Sequential and Adaptive Analysis with Time-to-Event Endpoints

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

Webinar
ASA Biopharmaceutical Section
April 18, 2017

Where Am I Going?

Overview and Organization of the Course

Science and Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)

- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)

- In RCT, we are trying to prove the effect of some treatment
  - What do we need to consider as we strive to meet the burden of proof with adaptive modification of a RCT design?

- Does time to event data affect those issues?
  - Short answer: No, UNLESS subject to censoring
  - So, true answer: Yes.

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

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- FDA Draft guidance on adaptive designs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

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

Course Organization

- Overview:
  - What do we know about survival analysis?
  - RCT setting
- 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

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

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)

Right Censored Data

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

$$\text{Unobserved :}$$

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

$$\text{Censoring Times : } \left\{C_1, C_2, \ldots, \mathrm{C}_n\right\}$$

$$\text{Observed data :}$$

$$\text{Observatio n Times : } T_i = \min\left(T_i^0, C_i\right)$$

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

$$\text{Group membership : } \mathrm{X}_i$$

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

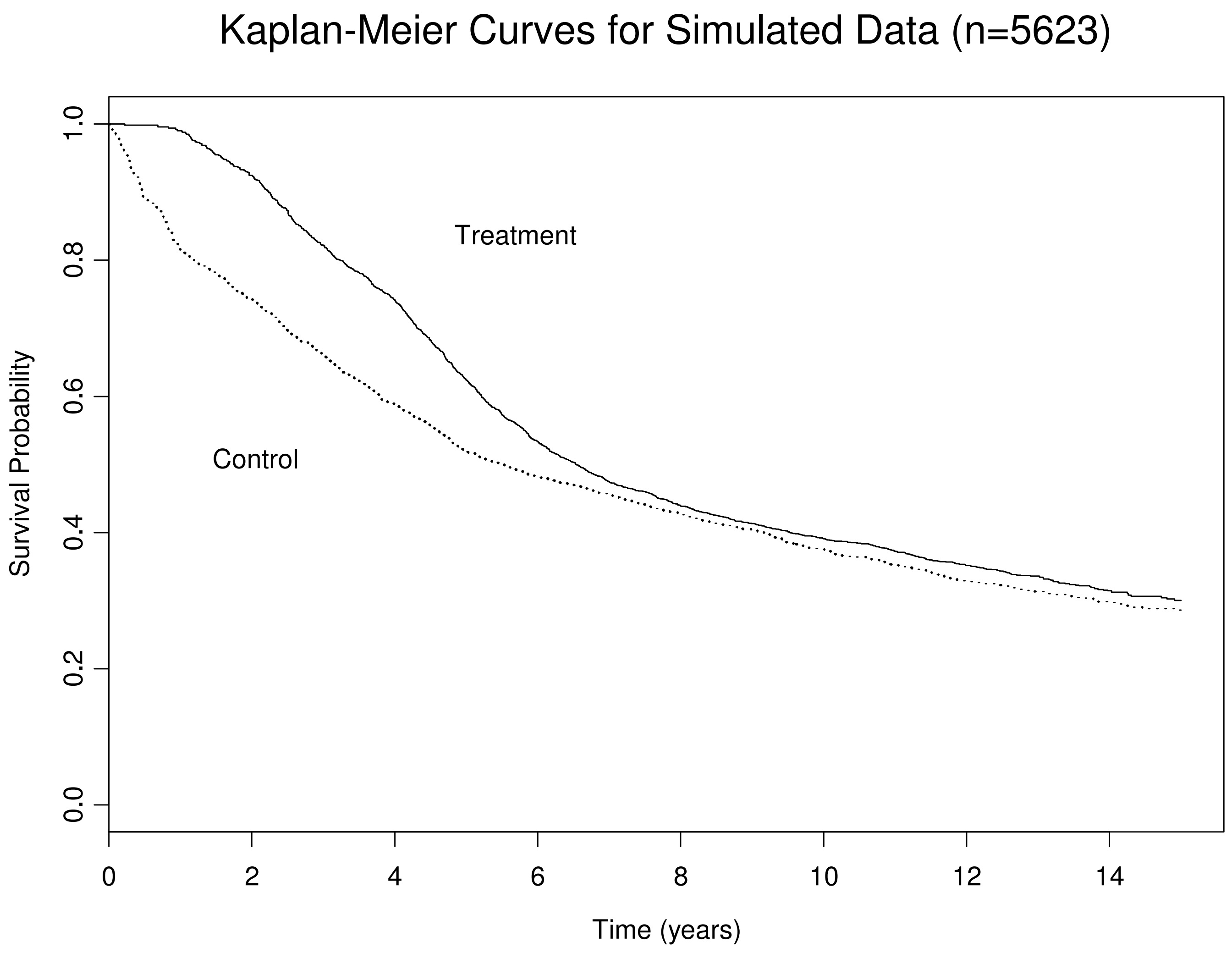
Hypothetical Example: Setting

- 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

Hypothetical Example: KM Curves



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?

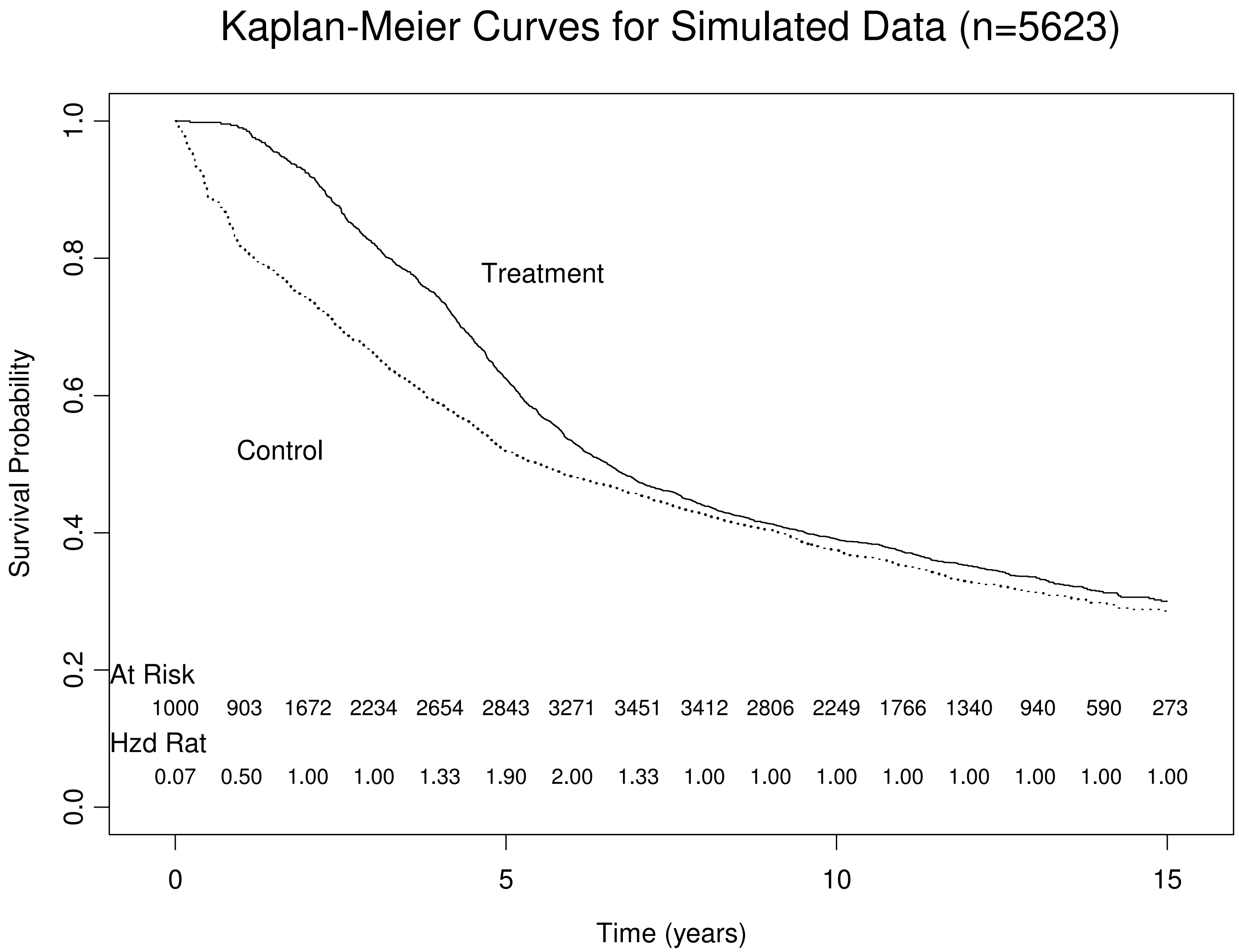
Who Wants To Be A Millionaire?

- 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?

Hypothetical Example: KM Curves



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

Partial Likelihood Based Score

- Logrank statistic

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

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

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

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

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)

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

Evidence Based Medicine

- Decisions about treatments should consider PICO
  - Patient (population)
  - Intervention
  - Comparators
  - Outcome

- There is a need for estimates of safety, effect

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?

