SESSION 4:
SELECTED TOPICS

Module 16: Survival Analysis for Clinical Trials
Summer Institute in Statistics for Clinical Research
University of Washington
June, 2017

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
  - Time-dependent covariates

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

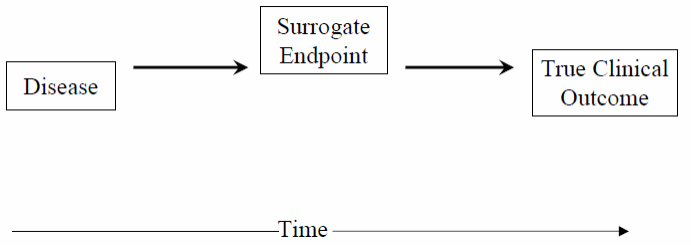
- Typically use observational data to find risk factors for clinical outcome
- Treatments attempt to intervene on those risk factors
- Surrogate endpoint for the treatment effect is then a change in the risk factor
- Establishing biologic activity does not always translate into effects on the clinical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

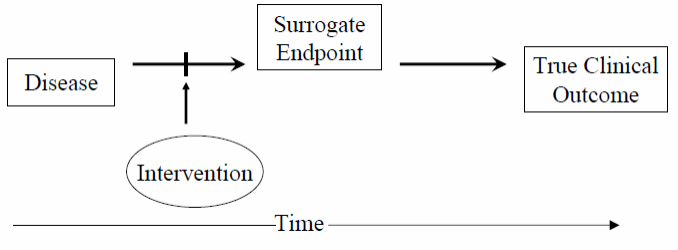
IDEAL SURROGATE

- 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

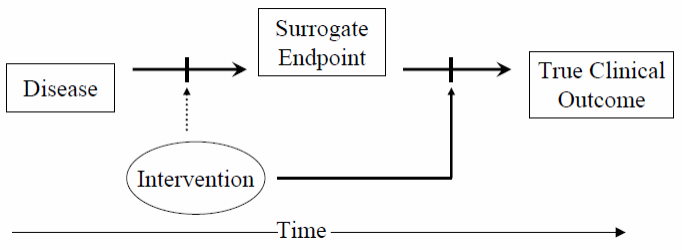


Typically

Too good to be true

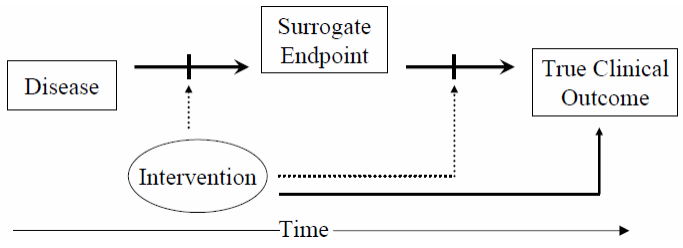
INEFFICIENT SURROGATE

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



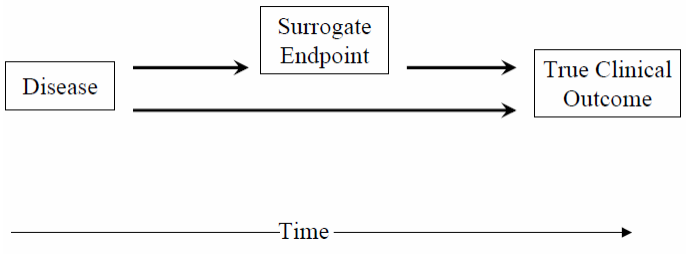
DANGEROUS SURROGATE

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



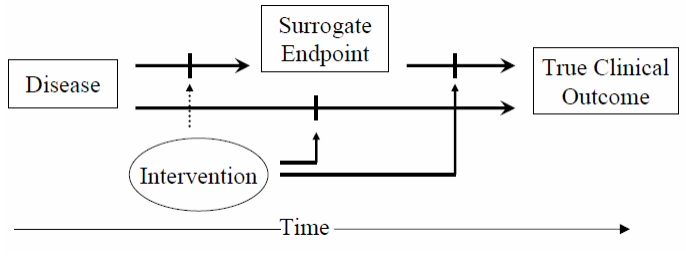
ALTERNATE PATHWAYS

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



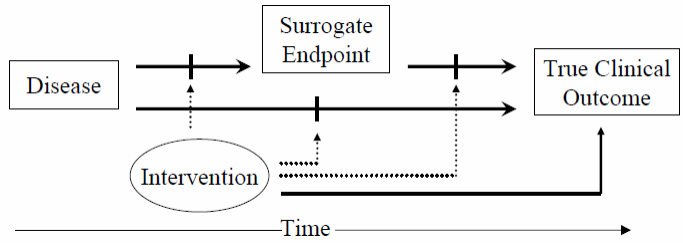
INEFFICIENT SURROGATE

