - E.g., death from CVD in cancer study
  - Further observation irrelevant
    - E.g., patient advances to other therapy (transplant)
- Methods
  - Event free survival: time to earliest event
  - Time to progression: censor competing risks (???)
  - All cause mortality



COMPETING RISKS

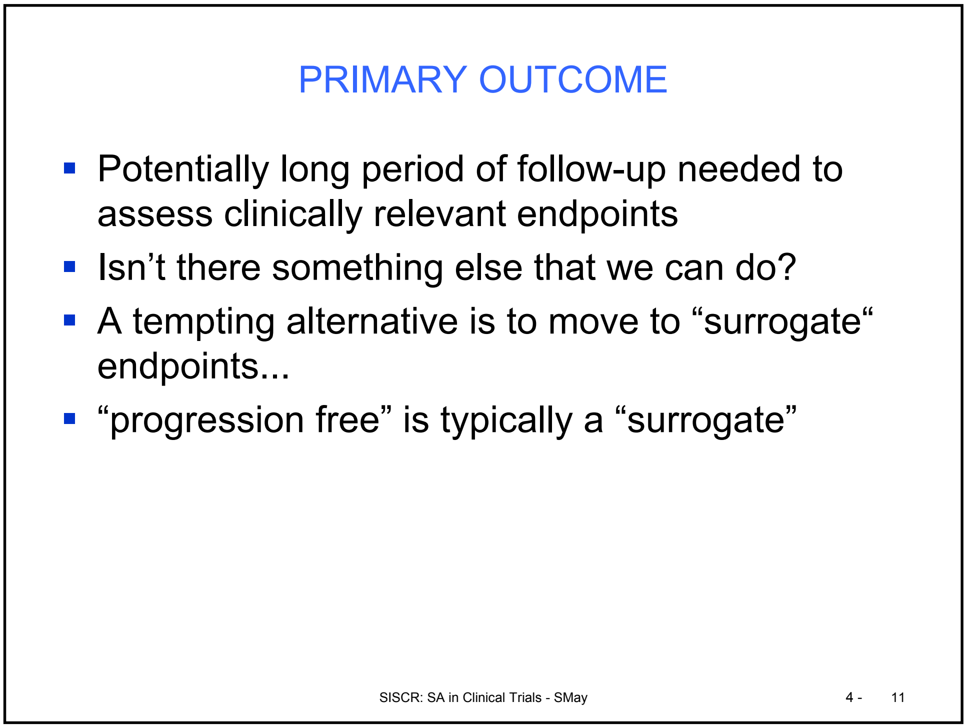
- Why not just censor observations that die from a different cause?

- Answer:



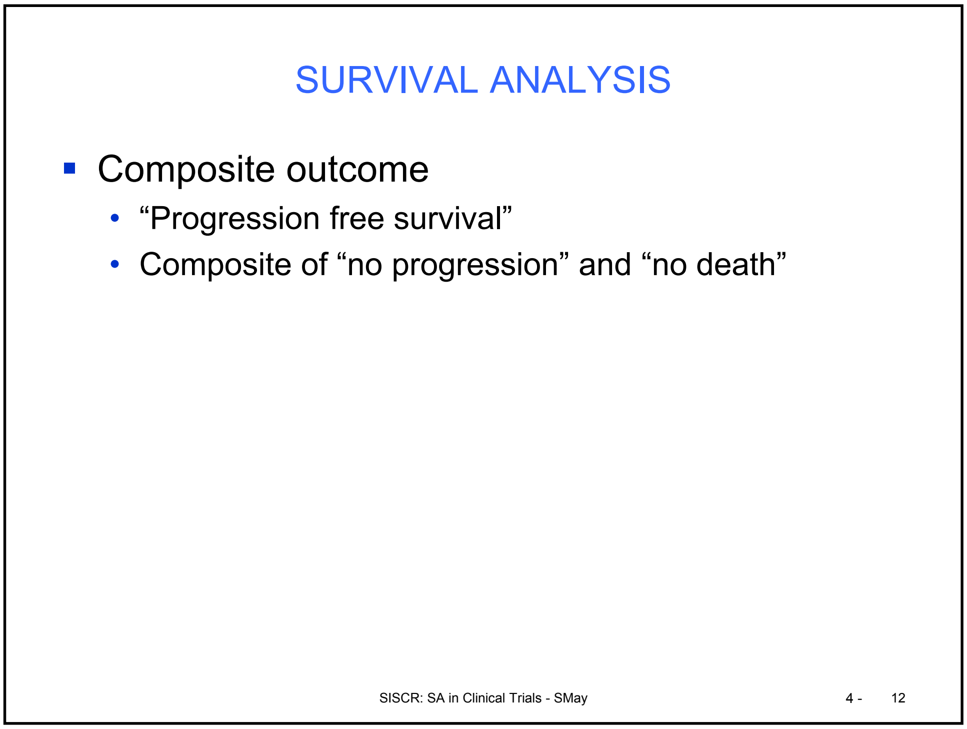
COMPETING RISKS

- Competing risks produce missing data on the event of greatest interest
  - There is nothing in your data that can tell you whether your actions are appropriate… but you might suspect that they are not….
- Are subjects with competing risk more or less likely to have event of interest?



PRIMARY OUTCOME

- Potentially long period of follow-up needed to assess clinically relevant endpoints
- Isn’t there something else that we can do?
- A tempting alternative is to move to “surrogate“ endpoints...
- “progression free” is typically a “surrogate”



SURVIVAL ANALYSIS

- Composite outcome
  - “Progression free survival”
  - Composite of “no progression” and “no death”

SURROGATE ENDPOINTS

- Hypothesized role of surrogate endpoints
  - Find a biological endpoint which
    - can be measured in a shorter timeframe,
    - can be measured precisely, and
    - is predictive of the clinical outcome
  - Use of such an endpoint as the primary measure of treatment effect will result in more efficient trials
- Treatment effects on Biomarkers
  - Establish *Biological Activity*
  - But not necessarily *overall Clinical Efficacy*
    - Ability to conduct normal activities
    - Quality of Life
    - Overall Survival

SURROGATE ENDPOINTS

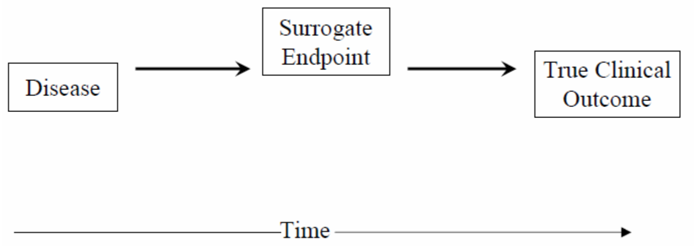
- Typically use observational data to find risk factors for clinical outcome
- Treatments attempt to intervene on those risk factors
- Surrogate endpoint for the treatment effect is then a change in the risk factor
- Establishing biologic activity does not always translate into effects on the clinical outcome
- May be treating the symptom, not the disease

EXAMPLES

- Example of surrogate endpoints
  - Cancer: tumor shrinkage
  - Coronary heart disease: cholesterol, nonfatal MI, blood pressure
  - Congestive heart failure: cardiac output
  - Arrhythmia: atrial fibrillation
  - Osteoporosis: bone mineral density
- Future surrogates?
  - Gene expression
  - Proteomics

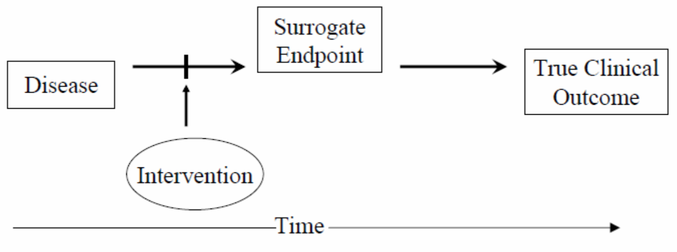
IDEAL SURROGATE

- Disease progresses to Clinical Outcome only through the Surrogate Endpoint



IDEAL SURROGATE USE

- The intervention’s effect on the Surrogate Endpoint accurately reflects its effect on the Clinical Outcome