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)

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

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

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

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

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

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:

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

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

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

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

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

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

Group Sequential Designs

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined

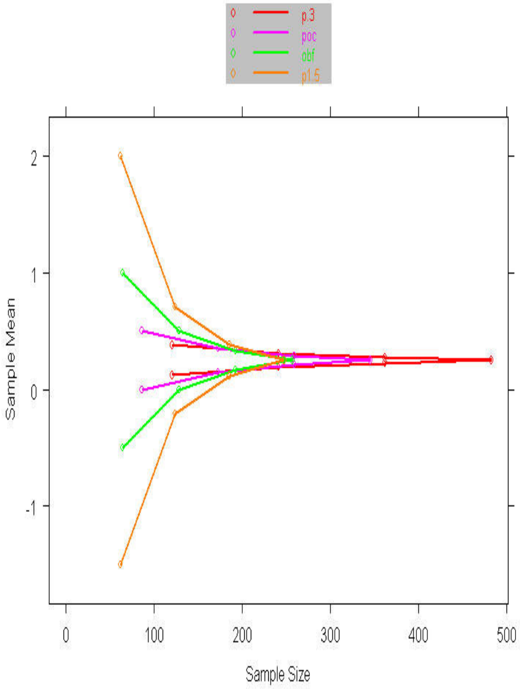
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$

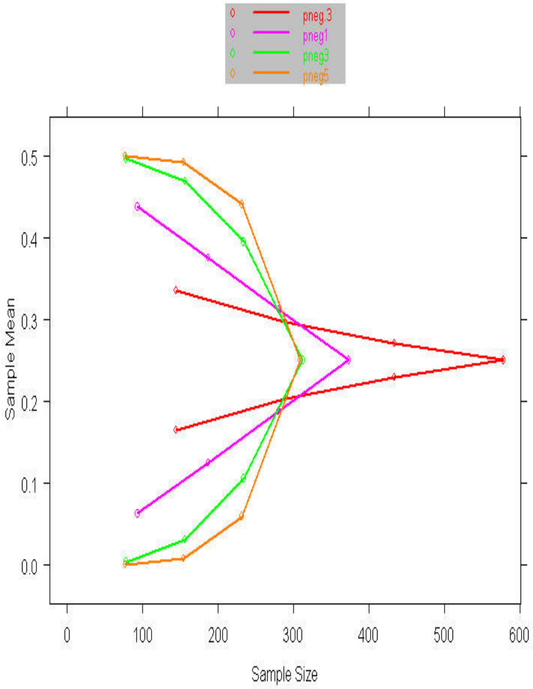
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue

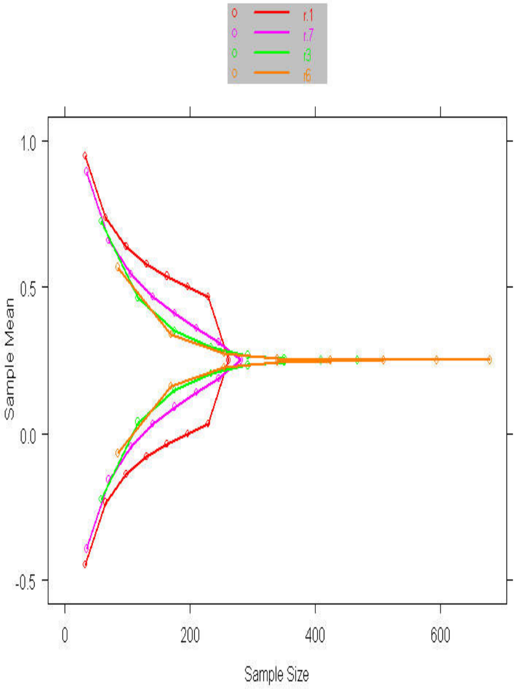
- Boundaries chosen to protect 2 of 3 operating characteristics
  - Type 1 error, power
  - Type 1 error, power, maximal sample size

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

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

Notation

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

Probabilit y 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 :

Approximat e distn : $\hat{\theta}_j = \hat{\theta}_{N_j} \overset{\cdot}{\sim} N(\theta, V / N_j)$

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

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

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

Typically:
$$se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}}$$

(V related to average "statistical information")

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

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

Ex: Difference of Indep Means

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

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

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

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

Ex: Hazard Ratios

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

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

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

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

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

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?)

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

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 : $$n = \frac{\left(\delta_{\alpha\beta}\right)^2 V}{\left(\theta_1 - \theta_0\right)^2}$$

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

When Sample Size Constrained

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

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

  - Compute alternative detected with high power

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

Increasing Precision

- Options

  - Increase sample size
    - Time to event: Accrue more patients

  - 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)

  - (Decrease confidence level)

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

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

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)

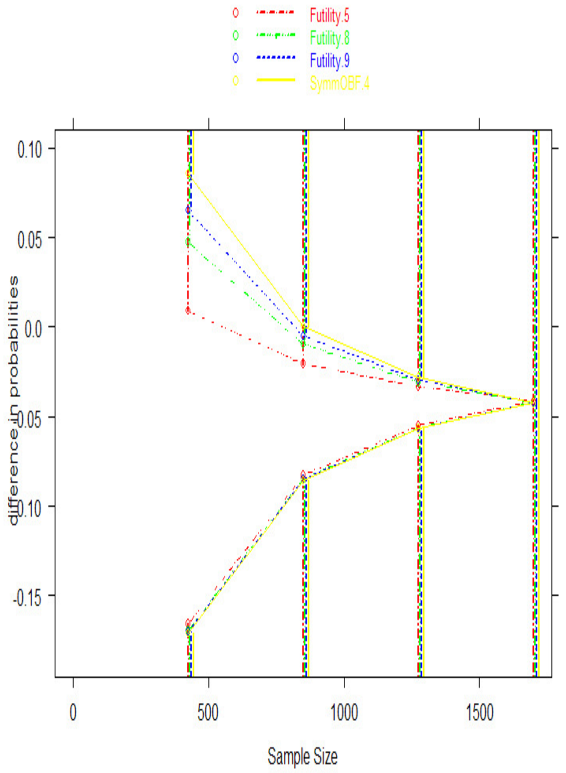
Efficiency / Unconditional Power

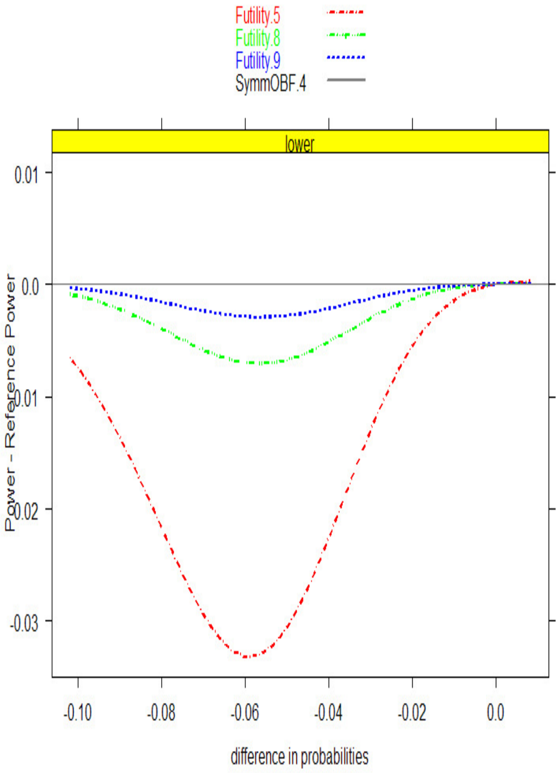
- Tradeoffs between early stopping and loss of power

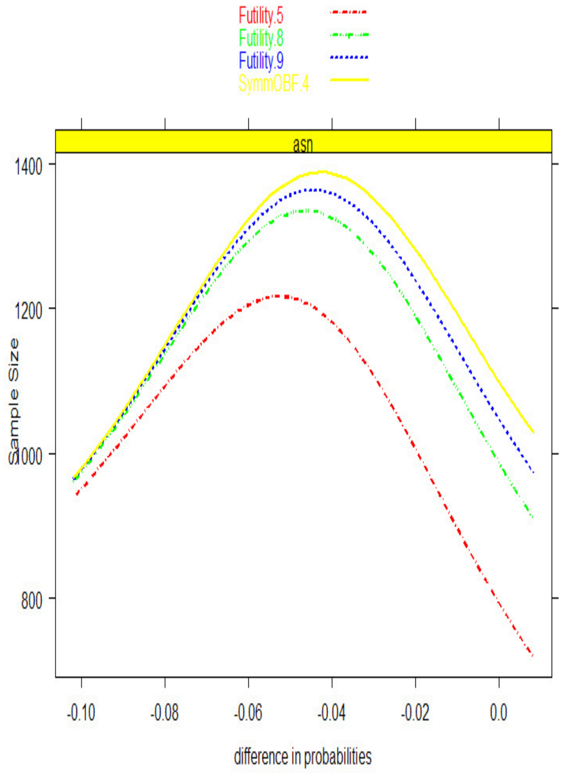
Boundaries

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

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

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

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

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)

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

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

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)

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

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

Incremental Statistics

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

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

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

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

Conditional Distribution

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

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

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

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)

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

Unconditional Distribution

- Under the null
  - SDCT: Standard normal
  - Bauer & Kohne: Sum of exponentials

