Module 18:

Sequential and Adaptive Analysis with Time-to-Event Endpoints

Daniel L. Gillen, Ph.D.
Department of Statistics
University of California, Irvine

Scott S. Emerson, M.D., Ph.D.
Department of Biostatistics
University of Washington

Summer Institute in Statistics for Clinical Research
July 29, 2016

1

Where Am I Going?

Overview and Organization of the Course

2

Science and Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)
- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)
- In RCT, we are trying to prove the effect of some treatment
  - What do we need to consider as we strive to meet the burden of proof with adaptive modification of a RCT design?
- Does time to event data affect those issues?
  - Short answer: No, UNLESS subject to censoring
  - So, true answer: Yes.

3

Overview: Time-to-Event

- Many confirmatory phase 3 RCTs compare the distribution of time to some event (e.g., time to death or progression free survival).
- Common statistical analyses: Logrank test and/or PH regression
- Just as commonly: True distributions do not satisfy PH
- Providing users are aware of the nuances of those methods, such departures need not preclude the use of those methods

4

Overview: Sequential, Adaptive RCT

- Increasing interest in the use of sequential, adaptive RCT designs
- FDA Draft guidance on adaptive designs
  - "Well understood" methods
    - Fixed sample
    - Group sequential
    - Blinded adaptation
  - "Less well understood" methods
    - Adaptive sample size re-estimation
    - Adaptive enrichment
    - Response-adaptive randomization
    - Adaptive selection of doses and/or treatments

5

Overview: Premise

- Much of the concern with "less well understood" methods has to do with "less well understood" aspects of survival analysis in RCT
- Proportional hazards holds under strong null
  - But weak null can be important (e.g., noninferiority)
- Log linear hazard may be close to linear in log time over support of censoring distribution ➔ approximately Weibull
  - A special case of PH only when shape parameter is constant
- Hazard ratio estimate can be thought of a weighted time-average of ratio of hazard functions
  - *But* in Cox regression, weights depend on censoring distribution
  - *And* in sequential RCT, censoring distribution keeps changing

6

Course Organization

- Overview:
  - RCT setting
  - What do we know about survival analysis?
- Group sequential methods with time-to-event endpoints
  - Evaluation of RCT designs
  - Monitoring: implementation of stopping rules
- Adaptive methods for sample size re-estimation with PH
  - Case study: Low event rates, extreme effects
- Time to event analyses in presence of time-varying effects
- Special issues with adaptive RCT in time-to-event analyses

7

Overview

RCT setting

Where am I going?

It is important to keep in mind the overall goal of RCTs

I briefly describe some issues that impact our decisions in the design, monitoring, and analysis of RCTs

8

Overall Goal: "Drug Discovery"

- More generally
  - a therapy / preventive strategy or diagnostic / prognostic procedure
  - for some disease
  - in some population of patients

- A ***sequential***, ***adaptive*** series of experiments to establish
  - Safety of investigations / dose (phase 1)
  - Safety of therapy (phase 2)
  - Measures of efficacy (phase 2)
    - Treatment, population, and outcomes
  - Confirmation of efficacy (phase 3)
  - Confirmation of effectiveness (phase 3, post-marketing)

9

Science: Treatment "Indication"

- Disease
  - Therapy: Putative cause vs signs / symptoms
    - May involve method of diagnosis, response to therapies
  - Prevention / Diagnosis: Risk classification
- Population
  - Therapy: Restrict by risk of AEs or actual prior experience
  - Prevention / Diagnosis: Restrict by contraindications
- Treatment or treatment strategy
  - Formulation, administration, dose, frequency, duration, ancillary therapies
- Outcome
  - Clinical vs surrogate; timeframe; method of measurement

10

Evidence Based Medicine

- Decisions about treatments should consider PICO
  - Patient (population)
  - Intervention
  - Comparators
  - Outcome
- There is a need for estimates of safety, effect

11

Clinical Trials

- Experimentation in human volunteers
- Investigates a new treatment/preventive agent
  - Safety:
    - Are there adverse effects that clearly outweigh any potential benefit?
  - Efficacy:
    - Can the treatment alter the disease process in a beneficial way?
  - Effectiveness:
    - Would adoption of the treatment as a standard affect morbidity / mortality in the population?

12

Carrying Coals to Newcastle

- Wiley Act (1906)
  - Labeling
- Food, Drug, and Cosmetics Act of 1938
  - Safety
- Kefauver – Harris Amendment (1962)
  - Efficacy / effectiveness
    - " [If] there is a lack of substantial evidence that the drug will have the effect ... shall issue an order refusing to approve the application. "
    - "...The term 'substantial evidence' means evidence consisting of **adequate and well-controlled investigations, including clinical investigations**, by experts qualified by scientific training"
- FDA Amendments Act (2007)
  - Registration of RCTs, Pediatrics, Risk Evaluation and Mitigation Strategies (REMS)

13

Medical Devices

- Medical Devices Regulation Act of 1976
  - Class I: General controls for lowest risk
  - Class II: Special controls for medium risk - 510(k)
  - Class III: Pre marketing approval (PMA) for highest risk
    - **"…valid scientific evidence** for the purpose of determining the safety or effectiveness of a particular device … adequate to support a determination that there is reasonable assurance that the device is safe and effective for its conditions of use…"
    - "Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, **from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness…**"
- Safe Medical Devices Act of 1990
  - Tightened requirements for Class 3 devices

14

Clinical Trial Design

- Finding an approach that best addresses the often competing goals: Science, Ethics, Efficiency
  - Basic scientists: focus on mechanisms
  - Clinical scientists: focus on overall patient health
  - Ethical: focus on patients on trial, future patients
  - Economic: focus on profits and/or costs
  - Governmental: focus on safety of public: treatment safety, efficacy, marketing claims
  - Statistical: focus on questions answered precisely
  - Operational: focus on feasibility of mounting trial

15

Sequential RCT

- Ethical and efficiency concerns can be addressed through sequential sampling
- During the conduct of the study, data are analyzed at periodic intervals and reviewed by the DMC
- Using interim estimates of treatment effect decide whether to continue the trial
- If continuing, decide on any modifications to
  - scientific / statistical hypotheses and/or
  - sampling scheme

16

Design: Distinctions without Differences

- There is no such thing as a “Bayesian design”
- Every RCT design has a Bayesian interpretation
  - (And each person may have a different such interpretation)
- Every RCT design has a frequentist interpretation
  - (In poorly designed trials, this may not be known exactly)
- I focus on the use of both interpretations
  - Phase 2: Bayesian probability space
  - Phase 3: Frequentist probability space
  - Entire process: Both Bayesian and frequentist optimality criteria

17

Application to Drug Discovery

- We consider a population of candidate drugs
- We use RCT to “diagnose” truly beneficial drugs
- Use both frequentist and Bayesian optimality criteria
  - Sponsor:
    - High probability of adopting a beneficial drug (frequentist power)
  - Regulatory:
    - Low probability of adopting ineffective drug (freq type 1 error)
    - High probability that adopted drugs work (posterior probability)
  - Public Health (frequentist sample space, Bayes criteria)
    - Maximize the number of good drugs adopted
    - Minimize the number of ineffective drugs adopted

18

Frequentist vs Bayesian: Bayes Factor

- Frequentist and Bayesian inference truly complementary
  - Frequentist: Design so the same data not likely from null / alt
  - Bayesian: Explore updated beliefs based on a range of priors
- Bayes rule tells us that we can parameterize the positive predictive value by the type I error and prevalence
  - Maximize new information by maximizing Bayes factor
  - With simple hypotheses:

$$PPV = \frac{power \times prevalence}{power \times prevalence \quad + \quad type\ I\ err \times (1 - prevalence)}$$

$$\frac{PPV}{1 - PPV} = \frac{power}{type\ I\ err} \times \frac{prevalence}{1 - prevalence}$$

19

$$posterior\ \ odds = Bayes\ Factor \times prior\ \ odds$$

Adaptive Sampling: General Case

- At each interim analysis, possibly modify statistical or scientific aspects of the RCT
- Primarily statistical characteristics
  - Maximal statistical information (UNLESS: impact on MCID)
  - Schedule of analyses (UNLESS: time-varying effects)
  - Conditions for stopping (UNLESS: time-varying effects)
  - Randomization ratios (UNLESS: introduce confounding)
  - Statistical criteria for credible evidence
- Primarily scientific characteristics
  - Target patient population (inclusion, exclusion criteria)
  - Treatment (dose, administration, frequency, duration)
  - Clinical outcome and/or statistical summary measure

20

FDA Guidance on Adaptive RCT Designs

- Distinctions by role of trial
  - "Adequate and well-controlled" (Kefauver-Harris wording)
  - "Exploratory"
- Distinctions by adaptive methodology
  - "Well understood"
    - Fixed sample design
    - Blinded adaptation
    - Group sequential with pre-specified stopping rule
  - "Less well understood"
    - "Adaptive" designs with a prospectively defined opportunity to modify specific aspects of study designs based on review of unblinded interim data
  - "Not within scope of guidance"
    - Modifications to trial conduct based on unblinded interim data that are not prospectively defined

21

FDA Concerns

- Statistical errors: Type 1 error; power
- Bias of estimates of treatment effect
  - Definition of treatment effect
  - Bias from multiplicity
- Information available for subgroups, dose response, secondary endpoints
- Operational bias from release of interim results
  - Effect on treatment of ongoing patients
  - Effect on accrual to the study
  - Effect on ascertainment of outcomes

22

Group Sequential Designs

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
- Boundaries chosen to protect 2 of 3 operating characteristics
  - Type 1 error, power
  - Type 1 error, power, maximal sample size

23

Typical Adaptive Design

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
- At penultimate analysis (*J-1*), use unblinded interim test statistic to choose final sample size N_J

24

Adaptive Control of Type 1 Errors

- Proschan and Hunsberger (1995)
  - Adaptive modification of RCT design at a single interim analysis can more than double type 1 error unless carefully controlled
- Those authors describe adaptations to maintain experimentwise type I error and increase conditional power
  - Must prespecify a conditional error function

$$\int_{-\infty}^{\infty} A(z)\,\phi(z)\,dz = \alpha.$$

  - Often choose function from some specified test

$$A(z) = Pr_{\delta=0}(Z_2 \geq \Phi^{-1}(1-\alpha)\,|\,\tilde{Z}_1 = z, \tilde{n}_2 = n_2 - n_1),$$

  - Find critical value to maintain type I error

$$Pr_{\delta=0}\left(Z_2^* \geq c(\tilde{n}_2^*, \tilde{z}_1)\,|\,\tilde{n}_2^*(\tilde{z}_1)\right) = A(\tilde{z}_1).$$

25

Alternative Approaches

- Combining P values (Bauer & Kohne, 1994)
  - Based on R.A. Fisher's method
  - Extended to weighted combinations
- Cui, Hung, and Wang (1999)
  - Maintain conditional error from pre-specified design
- Self-designing Trial (Fisher, 1998)
  - Combine arbitrary test statistics from sequential groups using weighting of groups pespecified "just in time"

26

Overview

What do we know about time-to-event analyses?

Where am I going?

I present some examples where the behavior of standard analysis methods for time-to-event data are not well understood

27

Time to Event

- In time to event data, a common treatment effect across stages is reasonable under some assumptions
  - Strong null hypothesis (exact equality of distributions)
  - Strong parametric or semi-parametric assumptions
- The most common methods of analyzing time to event data will often lead to varying treatment effect parameters across stages
  - Proportional hazards regression with non proportional hazards data
  - Weak null hypotheses of equality of summary measures (e.g., medians, average hazard ratio)

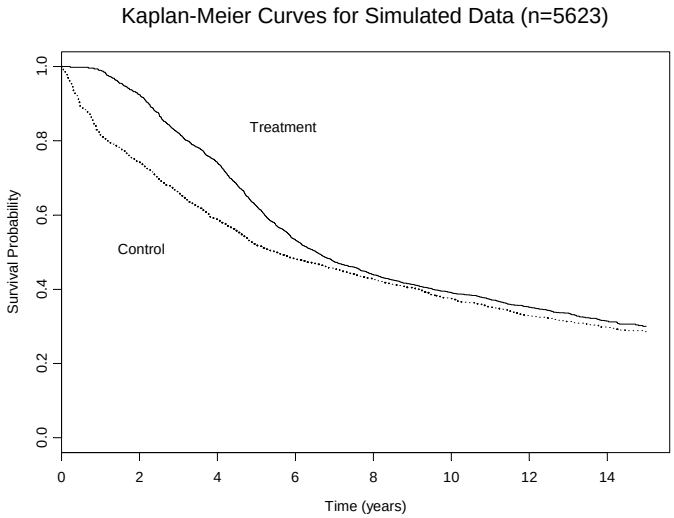
28

Hypothetical Example: Setting

- Consider survival with a particular treatment used in renal dialysis patients
- Extract data from registry of dialysis patients
- To ensure quality, only use data after 1995
  - Incident cases in 1995: Follow-up 1995 – 2002 (8 years)
  - Prevalent cases in 1995: Data from 1995 - 2002
    - Incident in 1994: Information about 2nd – 9th year
    - Incident in 1993: Information about 3rd – 10th year
    - …
    - Incident in 1988: Information about 8th – 15th year

29

Hypothetical Example: KM Curves



30

Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  A: 2.07 (logrank P = .0018)
  B: 1.13 (logrank P = .0018)
  C: 0.87 (logrank P = .0018)
  D: 0.48 (logrank P = .0018)

  - Lifelines:
    - 50-50? Ask the audience? Call a friend?

31

Who Wants To Be A Millionaire?

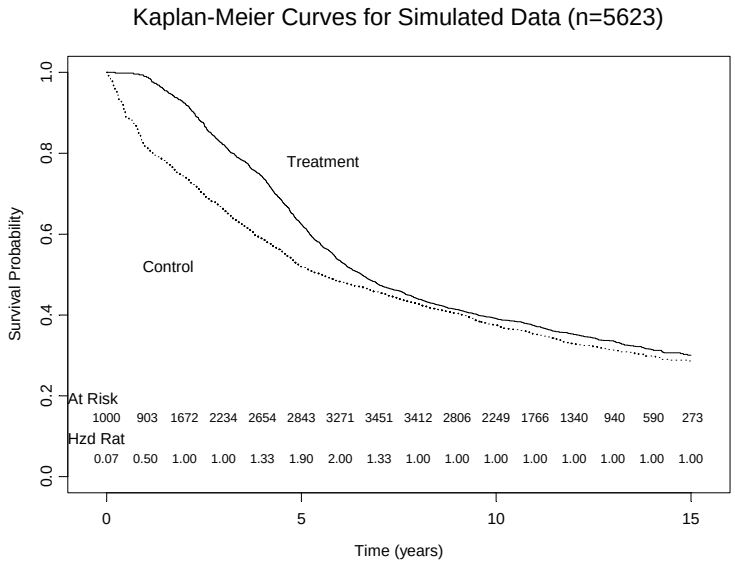
- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  B: 1.13 (logrank P = .0018)
  C: 0.87 (logrank P = .0018)

  - Lifelines:
    - 50-50? Ask the audience? Call a friend?

32

Hypothetical Example: KM Curves



33

Who Wants To Be A Millionaire?

Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

B: 1.13 (logrank P = .0018)

The weighting using the risk sets made no scientific sense

- Statistical precision to estimate a meaningless quantity is meaningless

34

Partial Likelihood Based Score

- Logrank statistic

$$U(\beta)=\frac{\partial}{\partial\beta}\log L(\beta)=\sum_{i=1}^{n}D_i\left[X_i-\frac{\sum_{j:T_j\geq T_i}X_j\exp\{X_j\beta\}}{\sum_{j:T_j\geq T_i}\exp\{X_j\beta\}}\right]$$

$$=\sum_t\left[d_{1t}-\frac{n_{1t}e^{\beta}}{n_{0t}+n_{1t}e^{\beta}}(d_{0t}+d_{1t})\right]$$

$$=\sum_t\frac{n_{0t}n_{1t}}{n_{0t}+n_{1t}}\left[\hat{\lambda}_{1t}-e^{\beta}\hat{\lambda}_{0t}\right]$$

35

Weighted Logrank Statistics

- Choose additional weights to detect anticipated effects

$$W(\beta)=\sum_t w(t)\frac{n_{0t}n_{1t}}{n_{0t}+n_{1t}}\left[\hat{\lambda}_{1t}-e^{\beta}\hat{\lambda}_{0t}\right]$$

$$n_{kt}=N_k\times\Pr(T\geq t,Cens\geq t)\overset{ind}{=}N_kS_k(t)\times\Pr(Cens\geq t)$$

$G^{\rho\gamma}$ Family of weighted logrank statistics:

$$w(t)=\left[\hat{S}_{\bullet}(t)\right]^{\rho}\left[1-\hat{S}_{\bullet}(t)\right]^{\gamma}$$

36

A Further Example

Biometrics 64, 733–740
September 2008

DOI: 10.1111/j.1541-0420.2007.00975.x

Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation

Brent R. Logan,* John P. Klein, and Mei-Jie Zhang

Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226, U.S.A.

**email:* blogan@mcw.edu

Summary. In some clinical studies comparing treatments in terms of their survival curves, researchers may anticipate that the survival curves will cross at some point, leading to interest in a long-term survival comparison. However, simple comparison of the survival curves at a fixed point may be inefficient, and use of a weighted log-rank test may be overly sensitive to early differences in survival. We formulate the problem as one of testing for differences in survival curves after a prespecified time point, and propose a variety of techniques for testing this hypothesis. We study these methods using simulation and illustrate them on a study comparing survival for autologous and allogeneic bone marrow transplants.

Key words: Censored data; Crossing hazard functions; Generalized linear models; Log-rank test; Pseudo-value approach; Weibull distribution; Weighted Kaplan–Meier statistic.

37

Logan, et al.: Motivation

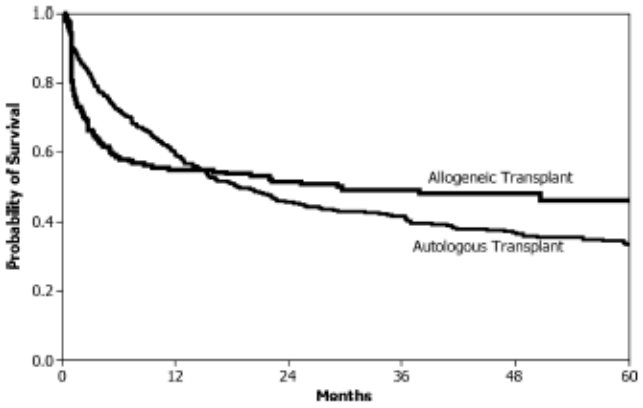


Figure 1. Kaplan–Meier estimate of DFS for follicular lymphoma example, by stem cell source.

38

Logan, et al.: Comparisons

- Logrank starting from time 0
- Weighted logrank test (rho=0, gamma=1) from time 0
- Survival at a single time point after time t_0
- Logrank starting from time t_0
- Weighted area between survival curves (restricted mean)
  - Most weight after time t_0
- Pseudovalues after time t_0
- Combination tests (linear and quadratic)
  - Compare survival at time t_0
  - Compare hazard ratio after time t_0

39

Logan, et al.: Simulations

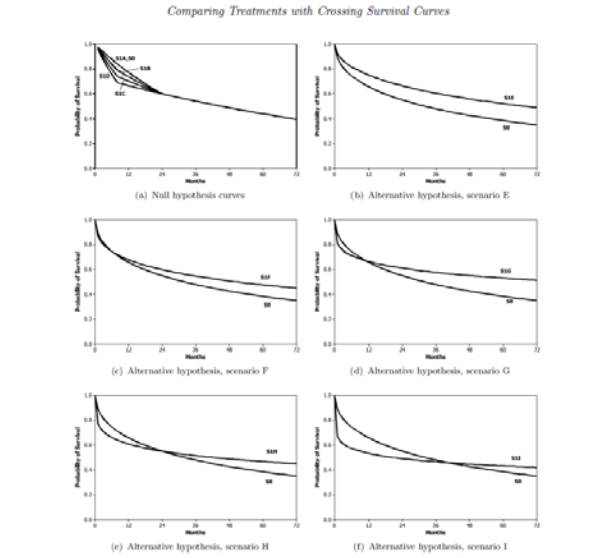


Figure 2. Survival curves for treatment (S1) and control (S0) groups used in simulations. Curves for the null hypoth simulations are shown in (a) for each of the four scenarios, and curves for the alternative hypothesis simulations are show (b)–(f) for the five scenarios.

40

Logan, et al.: Results

Table 2

Average rejection rates for 11 tests adjusted using ANOVA for censoring pattern. Rejection rates given by scenario using model (12). The last two rows refer to the log-rank (LR) test and weighted log-rank (WLR) tests starting at time 0. $t_0 = 24$.

		Scenario				
Method	Equation	E	F	G	H	I
$Z_{CLL}(24)$	(1)	62.4	15.3	21.1	4.7	21.8
$Z_{CLL}(48)$	(1)	70.1	32.9	65.1	21.5	6.8
$Z_{CLL}(72)$	(1)	71.2	44.5	85.1	46.1	25.9
$Z_{WKM}(t_0)$	(2)	75.8	35.0	66.3	20.3	6.0
$\chi^2_{PSV}(t_0)$	(3)	74.8	32.0	61.2	16.4	4.8
$Z_{LR}(t_0)$	(4)	30.7	36.5	85.4	71.7	82.6
$Z_{OLS}(t_0)$	(5)	74.7	43.9	84.1	43.4	23.6
$Z_{SP,P}(t_0)$	(6)	76.9	40.2	74.8	29.6	10.7
$\chi^2(t_0)$	(7)	67.2	36.7	83.1	61.1	81.0
Log rank		78.0	28.9	47.0	8.6	22.2
Weighted log rank $\rho = 0, \gamma = 1$		64.7	49.7	93.8	70.0	64.6

41

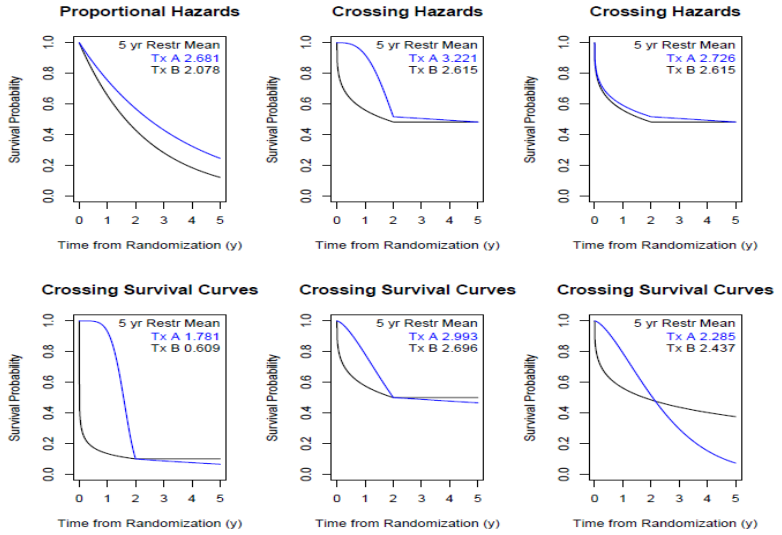
Logan, et al.: Critique

- In considering the combination tests, crossing survival curves might have
  - No difference at time t_0 (perhaps we are looking for equivalence)
  - Higher hazard after time t_0
- Presumably, the authors are interested in the curve that is higher at longer times post treatment
  - The authors did not describe how to use their test in a one-sided setting
- PROBLEM: The authors do not seem to be considering the difference between crossing survival curves and crossing hazard functions
  - Higher hazard over some period of time does not imply lower survival curves

42

Logan, et al.: Critique

- Additional scenarios that are of interest



43

Logan, et al.: Critique

- How might a naïve investigator use this test?
  - If the observed survival curves cross and the hazard is significantly higher after that point, the presumption might be that we have significant evidence that the group with higher hazard at later times has worse survival at those times
- "But it would be wrong" (Richard Nixon, March 21, 1973)
- We can create a scenario in which
  - Survival curves are truly stochastically ordered $S_A(t) > S_B(t) \; \forall t>0$
  - The probability of observing estimated curves that cross at t_0 is arbitrarily close to 50%
  - The probability of obtaining statistically significant higher hazards for group A after t_0 is arbitrarily close to 100%
  - Thus, the one-sided type 1 error is arbitrarily close to 50%

44

Relevance to Today

- Even experts in survival analysis sometimes lose track of the way that time to event analyses behave, relative to our true goals

Lifetime Data Anal (2015) 21:218–240
DOI 10.1007/s10985-014-9298-4

Group sequential tests for long-term survival comparisons

Brent R. Logan · Shuyuan Mo

Received: 6 February 2014 / Accepted: 2 July 2014 / Published online: 23 July 2014

Abstract Sometimes in clinical trials, the hazard rates are anticipated to be nonproportional, resulting in potentially crossing survival curves. In these cases, researchers are usually interested in which treatment has better long-term survival. The log-rank test and the weighted log-rank test may not be appropriate or efficient to use here, because they are sensitive to differences in survival at any time and don't just focus on long-term outcomes. Also in a prospective clinical trial, patients are entered sequentially over calendar time, so that group sequential designs may be considered for ethical, administrative and economic concerns. Here we develop group sequential methods for testing the null hypothesis that the survival curves are identical after a prespecified time point. Several classes of tests are considered, including an integrated difference in survival probabilities after this time point, and linear or quadratic combinations of two component test statistics (pointwise comparisons of survival at the time point and comparisons of hazard rates after the time point). We examine the type I errors, stopping probabilities, and powers of these tests through simulation studies under the null and different alternatives, and we apply them to a real bone marrow transplant clinical trial.

Keywords Crossing hazards · Crossing survival curves · Late survival difference · Group sequential test · Error-spending methods

45

Final Comments

- There is still much for us to understand about the implementation of adaptive designs
- Most often the "less well understood" part is how they interact with particular data analysis methods
  - In particular, the analysis of censored time to event data has many scientific and statistical issues
- How much detail about accrual patterns, etc. do we want to have to examine for each RCT?
- How much do we truly gain from the adaptive designs?
  - (Wouldn't it be nice if statistical researchers started evaluating their new methods in a manner similar to evaluation of new drugs?)

46

Bottom Line

- There is no substitute for planning a study in advance
  - At Phase 2, adaptive designs may be useful to better control parameters leading to Phase 3
    - Most importantly, learn to take "NO" for an answer
  - At Phase 3, there seems little to be gained from adaptive trials
    - We need to be able to do inference, and poorly designed adaptive trials can lead to some very perplexing estimation methods
- **"Opportunity is missed by most people because it is dressed in overalls and looks like work."** -- Thomas Edison
- In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

47

Really Bottom Line

"You better think (think)
about what you're
trying to do…"

-Aretha Franklin, "Think"

48

Sequential and Adaptive Analysis with Time-to-Event Endpoints

Session 2 - Group Sequential Designs for Time-to-Event Endpoints

Presented July 29, 2016

Scott S. Emerson
Department of Biostatistics
University of Washington

Daniel L. Gillen
Department of Statistics
University of California, Irvine

Overview of group sequential designs

Statistical framework for trial monitoring: Statistical design of the fixed-sample trial

- The statistical decision criteria are referenced to the trial's design hypotheses. For example:
  - One-sided superiority test (assume small θ favors new treatment):

$$\begin{aligned} \text{Null:} \quad & \theta \geq \theta_\emptyset \\ \text{Alternative:} \quad & \theta \leq \theta_+ \end{aligned}$$

    with $\theta_+ < \theta_\emptyset$, and θ_+ is chosen to represent the smallest difference that is clinically important.

  - Two-sided (equivalence) test:

$$\begin{aligned} \text{Null:} \quad & \theta = \theta_\emptyset \\ \text{Lower Alternative:} \quad & \theta \leq \theta_- \\ \text{Upper Alternative:} \quad & \theta \geq \theta_+ \end{aligned}$$

    with $\theta_- < \theta_\emptyset < \theta_+$. θ_- and θ_+ denote the smallest important differences.

Overview of group sequential designs

Statistical framework for trial monitoring: Selecting decision criteria

- A decision to stop needs to consider what has or has not been ruled out. For example
  - One-sided superiority test (assume small θ favors new treatment):
    - Stop for superiority when any harm ($\theta \geq \theta_\emptyset$) has been ruled out.
    - Stop for futility when important benefits ($\theta \leq \theta_+$) have been ruled out.
  - Two-sided (equivalence) test:
    - Stop for treatment *A* better than treatment *B* when inferiority of *A* ($\theta \leq \theta_\emptyset$) has been ruled out.
    - Stop for treatment *B* better than treatment *A* when inferiority of *B* ($\theta \geq \theta_\emptyset$) has been ruled out.
    - Stop for equivalence when important differences (either $\theta \geq \theta_+$ or $\theta \leq \theta_-$) have been ruled out.
- The hypotheses that have been ruled in/out are given by the interval estimate.

SISCR
UW - 2016

Overview of group sequential designs

Statistical framework for trial monitoring: Group sequential designs (superiority trial)

- Suppose that the trial is planned for $j = 1, ..., J$ interim analyses.
- Let $\hat{\theta}_j$ denote the estimated treatment effect at the *j*th analysis.
- Consider stopping criteria $a_j < d_j$ with:

$$\begin{aligned} \hat{\theta}_j \leq a_j &\Rightarrow \text{Decide new treatment is superior} \\ \hat{\theta}_j \geq d_j &\Rightarrow \text{Decide new treatment is not superior} \\ a_j < \hat{\theta}_j < d_j &\Rightarrow \text{Continue trial} \end{aligned}$$

  Set $a_J = d_J$ so that the trial stops by the *J*th analysis.
- How should we choose these critical values?

SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Statistical framework for trial monitoring

Inadequacy of Fixed Sample Methods

- Suppose we simply ignore the fact that we are repeatedly testing our hypothesis
- We can quickly see the impact of this via simulation
  - Let $X_i \sim_{\text{iid}} \mathcal{N}(\theta, \sigma^2)$
  - $j = 1, ..., 4$ equally spaced analyses at 25, 50, 75, and 100 observations
  - Test statistic after n_j observations have been accrued

$$\bar{X}_{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i$$

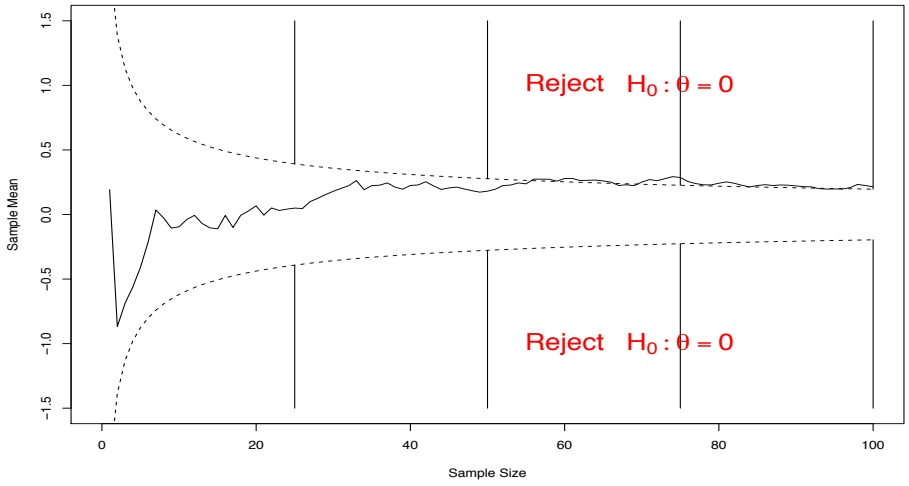
  - Test $H_0 : \theta = 0$ with level $\alpha = .05$
- Fixed sample methods (2-sided test): Reject H_0 first time

$$|\bar{X}_{n_j}| > z_{1-\alpha/2} \frac{\sigma}{\sqrt{n_j}}, \qquad j = 1, 2, 3, 4$$

Statistical framework for trial monitoring

Inadequacy of Fixed Sample Methods : Simulation

- Consider the sample path of the statistic for a single simulated trial



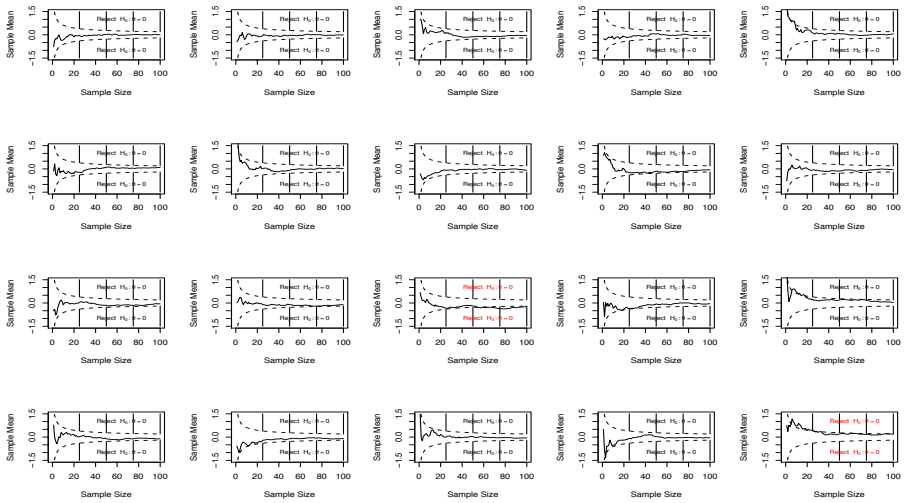
SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Statistical framework for trial monitoring

Inadequacy of Fixed Sample Methods : Simulation

- Consider the sample path of the statistic for 20 randomly sampled trials



SISCR
UW - 2016

Statistical framework for trial monitoring

Inadequacy of Fixed Sample Methods : Simulation

- Simulated type I error rate using fixed sample methods
- Based on 100,000 simulations

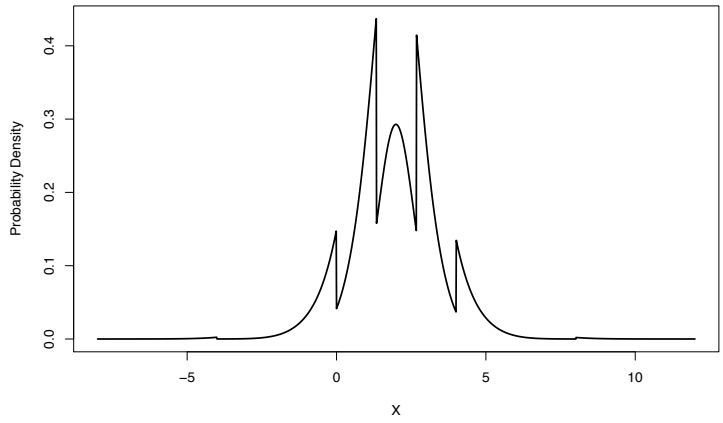
Significant at	Proportion Significant	Number Significant	Proportion Significant
Analysis 1	0.05075	Exactly 1	0.07753
Analysis 2	0.04978	Exactly 2	0.02975
Analysis 3	0.05029	Exactly 3	0.01439
Analysis 4	0.05154	All 4	0.00554
		Any	0.12721

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Interim analyses require special methods

Sampling density for sequentially-monitored test statistic

- The filtering due to interim analyses creates non-standard sampling densities as the basis for inference.
- Sampling density depends on the stopping rule.
- In order to correct the type 1 error rate, we must be able to compute the density of the statistic that accounts for the possibility of stopping at interim analyses



Sampling density for sequentially sampled test statistic

- Let C_j denote the continuation set at the jth interim analysis.
- Let (M, S) denote the bivariate statistic where M denotes the stopping time ($1 \leq M \leq J$) and $S = S_M$ denotes the value of the partial sum statistic at the stopping time.
- The sampling density for the observation $(M = m, S = s)$ is:

$$p(m, s; \theta) = \begin{cases} f(m, s; \theta) & s \notin C_m \\ 0 & else \end{cases}$$

where the (sub)density function $f(j, s; \theta)$ is recursively defined as

$$\begin{aligned} f(1, s; \theta) &= \frac{1}{\sqrt{n_1 V}} \phi\left(\frac{s - n_1\theta}{\sqrt{n_1 V}}\right) \\ f(j, s; \theta) &= \int_{C_{(j-1)}} \frac{1}{\sqrt{n_j V}} \phi\left(\frac{s - u - n_j\theta}{\sqrt{n_j V}}\right) f(j-1, u; \theta)\, du, \\ & \quad j = 2, \ldots, m \end{aligned}$$

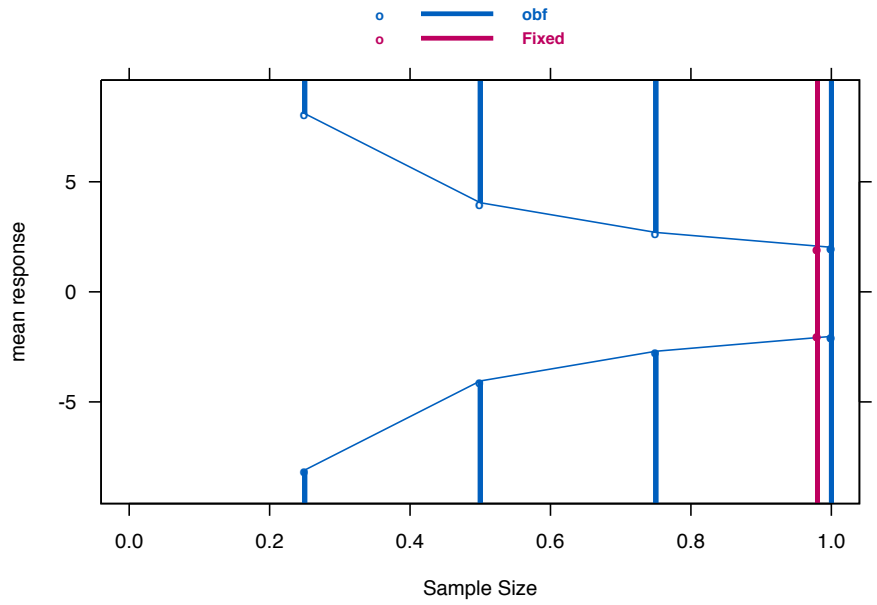
SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Types of group sequential designs

Example: O'Brien-Fleming (OBF) 2-sided design

- Using the correct sampling density, we can choose boundary values that maintain experiment wise Type I error



SISCR
UW - 2016

Types of group sequential designs

Example: O'Brien-Fleming (OBF) 2-sided design

- Simulated type I error rate using fixed sample methods
- Based on 100,000 simulations

Significant at	Proportion Significant	Number Significant	Proportion Significant
Analysis 1	0.00006	Exactly 1	0.03610
Analysis 2	0.00409	Exactly 2	0.01198
Analysis 3	0.01910	Exactly 3	0.00210
Analysis 4	0.04315	All 4	0.00001
		Any	0.05019

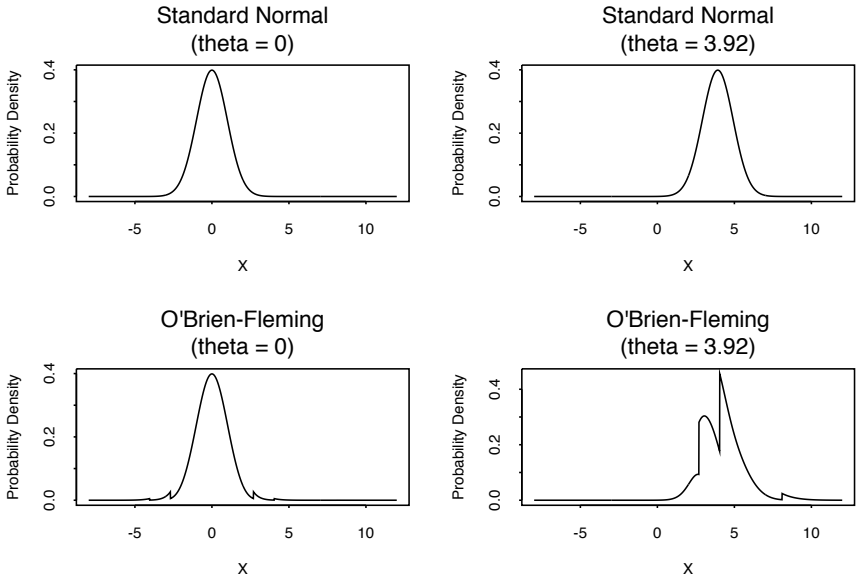
SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Types of group sequential designs

Example: O'Brien-Fleming (OBF) 2-sided design

- Sampling density for OBF boundaries with $\theta = 0$ and $\theta = 3.92$ (corresponding Normal sampling density for comparison):



Types of group sequential designs

Boundary shape functions

- There are an infinite number of stopping boundaries to choose from that will maintain a given family-wise error
  - They will differ in required sample size and power
- Kittelson and Emerson (1999) described a "unified family" of designs that are parameterized by three parameters (A, R, and P)
- Parameterization of boundary shape function includes many previously described approaches
  - Wang & Tsiatis boundary shape functions:
    - $A = 0, R = 0$, and $P > 0$
    - $P = 0.5$: Pocock (1977)
    - $P = 1.0$: O'Brien-Fleming (1979)
  - Triangular Test boundary shape functions (Whitehead):
    - $A = 1, R = 0$, and $P = 1$
  - Sequential Conditional Probability Ratio Test (Xiong):
    - $R = 0.5$, and $P = 0.5$

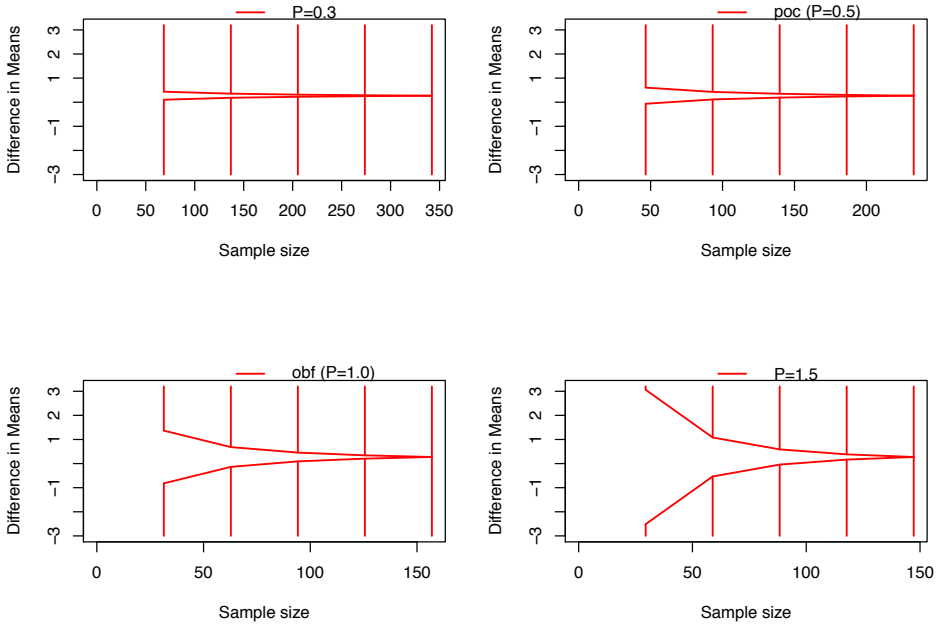
SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

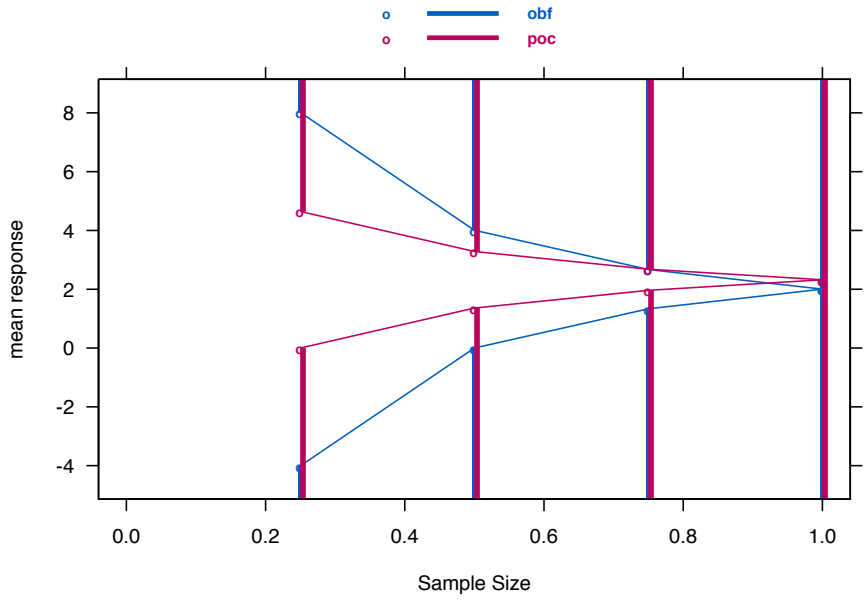
Types of group sequential designs

Boundary shape functions

- Consider differing choices of P



Example: OBF (P=1) versus Pocock (P=0.5) 1-sided designs

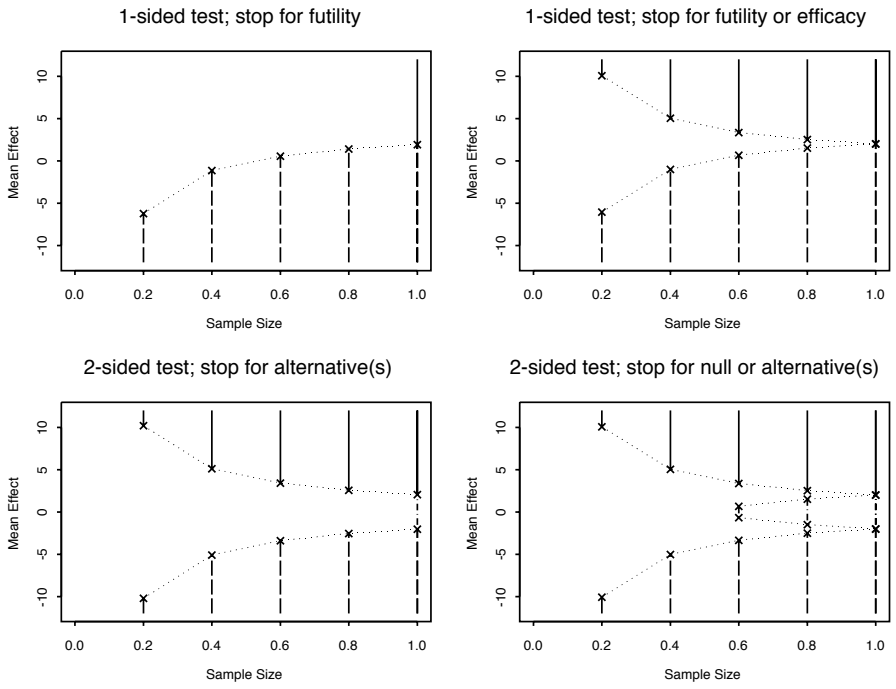


Types of group sequential designs

Group sequential designs can be formulated for various hypotheses

- Four design categories:
  - One-sided test; One-sided stopping
    (allow stopping for efficacy *or* futility, but not both)
  - One-sided test; Two-sided stopping
    (allow stopping for either efficacy or futility)
  - Two-sided test; One-sided stopping
    (allow stopping only for the alternative(s))
  - Two-sided test; Two-sided stopping
    (allow stopping for either the null or the alternative)

Four general design categories



Types of group sequential designs

So how should we choose a stoping rule?

- Consider appropriate type of hypothesis to test
- Maintain statistical design criteria of the fixed sample trial:
  - Type I error rate of $\alpha = 0.025$ (one-sided test) or $\alpha = 0.05$ (two-sided test).
  - Maintain maximal sample size (with potential loss of power)
  - Maintain power (with larger maximal sample size)
- Other considerations when selecting critical values:
  - Number of interim analyses
  - Timing of interim analyses
  - Degree of early conservatism
  - Characteristics of the sample size distribution:
    - Expected sample size (Average Sample Number; ASN)
    - Quantiles of the sample size distribution
    - Maximal sample size
    - Stopping probabilities at each of the interim analyses

Interim analyses require special methods

Characteristics of the group sequential sampling density

- Density is not shift invariant
- Jump discontinuities
- Requires numerical integration
- Sequential testing introduces bias:

	$E(\hat{\theta})$	
θ	OBF	Pocock
0.00	-0.29	-0.48
1.96	1.95	1.82
3.92	4.21	4.38

Case Study : Hodgkin's Trial

Background

- Hodgkin's lymphoma represents a class of neoplasms that start in lymphatic tissue
- Approximately 7,350 new cases of Hodgkin's are diagnosed in the US each year (nearly equally split between males and females)
- 5-year survival rate among stage IV (most severe) cases is approximately 60-70%

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Background (cont.)

- Common treatments include the use of chemotherapy, radiation therapy, immunotherapy, and possible bone marrow transplantation
- Treatment typically characterized by high rate of initial response followed by relapse
- Hypothesize that experimental monoclonal antibody in addition to standard of care will increase time to relapse among patients remission

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Definition of Treatment

- Administered via IV once a week for 4 weeks
- Patients randomized to receive standard of care plus active treatment or placebo (administered similarly)
- Treatment discontinued in the event of grade 3 or 4 AEs
- Primary analysis based upon intention-to-treat

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Defining the target population

- Histologically confirmed Hodgkin's lymphoma Grade 1-3
- Progressive disease requiring treatment after at least 1 prior chemotherapy
- Recovered fully from any significant toxicity associated with prior surgery, radiation treatments, chemotherapy, biological therapy, autologous bone marrow or stem cell transplant, or investigational drugs

Case Study : Hodgkin's Trial

Defining the Comparison Group

- Scientific credibility for regulatory approval
- Concurrent comparison group
  - inclusion / exclusion criteria may alter baseline rates from historical experience
  - crossover designs impossible
- Final Decision
  - Single comparison group treated with placebo
    - not interested in studying dose response
    - no similar current therapy
    - avoid bias with assessment of softer endpoints
  - Randomize
    - allow causal inference

Case Study : Hodgkin's Trial

Defining the Outcomes of Interest

- **Goals**:
  - Primary: Increase relapse-free survival
    - Long term (always best)
    - Short term (many other processes may intervene)
  - Secondary: Decrease morbidity
- **Refinement of the primary endpoint**
  - Definition of event
    - First occurrence of death or relapse (relapse defined as presence of measurable lesion at 3-month scheduled visits)
  - Possible primary endpoints
    - Event rate at fixed point in time
    - Quantile of time to event distribution
    - Hazard of event

Case Study : Hodgkin's Trial

Refinement of the primary endpoint

Final Choice: Comparison of hazards for event (censored continuous data)

- Duration of followup
  - Wish to compare relapse-free survival over 4 years
  - Patients accrued over 3 years in order to guarantee at least one year of followup for all patients
- Measures of treatment effect (comparison across groups)
  - Hazard ratio (Cox estimate; implicitly weighted over time)
  - No adjustment for covariates
  - Statistical information dictated by number of events (under proportional hazards, statistical information is approximately D/4)

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Definition of statistical hypotheses

Null hypothesis

- Hazard ratio of 1 (no difference in hazards)
- Estimated baseline survival
  - Median progression-free survival approximately 9 months
  - (needed in this case to estimate variability)

Alternative hypothesis

- One-sided test for decreased hazard
  - Unethical to prove increased mortality relative to comparison group in placebo controlled study (always??)
- 33% decrease in hazard considered clinically meaningful
  - Corresponds to a difference in median survival of 4.4 months assuming exponential survival

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Criteria for statistical evidence

- Type I error: Probability of falsely rejecting the null hypothesis Standards:
  - Two-sided hypothesis tests: 0.050
  - One-sided hypothesis test: 0.025
- Power: Probability of correctly rejecting the null hypothesis (1-type II error) Popular choice:
  - 80% power

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Determination of sample size

- Sample size chosen to provide desired operating characteristics
  - Type I error : 0.025 when no difference in mortality
  - Power : 0.80 when 33% reduction in hazard
- Expected number of events determined by assuming
  - Exponential survival in placebo group with median survival of 9 months
  - Uniform accrual of patients over 3 years
  - Negligible dropout

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Determination of sample size

- General sample size formula:
  - δ = standardized alternative
  - Δ = log-hazard ratio
  - π_i = proporiton of patients in group i, $i = 0, 1$
  - D = number of sampling units (events)

$$D = \frac{\delta^2}{\pi_0 \pi_1 \Delta^2}$$

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Determination of sample size

- Fixed sample test (no interim analyses):
  - $\delta = (z_{1-\alpha} + z_\beta)$ for size α and power β
- For current study, we assume 1:1 randomization
  - $\pi_0 = \pi_1 = 0.5$
- Number of events for planned trial:

$$D = \frac{(1.96 + 0.84)^2}{0.5^2 \times [\log(.67)]^2} = 195.75$$

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Specification of fixed sample design using RCTdesign

- Again, we can use the function `seqDesign()` for specifying the fixed sample design (`prob.model="hazard"`)

```
> survFixed <- seqDesign( prob.model = "hazard", arms = 2,
                          null.hypothesis = 1, alt.hypothesis = 0.67,
                          ratio = c(1, 1), nbr.analyses = 1,
                          test.type = "less",
                          power = 0.80, alpha = 0.025 )


> survFixed
Call:
seqDesign(prob.model = "hazard", arms = 2, null.hypothesis = 1,
    alt.hypothesis = 0.67, ratio = c(1, 1), nbr.analyses = 1,
    test.type = "less", power = 0.8, alpha = 0.025)

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >= 1.00    (size  = 0.025)
    Alternative hypothesis : Theta <= 0.67    (power = 0.800)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                              a      d
    Time 1 (N= 195.75) 0.7557 0.7557
```


Case Study : Hodgkin's Trial

Determination of sample size (cont.)

- In general, it necessary to know the expected number of patients required to obtain the desired operating characteristics
- This is given by:

$$N = \frac{D}{\pi_0 \Pr_0[\text{Event}] + \pi_1 \Pr_1[\text{Event}]}$$

where D is the total number of required events and π_i is the proportion of patients allocated to group i

Case Study : Hodgkin's Trial

Determination of sample size (cont.)

- Under proportional hazards, Pr[Event] for each group depends upon
  1. The total followup (T_L) and accrual (T_A) time
  2. The underlying survival distribution
  3. The accrual distribution
  4. Drop-out

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Determination of sample size (cont.)

- From the above, if we assume a uniform accrual pattern we have:

$$\begin{aligned}
\Pr[\text{Event}] &= \int_0^{T_A} \Pr[\text{Event \& Entry at } t]dt \\
&= \int_0^{T_A} \Pr[\text{Event} \mid \text{Entry at } t]\Pr[\text{Entry at } t]dt \\
&= 1 - \int_0^{T_A} \Pr[\text{No Event} \mid \text{Entry at } t]\Pr[\text{Entry at } t]dt \\
&= 1 - \frac{1}{T_A}\int_0^{T_A} \Pr[\text{No Event} \mid \text{Entry at } t]dt \qquad \text{(unif acc)} \\
&= 1 - \frac{1}{T_A}\int_0^{T_A} S(T_L - t)dt
\end{aligned}$$

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Specification of fixed sample design using RCTdesign

- In RCTdesign this is automated assuming exponential survival using the function `seqPHSubjects()`
- For the Hodgkin's trial we assumed
  - Median survival in the control arm of 9 months
  - Uniform accrual over 3 years with one additional year of followup

```
> seqPHSubjects( survFixed, controlMedian=0.75,
                 accrualTime=3, followupTime=1 )

  accrualTime followupTime   rate hazardRatio controlMedian nSubjects
1           3            1 75.364        1.00          0.75    226.09
2           3            1 80.497        0.67          0.75    241.49
```


Case Study : Hodgkin's Trial

SISCR
UW - 2016

Determination of sample size (cont.)

- Interpretation:
  - In order to desire the required number of patients we would need to accrue:
    - N=76 patients per year for 3 years if the null hypothesis were true (Total of 228 patients)
    - N=81 patients per year for 3 years if the alternative hypothesis were true (Total of 243 patients)

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Evaluating the operating characteristics

1. Critical values
   - Observed value which rejects the null
   - Point estimate of treatment effect (clinical and marketing relevance?
2. Confidence interval at the critical value
   - Set of hypothesized treatment effects which might reasonably generate data like that observed
     - Have we excluded all scientifically meaningful alternatives with a negative study?
3. Statistical power across various alternatives
4. Bayesian posterior probabilities at the critical value (more later)
5. Sensitivity to design assumptions (sample size and/or baseline survival)

Case Study : Hodgkin's Trial

Frequentist inference at the boundaries using RCTdesign

- In RCTdesign frequentist inference can be obtained with the `seqInference()` function
- Only required argument is the design to be used

```
> seqInference( survFixed )
       Ordering                *** a Boundary ***  *** d Boundary ***
Time 1          Boundary              0.756               0.756
                MLE                   0.756               0.756
                BAM                   0.756               0.756
                RBadj                 0.756               0.756
           Mean MUE                   0.756               0.756
           Mean P-value               0.025               0.025
           Mean 95% Conf Int          (0.571, 1)          (0.571, 1)
           Time MUE                   0.756               0.756
           Time P-value               0.025               0.025
           Time 95% Conf Int          (0.571, 1)          (0.571, 1)
```

Case Study : Hodgkin's Trial

Statistical power using RCTdesign

- Power can be computed using `seqOC()` or plotted using `seqPlotPower()`

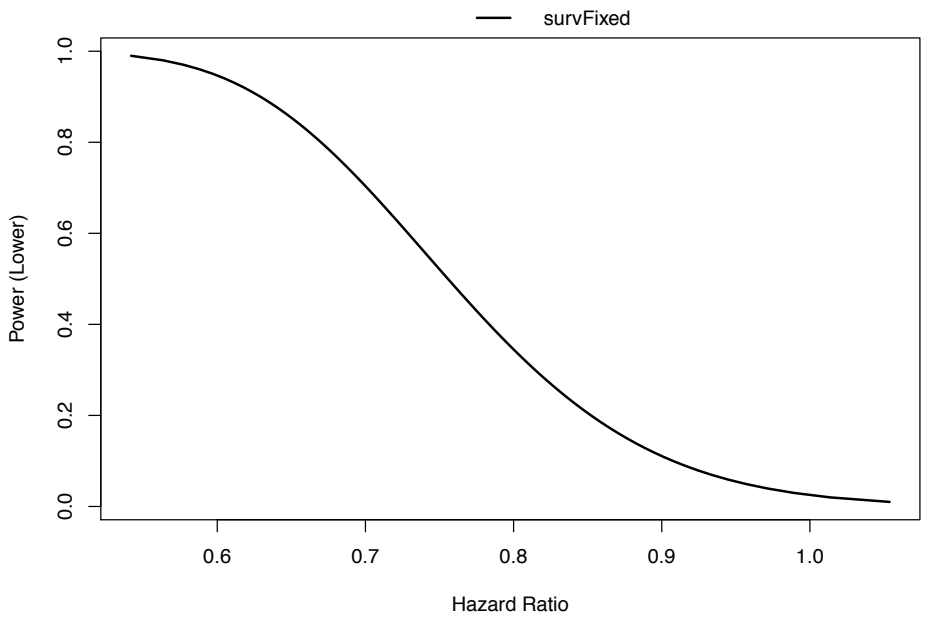
```
> seqOC(survFixed, theta=seq(.4,1,by=.05) )
Operating characteristics
 Theta     ASN Power.lower
  0.40 195.75      1.0000
  0.45 195.75      0.9999
  0.50 195.75      0.9981
  0.55 195.75      0.9869
  0.60 195.75      0.9467
  0.65 195.75      0.8540
  0.70 195.75      0.7037
  0.75 195.75      0.5210
  0.80 195.75      0.3450
  0.85 195.75      0.2052
  0.90 195.75      0.1107
  0.95 195.75      0.0547
  1.00 195.75      0.0250

Fixed design (one analysis time)


> seqPlotPower( survFixed, dsnLbls=c("survFixed") )
```

SISCR
UW - 2016

Case Study : Hodgkin's Trial

Statistical power using RCTdesign

- Power can be computed using `seqOC()` or plotted using `seqPlotPower()`



SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Re-designing the study

- Sponsor felt that attaining 75-80 patients per year would be unrealistic
- Wished to consider design operating characteristics assuming approximately uniform accrual of 50 patients per year while maintaining the same accrual time and follow up
- Problem: Need to determine the expected number of events if 50 subjects were accrued per year
- Solution: Solve backwards using the `nEvents` argument in `seqPHSubjects()`, substituting various numbers of events

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Re-designing the study

- After a (manual) iterative search, we find that if roughly 50 patients are accrued yearly (under the alternative), 121 events would be expected

```
> seqPHSubjects( survFixed, controlMedian = 0.75, accrualTime = 3,
                 followupTime = 1, nEvents = 121 )

  accrualTime followupTime   rate hazardRatio controlMedian nSubjects
1           3            1 46.584        1.00          0.75    139.75
2           3            1 49.757        0.67          0.75    149.27
```

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Re-designing the study

- Use the `update()` function in RCTdesign to update to the new sample size and compare operating characteristics