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



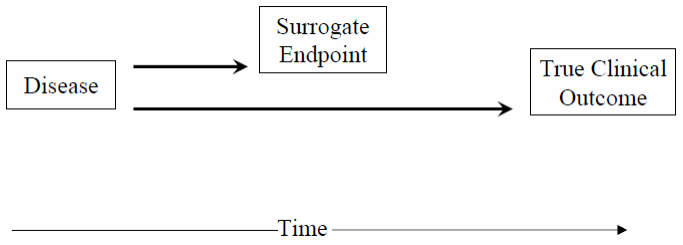
DANGEROUS SURROGATE

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



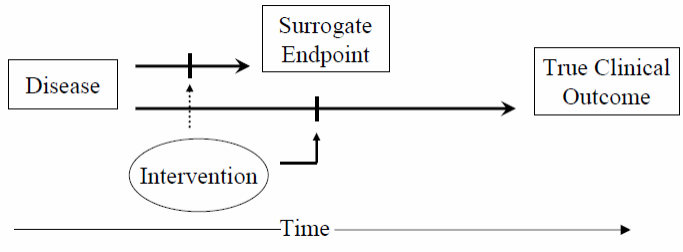
MARKER

- Disease causes Surrogate Endpoint and Clinical Outcome via different mechanisms



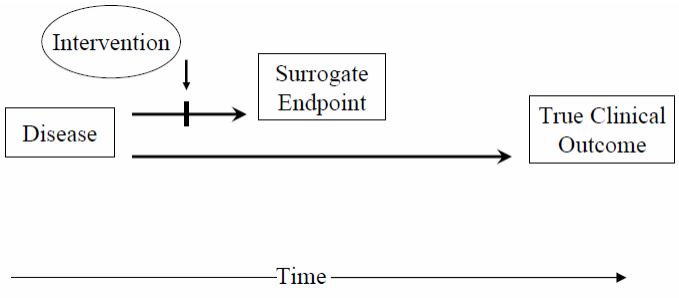
INEFFICIENT SURROGATE

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



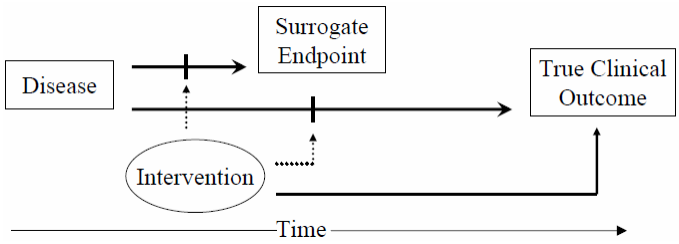
MISLEADING SURROGATE

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



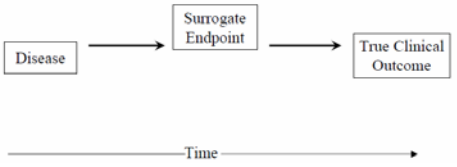
DANGEROUS SURROGATE

- 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
  - Time-dependent covariates

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

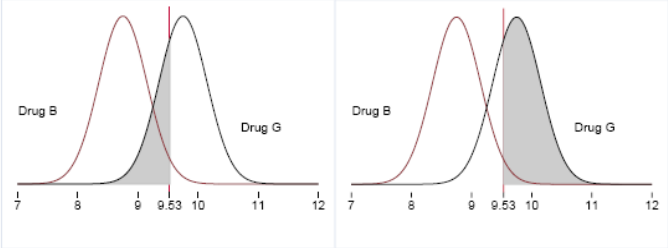
- Ethical issues
- Financial issues

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)

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

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

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

EXAMPLE

- Assume: HR = 0.75
- Alpha = 0.05
- Power = 80%
- $\beta = 0.2$
- ⇨ $379.5 = \frac{4(1.96 + 0.842)^2}{\left[\ln(0.75)\right]^2}$

- 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

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

- 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.