- Under the alternative
  - Unknown unless prespecified adaptations

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

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

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

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

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

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

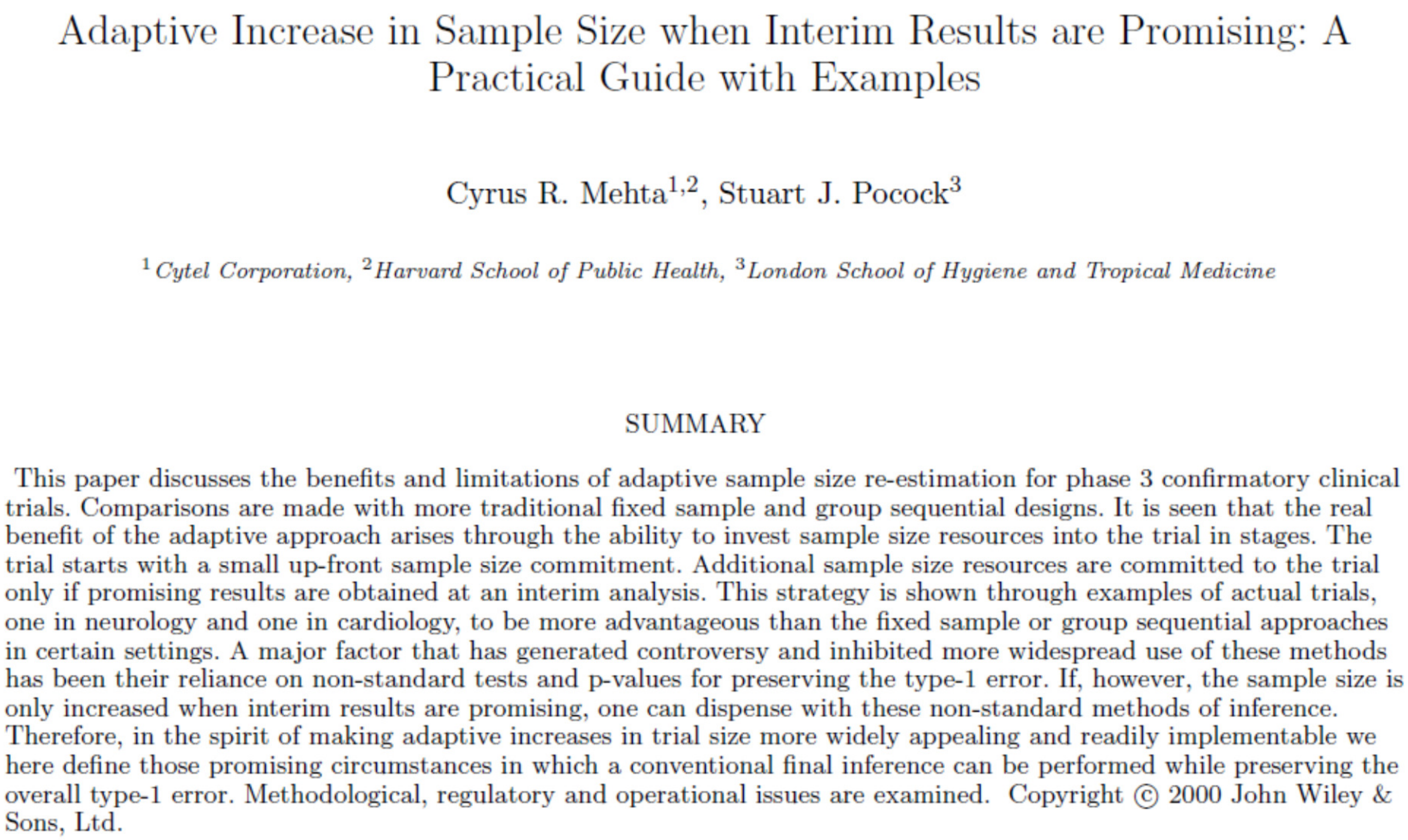
Proportional Hazards

Sample Size Re-estimation (SSRE)

Where am I going?

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

Example



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

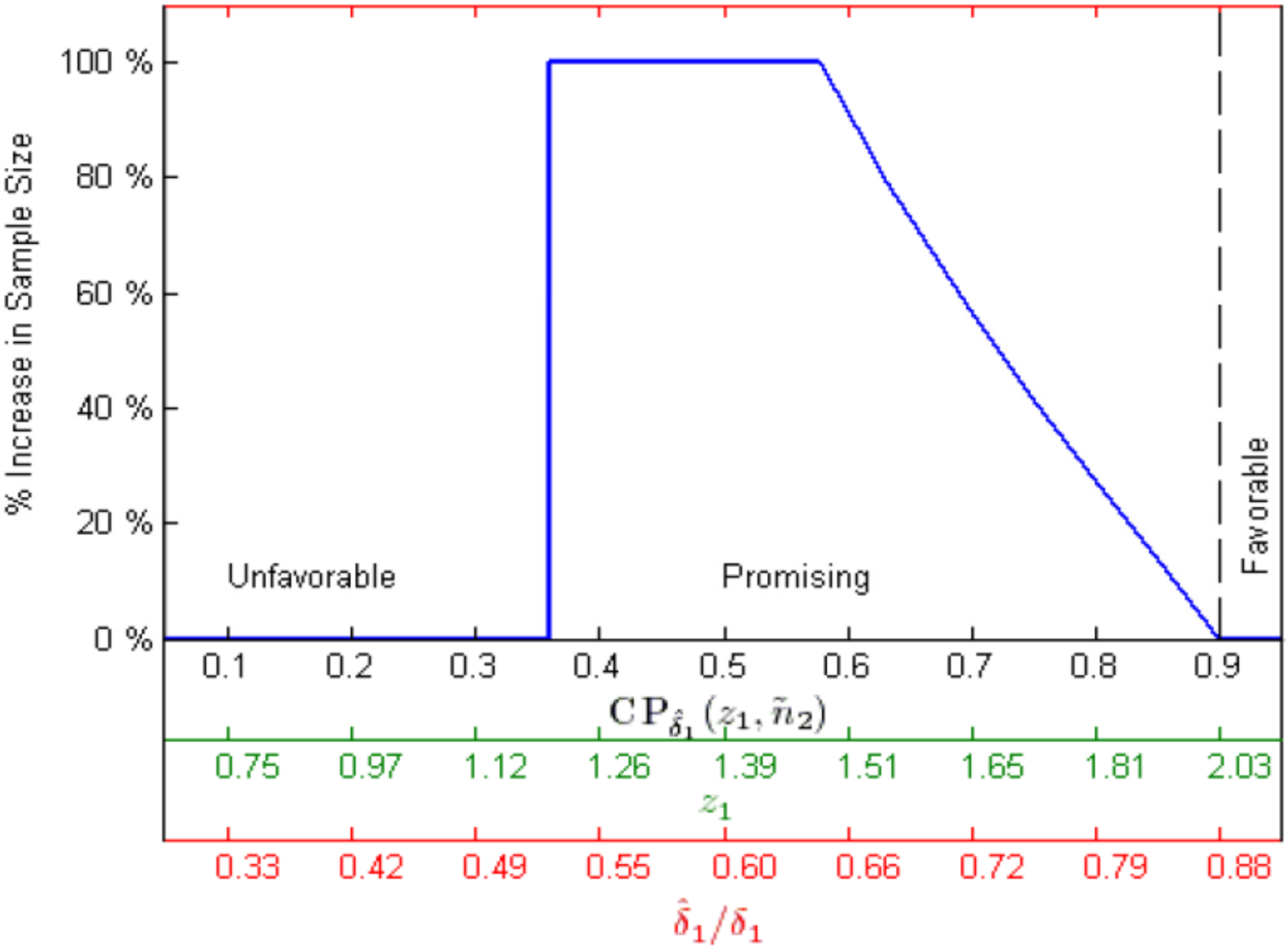
Cyrus R. Mehta[1,2], Stuart J. Pocock[3]

[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. Copyright © 2000 John Wiley & Sons, Ltd.