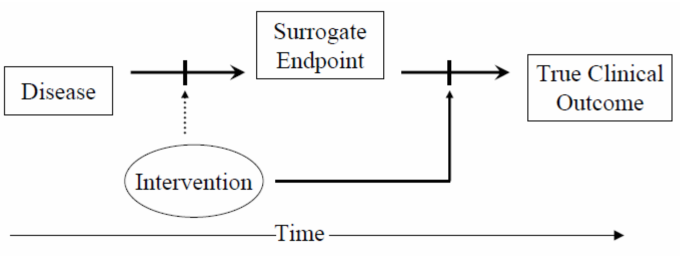


Typically

Too good to be true

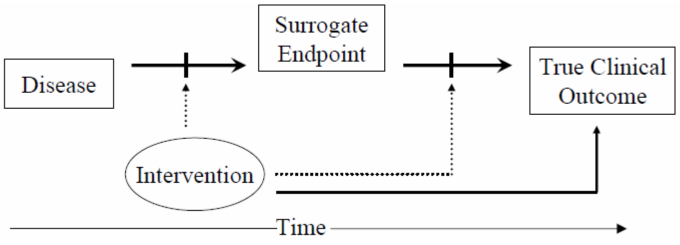
INEFFICIENT SURROGATE

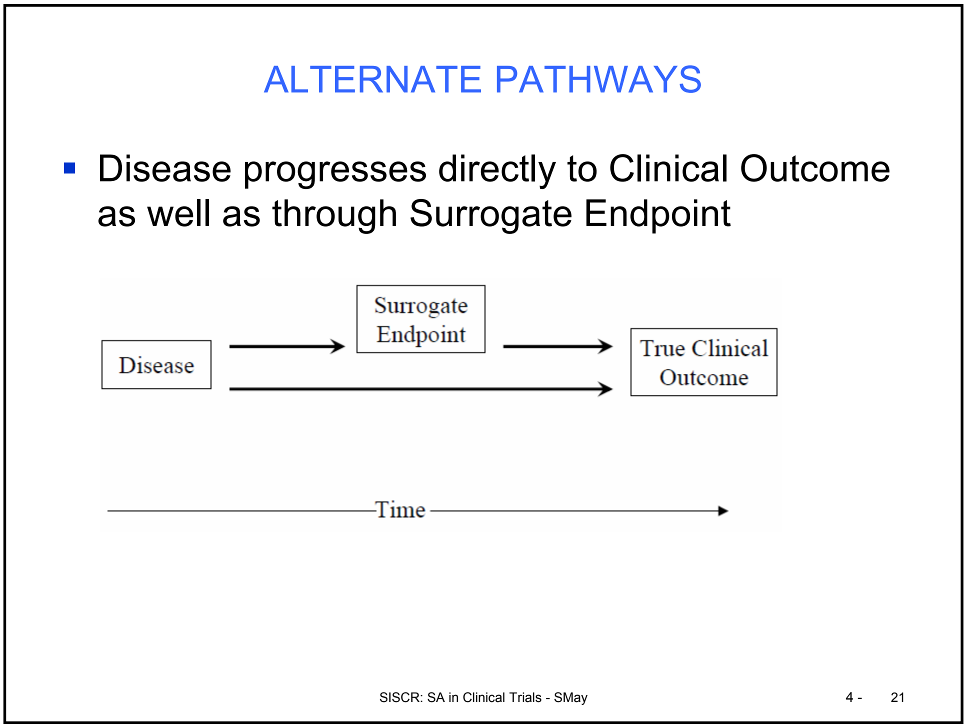
- The intervention's effect on the Surrogate Endpoint understates its effect on the Clinical Outcome



DANGEROUS SURROGATE

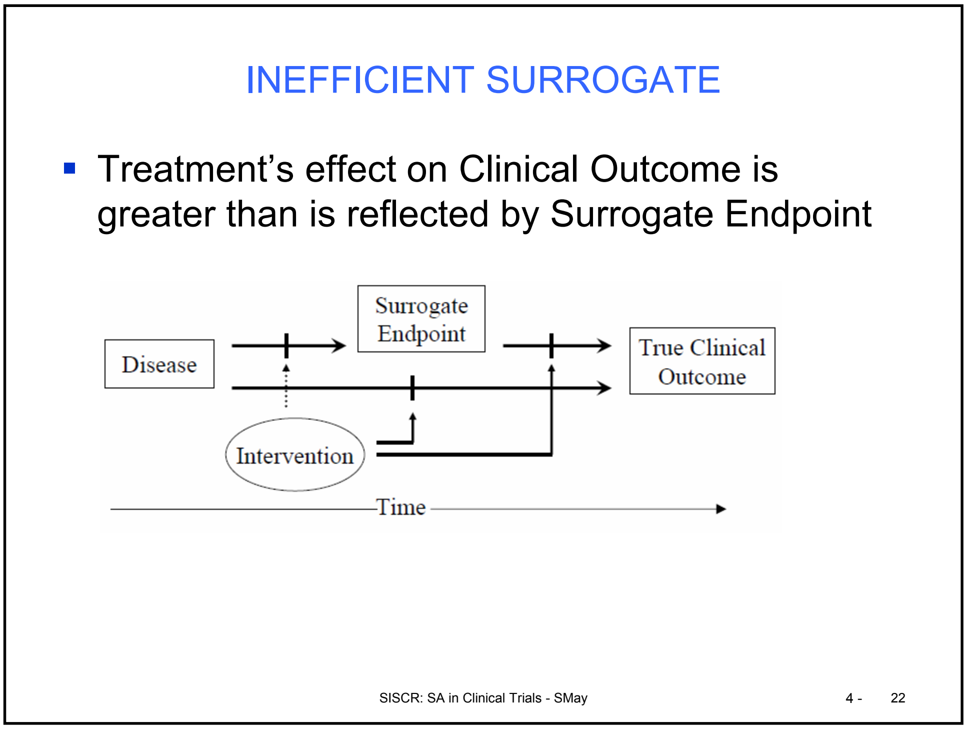
- Effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)





ALTERNATE PATHWAYS

- Disease progresses directly to Clinical Outcome as well as through Surrogate Endpoint

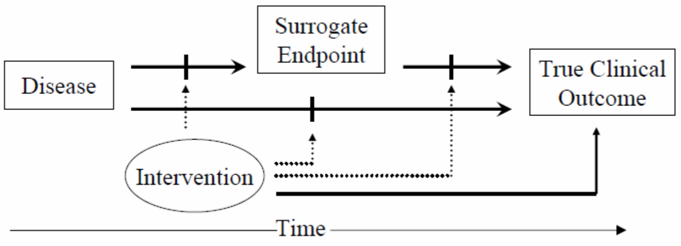


INEFFICIENT SURROGATE

- Treatment's effect on Clinical Outcome is greater than is reflected by Surrogate Endpoint

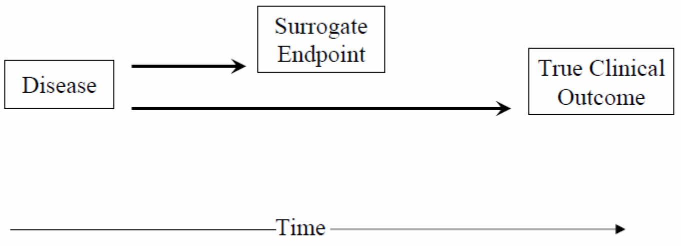
DANGEROUS SURROGATE

- The effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)



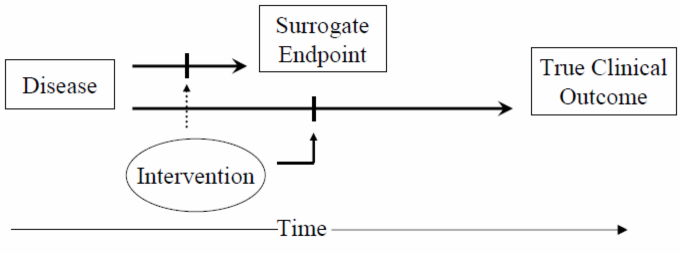
MARKER

- Disease causes Surrogate Endpoint and Clinical Outcome via different mechanisms



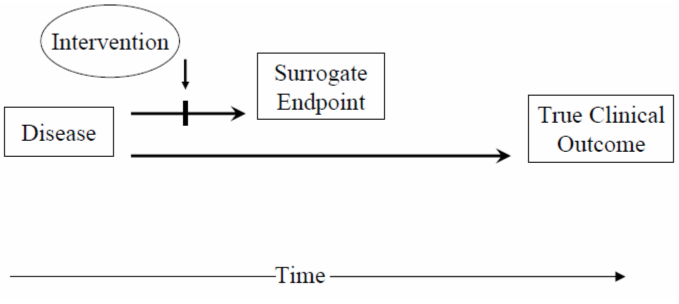
INEFFICIENT SURROGATE

- Treatment's effect on Clinical Outcome is greater than is reflected by Surrogate Endpoint



