Module 12:

Adaptive Sequential RCT – Advanced Topics

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

Summer Institute in Statistics for
Clinical & Epidemiological Research
July 25, 2019

Abstract

Adaptive sequential sampling plans are often used in the monitoring of clinical trials in order to address the ethical and efficiency issues inherent in human testing of a new treatment or preventive agent for disease. Group sequential stopping rules are perhaps the most commonly used approaches, but in recent years, a number of authors have proposed more flexible adaptive methods such as unblinded sample size re-estimation, adaptive enrichment, and response adaptive randomization. This module describes some of the special issues that can arise from a statistical and/or regulatory standpoint in any sequential RCT. Particular emphasis is placed on issues that can arise with time to event endpoints. Topics include the

- analysis of RCT data in the presence of adaptive enrichment, response adaptive randomization or adaptive selection of doses,
- the use of sequential designs when treatment effects might be time-varying (e.g. non proportional hazards), and
- the avoidance of operational bias with unblinded adaptation

2

Where Am I Going?

Overview and Organization of the Course

3

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 delayed measured of outcome data affect those issues?
  - Short answer: No, UNLESS subject to censoring
  - So, true answer: Yes.

4

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
- Note: Longitudinal data analyses based on (average) disease status at (over) some time period have many of these same issues)
  - I will refer to these similarities at times

5

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

6

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

7

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
  - Tradeoffs between number accrued and length of follow-up
  - 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

8

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

9

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

10

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)

11

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

12

Evidence Based Medicine

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

13

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?

14

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)

15

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

16

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

17

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

18

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

19

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

20

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:

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

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

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

21

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

22

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

23

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

24

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

25

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

26

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

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

27

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"

28

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

29

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)

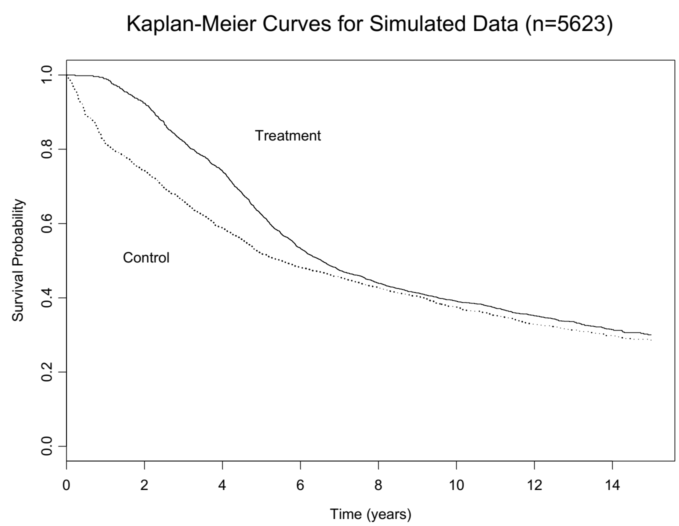
30

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

31

Hypothetical Example: KM Curves



32

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?

33

Who Wants To Be A Millionaire?

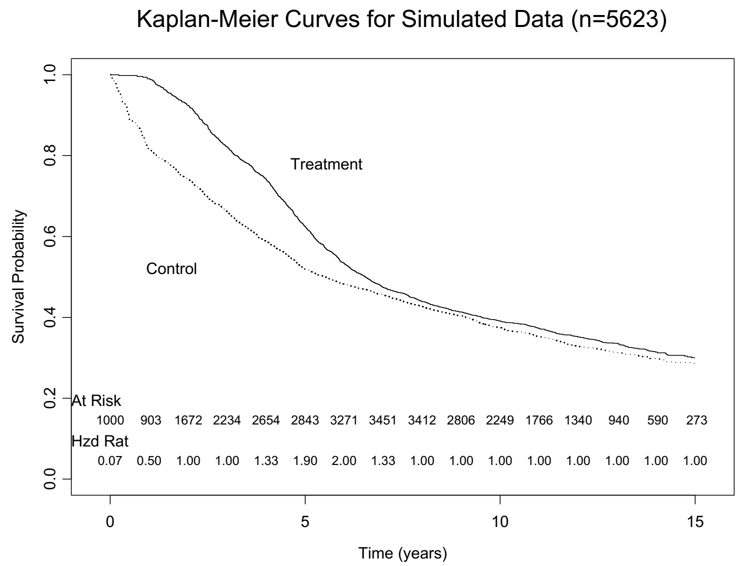
- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

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

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

34

Hypothetical Example: KM Curves



35

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

36

Partial Likelihood Based Score

- Logrank statistic

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

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

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

37

Weighted Logrank Statistics

- Choose additional weights to detect anticipated effects

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

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

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

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

38

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.

39

Logan, et al.: Motivation

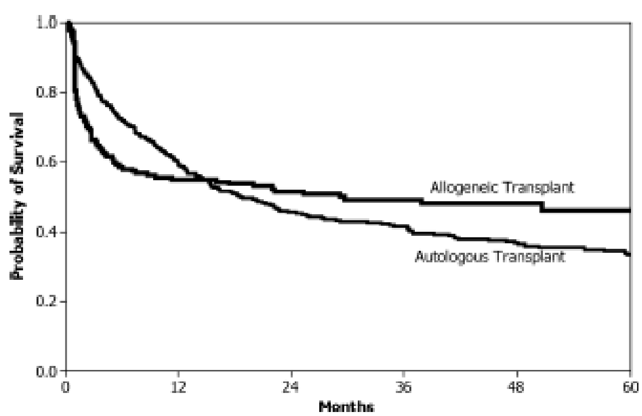


Figure 1. Kaplan–Meier estimate of DFS for follicular lymphoma example, by stem cell source.

40

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

41

Logan, et al.: Simulations

Comparing Treatments with Crossing Survival Curves

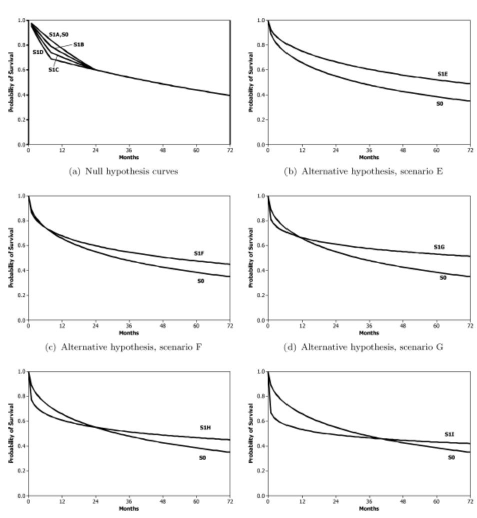


Figure 2. Survival curves for treatment (S1) and control (S0) groups used in simulations. Curves for the null hypot simulations are shown in (a) for each of the four scenarios, and curves for the alternative hypothesis simulations are show (b)–(f) for the five scenarios.

42

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
$Z_{\mathrm{CLL}}(24)$	(1)	62.4	15.3	21.1	4.7	21.8
$Z_{\mathrm{CLL}}(48)$	(1)	70.1	32.9	65.1	21.5	6.8
$Z_{\mathrm{CLL}}(72)$	(1)	71.2	44.5	85.1	46.1	25.9
$Z_{\mathrm{WKM}}(t_0)$	(2)	75.8	35.0	66.3	20.3	6.0
$\chi^2_{\mathrm{PSV}}(t_0)$	(3)	74.8	32.0	61.2	16.4	4.8
$Z_{\mathrm{LR}}(t_0)$	(4)	30.7	36.5	85.4	71.7	82.6
$Z_{\mathrm{OLS}}(t_0)$	(5)	74.7	43.9	84.1	43.4	23.6
$Z_{\mathrm{SP,P}}(t_0)$	(6)	76.9	40.2	74.8	29.6	10.7
$\boxed{\chi^2(t_0)}$	(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

43

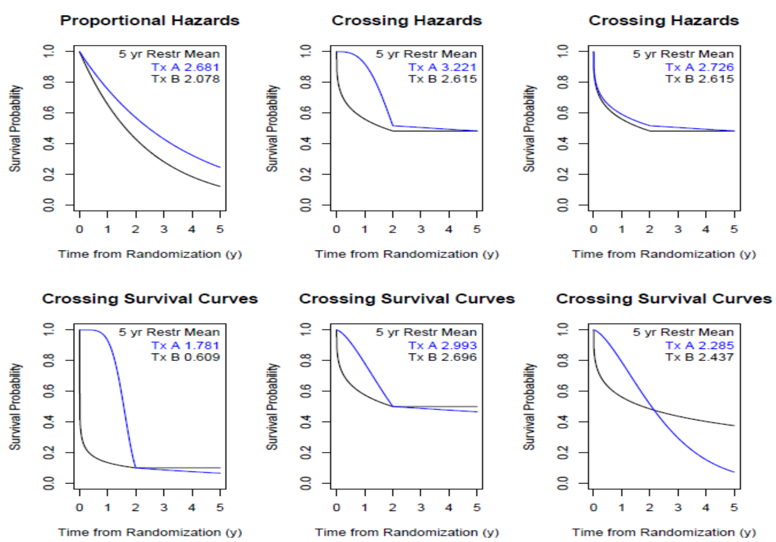
Logan, et al.: Critique

- 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

44

Logan, et al.: Critique

- Additional scenarios that are of interest



45

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%

46

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

47

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?)

48

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.

49

Really Bottom Line

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

-Aretha Franklin, "Think"

50

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

51

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)

52

Right Censored Data

- Incompete data: Some events have not occurred at time of data analysis
- Notation:

Unobserved:

True times to event: $\left\{T_1^0, T_2^0, \ldots, T_n^0\right\}$

Censoring Times: $\left\{C_1, C_2, \ldots, C_n\right\}$

Observed data:

Observation Times: $T_i = \min\left(T_i^0, C_i\right)$

Event indicators: $D_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

Group membership: X_i

53

Hypothetical Example: Analysis

- Choice of summary measure
  - Survival at fixed point in time
  - Median, other quantiles
  - Mean (or restricted mean)
  - Hazard ratio (or weighted average of hazard ratio over time)
- Choice of methods
  - Parametric, semiparametric, nonparametric

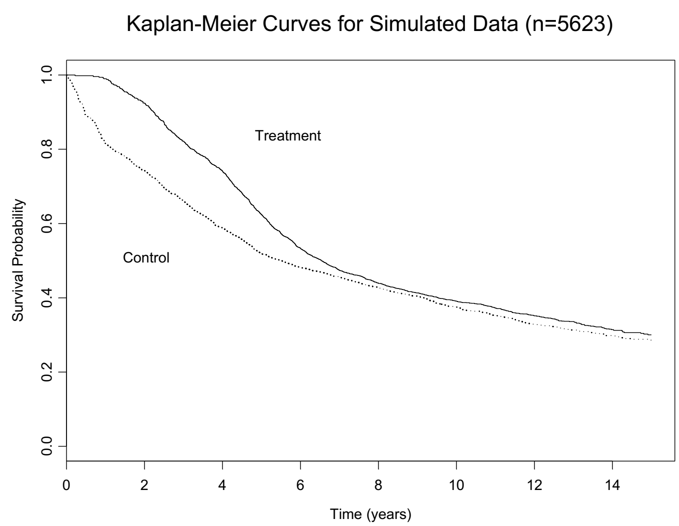
54

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

55

Hypothetical Example: KM Curves



56

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?

57

Who Wants To Be A Millionaire?

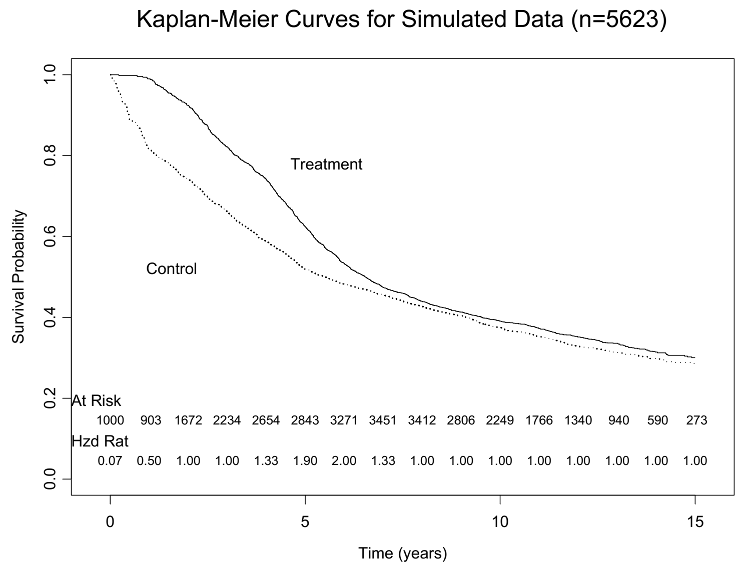
- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

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

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

58

Hypothetical Example: KM Curves



59

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

60

Partial Likelihood Based Score

- Logrank statistic

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

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

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

61

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

62

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)

63

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

64

Evidence Based Medicine

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

65

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?

66

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)

67

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

68

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

69

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

70

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

71

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

72

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:

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

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

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

73

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

74

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

75

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

76

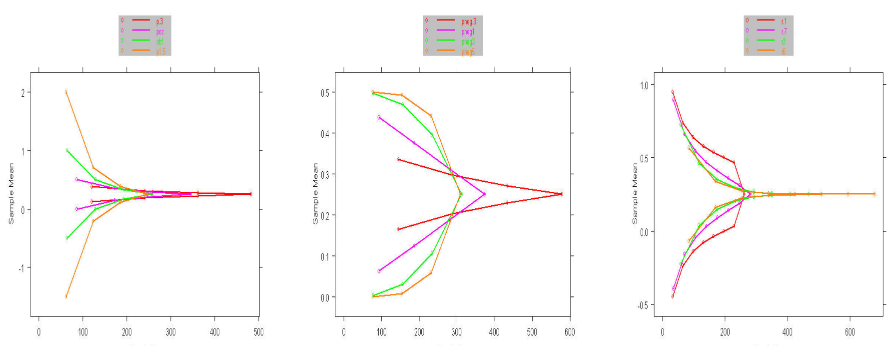
Group Sequential Designs

- Perform analyses when sample sizes $N_1 . . . N_J$
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 . . . X_{Nj})$
  - 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

77

Spectrum of Boundary Shapes

- All of the rules depicted have the same type I error and power to detect the design alternative



RCT Design to Address Variability

- At the end of the study we perform frequentist and/or Bayesian data analysis to assess the credibility of clinical trial results
  - Estimate of the treatment effect
    - Single best estimate
    - Precision of estimates
  - Decision for or against hypotheses
    - Binary decision
    - Quantification of strength of evidence

79

Measures of Precision

- Estimators are less variable across studies
  - Standard errors are smaller
- Estimators typical of fewer hypotheses
  - Confidence intervals are narrower
- Able to statistically reject false hypotheses
  - Z statistic is higher under alternatives

80

Notation

Potential data : $Y_1, Y_2, Y_3, \ldots, Y_{N_J}$

Probability model : $Y_i \overset{iid}{\sim} (\theta, V)$

Interim estimates : $\hat{\theta}_{N_j} = \hat{\theta}(Y_1, \ldots, Y_{N_j})$

Without sequential sampling :

Approximate distn : $\hat{\theta}_j = \hat{\theta}_{N_j} \dot{\sim} N(\theta, V / N_j)$

Indep increments : $Cov(\hat{\theta}_{N_j}, \hat{\theta}_{N_{j+1}}) = V / N_{j+1}$

Interim test statistics : $Z_j = Z_{N_j} = \dfrac{\hat{\theta}_j - \theta_0}{\sqrt{V / N_j}}$

81

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

Typically : $se(\hat{\theta}) = \sqrt{\dfrac{V}{n}}$

(V related to average "statistical information")

Width of CI : $2 \times (crit\ val) \times se(\hat{\theta})$

Test statistic : $Z = \dfrac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$

82

Ex: Difference of Indep Means

$$ind\ Y_{ij} \sim \left(\mu_i, \sigma_i^2\right), i = 1,2;\ j = 1,\ldots,n_i$$

$$n = n_1 + n_2; \quad r = n_1 / n_2$$

$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

$$V = (r+1)\left[\sigma_1^2 / r + \sigma_2^2\right] \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

83

Ex: Hazard Ratios

- With **noninformative** censoring, proportional hazards
  - Statistical information involves probability of censoring

$$ind \text{ censored time to event } \left(T_{ij}, \delta_{ij}\right),$$

$$i = 1,2;\ j = 1,\ldots,n_i; n = n_1 + n_2; \quad r = n_1 / n_2$$

$$\theta = \log(HR) \qquad \hat{\theta} = \hat{\beta} \text{ from PH regression}$$

$$V = \frac{(1+r)(1/r+1)}{\Pr\left[\delta_{ij} = 1\right]} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{(1+r)(1/r+1)}{d}}$$

84

Time to Event Analyses

- Sample size computation usually presumes PH
  - Perhaps attenuation of effect due to cross-over
  - Perhaps precision gained by deattenuating HR with adjustment for prognostic baseline variables
- Formula leads to number of events
- Accrual size based on
  - Control event rate
  - Hypothesized treatment effect (null vs alternative)
  - Accrual time
  - Follow-up after accrual ends
  - (Censoring due to loss to follow-up?)

85

Sample Size Determination

- Based on sampling plan, statistical analysis plan, and estimates of variability, compute
  - Sample size that discriminates hypotheses with desired power,

    OR

  - Hypothesis that is discriminated from null with desired power when sample size is as specified, or

    OR

  - Power to detect the specific alternative when sample size is as specified

86

Sample Size Computation

Standardized level α test (n = 1): $\delta_{\alpha\beta}$ detected with power β

Level of significance α when $\theta = \theta_0$

Design alternative $\theta = \theta_1$

Variability V within 1 sampling unit

Required sampling units: $$n = \frac{(\delta_{\alpha\beta})^2 V}{(\theta_1 - \theta_0)^2}$$

(Fixed sample test: $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_\beta$)

87

When Sample Size Constrained

- Often (usually?) logistical constraints impose a maximal sample size
  - Compute power to detect specified alternative

Find β such that $$\delta_{\alpha\beta} = \sqrt{\frac{n}{V}}(\theta_1 - \theta_0)$$

  - Compute alternative detected with high power

$$\theta_1 = \theta_0 + \delta_{\alpha\beta}\sqrt{\frac{V}{n}}$$

88

Ex: Difference of Indep Means

$$ind\ Y_{ij} \sim \left(\mu_i, \sigma_i^2\right), i = 1,2;\ j = 1, \ldots, n_i$$

$$n = n_1 + n_2; \quad r = n_1 / n_2$$

$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

$$V = (r+1)\left[\sigma_1^2 / r + \sigma_2^2\right] \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

89

Ex: Hazard Ratios

ind censored time to event $\left(T_{ij}, \delta_{ij}\right)$,

$$i = 1,2;\ j = 1, \ldots, n_i; n = n_1 + n_2; \quad r = n_1 / n_2$$

$\theta = \log(HR) \qquad \hat{\theta} = \hat{\beta}$ from PH regression

$$V = \frac{(1+r)(1/r+1)}{\Pr\left[\delta_{ij} = 1\right]} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{(1+r)(1/r+1)}{d}}$$

90

Increasing Precision

- Options
  - Increase sample size
    - Time to event: Accrue more patients or follow for longer
  - Decrease V
    - Improve reliability of measurements
      - Time to event: Decrease probability of censoring
    - Alter study design (e.g., cross-over)
    - (Alter eligibility to decrease heterogeneity)
    - (Alter clinical endpoint)
  - Time to event: Adjust (control) for precision variables
    - De-attenuate the hazard ratio
  - (Decrease confidence level)

91

Simulations

- Unadjusted and adjusted estimates of effect of binary POI as a function of
  - Effect of a covariate on summary of outcome (mean, odds, …)
  - Sampling scheme: Association between covariate and POI
    - Difference in mean covariate
    - Difference in median covariate

$$\text{Sampling} \quad : \quad E[W_i \mid X_i] = \alpha_0 + \alpha_1 \times X_i$$

$$\hat{\alpha}_1 = r_{XW}\frac{s_W}{s_X} = \overline{W}_{X=1} - \overline{W}_{X=0}$$

$$\text{Unadjusted} : \quad g[\theta \mid X_i] = \beta_0 + \beta_1 \times X_i$$

$$\text{Adjusted} \quad : \quad g[\theta \mid X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

92

Linear Regression

- Simulation results

	Truth					Avg Estimates (SE)	
	ΔMdn	α_1	r_{XW}	γ_2	γ_1	β_1	γ_1
Irrelevant	0.0	0.0	0.00	0.0	0.0	0.0 (0.20)	0.0 (0.20)
Precision	0.0	0.0	0.00	1.0	0.0	0.0 (0.28)	0.0 (0.19)
Precision	- 0.3	0.0	0.00	1.0	0.0	0.0 (0.28)	0.0 (0.20)
Precision	0.0	0.0	0.00	1.0	1.0	1.0 (0.28)	1.0 (0.20)
Confound	0.3	0.3	0.15	1.0	0.0	0.3 (0.28)	0.0 (0.21)
Confound	0.0	0.3	0.15	1.0	0.0	0.3 (0.29)	0.0 (0.21)
Var Inflatn	0.0	1.0	0.45	0.0	0.0	0.0 (0.20)	0.0 (0.22)

93

Proportional Hazards Regression

- Simulation results

	Truth					Avg Estimates (SE)	
	ΔMdn	α_1	r_{XW}	γ_2	γ_1	β_1	γ_1
Irrelevant	0.0	0.0	0.00	0.0	0.0	0.0 (0.20)	0.0 (0.20)
Precision	0.0	0.0	0.00	1.0	0.0	0.0 (0.21)	0.0 (0.22)
Precision	- 0.3	0.0	0.00	1.0	0.0	0.0 (0.21)	0.0 (0.21)
Precision	0.0	0.0	0.00	1.0	1.0	0.7 (0.21)	1.0 (0.22)
Confound	0.3	0.3	0.15	1.0	0.0	0.2 (0.21)	0.0 (0.21)
Confound	0.0	0.3	0.15	1.0	0.0	0.1 (0.20)	0.0 (0.22)
Var Inflatn	0.0	1.0	0.45	0.0	0.0	0.0 (0.20)	0.0 (0.23)

Next

94

Subjects vs Time

- Sample size computations for time to event generally return the number of events that are needed
- We then have to decide on
  - The length of accrual time
  - The pattern of accrual (e.g., uniform vs increasing rates)
  - How many subjects to accrue
  - The follow-up time after accrual ends
  - Any noninformative loss to follow-up
  - The event rates (over time)
- Note that the event rates are presumed to differ according to treatment effects
  - Different rates of statistical information accrual under null and alternative

95

Time to Event Endpoints

Estimating Subject Accrual in RCTdesign

Where am I going?