Example Modification Plan



Comparisons Unconditional Power

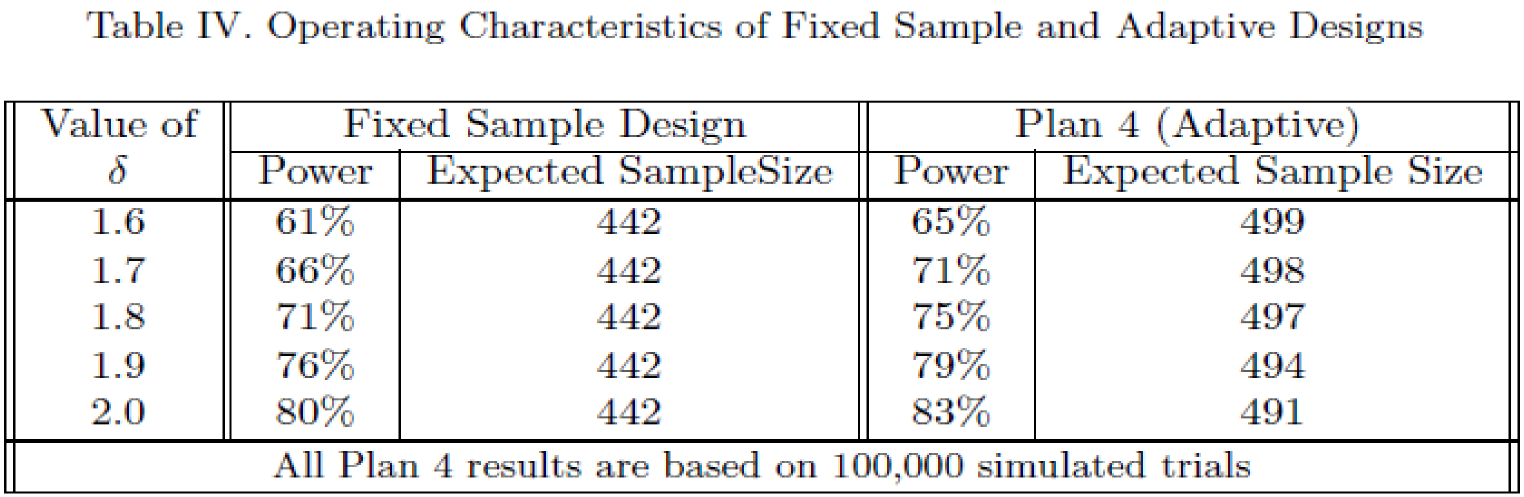
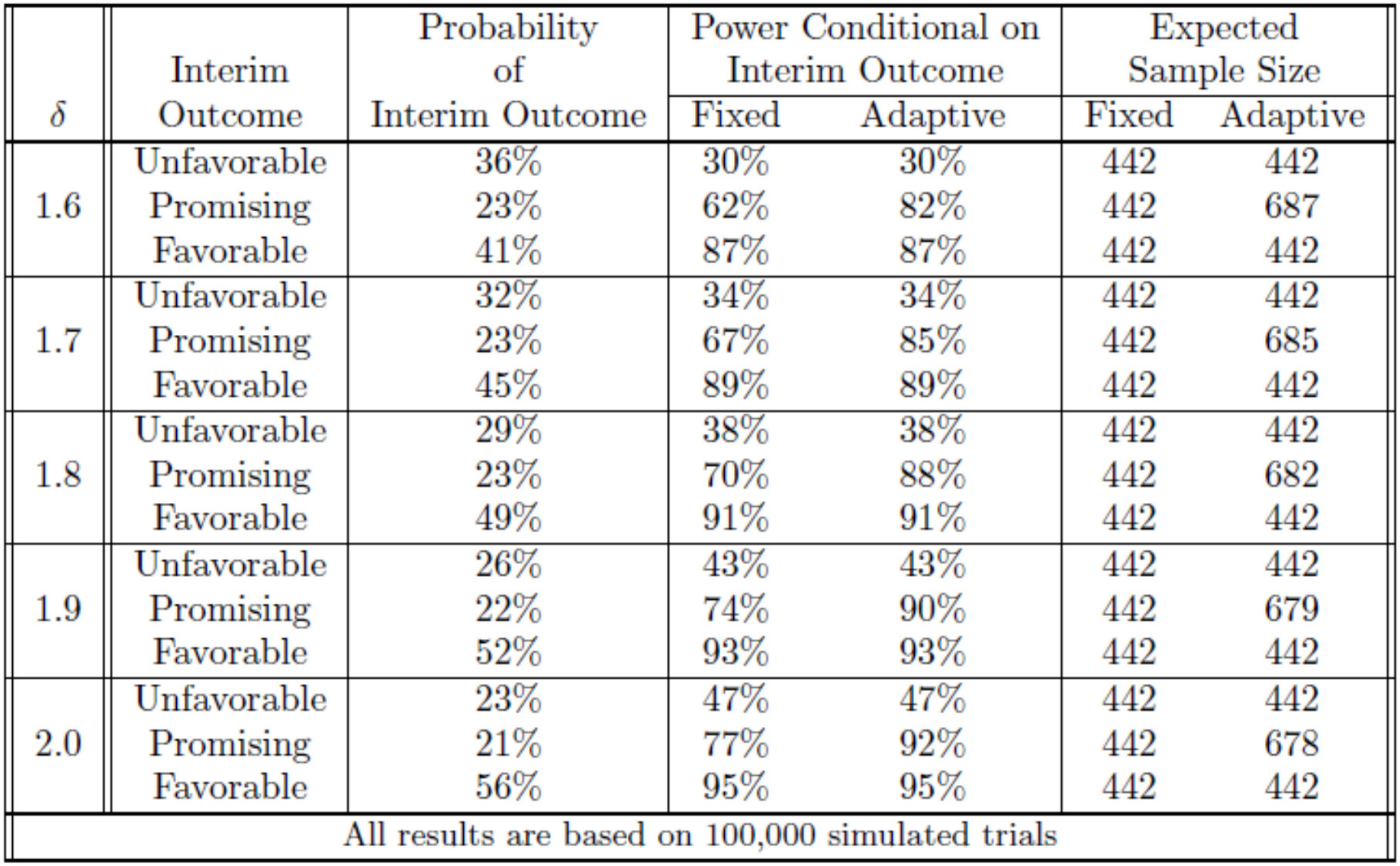


Table IV. Operating Characteristics of Fixed Sample and Adaptive Designs

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				

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		Expected Sample Size	
			Fixed	Adaptive	Fixed	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

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)

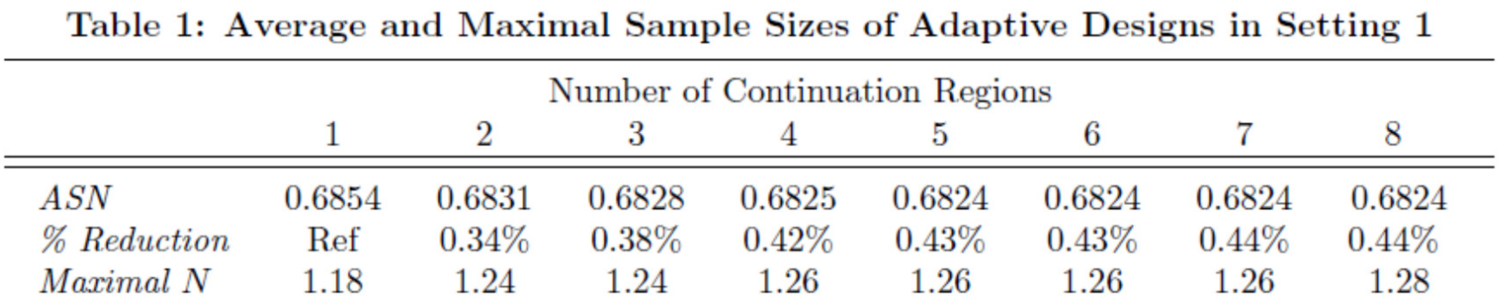


Table 1: Average and Maximal Sample Sizes of Adaptive Designs in Setting 1

	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

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)