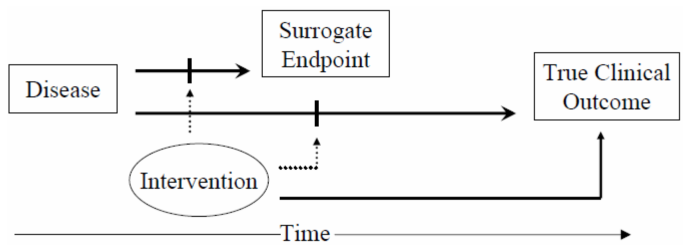
MISLEADING SURROGATE

- Effect on Surrogate Endpoint does not reflect lack of effect on Clinical Outcome



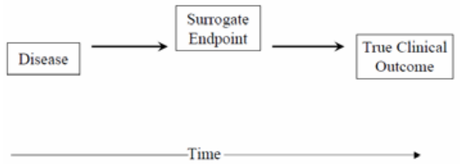
DANGEROUS SURROGATE

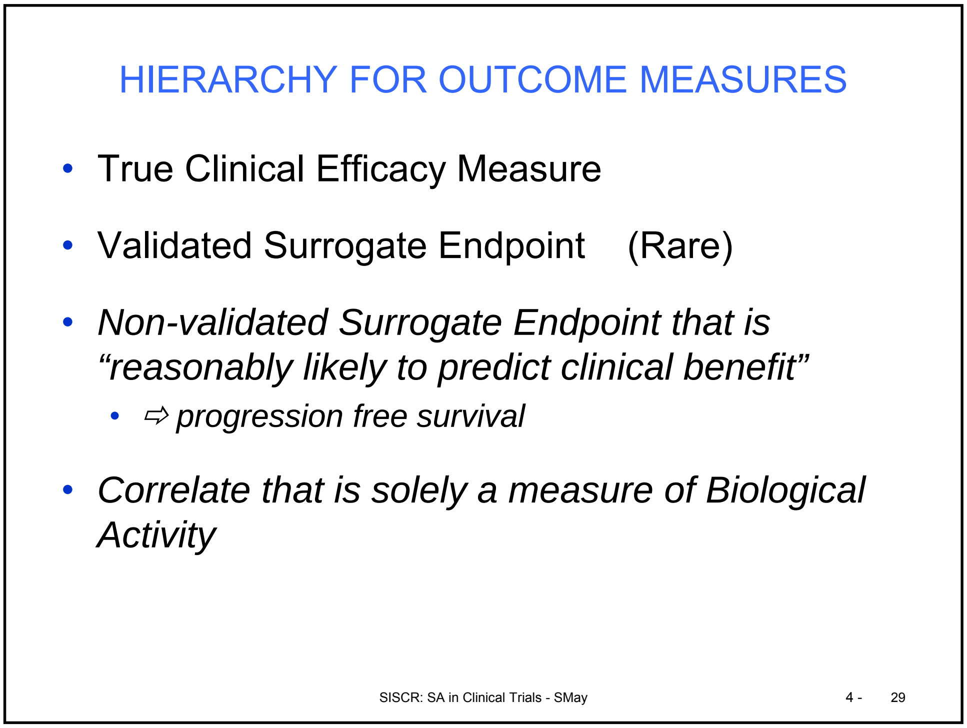
- Effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)



VALIDATION OF SURROGATE

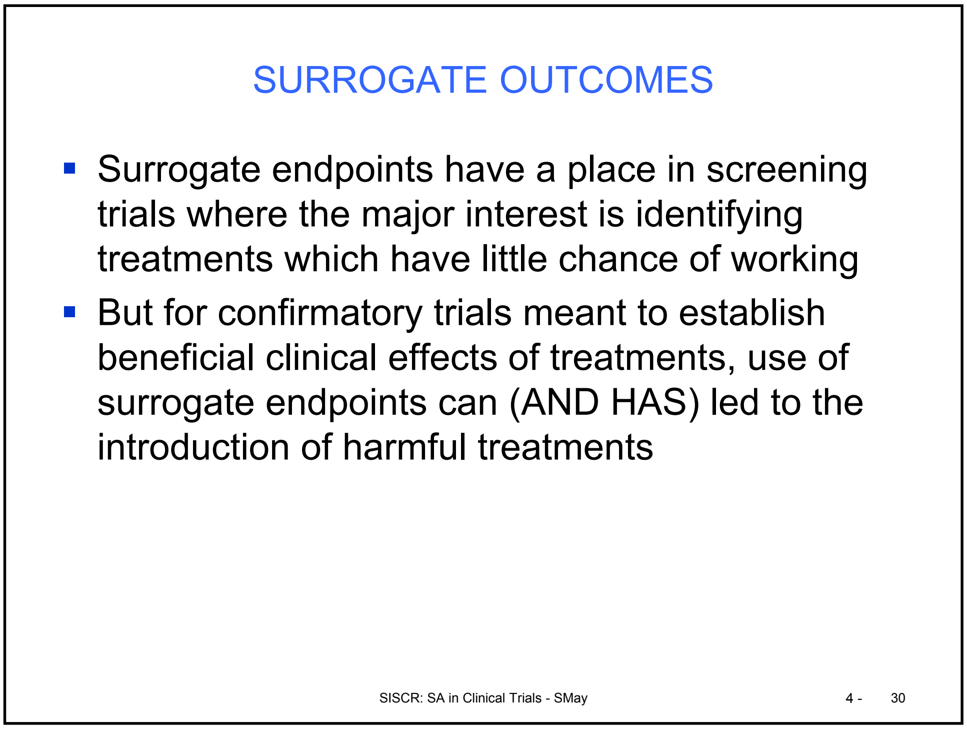
- Prentice criteria (Stat in Med, 1989)
- To be a direct substitute for a clinical benefit endpoint on inferences of superiority and inferiority
  - The surrogate endpoint must be correlated with the clinical outcome
  - The surrogate endpoint must fully capture the net effect of treatment on the clinical outcome





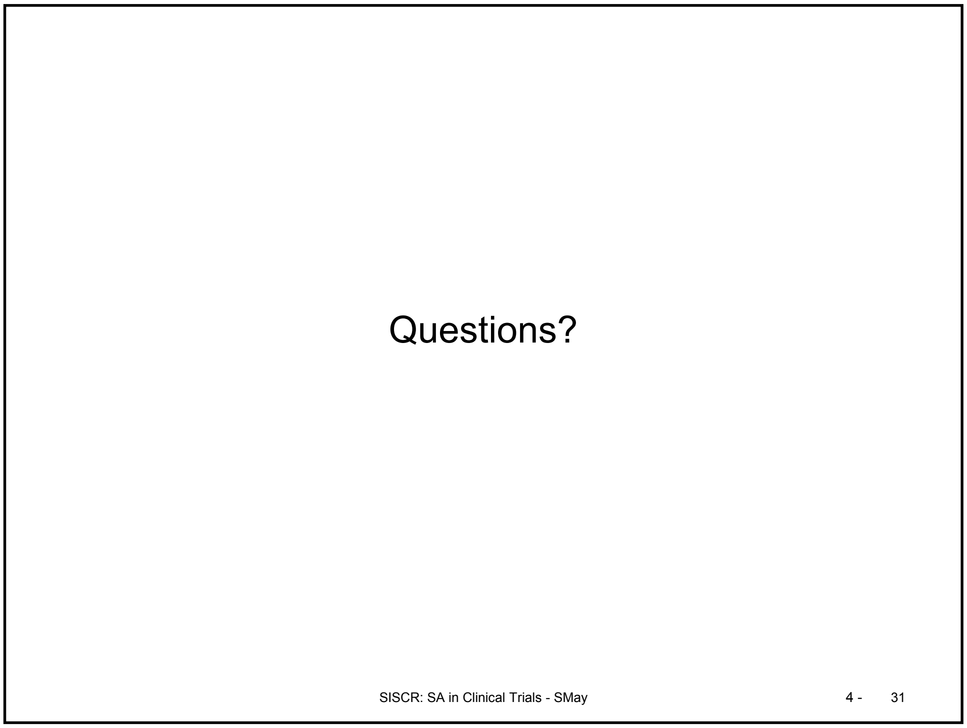
HIERARCHY FOR OUTCOME MEASURES

- True Clinical Efficacy Measure
- Validated Surrogate Endpoint (Rare)
- *Non-validated Surrogate Endpoint that is "reasonably likely to predict clinical benefit"*
    - ⇨ *progression free survival*
- *Correlate that is solely a measure of Biological Activity*