In time to event analyses, statistical information is roughly proportional to the number of events.

Additional consideration must be given to accrual of subjects.

Time to Event Endpoints

- RCTdesign allows specification of the "hazard" probability model for time to event data
  - Logrank statistic
  - Estimates of hazard ratio using Cox model
- "Sample size" computations return number of events primarily
- Additional accrual models are used to estimate
  - Number of subjects to accrue
  - Calendar time of interim analyses

97

Case Study: Stopping Rule

- Design of RCT to test a new drug for NSCLC
  - One-sided type 1 error: 0.025 for null of HR=1.0
  - Power: 90% to detect HR= 0.77
  - Four interim analyses with OBF efficacy, intermediate futility

```
> tte <- seqDesign("hazard",alt=0.77,power=0.9,
+   nbr=4,P=c(1,0.8))
> tte
Call:
seqDesign(prob.model = "hazard", alt.hypothesis = 0.77, nbr.analyses = 4,
    power = 0.9, P = c(1, 0.8))

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 <= 0.77    (power = 0.900)

STOPPING BOUNDARIES: Sample Mean scale
                       Efficacy Futility
    Time 1 (NEv= 163.7)   0.5372   1.1891
    Time 2 (NEv= 327.4)   0.7329   0.9651
    Time 3 (NEv= 491.1)   0.8129   0.8927
    Time 4 (NEv= 654.8)   0.8561   0.8561
```

98

Specification of Accrual

- Need to observe 655 events
- Accrual pattern can be specified in a variety of ways
  - Accrual and additional follow-up time
    - Calculates accrual rate
  - Accrual rate and study time
    - Calculates accrual time
- Wide variety of distributions for
  - Accrual (uniform, beta, rates for each time period)
  - Loss to follow-up (exponential, Weibull, piecewise exponential)
  - Time to event (exponential, Weibull, piecewise exponential)
- Accrual parameters can be specified in seqDesign()

99

Case Study: Accrual

- Need to observe 655 events
- Assume accrual over 3 years
  - Use uniform accrual
- Assume additional follow-up of 1 year
  - Total study time of 4 years
- Assume median survival of 3 years
  - Use exponential survival distribution

100

Case Study: Accrual

- tte <- update(tte, accrualTime= 3, studyTime= 4, eventQuantiles= 3)

```
Accrual summary table:
            theta Scenario NAccrual accrualRate accrualTime studyTime
alternative  0.77        1     1684       561.3           3         4
null         1.00        1     1537       512.3           3         4

Timing of analyses:

  Theta = 0.77    Scenario  1
              Analysis 1 Analysis 2 Analysis 3 Analysis 4
Analysis Time      1.801      2.615      3.269        4.0
N Accrued       1006.224   1465.501   1684.000     1684.0
N Events         163.700    327.400    491.100      654.8

  Theta = 1    Scenario  1
              Analysis 1 Analysis 2 Analysis 3 Analysis 4
Analysis Time      1.781      2.589      3.255        4.0
N Accrued        915.222   1329.520   1537.000     1537.0
N Events         163.700    327.400    491.100      654.8
```

101

Case Study: Accrual (months)

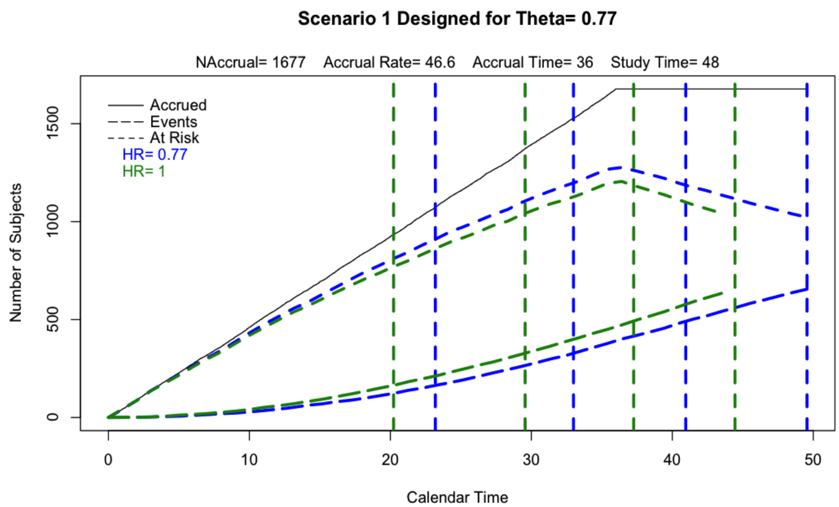
- tte <- update(tte, accrualTime= 36, studyTime= 48, eventQuantiles= 36)

```
Accrual summary table:
            theta Scenario NAccrual accrualRate accrualTime studyTime
alternative  0.77        1     1677       46.58          36        48
null         1.00        1     1528       42.44          36        48

Timing of analyses:

  Theta = 0.77    Scenario  1
              Analysis 1 Analysis 2 Analysis 3 Analysis 4
Analysis Time      21.53      31.19      39.18       48.0
N Accrued        1005.16    1454.43    1677.00     1677.0
N Events          163.70     327.40     491.10      654.8

  Theta = 1    Scenario  1
              Analysis 1 Analysis 2 Analysis 3 Analysis 4
Analysis Time      21.43      31.21       39.2       48.0
N Accrued         913.93    1324.33     1528.0     1528.0
N Events          163.70     327.40      491.1      654.8
```

102

Case Study: Accrual Plots

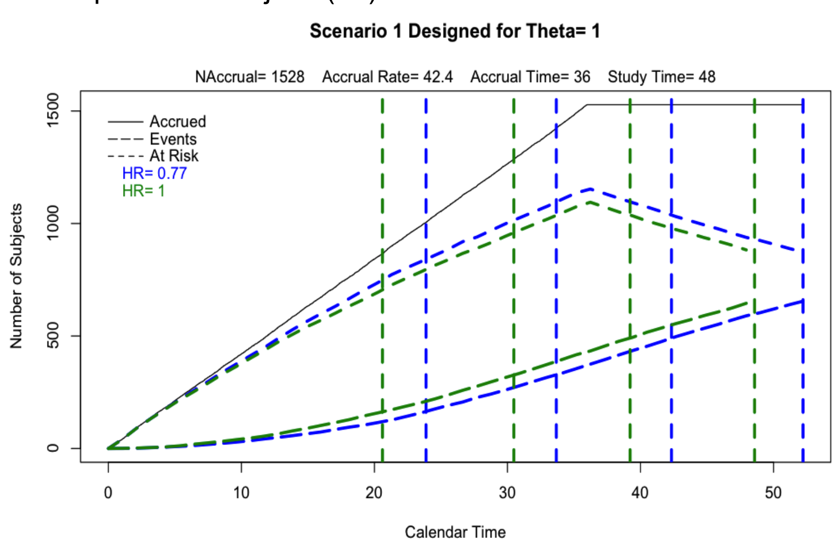
- seqPlotPHNSubjects(tte)



103

Case Study: Accrual Plots

- seqPlotPHNSubjects(tte)



104

Evaluation of Designs

- Process of choosing a trial design
  - Define candidate design
    - Usually constrain two operating characteristics
      - Type I error, power at design alternative
      - Type I error, maximal sample size
  - Evaluate other operating characteristics
    - Different criteria of interest to different investigators
  - Modify design
  - Iterate

105

Collaboration of Disciplines

Discipline	Collaborators	Issues
Scientific	Epidemiologists Basic Scientists Clinical Scientists	Hypothesis generation Mechanisms Clinical benefit
Clinical	Experts in disease / treatment Experts in complications	Efficacy of treatment Adverse experiences
Ethical	Ethicists	Individual ethics Group ethics
Economic	Health services Sponsor management Sponsor marketers	Cost effectiveness Cost of trial / Profitability Marketing appeal
Governmental	Regulators	Safety Efficacy
Statistical	Biostatisticians	Estimates of treatment effect Precision of estimates
Operational	Study coordinators Data management	Collection of data Study burden Data integrity

106

Which Operating Characteristics

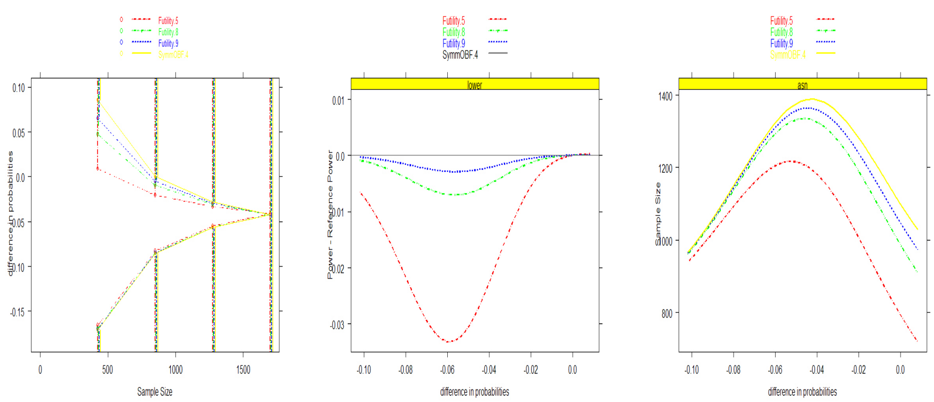
- The same regardless of the type of stopping rule
- Frequentist power curve
  - Type I error (null) and power (design alternative)
- Sample size requirements
  - Maximum, average, median, other quantiles
  - Stopping probabilities
- Inference at study termination (at each boundary)
  - Frequentist or Bayesian (under spectrum of priors)
- (Futility measures
  - Conditional power, predictive power)

107

Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power

Boundaries Loss of Power Avg Sample Size



At Design Stage

- In particular, at design stage we can know
  - Conditions under which trial will continue at each analysis
    - Estimates
      - » (Range of estimates leading to continuation)
    - Inference
      - » (Credibility of results if trial is stopped)
    - Conditional and predictive power
  - Tradeoffs between early stopping and loss in unconditional power

109

Operating Characteristics

- For any pre-specified stopping rule, however, we can compute the correct sampling distribution with specialized software
- From the computed sampling distributions we then compute
  - Bias adjusted estimates
  - Correct (adjusted) confidence intervals
  - Correct (adjusted) P values
- Candidate designs are then compared with respect to their operating characteristics

110

But What If …?

- Possible motivations for adaptive designs
  - Changing conditions in medical environment
    - Approval / withdrawal of competing / ancillary treatments
    - Diagnostic procedures
  - New knowledge from other trials about similar treatments
  - Evidence from ongoing trial
    - Toxicity profile (therapeutic index)
    - Interim estimates of primary efficacy / effectiveness endpoint
      - Overall
      - Within subgroups
    - Interim alternative analyses of primary endpoints
    - Interim estimates of secondary efficacy / effectiveness endpoints

111

Adaptive Sampling Plans

- At each interim analysis, possibly modify
  - Maximal statistical information
  - Schedule of analyses
  - Conditions for early stopping
  - Randomization ratios
  - Statistical criteria for credible evidence
  - Scientific and statistical hypotheses of interest

112

Adaptive Sampling: Examples

- Response adaptive modification of sample size
  - Proschan & Hunsberger (1995); Cui, Hung, & Wang (1999)
- Response adaptive randomization
  - Play the winner (Zelen, 1979)
- Adaptive enrichment of promising subgroups
  - Wang, Hung & O'Neill (2009)
- Adaptive modification of endpoints, eligibility, dose, …
  - Bauer & Köhne (1994); LD Fisher (1998)

113

Adaptive Sampling: Issues

- How do the newer adaptive approaches relate to the constraint of human experimentation and scientific method?
- Effect of adaptive sampling on trial ethics and efficiency
  - Avoiding unnecessarily exposing subjects to inferior treatments
  - Avoiding unnecessarily inflating the costs (time / money) of RCT
- Effect of adaptive sampling on scientific interpretation
  - Exploratory vs confirmatory clinical trials
- Effect of adaptive sampling on statistical credibility
  - Control of type I error in frequentist analyses
  - Promoting predictive value of "positive" trial results

114

Typical Adaptive Design

- Perform analyses when sample sizes $N_1 ... N_J$
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 ... X_{Nj})$
  - 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 or to modify other aspects of RCT

115

Proschan & Hunsberger

- Worst case type I error of two stage design

$$\alpha_{worst} = 1 - \Phi\left(a_2^{(Z)}\right) + \frac{\exp\left(-\left(a_2^{(Z)}\right)^2 / 2\right)}{4},$$

- Can be more than two times the nominal
  - a_2 = 1.96 gives type I error of 0.0616
  - (Compare to Bonferroni results)

116

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

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

117

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 : $\hat{\theta}_k^* \quad Z_k^* \quad P_k^*$

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

118

Conditional Distribution

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

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

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

119

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)

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

120

Unconditional Distribution

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

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

121

Approaches for Testing

- If modify sample size at second stage (Cui, Hung, & Wang)

$$\tilde{N}_2^* = \tilde{N}_2^*(Z_1) \qquad \tilde{Z}_2^* \text{ incremental statistic with } \tilde{N}_2^*$$

$$\tilde{Z}_2 = \sqrt{\frac{N_1}{N_2}} Z_1 + \sqrt{\frac{N_2^*}{N_2}} \tilde{Z}_2^* \overset{H_0}{\sim} N(0,1)$$

- Equivalently, calculate *Z* statistic as usual and use different critical value

$$\textit{reject } H_0 \iff \tilde{Z}_2 = \sqrt{\frac{N_1}{\tilde{N}_2}} Z_1 + \sqrt{\frac{\tilde{N}_2^*}{\tilde{N}_2}} \tilde{Z}_2^* > b\left(Z_{1,} \tilde{N}_2^*\right)$$

$$b\left(Z_{1,} \tilde{N}_2^*\right) = \frac{1}{\sqrt{\tilde{N}_2^*}} \left[\sqrt{\frac{\tilde{N}_2^*}{N_2^*}} \left(z_{1-\alpha} \sqrt{N_2} - Z_1 \sqrt{N_1} \right) + Z_1 \sqrt{N_1} \right]$$

122

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

123

Topics of Special Interest

- Proportional Hazards
  - Sample size re-estimation
    - General case and in presence of an extreme effect
  - Surrogate information
- Nonproportional hazards
  - Weighted logrank statistics
  - Crossing survival curves

124

Proportional Hazards

Sample Size Re-estimation (SSRE)

Where am I going?

Some investigators desire to modify sample size more flexibly than allowed with GST

125

Example

Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples

Cyrus R. Mehta[1,2], Stuart J. Pocock[3]

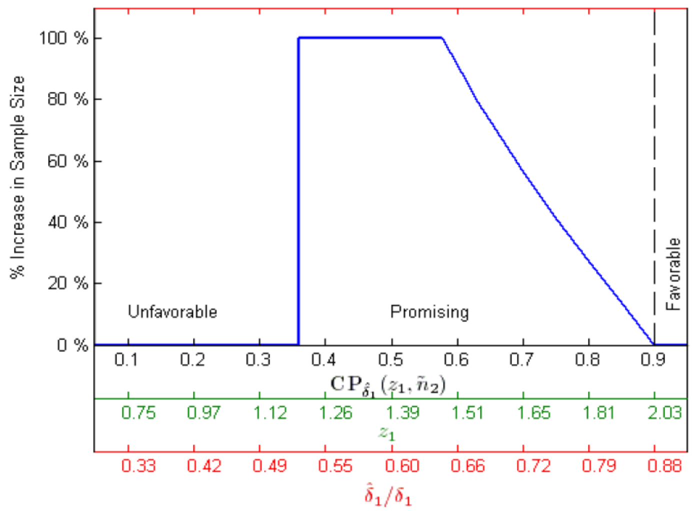
[1] *Cytel Corporation*, [2] *Harvard School of Public Health*, [3] *London School of Hygiene and Tropical Medicine*

SUMMARY

- This paper discusses the benefits and limitations of adaptive sample size re-estimation for phase 3 confirmatory clinical trials. Comparisons are made with more traditional fixed sample and group sequential designs. It is seen that the real benefit of the adaptive approach arises through the ability to invest sample size resources into the trial in stages. The trial starts with a small up-front sample size commitment. Additional sample size resources are committed to the trial only if promising results are obtained at an interim analysis. This strategy is shown through examples of actual trials, one in neurology and one in cardiology, to be more advantageous than the fixed sample or group sequential approaches in certain settings. A major factor that has generated controversy and inhibited more widespread use of these methods has been their reliance on non-standard tests and p-values for preserving the type-1 error. If, however, the sample size is only increased when interim results are promising, one can dispense with these non-standard methods of inference. Therefore, in the spirit of making adaptive increases in trial size more widely appealing and readily implementable we here define those promising circumstances in which a conventional final inference can be performed while preserving the overall type-1 error. Methodological, regulatory and operational issues are examined.

126

Example Modification Plan



127

Comparisons Unconditional Power

Table IV. Operating Characteristics of Fixed Sample and Adaptive Designs

Value of δ	Fixed Sample Design		Plan 4 (Adaptive)	
	Power	Expected SampleSize	Power	Expected Sample Size
1.6	61%	442	65%	499
1.7	66%	442	71%	498
1.8	71%	442	75%	497
1.9	76%	442	79%	494
2.0	80%	442	83%	491
All Plan 4 results are based on 100,000 simulated trials				

128

Comparisons Conditional Power

Table V. Operating Characteristics of the Fixed Sample and Adaptive Designs, Conditional on Interim Outcome

δ	Interim Outcome	Probability of Interim Outcome	Power Conditional on Interim Outcome: Fixed	Power Conditional on Interim Outcome: Adaptive	Expected Sample Size: Fixed	Expected Sample Size: Adaptive
1.6	Unfavorable	36%	30%	30%	442	442
	Promising	23%	62%	82%	442	687
	Favorable	41%	87%	87%	442	442
1.7	Unfavorable	32%	34%	34%	442	442
	Promising	23%	67%	85%	442	685
	Favorable	45%	89%	89%	442	442
1.8	Unfavorable	29%	38%	38%	442	442
	Promising	23%	70%	88%	442	682
	Favorable	49%	91%	91%	442	442
1.9	Unfavorable	26%	43%	43%	442	442
	Promising	22%	74%	90%	442	679
	Favorable	52%	93%	93%	442	442
2.0	Unfavorable	23%	47%	47%	442	442
	Promising	21%	77%	92%	442	678
	Favorable	56%	95%	95%	442	442

All results are based on 100,000 simulated trials

Adaptation to Gain Efficiency?