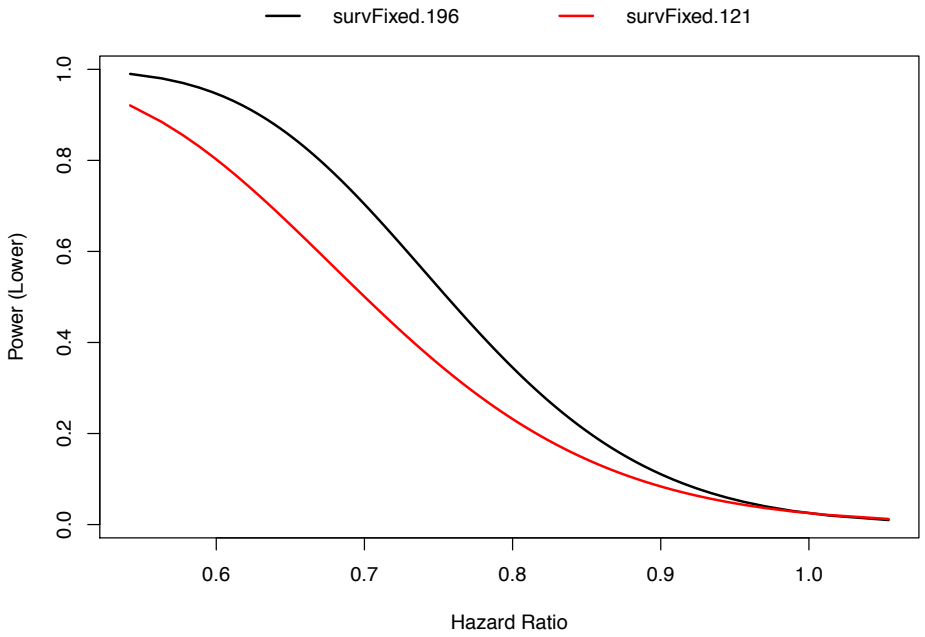
```
> survFixed.121 <- update( survFixed, sample.size=121,
                           power="calculate" )
> survFixed.121
Call:
seqDesign(prob.model = "hazard", arms = 2, null.hypothesis = 1,
    alt.hypothesis = 0.67, ratio = c(1, 1), nbr.analyses = 1,
    sample.size = 121, test.type = "less", power = "calculate",
    alpha = 0.025)

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.00    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.67    (power = 0.5959)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                         a      d
    Time 1 (N= 121) 0.7002 0.7002
```

SISCR
UW - 2016

Case Study : Hodgkin's Trial

Statistical power using RCTdesign

- Compare power curves using `seqPlotPower()`

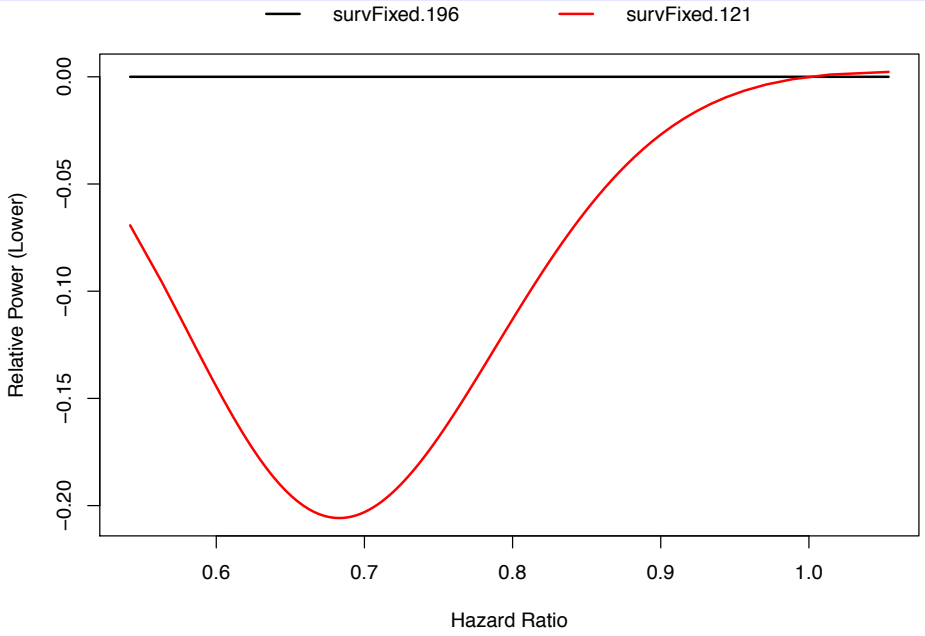


SISCR
UW - 2016

Case Study : Hodgkin's Trial

Statistical power using RCTdesign

- Often more useful to compare differences between power curves
- Use the `reference` argument in `seqPlotPower()`



Case Study : Hodgkin's Trial

Candidate group sequential designs

- Principles in guiding initial choice of stopping rule
  - Early conservatism
    - Long-term benefit of high importance
    - Early stopping precludes the observation of long-term safety data
  - Ability to stop early for futility
    - Safety concerns
    - Logistical considerations (monetary)
  - Number and timing of interim analyses
    - Trade-off between power and sample size
    - Determined by information accrual (events) but ultimately scheduled on calendar time

Case Study : Hodgkin's Trial

Candidate group sequential designs

- `SymmOBF.2, SymmOBF.3, SymmOBF.4`
  - One-sided symmetric stopping rules with O'Brien-Fleming boundary relationships having 2, 3, and 4 equally spaced analyses,respectively, and a max sample size of 196 events
- `SymmOBF.Power`
  - One-sided symmetric stopping rule with O'Brien-Fleming boundary having 4 equally spaced analyses, and 80% under the alternative hypothesis (HR=0.67)
- `Futility.5, Futility.8, Futility.9`
  - One-sided stopping rules from the unified family [5] with a total of 4 equally spaced analyses, with a maximal sample size of 196 events, and having O'Brien-Fleming lower (efficacy) boundary relationships and upper (futility) boundary relationships corresponding to boundary shape parameters P = 0.5, 0.8, and 0.9, respectively. P = 0.5 corresponds to Pocock boundary shape functions, and P = 1.0 corresponds to O'Brien-Fleming boundary relationships

Case Study : Hodgkin's Trial

Candidate group sequential designs

- `Eff11.Fut8, Eff11.Fut9`
  - One-sided stopping rules from the unified family with a total of 4 equally spaced analyses, with a maximal sample size of 196 events, and having lower (efficacy) boundary relationships corresponding to boundary shape parameter P = 1.1 and upper (futility) boundary relationships corresponding to boundary shape parameters P = 0.8, and 0.9, respectively. P = 0.5 corresponds to Pocock boundary shape functions, and P = 1.0 corresponds to O'Brien-Fleming boundary relationships
- `Fixed.Power`
  - A fixed sample study which provides the same power to detect the alternative (HR=0.67) as the `Futility.8` trial design

Case Study : Hodgkin's Trial

Candidate group sequential designs

- Specification of candidate designs using `update()`

```
> Fixed <- survFixed
>
> SymmOBF.2 <- update( Fixed, nbr.analyses=2, P=c(1,1),
                       sample.size=196, power="calculate" )
> SymmOBF.3 <- update( SymmOBF.2, nbr.analyses = 3, P=c(1,1) )
> SymmOBF.4 <- update( SymmOBF.2, nbr.analyses = 4, P=c(1,1) )
> SymmOBF.Power <- update( SymmOBF.4, power = 0.80 )
>
> Futility.5 <- update( SymmOBF.4, P=c(1,.5) )
> Futility.8 <- update( SymmOBF.4, P=c(1,.8) )
> Futility.9 <- update( SymmOBF.4, P=c(1,.9) )
>
> Eff11.Fut8 <- update( SymmOBF.4, P=c(1.1,.8) )
> Eff11.Fut9 <- update( SymmOBF.4, P=c(1.1,.9) )
>
> Fixed.Power <- update( SymmOBF.2, nbr.analyses=1, power=0.7767 )
```


Case Study : Hodgkin's Trial

SISCR
UW - 2016

Candidate group sequential designs

- Stopping boundaries for `SymmOBF.4`

```
> SymmOBF.4
Call:
seqDesign(prob.model = "hazard", arms = 2, null.hypothesis = 1,
    alt.hypothesis = 0.67, ratio = c(1, 1), nbr.analyses = 4,
    sample.size = 196, test.type = "less", power = "calculate",
    alpha = 0.025, P = c(1, 1))

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >= 1.00    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.67    (power = 0.7837)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                        a      d
    Time 1 (N=  49) 0.3183 1.7724
    Time 2 (N=  98) 0.5642 1.0000
    Time 3 (N= 147) 0.6828 0.8263
    Time 4 (N= 196) 0.7511 0.7511
```

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

Boundaries on various design scales

- Normalized Z statistic: $Z_j = z_j = (\hat{\theta}_j - \theta_0)/se(\hat{\theta}_j)$

```
> seqBoundary( SymmOBF.4, scale="Z" )
STOPPING BOUNDARIES: Normalized Z-value scale
                          a       d
    Time 1 (N=  49) -4.0065  2.0032
    Time 2 (N=  98) -2.8330  0.0000
    Time 3 (N= 147) -2.3131 -1.1566
    Time 4 (N= 196) -2.0032 -2.0032
```


Case Study : Hodgkin's Trial

SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Boundaries on various design scales

- Fixed sample P value statistic: $P_j = \Phi(z_j)$

```
> 1-seqBoundary( SymmOBF.4, scale="P" )
STOPPING BOUNDARIES: Fixed Sample P-value scale
                        a      d
    Time 1 (N=  49) 0.0000 0.9774
    Time 2 (N=  98) 0.0023 0.5000
    Time 3 (N= 147) 0.0104 0.1237
    Time 4 (N= 196) 0.0226 0.0226
```

Case Study : Hodgkin's Trial

Boundaries on various design scales

- Error spending statistic:

$$E_{aj} = \frac{1}{\alpha_L}\left(\Pr\left[S_j \le s_j, \bigcap_{k=1}^{j-1} S_k \in C_k \mid \theta = \theta_0\right] + \sum_{\ell=1}^{j-1} \Pr\left[S_\ell \le a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k \mid \theta = \theta_0\right]\right),$$

where α_L is the lower type I error of the stopping rule defined by

$$\alpha_L = \sum_{\ell=1}^{J} \Pr\left[S_\ell \le a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k | \theta = \theta_0\right].$$

```
> seqBoundary( SymmOBF.4, scale="E" )
STOPPING BOUNDARIES: Error Spending Function scale
                          a      d
    Time 1 (N=  49) 0.0012 0.0012
    Time 2 (N=  98) 0.0927 0.0927
    Time 3 (N= 147) 0.4470 0.4470
    Time 4 (N= 196) 1.0000 1.0000
```


Case Study : Hodgkin's Trial

Boundaries on various design scales

- Error spending statistic:

$$E_{aj} = \frac{1}{\alpha_L}\left(\Pr\left[S_j \le s_j, \bigcap_{k=1}^{j-1} S_k \in C_k \mid \theta = \theta_0\right] + \sum_{\ell=1}^{j-1} \Pr\left[S_\ell \le a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k \mid \theta = \theta_0\right]\right),$$

where α_L is the lower type I error of the stopping rule defined by

$$\alpha_L = \sum_{\ell=1}^{J} \Pr\left[S_\ell \le a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k | \theta = \theta_0\right].$$

```
> seqBoundary( SymmOBF.4, scale="E" )*.025
STOPPING BOUNDARIES: Error Spending Function scale
                          a      d
    Time 1 (N=  49) 0.0000 0.0000
    Time 2 (N=  98) 0.0023 0.0023
    Time 3 (N= 147) 0.0112 0.0112
    Time 4 (N= 196) 0.0250 0.0250
```

SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

Case Study : Hodgkin's Trial

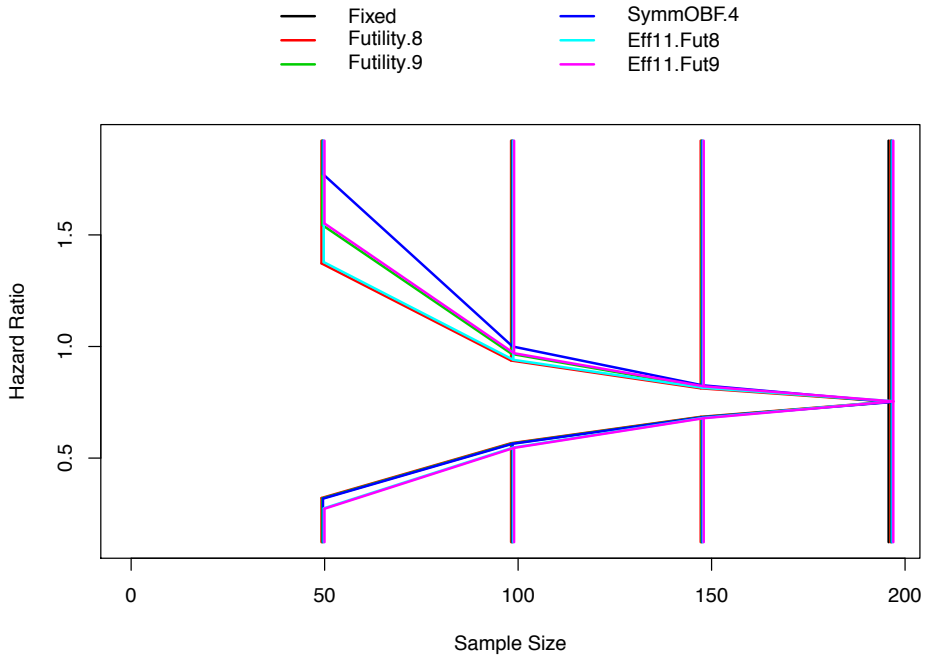
Boundaries on various design scales

- `RCTdesign` also has the ability to incorporate prior distributions for treatment effects in order to evaluate:
  - Bayesian posterior probabilities
  - Bayesian predictive probabilities

Case Study : Hodgkin's Trial

Visual comparison of stopping boundaries

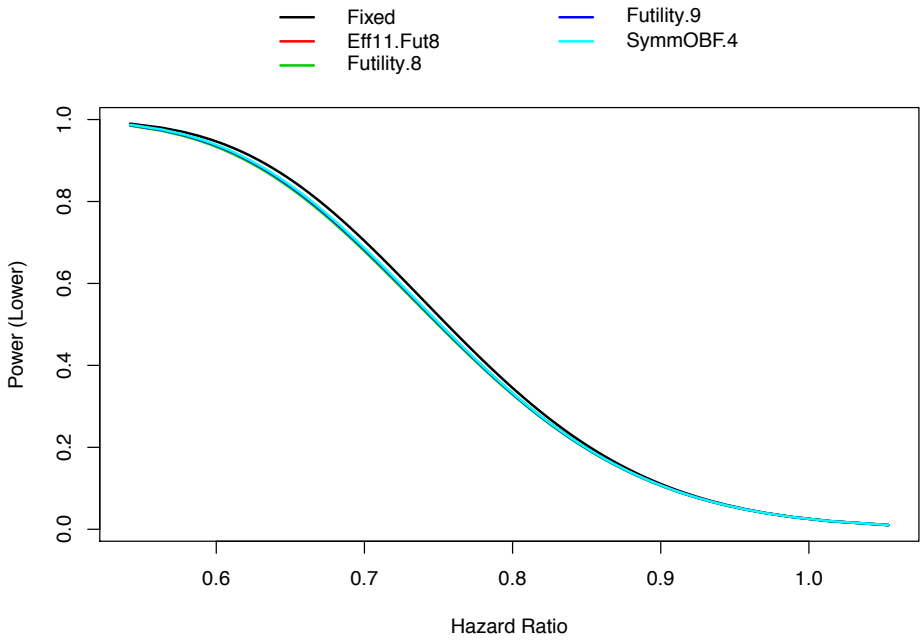
- Stopping boundaries can be plotted using `seqPlotBoundary()`



Case Study : Hodgkin's Trial

Visual comparison of statistical power for selected designs

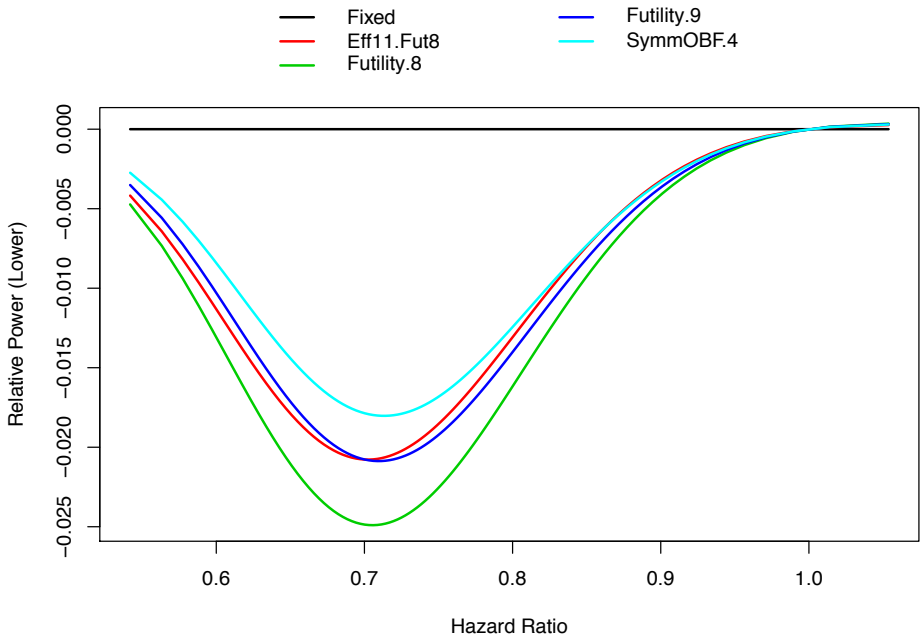
- Power curves (or differences) can be plotted with `seqPlotPower()`



Case Study : Hodgkin's Trial

Visual comparison of statistical power for selected designs

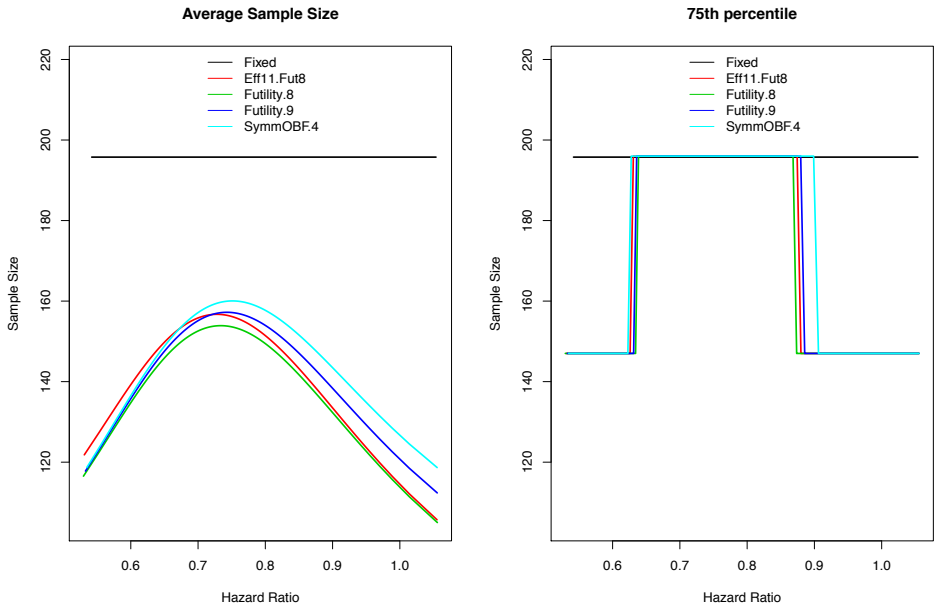
- As before, power curves (or differences) can be plotted with `seqPlotPower()`



Case Study : Hodgkin's Trial

Comparison of sample size distributions

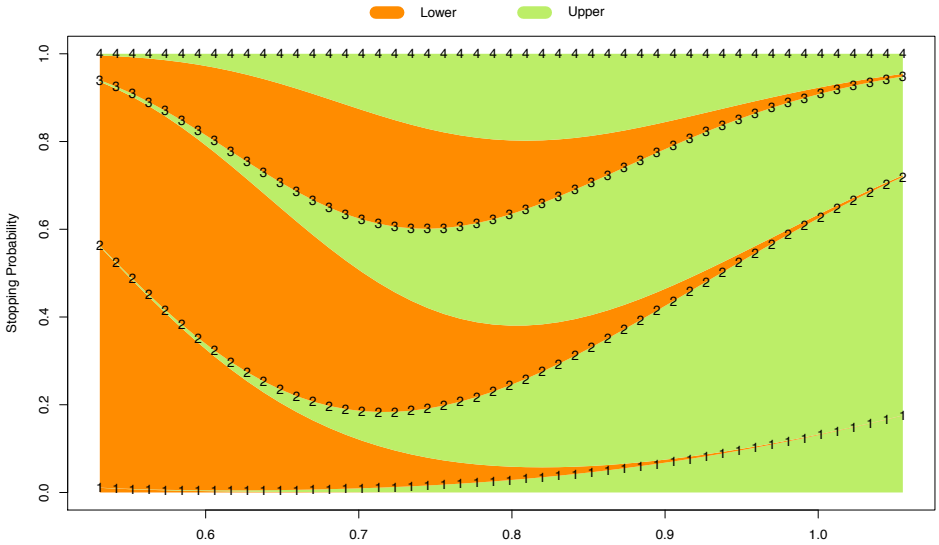
- Mean and quantiles of the sample size distribution can be plotted with `seqPlotASN()`



Case Study : Hodgkin's Trial

Stopping probabilities at each analysis for design `Eff11.Fut8`

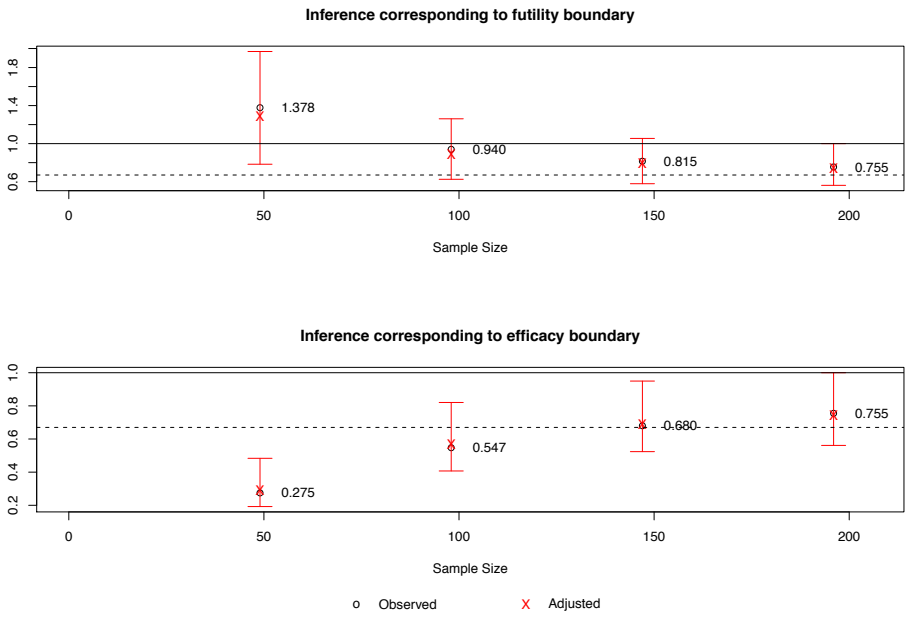
- Plot stopping probabilities using the `seqPlotStopProb()` function



Case Study : Hodgkin's Trial

Inference at each analysis for design `Eff11.Fut8`

- Plot inference on the boundaries using the `seqPlotStopProb()` function



Case Study : Hodgkin's Trial

Tabulation of operating characteristics for design `Eff11.Fut8`

- Computed operating characteristics can be obtained with the `seqOC()` function

```
> seqOC( Eff11.Fut8,  theta=seq(.6,1,by=.2) )
Operating characteristics
 Theta     ASN Power.lower
   0.6 139.24      0.9354
   0.8 151.43      0.3319
   1.0 114.51      0.0250

Stopping Probabilities:
 Theta Time 1 Time 2 Time 3 Time 4
   0.6 0.0049 0.3339 0.4757 0.1855
   0.8 0.0286 0.2174 0.3891 0.3649
   1.0 0.1308 0.4939 0.2830 0.0923
```

seqDesign() for extended investigation of accrual patterns

seqDesign()

- Recall that `seqPHSubjects()` can be used to estimate accrual and event rates under the assumption of
  - Exponential baseline survival
  - Proportional hazards treatment effect
  - Uniform accrual
  - Negligible dropout
- For survival studies, `seqDesign()` incorporates accrual assumptions into the `seqDesign()` object and allows for added flexibility in the definition of accrual / event rates

seqDesign() for extended investigation of accrual patterns

SISCR
UW - 2016

Group Sequential Designs
Statistical framework for trial monitoring
Types of group sequential designs
Case Study: Design of Hodgkin's Trial
Background
Fixed Sample Design
Group sequential design evaluations
Extended investigation of accrual patterns

seqDesign()

- `seqDesign()` provides added flexibility
  - Baseline survival : exponential, weibull, piecewise exponential, pilot data
  - Accrual : uniform, beta, piecewise uniform, pilot data
  - Dropout : exponential, weibull, piecewise exponential, pilot data
- `seqDesign()` relies upon simulation for estimation of accrual / event rates

Output from `seqDesign()`

Ex: Hodgkin's trial

- As an example of `seqDesign()`, again consider the Hodgkin's trial
- There we assumed:
  - Median survival in the control arm of 9 months
  - Uniform accrual over 3 years with one additional year of followup
- Let's consider the event rates/timing of analyses when accrual is:
  - Early (Beta(2,1))
  - Late (Beta(1,2))

Output from `seqDesign()`

Ex: Hodgkin's trial

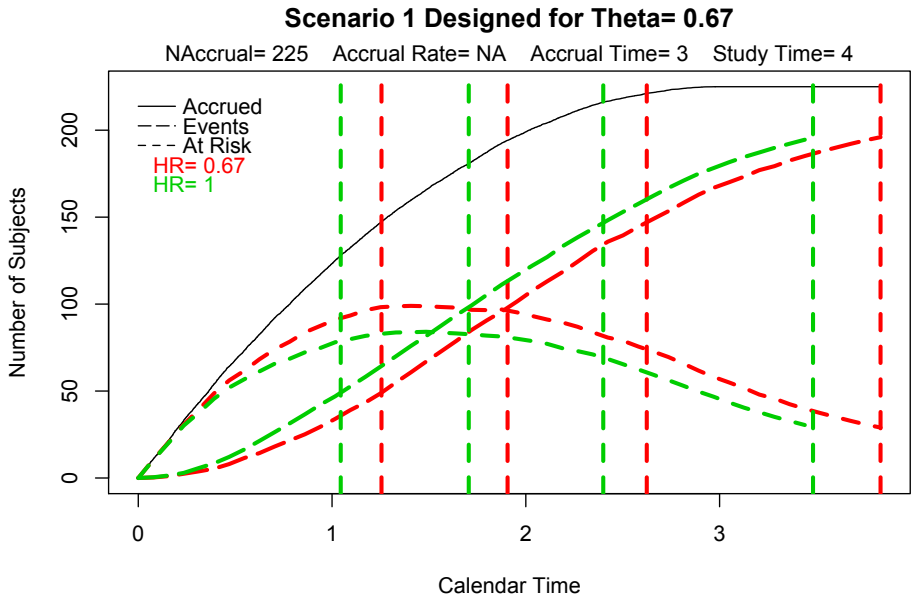
- Call to `seqDesign()` defining the `Eff11.Fut8` design:

```
##
#####        Exploration of analysis timing and total number
#####        of subjects accrued if total study time fixed at 4
##
##        Fast early accrual
##
Eff11.Fut8Extd.early <- seqDesign(prob.model = "hazard", arms = 2,
     null.hypothesis = 1.,  alt.hypothesis = 0.67, ratio = c(1., 1.),
     nbr.analyses = 4,  test.type = "less", alpha = 0.025,
     sample.size=196, power="calculate", P=c(1.1,.8), accrualTime=3,
     studyTime=4,  bShapeAccr=2,  eventQuantiles=.75,
     nPtsSim=10000, seed=0)

##
##        Slow early accrual
##
Eff11.Fut8Extd.late <- seqDesign(prob.model = "hazard", arms = 2,
     null.hypothesis = 1., alt.hypothesis = 0.67, ratio = c(1., 1.),
     nbr.analyses = 4,  test.type = "less", alpha = 0.025,
     sample.size=196, P=c(1.1,.8),  accrualTime=3, studyTime=4,
     aShapeAccr=2, eventQuantiles=.75, nPtsSim=10000, seed=0)
```

Output from `seqDesign()`

Sensitivity to the accrual distribution

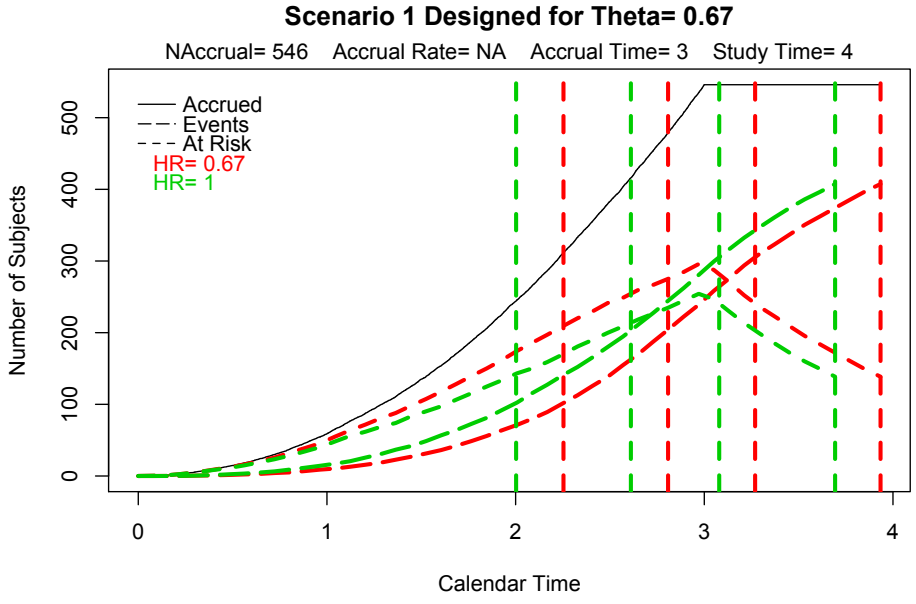
- Plot timing of analyses under early accrual
  - `seqPlotPHNSubjects(Eff11.Fut8Extd.early)`



Output from `seqDesign()`

Sensitivity to the accrual distribution

- Plot timing of analyses under late accrual
  - `seqPlotPHNSubjects(Eff11.Fut8Extd.late)`



Sequential and Adaptive Analysis with Time-to-Event Endpoints

Session 3 - Monitoring Group Sequential Designs with Time-to-Event Endpoints

Presented July 29, 2016

Scott S. Emerson
Department of Biostatistics
University of Washington

Daniel L. Gillen
Department of Statistics
University of California, Irvine

Monitoring group sequential trials

Operating characteristics to consider at the design stage

1. Standard for evidence and efficiency of designs
   - Type I error
   - Power at various alternatives
   - Average sample number (ASN) / stopping probabilities
2. Point estimates of treatment effect corresponding to boundary decisions in favor of
   - Efficacy – Futility – Harm
3. Frequentist/Bayesian/Likelihood inference on the boundaries
4. Conditional futility/reversal of decision corresponding to boundary decisions

All dependent on the sampling density of the test statistic...