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

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

Alternative Approaches

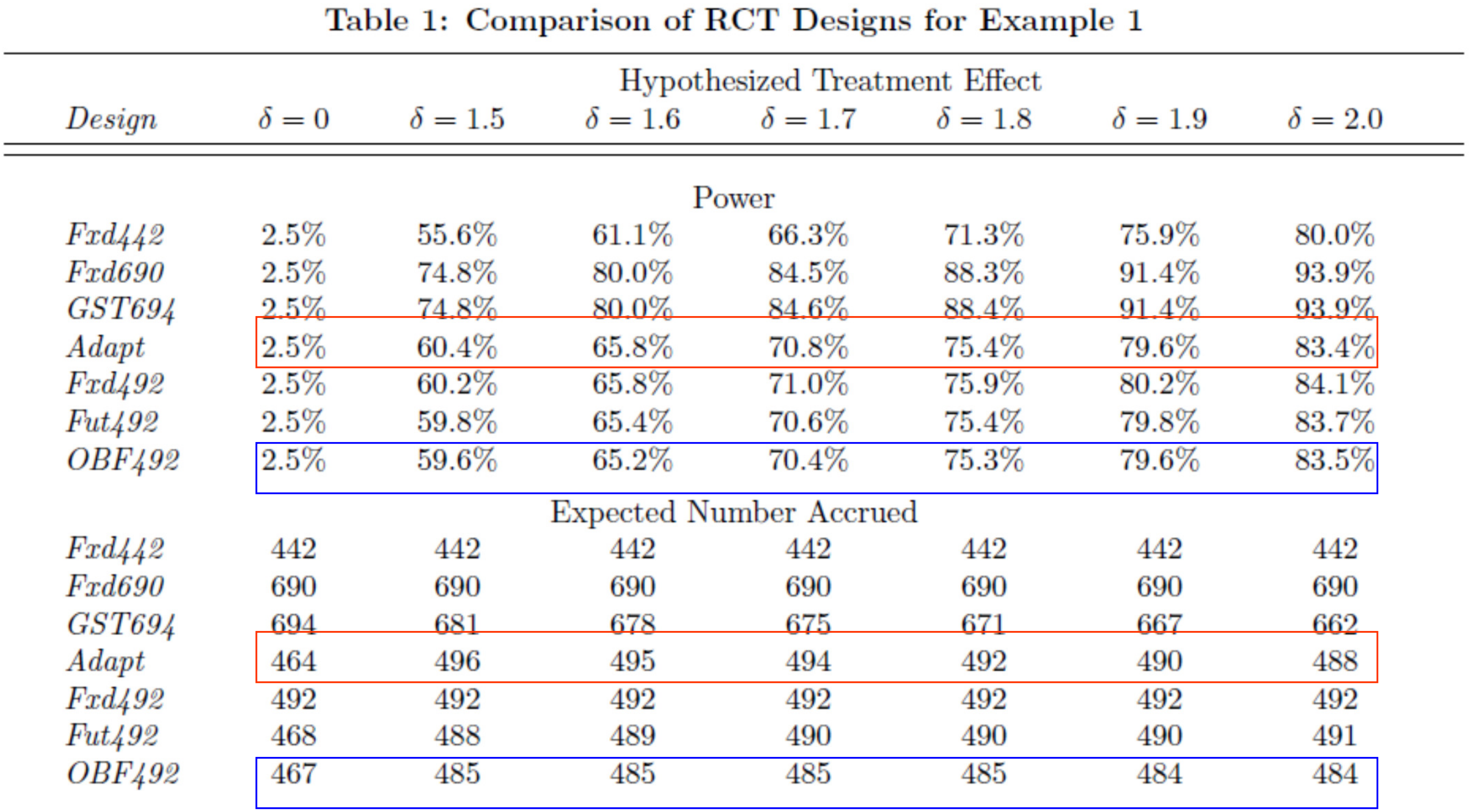


Table 1: Comparison of RCT Designs for Example 1

	Hypothesized Treatment Effect						
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

Alternative Approaches

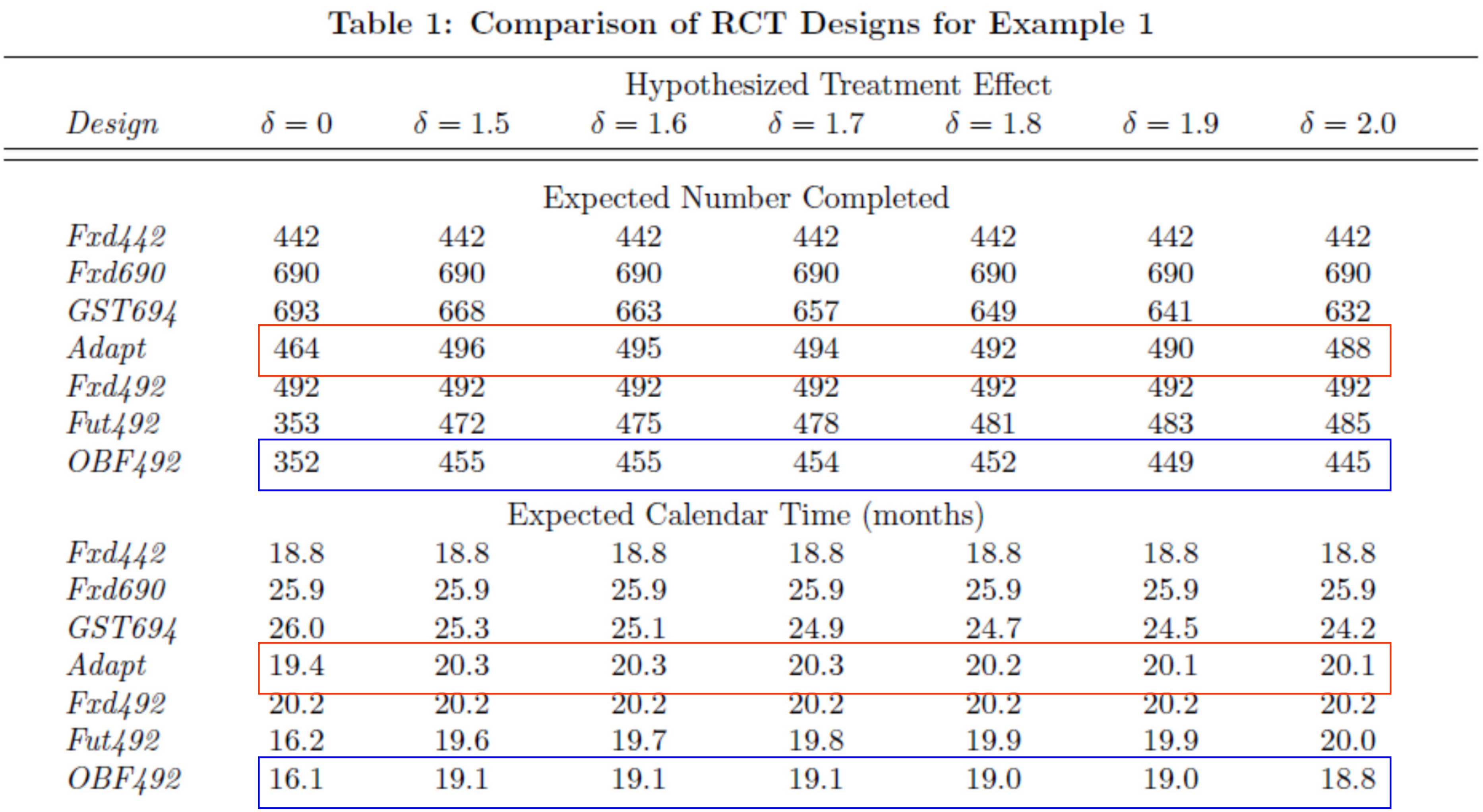


Table 1: Comparison of RCT Designs for Example 1

	Hypothesized Treatment Effect						
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

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

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

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.