- Consider adaptation merely to repower study
  - “We observed a result that was not as good as we had anticipated”
- All GST are within family of adaptive designs
  - Don’t we have to be at least as efficient?
- Issues
  - Unspecified adaptations
  - Comparing apples to apples

130

Apples with Apples

- Can adapting beat a GST with the same number of analyses?
  - Fixed sample design: N=1
  - Most efficient symmetric GST with two analyses
    - N = 0.5, 1.18
    - ASN = 0.6854
  - Most efficient adaptive design with two possible N
    - N = 0.5 and either 1.06 or 1.24
    - ASN = 0.6831 (0.34% more efficient)
  - "Most efficient" adaptive design with four possible N
    - N = 0.5 and either 1.01, 1.10, 1.17, or 1.31
    - ASN = 0.6825 (0.42% more efficient)

Table 1: Average and Maximal Sample Sizes of Adaptive Designs in Setting 1

	Number of Continuation Regions							
	1	2	3	4	5	6	7	8
ASN	0.6854	0.6831	0.6828	0.6825	0.6824	0.6824	0.6824	0.6824
% Reduction	Ref	0.34%	0.38%	0.42%	0.43%	0.43%	0.44%	0.44%
Maximal N	1.18	1.24	1.24	1.26	1.26	1.26	1.26	1.28

131

Apples with Apples (continued)

- GST with more analyses?
  - Fixed sample design: N=1
  - Most efficient symmetric GST with two analyses
    - N = 0.5, 1.18
    - ASN = 0.6854
  - GST with same three analyses
    - N = 0.5,1.06 and 1.24
    - ASN = 0.6666 (2.80% more efficient)
  - GST with same five analyses
    - N = 0.5, 1.01, 1.10, 1.17, or 1.31
    - ASN = 0.6576 (4.20% more efficient)

132

Comments re Conditional Power

- Many propose adaptations based on conditional /predictive power
- Neither have good foundational motivation
  - Frequentists should use Neyman-Pearson paradigm and consider optimal unconditional power across alternatives
    - And conditional/predictive power is not a good indicator in loss of unconditional power
  - Bayesians should use posterior distributions for decisions
- Difficulty understanding conditional / predictive power scales can lead to bad choices for designs

133

Comparisons of Designs

- The example used here was a longitudinal study, rather than time to event, though the same issues obtain
- Statistical power
- Sample size accrued
  - With time to event, often all subjects have been accrued when half the statistical information is not yet available
- Calendar time
  - Number of events is more a surrogate for savings in time monitoring subjects and marketing time lost

134

Alternative Approaches

Table 1: Comparison of RCT Designs for Example 1

Design	$\delta = 0$	$\delta = 1.5$	$\delta = 1.6$	$\delta = 1.7$	$\delta = 1.8$	$\delta = 1.9$	$\delta = 2.0$
	Power						
Fxd442	2.5%	55.6%	61.1%	66.3%	71.3%	75.9%	80.0%
Fxd690	2.5%	74.8%	80.0%	84.5%	88.3%	91.4%	93.9%
GST694	2.5%	74.8%	80.0%	84.6%	88.4%	91.4%	93.9%
Adapt	2.5%	60.4%	65.8%	70.8%	75.4%	79.6%	83.4%
Fxd492	2.5%	60.2%	65.8%	71.0%	75.9%	80.2%	84.1%
Fut492	2.5%	59.8%	65.4%	70.6%	75.4%	79.8%	83.7%
OBF492	2.5%	59.6%	65.2%	70.4%	75.3%	79.6%	83.5%
	Expected Number Accrued						
Fxd442	442	442	442	442	442	442	442
Fxd690	690	690	690	690	690	690	690
GST694	694	681	678	675	671	667	662
Adapt	464	496	495	494	492	490	488
Fxd492	492	492	492	492	492	492	492
Fut492	468	488	489	490	490	490	491
OBF492	467	485	485	485	485	484	484

Hypothesized Treatment Effect

135

Alternative Approaches

Table 1: Comparison of RCT Designs for Example 1

Design	$\delta = 0$	$\delta = 1.5$	$\delta = 1.6$	$\delta = 1.7$	$\delta = 1.8$	$\delta = 1.9$	$\delta = 2.0$
	Expected Number Completed						
Fxd442	442	442	442	442	442	442	442
Fxd690	690	690	690	690	690	690	690
GST694	693	668	663	657	649	641	632
Adapt	464	496	495	494	492	490	488
Fxd492	492	492	492	492	492	492	492
Fut492	353	472	475	478	481	483	485
OBF492	352	455	455	454	452	449	445
	Expected Calendar Time (months)						
Fxd442	18.8	18.8	18.8	18.8	18.8	18.8	18.8
Fxd690	25.9	25.9	25.9	25.9	25.9	25.9	25.9
GST694	26.0	25.3	25.1	24.9	24.7	24.5	24.2
Adapt	19.4	20.3	20.3	20.3	20.2	20.1	20.1
Fxd492	20.2	20.2	20.2	20.2	20.2	20.2	20.2
Fut492	16.2	19.6	19.7	19.8	19.9	19.9	20.0
OBF492	16.1	19.1	19.1	19.1	19.0	19.0	18.8

Hypothesized Treatment Effect

136

Alternative Approaches

- The authors plan for adaptation could increase sample size by 100%
- Using their adaptive plan, the probability of continuing until a 25% increase in maximal sample size
  - .064 under null hypothesis
  - .162 if treatment effect is new target of 1.6
  - .142 if treatment effect is old target of 2.0
- By way of contrast
  - A fixed sample test with 11% increase in sample size has same power
  - A group sequential test with 11% increase in maximal sample size has same power and better ASN

137

Apparent Problem

- The authors chose extremely inefficient thresholds for conditional power
  - Adaptation region $0.365 < CP_{est} < 0.8$
  - From optimal test, $0.049 < CP_{est} < 0.8$ is optimal
- Of course, we do not always choose the most efficient designs
  - O'Brien-Fleming designs are markedly inefficient for primary endpoint, but do allow adequate sample size for safety and secondary endpoints
- But more careful evaluation can allow us to choose adaptations that satisfy desired operating characteristics

138

The Cost of Planning Not to Plan

- Hypothesis testing of a null with fully adaptive trials
  - Statistics: type I error is controlled
  - Game theory: chance of "winning" with completely ineffective therapy is controlled
  - Science:
    - Discrimination of clinically relevant hypothesis may be impaired
    - May be ncertainty as to what the treatment has effect on
- Frequentist estimation: (Levin, Emerson, Emerson, 2012)
  - Ideally pre-specify the adaptive rule
    - GST methods can be extended to adaptive sampling density
  - When fully adaptive, Brannath, Mehta, Posch (2009) have proposed a very clever method that works reasonably well.

139

Proportional Hazards

Accrual vs Follow-up in Time to Event Analyses

Where am I going?

We can sometimes gain from unblinded adaptation by changing the number accrued

140

Comparing Sequential Designs: Group Sequential vs. Adaptive

In general setting (when statistical information is proportional to number of subjects) the advantage of adaptive designs is questionable. Claims made for adaptive sequential designs focus on:

- Increased flexibility: fix a trial designed with an incorrect alternative
- Possibly improve efficiency

Comparing Sequential Designs: Group Sequential vs. Adaptive

But it has also been shown that certain forms of adaptive designs are uniformly inferior to standard group sequential designs (Tsiatis and Mehta, *Biometrika*, 2003; Jennison and Turnbull, *Biometrika*, 2006).

Other issues with the use of adaptive designs include (Emerson, 2006):

- Changing scientific question
- Relevance of experimentwise error
- Full inference is difficult or impossible
- Loss of efficiency with statistics not based on sufficient statistics
- Maintaining blind to interim results

Motivating Example

- An industry sponsor desired sample size re-estimation based on estimated treatment effect.
- At an interim analysis, the sample statistic is computed and then used to determine the number and sample sizes of future analyses.
- Claims of designs better addressing trial costs in time-to-event analyses

Question:

- Does sample size vs number of events question modify more general results?
- Is this an issue of what parameters are being constrained?
  - Maximal sample size
  - Number and timing of analyses
  - Conservatism at early analyses

Motivating Example: Adaptive Design

We considered the following design model:

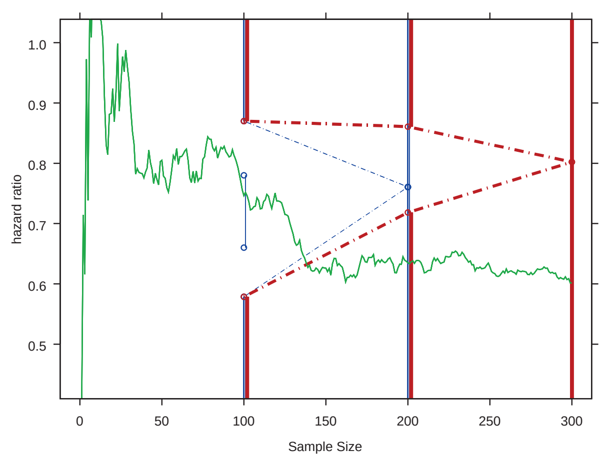
- One-sided test of inferiority at level $\alpha = 0.025$ constructed from two group sequential designs:
  - Design A with analyses at 100 and 200 events
  - Design B with analyses at 100, 200, and 300 events, constructed to have the same stopping boundaries as Design A at the first analysis.

Motivating Example: Adaptive Design

- Choose parameters $A, D : a_1 < A < D < d_1$.
- If the statistic at the first analysis is in the range (a_1, A) or (D, d_1), continue the trial using the Design A stopping rule.
- If the statistic at the first analysis is in the range (A, D), continue the trial using an appropriately adjusted Design B stopping rule:
  - Maintain the type 1 error for adaptive procedure at $\alpha = 0.025$ by adjusting the level of Design B, dependent upon values of A and D, resulting in Design $B^*(A, D)$
- Let ASD(A, D) denote the resulting adaptive design procedure that switches between Design A and Design $B^*(A, D)$.

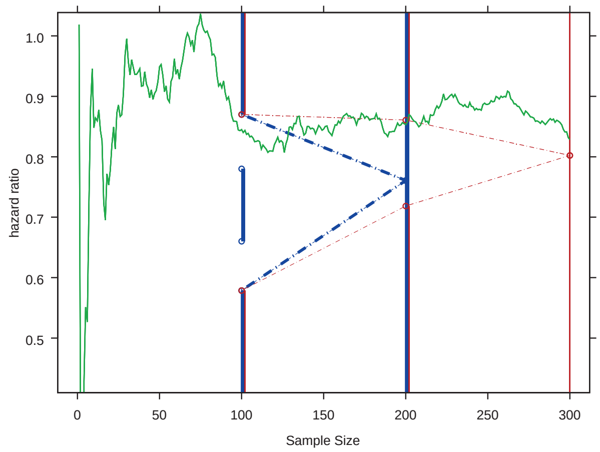
Motivating Example: Adaptive Design

Graphical representation of adaptive stopping rule



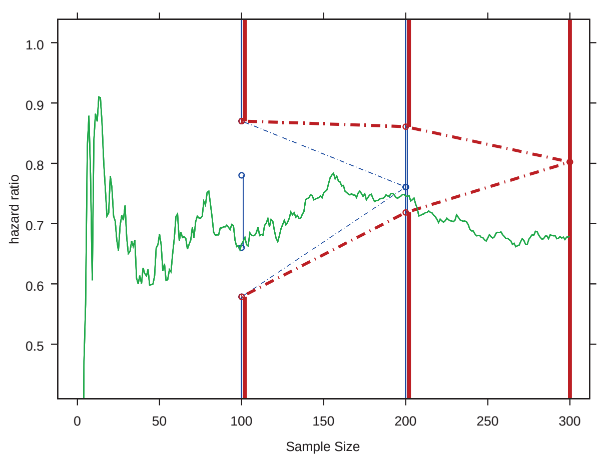
Motivating Example: Adaptive Design

Graphical representation of adaptive stopping rule



Motivating Example: Adaptive Design

Graphical representation of adaptive stopping rule



Finding Hybrid Design

$$P\left(\hat{\theta}_1 \le a_1 \;\middle|\; \theta = 1\right) + P\left(a_1 \le \hat{\theta}_1 \le d_1, \hat{\theta}_2 \le a_2 \;\middle|\; \theta = 1\right) = \alpha.$$

Now, suppose that if we do not terminate the clinical trial at the first analysis, we want to switch to an alternative stopping rule whenever $\hat{\theta}_1$ is observed between pre-specified values of A and D satisfying

$$a_1 \le A \le D \le d_1.$$

If $a_1 < \hat{\theta}(N_1) < A$ or $D < \hat{\theta}(N_1) < d_1$, we will continue to use the sampling plan that specified a maximal sample size of N_2, with a threshold for statistical significance of a_2 at that last analysis.

Finding Hybrid Design

Based on the pre-specified values of A and D, we further prospectively identify a group sequential design $\mathrm{B}^*(A,D)$ having continuation sets $\mathscr{C}_2^* = (a_2^*, d_2^*)$ and $\mathscr{C}_3^* = \emptyset$ for $\hat{\theta}(N_2^*)$ and $\hat{\theta}(N_3^*)$, respectively, computed at analyses performed when the accrued sample sizes are $N_2^* = N_1 + n_2^*$ and $N_3^* = N_2^* + n_3^*$, respectively. Values of $\hat{\theta}(N_2^*) \le a_2^*$ or $\hat{\theta}(N_3^*) \le a_3^*$ will be judged cause to reject the null hypothesis. Hence, we need to pre-specify values of a_2^*, d_2^*, and a_3^* that, when used in conjunction with the group sequential design A and the adaptation pre-specified through the choice of A and D, will preserve the experimentwise error of α:

Finding Hybrid Design

$$
\begin{aligned}
P\left(\text{Reject } H_0 \,\middle|\, \theta = 1\right) \;=\;& P\left(\hat{\theta}(N_1) \le a_1 \,\middle|\, \theta = 1\right) \\
&+ P\left(D \le \hat{\theta}(N_1) < d_1,\, \hat{\theta}(N_2) \le a_2 \,\middle|\, \theta = 1\right) \\
&+ P\left(a_1 < \hat{\theta}(N_1) \le A,\, \hat{\theta}(N_2) \le a_2 \,\middle|\, \theta = 1\right) \\
&+ P\left(A < \hat{\theta}(N_1) < D,\, \hat{\theta}(N_2^*) \le a_2^* \,\middle|\, \theta = 1\right) \\
&+ P\left(A < \hat{\theta}(N_1) < D,\, a_2^* < \hat{\theta}(N_2^*) < d_2^*,\, \hat{\theta}(N_3^*) \le a_3^* \,\middle|\, \right. \\
=\;& \alpha.
\end{aligned}
$$

Finding Hybrid Design

Now using the fact that the original specification of group sequential design A was a level α test, and the fact that $(a_1, A]$, (A, D), and $[D, d_1)$ form a partition of (a_1, d_1), we have that we only need

$$
\begin{aligned}
P\left(\text{Reject } H_0 \,\middle|\, \theta = 1\right) \;=\;& \alpha - P\left(A < \hat{\theta}(N_1) < D,\, \hat{\theta}(N_2) \le a_2 \,\middle|\, \theta = 1\right) \\
&+ P\left(A < \hat{\theta}(N_1) < D,\, \hat{\theta}(N_2^*) \le a_2^* \,\middle|\, \theta = 1\right) \\
&+ P\left(A < \hat{\theta}(N_1) < D,\, a_2^* < \hat{\theta}(N_2^*) < d_2^*,\, \hat{\theta}(N_3^*) \le a_3^* \,\middle|\, \right. \\
=\;& \alpha
\end{aligned}
$$

Finding Hybrid Design

which in turn yields that we only need find a_2^*, d_2^*, and a_3^* to satisfy

$$P\left(A < \hat{\theta}(N_1) < D,\, \hat{\theta}(N_2) \geq d_2 \;\middle|\; \theta = 1\right) = P\left(A < \hat{\theta}(N_1) < D,\, \hat{\theta}(N_2^*) \leq a_2^* \;\middle|\; \theta = 1\right)$$
$$+ P\left(A < \hat{\theta}(N_1) < D,\, a_2^* < \hat{\theta}(N_2^*) < d_2^*\right.$$

Finding Hybrid Design

In particular, we can define a group sequential design using the constrained boundary approach of Burington and Emerson (2003) in which analyses are performed at sample sizes N_1, N_2^*, and N_3^*, the continuation set (a_1^*, d_1^*) at the first analysis is constrained to be $a_1^* = A$ and $d_1^* = D$, and a_2^*, d_2^*, and a_3^* can be chosen as any resulting group sequential design that has type I error of

$$\alpha^* = P\left(\hat{\theta}_1 \leq A \;\middle|\; \theta = 1\right) + P\left(A < \hat{\theta}(N_1) < D,\, \hat{\theta}(N_2) \leq a_2 \;\middle|\; \theta = 1\right).$$

Any such choice will thus preserve an experimentwise error of α for the adaptive procedure.

Sampling Density

For a particular value of μ, we desire to find the sampling density $p(m,s;\mu)$ for the test statistic $(M = m, S = s)$, $m = 1,\ldots,J$, $s \in (-\infty,\infty)$, as defined by eqn (**??**). This can be shown to be (Armitage, McPherson, and Rowe, 1969)

$$p(m,s;\mu) = \begin{cases} f(m,s;\mu) & s \notin \mathscr{C}_{Sm}, \text{and} \\ 0 & \text{else} \end{cases} \tag{1}$$

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

$$\begin{aligned} f(1,s;\mu) &= \frac{1}{\sqrt{n_1}\sigma}\phi\left(\frac{s-n_1\mu}{\sqrt{n_1}\sigma}\right) \\ f(j,s;\mu) &= \int_{\mathscr{C}_{S(j-1)}} \frac{1}{\sqrt{n_j}\sigma}\phi\left(\frac{s-u-n_j\mu}{\sqrt{n_j}\sigma}\right) f(j-1,u;\mu)\,du, \quad j = 2,\ldots,m \end{aligned} \tag{2}$$

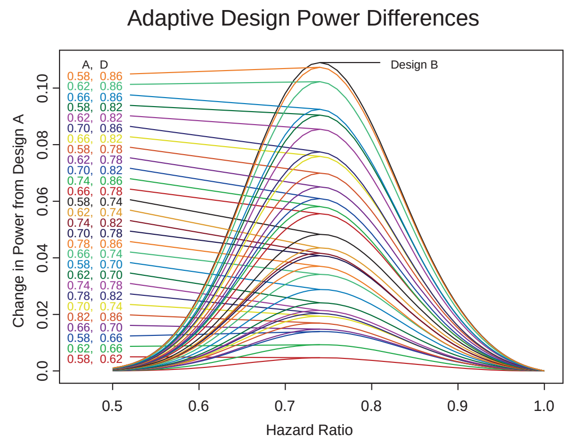
where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the density for the standard normal distribution and $n_1 = N_1$ and $n_j = N_j - N_{j-1}$ for $j = 2,\ldots,J$ denote the size of the groups accrued between successive analyses.

Motivating Example: Adaptive Design

- To explore the spectrum of behavior exhibited by this family of adaptive designs, we consider a full range of A and D values.
- In practice, A and D might be chosen based on desired operating charactistics, symmetry constraints, or conditional power considerations
- The boundary at analysis time 1 is $a_1 = 0.5792, d_1 = 0.8645$, so we present results for $A, D \in \{0.58, 0.62, 0.66, 0.70, 0.74, 0.78, 0.82, 0.86\}$ with the constraint that $A < D$.
- Note that if $A = D$, the adaptive design reduces to Design A, and if $A = a_1, D = d_1$ the adaptive design reduces to Design B.

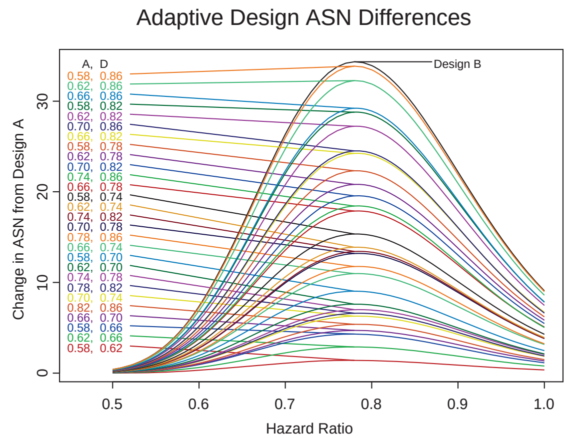
Standard group sequential software can be used to calculate operating characteristics of this design.