Percent Lost (per/ year)	Length of Recruit-ment Pe-riod	Hazard Ratio		
		0.75	0.5	0.25
		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

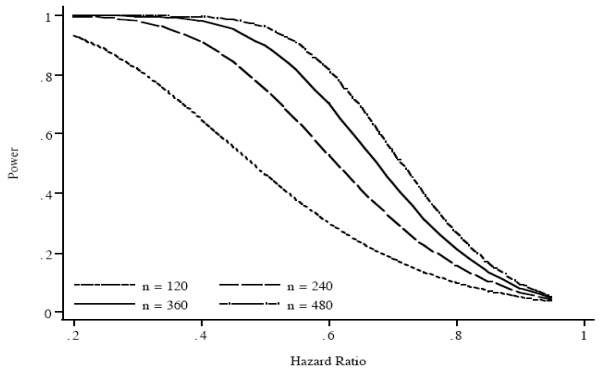
SAMPLE SIZE / POWER

- 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 $\alpha = 0.05$ ⇔ one-sided $\alpha = 0.025$

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

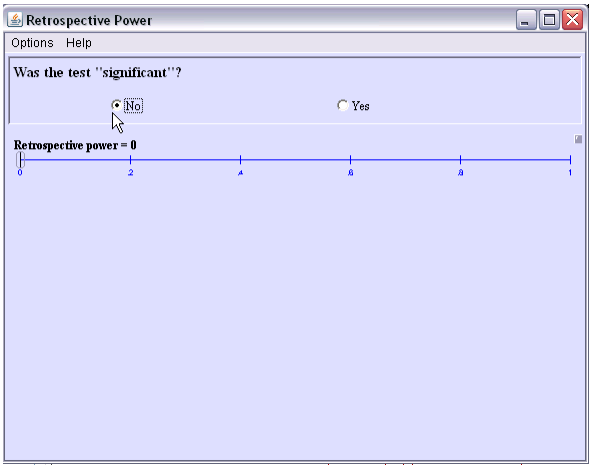
- **How to determine the "target" effect size?**
- Clinically meaningful
- Achievable

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 ….?

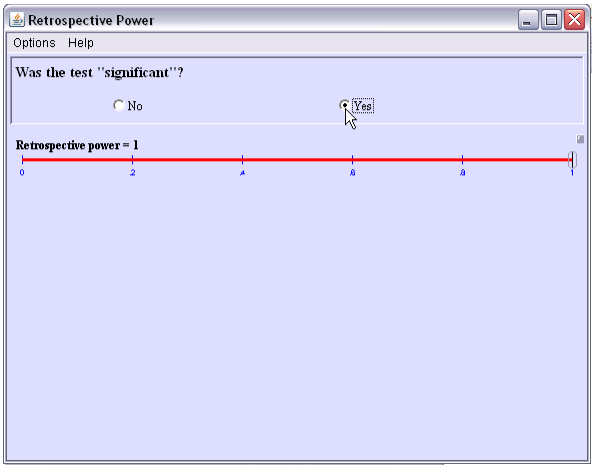
POST-HOC POWER

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



POST-HOC POWER

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



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
  - Time-dependent covariates

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

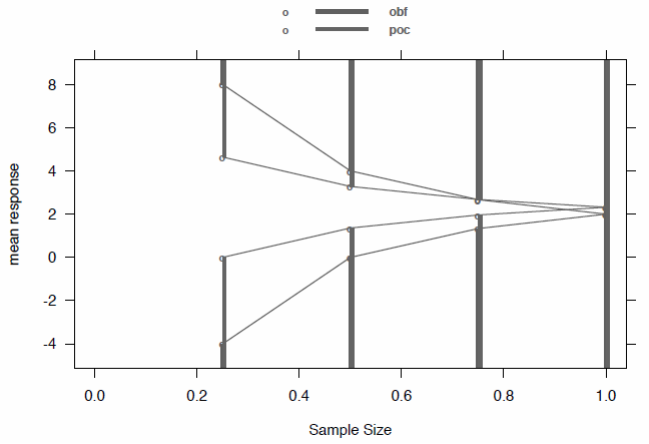
- How are trials monitored?
  - Investigator knowledge of interim results can lead to biased results:
    - Negative results may lead to loss of enthusiasm
    - Positive interim results may lead to inappropriate early publication
    - Either result may cause changes in the types of subjects who are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