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

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

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

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

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

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

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

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

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

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

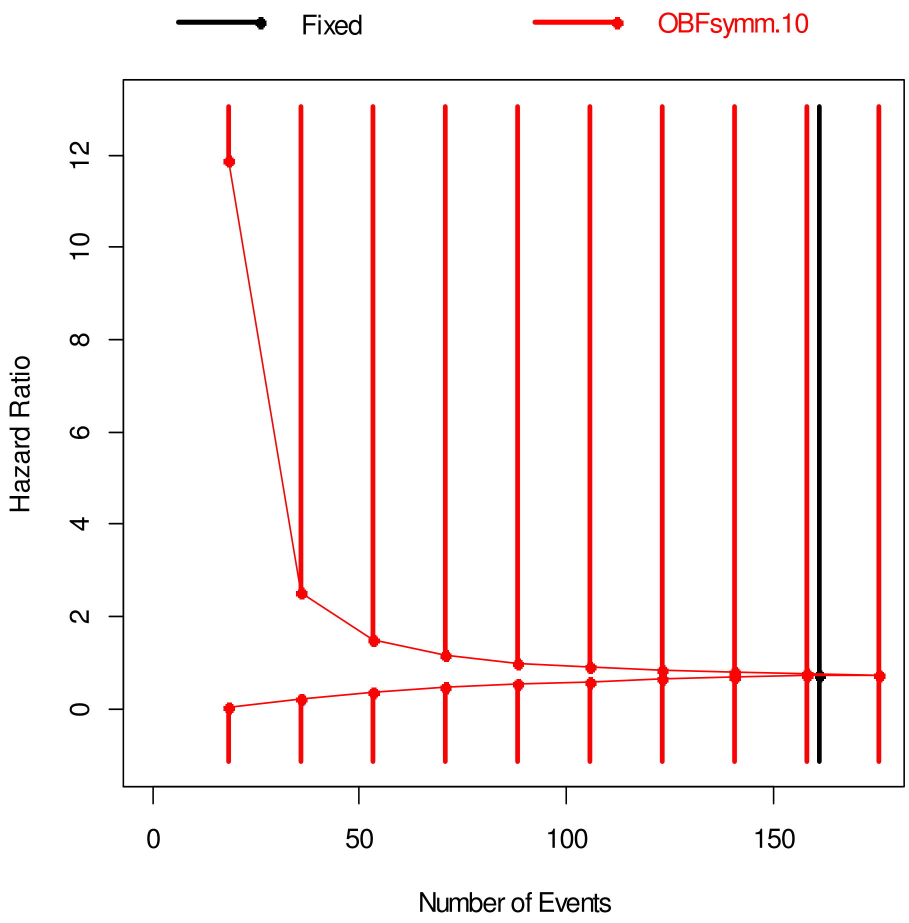
Group Sequential Design

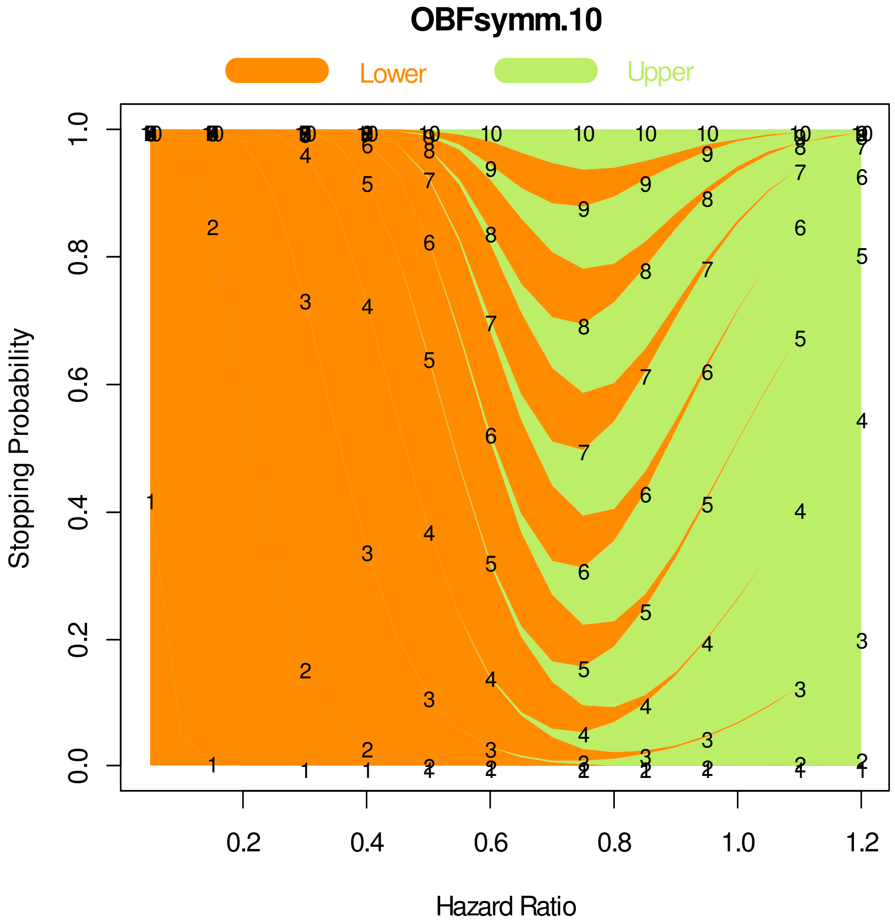
- Instead of a fixed sample design, pre-specify a group sequential design with, say, 10 possible analyses
  - Example: level 0.025, 90% power to detect HR=0.6

```
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
```

Group Sequential Design

- Stopping boundaries, stopping probabilities





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.

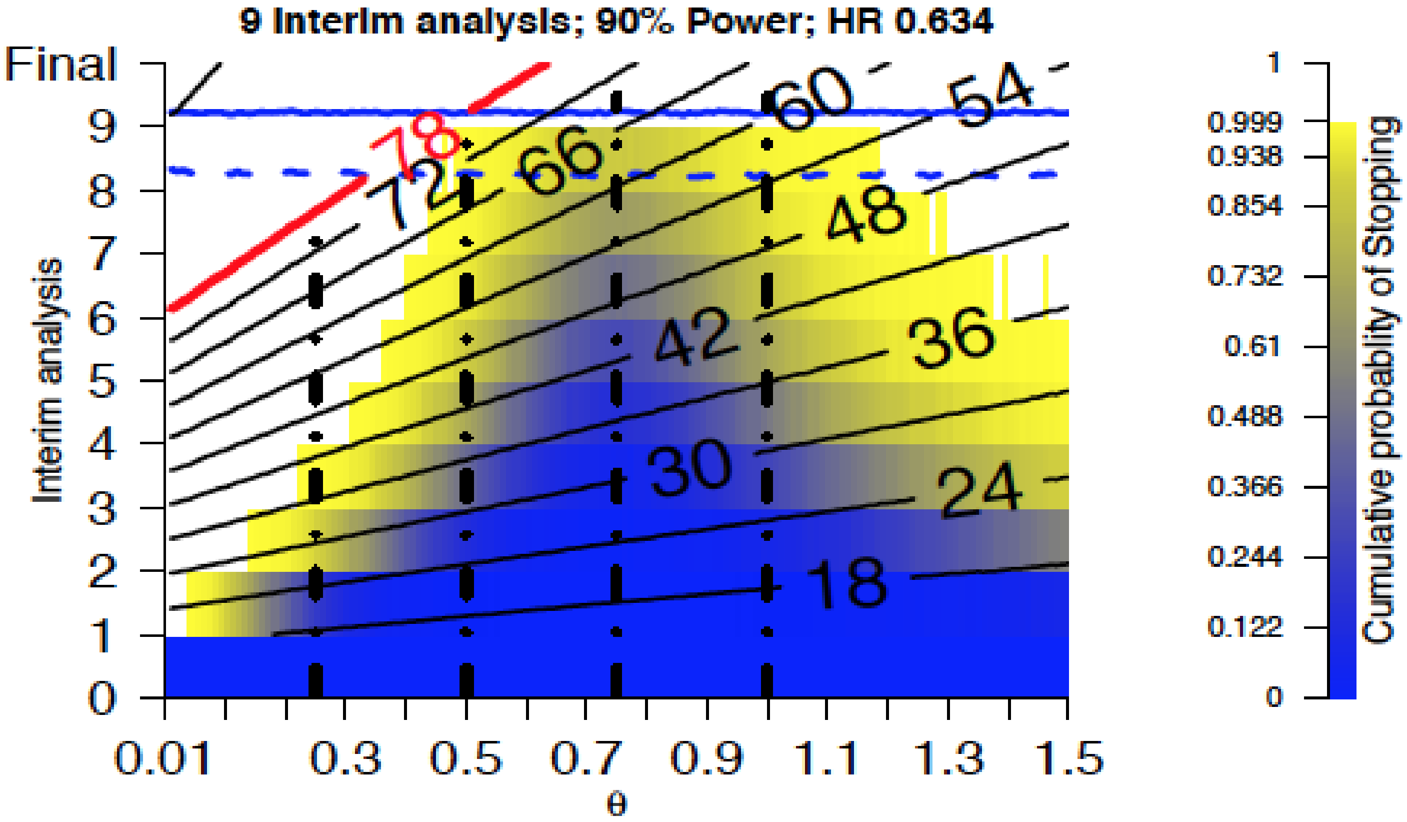
- Suppose, a slow accrual of events is due solely to a highly effective treatment
  - Placebo has the planned event rate, Experimental treatment has extremely low event rate

- 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

Calendar Time: Half Event Rate

- Stopping probabilities under planned event rate



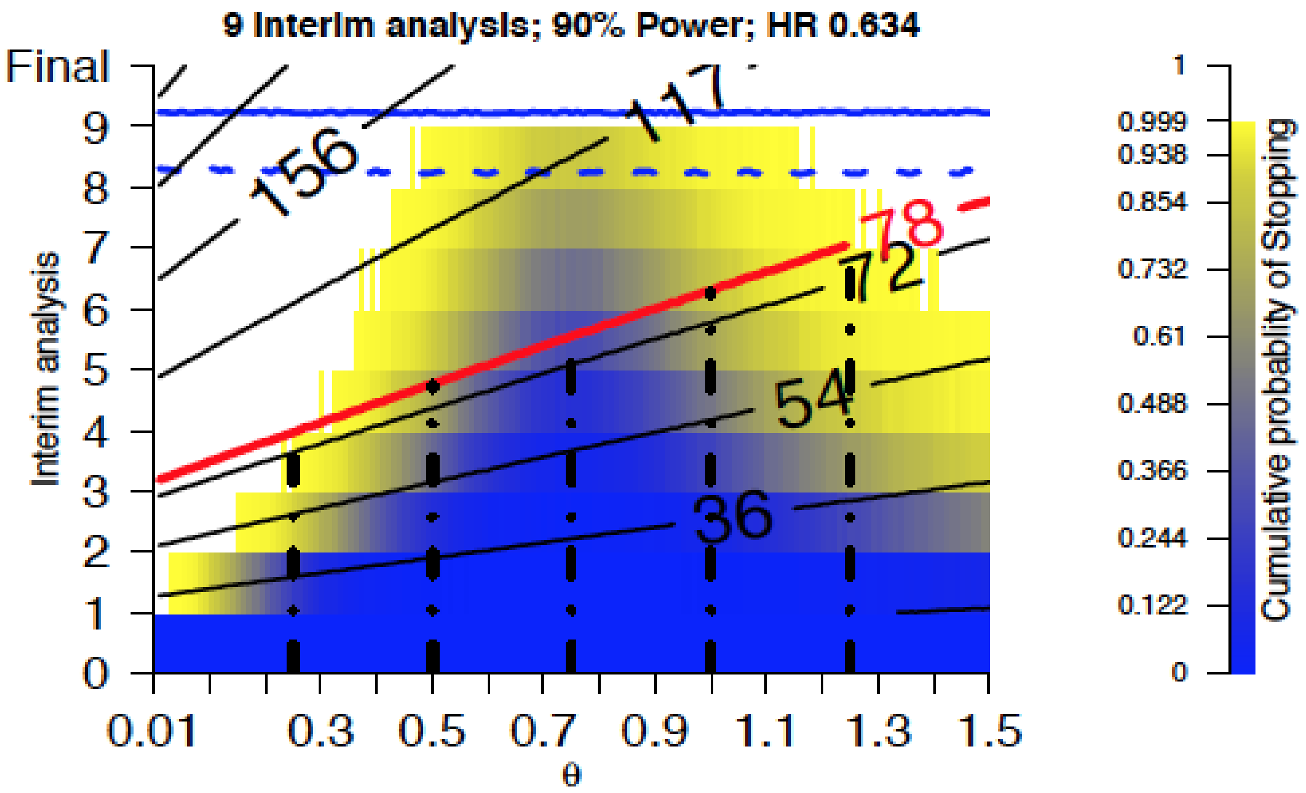
Incorporating Lower Event Rates

- 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.”