Adaptive Design: Power Differences



Increase in power over Design A for the adaptive designs considered.

Adaptive Design: ASN Differences



Increase in ASN over Design A for the adaptive designs considered.

Comparing Sequential Designs

In comparing different types of sequential designs, we must select criteria that we wish to constrain or optimize. Possibilities include:

- Maximal possible sample size
- Average sample size (ASN) across a range of alternatives
- Power across a range of alternatives
- Probability of using more than S subjects
- Median sample size (or any other quantile)

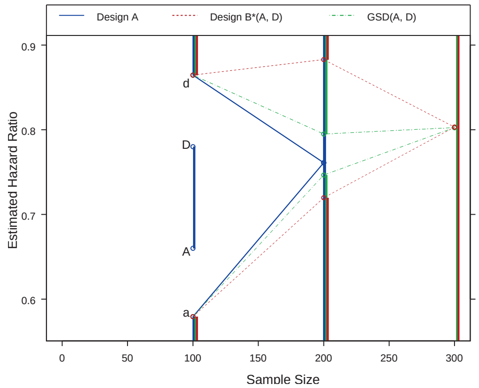
We will focus on ASN and power, with the additional constraints of equal maximal sample size and matching boundaries at the first analysis.

Motivating Example: Matching Group Sequential Designs

We attempted to identify group sequential designs that closely matched the power curve of the adaptive design. We discovered that this could be accomplished by:

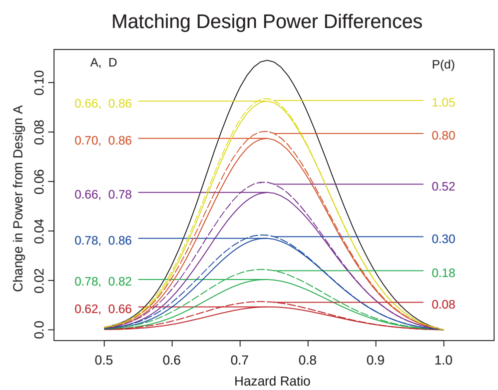
- Group sequential design with 3 analyses at 100, 200, and 300 subjects
- Match the Design B stopping boundary at analyses 1 and 3
- Vary the shape of the design by choosing the P parameter in the unified family of group sequential designs ($P = 0.5$ results in a Pocock design; $P = 1$ produces an O'Brien-Fleming design):

Motivating Example: Matching Group Sequential Designs



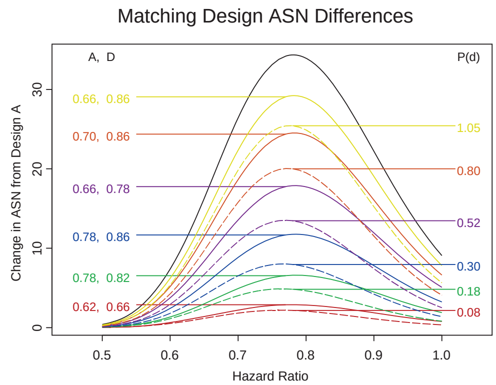
Stopping boundaries for adaptive and group sequential designs. Design A is shown in blue; Design $B^*(A,D)$ is shown in red. The green boundary is the group sequential design GSD(A,D) that matches the adaptive design.

Motivating Example: Matching Group Sequential Designs



Change in power (increase from power of Design A) curves for a selection of adaptive designs along with corresponding group sequential design. Numbers on the left are values of A and D for the adaptive design; numbers on the right are the values of the P parameter for the d boundary of the group sequential design.

Motivating Example: Matching Group Sequential Designs



Change in ASN (increase from ASN of Design A) curves for a selection of adaptive designs along with corresponding group sequential design. Numbers on the left are values of A and D for the adaptive design; numbers on the right are the values of the P parameter for the d boundary of the group sequential design.

Motivating Example: Matching Group Sequential Designs

- In each matched pair (ASD(A, D) and GSD(A, D)) the group sequential design has equal or slightly higher power than the adaptive design.
- In an uncensored setting it is clear that this adaptive design offers no benefit.
- Similar results are found for trials involving a normal mean endpoint.

Survival Analysis Setting

Issues involved in sequential trials are more complicated for time-to-event endpoint:

- Statistical information proportional to number of events observed rather than number of subjects
- Subjects accrued to trial for a certain amount of time: the accrual period of length t
- After accrual ends there may be a follow-up period of length t_F, when no more subjects are allowed to join the trial, but current subjects are still being followed
- Subjects followed until an event is observed, until they drop out, or until the trial ends
- Length of the trial and the number of subjects accrued are efficiency concerns, along with power and ASN (ASN in this case is the average number of events)

Survival Analysis: Group Sequential vs. Adaptive Designs

- Adaptive designs that limit the maximal possible sample size may offer unique benefits in the survival analysis setting.
- If it is known at an early point in the trial that the number of events needed will be limited to a certain number, accrual of subjects may be stopped early.
- This has the potential to reduce trial costs by limiting number of patients involved as compared to a group sequential design, even when the group sequential design requires fewer events on average.

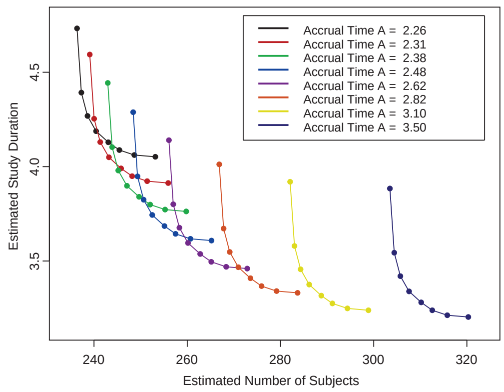
Motivating Example: Time and Number of Subject Calculations

- To compare the cost behavior of the adaptive designs, we fixed values for A, D, median event rate on control arm, and accrual rate.
- Accrual model:
  - Uniform accrual over accrual time period at rate r per unit time
  - No drop-out
  - Exponential event rates in both arms

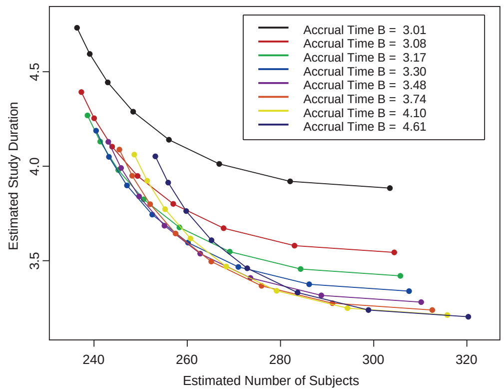
Motivating Example: Time and Number of Subject Calculations

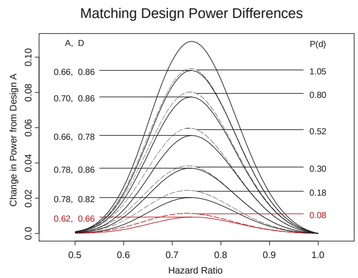
- We must consider two independent accrual times t_A and t_B depending on whether Design A or Design B is adaptively chosen.
- Each combination of t_A and t_B produces an estimated number of subjects and trial duration. Factors involved in choosing t_A and t_B:
  - Each of t_A and t_B must be larger than the time of the first analysis. We consider the case of accrual finishing before the first analysis separately.
  - t_A and t_B are required to be large enough to obtain at least 200 and 300 subjects respectively.
  - The maximum accrual time considered for each design was chosen to be the accrual time that resulted in zero follow-up time after the end of accrual for that design.

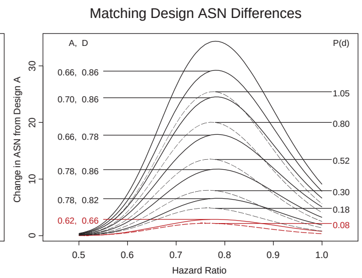
Motivating Example: Time and Number of Subject Calculations

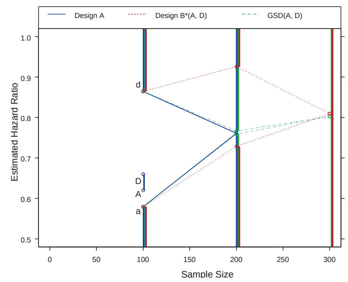


Motivating Example: Time and Number of Subject Calculations



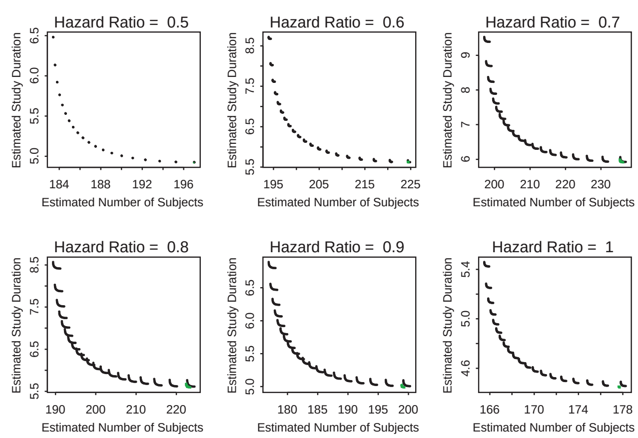






Motivating Example: Accrual Results

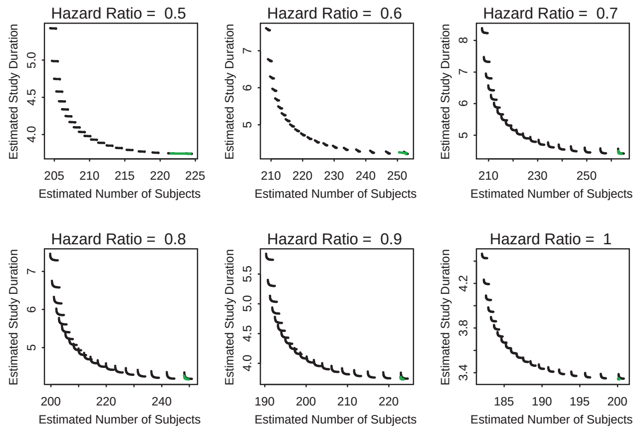
A = 0.62, D = 0.66, Control Median = 1, Accrual Rate = 40



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results

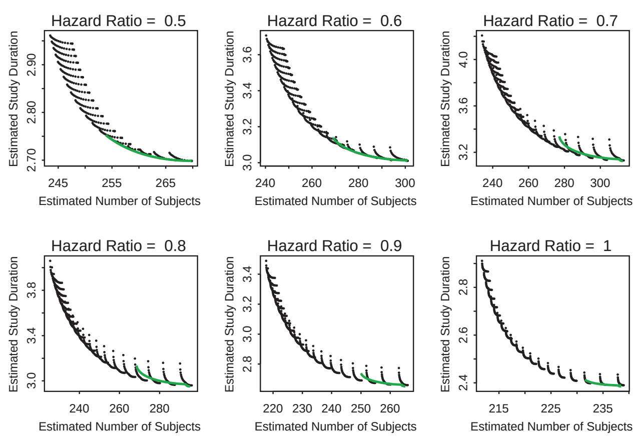
A = 0.62, D = 0.66, Control Median = 1, Accrual Rate = 60



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results

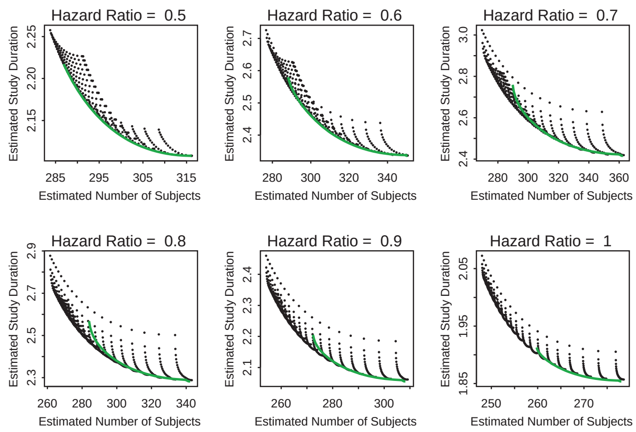
A = 0.62, D = 0.66, Control Median = 1, Accrual Rate = 100



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results

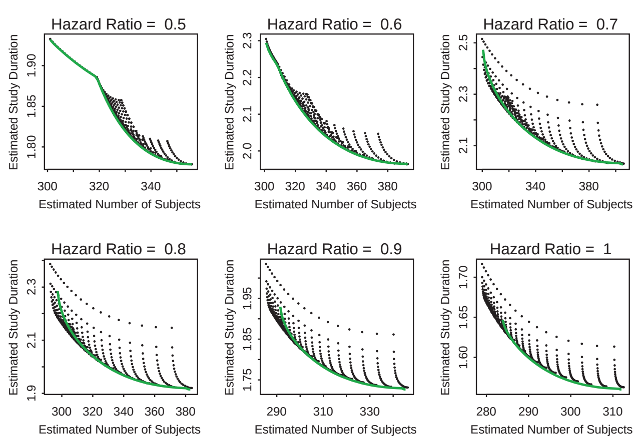
A = 0.62, D = 0.66, Control Median = 1, Accrual Rate = 150



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results

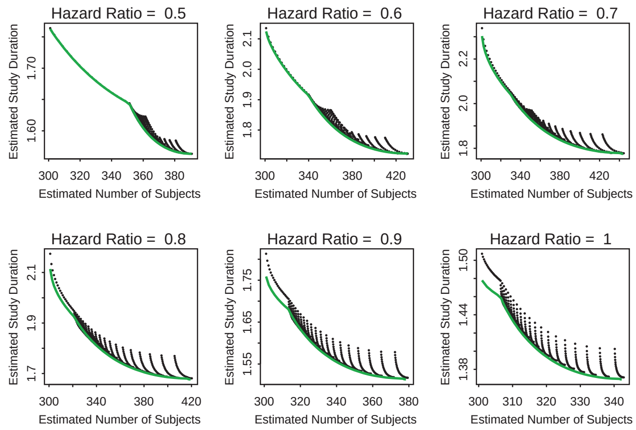
A = 0.62, D = 0.66, Control Median = 1, Accrual Rate = 200



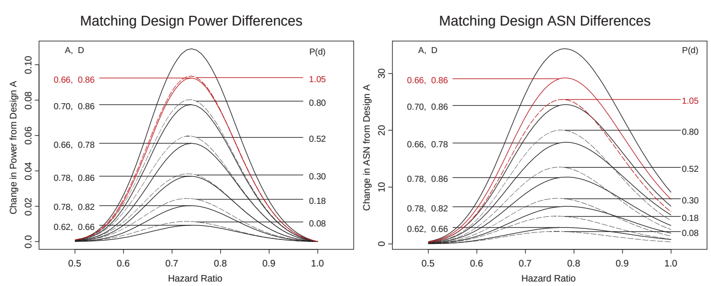
The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

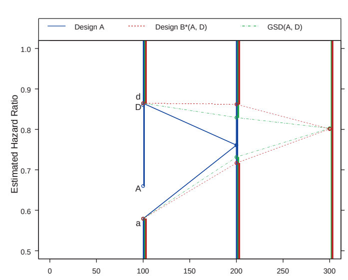
Motivating Example: Accrual Results

A = 0.62, D = 0.66, Control Median = 1, Accrual Rate = 250

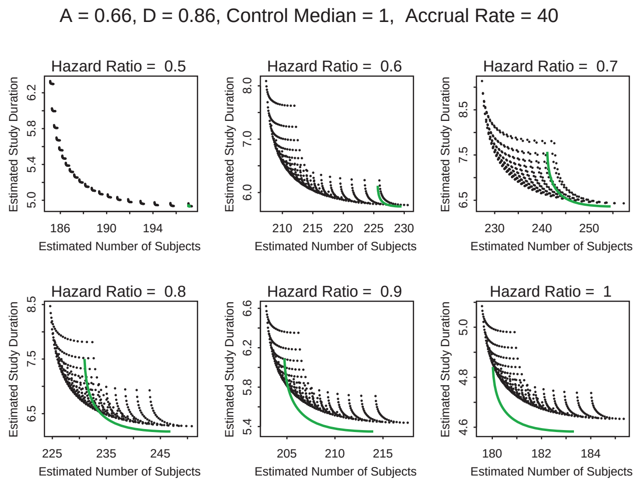


The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.



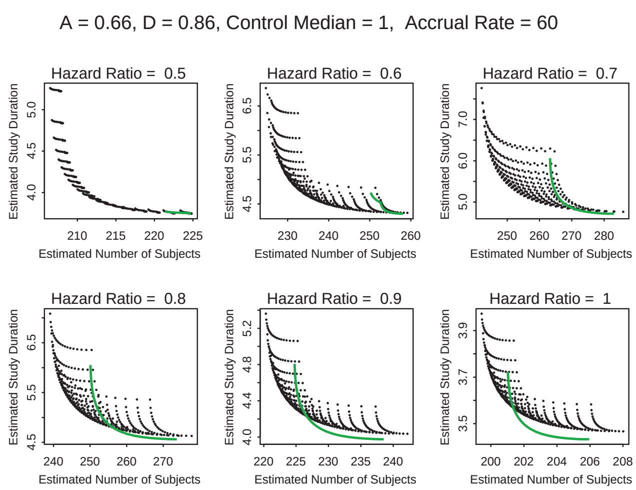


Motivating Example: Accrual Results



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

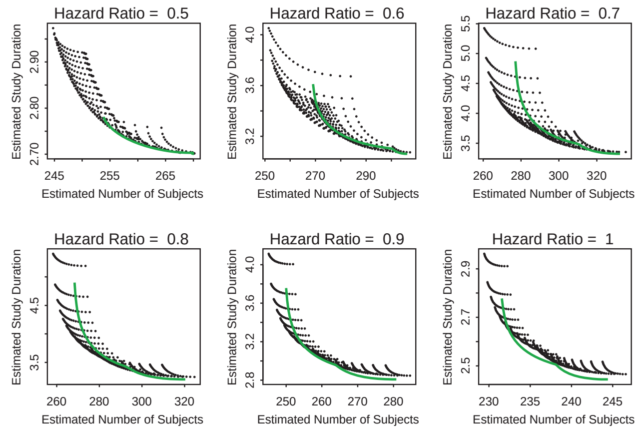
Motivating Example: Accrual Results



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results

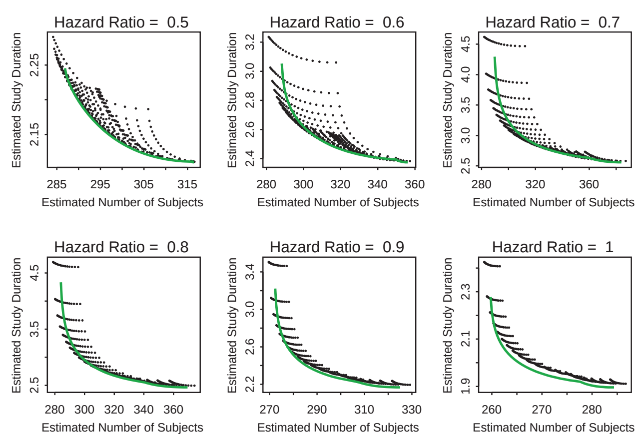
A = 0.66, D = 0.86, Control Median = 1, Accrual Rate = 100



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

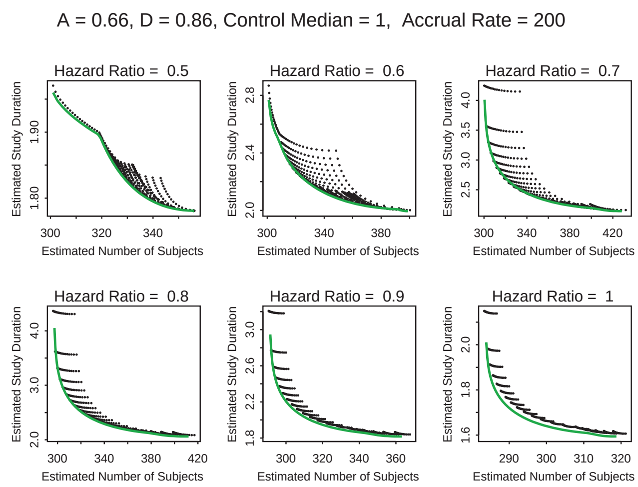
Motivating Example: Accrual Results

A = 0.66, D = 0.86, Control Median = 1, Accrual Rate = 150



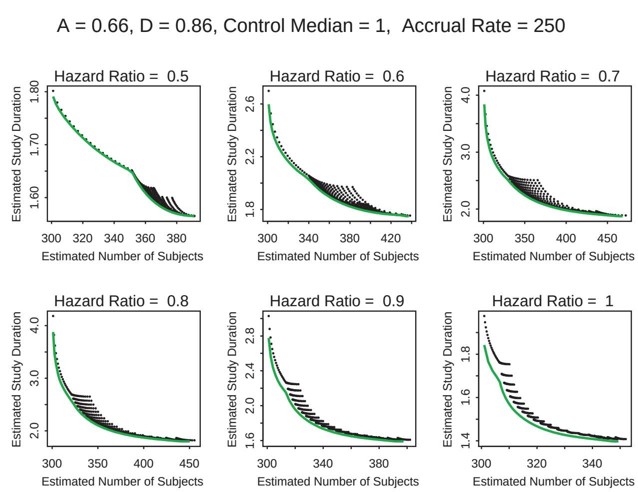
The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results



The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

Motivating Example: Accrual Results

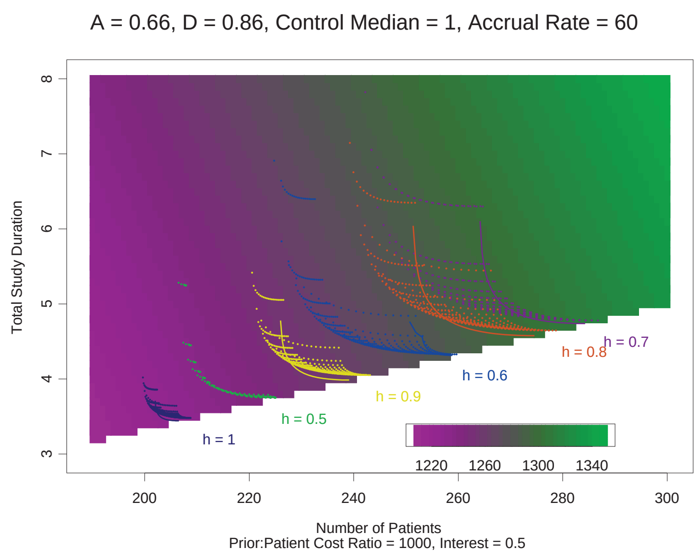


The adaptive design ASD(A, D) is represented by the black dots, and the matching group sequential design GSD(A, D) is represented by the green line.