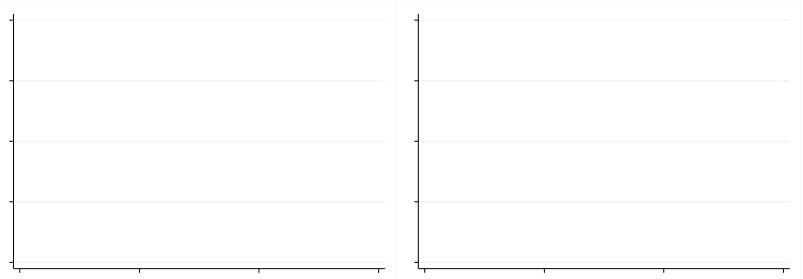
MONITORING BOUNDARIES

- Example of monitoring boundaries – note: scale



TYPICAL (NON-SURVIVAL) TRIAL

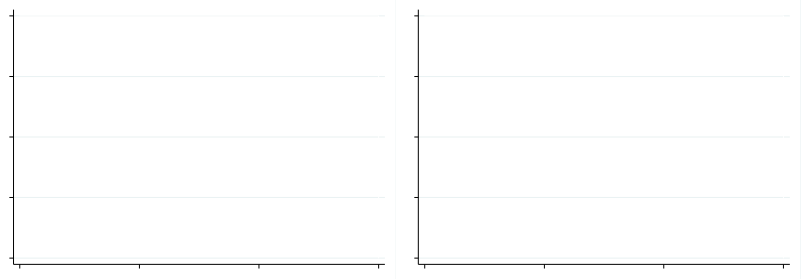
- Accrual pattern and information growth



Time Time

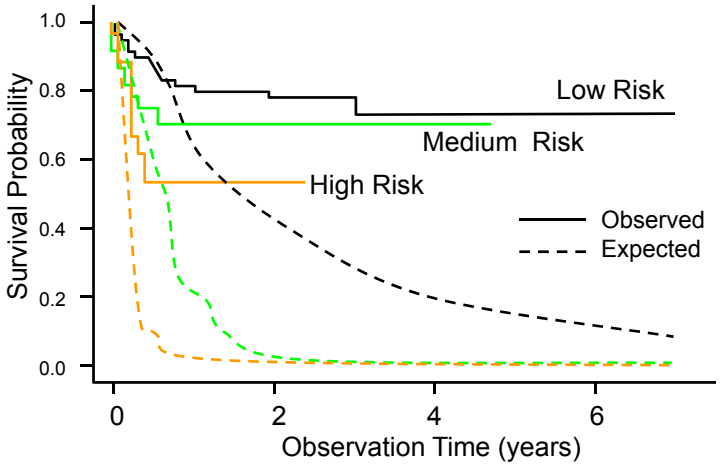
TRIAL WITH SURVIVAL ANALYSIS

- Accrual pattern and information growth



Time Time

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?

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
  - Time-dependent covariates

Time dependent covariates

TIME DEPENDENT COVARIATES

- The proportional hazards model
  - With fixed covariates

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{x})$$

$$\boldsymbol{\beta}'\mathbf{x} = \beta_1 x_1 + \ldots + \beta_k x_k$$

  - With time-dependent covariates

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}(t))$$

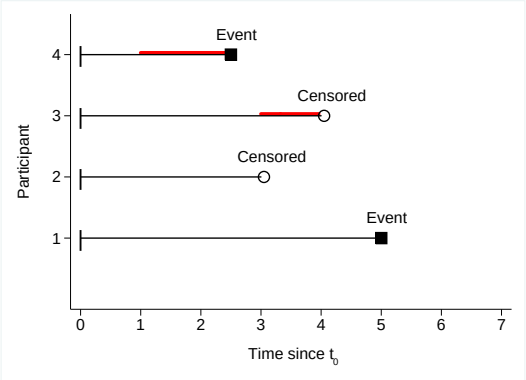
$$\boldsymbol{\beta}'\mathbf{x}(t) = \beta_1 x_1(t) + \ldots + \beta_k x_k(t)$$