SISCR
UW - 2016

Monitoring group sequential trials

RECALL: Group sequential sampling density

- Consider independent observations $X_1, \ldots, X_{n_J}$ with $E[X_i] = \theta$, $i = 1, \ldots, n_J$
- Interested in testing $H_0 : \theta = \theta_0$ based upon a maximum of J analyses
- Let S_j denote the test statistic computed at interim analysis j using observations $1, \ldots, n_j$, and suppose that $S_j \dot{\sim} N(\theta V_j, V_j)$, $j = 1, \ldots, J$
- At each analysis we partition the outcome space for statistic S_j into *stopping set* $\mathcal{S}_j$ and *continuation set* $\mathcal{C}_j$
  - If $S_j \in \mathcal{S}_j$, the trial is stopped.
  - Otherwise, $S_j \in \mathcal{C}_j$ and the study continues to gather additional observations.

SISCR
UW - 2016

Monitoring group sequential trials

RECALL: Group sequential sampling density

- Under an independent increments covariance structure, the sampling density of the bivariate group sequential statistic (M, S_M), where $M = \min\{j : S_j \notin \mathcal{C}_j\}$ is given by

$$p(m, s; \theta) = \begin{cases} f(m, s; \theta) & s \notin \mathcal{C}_m \\ 0 & \text{otherwise} \end{cases},$$

where the function $f(j, s; \theta)$ is given recursively by,

$$f(1, s; \theta) = \frac{1}{\sqrt{V_1}} \phi\left(\frac{s - \theta V_1}{\sqrt{V_1}}\right)$$

$$f(j, s; \theta) = \int_{C_{j-1}} \sqrt{v_j} \phi\left(\frac{s - u - v_j}{\sqrt{v_j}}\right) f(j-1, u; \theta) du, j = 2, ..., m$$

with $v_j = V_j - V_{j-1}$ and $\phi(x) = \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}$.

SISCR
UW - 2016

Monitoring group sequential trials

Operating characteristics condition upon exact timing

- When S_j represents the score statistic resulting from a parametric probability model, $Var[S_j] = V_j = \mathcal{I}_j$ is Fisher Information
- The group sequential density (and hence all of the previously mentioned operating characteristics) will depend upon the timing of analyses as measured by the information accrued
- Most commonly, we carry out *maximal information trials*
  - Specify the maximum information that will be entertained
    - Usually in order to guarantee a specified power at a clinically relevant alternative
  - Interim analyses are then planned according to the proportion of the maximal sample size that has been accrued to the trial ($\Pi_j \equiv V_j / V_J$)

Monitoring group sequential trials

SISCR
UW - 2016

Operating characteristics condition upon exact timing

- During the conduct of a study the timing of analyses may change because:
  - Monitoring scheduled by calendar time
  - Slow (or fast) accrual
  - External causes (should not be influenced by study results)
  - Statistical information from a sampling unit may be different than originally estimated
    - Variance of measurements
    - Baseline event rates (binary outcomes)
    - Censoring and survival distributions (weighted survival statistics)
- Consequences of these changes can include
  - Change in nominal type I error rate from originally planned design
  - Change in power from originally planned design

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Example: Stopping rule chosen at design

- Test of normal mean:
  - $H_0 : \theta \leq 0.0$
  - $H_1 : \theta \geq 0.5$
- One-sided symmetric test
  - Size .025, Power .975
  - Four equally spaced analyses
  - Pocock (1977) boundary relationships

Example: Stopping rule chosen at design

```
> dsn <- seqDesign( prob.model="normal", arms=1, null.hypothesis=0,
+          alt.hypothesis=0.5, test.type="greater", variance=4,
+          power=0.975, P=0.5, nbr.analyses=4, early.stopping="both" )

> dsn

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
           Null hypothesis : Theta <= 0.0    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                       Futility Efficacy
    Time 1 (N=  86.31)   0.0000   0.5000
    Time 2 (N= 172.62)   0.1464   0.3536
    Time 3 (N= 258.92)   0.2113   0.2887
    Time 4 (N= 345.23)   0.2500   0.2500
```

Monitoring group sequential trials

Analyses after 40%, 60%, 80%, 100% (maintain power)

```
> dsn.late.power <- update(dsn, sample.size=c(.4,.6,.8,1) )

> dsn.late.power

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
           Null hypothesis : Theta <= 0.0    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Futility Efficacy
    Time 1 (N= 131.97)   0.1047   0.3953
    Time 2 (N= 197.95)   0.1773   0.3227
    Time 3 (N= 263.93)   0.2205   0.2795
    Time 4 (N= 329.91)   0.2500   0.2500
```


Monitoring group sequential trials

SISCR
UW - 2016

Analyses after 40%, 60%, 80%, 100% (maintain max sample size)

```
> dsn.late.n <- update(dsn,
        sample.size=c(.4,.6,.8,1)*max(dsn$parameters$sample.size),
        alt.hypothesis="calculate" )

> dsn.late.n

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
           Null hypothesis : Theta <= 0.0000    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.4888    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Futility Efficacy
    Time 1 (N= 138.09)   0.1024   0.3864
    Time 2 (N= 207.14)   0.1733   0.3155
    Time 3 (N= 276.19)   0.2155   0.2732
    Time 4 (N= 345.23)   0.2444   0.2444
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Changes in the number of analyses

- During the conduct of a study, the number of analyses may also be different from design stage
  - Monitoring scheduled by calendar time
  - Slow (or fast) accrual
  - External causes (should not be influenced by study results)
- This will also result in changes to design operating characteristics

Example: Stopping rule chosen at design (cont'd)

```
> dsn <- seqDesign( prob.model="normal", arms=1, null.hypothesis=0,
+         alt.hypothesis=0.5, test.type="greater", variance=4,
+         power=0.975, P=0.5, nbr.analyses=4, early.stopping="both" )

> dsn

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
          Null hypothesis : Theta <= 0.0    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                       Futility Efficacy
    Time 1 (N=  86.31)   0.0000   0.5000
    Time 2 (N= 172.62)   0.1464   0.3536
    Time 3 (N= 258.92)   0.2113   0.2887
    Time 4 (N= 345.23)   0.2500   0.2500
```

Monitoring group sequential trials

Analyses after 20%, 40%, 60%, 80%, 100% (maintain power)

```
> dsn.5.power <- update(dsn, sample.size=c(.2,.4,.6,.8,1) )

> dsn.5.power

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
           Null hypothesis : Theta <= 0.0    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                       Futility Efficacy
    Time 1 (N=  72.10)  -0.0590   0.5590
    Time 2 (N= 144.20)   0.1047   0.3953
    Time 3 (N= 216.31)   0.1773   0.3227
    Time 4 (N= 288.41)   0.2205   0.2795
    Time 5 (N= 360.51)   0.2500   0.2500
```


Monitoring group sequential trials

SISCR
UW - 2016

Analyses after 20%, 40%, 60%, 80%, 100% (maintain max sample size)

```
> dsn.5.n <- update(dsn,
        sample.size=c(.2,.4,.6,.8,1)*max(dsn$parameters$sample.size),
        alt.hypothesis="calculate" )

> dsn.5.n

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
           Null hypothesis : Theta <= 0.0000    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5109    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                       Futility Efficacy
    Time 1 (N=  69.05)  -0.0603   0.5713
    Time 2 (N= 138.09)   0.1070   0.4039
    Time 3 (N= 207.14)   0.1811   0.3298
    Time 4 (N= 276.19)   0.2253   0.2856
    Time 5 (N= 345.23)   0.2555   0.2555
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Monitoring group sequential trials

Result of changing schedule of analyses

- Summary for Pocock boundary relationships

Analysis Times	Alt	Max N	Bound
.25, .50, .75, 1.00	.500	345.23	.2500
.40, .60, .80, 1.00	.500	329.91	.2500
.40, .60, .80, 1.00	.489	345.23	.2444
.20, .40, .60, .80, 1.00	.500	360.51	.2500
.20, .40, .60, .80, 1.00	.511	345.23	.2555

Monitoring group sequential trials

SISCR
UW - 2016

Result of changing schedule of analyses

- Summary for O'Brien-Fleming boundary relationships

Analysis Times	Alt	Max N	Bound
.25, .50, .75, 1.00	.500	256.83	.2500
.40, .60, .80, 1.00	.500	259.44	.2500
.40, .60, .80, 1.00	.503	256.83	.2513
.20, .40, .60, .80, 1.00	.500	259.45	.2500
.20, .40, .60, .80, 1.00	.503	256.83	.2513

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Constrained Boundaries Example

Constrained O'Brien-Fleming Design

- It is often desirable to modify a stopping rule at the design stage to maintain a particular set of boundary constraints
- For example, an O'Brien-Fleming stopping rule is known for extreme conservatism at early analysis
  - One-sided level .025 test of a normal mean with four equally spaced analyses
  - Stopping at first analysis for efficacy requires a fixed sample P-value of less than .0001

```
> obf <- seqDesign( prob.model="normal", arms=1, null.hypothesis=0,
+          alt.hypothesis=0.5, test.type="greater", variance=4,
+          power=0.975, P=1, nbr.analyses=4, early.stopping="both" )
```


Constrained Boundaries Example

SISCR
UW - 2016

Constrained O'Brien-Fleming Design

```
> obf

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
          Null hypothesis : Theta <= 0.0    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5    (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Futility Efficacy
    Time 1 (N=  64.21)  -0.5000   1.0000
    Time 2 (N= 128.41)   0.0000   0.5000
    Time 3 (N= 192.62)   0.1667   0.3333
    Time 4 (N= 256.83)   0.2500   0.2500

> seqBoundary(obf, scale="P")
STOPPING BOUNDARIES: Fixed Sample P-value scale
                     Futility Efficacy
    Time 1 (N=  64.21)   0.9774   0.0000
    Time 2 (N= 128.41)   0.5000   0.0023
    Time 3 (N= 192.62)   0.1237   0.0104
    Time 4 (N= 256.83)   0.0226   0.0226
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Constrained Boundaries Example

Constrained O'Brien-Fleming Design

- Some sponsor's wish for the operating characteristics of an O'Brien-Fleming design but desire a slightly less conservative first boundary
- One possibility is to constrain the O'Brien-Fleming design at the first analysis so that the efficacy bound corresponds to a P-value of 0.0005
- In order to maintain the overall type I error rate, the value of G must be re-computed using this constraint
- This can be done using an `exact.constraint`:

```
> bnd.const <- as.seqBoundary( cbind(matrix(NA,nrow=4,ncol=3),
                c(.0005,rep(NA,3))), scale="P" )
> bnd.const
STOPPING BOUNDARIES: Fixed Sample P-value scale
        a  b  c     d
Time 1 NA NA NA 5e-04
Time 2 NA NA NA    NA
Time 3 NA NA NA    NA
Time 4 NA NA NA    NA
```

SISCR
UW - 2016

Constrained Boundaries Example

Constrained O'Brien-Fleming Design

```
> obf.const <- update( obf, exact.constraint=bnd.const )
> obf.const

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a greater alternative:
          Null hypothesis : Theta <= 0.0    (size  = 0.025)
    Alternative hypothesis : Theta >= 0.5    (power = 0.975)

STOPPING BOUNDARIES: Sample Mean scale
                     Futility Efficacy
    Time 1 (N=  64.31)  -0.4990   0.8207
    Time 2 (N= 128.61)   0.0005   0.5005
    Time 3 (N= 192.92)   0.1670   0.3337
    Time 4 (N= 257.23)   0.2502   0.2502

> seqBoundary(obf.const, scale="P")
STOPPING BOUNDARIES: Fixed Sample P-value scale
                     Futility Efficacy
    Time 1 (N=  64.31)   0.9773   0.0005
    Time 2 (N= 128.61)   0.4989   0.0023
    Time 3 (N= 192.92)   0.1231   0.0102
    Time 4 (N= 257.23)   0.0224   0.0224
```

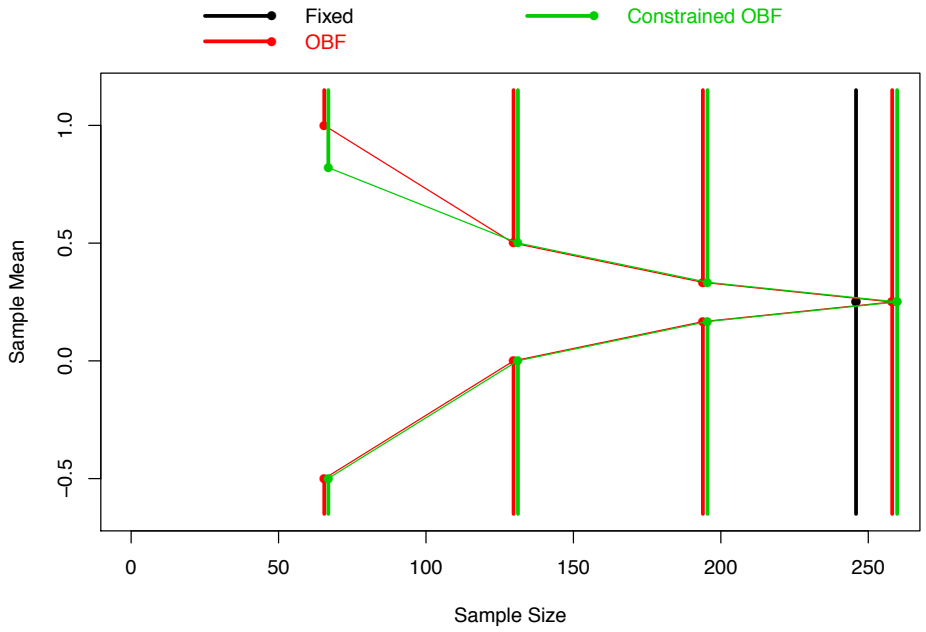
SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Constrained Boundaries Example

Constrained O'Brien-Fleming Design

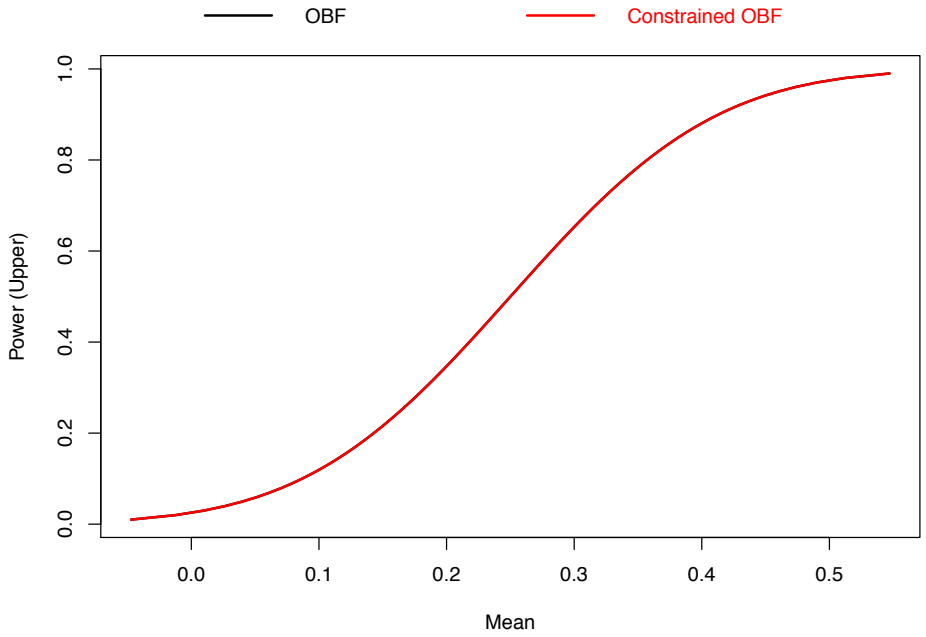
- Comparison of stopping boundaries (sample mean scale)



Constrained Boundaries Example

Constrained O'Brien-Fleming Design

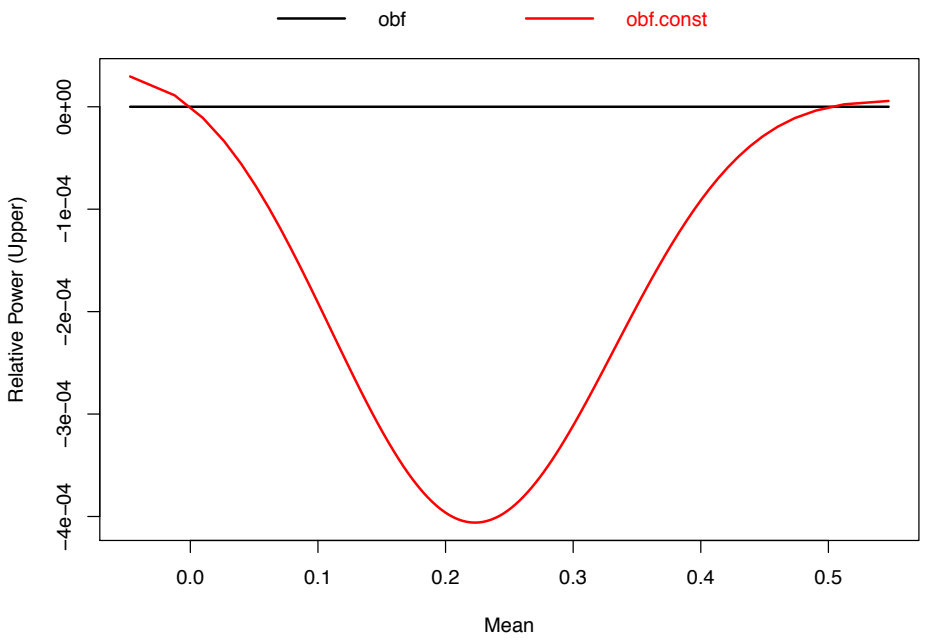
- Comparison of statistical power



Constrained Boundaries Example

Constrained O'Brien-Fleming Design

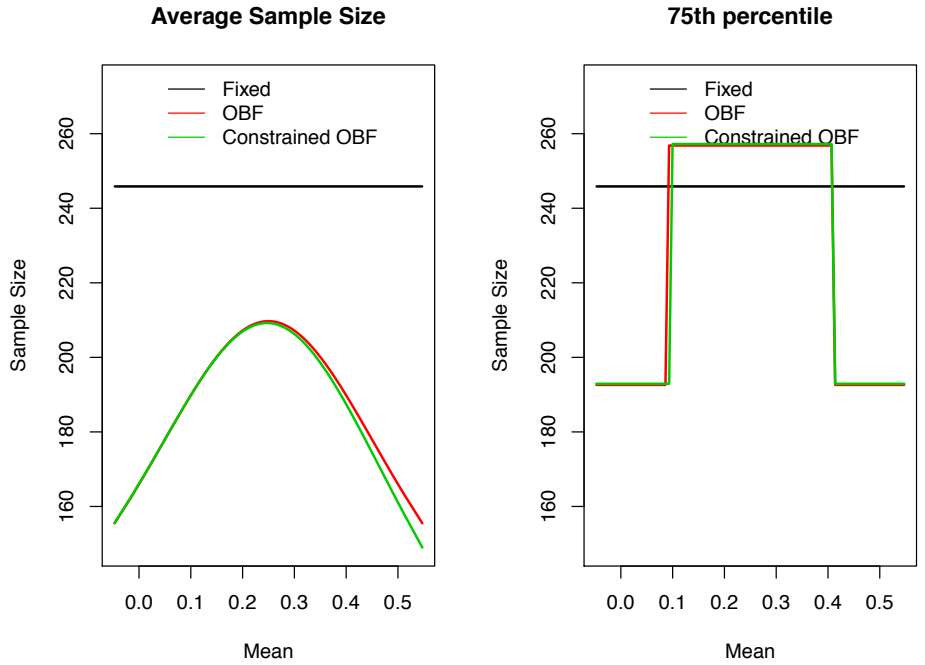
- Comparison of statistical power



Constrained Boundaries Example

Constrained O'Brien-Fleming Design

- Comparison of sample size distribution



Result of changing schedule of analyses

- As previously noted, during the conduct of a study the timing of analyses may change because:
  - Monitoring scheduled by calendar time
  - Slow (or fast) accrual
  - External causes (should not be influenced by study results)
  - Statistical information from a sampling unit may be different than originally estimated
    - Variance of measurements
    - Baseline event rates (binary outcomes)
    - Censoring and survival distributions (weighted survival statistics)

Monitoring group sequential trials

SISCR
UW - 2016

Result of changing schedule of analyses

- Need methods that allow flexibility in determining number and timing of analyses
- Should maintain some (but not, in general, all) desired operating characteristics, e.g.:
  - Type I error
  - Type II error
  - Maximal sample size
  - Futility properties
  - Bayesian properties

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Monitoring group sequential trials

Popular methods for flexible implementation of group sequential boundaries

1. Christmas tree approximation for triangular tests: Whitehead and Stratton (1983)
2. Error spending functions: Lan and DeMets (1983); Pampallona, Tsiatis, and Kim (1995)
3. Constrained boundaries in unified design family: Emerson (2000); Burrington & Emerson (2003)

Monitoring group sequential trials

Common features

- Stopping rule specified at design stage parameterizes the boundary for some statistic (boundary scale)
  - Error spending family (Lan & Demets, 1983) $\rightarrow$ proportion of type I error spent
  - Unified family (Emerson & Kittelson, 1999) $\rightarrow$ point estimate (MLE)
- At the first interim analysis, parametric form is used to compute the boundary for actual time on study
- At successive analyses, the boundaries are recomputed accounting for the exact boundaries used at previously conducted analyses
- Maximal sample size estimates may be updated to maintain power
  - For binary outcomes, generally use pooled estimate of event rates to withhold treatment effect from study sponsor

SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Error spending functions

Implementing error spending functions

- *Error spending* (also known as α-spending) allow flexible implementation by pre-specifying a rate at which the type I error will be “spent" at each interim analysis; specifically:
  - Let α denote the type I error probability for the trial.
  - Use the group sequential sampling density to calculate the stopping probabilities (α_j) over the prior interim analyses.
  - Let α_j denote the probability of rejecting the null hypothesis at the *j*th interim analysis (then $\alpha = \sum_j \alpha_j$).
  - *Error spending function:* Let $\alpha(\Pi)$ denote a function that constrains the probability of rejecting the null hypothesis at or before $100 \times \Pi$% of the total information; that is:

$$\alpha(\Pi) = \frac{1}{\alpha} \sum_{j:\Pi_j < \Pi} \alpha_j \qquad (1)$$

Thus, $\alpha(\Pi)$ is the proportion of the total type I error that has been “spent" when there is Π information in the trial.

SISCR
UW - 2016

Error spending functions

Implementing error spending functions

- Examples of error spending functions:

$$\begin{aligned}
\text{Constant spending:} \quad & \alpha(\Pi) = \Pi \\
\text{Power family:} \quad & \alpha(\Pi) = \Pi^P, \; P > 1 \\
\text{Approximate O'Brien-Fleming:} \quad & \alpha(\Pi) = \Phi\left(\frac{Z_{\alpha/2}}{\sqrt{\Pi}}\right) \\
\text{Approximate Pocock:} \quad & \alpha(\Pi) = \mathit{ln}[1 + (e-1)\Pi] \\
\text{Hwang, Shih, Decani, 1990:} \quad & \alpha(\Pi) = \frac{1 - e^{-\gamma\Pi}}{1 - e^{-\gamma}}, \; \gamma \neq 0
\end{aligned}$$

where $\Phi()$ is the standard normal cdf.

SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Error spending functions

Implementing error spending functions - Sepsis trial

- Critically ill patients often get overwhelming bacterial infection (sepsis), after which mortality is high
- Gram negative sepsis is often characterized by production of endotoxin, which is thought to be the cause of much of the ill effects of gram negative sepsis
- Hypothesis: Administering antibody to endotoxin may decrease morbidity and mortality
- Binary primary endpoint : 28 mortality (difference)

Error spending functions

Implementing error spending functions - Sepsis trial

- Consider a group sequential design with four equally spaced analyses utilizing an O'Brien-Fleming stopping rule (efficacy and futility)
  - Baseline event rate assumed to be 30%
  - Design alternative : 5% absolute decrease
  - One-sided type I error .025
  - N=1700 maximal patients

```
> sepsis.fix <- seqDesign(prob.model="proportions", arms=2,
                        size=.025, power="calculate",
                        null.hypothesis= c(.30, .30),
                        alt.hypothesis=c(0.25,0.30),
                        sample.size=1700, test.type="less")

> #****** pre-trial monitoring plan
> sepsis.obf <- update(sepsis.fix,nbr.analyses=4,P=1)
> sepsis.obf

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N=  425)  -0.1733   0.0866
    Time 2 (N=  850)  -0.0866   0.0000
    Time 3 (N= 1275)  -0.0578  -0.0289
    Time 4 (N= 1700)  -0.0433  -0.0433
```

Error spending functions

Implementing error spending functions - Sepsis trial

- Pre-trial analysis timing in terms of information:
  - Recall $V = 0.25 \times 0.75 + 0.3 \times 0.7$
  - Pre-trial planned information:

$$I = \frac{N_J/2}{V} = \frac{850}{0.3975} = 2138.4$$

  - Pre-trial plan for analysis timing:

Π_j	N_j	Information: $\frac{N_j}{2V}$
0.25	425	534.6
0.50	850	1069.2
0.75	1275	1603.8
1.00	1700	2138.4

Error spending functions

SISCR
UW - 2016

Implementing error spending functions - Sepsis trial

- Suppose the first interim analysis was conducted after data on 520 subjects (263 on the antibody arm, 257 on the placebo arm)
- Further suppose that 52 deaths were observed on the antibody arm and 65 deaths were observed on the placebo arm

$$\widehat{\theta}_1 = \frac{52}{263} \qquad \widehat{\theta}_0 = \frac{65}{257}$$

- Observed information at first interim analysis:

$$\begin{aligned} \widehat{S}_1 &= \frac{\hat{\theta}_1(1-\hat{\theta}_1)}{263} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{257} = 0.0013384 \\ \frac{1}{\widehat{S}_1} &= 747.2 \\ \Pi &= 747.2/2138.4 = 0.34942 \end{aligned}$$

Thus, we estimate that the first interim analysis has occurred at 34.9% of the planned total information.

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Error spending functions

Implementing error spending functions - Sepsis trial

- Pre-trial error-spending function:
  - Use `seqOC(sepsis.obf,theta=0)` to get the lower stopping probabilities at the interim analyses. These are the values of α_j. The pretrial error-spending function, $\alpha(\Pi)$ has values at Π_j defined by equation (1).

Π_j	a_j	Stopping Prob (α_j)	Cumulative type I error	Error spending function $\alpha(\Pi_j)$
0.25	-0.1733	0.00003	0.00003	0.00123
0.50	-0.0866	0.00229	0.00232	0.09274
0.75	-0.0578	0.00886	0.01176	0.44703
1.00	-0.0433	0.01382	0.02500	1.00000

  - To get values of $\alpha(\Pi)$ for $\Pi \neq \Pi_j$ we can either:
    - Use an error-spending function that approximates the pre-trial plan
    - Use linear interpolation

Error spending functions

Implementing error spending functions - Sepsis trial

- Using linear interpolation to find the critical value at 34.9% of total information:

$$\begin{aligned}\alpha(0.349) &= \alpha(0.25) + [\alpha(0.50) - \alpha(0.25)]\frac{0.349 - 0.25}{0.50 - 0.25} \\ &= 0.00003 + 0.00229 \times \frac{0.099}{0.25} \\ &= 0.00091872\end{aligned}$$

- Because this is the first interim analysis, we can calculate the revised value for a_1 directly from the normal density:

$$\begin{aligned}\frac{a_1}{\sqrt{\widehat{S}_1}} &= \Phi^{-1}(0.00091872) \\ &= -3.1153\end{aligned}$$

Thus, $a_1 = -3.1938\sqrt{0.0013384} = -0.11397$, and so we would continue because $\widehat{\theta}^{(1)} = -0.0552 > -0.11397$.

Error spending functions

Implementing error spending functions

- Notes:
  - At subsequent interim analyses we would repeat this process, but would need to account for the decision criteria used at earlier interim analyses to determine how much error should be spent and what the critical value should be.
  - We can develop analogous stopping criteria for the futility (d_j) boundary using a β-spending function.
  - I am not illustrating the above points because:
    - Error-spending scales do not directly elucidate the scientific/clinical aspects of the stopping criteria.
    - Error-spending scales do not do directly address changes in the estimated standard deviation at subsequent interim analyses.
  - (Note: any scale can be expressed on the sample mean scale, so you can (and should) consider the inference on the boundary when evaluating error-spending decision criteria.)

Error spending functions

Implementing error spending functions

- Error spending families have been implemented in `RCTdesign`
  - To get the error spending function from an existing design:

```
> update(sepsis.obf,display.scale="E")
```

  - To design a monitoring plan in the error spending scale:

```
> update(sepsis.obf,design.scale="E",
P=-1,display.scale="E")
> update(sepsis.obf,design.scale="E",
P=-1,display.scale="X")
```

  - This implements the power family of error spending functions described above: $\alpha(\Pi) = \Pi^P \times \alpha$

Constrained boundaries

- Constrained boundaries allow the same flexibility as error spending functions, but are constructed in the scale of the estimated treatment effects (or any scale desired).
- Overview:
  - Calculate the estimated information at the interim analysis as a proportion of the total information.
  - Calculate a revised group sequential design:
    - Use the values of a_ℓ and d_ℓ that were actually used at earlier interim analyses ($\ell < j$).
    - Calculate the new future values for a_ℓ and d_ℓ for $\ell \geq j$ using the original boundary shape function.
    - Find the value of G that maintains the desired operating characteristics.
    - (Implemented in the function `seqMonitor`).

Constrained boundaries - Sepsis example

- Recall the pre-trial interim analysis stopping rules:
  - With a “less than" alternative hypothesis:

$$a_j = -G\Pi_j^{-1}\sqrt{\frac{V}{850}}$$
$$d_j = (-2G + G\Pi_j^{-1})\sqrt{\frac{V}{850}}$$

- Pre-trial design ($\Pi_j = (0.25, 0.50, 0.75, 1.0)$, $G = 2.0032$):

Π_j	a_j	d_j
0.25	-0.1733	0.0866
0.50	-0.0866	0.0000
0.75	-0.0578	-0.0289
1.00	-0.0433	-0.0433

Constrained Boundaries

Constrained boundaries - Sepsis example

- Suppose we observe $\widehat{\theta}^{(1)} = -0.0552$ at 34.9% of total information.
- Calculate the revised design:
  - Use the same boundary shape function, but update as follows:

```
sepsis.IA1 <- update(sepsis.obf,
    sample.size=c(520,850,1275,1700),
    null.hypothesis=c(65/257,65/257),
    alt.hypothesis=c(52/263,65/257))
```

  - Now $G = 2.0036$ and the new stopping boundaries are:

Π_j	a_j	d_j
520	-0.1325	0.0514
850	-0.0810	0.0000
1275	-0.0541	-0.0270
1700	-0.0405	-0.0405

- Decision: continue the trial because $a_1 < \widehat{\theta}^{(1)} < d_1$.

Constrained Boundaries

Constrained boundaries - Sepsis example

- This approach can be automated using the (`seqMonitor()` function):
  - Create a vector of the results at the first interim analysis:

```
Y.1 <- c(rep(1,52),rep(0,263-52),rep(1,65),rep(0,257-65))
tx.1 <- c(rep(1,263),rep(0,257))
```

  - Determine revised boundaries and a stopping decision:

```
IA1 <- seqMonitor(sepsis.obf,response=Y.1,
        treatment=tx.1,future.analyses=c(850,1275,1700))
```

  - Results include:
    - Recommendation (continue)
    - Estimate ($\hat{\theta}_1 = -0.055$)
    - Revised stopping boundaries:

Π_j	a_j	d_j
520	-0.1325	0.0514
850	-0.0810	0.0000
1275	-0.0541	-0.0270
1700	-0.0405	-0.0405

SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Case Study : Hodgkin's Trial

Challenges in monitoring the Hodgkin's trial

- For a more complete example, let's consider monitoring the Hodgkin's trial from Session 2
- Recall that the primary endpoint was time to death with possible right-censoring
- Testing for group differences was based upon the logrank statistic (score test for the proportional hazards model)
- Under the proportional hazards model, statistical information is directly proportional to the number of observed events
- One complication in monitoring such a trial is to translate the from events to calendar time so that analyses/meetings can be scheduled

Case Study : Hodgkin's Trial

Chosen design

- `Eff11.Fut8` : P=1.1 efficacy bound with P=0.8 futility bound (Unified Family)

```
PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.00    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.67    (power = 0.7804)

STOPPING BOUNDARIES: Sample Mean scale
                          a      d
    Time 1 (N=  49) 0.2748 1.3782
    Time 2 (N=  98) 0.5474 0.9403
    Time 3 (N= 147) 0.6799 0.8151
    Time 4 (N= 196) 0.7549 0.7549
```

Case Study : Hodgkin's Trial

Chosen design

- `Eff11.Fut8` : P=1.1 efficacy bound with P=0.8 futility bound (Unified Family)

	Efficacy Bound			Futility Bound		
	lo.hr	lo.ztat	lo.pval	up.hr	up.zstat	up.pval
Time 1	0.275	-4.521	0.000	1.378	1.123	0.869
Time 2	0.547	-2.983	0.001	0.940	-0.305	0.380
Time 3	0.680	-2.339	0.010	0.815	-1.239	0.108
Time 4	0.755	-1.968	0.025	0.755	-1.968	0.025

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Timing of analyses

- Assumed
  - Uniform accrual over 3 years
  - One additional year of followup
  - Median survival in control arm of 9 months

```
> seqPHSubjects( Eff11.Fut8, controlMedian=0.75,
                      accrualTime=3, followupTime=1 )
  accrualTime followupTime   rate hazardRatio controlMedian nSubjects
1           3            1 75.459        1.00          0.75    226.38
2           3            1 80.598        0.67          0.75    241.79

  analysisTimes.1 analysisTimes.2 analysisTimes.3 analysisTimes.4
1          1.4474          2.2448          2.9599          4.0000
2          1.5033          2.3067          3.0142          4.0000
```

Case Study : Hodgkin's Trial

Timing of analyses

- Hypothetical data
  - Uniform accrual (80 subjects per year)
  - Median survival in the control arm of 1 year
  - True hazard ratio of 0.70
- Result
  - Longer median survival in control arm will result in longer time to accrue specified events
- Based upon initial estimates data is analyzed at 1.5 years of followup for DSMB meeting

Case Study : Hodgkin's Trial

SISCR
UW - 2016

1st interim analysis

- Monitoring at first interim analysis
  - Data stored in data frame `hodgData`

`grp`	:	Indicator of treatment group (0=control, 1=treatment)
`obsSurv`	:	Observed survival times
`event`	:	Indicator of mortality

  - Define response as a survival object

```
resp <- Surv( hodgData$obsSurv, hodgData$event )
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Case Study : Hodgkin's Trial

1st interim analysis

- Monitoring at first interim analysis
  - Specify remaining analysis at intended schedule to (roughly) maintain power (98, 147, 196)
  - Use function `seqMonitor()` to analyze current data and produce constrained boundaries

Case Study : Hodgkin's Trial

1st interim analysis

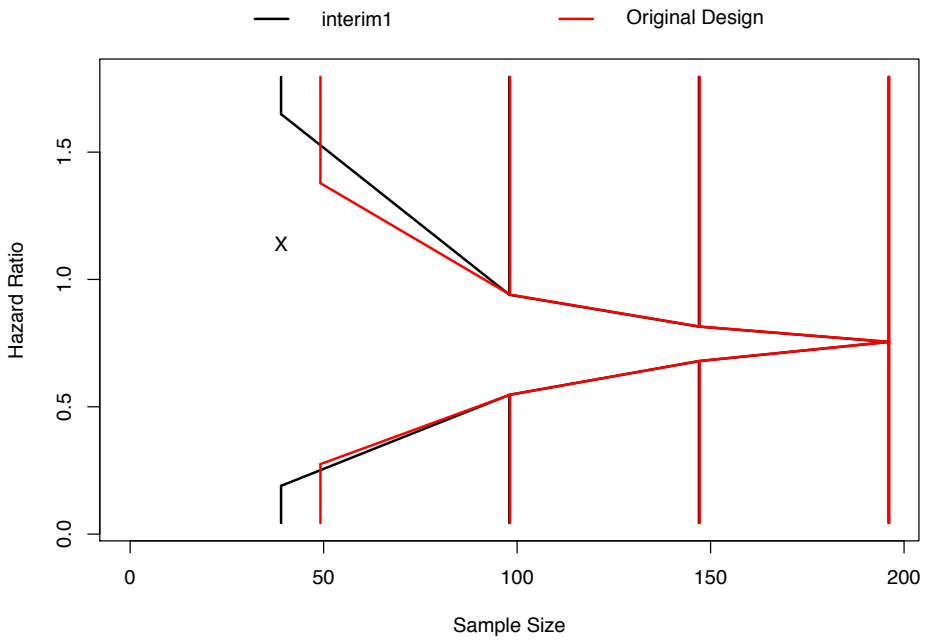
- Result of `seqMonitor()` at 1st analysis

```
RECOMMENDATION:
     Continue

OBSERVED STATISTICS:
 Sample Size Crude Estimate Z Statistic
          39          1.139      0.4062

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
         Null hypothesis : Theta >= 1.0000    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.6696    (power = 0.7804)

STOPPING BOUNDARIES: Sample Mean scale
                          a      d
    Time 1 (N=  39) 0.1895 1.6495
    Time 2 (N=  98) 0.5468 0.9399
    Time 3 (N= 147) 0.6795 0.8147
    Time 4 (N= 196) 0.7546 0.7546
```

Case Study : Hodgkin's Trial

Timing of 1st analysis

- Plot or monitoring result at 1st analysis



Case Study : Hodgkin's Trial

1st interim analysis

- Monitoring at first interim analysis
  - Notice that because of the longer median survival, the number of events at the first analysis are lower than expected (39 vs 49)
  - Would like to stick to original analysis schedule and accrual rate
  - Need to estimate event rates using POOLED data and estimate new analysis times

SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Case Study : Hodgkin's Trial

Estimate pooled survival at 1st analysis

- Estimate hazard from pooled data based upon exponential fit

```
> expFit <- survReg(Surv(obsSurv, event) ~ 1,
                    dist = "exponential", data = hodgData)
> estHaz <- exp( - expFit$coef )
```

Estimate event rates

- Estimate timing of future analyses based upon new pooled survival estimate

```
> seqPHSubjects( Eff11.Fut8, controlMedian=log(2)/estHaz,
        accrualTime=3, followupTime=1 )
  accrualTime followupTime   rate hazardRatio cntrlMedian nSubjects
1           3            1 87.999        1.00      1.1665  263.9991
2           3            1 96.757        0.67      1.1665  290.2737
  analysisTimes.1 analysisTimes.2 analysisTimes.3 analysisTimes.4
1        1.582587        2.389780        3.086729        4.000000
2        1.626356        2.436201        3.127887        4.000000
```


Case Study : Hodgkin's Trial

SISCR
UW - 2016

Estimate pooled survival at 1st analysis

- Determine the amount of additional followup needed in order to obtain desired events while maintaining accrual of 80 patients per year for 3 years

```
  accrualTime followupTime rate hazardRatio controlMedian nSubjects
1           3     1.572187   80        1.00      1.166507       240
2           3     2.215662   80        0.67      1.166507       240

  analysisTimes.1 analysisTimes.2 analysisTimes.3 analysisTimes.4
1        1.672433        2.534704        3.312677        4.572187
2        1.813171        2.733575        3.630260        5.215662
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Case Study : Hodgkin's Trial

Timing of 2nd interim analysis

- Monitoring at second interim analysis
  - Based upon previous estimates of pooled survival, next analysis conducted at 2.75 years
  - Specify remaining analysis at intended schedule to (roughly) maintain power (147, 196)
  - Use function `seqMonitor()` to analyze current data and produce constrained boundaries

Case Study : Hodgkin's Trial

SISCR
UW - 2016

2nd interim analysis

- Result of `seqMonitor()` at 2nd analysis

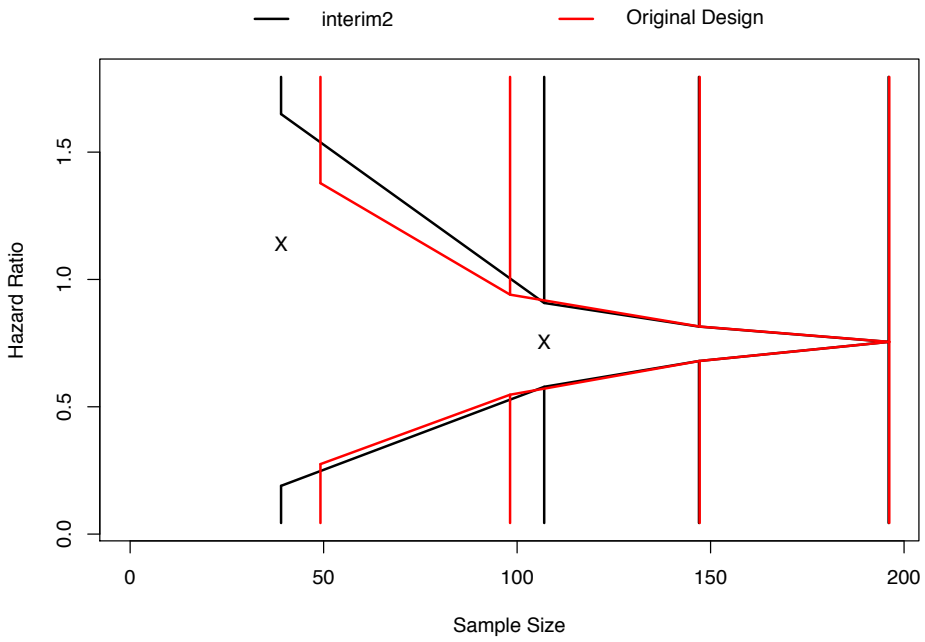
```
RECOMMENDATION:
     Continue

OBSERVED STATISTICS:
 Sample Size Crude Estimate Z Statistic
          39         1.1395      0.4062
         107         0.7571     -1.4233

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.0000    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.6698    (power = 0.7804)

STOPPING BOUNDARIES: Sample Mean scale
                          a      d
    Time 1 (N=  39) 0.1895 1.6495
    Time 2 (N= 107) 0.5784 0.9077
    Time 3 (N= 147) 0.6797 0.8149
    Time 4 (N= 196) 0.7548 0.7548
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Case Study : Hodgkin's Trial

Timing of 2nd analysis

- Plot or monitoring result at 2nd analysis



Case Study : Hodgkin's Trial

Estimate timing for future analyses

- Based upon new pooled event rates, determine the amount of additional followup needed in order to obtain desired events while maintaining accrual of 80 patients per year for 3 years

```
  accrualTime followupTime rate hazardRatio controlMedian nSubjects
1           3     1.753815   80        1.00      1.246134       240
2           3     2.446173   80        0.67      1.246134       240

  analysisTimes.1 analysisTimes.2 analysisTimes.3 analysisTimes.4
1        1.719462        2.599327        3.408330        4.753815
2        1.864917        2.805022        3.751868        5.446173
```

Case Study : Hodgkin's Trial

Timing of 3rd interim analysis

- Monitoring at 3rd interim analysis
  - Based upon previous estimates of pooled survival, next analysis conducted at 3.5 years
  - Specify remaining analysis at intended schedule to (roughly) maintain power (196)
  - Use function `seqMonitor()` to analyze current data and produce constrained boundaries

Case Study : Hodgkin's Trial

SISCR
UW - 2016

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

3rd interim analysis

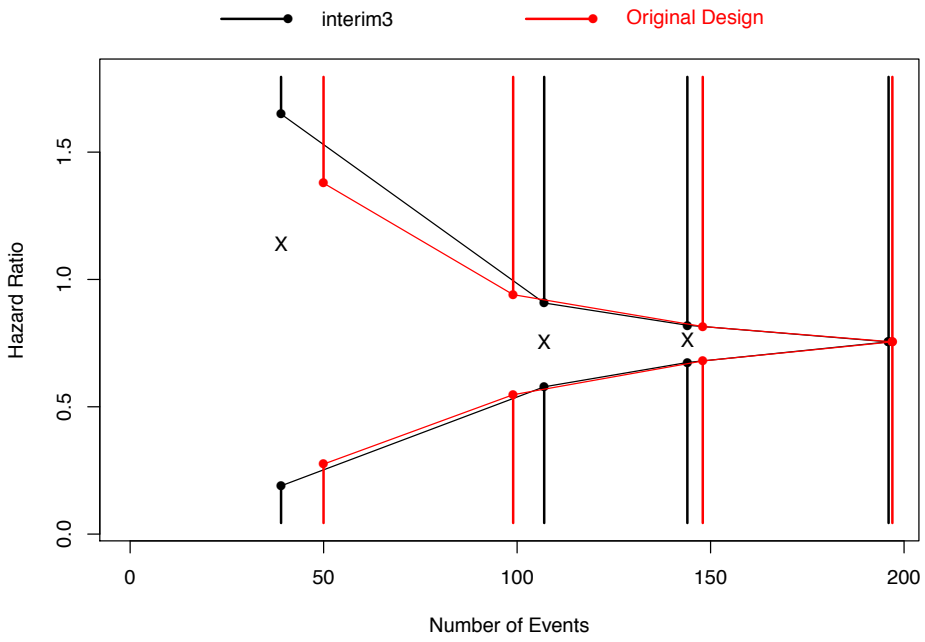
- Result of `seqMonitor()` at 3rd analysis

```
RECOMMENDATION:
     Continue

OBSERVED STATISTICS:
 Sample Size Crude Estimate Z Statistic
          39         1.1395      0.4062
         107         0.7571     -1.4233
         144         0.7648     -1.6044

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.00    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.67    (power = 0.7804)

STOPPING BOUNDARIES: Sample Mean scale
                          a      d
    Time 1 (N=  39) 0.1895 1.6495
    Time 2 (N= 107) 0.5784 0.9077
    Time 3 (N= 144) 0.6739 0.8201
    Time 4 (N= 196) 0.7549 0.7549
```

Case Study : Hodgkin's Trial

Timing of 3rd analysis

- Plot or monitoring result at 3rd analysis



Case Study : Hodgkin's Trial

SISCR
UW - 2016

Estimate timing for future analyses

- Based upon new pooled event rates, determine the amount of additional followup needed in order to obtain desired events while maintaining accrual of 80 patients per year for 3 years

```
  accrualTime followupTime rate hazardRatio controlMedian nSubjects
1           3     1.933717   80        1.00      1.324366       240
2           3     2.673878   80        0.67      1.324366       240

  analysisTimes.1 analysisTimes.2 analysisTimes.3 analysisTimes.4
1        1.764297        2.661064        3.503763        4.933717
2        1.914171        2.873225        3.872611        5.673878
```

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Case Study : Hodgkin's Trial

Timing of final analysis

- Monitoring at final analysis
  - Based upon previous estimates of pooled survival, next analysis conducted at 5 years
  - Omit the `future.analyses` option
  - Use function `seqMonitor()` to analyze final data

Case Study : Hodgkin's Trial

Final analysis

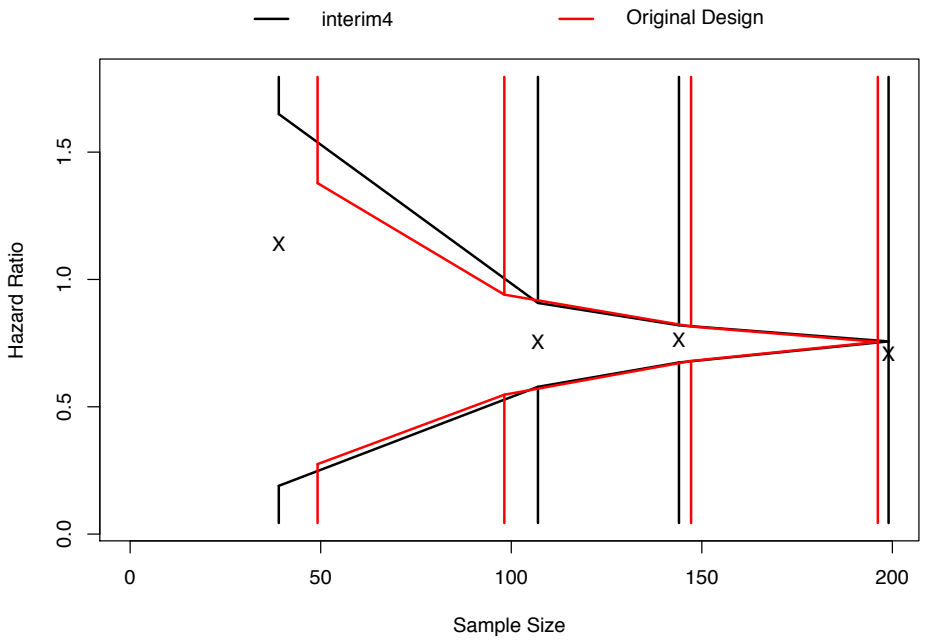
- Result of `seqMonitor()` at final analysis

```
RECOMMENDATION:
     Stop with decision for Lower Alternative Hypothesis

OBSERVED STATISTICS:
 Sample Size Crude Estimate Z Statistic
          39         1.1395      0.4062
         107         0.7571     -1.4233
         144         0.7648     -1.6044
         199         0.7067     -2.4489

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.0000    (size  = 0.0250)
    Alternative hypothesis : Theta <= 0.6714    (power = 0.7804)

STOPPING BOUNDARIES: Sample Mean scale
                        a      d
    Time 1 (N=  39) 0.1895 1.6495
    Time 2 (N= 107) 0.5784 0.9077
    Time 3 (N= 144) 0.6739 0.8201
    Time 4 (N= 199) 0.7567 0.7567
```

Case Study : Hodgkin's Trial

Timing of final analysis

- Plot or monitoring result at final analysis



Case Study : Hodgkin's Trial

Final analysis

- Result of `seqMonitor()` at final analysis

```
INFERENCE:

Adjusted estimates based on observed data:
  analysis.index observed    MLE    BAM RBadj
1              4   0.7067 0.7067 0.7099 0.728

Inferences based on Analysis Time Ordering:
     MUE P-value     **** CI ****
1 0.7166 0.01299 (0.5381, 0.9599)

Inferences based on Mean Ordering:
     MUE P-value     **** CI ****
1 0.7166 0.01299 (0.5381, 0.9599)
```

Estimation of Statistical Information

Design stage vs. implementation stage

- At time of study design
  - Sample size (power, alternative) calculations based on specifying statistical information available from each sampling unit
- During conduct of study
  - Statistical information from a sampling unit may be different than originally estimated
    - Variance of measurements
    - Baseline event rates
    - (Altered sampling distribution for treatment levels)

Estimation of Statistical Information

SISCR
UW - 2016

Computation of sample size

- Sample size formulas used in group sequential test design

$$N = \frac{\delta_1^2 V}{(\Delta_1 - \Delta_0)^2}$$

  - N : maximal number of sampling units
  - δ_1 : alternative for which a standardized form of a level α test has power β
  - $1/V$: statistical information contributed by each sampling unit

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Estimation of Statistical Information

Computation of sample size

- Sample size formulas used in group sequential test design are completely analogous to those used in fixed sample studies

$$N = \frac{\delta_1^2 V}{(\Delta_1 - \Delta_0)^2}$$

  - In a fixed sample two arm test of an (approximately) normal mean we have
    - $\delta_1 = z_{1-\alpha/2} + z_\beta$
    - $V = 2\sigma^2$

Estimation of Statistical Information

SISCR
UW - 2016

Incorrect estimates of information at design stage

- Effect of using incorrect estimates of statistical information at the design stage
  - Using the specified sample size, the design alternative will not be detected with the desired power
  - Using the specified sample size, the alternative detected with the desired power will not be the design alternative
  - In order to detect the design alternative with the desired power, a different sample size is needed

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Estimation of Statistical Information

Maintaining maximal sample size or power

- If maximal sample size is maintained, the study discriminates between null hypothesis and an alternative measured in units of statistical information

$$N = \frac{\delta_1^2 V}{(\Delta_1 - \Delta_0)^2} = \frac{\delta_1^2}{\left(\frac{(\Delta_1 - \Delta_0)^2}{V}\right)}$$

- If statistical power is maintained, the study sample size is measured in units of statistical information

$$\frac{N}{V} = \frac{\delta_1^2}{(\Delta_1 - \Delta_0)^2}$$

Estimation of Statistical Information

SISCR
UW - 2016

Measuring study time

- Flexible methods compute boundaries at an interim analysis according to study time at that analysis
- Study time can be measured by
  - Proportion of planned number of subjects accrued (maintains maximal sample size)
  - Proportion of planned statistical information accrued (maintains statistical power)
  - (Calendar time– not really advised)

Estimation of Statistical Information

Measuring study time

- In either case, we must decide how we will deal with estimates of statistical information at each analysis when constraining boundaries
- Statistical information in clinical trials typically has two parts
  - V = variability associated with a single sampling unit
  - The distribution of sampled levels of treatment
- In many clinical trials, the dependence on the distribution of treatment levels across analyses is only on the sample size N

Estimation of Statistical Information

SISCR
UW - 2016

Possible approaches

- At each analysis estimate the statistical information available, and use that estimate at all future analyses
  - Theoretically, this can result in estimates of negative information gained between analyses
- At each analysis use the sample size with the current best estimate of V
  - The 1:1 correspondence between boundary scales (see Session 2) is broken at previously conducted analyses

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Estimation of Statistical Information

Possible approaches

- In `RCTdesign`, all probability models have statistical information directly proportional to sample size for block randomized experiments, thus we chose to update V at all analyses using the current best estimate
- Other statistical packages (PEST, EaSt) constrain boundaries using the estimate of statistical information available at the previous analyses.
- There is no clear best approach

Estimation of Statistical Information

SISCR
UW - 2016

Possible approaches

- Overall, I think it makes more sense to use the best estimate of the variance of an observation when estimating a sampling distribution.
- This avoids the possibility of negative information, but allows the conflicting results described above.

Impact of Changing the Number and Timing of Analyses
Background
Example : Constrained OBF design
Flexible Trial Monitoring
Error Spending Functions
Constrained Boundaries
Case Study: Monitoring of Hodgkin's Trial
Issues When Monitoring a Trial
Estimation of statistical information
Measuring study time

Sequential and Adaptive Analysis with Time-to-Event Endpoints

Session 4 - Time-Varying Treatment Effects

Presented July 29, 2016

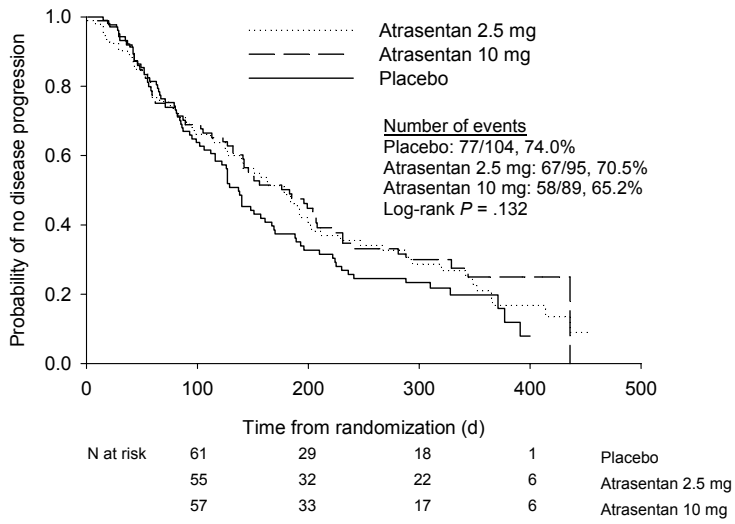
Scott S. Emerson
Department of Biostatistics
University of Washington

Daniel L. Gillen
Department of Statistics
University of California, Irvine

Motivating example

Atrasentan for the treatment of hormone-refractory prostate cancer

- Phase II results for time to progression of disease



SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Motivating example

Atrasentan for the treatment of hormone-refractory prostate cancer

- From the ODAC briefing document:

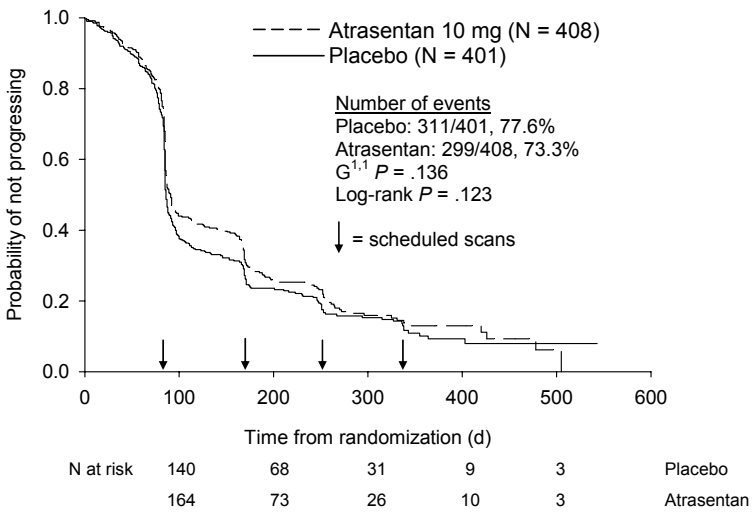
  "In study M96-594, an exploratory analysis of time to disease progression had been performed using the $G^{1,1}$ test statistic, a variant of the log-rank test described by Fleming et al. The $G^{1,1}$ test statistic reduces the weight given to events that occur very early or very late in time-to-progression distributions. This statistic was chosen due to the shape of the disease progression curve (greatest separation between treatment at the median) as observed in study M96-594."

Motivating example

SISCR
UW - 2016

Atrasentan for the treatment of hormone-refractory prostate cancer

- Phase III results for time to progression of disease



Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Motivating example

Atrasentan for the treatment of hormone-refractory prostate cancer

- From the ODAC briefing document (next paragraph):

  "Based on the anticipation that the time to disease progression curve would be similar in study M00-211, the $G^{1,1}$ statistic was the protocol-specified primary analysis for the endpoint of time to disease progression. Unfortunately, the impact of the protocol-defined 12-week scheduling of radiographic scans resulted in approximately 50% of patients completing the study at the time of their first scan (around 12 weeks). Thus, in retrospect, the $G^{1,1}$ statistic was no longer optimal and the median statistic is not a good indicator of the treatment effect of atrasentan. To present results in a more clinically relevant fashion, Cox proportional hazards modeling, which describes the relative risk across the entire distribution of events, was used."