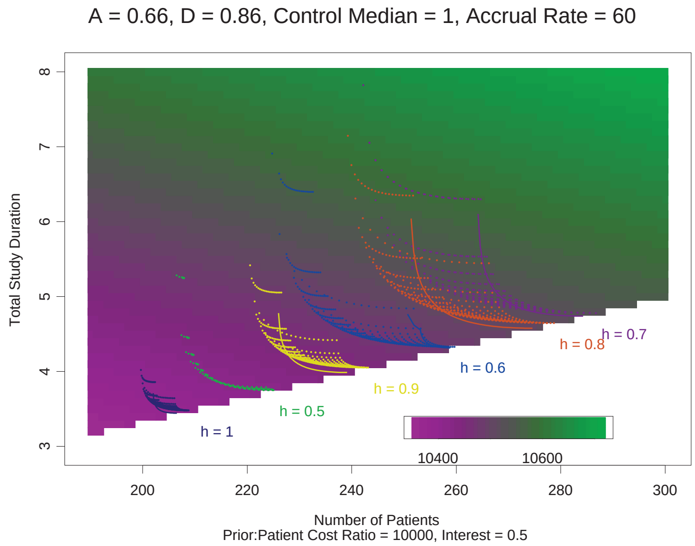
Motivating Example: Cost Estimation

- To explore tradeoffs between sample size and study duration we introduce a simple model for trial costs
- Trial costs based on a per-patient cost p and a prior or start-up cost c.
- Duration costs incorporated with an interest rate ω per unit of time, calculated based on total trial costs so far.
  - Surrogate for all time-related costs such as maintaining databases, personnel, and the cost of money.
- Interested in ratio of minimal group sequential design cost to minimal adaptive design cost, over a range of parameter values, so cost presented in standardized units.

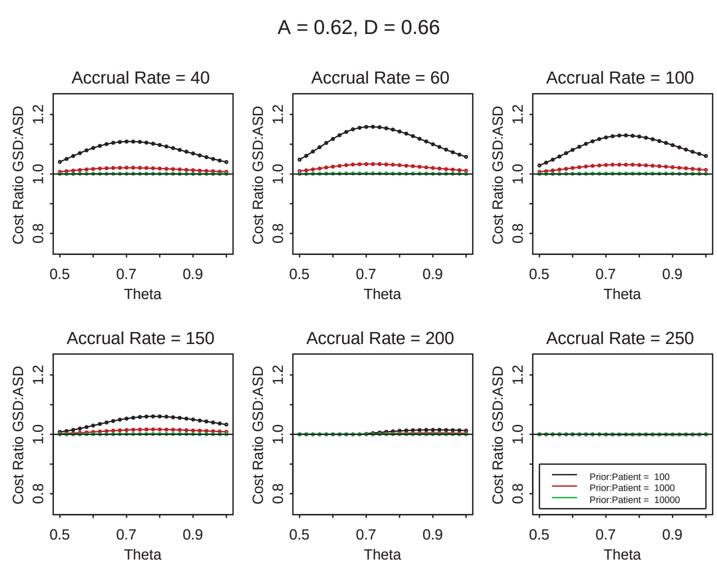
Motivating Example: Cost Estimation Results



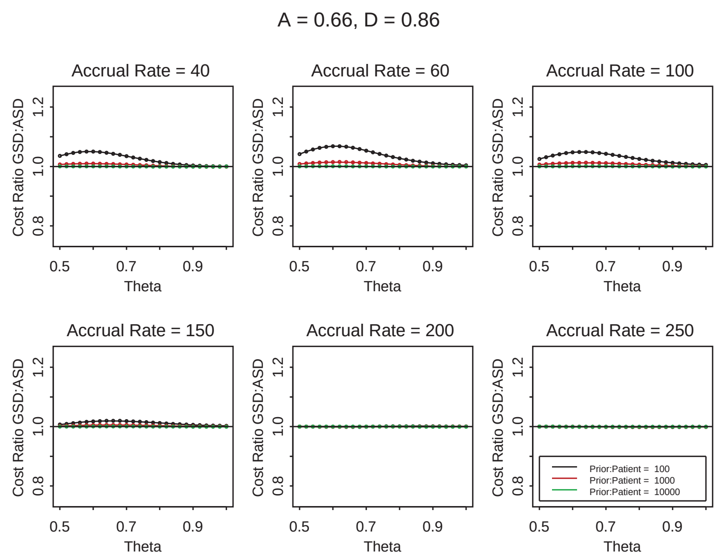
Motivating Example: Cost Estimation Results



Motivating Example: Cost Estimation Results



Motivating Example: Cost Estimation Results

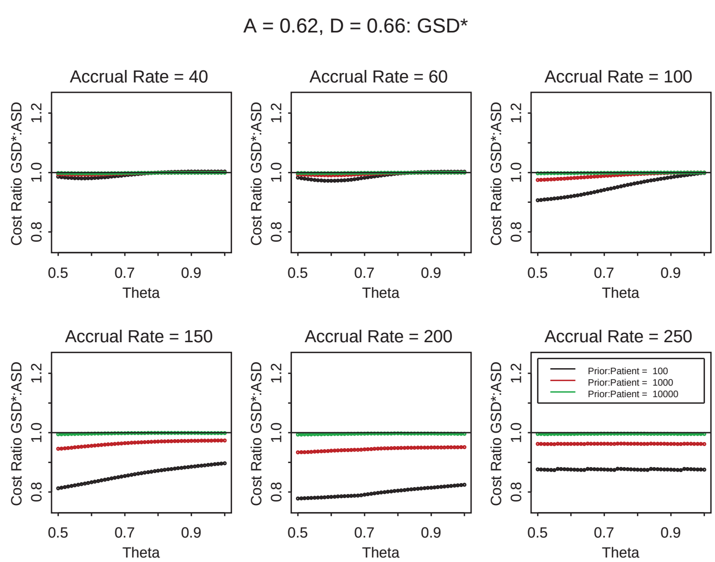


Discussion

- It is clear that this adaptive design offers no advantages over the matched group sequential design in uncensored settings
- There are some cost advantages to the adaptive design in certain survival analysis settings.
- The degree of benefit depends on the distributions of event times and accrual rate as well as on the particular adaptive design under consideration.
- This suggests it may therefore be worth considering the cost-effectiveness of using such an adaptive design in time-to-event endpoints.
- There remain difficulties with inference following the use of such an adaptive design, so in cases where there is questionable or insignificant gain from the adaptive design it may be wiser to continue to use a standard group sequential design.

Relaxing Constraints

- We also considered relaxing the constraint that the adaptive and group sequential designs should have the same maximal sample size, but maintaining the boundary at the first analysis.
  - Explored a range of possible maximal sample sizes from 205 to 300 in steps of 5
  - Explored a range of P parameters
- We found that a design GSD$^*(0.62, 0.66)$ with analyses at 100, 155, and 210 events, with $P = (P_a, P_d) = (1.3, 1.3)$ matched the power curve of ASD$(0.62, 0.66)$ while dramatically improving ASN.
- Cost ratio plots comparing GSD$^*(0.62, 0.66)$ to ASD$(0.62, 0.66)$ show a significant reduction in cost using the group sequential design.



Bottom Line

- The dimensionality of stopping rules is quite high:
  - Number and timing of analyses
  - Number of boundaries allowing early stopping
  - Degree of early conservatism
  - Type I and 2 errors
- Often comparisons made between standard group sequential designs and adaptive designs unnecessarily constrain parameters
- Efficiency plays a role, but many other issues need to be considered
- Full evaluation of designs is all important

"You better think (think), think about what you are trying to do..."
- Aretha Franklin

Proportional Hazards

SSRE with Extreme Treatment Effects

Where am I going?

Design of a RCT is based on a variety of assumptions that may not obtain in practice

Investigators then may have an interest in adjusting the RCT design to better address the actual conditions

141

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

142

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

143

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

144

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

145

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

146

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

147

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

148

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

149

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 end of study

150

Group Sequential Design

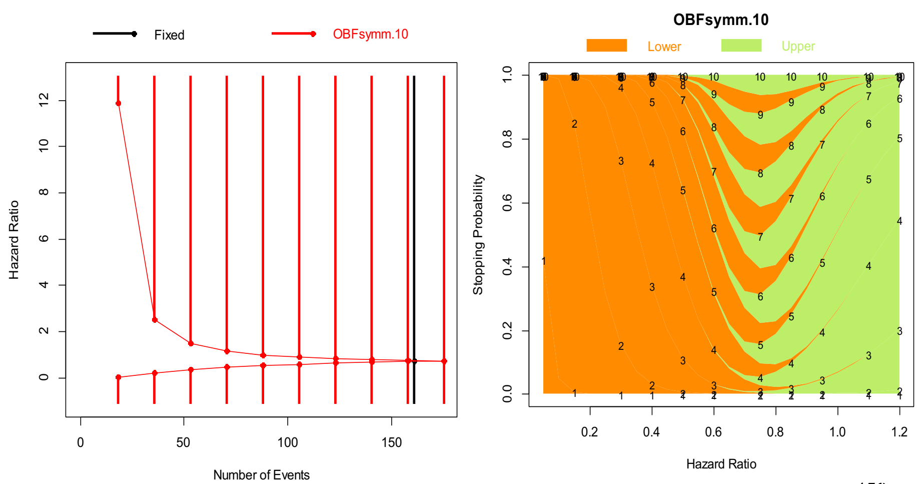
- Instead of a fixed sample design, pre-specify a group sequential design with, say, 10 possible analyses
  - Example: level 0.025, 90% power to detect HR=0.6

```
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
```

151

Group Sequential Design

- Stopping boundaries, stopping probabilities



152

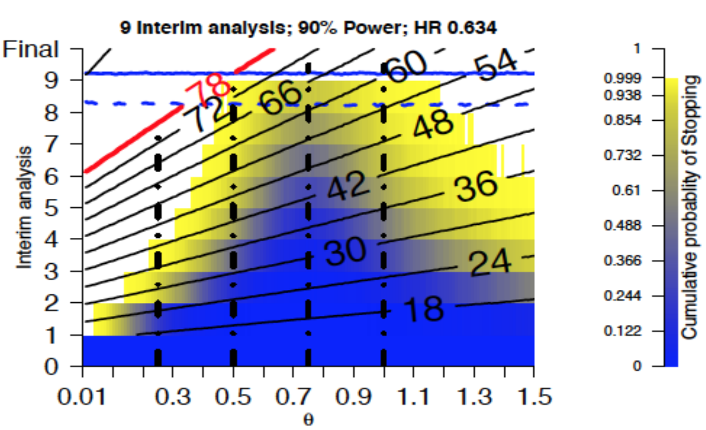
Group Sequential Design

- 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

153

Calendar Time: Half Event Rate

- Stopping probabilities under planned event rate



154

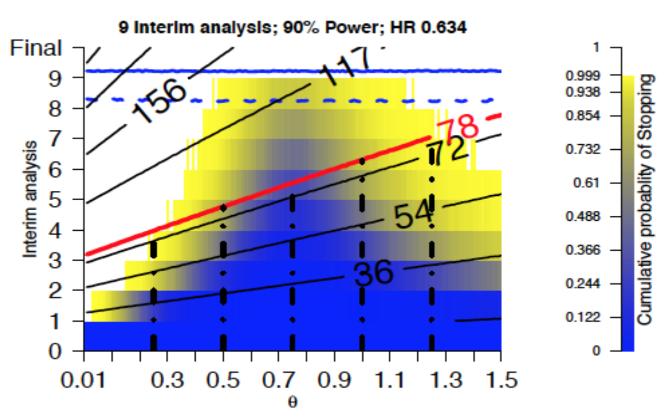
Incorporating Lower Event Rates

- 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.”

155

Calendar Time: Half Event Rate

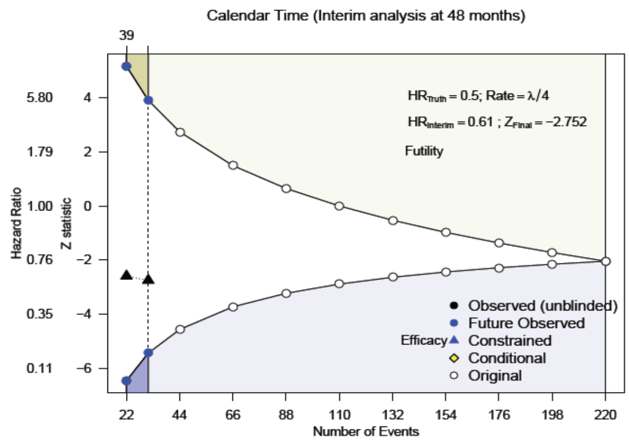
- If control group event rate is halved
  - Power is affected relatively little



156

The Escape Clause

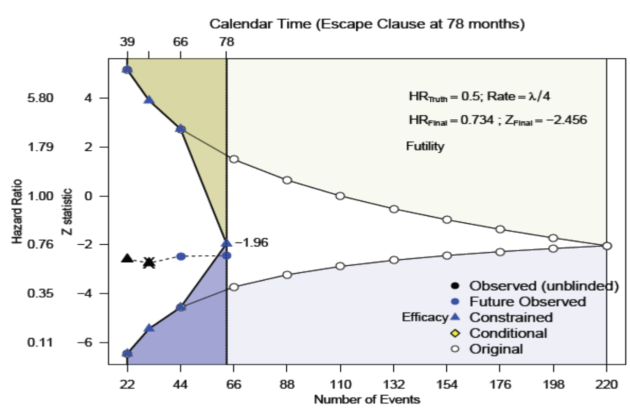
- Prior to pre-specified maximal calendar time, perform group sequential test as usual



157

The Escape Clause

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



158

Terminate for efficacy at 78 months

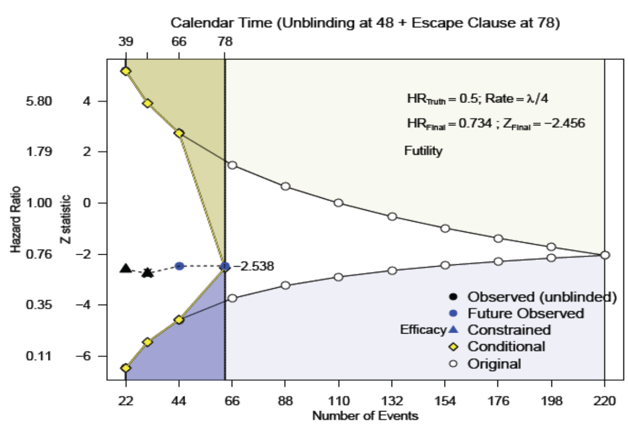
Unblinded Adaptation

- 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

159

What if Unblinded?

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



160

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

162

Proportional Hazards

Availability of Surrogate Data

Where am I going?

Methods for preserving type 1 errors presume an accurate representation of the statistical information available at the adaptive analysis

With time to event data (as well as other longitudinal endpoints), however, we may have information on surrogate prognostic endpoints.

To the extent that those surrogate endpoints inform the adaptation of the clinical trial, we may not be adequately preserving the type 1 error

163

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?

164

Approaches for Testing

- If modify sample size at second stage (Cui, Hung, & Wang)

$$\tilde{N}_2^* = \tilde{N}_2^*(Z_1) \qquad \tilde{Z}_2^* \text{ incremental statistic with } \tilde{N}_2^*$$

$$\tilde{Z}_2 = \sqrt{\frac{N_1}{N_2}} Z_1 + \sqrt{\frac{N_2^*}{N_2}} \tilde{Z}_2^* \overset{H_0}{\sim} N(0,1)$$

- Equivalently, calculate *Z* statistic as usual and use different critical value

$$reject\ H_0 \quad \Leftrightarrow \quad \tilde{Z}_2 = \sqrt{\frac{N_1}{\tilde{N}_2}} Z_1 + \sqrt{\frac{\tilde{N}_2^*}{\tilde{N}_2}} \tilde{Z}_2^* > b(Z_1, \tilde{N}_2^*)$$

$$b(Z_1, \tilde{N}_2^*) = \frac{1}{\sqrt{\tilde{N}_2^*}} \left[\sqrt{\frac{\tilde{N}_2^*}{N_2^*}} \left(z_{1-\alpha} \sqrt{N_2} - Z_1 \sqrt{N_1} \right) + Z_1 \sqrt{N_1} \right]$$

165

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^*	

166

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

167

Estimands by Stage: Time to Event

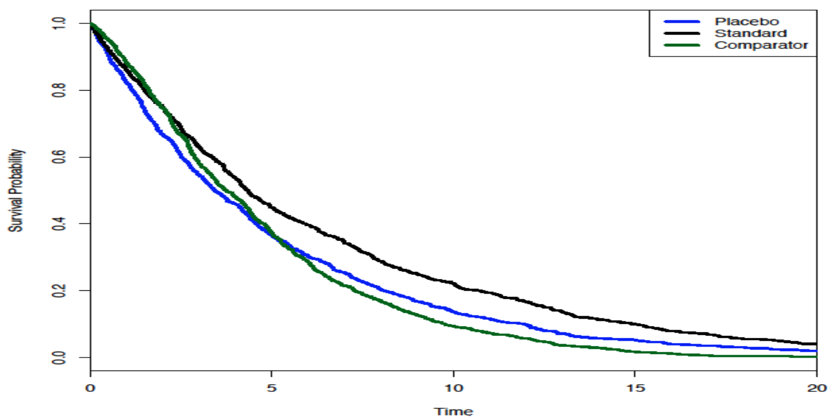
- In time to event data, a common treatment effect across stages is reasonable under some assumptions
  - Strong null hypothesis (exact equality of distributions)
  - Strong parametric or semi-parametric assumptions

- 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)
    - E.g., noninferiority trials

168

Impact on Noninferiority Trials

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



169

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

170

Protecting Type I Error

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

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

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

171

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
  - E.g., tumor progression and overall survival
- This can be a large problem if we allow adaptation to a much smaller sample size
  - Data quite often becomes available between database lock and a DSMB meetin

172

Complications: Longitudinal Outcomes

- 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

173

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^*	

174

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 kth 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)

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

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

175

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

176

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

177

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

178

Nonproportional Hazards

Weighted Logrank Statistics

Where am I going?

Early phase clinical trials sometimes show treatment effects that are more pronounced early or more pronounced late

Weighted versions of the logrank statistic have been proposed to accentuate those portions of the survival curve that are most plausibly different

179

Nonproportional Hazards

Adaptation with Weighted Logrank Statistics

Where am I going?