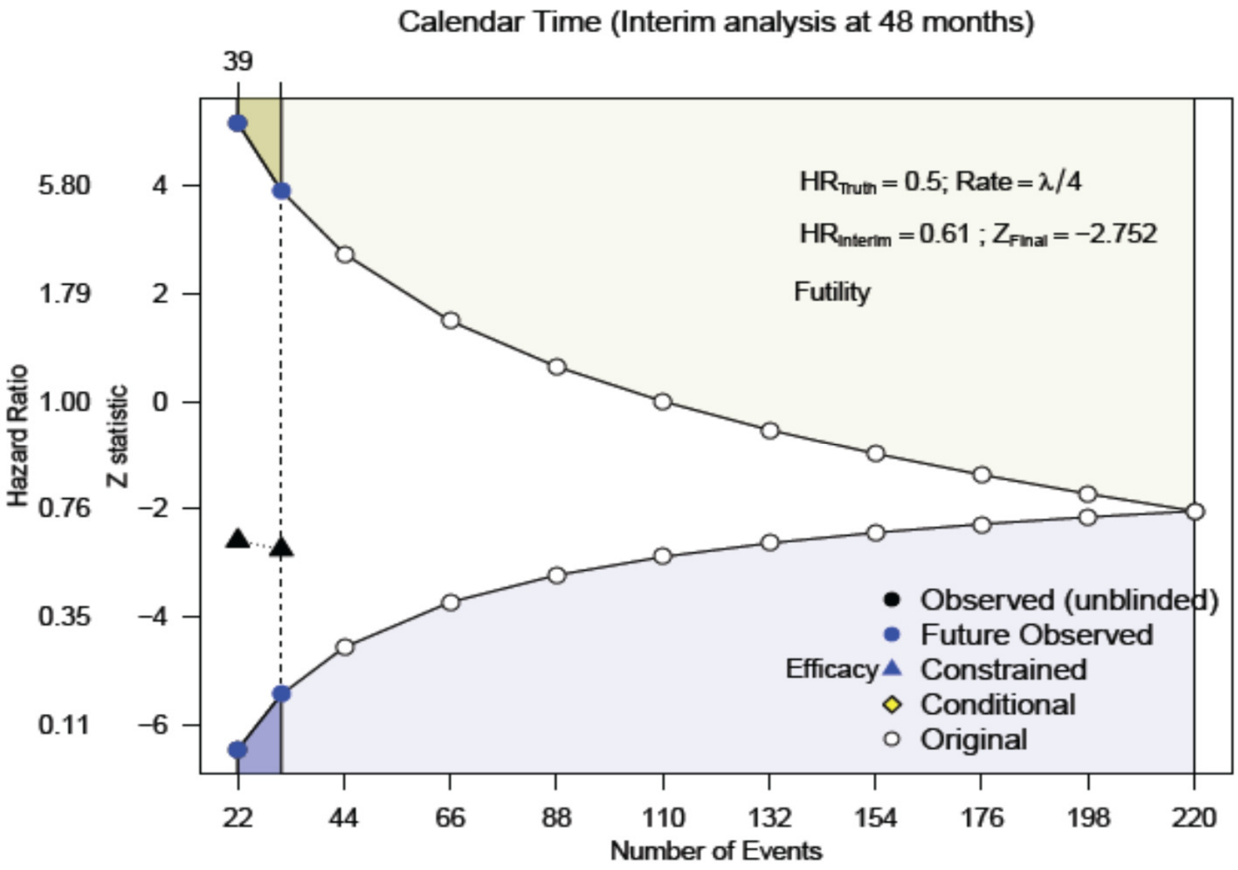
Calendar Time: Half Event Rate

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



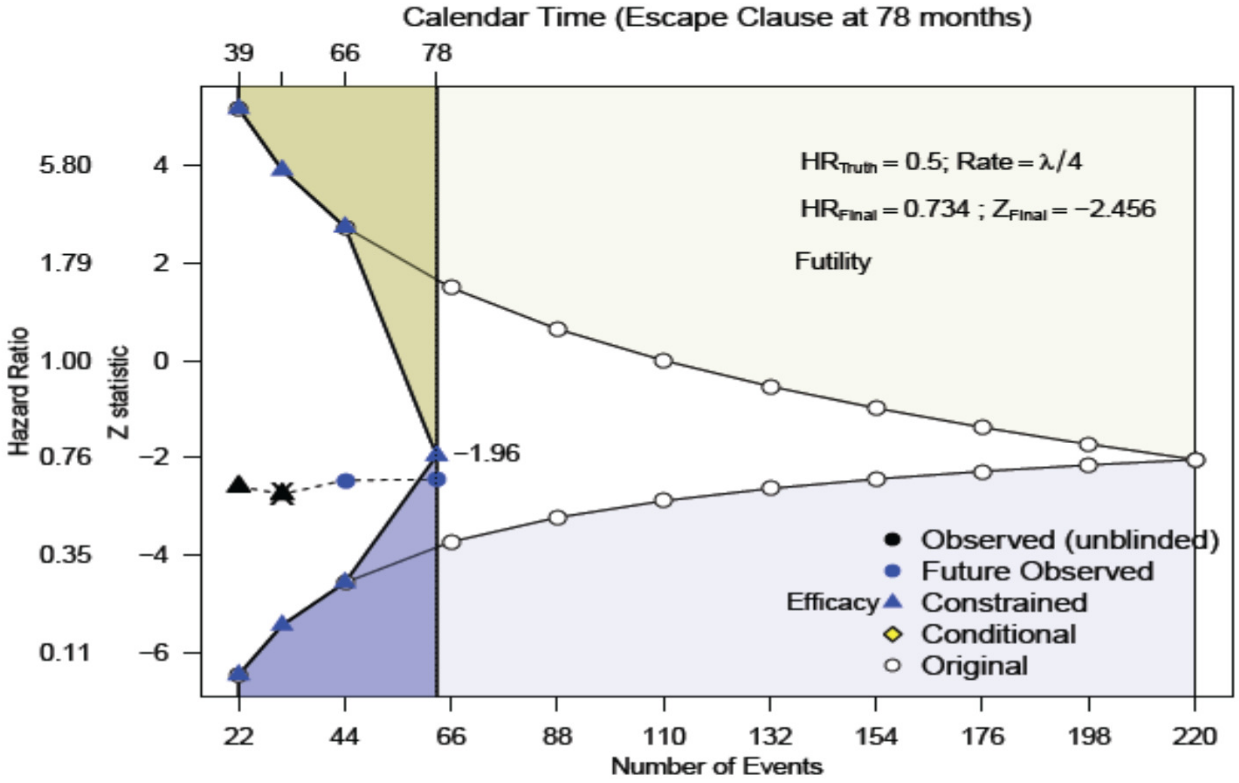
The Escape Clause

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



The Escape Clause

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



Terminate for efficacy at 78 months

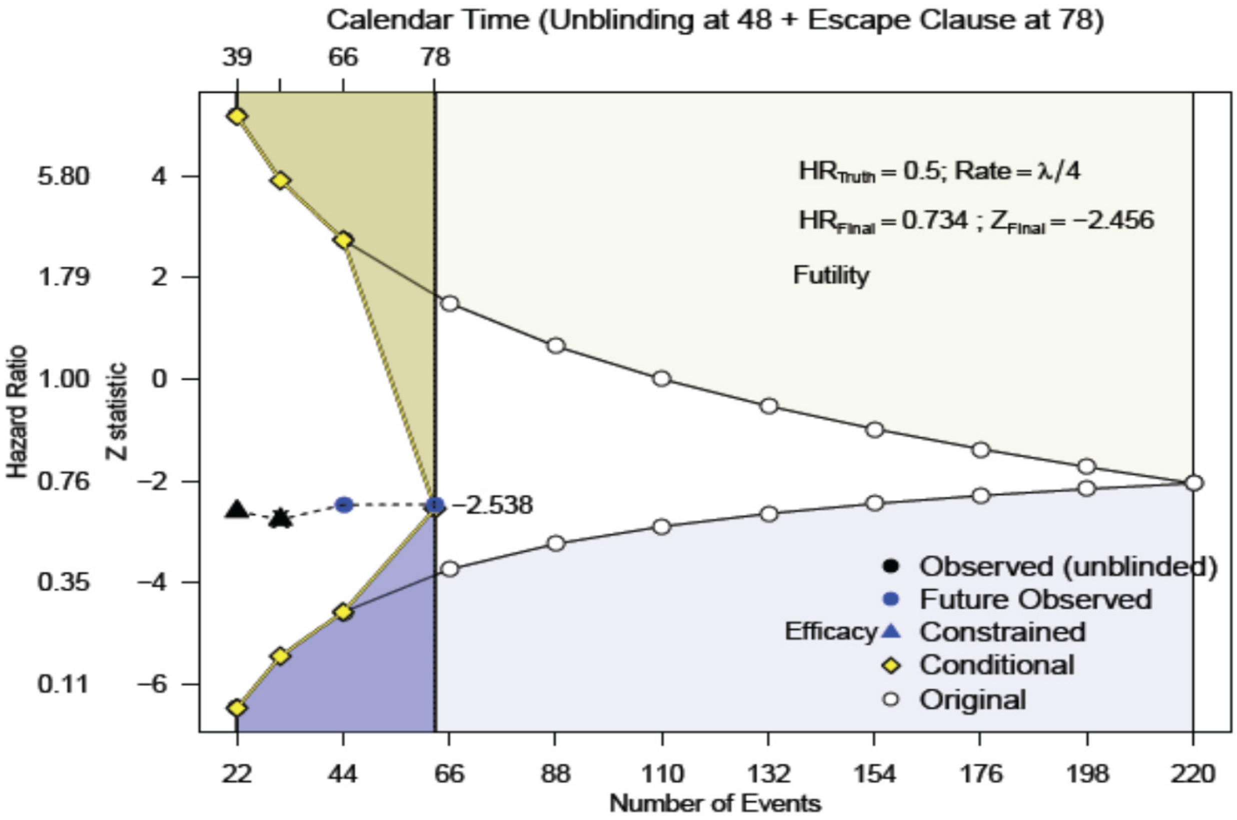
Unblinded Adaptation

- 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

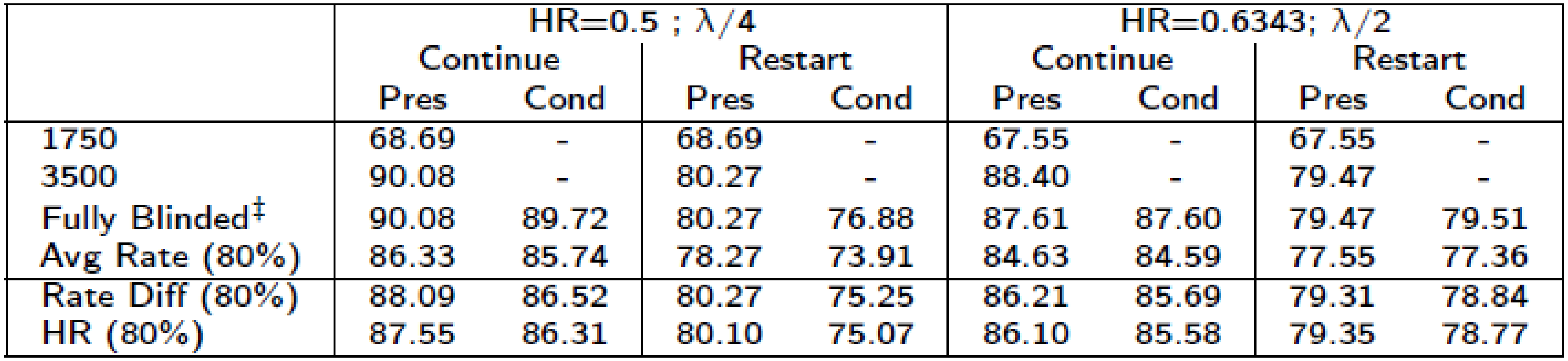
What if Unblinded?

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



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

Simulations



	HR=0.5 ; λ/4				HR=0.6343; λ/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

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

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?

Approaches for Testing

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

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

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

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

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

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

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	$\vec{X}_k^*$	$\vec{X}_k = \left(\vec{X}_1^*, \ldots, \vec{X}_k^*\right)$
1° outcome data	$\vec{Y}_k^*$	$\vec{Y}_k = \left(\vec{Y}_1^*, \ldots, \vec{Y}_k^*\right)$
2° outcome data	$\vec{W}_k^*$	$\vec{W}_k = \left(\vec{W}_1^*, \ldots, \vec{W}_k^*\right)$

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

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

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)

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

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

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

Conditional distributions are totally independent under the null hypothesis

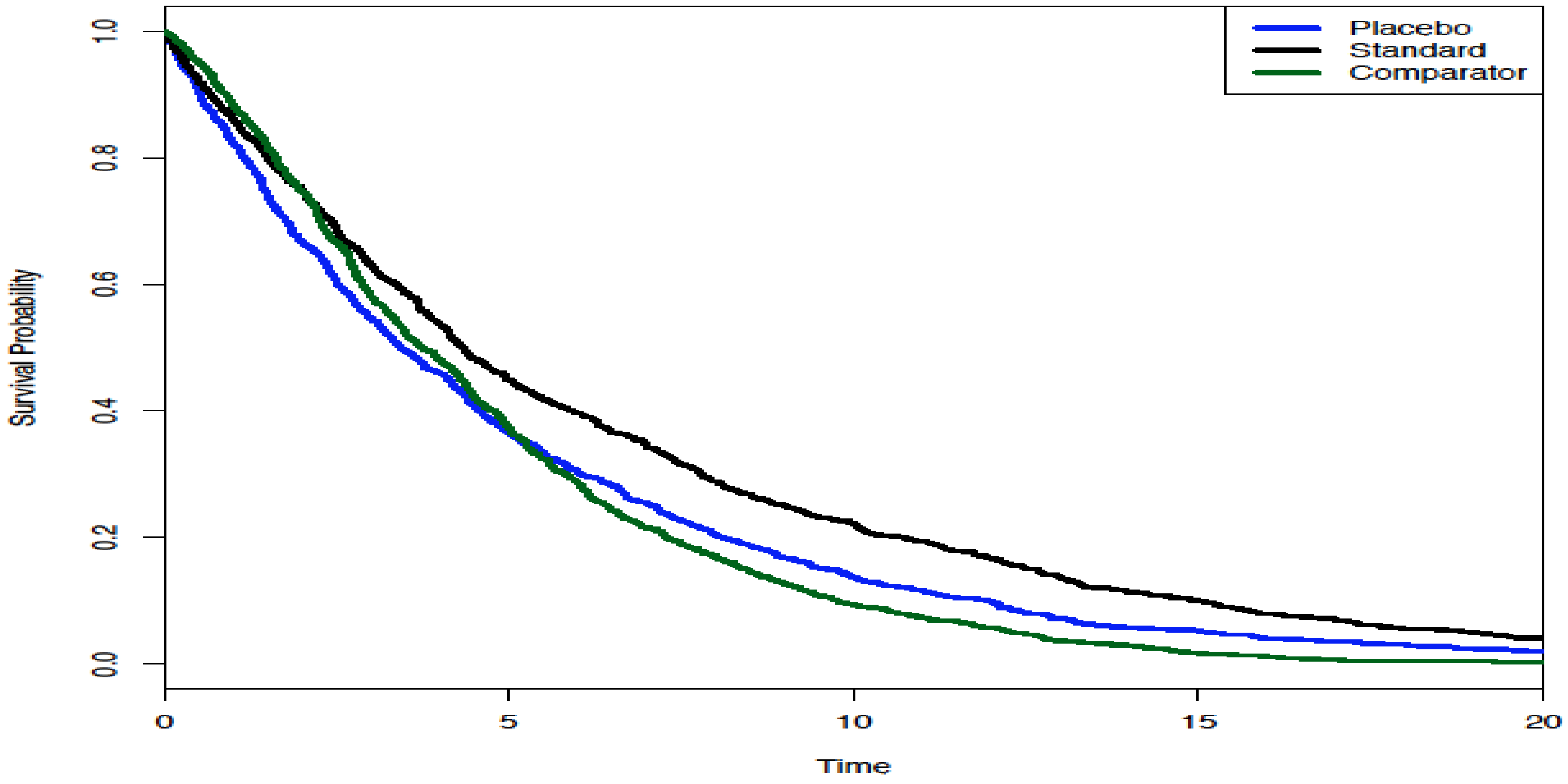
Estimands by Stage: Time to Event

- 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

Impact on Noninferiority Trials