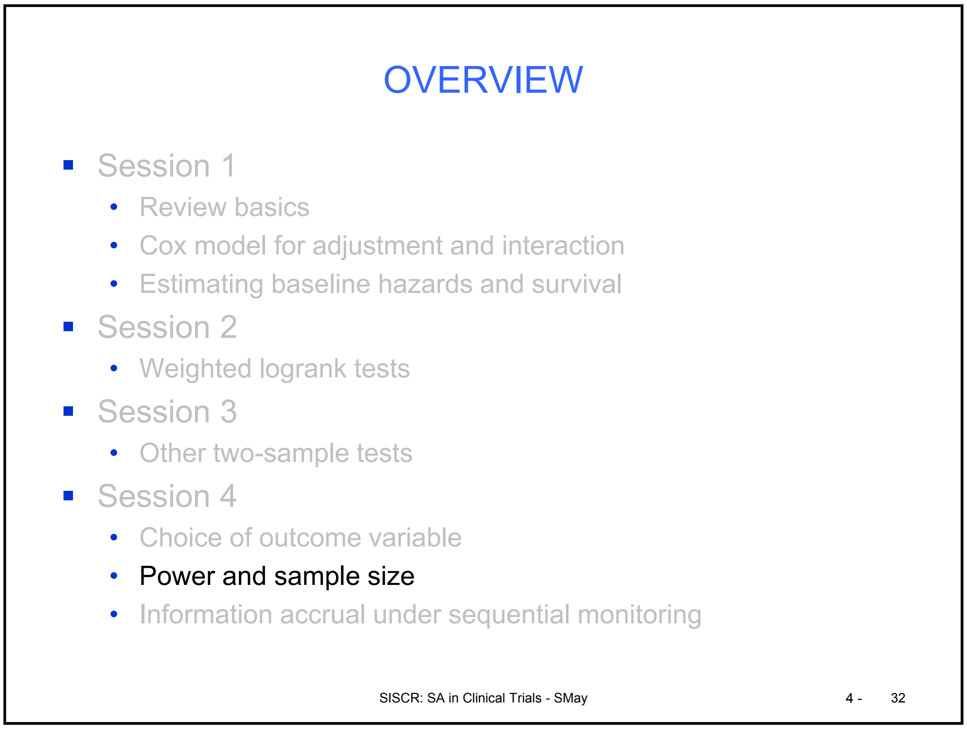


SURROGATE OUTCOMES

- Surrogate endpoints have a place in screening trials where the major interest is identifying treatments which have little chance of working
- But for confirmatory trials meant to establish beneficial clinical effects of treatments, use of surrogate endpoints can (AND HAS) led to the introduction of harmful treatments



Questions?

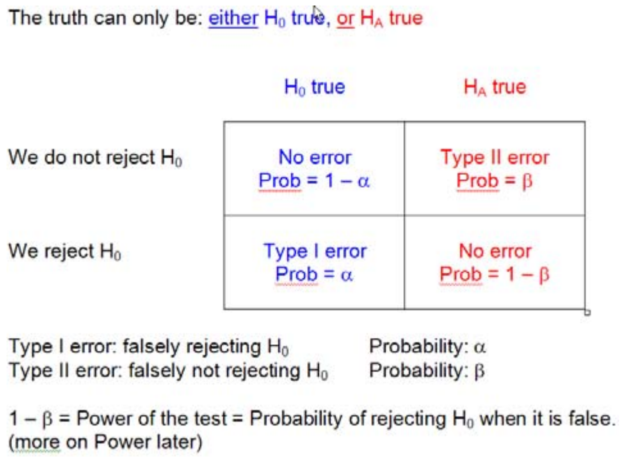


OVERVIEW

- 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

SAMPLE SIZE / POWER

- Hypothesis testing



The truth can only be: either H_0 true, or H_A true

	H_0 true	H_A true
We do not reject H_0	No error Prob = $1 - \alpha$	Type II error Prob = β
We reject H_0	Type I error Prob = α	No error Prob = $1 - \beta$

Type I error: falsely rejecting H_0 Probability: α
Type II error: falsely not rejecting H_0 Probability: β

$1 - \beta$ = Power of the test = Probability of rejecting H_0 when it is false.
(more on Power later)

GOAL

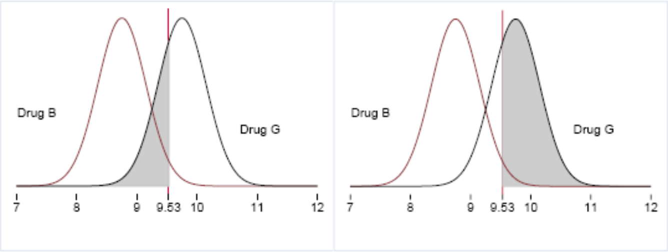
- Main goals of power / sample size calculations
- Avoid sample size that is TOO small
- Avoid sample size that is TOO large
- Ethical issues
- Financial issues

SAMPLE SIZE / POWER

- Normally distributed outcome

Shaded area represents β,
the probability of type II error

$$n = \sigma^2 \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\mu_a - \mu_0\right)^2}$$



Shaded area represents $1-\beta$,
the power of the test.

SAMPLE SIZE / POWER

- How does this change for survival analysis?
  - Because of censoring
  - Two-step process
  - Determine total number of events
    - Specify hypothesis in terms of statistical parameters, their estimators and variance
    - Clinically important change in the parameters
    - Specify Type I and Type II error probabilities
    - Solve for sample size
  - Determine total number of observations
  - Length of recruitment and follow-up

SAMPLE SIZE / POWER

- Schoenfeld (1983)

$$m = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\theta^2 \pi \left(1 - \pi\right)} \qquad HR = \exp\left(\theta\right)$$

- $z_{\alpha/2}$ corresponding percentage points from
  z_{β} the standard normal
  π fraction of subjects in the first group

With equal allocation ($m_1 = m_2$) $m = \dfrac{4\left(z_{\alpha/2} + z_{\beta}\right)^2}{\theta^2}$

EXAMPLE

- Assume: HR = 0.75
- Alpha = 0.05
- Power = 80%
- $\beta = 0.2$
- ⇨ $379.5 = \dfrac{4\left(1.96 + 0.842\right)^2}{\left[\ln\left(0.75\right)\right]^2}$
- Would be the right sample size if 380 subjects are randomized at time zero and all followed until the event occurs ⇨ not realistic

EXAMPLE

- Need to adjust *m* by dividing by an estimate of the overall probability of death by the end of the study
- Might have an estimate from past studies?
- Might have K-M estimate of baseline survival function
  $\hat{S}_0(t)$
- Estimate can be used to approximate the survival function under the new treatment and a PH model $\hat{S}_1(t) = \left[\hat{S}_0(t)\right]^{\exp(\theta)}$

EXAMPLE

- If subjects uniformly recruited over the first "a" years
- And then followed for an additional "f" years
- An estimate of the probability of death at the end of the study a + f is

$$\bar{F}(a+f) = 1 - \frac{1}{6}\left[\bar{S}(f) + 4\bar{S}(0.5a+f) + \bar{S}(a+f)\right]$$

$$\bar{S}(t) = \pi \times \hat{S}_0(t) + (1-\pi) \times \hat{S}_1(t)$$

- π fraction of subjects in the standard tx

EXAMPLE

- The estimated number of subjects that must be followed is

$$n = \frac{m}{\bar{F}(a+f)}$$

$$= \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\bar{F}(a+f)\theta^2\pi(1-\pi)}$$

SAMPLE SIZE / POWER

- Suppose we enroll subjects for 2 years
- And then follow them for an additional 3 years
- Also, we know (from previous research)

$$\hat{S}_0(3) = 0.7, \hat{S}_0(4) = 0.65 \text{ and } \hat{S}_0(5) = 0.55$$

- Then $\hat{S}_1(3) = 0.765 = [0.7]^{0.75}$

$$\hat{S}_1(4) = 0.724 = [0.65]^{0.75}$$

$$\hat{S}_1(5) = 0.639 = [0.55]^{0.75}$$

- And the average survival probabilities at these three time points are

$$\bar{S}_0(3) = 0.733, \bar{S}_0(4) = 0.687 \text{ and } \bar{S}_0(5) = 0.595$$

EXAMPLE

- The average probability of death at the end of the study is estimated as

$$\bar{F}(5) = 0.321 = 1 - \frac{1}{6}\left[0.733 + 4 \times 0.687 + 0.595\right]$$