Early phase clinical trials sometimes show treatment effects that are more pronounced early or more pronounced late

Weighted versions of the logrank statistic have been proposed to accentuate those portions of the survival curve that are most plausibly different

180

Sequential and Adaptive Analysis with Time-to-Event Endpoints

Time-Varying Treatment Effects

Presented July 25, 2019

Scott S. Emerson
Department of Biostatistics
University of Washington

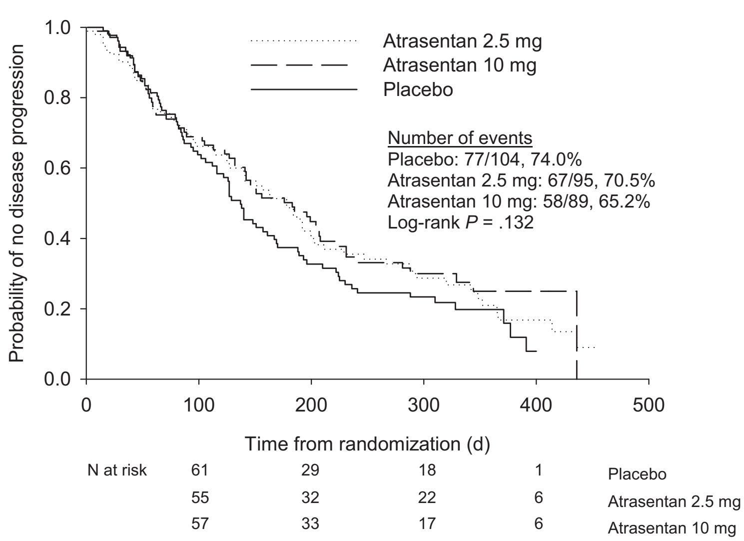
Daniel L. Gillen
Department of Statistics
University of California, Irvine

SISCR

Motivating example

Atrasentan for the treatment of hormone-refractory prostate cancer

- Phase II results for time to progression of disease



SISCR

Motivating example

Atrasentan for the treatment of hormone-refractory prostate cancer

- From the ODAC briefing document:

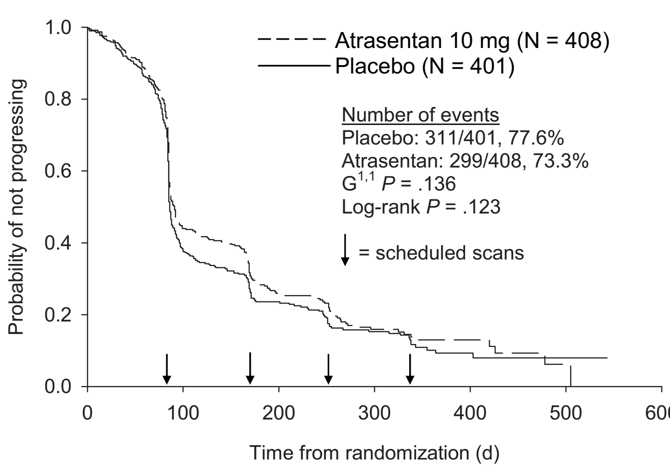
  “In study M96-594, an exploratory analysis of time to disease progression had been performed using the $G^{1,1}$ test statistic, a variant of the log-rank test described by Fleming et al. The $G^{1,1}$ test statistic reduces the weight given to events that occur very early or very late in time-to-progression 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
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

Atrasentan for the treatment of hormone-refractory prostate cancer

- Phase III results for time to progression of disease



N at risk						
	140	68	31	9	3	Placebo
	164	73	26	10	3	Atrasentan

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

SISCR

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

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

$$\begin{aligned} M_i &= \text{number of subjects initially at risk in group } i, \quad i = 01 \\ Y_i(t) &= \text{number of subjects at risk in group } i \text{ at time } t \\ N_i(t) &= \text{the counting process for group } i \text{ at time } t \end{aligned}$$

The logrank statistic

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 $w(t) = \frac{Y_1(t)Y_0(t)}{Y_1(t)+Y_0(t)}$

SISCR

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

SISCR

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)

SISCR

The logrank statistic

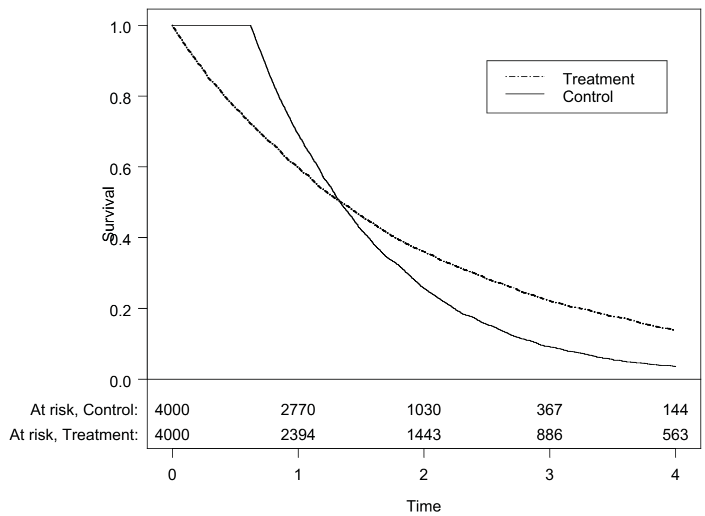
- Under proportional hazards
  - Terms composing the logrank statistic are roughly constant (in a neighborhood of the null hypothesis of equal hazards)
- Under nonproportional hazards
  - Differences in hazards (likely to) change with time
  - As the weights change, what we are estimating/testing changes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

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

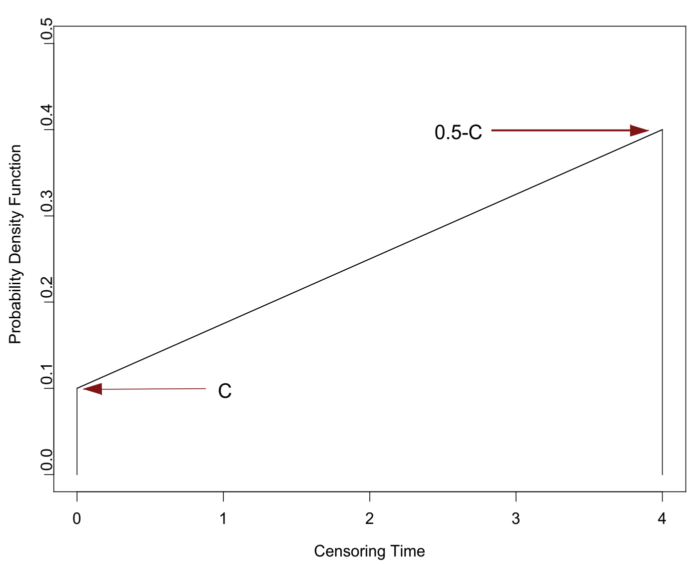
- Grossly exaggerated depiction of a non-proportional hazards treatment effect in the absence of censoring



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



SISCR

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

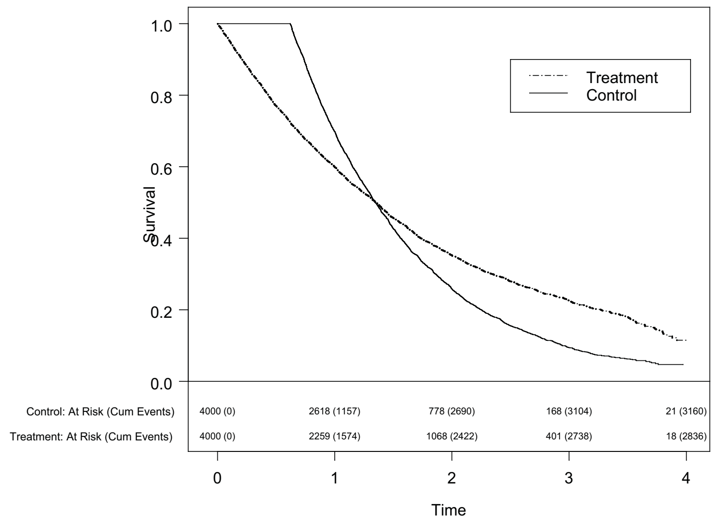
Ex: Sensitivity of operating characteristics to the censoring distribution

RCTdesign implementation of group sequential rules

The logrank statistic

Example 1: Sensitivity to the censoring distribution

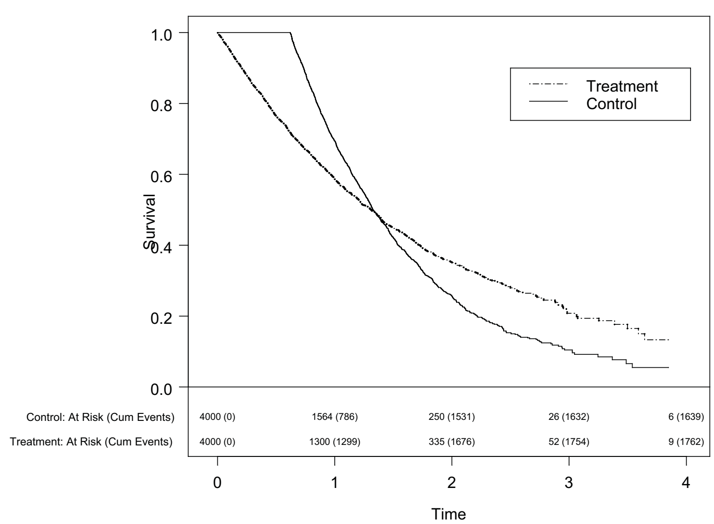
- Estimated survival curves when $C = 0$ (heavy early accrual)



The logrank statistic

Example 1: Sensitivity to the censoring distribution

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



SISCR

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

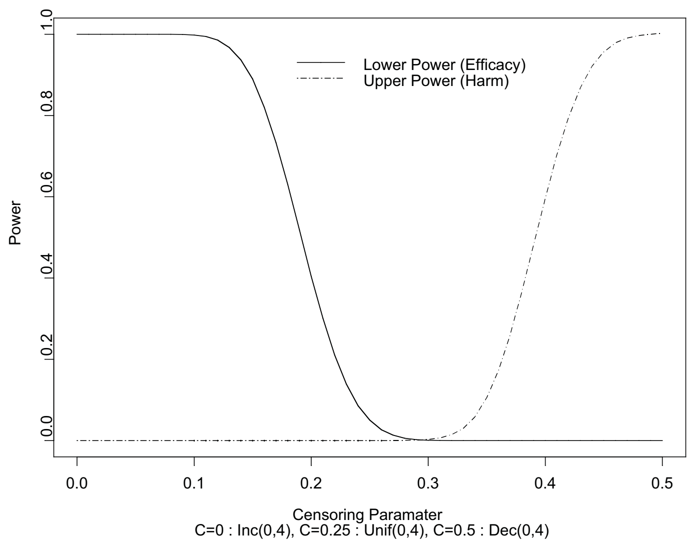
Information growth for weighted LR statistics

Ex: Sensitivity of operating characteristics to the censoring distribution

RCTdesign implementation of group sequential rules

Example 1: Sensitivity to the censoring distribution

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



The logrank statistic

SISCR

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?

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

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

SISCR

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

- 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

The logrank statistic

Example 2: Sensitivity to the censoring distribution

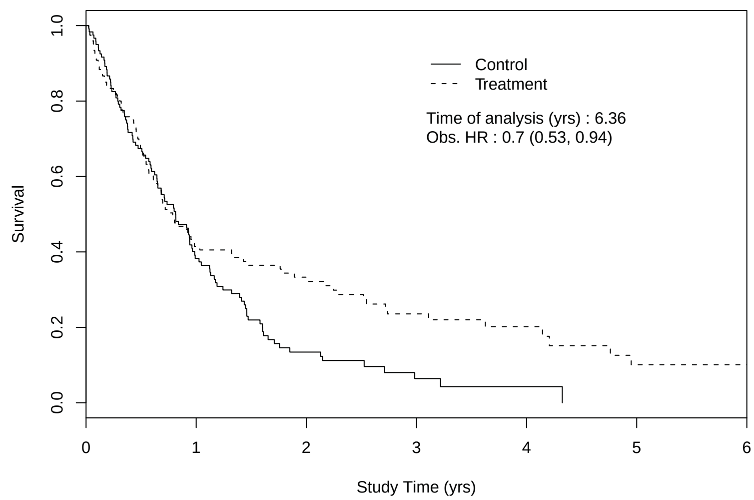
- From the above, if we assume a uniform accrual pattern we have:

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

The logrank statistic

Example 2: Sensitivity to the censoring distribution

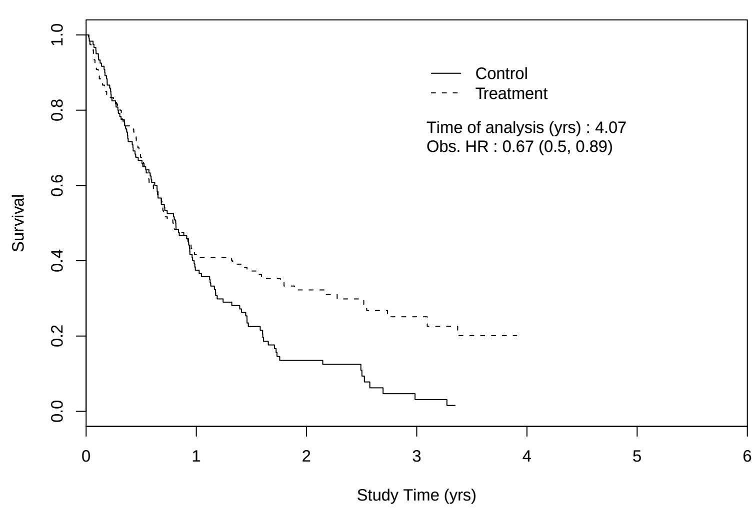
- Accrual of 40 patients per year for 6 years
  - 196th event occurs at 6.36 yrs after first enrollment
  - HR estimate of 0.70 (0.53,0.94)



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)



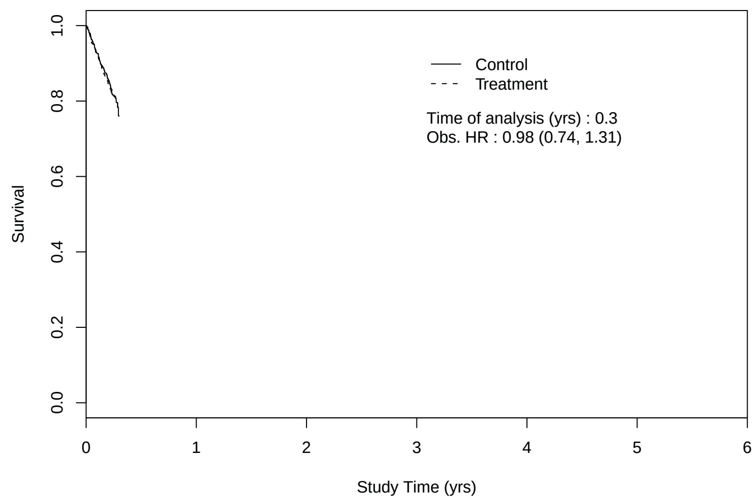
SISCR

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
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

$G^{\rho,\gamma}$ 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

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

SISCR

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

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

SISCR

Evaluation of designs when testing with a WLR statistic

`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

SISCR

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

SISCR

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

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 $\vec{\delta} = (\vec{\delta}_0, \vec{\delta}_1)$.

Output from `seqOCWLR()`

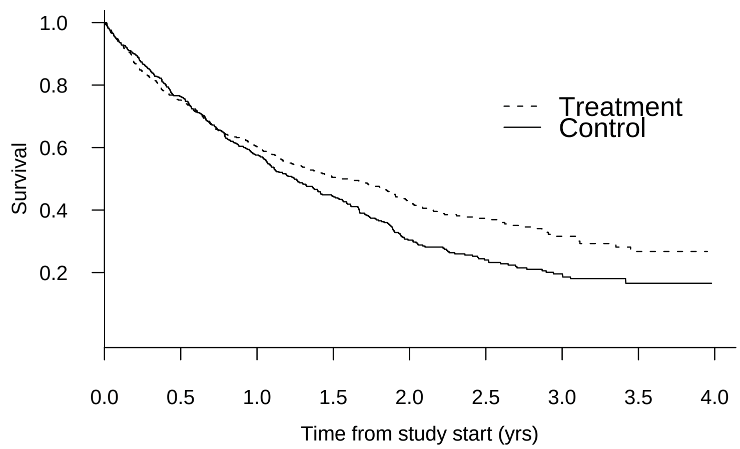
Output from `seqOCWLR()`

- `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

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 $A = 0.3$

SISCR

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

Output from `seqOCWLR()`

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

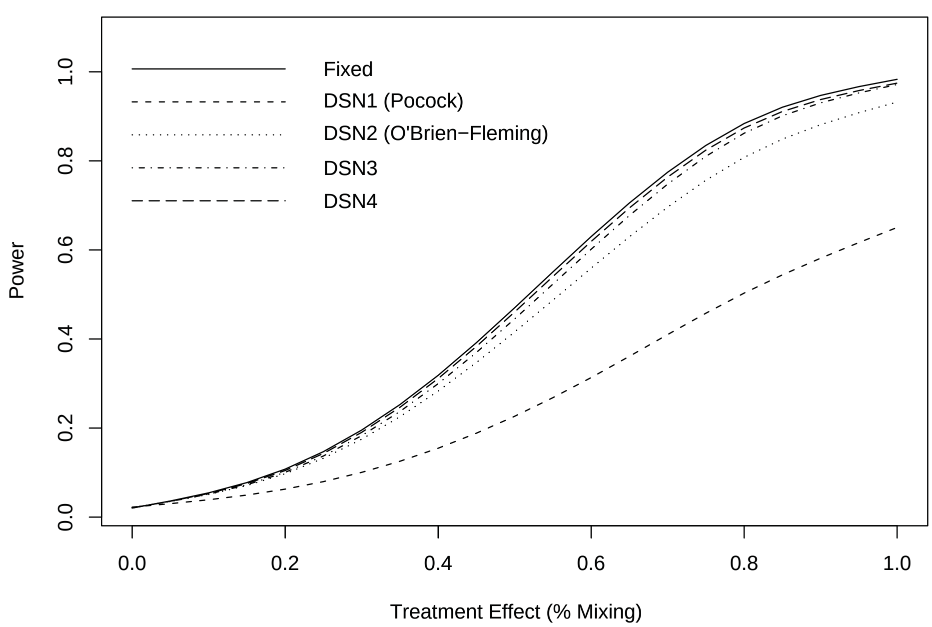
- Potential point estimates that could be observed on the boundary of a symmetric O'Brien-Fleming design `(DSN1)`