Motivating example

Atrasentan for the treatment of hormone-refractory prostate cancer

- A few take-home messages:

  1. "Past performance may not be indicative of future results"
     -Any TV channel randomly selected at 3am
  2. The choice of summary measure has great impact and should be chosen based upon (in order of importance):
     - Most clinically relevant summary measure
     - Summary measure most likely to be affected by the intervention
     - Summary measure affording the greatest statistical precision
  3. Outside of an assumed semi-parametric framework, the censoring (accrual) distribution plays a key role in the estimation of effects on survival

The logrank statistic

Notation

- The logrank statistic is given by

$$LR = \left(\frac{M_1 + M_0}{M_1 M_0}\right)^{1/2} \int_0^\infty \left\{\frac{Y_1(t)Y_0(t)}{Y_1(t)+Y_0(t)}\right\}\left\{\frac{dN_1(t)}{Y_1(t)} - \frac{dN_0(t)}{Y_0(t)}\right\}$$

with

M_i = number of subjects initially at risk in group i, $i = 01$
$Y_i(t)$ = number of subjects at risk in group i at time t
$N_i(t)$ = the counting process for group i at time t

The logrank statistic

SISCR
UW - 2016

The logrank statistic

- The logrank statistic can be rewritten as the sum, over all failure times, of the weighted difference in estimated hazards

$$LR = \left(\frac{M_1 + M_0}{M_1 M_0}\right)^{1/2} \sum_{t\in\mathcal{F}} w(t)\left[\hat{\lambda}_1(t) - \hat{\lambda}_0(t)\right]$$

with $\hat{\lambda}_i = dN_i(t)/Y_i(t)$ and $w(t) = \frac{Y_1(t)Y_0(t)}{Y_1(t)+Y_0(t)}$

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

The logrank statistic

- Weights are determined by the number of subjects at risk at each failure time
- Number of subjects at risk is determined by:
  - Number initially at risk
  - The censoring distribution (accrual and dropout distributions)
  - The survival distribution

$$Y_i(t) = M_i \times S_i(t) \times (1 - F_C(t))$$

with S_i the survival distribution of group i and F_C the cdf of the censoring distribution (potentially group-specific)

The logrank statistic

The logrank statistic

- Under proportional hazards
  - Terms composing the logrank statistic are roughly constant (in a neighborhood of the null hypothesis of equal hazards)
- Under nonproportional hazards
  - Differences in hazards (likely to) change with time
  - As the weights change, what we are estimating/testing changes
  - As the censoring distribution changes, what we are estimating/testing changes
  - Need to consider sensitivity to the accrual/dropout distribution

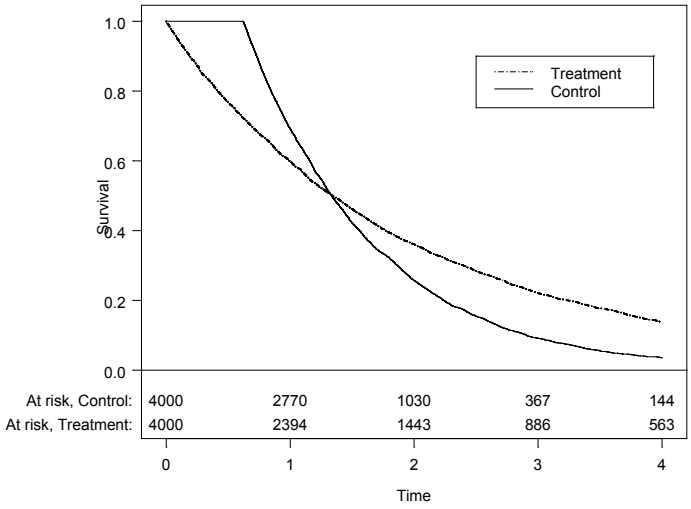
SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

The logrank statistic

Example 1: Sensitivity to the censoring distribution

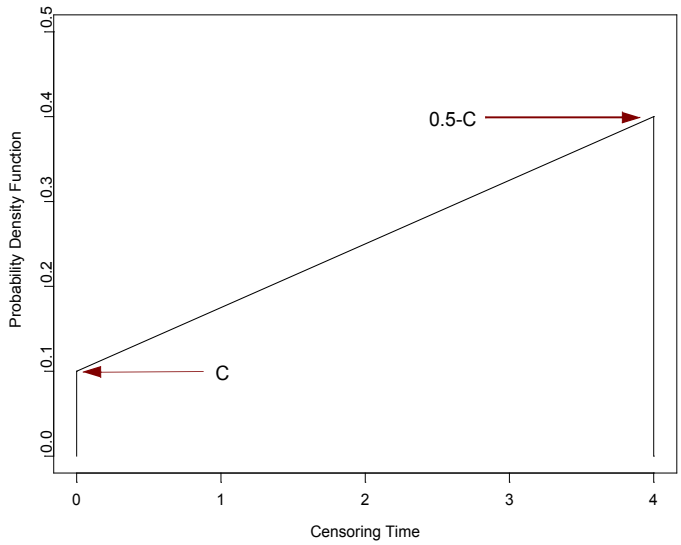
- Grossly exaggerated depiction of a non-proportional hazards treatment effect in the absence of censoring



The logrank statistic

Example 1: Sensitivity to the censoring distribution

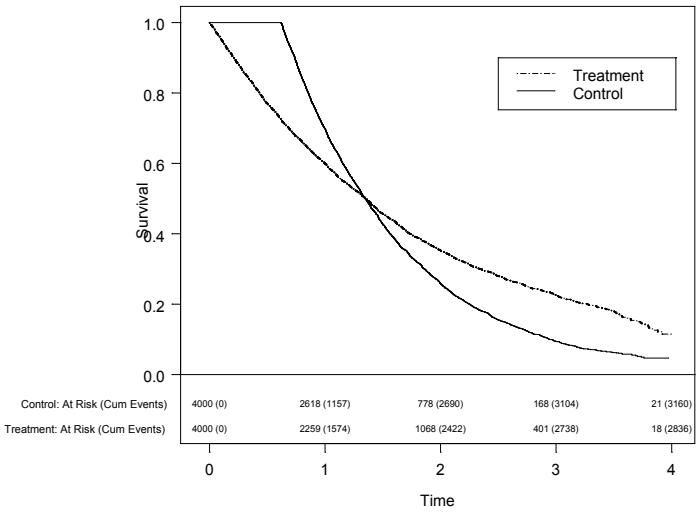
- Simple example of parametric censoring distribution
  - $C = 0 \Rightarrow$ Heavy early accrual
  - $C = 0.25 \Rightarrow$ Uniform accrual
  - $C = 0.5 \Rightarrow$ Slow early accrual



The logrank statistic

Example 1: Sensitivity to the censoring distribution

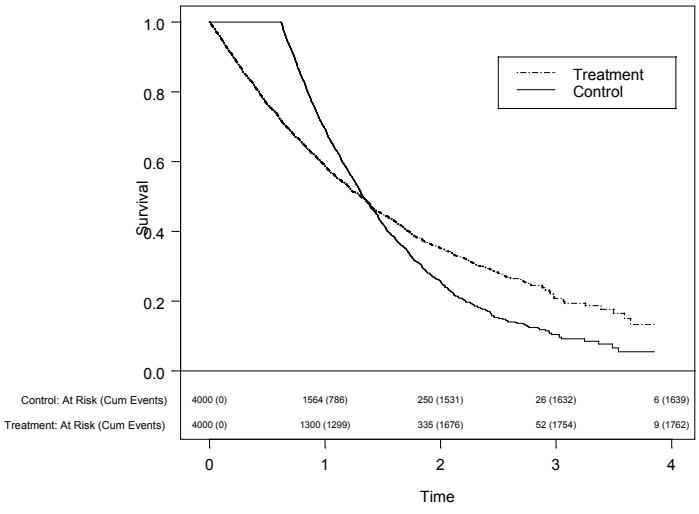
- Estimated survival curves when $C = 0$ (heavy early accrual)



The logrank statistic

Example 1: Sensitivity to the censoring distribution

- Estimated survival curves when $C = 0.5$ (slow early accrual)



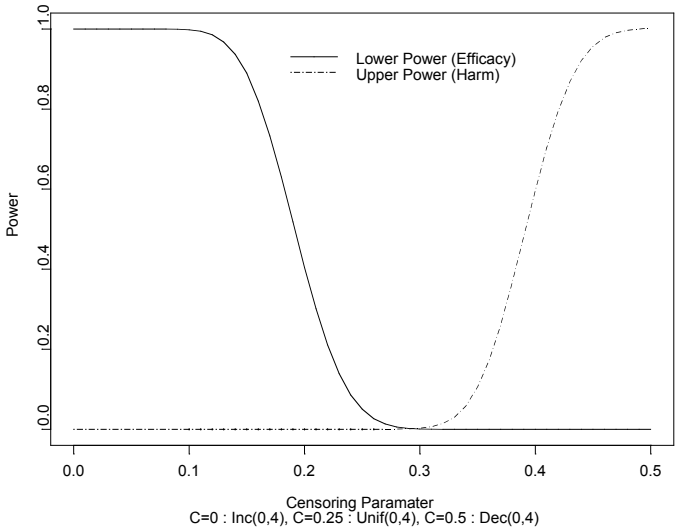
SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

The logrank statistic

Example 1: Sensitivity to the censoring distribution

- Upper (harm) and lower (efficacy) power as a function of C



The logrank statistic

Example 2: Sensitivity to the censoring distribution

- Consider the Hodgkin's trial
  - Suppose that there was a delayed treatment effect
    - No change in survival over the first year
    - Hazard ratio of 0.4 after first year
    - (Subset of sickest patients that could not be helped)
  - What would we estimate if we uniformly accrued
    - 40 patients per year for 6 years?
    - 80 patients per year for 3 years?
    - 1000 patients for 1 month?

SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Example 2: Sensitivity to the censoring distribution

- Sample size chosen to provide desired operating characteristics
  - Type I error : 0.025 when no difference in mortality
  - Power : 0.80 when 33% reduction in hazard
- Expected number of events determined by assuming
  - Exponential survival in placebo group with median survival of 9 months
  - Uniform accrual of patients over 3 years
  - Negligible dropout

Example 2: Sensitivity to the censoring distribution

- General sample size formula:
  - δ = standardized alternative
  - Δ = log-hazard ratio
  - π_i = proporiton of patients in group i, $i = 0, 1$
  - D = number of sampling units (events)

$$D = \frac{\delta^2}{\pi_0 \pi_1 \Delta^2}$$

The logrank statistic

Example 2: Sensitivity to the censoring distribution

- Fixed sample test (no interim analyses):
  - $\delta = (z_{1-\alpha} + z_{\beta})$ for size α and power β
- For current study, we assume 1:1 randomization
  - $\pi_0 = \pi_1 = 0.5$
- Number of events for planned trial:

$$D = \frac{(1.96 + 0.84)^2}{0.5^2 \times [\log(.67)]^2] } = 195.75$$

The logrank statistic

SISCR
UW - 2016

Example 2: Sensitivity to the censoring distribution

- In general, it necessary to know the expected number of patients required to obtain the desired operating characteristics
- This is given by:

$$N = \frac{D}{\pi_0 \Pr_0[\text{Event}] + \pi_1 \Pr_1[\text{Event}]}$$

where D is the total number of required events and π_i is the proportion of patients allocated to group i

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Example 2: Sensitivity to the censoring distribution

- Under proportional hazards, Pr[Event] for each group depends upon
  1. The total followup (T_L) and accrual (T_A) time
  2. The underlying survival distribution
  3. The accrual distribution
  4. Drop-out

Example 2: Sensitivity to the censoring distribution

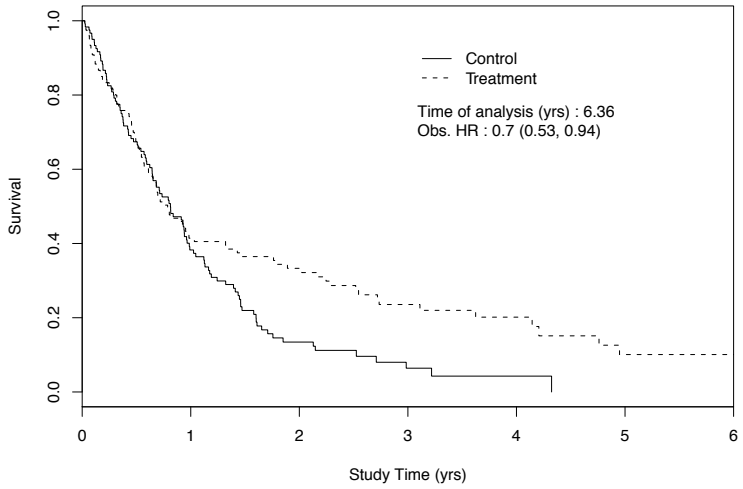
- From the above, if we assume a uniform accrual pattern we have:

$$\begin{aligned}
\Pr[\text{Event}] &= \int_0^{T_A} \Pr[\text{Event \& Entry at } t]dt \\
&= \int_0^{T_A} \Pr[\text{Event} \mid \text{Entry at } t] \Pr[\text{Entry at } t]dt \\
&= 1 - \int_0^{T_A} \Pr[\text{No Event} \mid \text{Entry at } t] \Pr[\text{Entry at } t]dt \\
&= 1 - \int_0^{T_A} S(T_L - t) f_E(t)dt
\end{aligned}$$

The logrank statistic

Example 2: Sensitivity to the censoring distribution

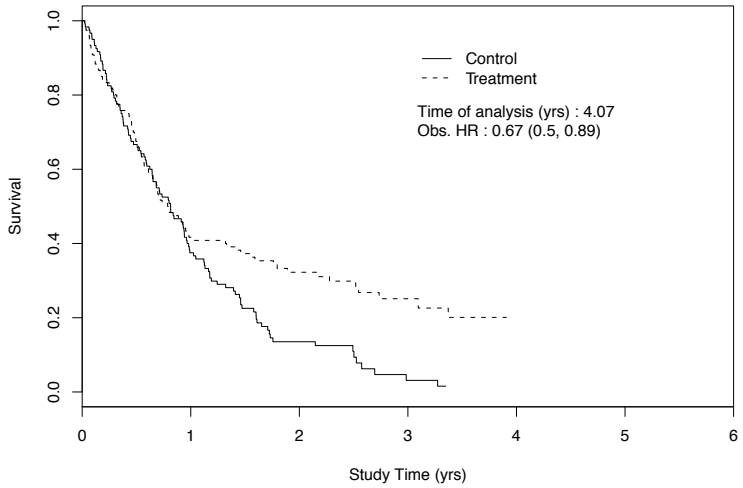
- Accrual of 40 patients per year for 6 years
  - 196th event occurs at 6.36 yrs after first enrollment
  - HR estimate of 0.70 (0.53,0.94)



The logrank statistic

Example 2: Sensitivity to the censoring distribution

- Accrual of 80 patients per year for 3 years
  - 196th event occurs at 4.07 yrs after first enrollment
  - HR estimate of 0.67 (0.50,0.89)



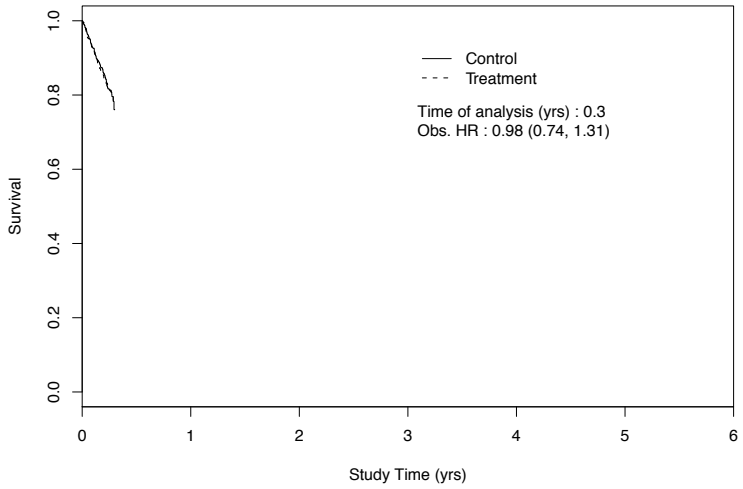
SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

The logrank statistic

Example 2: Sensitivity to the censoring distribution

- Accrual of 1000 patients for 1 month
  - 196th event occurs at 0.3 yrs after first enrollment
  - HR estimate of 0.98 (0.74,1.31)



The logrank statistic

Sensitivity to the censoring distribution

- Bottom line
  - Under a hypothesized nonproportional hazards alternative, need to assess sensitivity to the censoring (accrual and dropout) distribution
  - Consider the usual operating characteristics under variations
    - Sample size
    - Power curve
    - Estimates corresponding to boundary decisions (HR?)
  - Need to ask whether the hazard ratio is the best functional to test
    - Alternatives?

SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

The logrank statistic

Sensitivity to the censoring distribution

- Problem gets even more difficult when moving to group sequential testing
  - Interim analyses truncate the length of observed support
  - Analyses are scheduled based upon the number of observed events
    - Number of events is partially determined by accrual rate
    - Faster/slower accrual implies shorter/longer support
    - If hazard ratio is changing with time, what will be tested at each analysis?

Weighted LR statistics

SISCR
UW - 2016

$G^{\rho,\gamma}$ statistic

- When a non-proportional hazards treatment effect is hypothesized some have suggested the use of weighted logrank statistics
  - Potential for increased power by up-weighting areas of survival where largest (most clinically relevant?) effects are hypothesized to occur
- $G^{\rho,\gamma}$ family of weighted logrank statistics (Fleming & Harrington, 1991)

$$G^{\rho,\gamma} = \left(\frac{M_1 + M_0}{M_1 M_0}\right)^{1/2} \int_0^\infty w(t) \left\{\frac{Y_1(t)Y_0(t)}{Y_1(t)+Y_0(t)}\right\} \left\{\frac{dN_1(t)}{Y_1(t)} - \frac{dN_0(t)}{Y_0(t)}\right\}$$

with

$$w(t) = [\hat{S}(t-)]^\rho [1 - \hat{S}(t-)]^\gamma$$

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Weighted LR statistics

$G^{\rho,\gamma}$ statistic

- Can be rewritten as the sum, over all failure times, of the weighted difference in estimated hazards

$$G^{\rho,\gamma} = \left(\frac{M_1 + M_0}{M_1 M_0}\right)^{1/2} \sum_{t \in \mathcal{F}} w^*(t) \left[\hat{\lambda}_1(t) - \hat{\lambda}_0(t)\right]$$

with $\hat{\lambda}_i = dN_i(t)/Y_i(t)$ and

$$w^*(t) = \left\{\frac{Y_1(t)Y_0(t)}{Y_1(t) + Y_0(t)}\right\} [\hat{S}(t-)]^{\rho}[1 - \hat{S}(t-)]^{\gamma}$$

Evaluation of designs when testing with a WLR statistic

SISCR
UW - 2016

seqOCWLR()

- `seqOCWLR()` uses simulation to evaluate the operating characteristics of potential designs when a $G^{\rho,\gamma}$ statistic is used for testing survival effects
  - Relies upon user-inputted pilot data
  - Simulates alternatives in a non-parametric fashion
  - Considers sensitivity of other relevant summary statistics when testing based upon a WLR statistic

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from seqOCWLR()
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Evaluation of designs when testing with a WLR statistic

Definition of null survival distribution

- `seqOCWLR()` simulates alternatives by resampling repeatedly from a single set of Kaplan-Meier estimates of survival curves arising from user-supplied pilot data
- Two reasonable choices for the null survival distribution:
  1. 50-50 mixture of the estimated survival experience of the control and treatment samples from the pilot study
  2. control sample alone

Evaluation of designs when testing with a WLR statistic

Definition of alternatives

- Given the existence of pilot data, one natural alternative to the chosen null distribution is the observed survival experience of the comparison group
- Need to consider a variety of alternatives for evaluating operating characteristics, but outside of a parametric/semi-parametric model
- In `seqOCWLR()` we consider mixtures of the control and comparison Kaplan-Meier estimates of survival from the pilot data
  - 0% mixing : indicates no treatment effect on survival
  - 50% mixing : indicates a treatment effect where treated group represents a 50-50 mixture of the control and comparison survival experience from the pilot data
  - 100% mixing : corresponds to a treatment effect that results in a survival experience that is equivalent to that of the comparison sample in the pilot study

SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Evaluation of designs when testing with a WLR statistic

Algorithm for simulating operating characteristics

1. Compute the Kaplan-Meier estimate of the survival distribution for the control and treatment groups in the pilot study, $\hat{S}_0$ and $\hat{S}_1$, respectively.

2. Define the alternative via the percentage that the control and treatment groups are to be mixed, $0 \leq m \leq 1$.

3. For $i = 0, 1$ do

   3.1 Let $N_i = \text{ceiling}(N * |(1 - i) - m|)$.
   3.2 Sample N_i survival times $\vec{t}_i = (t_1^*, t_2^*, ..., t_{N_i}^*)$ with replacement from $(t_{1i}, t_{2i}, ..., t_{n_i i}, \infty)$ with probability $(1 - \hat{S}_i(t_{1i}), \hat{S}_i(t_{1i}) - \hat{S}_i(t_{2i}), ..., \hat{S}_i(t_{n_i i}) - 0)$.
   3.3 For $j = 1, ..., N_i$, if $t_j^* = \infty$ set $\delta_j = 0$, otherwise set $\delta_j = 1$.

4. Combine the sampled survival times $\vec{t} = (\vec{t}_0, \vec{t}_1)$ and event indicators $\vec{\delta} = (\vec{\delta}_0, \vec{\delta}_1)$.

SISCR
UW - 2016

Output from `seqOCWLR()`

Output from `seqOCWLR()`

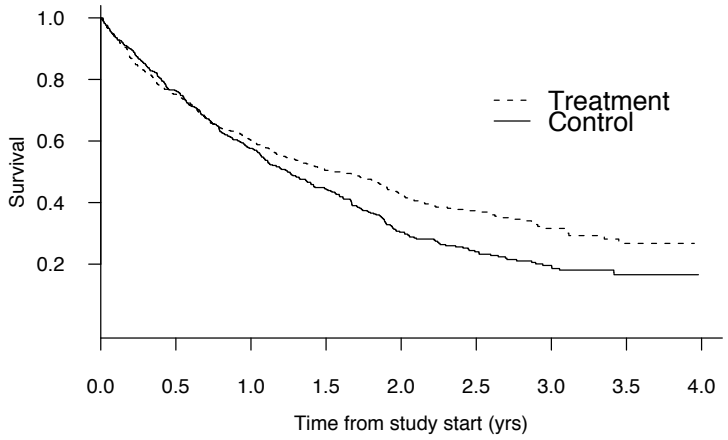
- `seqOCWLR()` produces similar operating characteristics as `seqOC()`
  - Point estimates on the boundary (min/max estimates for Cox estimate and others)
  - ASN
  - Power / Relative Power
  - Stopping probabilities
- All operating characteristics are reported as a function of mixings from the supplied pilot data

SISCR
UW - 2016

Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

- Example pilot data exhibiting a late-occurring treatment effect



Treatment	500 (0)	289 (212)	100 (323)	40 (351)	1 (356)
Control	500 (0)	302 (199)	142 (273)	47 (299)	1 (304)
Total	1000 (0)	591 (411)	242 (596)	87 (650)	2 (660)

Output from `seqOCWLR()`

Designs to consider

- `DSN1`: A one-sided level .025 Pocock stopping rule (corresponding to $P = .5$, $R = 0$, and $A = 0$) on both the lower (efficacy) and upper (futility) boundaries
- `DSN2`: A one-sided level .025 test utilizing the O'Brien-Fleming stopping rule (corresponding to $P = 1$, $R = 0$, and $A = 0$) on both the lower (efficacy) and upper (futility) boundaries
- `DSN3`: A one-sided level .025 test parameterized using an O'Brien-Fleming lower (efficacy) boundary corresponding to $P = 1.0$, $R = 0$, and $A = 0$, and an upper (futility) boundary corresponding to $P = 1.5$, $R = 0$, and $A = 0$
- `DSN4`: A one-sided level .025 test with lower (efficacy) boundary takes $P = 1.2$, $R = 0$, and $A = 0$ and upper (futility) boundary $P = 0$, $R = 0.5$, and $A = 0.3$

Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

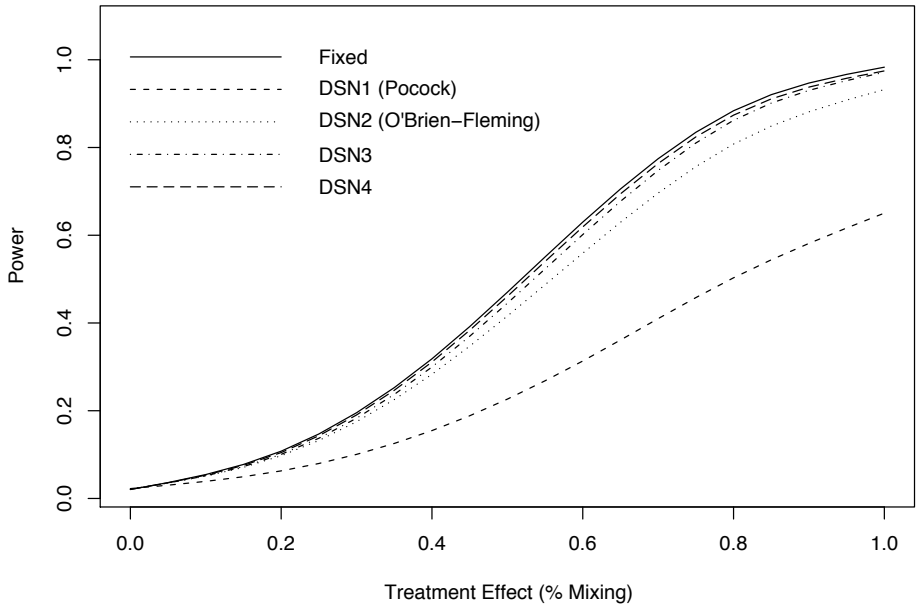
- Potential point estimates that could be observed on the boundary of a symmetric O'Brien-Fleming design `(DSN1)`

Summary Statistic	Efficacy (Min Effect)	Futility (Max Effect)
Analysis 1 ($\Pi_1 = .229$)		
Z statistic	-4.176	2.263
Hazard rato	–	1.009
Trimmed hazard ratio	–	0.873
Analysis 2 ($\Pi_2 = .510$)		
Z statistic	-2.797	-0.058
Hazard rato	0.930	0.856
Trimmed hazard ratio	0.872	0.718
Analysis 3 ($\Pi_3 = .687$)		
Z statistic	-2.411	-0.902
Hazard rato	0.969	0.817
Trimmed hazard ratio	0.904	0.734
Analysis 4 ($\Pi_4 = 1.00$)		
Z statistic	-1.998	-1.998
Hazard rato	0.988	0.801
Trimmed hazard ratio	0.929	0.708

Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

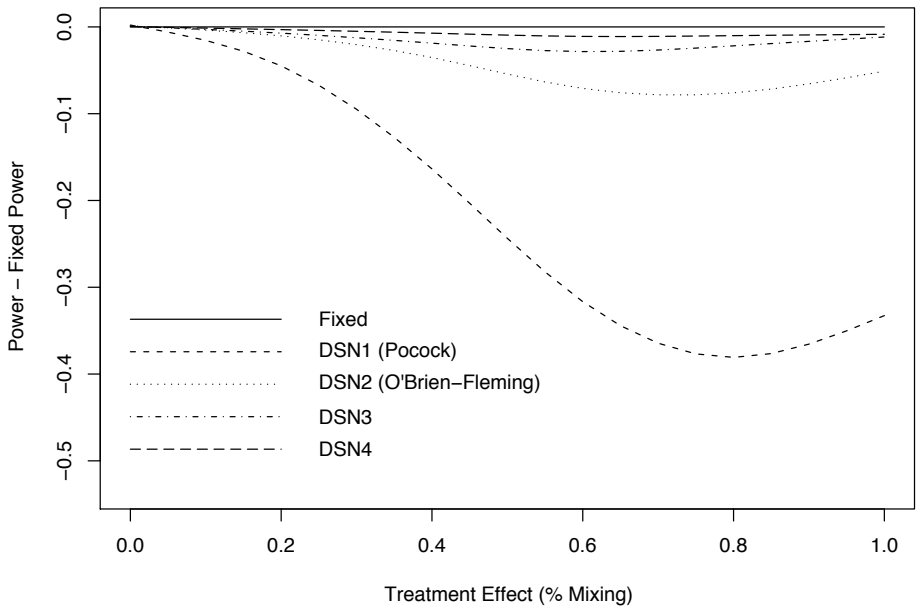
- Power as a function of % mixing



Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

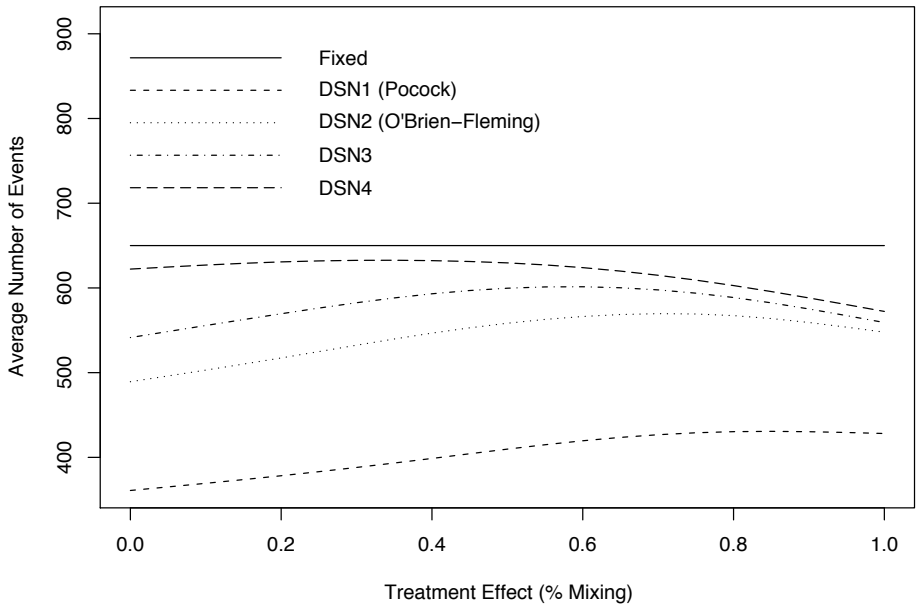
- Relative power as a function of % mixing



Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

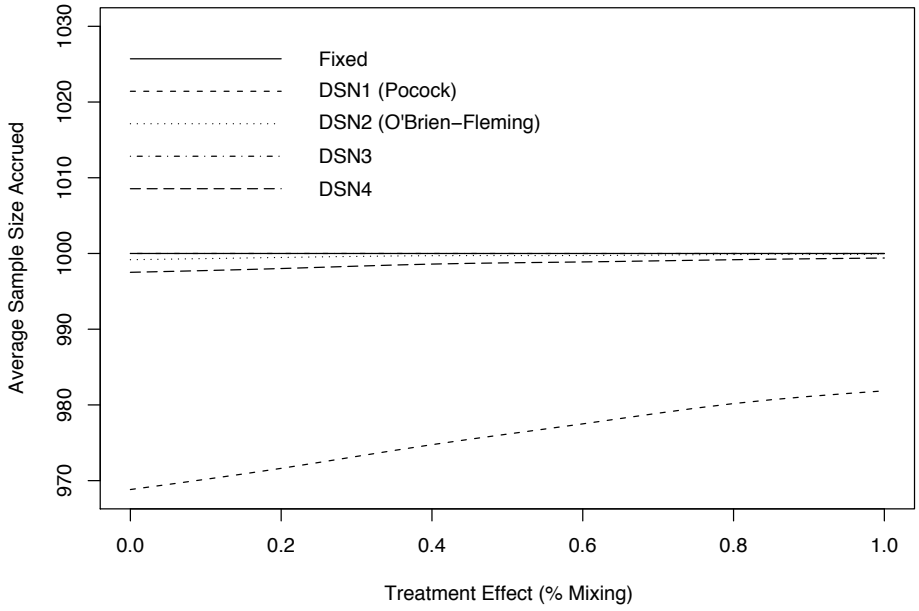
- Average number of events required as a function of % mixing



Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

- Average number of patients required as a function of % mixing

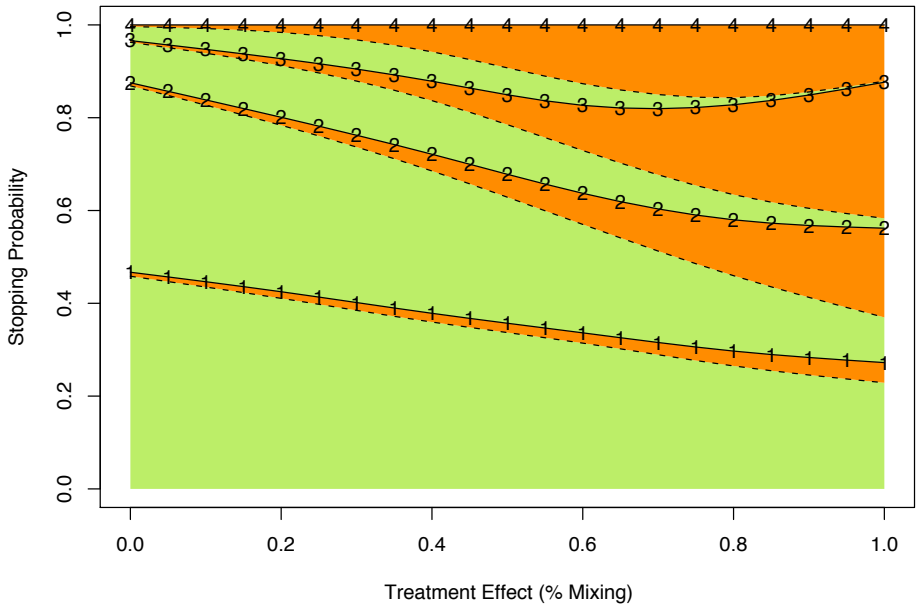


SISCR
UW - 2016

Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

- Stopping probabilities as a function of % mixing for `DSN1` (Pocock)

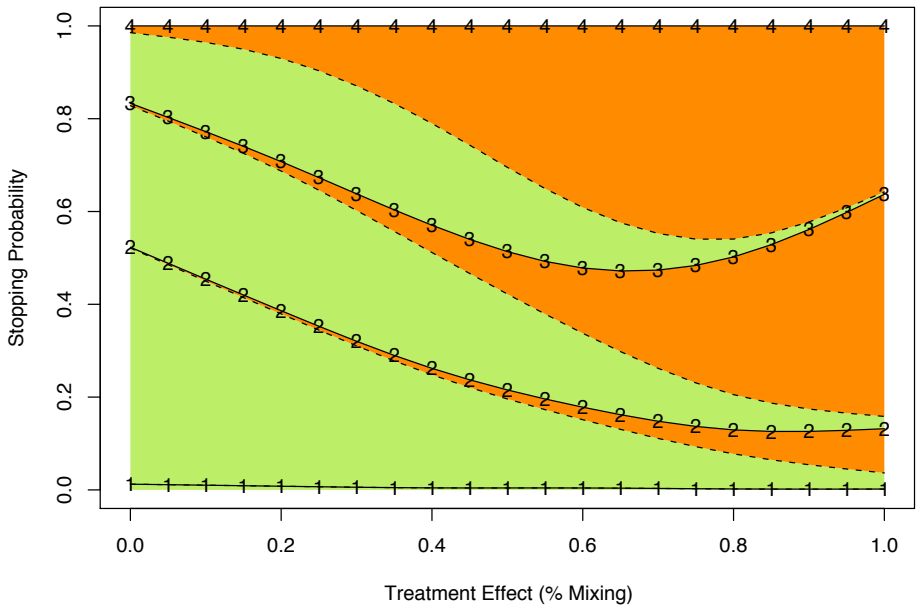


SISCR
UW - 2016

Output from `seqOCWLR()`

Operating characteristics under the $G^{1,1}$ statistic

- Stopping probabilities as a function of % mixing for `DSN2` (OBF)



SISCR
UW - 2016

Monitoring group sequential trials

Popular methods for flexible implementation of group sequential boundaries

1. Christmas tree approximation for triangular tests: Whitehead and Stratton (1983)
2. Error spending functions: Lan and DeMets (1983); Pampallona, Tsiatis, and Kim (1995)
3. Constrained boundaries in unified design family: Emerson (2000); Burrington & Emerson (2003)

2 and 3 implemented in RCTdesign via `seqMonitor()`

SISCR
UW - 2016

Monitoring group sequential trials

Common features

- Stopping rule specified at design stage parameterizes the boundary for some statistic (boundary scale)
  - Error spending family (Lan & Demets, 1983) $\rightarrow$ proportion of type I error spent
  - Unified family (Emerson & Kittelson, 1999) $\rightarrow$ point estimate (MLE)
- At the first interim analysis, parametric form is used to compute the boundary for actual time on study
- At successive analyses, the boundaries are recomputed accounting for the exact boundaries used at previously conducted analyses
- Maximal sample size estimates may be updated to maintain power

Monitoring group sequential trials

Use of constrained boundaries in flexible implementation of stopping rules

1. At the first analysis, compute stopping boundary (on some scale) from parametric family
2. At successive analyses, use parametric family with constraints (on some scale) for the previously conducted interim analyses

- When the error spending scale is used, this is just the error spending approach of Lan & DeMets (1983) or Pampallona, Tsiatis, & Kim (1995)

Group sequential testing in survival trials

Further considerations when considering survival endpoints

- Common to use the logrank statistic for testing survival differences
  - Locally efficient for proportional hazards alternatives
- In this case, translation between sample size and statistical information is trivial
  - Information is proportional to the number of observed events

Information growth for the $G^{\rho,\gamma}$ family

Information growth for the $G^{\rho,\gamma}$ family

- Under the null hypothesis $H_0 : S_0 = S_1$, the variance of the $G^{\rho,\gamma}$ statistic calculated at calendar time τ reduces to

$$\sigma^2 \propto \int_0^\tau w^2(t) F_E(\tau - t)[1 - F_C(t)]dS(t)$$

- Let σ_j^2 equal the estimated variance of the $G^{\rho,\gamma}$ statistic applied at interim analysis j. Then the proportion of information at analysis j, relative to the maximal analysis J, is given by

$$\prod_j \equiv \left(\frac{M_{1,j} + M_{0,j}}{M_{1,j} M_{0,j}}\right)^{-1} \sigma_j^2 \Bigg/ \left(\frac{M_{1,J} + M_{0,J}}{M_{1,J} M_{0,J}}\right)^{-1} \sigma_J^2,$$

Information growth for the $G^{\rho,\gamma}$ family

Example: Information Growth for the $G^{1,0}$ and $G^{1,1}$ statistics

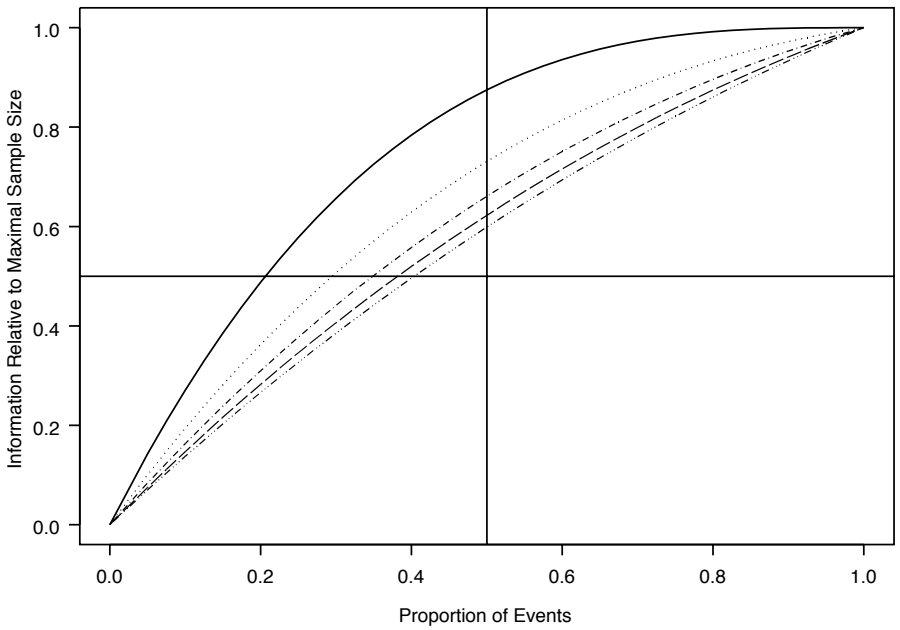
- Consider information growth for the $G^{1,0}$ and $G^{1,1}$ statistics as a function of observed events
- Assume
  - $S_1(t)$ and $S_0(t)$ are Exponential(1)
  - Assume accrual follows a “powered uniform" distribution

$$F_E(t) = \left(\frac{t}{\theta}\right)^r, \text{ with } \theta > 0, r > 0, 0 < t \leq \theta$$

  - Enrollment occurs over interval $(0, \theta)$
  - $r = 1 \Rightarrow$ Unif(0,θ) enrollment
  - $r \to 0 \Rightarrow$ Instantaneous enrollment at time 0
  - $r \to \infty \Rightarrow$ Instantaneous enrollment at time θ

Example: Difference in Information by Accrual for the $G^{1,0}$ Statistic

Effect of total censoring: No censoring (solid line) to 66% censoring

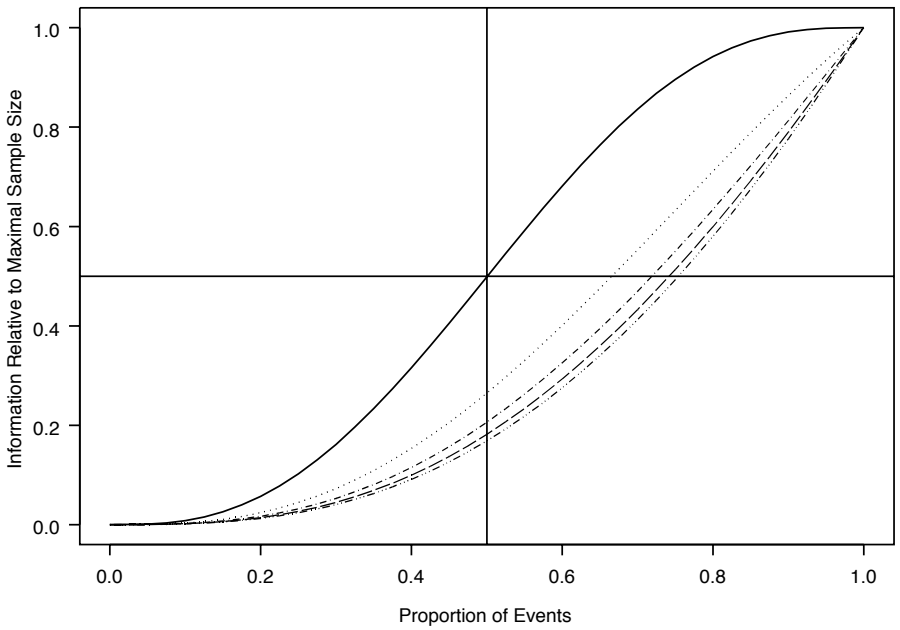


SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Example: Difference in Information by Accrual for the $G^{1,1}$ Statistic

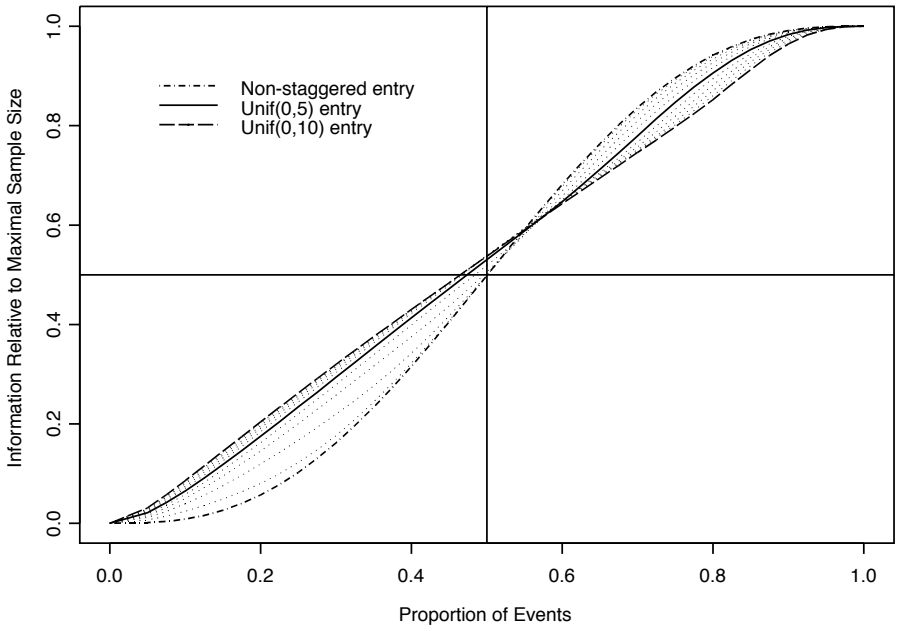
Effect of total censoring: No censoring (solid line) to 66% censoring



SISCR
UW - 2016

Example: Information Growth for the $G^{1,1}$ Statistic

Uniform accrual with no administrative censoring

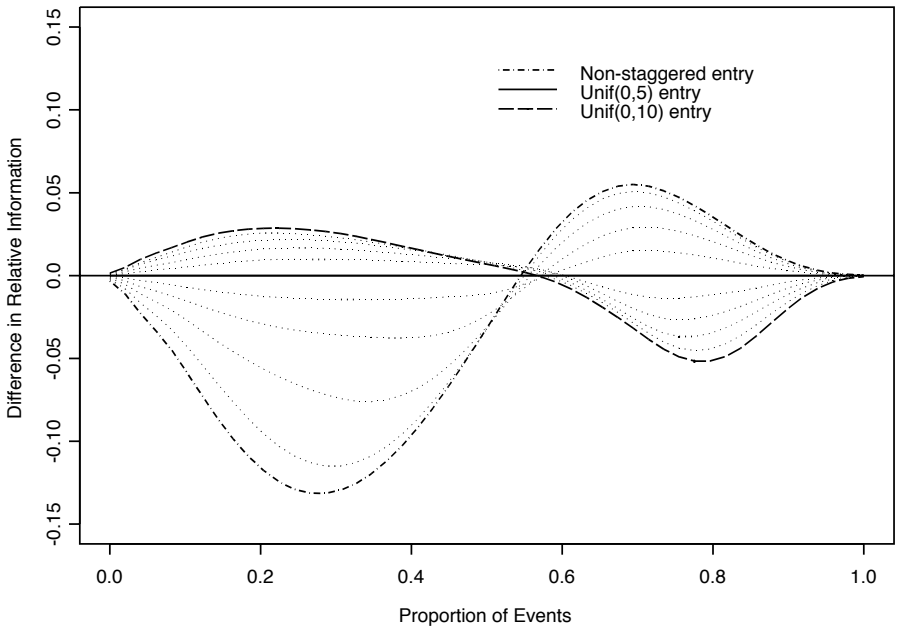


SISCR
UW - 2016

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

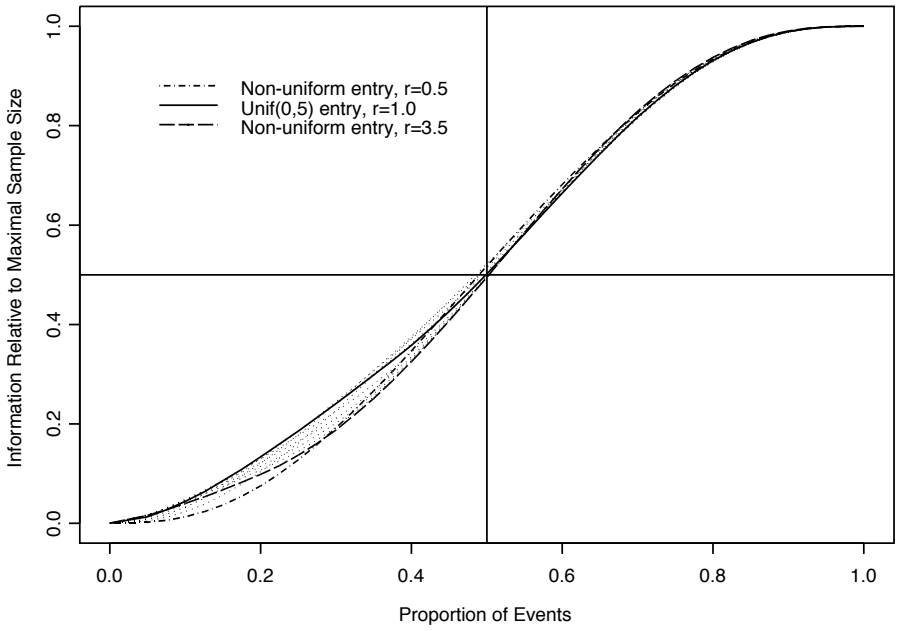
Example: Difference in Information by Accrual for the $G^{1,1}$ Statistic

Uniform accrual with no administrative censoring



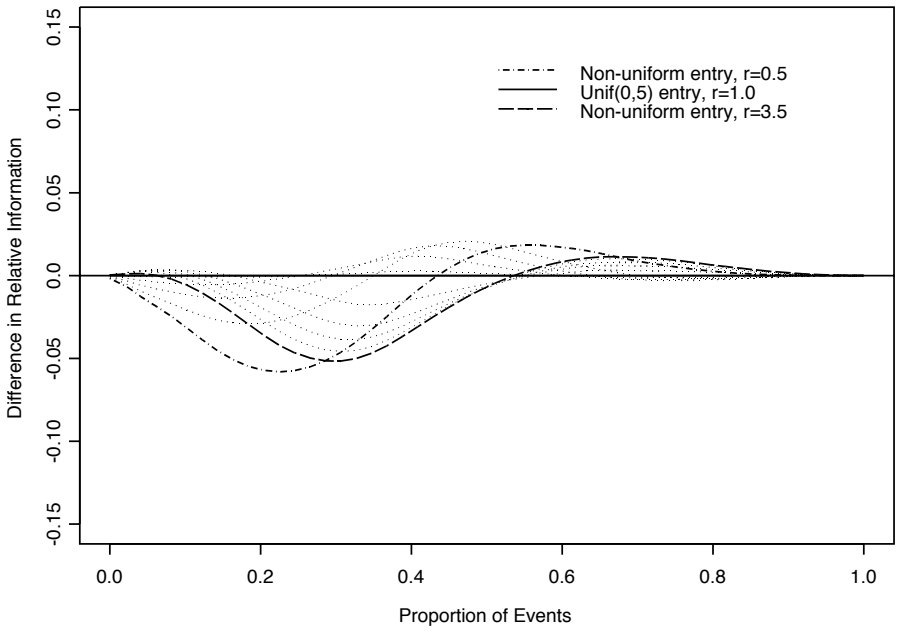
Example: Information Growth for the $G^{1,1}$ Statistic

Nonuniform accrual with no administrative censoring



Example: Difference in Information by Accrual for the $G^{1,1}$ Statistic

Nonuniform accrual with no administrative censoring



Example: Operating characteristics with misspecified accrual distribution

Example: Operating characteristics when testing with the $G^{1,1}$ Statistic

- Design
  - One-sided level .05 test
  - O'Brien-Fleming efficacy bound; Pocock futility bound
  - 4 analyses occurring at proportional information of .25, .50, .75, and 1
  - Power of .90 at alternative HR of .75 $\rightarrow$ 507 max events
- Assumed survival and accrual distributions
  - Pooled survival distributed Exponential(.4)
  - Accrual uniform over 3 years
- Suppose true accrual is uniform over 1 year

SISCR
UW - 2016

Example: Operating characteristics with misspecified accrual distribution

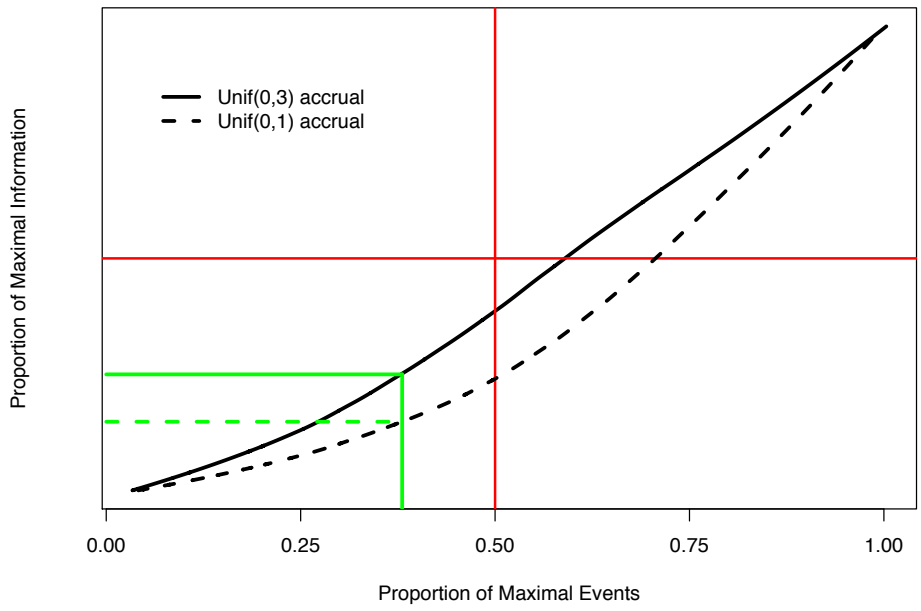
Example: Operating characteristics when testing with the $G^{1,1}$ Statistic

- Stopping boundaries for original design on Z-statistic scale

```
STOPPING BOUNDARIES: Normalized Z-value scale
                          efficacy  futility
    Time 1 (Pi_1= 0.25) -3.2642    0.2094
    Time 2 (Pi_2= 0.50) -2.3082   -0.5534
    Time 3 (Pi_3= 0.75) -1.8846   -1.1387
    Time 4 (Pi_4= 1.00) -1.6321   -1.6321
```


Example: Operating characteristics with misspecified accrual distribution

SISCR
UW - 2016



Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from seqOCWLR()
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Example: Operating characteristics with misspecified accrual distribution

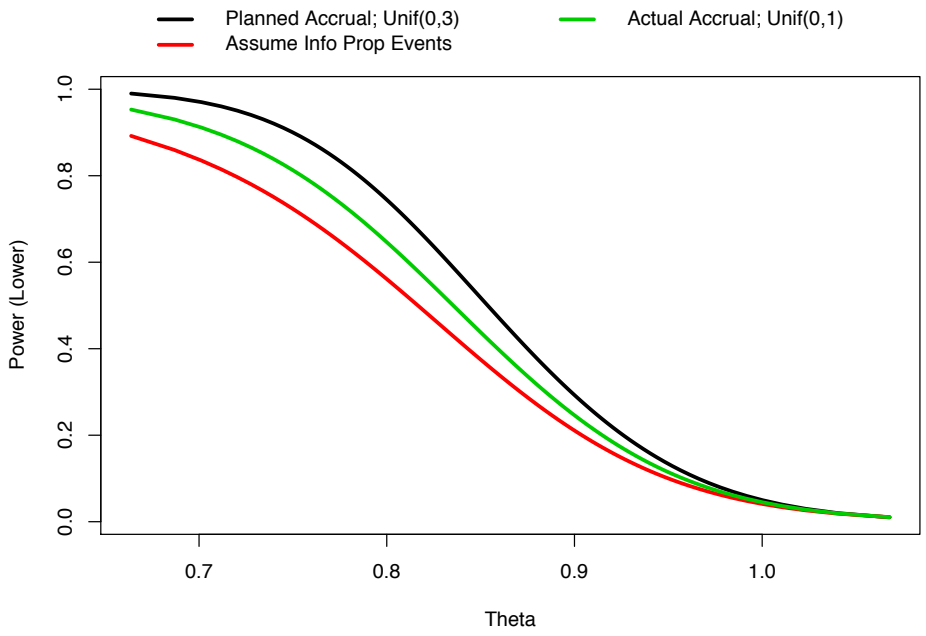
Example: Operating characteristics when testing with the $G^{1,1}$ Statistic

- Stopping boundaries if Unif(0,3) accrual assumed, but true accrual Unif(0,1)

```
STOPPING BOUNDARIES: Normalized Z-value scale
                          efficacy  futility
    Time 1 (Pi_1= 0.12) -3.2642    0.2094
    Time 2 (Pi_2= 0.36) -2.3082   -0.5534
    Time 3 (Pi_3= 0.66) -1.8846   -1.1387
    Time 4 (Pi_4= 1.00) -1.6321   -1.6321
```

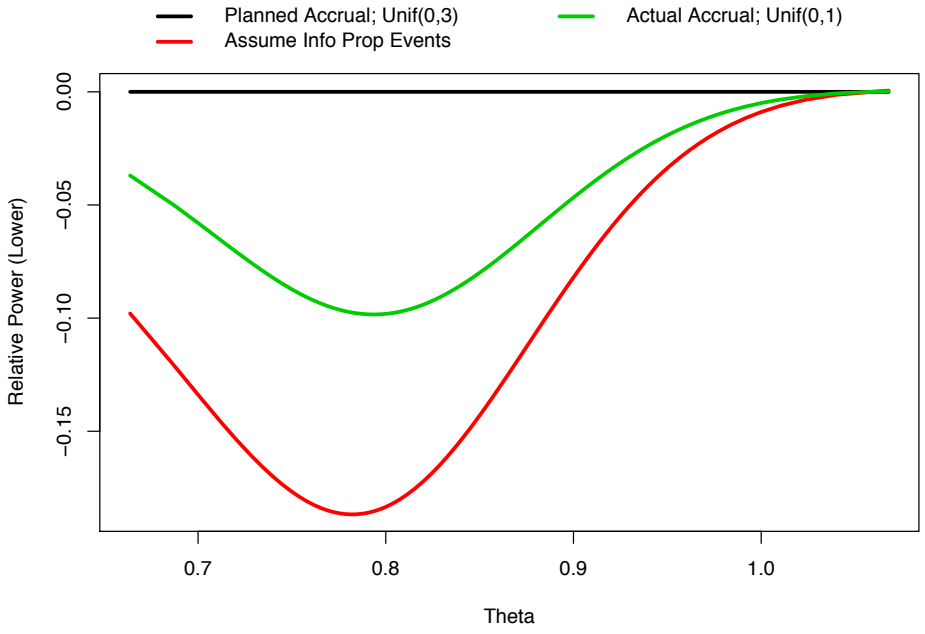

Example: Operating characteristics with misspecified accrual distribution

SISCR
UW - 2016



Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Example: Operating characteristics with misspecified accrual distribution



Implementation of group sequential rules

SISCR
UW - 2016

Goal: Maintain operating characteristics to be as close to design stage as possible

1. Need to choose between
   - maintaining maximal statistical information
   - maintaining statistical power
2. In addition, need to update our estimate of the information growth curve at each analysis
   - requires updating our estimate of $S(t)$ and $F_E(t)$ at each analysis

Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted LR statistics
Ex: Sensitivity of operating characteristics to the censoring distribution
RCTdesign implementation of group sequential rules

Implementation of group sequential rules

Algorithm as implemented in RCTdesign: Step 1

1. Specify original design using a parametric design family to satisfy desired operating characteristics
   - 1.1 specify timing of analyses
   - 1.2 assume $S(t)$ and $F_E(t)$
   - 1.3 estimate information growth curve
   - 1.4 map information increments to proportion of events for desired timing of first analysis

Implementation of group sequential rules

Algorithm as implemented in RCTdesign: Step 2

2. At first analysis,
   - 2.1 estimate $S(t)$ and $F_E(t)$ via parametric model
     - Use pooled data so that constraint does not depend on observed treatment effect
     - Estimate survival and accrual distributions via parametric models (weibull and scaled beta)
   - 2.2 re-estimate information growth curve
   - 2.3 map information increments to proportion of events for desired timing of future analyses
   - 2.4 constrain first boundary to exact timing (based upon current best estimate) and re-estimate future boundaries using pre-specified design family

Implementation of group sequential rules

Algorithm as implemented in RCTdesign: Step 3

3. At future analyses,

   3.1 re-estimate $S(t)$ and $F_E(t)$ via parametric model available data up to the analysis

   3.2 re-estimate information growth curve

   3.3 map information increments to proportion of events for desired timing of future analyses

   3.4 constrain previous boundaries to exact timing (based upon current best estimate) and re-estimate future boundaries using pre-specified design family

Module 18, Session 5:

Sequential and Adaptive Analysis with Time-to-Event Endpoints Sample Size Re-estimation with PH

Daniel L. Gillen, Ph.D.
Department of Statistics
University of California, Irvine

Scott S. Emerson, M.D., Ph.D.
Department of Biostatistics
University of Washington

Summer Institute in Statistics for Clinical Research
July 9, 2016

1

Sample Size Re-estimation

Proportional Hazards

2

Motivation

- Consider the design of an RCT that investigates prevention strategies in HIV / AIDS
- Our primary clinical endpoint is sero-conversion to HIV positive
- We will randomize individuals 1:1 experimental treatment to control

3

Recall

- In the presence of time to event endpoint that is subject to censoring, the most commonly used analyses are the logrank test and the proportional hazards regression model (Cox regression)
- When using PH regression with alternatives that satisfy the PH assumption, statistical information is proportional to the number of events
  - We can separately consider number accrued and calendar time of ending study
- Sample size calculations thus return the number of events that are necessary to obtain desired power
  - There are multiple ways that we can obtain that number of events as a function of
    - Number and timing of accrued subjects
    - Length of follow-up after start of study

4

Motivation

- Highly effective treatment and possibly low event rate
- HPTN052: 2011 scientific breakthrough of the year
  - Early vs Delayed ART is effective treatment in the prevention of HIV-1 transmission
  - Design: 188 events anticipated
    - based on (Placebo: 13.2% vs Treatment: 8.3%)
  - Blinded analysis: Total of 28 events
  - Unblinded analysis: 27 from the delayed ART arm
  - HR: 0.04 95% CI 0.01 - 0.27

5

Motivation

- Highly effective treatment and possibly low event rate
- Partners PrEP: 2012
  - Three arm double-blind trial of daily oral tenofovir (TDF) and emtricitabine/tenofovir (FTC/TDF)
    - 1:1:1 randomization of 4578 serodiscordant couples
  - Study halted 18 months earlier than planned due to demonstrated effectiveness in reduction of HIV-1 transmission
    - Of 78 infections, 18 in tenofovir, 13 in Truvada, 47 in control
    - Reduction in risk of infection 62% (95% CI 34-78%) in tenofovir, 73% (95% CI 49-85%); $p < 0.0001$ vs control
  - Special note: Placebo event rate was 1.99 per 100 PY rather than planned 2.75 per 100 PY

6

Issues

- In both of these trials the number of events observed was much lower than had been anticipated
- A priori, there are two reasons observed event rates could be lower than anticipated
  - Lower event rate in the control arm that had been guessed
  - Highly effective treatment leads to very few events in the experimental treatment
- In retrospect, both of these trials had both of these problems

7

Possible Solutions

- Well-understood methods
  - Wrong baseline event rate
    - Extend planned follow-up time
    - Live with lower power at planned calendar time EOS
    - Adaptive sample size re-estimation based on blinded results
      - Tradeoffs between accrual size and follow-up
  - Highly effective therapy
    - Group sequential design
- Less understood methods
  - Adaptive sample size re-estimation based on blinded results
    - Differentially revise maximum number of events and/or accrual/follow-up based on interim estimates of treatment effect