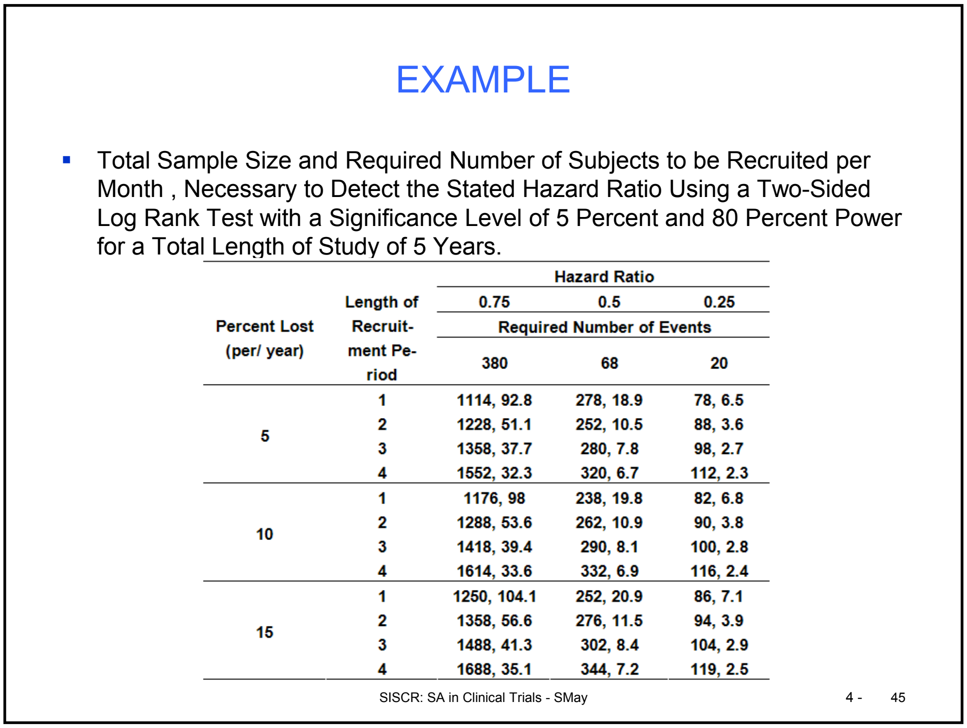
- And the total number of subjects that must be enrolled is

$$n_{total} = 1{,}183.8 = \frac{380}{0.321} \qquad n_{per-group} = 592$$

- ⇨ ~ 49-50 subjects per month need to be enrolled
- Slight differences in estimated numbers possible due to different approaches of different software packages

SAMPLE SIZE / POWER

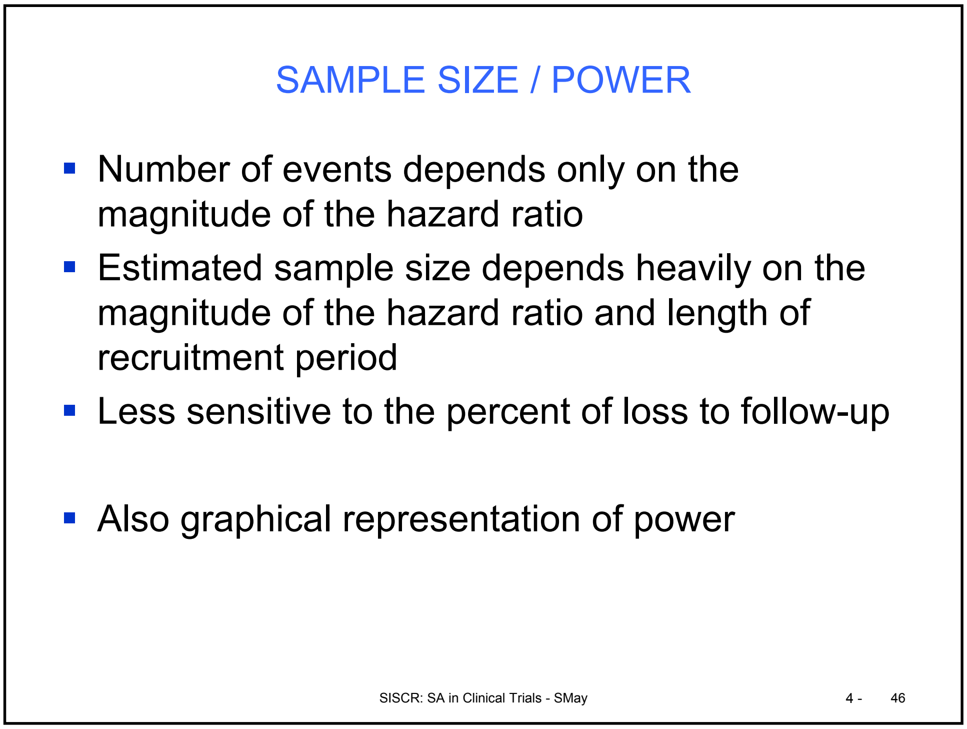
- Factors
  - Effect size
  - Allocation ratio
  - Alpha
  - Power
  - Baseline survival distribution
  - Length of recruitment
  - Length of follow-up period
  - Loss to follow-up
  - Number of events/censored observations



EXAMPLE

- Total Sample Size and Required Number of Subjects to be Recruited per Month , Necessary to Detect the Stated Hazard Ratio Using a Two-Sided Log Rank Test with a Significance Level of 5 Percent and 80 Percent Power for a Total Length of Study of 5 Years.

		Hazard Ratio		
	Length of	0.75	0.5	0.25
Percent Lost (per/ year)	Recruit-ment Pe-riod	Required Number of Events		
		380	68	20
5	1	1114, 92.8	278, 18.9	78, 6.5
	2	1228, 51.1	252, 10.5	88, 3.6
	3	1358, 37.7	280, 7.8	98, 2.7
	4	1552, 32.3	320, 6.7	112, 2.3
10	1	1176, 98	238, 19.8	82, 6.8
	2	1288, 53.6	262, 10.9	90, 3.8
	3	1418, 39.4	290, 8.1	100, 2.8
	4	1614, 33.6	332, 6.9	116, 2.4
15	1	1250, 104.1	252, 20.9	86, 7.1
	2	1358, 56.6	276, 11.5	94, 3.9
	3	1488, 41.3	302, 8.4	104, 2.9
	4	1688, 35.1	344, 7.2	119, 2.5

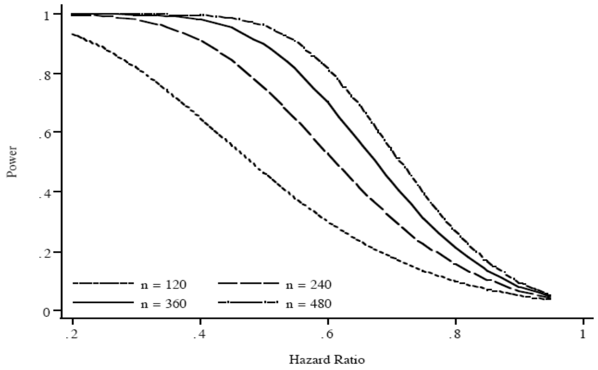


SAMPLE SIZE / POWER

- Number of events depends only on the magnitude of the hazard ratio
- Estimated sample size depends heavily on the magnitude of the hazard ratio and length of recruitment period
- Less sensitive to the percent of loss to follow-up
- Also graphical representation of power