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



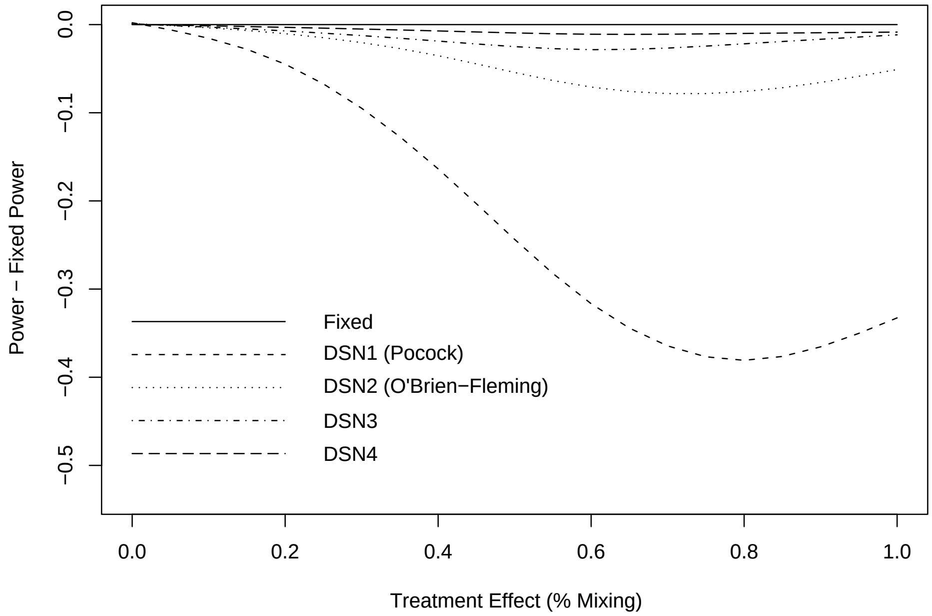
SISCR

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

Output from `seqOCWLR()`

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

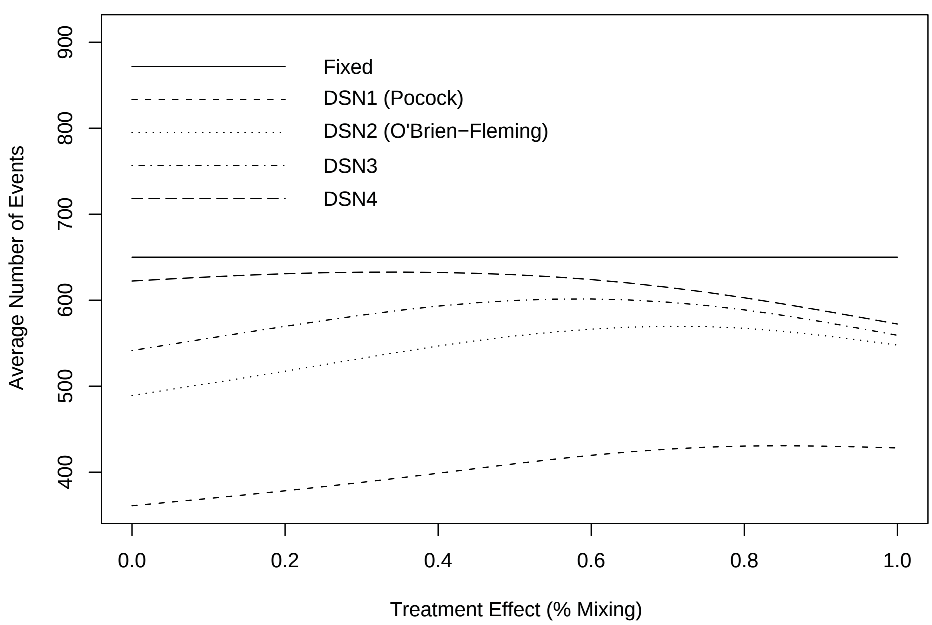
- Relative power as a function of % mixing



Output from `seqOCWLR()`

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

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



SISCR

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

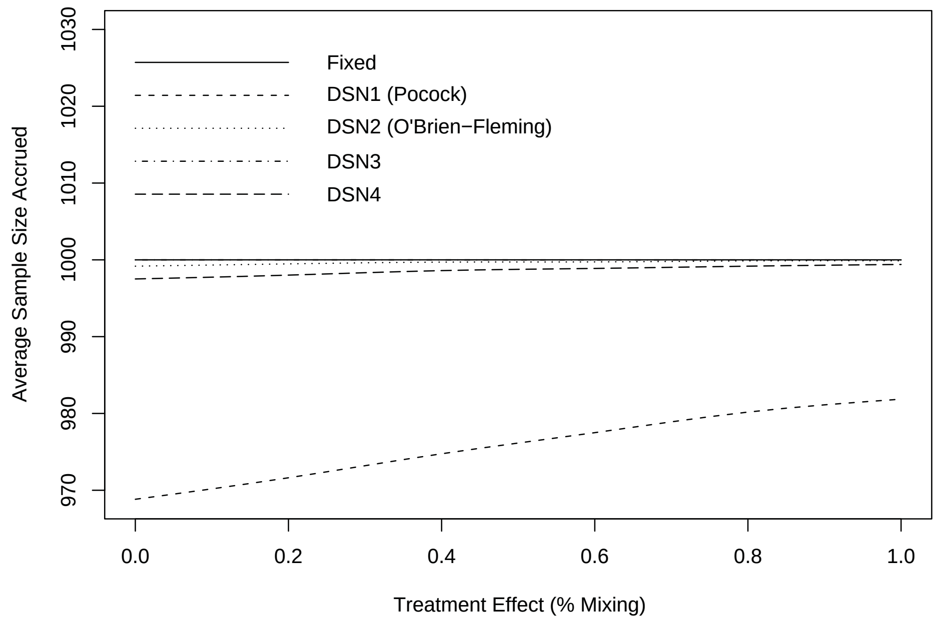
Ex: Sensitivity of operating characteristics to the censoring distribution

RCTdesign implementation of group sequential rules

Output from `seqOCWLR()`

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

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



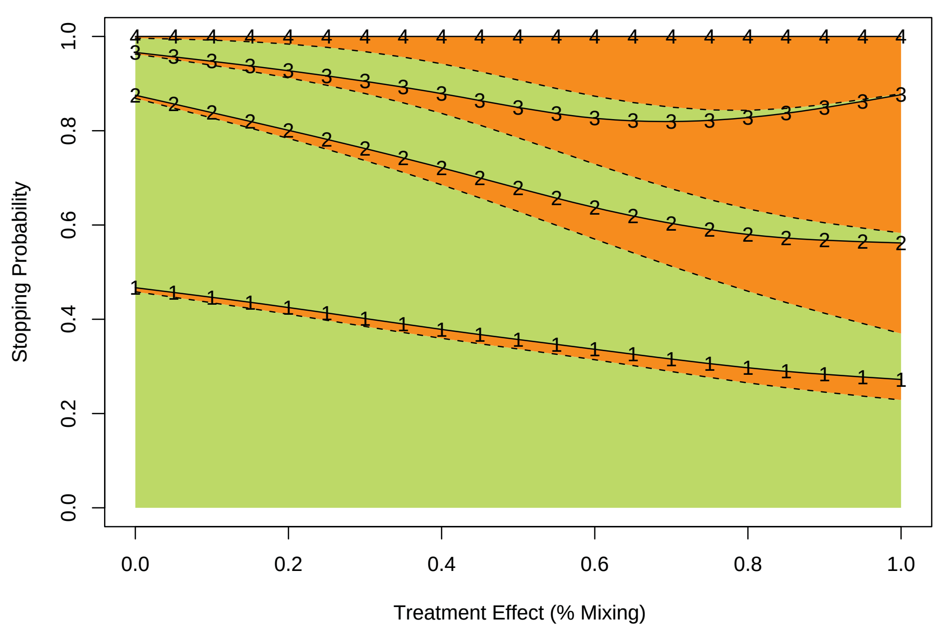
SISCR

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

Output from `seqOCWLR()`

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

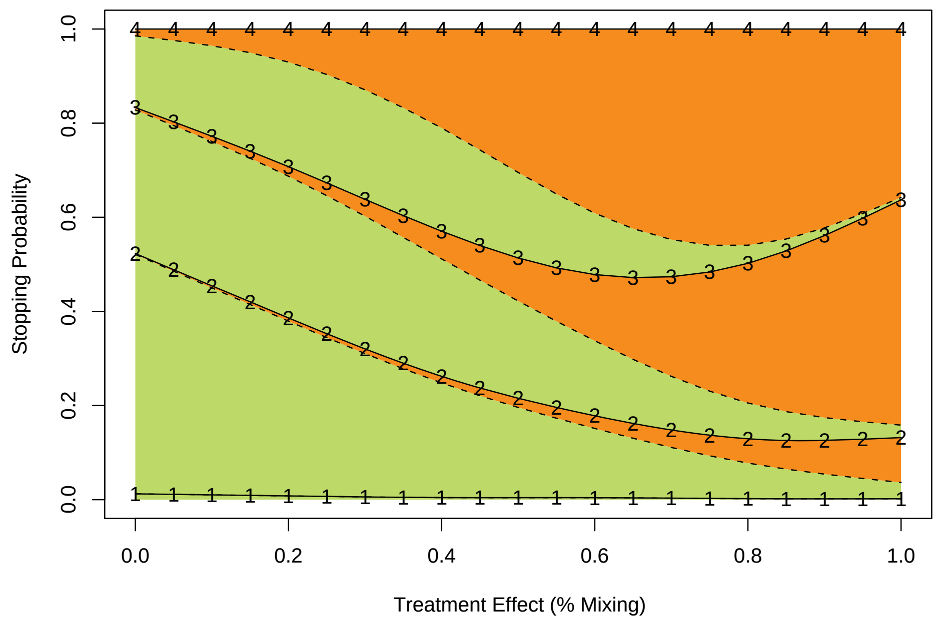
- Stopping probabilities as a function of % mixing for `DSN1` (Pocock)



Output from `seqOCWLR()`

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

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



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()`

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)

SISCR
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

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

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

SISCR

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

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

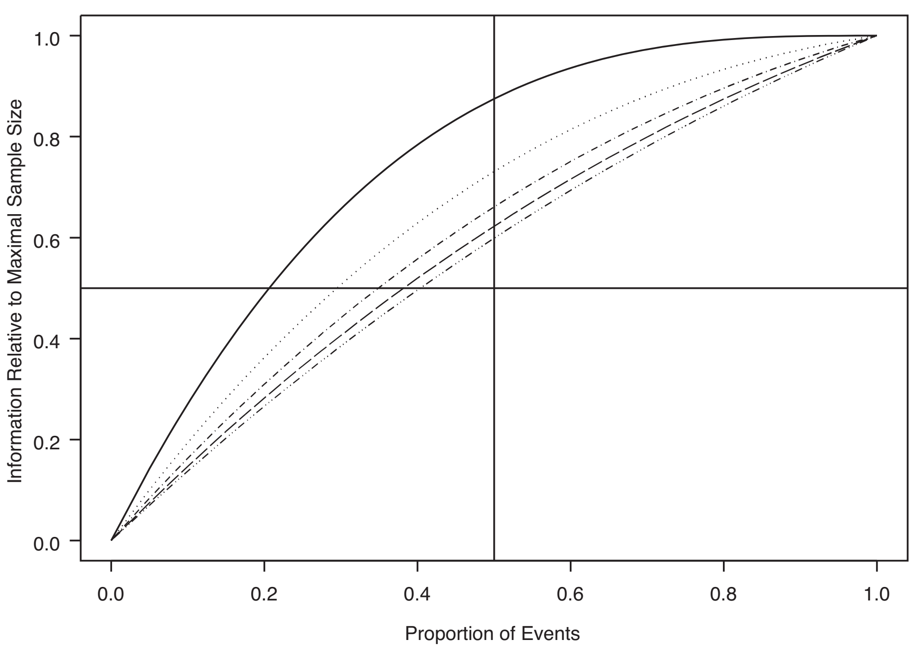
$$F_E(t) = \left(\frac{t}{\theta}\right)^r, \text{ with } \theta > 0, r > 0, 0 < t \leq \theta$$

  - 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

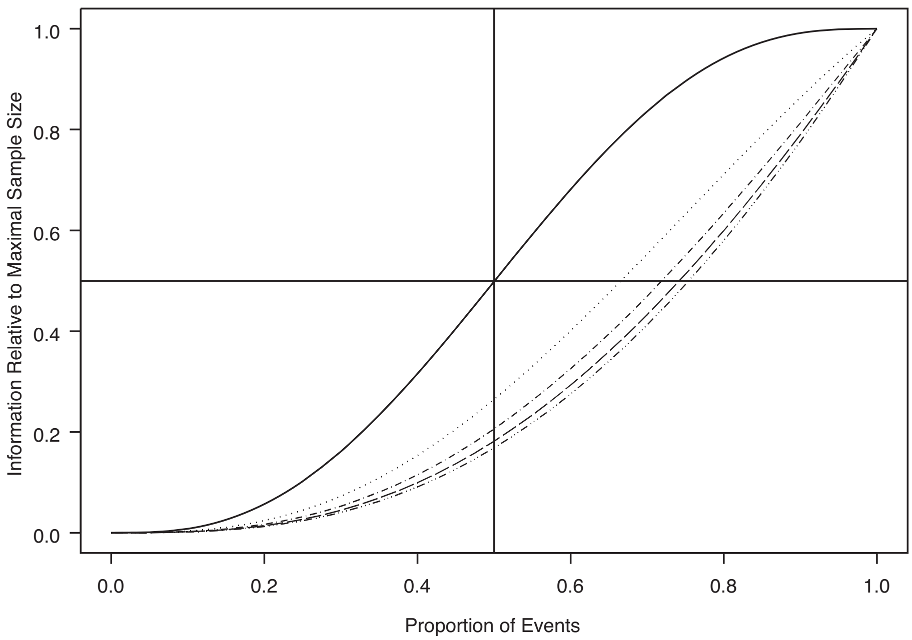
SISCR
Motivating Example
Sensitivity to Accrual Patterns
Impact of censoring on LR statistics
Evaluation of Designs When Testing with a WLR Statistic
Weighted LR statistics
Definition of alternatives
Output from `seqOCWLR()`
Monitoring Survival Trials with a WLR Statistic
Information growth for weighted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

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



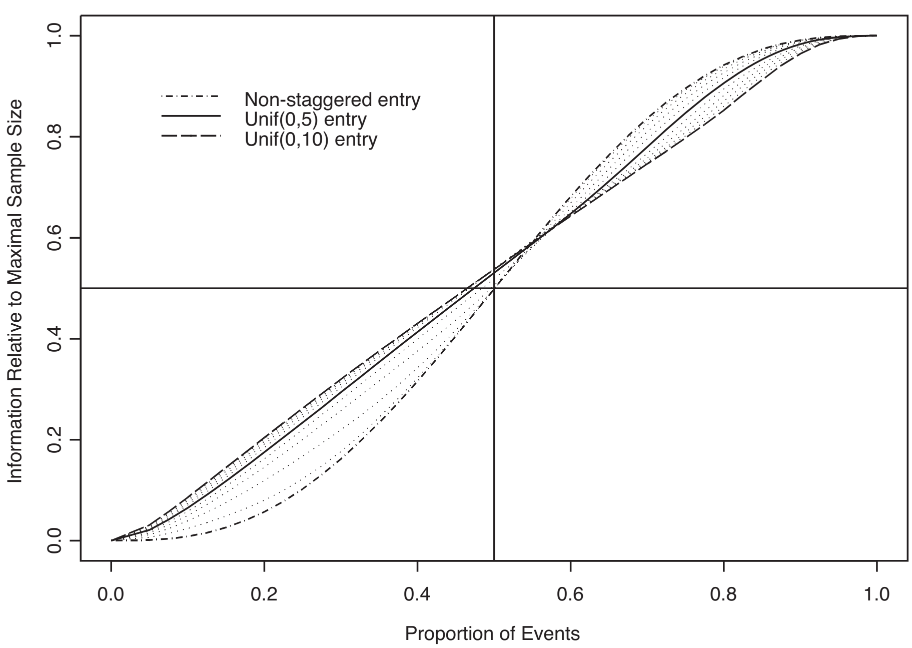
Example: Difference in Information by Accrual for the $G^{1,1}$ Statistic

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



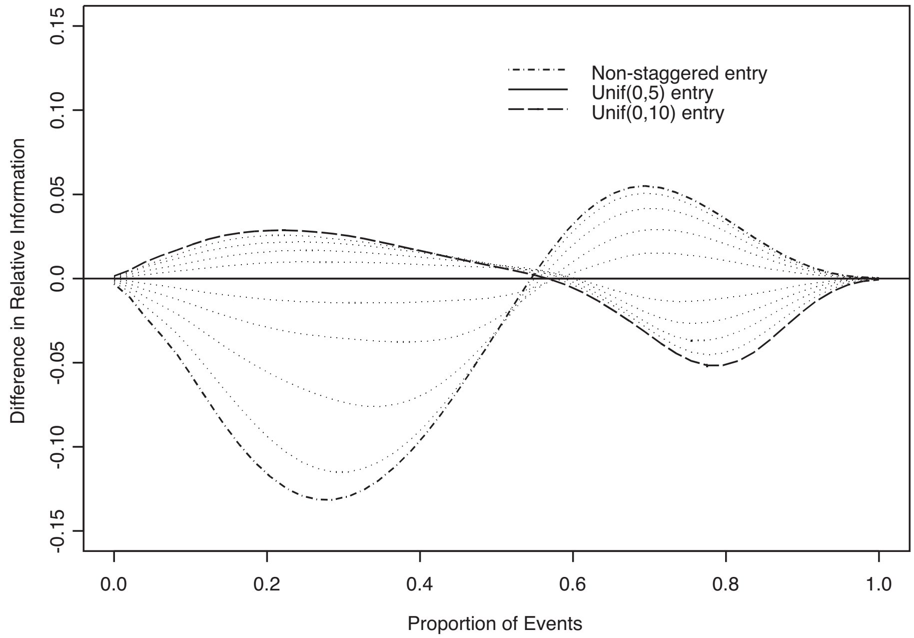
Example: Information Growth for the $G^{1,1}$ Statistic

Uniform accrual with no administrative censoring



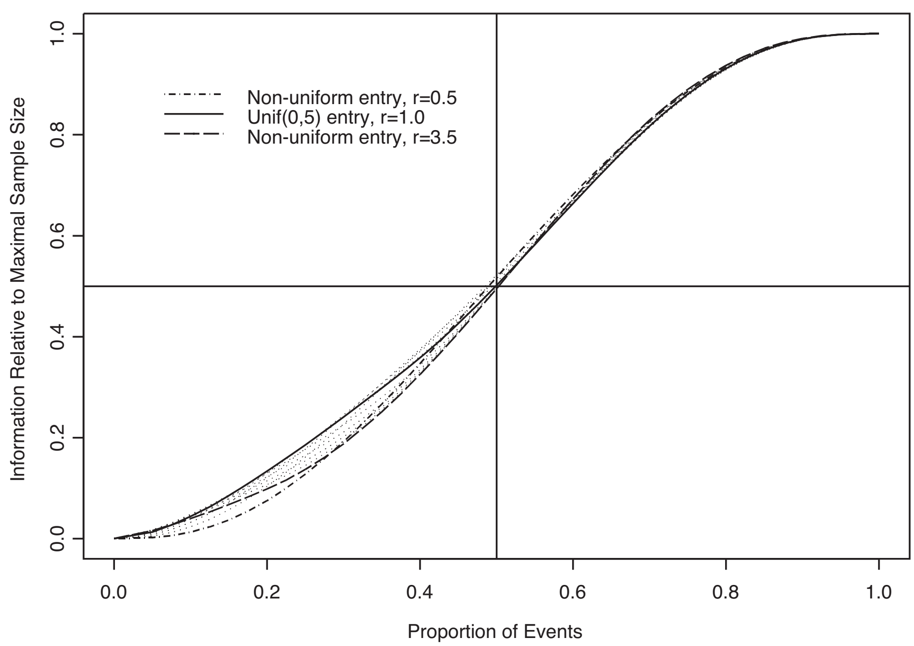
Example: Difference in Information by Accrual for the $G^{1,1}$ Statistic

Uniform accrual with no administrative censoring



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

Nonuniform accrual with no administrative censoring

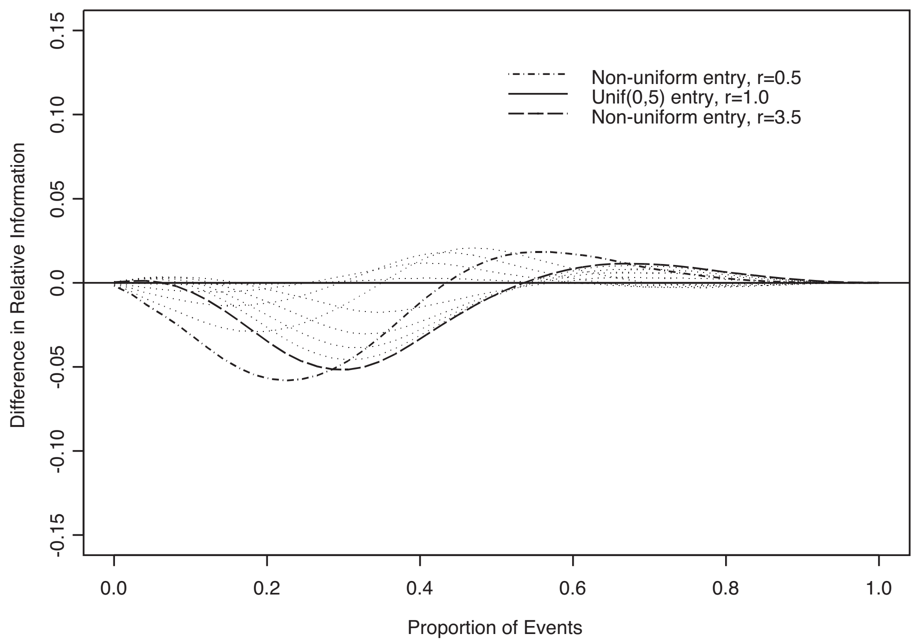


SISCR

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

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

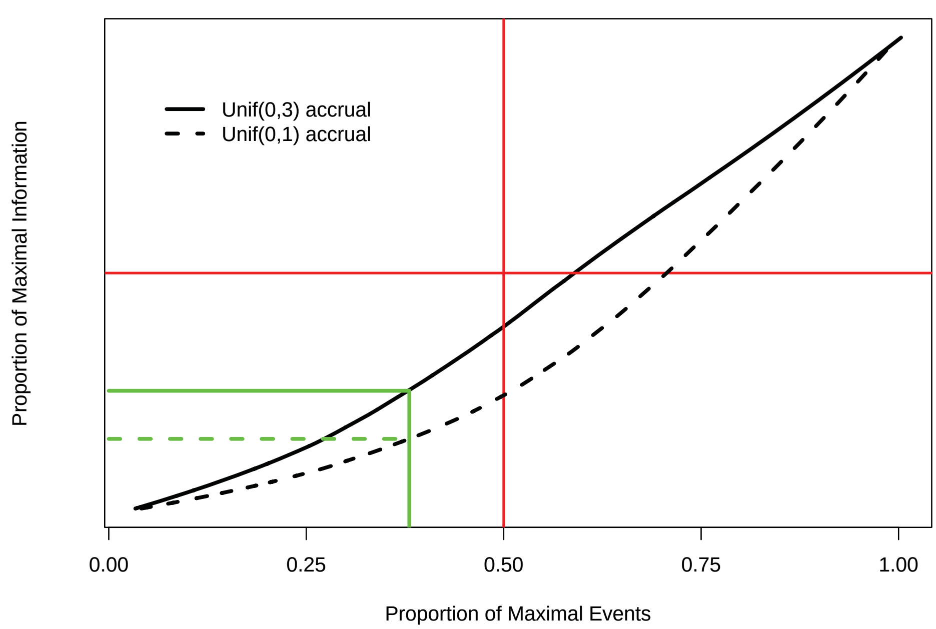
Example: Operating characteristics with misspecified accrual distribution