TIME DEPENDENT COVARIATES

- Status/values of factor change over time
  - Transplant and survival (from acceptance into program) of patients with heart disease
  - Development of depression during Alzheimer's trial
- Conceptual issues and technical issues
  - Special software
  - Computationally more intensive
  - Data management
  - Missing data
  - Conceptual issues

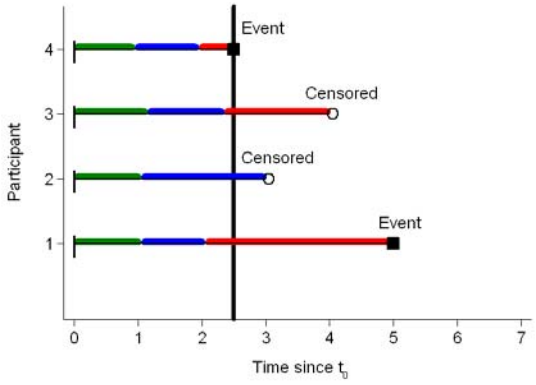
TIME DEPENDENT COVARIATES

- Example – Time varying indicator variable (here: switching on w/o switching off)



TIME DEPENDENT COVARIATES

- Evaluation at each event time



TIME DEPENDENT COVARIATES

- Evaluation of covariates at each event time
  - External
  - Internal (typically not available unless active follow-up / visits)
  - LOCF, imputation, interpolation
  - Computationally intensive
- Conceptual
  - Factor in causal pathway
  - Factors that change as result of “treatment”

TIME DEPENDENT COVARIATES – EXAMPLE

- Example: UMARU Impact Study (UIS).
- Outcome: time to return to drug use
- Treatment might have a time dependent effect. One might hypothesize that the treatment effect may simply be housing a subject where he/she has no access to drugs.
- We begin with a univariable model containing treatment.
- The estimated hazard ratio from a fit of this model for the longer versus the shorter duration of treatment is

  HR(long vs short treatment): 0.79 (95 % CIE 0.67, 0.94).

TIME DEPENDENT COVARIATES – EXAMPLE

- To examine the "under treatment" hypothesis, we create a time-varying dichotomous subject specific covariate

$$OFF\_TRT(t) = \begin{cases} 0 \text{ if } t \le LOT \\ 1 \text{ if } t > LOT \end{cases}$$

where LOT stands for the number of days the subject was on treatment.

- For example, suppose the survival time indexing the risk set is 30 days. Subjects in the risk set would have

$$OFF\_TRT(30) = 0$$

- if their value of LOT is greater than 30

TIME DEPENDENT COVARIATES – EXAMPLE

- The four estimated hazard ratios and their 95 percent confidence limits are shown in Table 7.3.

  - **Table 7.3 Estimated Hazard Ratios and 95 Percent Confidence Limit Estimates (CIE) for the Effect of Treatment and Being Off or On Treatment.**

Hazard Ratio for	Within Those	$\hat{HR}$	95% CIE
Long vs. Short	On Treatment	0.59	0.380, 0.922
Treatment Assignment	Off Treatment	1.10	0.910, 1.335
Off vs. On	Shorter Tx Duration	9.68	6.718, 13.955
Treatment	Longer Tx Duration	18.02	12.055, 26.927

TIME DEPENDENT COVARIATES – EXAMPLE

- The stated interpretations and conclusions comparing
  $OFF\_TRT(t) = 1$ versus $OFF\_TRT(t) = 0$
  require that the comparison is made for the same time *t*.

- If all patients were on treatment for exactly the same length of time and thus would go off treatment at exactly the same time, there would be no time point for which $OFF\_TRT(t) = 1$

  for some patients and for other patients $OFF\_TRT(t) = 0$

- In such a case, it would not make sense to estimate and interpret the hazard ratios presented in the last two rows of Table 7.3. In the UMARU Impact Study, the time points at which patients go off treatment vary greatly and the stated hazard ratios are valid for time points where some patients are on and others are off treatment.

Questions ?