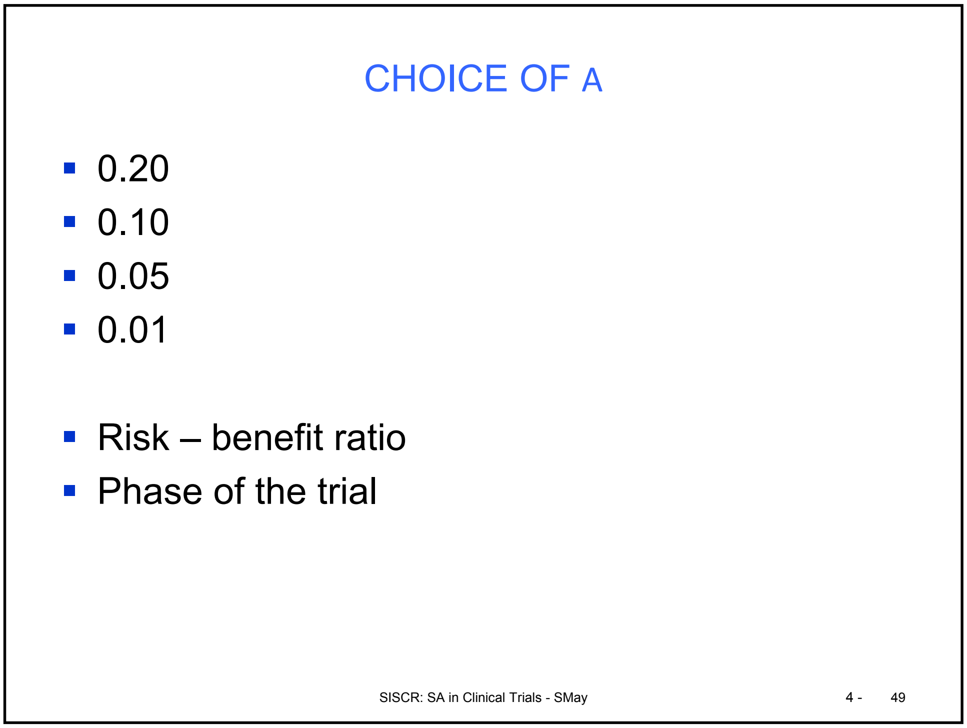
EXAMPLE

- Estimated power of a two sided five percent level of significance Log Rank test to detect the hazard ratio using the stated sample size



TWO-SIDED VS ONE-SIDED

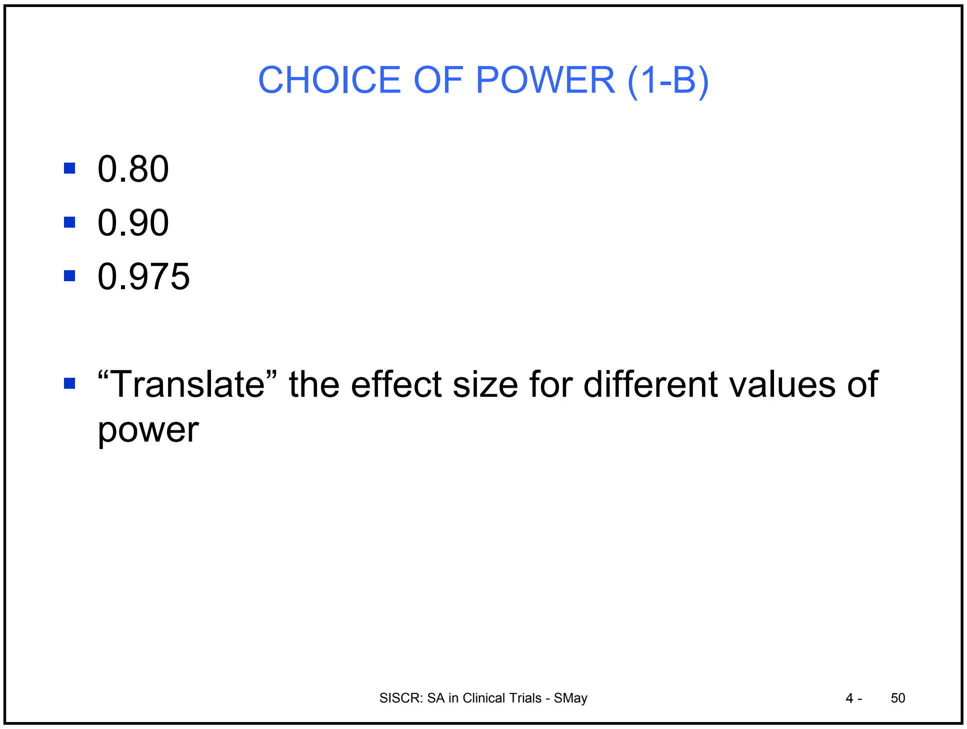
- Symmetry?
- Two-sided α = 0.05 ⇔ one-sided α = 0.025



CHOICE OF A

- 0.20
- 0.10
- 0.05
- 0.01

- Risk – benefit ratio
- Phase of the trial



CHOICE OF POWER (1-B)

- 0.80
- 0.90
- 0.975

- “Translate” the effect size for different values of power

EFFECT SIZE

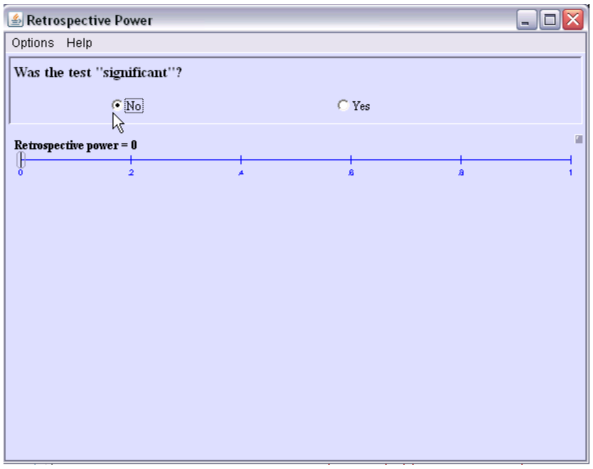
- **How to determine the “target” effect size?**
- Clinically meaningful
- Achievable

POST-HOC POWER

- After the study is done…. (usually) with a non-significant result….
- How much power did the study have to detect the result that was seen ….?

POST-HOC POWER

- <http://www.stat.uiowa.edu/~rlenth/Power/>



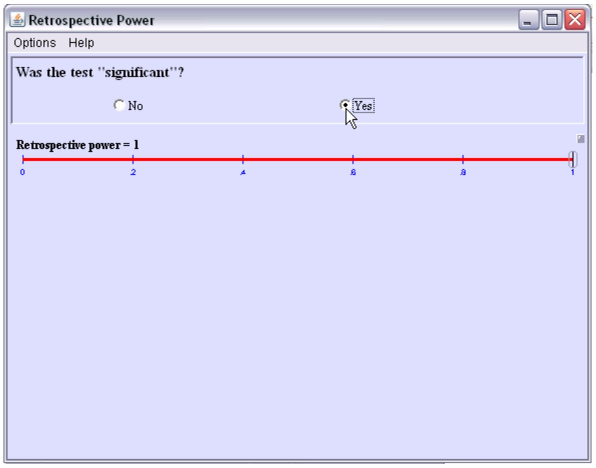
Was the test "significant"?

● No ○ Yes

Retrospective power = 0

POST-HOC POWER

- <http://www.stat.uiowa.edu/~rlenth/Power/>



Was the test "significant"?

○ No ● Yes

Retrospective power = 1

POST-HOC POWER

- Hoenig, John M. and Heisey, Dennis M. (2001), ``The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis,'' *The American Statistician,* **55**, 19-24.
- CIs obtained at the end of the study are much more informative than post hoc power!
- Probability of precipitation…
- “LA stories”… Steve Martin … pushing his car

OVERVIEW

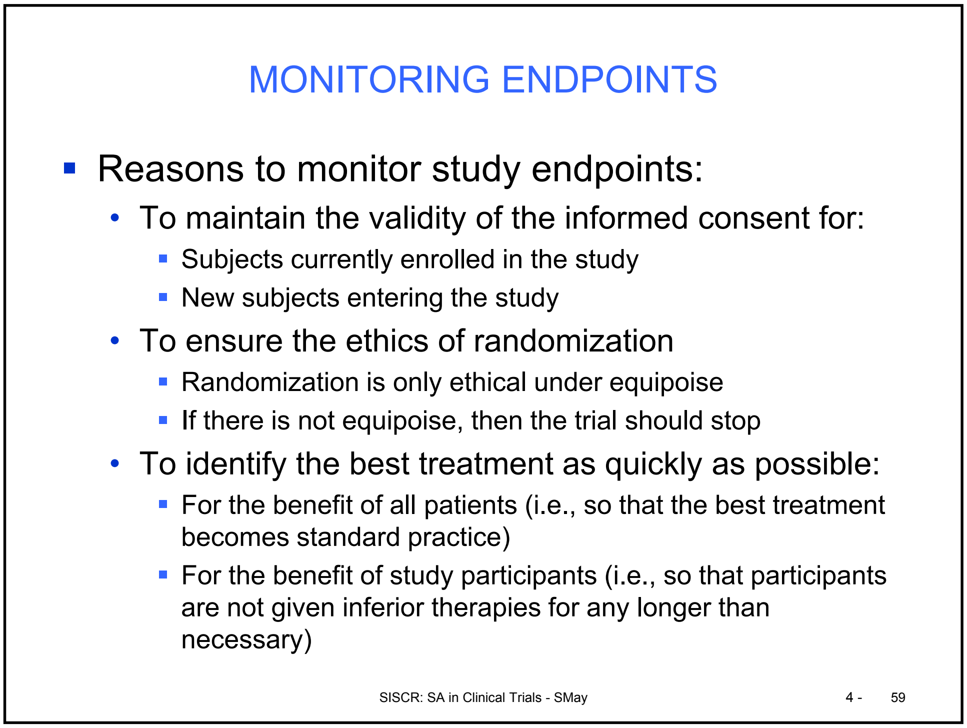
- Session 1
  - Review basics
  - Cox model for adjustment and interaction
  - Estimating baseline hazards and survival
- Session 2
  - Weighted logrank tests
- Session 3
  - Other two-sample tests
- Session 4
  - Choice of outcome variable
  - Power and sample size
  - Information accrual under sequential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