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

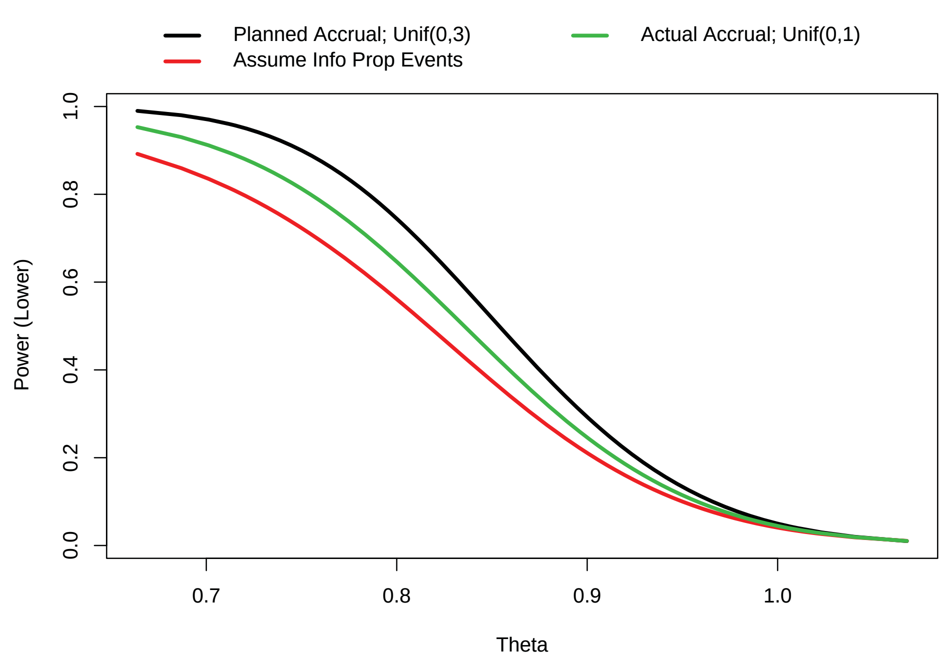
Example: Operating characteristics with misspecified accrual distribution

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
```

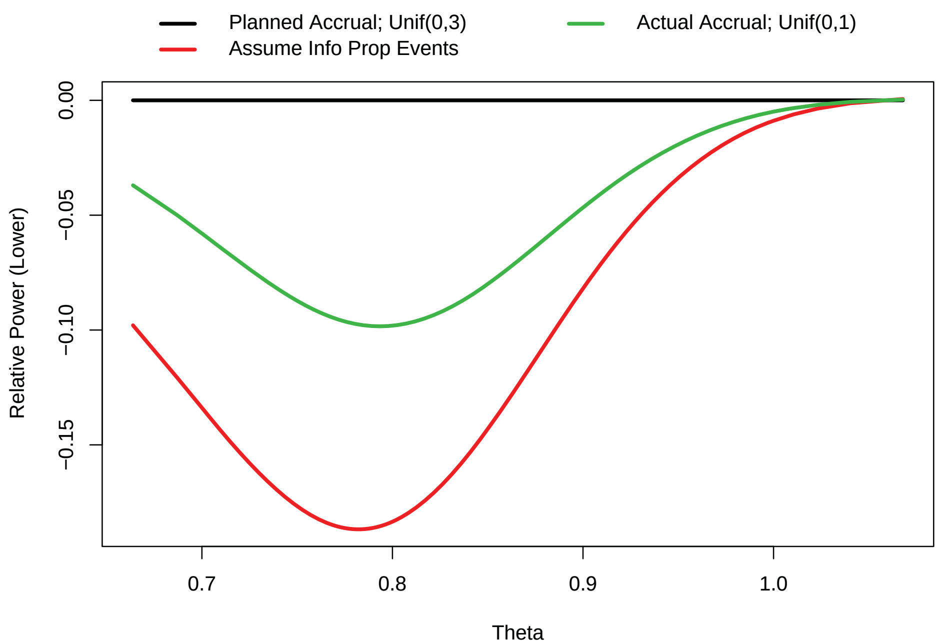

Example: Operating characteristics with misspecified accrual distribution



SISCR

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

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

SISCR

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

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

SISCR

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

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 :

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

181

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

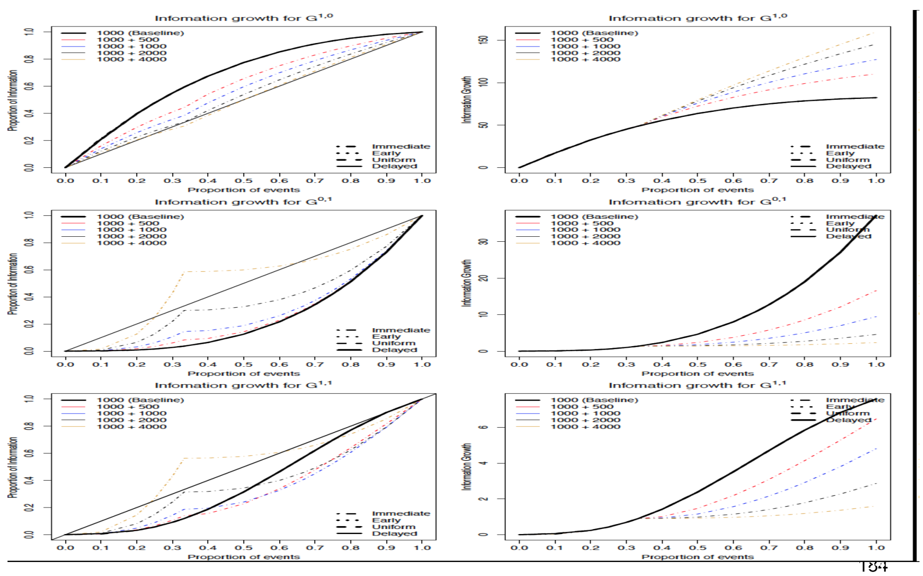
182

“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
    - **Information growth not linear in number of events**
  - Power largely unaffected, so PPV decreases

183

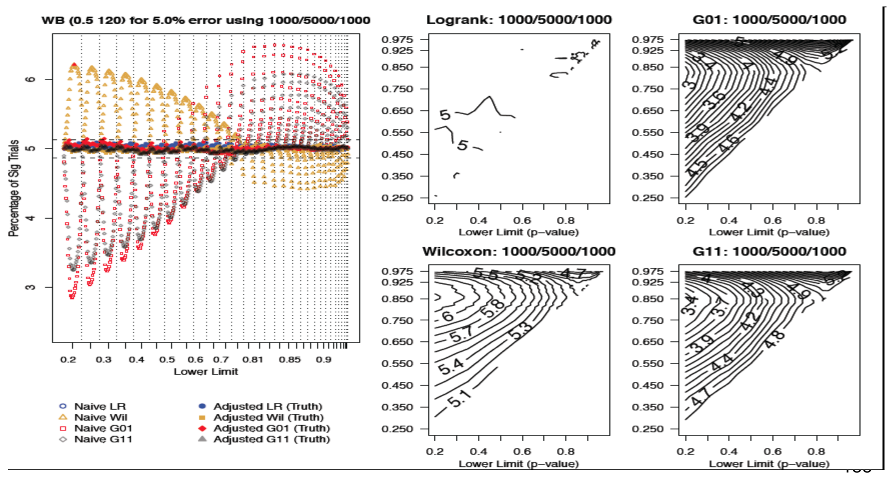
Information Growth with Adaptation



184

Inflation of Type 1 Error

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



Comments re WLR

- Hence, unblinded access to trial results can allow an investigator to inflate the type 1 error
- This might not be noticeable to a naïve audience if the number of events stays constant
- Proper handling of information growth can fix this
  - However, description of the information growth is often difficult with weighted log rank statistics

186

Nonproportional Hazards

Crossing Survival Curves

Where am I going?

Recently some authors have proposed sequential tests to be used in the presence of crossing survival curves

This example illustrates many of the difficulties inherent in applying time to event analyses

187

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.

188

Logan, et al.: Motivation

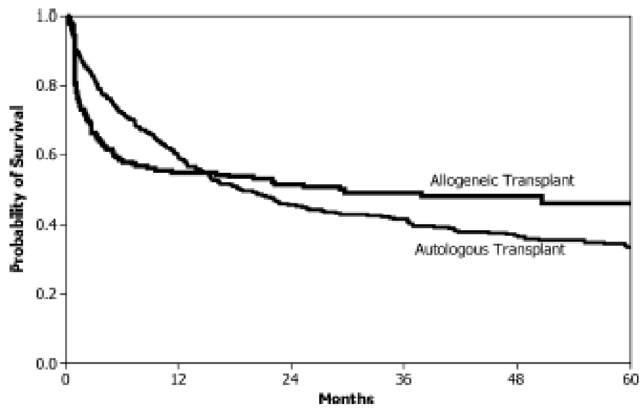


Figure 1. Kaplan–Meier estimate of DFS for follicular lymphoma example, by stem cell source.

189

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

190

Logan, et al.: Simulations

Comparing Treatments with Crossing Survival Curves

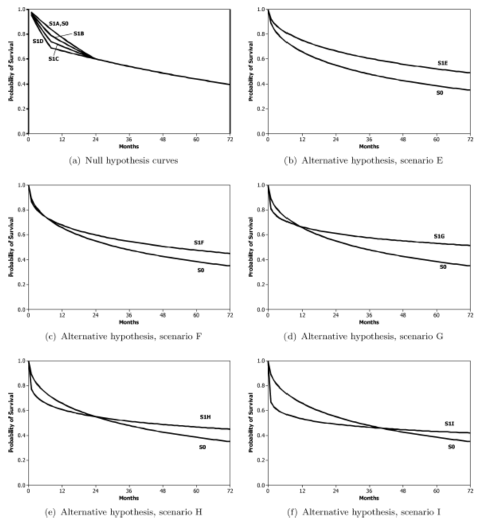


Figure 2. Survival curves for treatment (S1) and control (S0) groups used in simulations. Curves for the null hypotl simulations are shown in (a) for each of the four scenarios, and curves for the alternative hypothesis simulations are show (b)–(f) for the five scenarios.

191

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
$Z_{\mathrm{CLL}}(24)$	(1)	62.4	15.3	21.1	4.7	21.8
$Z_{\mathrm{CLL}}(48)$	(1)	70.1	32.9	65.1	21.5	6.8
$Z_{\mathrm{CLL}}(72)$	(1)	71.2	44.5	85.1	46.1	25.9
$Z_{\mathrm{WKM}}(t_0)$	(2)	75.8	35.0	66.3	20.3	6.0
$\chi^2_{\mathrm{PSV}}(t_0)$	(3)	74.8	32.0	61.2	16.4	4.8
$Z_{\mathrm{LR}}(t_0)$	(4)	30.7	36.5	85.4	71.7	82.6
$Z_{\mathrm{OLS}}(t_0)$	(5)	74.7	43.9	84.1	43.4	23.6
$Z_{\mathrm{SP,P}}(t_0)$	(6)	76.9	40.2	74.8	29.6	10.7
$\boxed{\chi^2(t_0)}$	(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

192

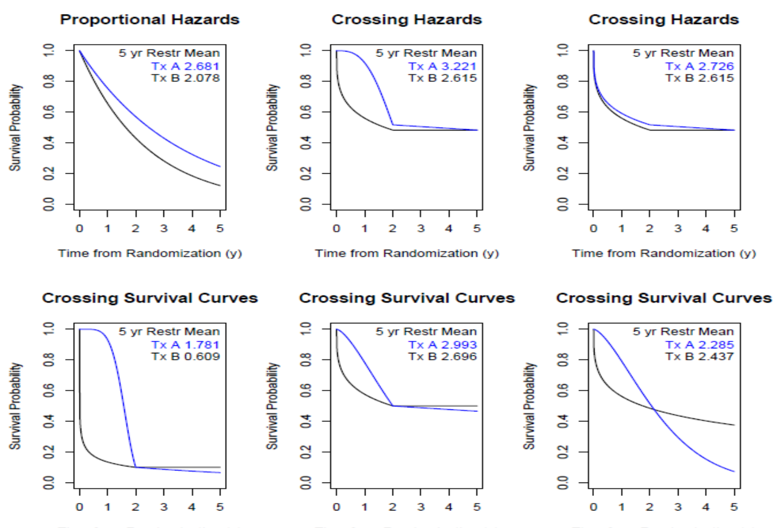
Logan, et al.: Critique

- 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

193

Logan, et al.: Critique

- Additional scenarios that are of interest



194

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%

195

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.

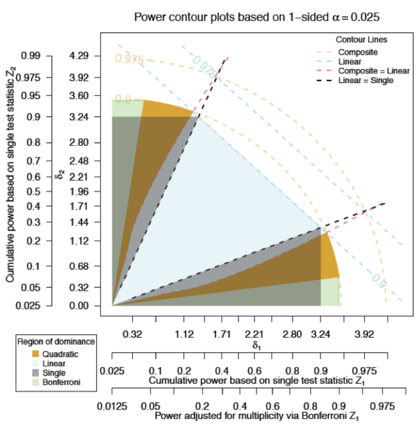
Keywords Crossing hazards · Crossing survival curves · Late survival difference · Group sequential test · Error-spending methods

196

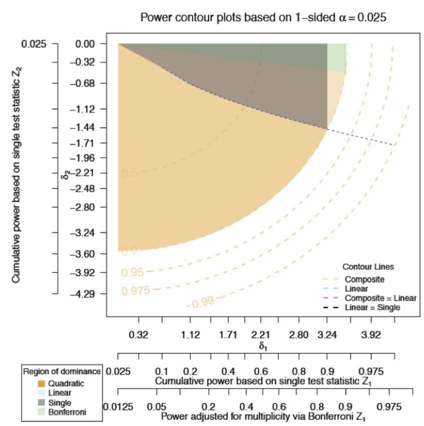
Combination statistics

- Dominance of methods of combining two statistics according to effect prior to pre-specified threshold time

Quadrants 1 & 3



Quadrant 4



Simulations: Proportional Hazards

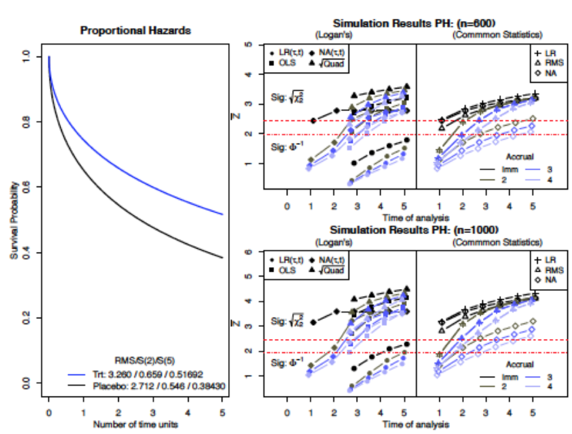


Figure 7: Survival curves under proportional hazards alternatives for n=600 and 1000. A positive Z corresponds to the treatment being superior relative to the placebo. Even with censoring, the alternatives are consistently positive (respectively in quadrant I) for the commonly used or composite statistics.

Simulations: Stochastic Ordering

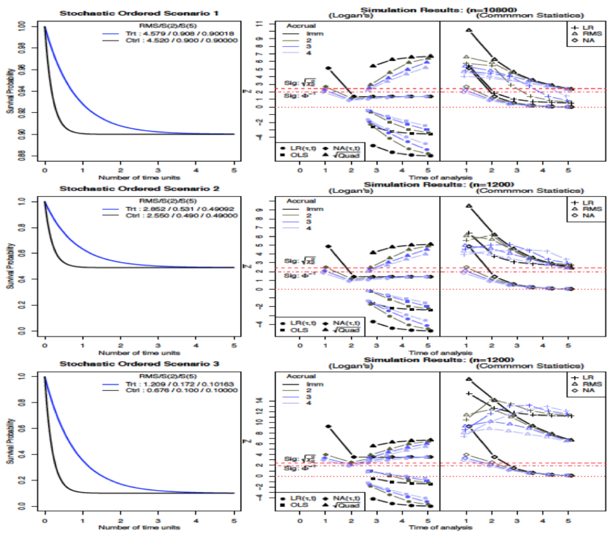


Figure 8: Survival curves under stochastic ordering with variations to either high/moderate/poor survival probability. The alternatives for each test statistic do not vary much across these 3 scenarios. While the commonly used statistics have alternatives depicting superiority of treatment for overall LR and RMS, the combination of alternatives for $NA(\tau, t)$ and $LR(\tau, t)$ belong to quadrant IV and describes conflicting conclusions prior to crossing and after crossing. The net effect for the linear composite statistics appear to conclude placebo is better than the treatment.

199

Simulations: Crossing Survival

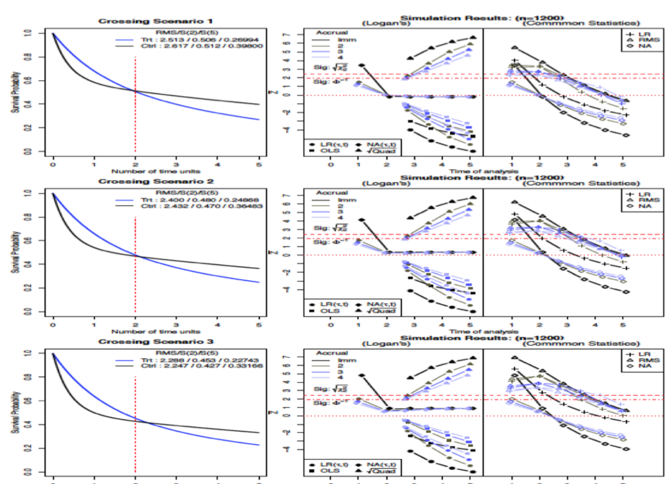


Figure 11: Simulated crossing hazards, crossing survival curves where we added variations to whether the curves crosses just before 2, and after 2. The combination of composite alternatives changes from Quadrant III (-,-), Quadrant III/IV(0, -), Quadrant IV (+,-) with the net result providing conclusion of preferring the placebo over treatment. The commonly used statistics is seen to have a changing alternative that switches from positive to negative.

Comments

- Sequential adaptive studies with crossing survival curves are fraught with peril
- One must *a priori* identify likely scenarios and ensure that any monitoring does not terminate before the best treatment is revealed
- More importantly, one must have a clear definition of what is best

201

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?)

202

Bottom Line

- There is no substitute for planning a study in advance
  - At Phase 2, adaptive designs are clearly useful to better control parameters leading to Phase 3
    - Most importantly, learn to take "NO" for an answer
  - At Phase 3, it is less clear whether much is gained from unblinded adaptation
    - And scientific / statistical credibility can suffer
- **"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.

203

Really Bottom Line

"You better think (think)

about what you're

trying to do…"

-Aretha Franklin, "Think"

204