8

Extending Time of Follow-Up

- Under “information time” monitoring, this presents no statistical issues when proportional hazards holds
  - And “information time” monitoring is the usual standard in prespecifying RCT design in the time to event setting, and we would be supposed to do this
- Sometimes, however, we are only willing to believe PH assumption over some shorter time of follow-up
  - National Lung Screening Trial
  - Vaccine trials where need for boosters is not known
- Always, calendar time is ultimately more costly than number of patients
  - Emerson SC, et al. considers tradeoffs between time and number of patients

9

Accepting Lower Power

- If the prespecified RCT design defined the maximal statistical information according to calendar time, there is no statistical issue
- Under “information time” monitoring, this represents an unplanned change in the maximal statistical information
  - When this decision is made without knowledge of the unblinded treatment effect, regulatory agencies will usually allow the reporting of a “conditional analysis”
  - But the sponsor will need to be able to convincingly establish that it was still blinded to treatment effect
- Ethics of performing a grossly underpowered study must be considered
  - The predictive value of a “positive” study is greatly reduced

10

Blinded Adaptation of Sample Size

- If the prespecified RCT design defined the maximal statistical information according to number of events, then we must be talking about blinded adaptation of accrual size
  - Under PH distribution with PH analysis, no statistical issue
- Under "calendar time" monitoring, this represents an unplanned change in the maximal statistical information
  - When this decision is made without knowledge of the unblinded treatment effect, regulatory agencies will usually allow the reporting of a "conditional analysis"
  - But the sponsor will need to be able to convincingly establish that it was still blinded to treatment effect
  - This is likely only credible if you were delaying EOS

11

Group Sequential Design

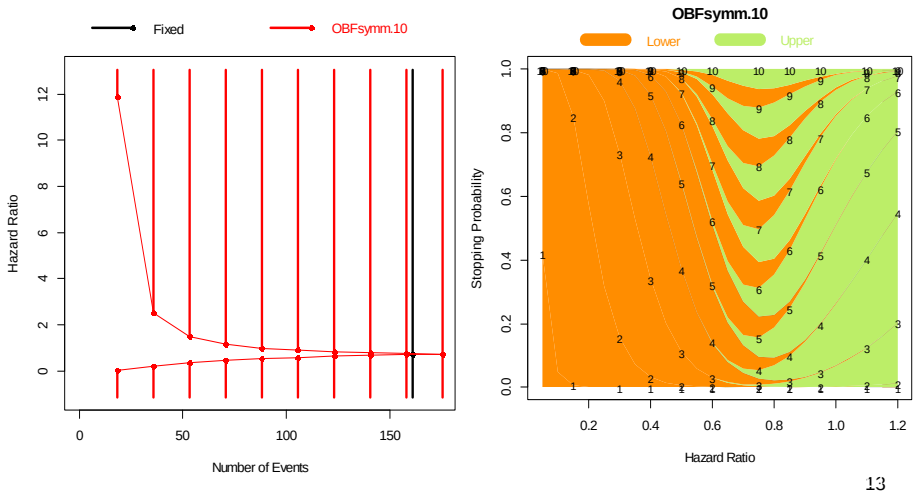
- Instead of a fixed sample design, pre-specify a group sequential design with, say, 10 possible analyses
  - Example: level 0.025, 90% power to detect HR=0.6

```
seqDesign(prob.model = "hazard", alt.hyp = 0.6, nbr.an = 10, power = 0.9)
PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.0    (size  = 0.025)
    Alternative hypothesis : Theta <= 0.6    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)
STOPPING BOUNDARIES: Sample Mean scale
                          Efficacy Futility
    Time  1 (NEv=  17.47)   0.0454  11.8598
    Time  2 (NEv=  34.95)   0.2132   2.5280
    Time  3 (NEv=  52.42)   0.3568   1.5101
    Time  4 (NEv=  69.90)   0.4617   1.1672
    Time  5 (NEv=  87.37)   0.5389   1.0000
    Time  6 (NEv= 104.85)   0.5974   0.9021
    Time  7 (NEv= 122.32)   0.6430   0.8381
    Time  8 (NEv= 139.79)   0.6795   0.7931
    Time  9 (NEv= 157.27)   0.7093   0.7597
    Time 10 (NEv= 174.74)   0.7341   0.7341
```

12

Group Sequential Design

- Stopping boundaries, stopping probabilities



13

Group Sequential Design

- Using this example, we see that if the true HR was 0.4 or less, we are virtually assured of stopping at the 4th analysis or earlier
- While the maximal number of events was 175, the 4th analysis occurs with 70 events.
- Suppose, a slow accrual of events is due solely to a highly effective treatment
  - Placebo has the planned event rate, Experimental treatment has extremely low event rate
- Relatively frequent monitoring will cause early termination long before the maximal event size needs to be observed
- We examine how calendar time might be affected

14

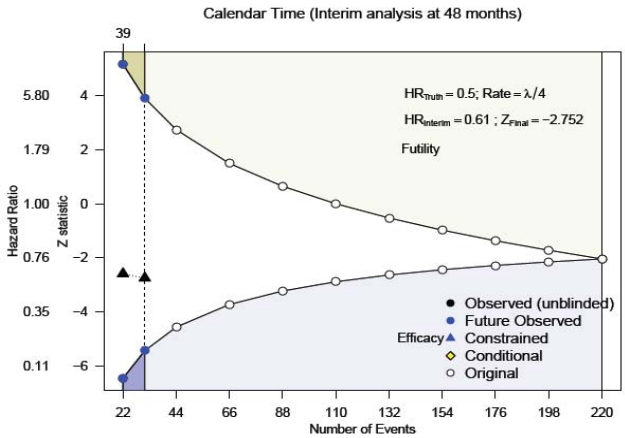
Incorporating Lower Event Rates

- We have not totally addressed problems that might arise with lower baseline event rates in the control group
  - If the treatment effect is not extreme, then the GSD might dictate that we proceed to the maximal sample size
- One approach is to build in an "escape clause" in the pre-specification of the RCT design
  - "The study will definitely terminate when we have 412 events or at 78 months after start of RCT, whichever comes first."

15

The Escape Clause

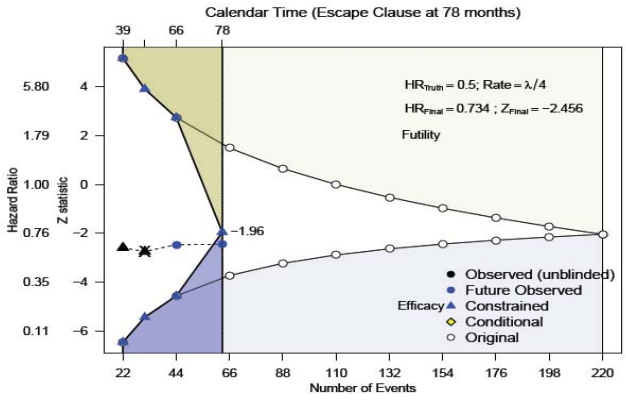
- Prior to pre-specified maximal calendar time, perform group sequential test as usual



16

The Escape Clause

- When the maximum calendar time is attained, modify the GST according to a constrained boundary approach / error spending function



17

Terminate for efficacy at 78 months

Unblinded Adaptation

- With unblinded adaptation, we can try to discriminate between
  - Strong treatment effect ➔ choose lower maximal event size
  - Low control event rate ➔ accrue more information
- We will have to decide whether to do adaptation prior to stopping accrual or whether to restart accrual
  - Early adaptation ➔ Less precise estimates of treatment effect
  - Late adaptation ➔ Have to restart accrual

18

Flexible Adaptive Designs

- Proschan and Hunsberger describe adaptations to maintain experimentwise type I error and increase conditional power
  - Must prespecify a conditional error function

$$\int_{-\infty}^{\infty} A(z)\,\phi(z)\,dz = \alpha.$$

  - Often choose function from some specified test

$$A(z) = Pr_{\delta=0}(Z_2 \geq \Phi^{-1}(1-\alpha)\,|\,\tilde{Z}_1 = z, \tilde{n}_2 = n_2 - n_1),$$

  - Find critical value to maintain type I error

$$Pr_{\delta=0}\left(Z_2^* \geq c(\tilde{n}_2^*, \tilde{z}_1)\,|\,\tilde{n}_2^*(\tilde{z}_1)\right) = A(\tilde{z}_1).$$

19

Other Approaches

- Self-designing Trial (Fisher, 1998)
  - Combine arbitrary test statistics from sequential groups
  - Prespecify weighting of groups “just in time”
    - Specified at immediately preceding analysis
  - Fisher’s test statistic is N(0,1) under the null hypothesis of no treatment difference on any of the endpoints tested
- Combining P values (Bauer & Kohne, 1994)
  - Based on R.A. Fisher’s method

20

Incremental Statistics

- Statistic at the j-th analysis a weighted average of data accrued between analyses

$$N_k^* = N_k - N_{k-1}$$

Statistics computed on kth increment : $\hat{\theta}_k^* \quad Z_k^* \quad P_k^*$

$$\hat{\theta}_j = \frac{\sum_{k=1}^{j} N_k^* \hat{\theta}_k^*}{N_j} \qquad Z_j = \frac{\sum_{k=1}^{j} \sqrt{N_k^*}\, Z_k^*}{\sqrt{N_j}}.$$

21

Conditional Distribution

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta, \frac{V}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\theta - \theta_0}{\sqrt{V / N_j^*}}, 1\right)$$

$$H_0$$

$$P_j^* \mid N_j^* \sim U(0,1).$$

22

Protecting Type I Error

- LD Fisher's variance spending method
  - Arbitrary hypotheses H_{0j}:$\theta_j = \theta_{0j}$
  - Incremental test statistics Z_j^*
  - Allow arbitrary weights W_j specified at stage *j-1*

$$Z_j = \frac{\sum_{k=1}^{J} \sqrt{W_k}\ Z_k^*}{\sqrt{\sum_{k=1}^{J} W_j}}$$

- RA Fisher's combination of P values (Bauer & Köhne)

$$P_j = \prod_{k=1}^{j} P_j^*$$

23

Unconditional Distribution

- Under the null
  - SDCT: Standard normal
  - Bauer & Kohne: Sum of exponentials
- Under the alternative
  - Unknown unless prespecified adaptations

$$\Pr\left(Z_j^* \le z\right) = \sum_{n=0}^{\infty} \Pr\left(Z_j^* \le z \mid N_j^*\right) \Pr\left(N_j^* = n\right).$$

24

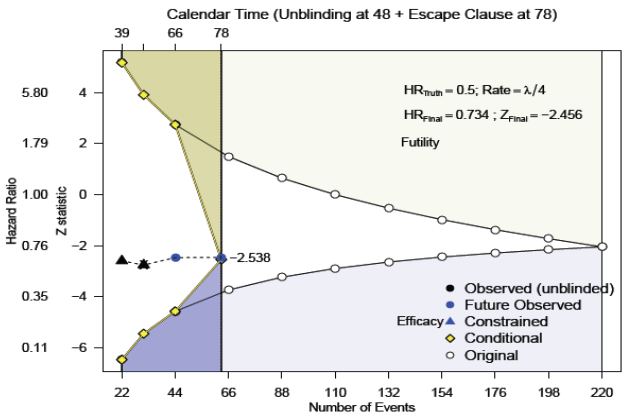
Sufficiency Principle

- It is easily shown that a minimal sufficient statistic is (Z, N) at stopping
- All methods advocated for adaptive designs are thus not based on sufficient statistics

25

What if Unblinded?

- When the maximum calendar time is attained, have to adjust the critical value according to the conditional error (CHW) or similar



26

Terminate for futility at 78 months (More conservative critical value)

Simulations

	HR=0.5 ; $\lambda/4$				HR=0.6343; $\lambda/2$			
	Continue		Restart		Continue		Restart	
	Pres	Cond	Pres	Cond	Pres	Cond	Pres	Cond
1750	68.69	-	68.69	-	67.55	-	67.55	-
3500	90.08	-	80.27	-	88.40	-	79.47	-
Fully Blinded‡	90.08	89.72	80.27	76.88	87.61	87.60	79.47	79.51
Avg Rate (80%)	86.33	85.74	78.27	73.91	84.63	84.59	77.55	77.36
Rate Diff (80%)	88.09	86.52	80.27	75.25	86.21	85.69	79.31	78.84
HR (80%)	87.55	86.31	80.10	75.07	86.10	85.58	79.35	78.77

- GSD (fully blinded procedures) almost efficient to the best *prespecified adaptive design* in context of $\lambda_{\text{Truth}} < \lambda_{\text{Planned}}$
- However, when integrity of the trial may be compromised and adjustments have to be used (CHW), we lose power
- The inefficient weighting scheme of CHW results in substantial loss of power particularly with late adaptations.

Final Comments

- The group sequential design definitely protects us from the extreme treatment effect
- In general, the group sequential design protected us from problems so long as the event rate was at least 25% of the planned rate
- There was definitely a price to pay when using the adaptive design
  - If the sponsor has access to unblinded results, adjustment for the adaptive analysis must be made
  - There is no allowance for the “escape clause” approach
  - Even more difficulty if non PH is possible

28

Module 19, Session 6:

Sequential and Adaptive Analysis with Time-to-Event Endpoints: Special Issues with Adaptive Methods

Daniel L. Gillen, Ph.D.
Department of Statistics
University of California, Irvine

Scott S. Emerson, M.D., Ph.D.
Department of Biostatistics
University of Washington

Summer Institute in Statistics for Clinical Research
July 29, 2016

1

Special Issues

- A basic premise of adaptive methods is that we can control the type 1 error, even when we have re-designed the trial based on interim estimates of the treatment effect
- Two special scenarios that we need to examine more closely
  - Do the interim statistics used in adjusting critical values truly contain all the information we had at our disposal?
  - Have we quantified the information growth correctly when using those statistics?

2

Control of Type 1 Errors

- Proschan and Hunsberger (1995)
  - Adaptive modification of RCT design at a single interim analysis can more than double type 1 error unless carefully controlled
- Those authors describe adaptations to maintain experimentwise type I error and increase conditional power
  - Must prespecify a conditional error function

$$\int_{-\infty}^{\infty} A(z)\,\phi(z)\,dz = \alpha.$$

  - Often choose function from some specified test

$$A(z) = Pr_{\delta=0}(Z_2 \geq \Phi^{-1}(1-\alpha)\,|\,\tilde{Z}_1 = z, \tilde{n}_2 = n_2 - n_1),$$

  - Find critical value to maintain type I error

$$Pr_{\delta=0}\left(Z_2^* \geq c(\tilde{n}_2^*, \tilde{z}_1)\,|\,\tilde{n}_2^*(\tilde{z}_1)\right) = A(\tilde{z}_1).$$

3

Alternative Approaches

- Combining P values (Bauer & Kohne, 1994)
  - Based on R.A. Fisher's method
  - Extended to weighted combinations
- Cui, Hung, and Wang (1999)
  - Maintain conditional error from pre-specified design
- Self-designing Trial (Fisher, 1998)
  - Combine arbitrary test statistics from sequential groups using weighting of groups pespecified "just in time"

4

Data at *j*-th Analysis: Immediate Outcome

- Subjects accrued at different stages are independent
- Statistics as weighted average of data accrued between analyses

At *k*th interim analysis	Incremental	Cumulative
Sample size (stat info)	N_k^*	$N_k = N_1^* + \cdots + N_k^*$
Baseline data	$\vec{X}_k^*$	$\vec{X}_k = (\vec{X}_1^*, \ldots, \vec{X}_k^*)$
1° outcome data	$\vec{Y}_k^*$	$\vec{Y}_k = (\vec{Y}_1^*, \ldots, \vec{Y}_k^*)$
2° outcome data	$\vec{W}_k^*$	$\vec{W}_k = (\vec{W}_1^*, \ldots, \vec{W}_k^*)$

Using $N_k^*, \vec{X}_k^*, \vec{Y}_k^*$:

Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*(N_k^*, \vec{X}_k^*, \vec{Y}_k^*)$	$\hat{\theta}_k = \frac{\sum_{j=1}^{k} N_j^* \hat{\theta}_j^*}{N_k}$
Normalized Z statistic	Z_k^*	$Z_k = \frac{\sum_{j=1}^{k} \sqrt{N_j^*} Z_j^*}{\sqrt{N_k}}$
Fixed sample P value	P_k^*	

5

Conditional Distn: Immediate Outcomes

- Sample size N_j^* and parameter θ_j can be adaptively chosen based on data from prior stages 1,…,*j-1*
  - (Most often we choose $\theta_j = \theta$ with immediate data)

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta_j, \frac{V(\theta_j)}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\hat{\theta}_j - \theta_{0j}}{\sqrt{V(\theta_j)/N_j^*}}, 1\right)$$

$$P_j^* \mid N_j^* \overset{H_0}{\sim} U(0,1).$$

Conditional distributions are totally independent under the null hypothesis

6

Estimands by Stage: Time to Event

- Most often we choose $\theta_j = \theta$ with immediate data
- In time to event data, a common treatment effect across stages is reasonable under some assumptions
  - Strong null hypothesis (exact equality of distributions)
  - Strong parametric or semi-parametric assumptions
- The most common methods of analyzing time to event data will often lead to varying treatment effect parameters across stages
  - Proportional hazards regression with non proportional hazards data
  - Weak null hypotheses of equality of summary measures (e.g., medians, average hazard ratio)

7

Partial Likelihood Based Score

- Logrank statistic

$$U(\beta) = \frac{\partial}{\partial \beta} \log L(\beta) = \sum_{i=1}^{n} D_i \left[X_i - \frac{\sum_{j:T_j \geq T_i} X_j \exp\{X_j \beta\}}{\sum_{j:T_j \geq T_i} \exp\{X_j \beta\}} \right]$$

$$= \sum_t \left[d_{1t} - \frac{n_{1t} e^{\beta}}{n_{0t} + n_{1t} e^{\beta}} (d_{0t} + d_{1t}) \right]$$

$$= \sum_t \frac{n_{0t} n_{1t}}{n_{0t} + n_{1t}} \left[\hat{\lambda}_{1t} - e^{\beta} \hat{\lambda}_{0t} \right]$$

8

Weighted Logrank Statistics

- Choose additional weights to detect anticipated effects

$$W(\beta)=\sum_t w(t)\frac{n_{0t}n_{1t}}{n_{0t}+n_{1t}}\left[\hat{\lambda}_{1t}-e^{\beta}\hat{\lambda}_{0t}\right]$$

$$n_{kt}=N_k\times\Pr(T\geq t, Cens\geq t)\overset{ind}{=}N_k S_k(t)\times\Pr(Cens\geq t)$$

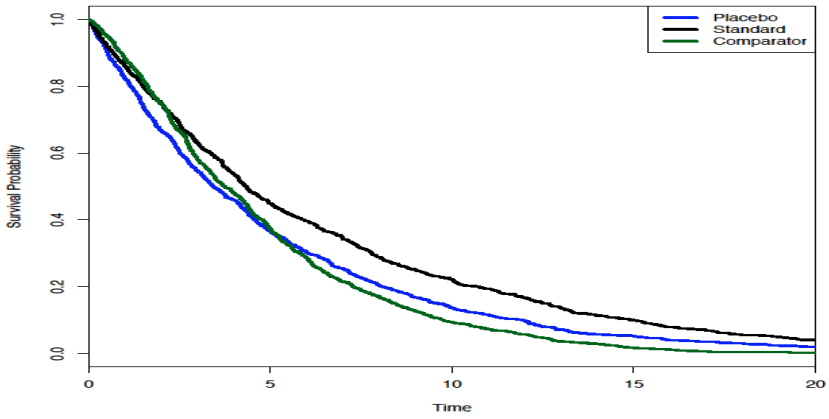
$G^{\rho\gamma}$ Family of weighted logrank statistics:

$$w(t)=\left[\hat{S}_{\bullet}(t)\right]^{\rho}\left[1-\hat{S}_{\bullet}(t)\right]^{\gamma}$$

9

Impact on Noninferiority Trials

- Weak null hypothesis is of greatest interest
  - Standard superior to placebo
  - Comparator (on average) equivalent to placebo



10

Conditional Distn: Immediate Outcomes

- Sample size N_j^* and parameter θ_j can be adaptively chosen based on data from prior stages 1,…,*j-1*
  - (Most often we choose $\theta_j = \theta$ with immediate data)

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta_j, \frac{V(\theta_j)}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\hat{\theta}_j - \theta_{0j}}{\sqrt{V(\theta_j)/N_j^*}}, 1\right)$$

$$P_j^* \mid N_j^* \overset{H_0}{\sim} U(0,1).$$

Conditional distributions are totally independent under the null hypothesis

11

Protecting Type I Error

- Test based on weighted averages of incremental test statistics
  - Allow arbitrary weights W_j specified by stage *j-1*

$$Z = \frac{\sum_{k=1}^{J} \sqrt{W_k}\ Z_k^*}{\sqrt{\sum_{k=1}^{J} W_j}} \qquad \overset{\bigcap_{k=1}^{J} H_{0j}}{\sim} \quad N(0,1)$$

$$Z = \frac{\sum_{k=1}^{J} \sqrt{W_k}\ \Phi^{-1}\left(1-P_k^*\right)}{\sqrt{\sum_{k=1}^{J} W_j}} \qquad \overset{\bigcap_{k=1}^{J} H_{0j}}{\sim} \quad N(0,1)$$

12

Complications: Longitudinal Outcomes

- Bauer and Posch (2004) noted that in the presence of incomplete data, partially observed outcome data may be informative of the later contributions to test statistics
- We need to make distinctions between
  - Independent subjects accrued at different stages
  - Statistical information about the primary outcome available at different analyses
- Owing to delayed observations, contributions to the primary test statistic at the *k*-th stage may come from subjects accrued at prior stages
  - Baseline and secondary outcome data available at prior analyses on those subject may inform the value of future data

13

Data at *j*-th Analysis: Delayed Outcome

- Subjects accrued at different stages are independent
- Some data is "missing"

At *k*th interim analysis	Incremental	Cumulative
Sample size (stat info)	N_k^*	$N_k = N_1^* + \cdots + N_k^*$
Baseline data	$\vec{X}_k^*$	$\vec{X}_k = \left(\vec{X}_1^*, \ldots, \vec{X}_k^*\right)$
1° outcome data (msng, observed)	$\vec{Y}_k^{*\mathrm{M}}, \vec{Y}_k^{*\mathrm{O}}$	$\vec{Y}_k^{\mathrm{M}}, \vec{Y}_k^{\mathrm{O}}$
2° outcome data	$\vec{W}_k^*$	$\vec{W}_k = \left(\vec{W}_1^*, \ldots, \vec{W}_k^*\right)$
Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*\left(N_k^*, \vec{X}_k^*, \vec{Y}_k^{*O}, \vec{Y}_{k-1}^M\right)$	$\hat{\theta}_k = \frac{\sum_{j=1}^{k} N_j^* \hat{\theta}_j^*}{N_k}$
Normalized Z statistic	Z_k^*	$Z_k = \frac{\sum_{j=1}^{k} \sqrt{N_j^*}\, Z_j^*}{\sqrt{N_k}}$
Fixed sample P value	P_k^*	

14

Major Problem: Delayed Outcome

- When sample size N_j^* and parameter θ_j adaptively chosen based on data from prior stages 1,…,*j-1,* some aspect of the "future" contributions may already be known

At *k*th interim analysis	Incremental	Cumulative
Sample size	$N_k^* = N_k^*\left(N_{k-1}, \vec{X}_{k-1}, \vec{W}_{k-1}, \vec{Y}_{k-1}^{*O}, \vec{Y}_{k-2?}^{M}\right)$	N_k
Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*\left(N_k^*, \vec{X}_k^*, \vec{Y}_k^{*O}, \vec{Y}_{k-1}^{M}\right)$	$\hat{\theta}_k = \frac{\sum_{j=1}^{k} N_j^* \hat{\theta}_j^*}{N_k}$

Impact: (One statistician's mean is another statistician's variance)

$corr(\vec{Y}_k^{*M}, \vec{W}_k^*) \neq 0$ or $corr(\vec{Y}_k^{*M}, \vec{X}_k^*) \neq 0 \quad \Rightarrow \quad \hat{\theta}_k^* \mid N_k^*$ not indep of $\hat{\theta}_{k+1}^* \mid N_{k+1}^*$

$\hat{\theta}_k^* \mid N_k^*$ is potentially biased for θ_k and not approximately normal

15

Potential Solutions

- Jenkins, Stone & Jennison (2010)
  - Only use data available at the *k*-th stage analysis
- Irle & Schaefer (2012)
  - Prespecify how the full *k*-th stage data will eventually contribute to the estimate of θ_k
- Magirr, Jaki, Koenig & Posch (2014, arXiv.org)
  - Assume worst case of full knowledge of future data and sponsor selection of most favorable P value

16

Comments: Burden of Proof Dilemma

- There is a contradiction of standard practices when viewing the incomplete data
  - We would never accept the secondary outcomes as validated surrogates
  - But we feel that we must allow for the possibility that the secondary outcomes were perfectly predictive of the eventual data
- We are in some sense preferring mini-max optimality criteria over a Bayes estimator

17

Comments: Impact on RCT Design

- The candidate approaches will protect the type 1 error, but the impact on power (and PPV) is as yet unclear
- Weighted statistics are not based on minimal sufficient statistics
  - But greatest loss in efficiency comes from late occurring adaptive analyses with large increases in maximal statistical information
  - Time to event will not generally have this
- The adaptation is based on imprecise estimates of the estimates that will eventually contribute to inference
- We may have to eventually either
  - Ignore some observed data (JS&S, I&S), or
  - Adjust for worst case multiple comparisons

18

What if No Adjustment?

- Many methods for adaptive designs seem to suggest that there is no need to adjust for the adaptive analysis if there were no changes to the study design
- However, changes to the censoring distribution definitely affect
  - Distribution-free interpretation of the treatment effect parameter
  - Statistical precision of the estimated treatment effect
  - Type 1 error when testing a weak null (e.g., noninferiority)
- Furthermore, “less understood” analysis models prone to inflation of type 1 error when testing a strong null
  - Information growth with weighted log rank tests is not always proportional to the number of events

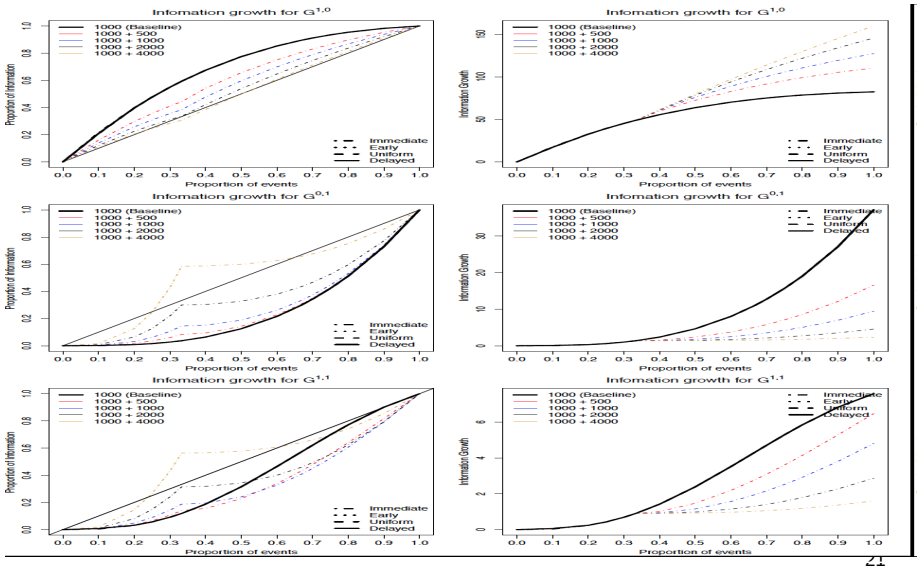
19

“Intent to Cheat” Zone

- At interim analysis, choose range of interim estimates that lead to increased accrual of patients
- How bad can we inflate type 1 error when holding number of events constant?
- Logrank test under strong null: Not at all
- Weighted logrank tests: Up to relative increase of 20%
  - Sequela of true information growth depends on more than number of events
  - Power largely unaffected, so PPV decreases

20

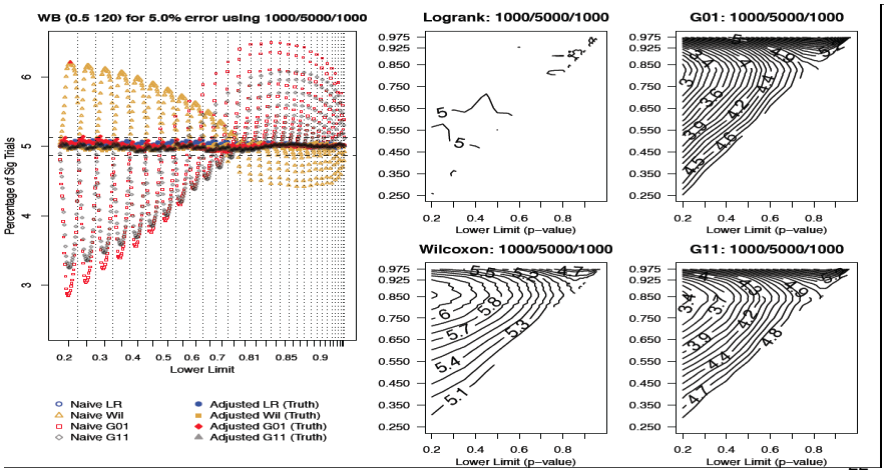
Information Growth with Adaptation



21

Inflation of Type 1 Error

- Function of definition of the adaptation zone
  - Varies according to weighted log rank test



Final Comments

- There is still much for us to understand about the implementation of adaptive designs
- Most often the "less well understood" part is how they interact with particular data analysis methods
  - In particular, the analysis of censored time to event data has many scientific and statistical issues
- How much detail about accrual patterns, etc. do we want to have to examine for each RCT?
- How much do we truly gain from the adaptive designs?
  - (Wouldn't it be nice if statistical researchers started evaluating their new methods in a manner similar to evaluation of new drugs?)

23

Bottom Line

- There is no substitute for planning a study in advance
  - At Phase 2, adaptive designs may be useful to better control parameters leading to Phase 3
    - Most importantly, learn to take "NO" for an answer
  - At Phase 3, there seems little to be gained from adaptive trials
    - We need to be able to do inference, and poorly designed adaptive trials can lead to some very perplexing estimation methods
- **"Opportunity is missed by most people because it is dressed in overalls and looks like work."** -- Thomas Edison
- In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

24

Really Bottom Line

"You better think (think)
about what you're
trying to do…"

-Aretha Franklin, "Think"

25