GOAL OF SEQUENTIAL MONITORING

- Develop a design for repeated data analyses
  - which satisfies the ethical need for early termination if initial results are extreme
  - while not increasing the chance of false conclusions

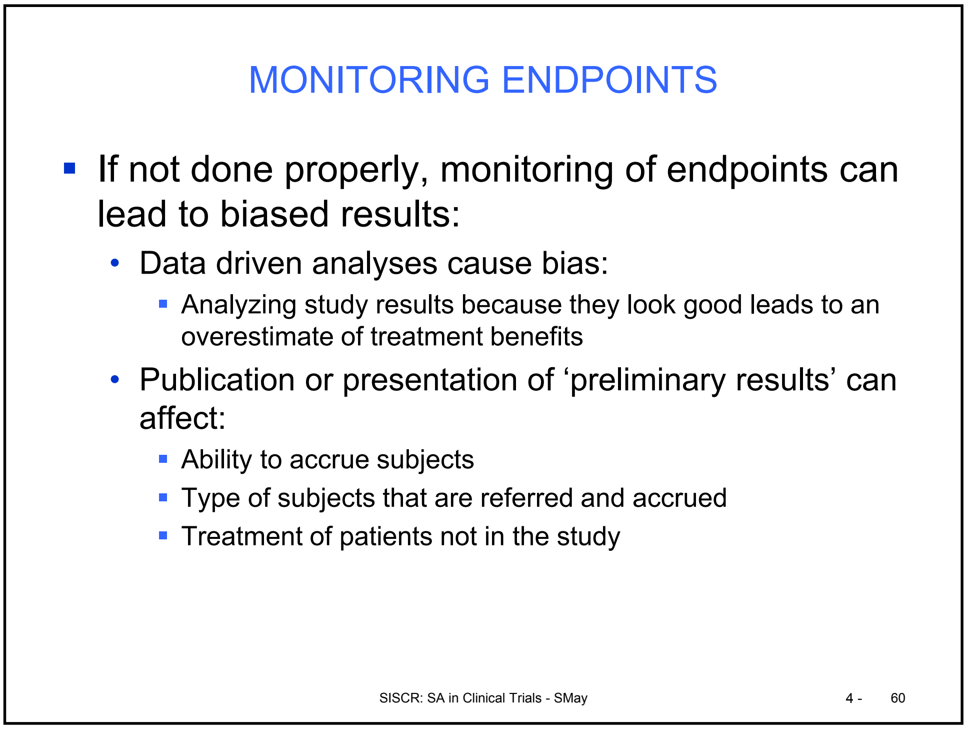
GROUP SEQUENTIAL MONITORING

- Motivation: Many trials have been stopped early:
  - Physician health study showed that aspirin reduces the risk of cardiovascular death.
  - A phase III study of tamoxifen for prevention of breast cancer among women at risk for breast cancer showed a reduction in breast cancer incidence.
  - A phase III study of anti-arrhythmia drugs for prevention of death in people with cardiac arrhythmia stopped due to excess deaths with the anti-arrhythmia drugs.
  - Women’s Health Initiative: Hormones cause heart disease.



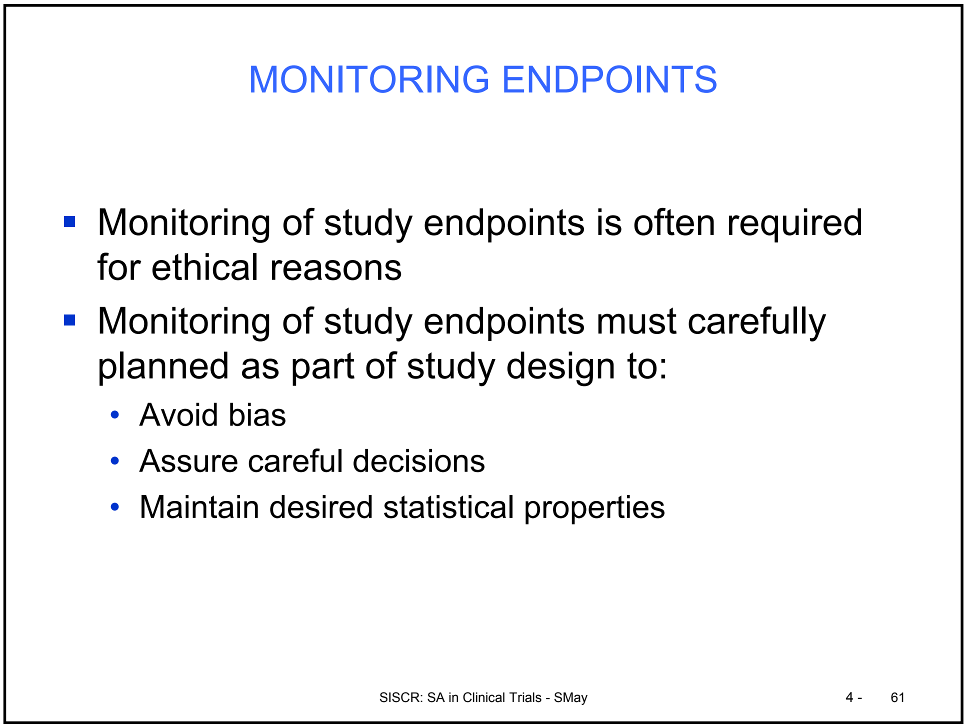
MONITORING ENDPOINTS

- Reasons to monitor study endpoints:
  - To maintain the validity of the informed consent for:
    - Subjects currently enrolled in the study
    - New subjects entering the study
  - To ensure the ethics of randomization
    - Randomization is only ethical under equipoise
    - If there is not equipoise, then the trial should stop
  - To identify the best treatment as quickly as possible:
    - For the benefit of all patients (i.e., so that the best treatment becomes standard practice)
    - For the benefit of study participants (i.e., so that participants are not given inferior therapies for any longer than necessary)



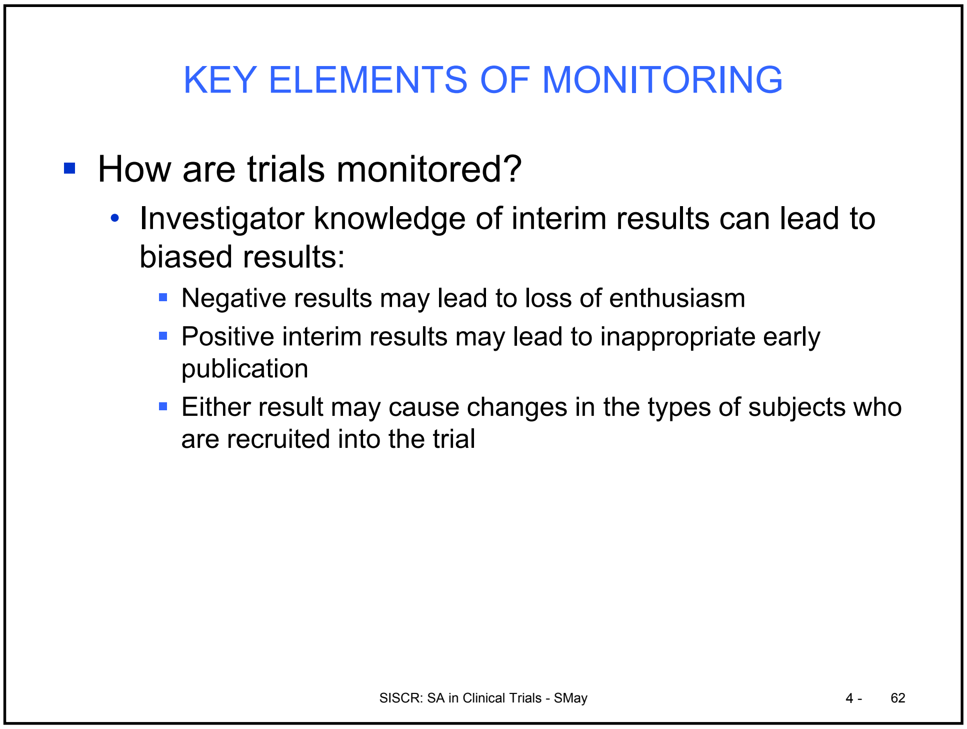
MONITORING ENDPOINTS

- If not done properly, monitoring of endpoints can lead to biased results:
  - Data driven analyses cause bias:
    - Analyzing study results because they look good leads to an overestimate of treatment benefits
  - Publication or presentation of ‘preliminary results’ can affect:
    - Ability to accrue subjects
    - Type of subjects that are referred and accrued
    - Treatment of patients not in the study



MONITORING ENDPOINTS

- Monitoring of study endpoints is often required for ethical reasons
- Monitoring of study endpoints must carefully planned as part of study design to:
  - Avoid bias
  - Assure careful decisions
  - Maintain desired statistical properties



KEY ELEMENTS OF MONITORING

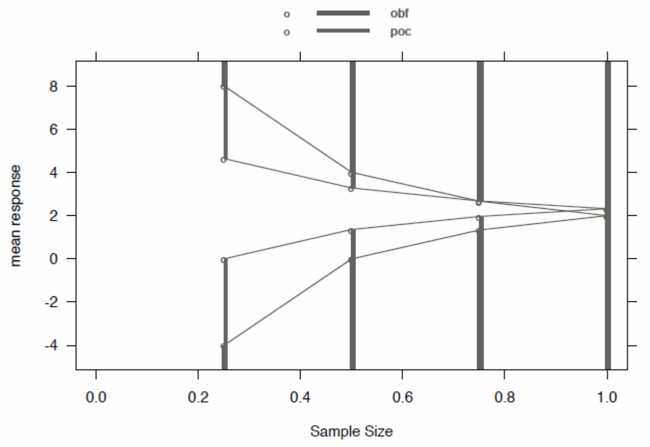
- How are trials monitored?
  - Investigator knowledge of interim results can lead to biased results:
    - Negative results may lead to loss of enthusiasm
    - Positive interim results may lead to inappropriate early publication
    - Either result may cause changes in the types of subjects who are recruited into the trial

INTERIM STATISTICAL ANALYSIS PLAN

- Typical content for ISAP:
  - Safety monitoring plan (if there are formal safety interim analyses)
    - Decision rules for formal safety analyses
    - Evaluation of decision rules (power, expected sample size, stopping probability)
    - Methods for modifying rules (changes in timing of analyses)
    - Methods for inference (bias adjusted inference)

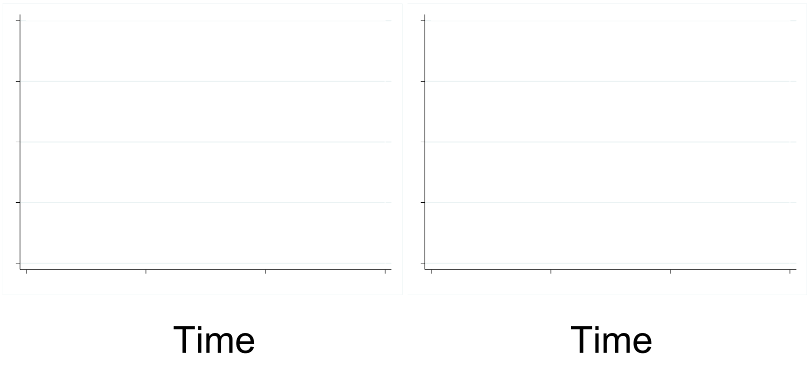
MONITORING BOUNDARIES

- Example of monitoring boundaries – note: scale



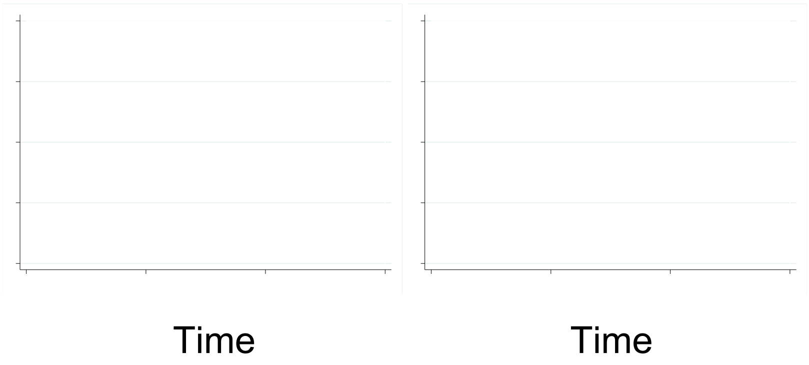
TYPICAL (NON-SURVIVAL) TRIAL

- Accrual pattern and information growth

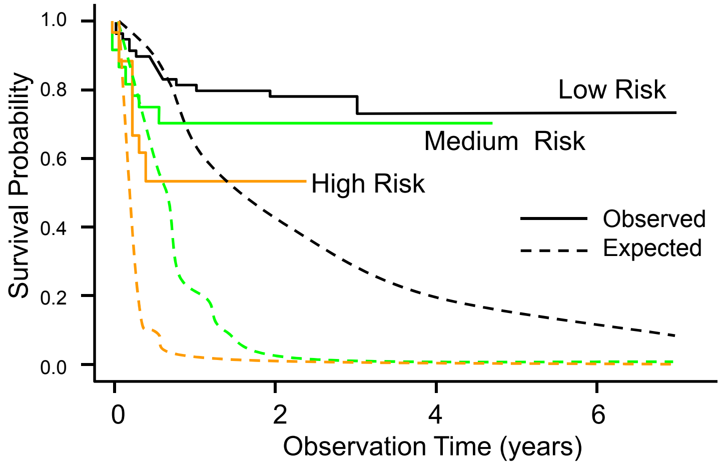


TRIAL WITH SURVIVAL ANALYSIS

- Accrual pattern and information growth

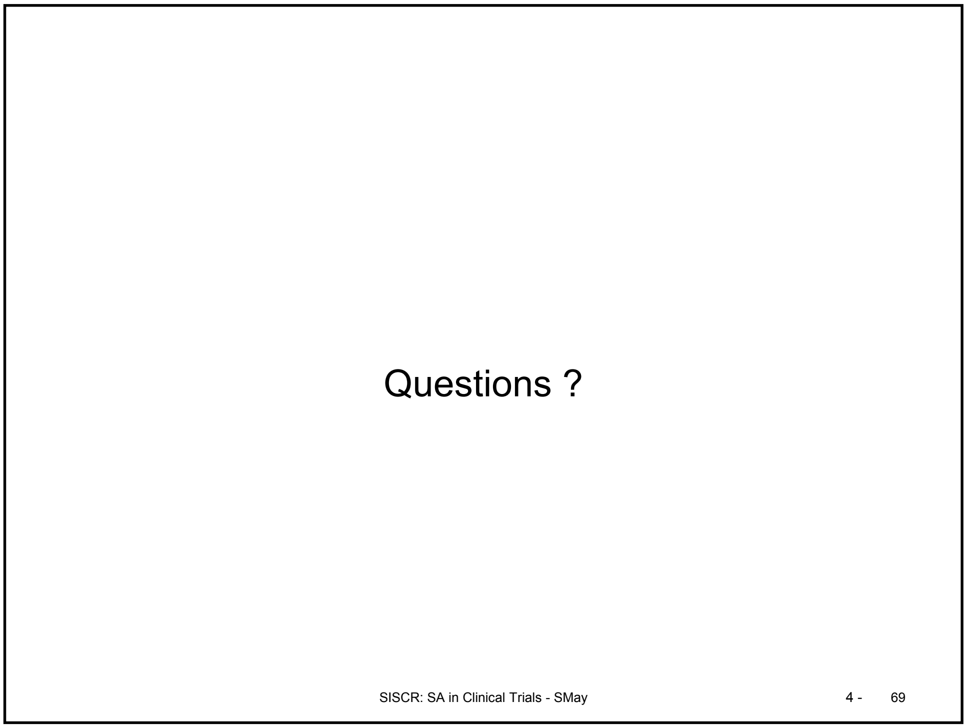


EXAMPLE



SAMPLE SIZE

- If the event rate of a trial is much lower than expected, and sample size adjustments are made to increase the number of individuals enrolled, will this affect the power of the study?



Questions ?