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



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

Protecting Type I Error

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

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

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

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

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

Data at *j*-th Analysis: Delayed Outcome

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

At *k*th interim analysis	Incrementa l	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^{o} outcome data (msng, observed)	$\vec{Y}_k^{*\text{M}}, \vec{Y}_k^{*\text{O}}$	$\vec{Y}_k^{\text{M}}, \vec{Y}_k^{\text{O}}$
2^{o} outcome data	$\vec{W}_k^*$	$\vec{W}_k = \left(\vec{W}_1^*, \ldots, \vec{W}_k^*\right)$
Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*\left(N_k^*, \vec{X}_k^*, \vec{Y}_k^{*O}, \vec{Y}_{k-1}^{M}\right)$	$\hat{\theta}_k = \dfrac{\sum_{j=1}^{k} N_j^* \hat{\theta}_j^*}{N_k}$
Normalized Z statistic	Z_k^*	$Z_k = \dfrac{\sum_{j=1}^{k} \sqrt{N_j^*}\, Z_j^*}{\sqrt{N_k}}$
Fixed sample P value	P_k^*	

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

<u>At *k*th interim analysis</u>	<u>Incrementa l</u>	<u>Cumulative</u>
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	$\hat{\theta}_k^* = \hat{\theta}_k^*\left(N_k^*, \vec{X}_k^*, \vec{Y}_k^{*O}, \vec{Y}_{k-1}^{M}\right)$	$\hat{\theta}_k = \dfrac{\sum_{j=1}^{k} N_j^* \hat{\theta}_j^*}{N_k}$

<u>Impact :</u> (One statistici an's mean is another statistici an's variance)

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

$\hat{\theta}_k^* \mid N_k^*$ is potentiall y biased for θ_k and not approximat ely normal

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

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

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

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

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 :

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

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

“Intent to Cheat” Zone

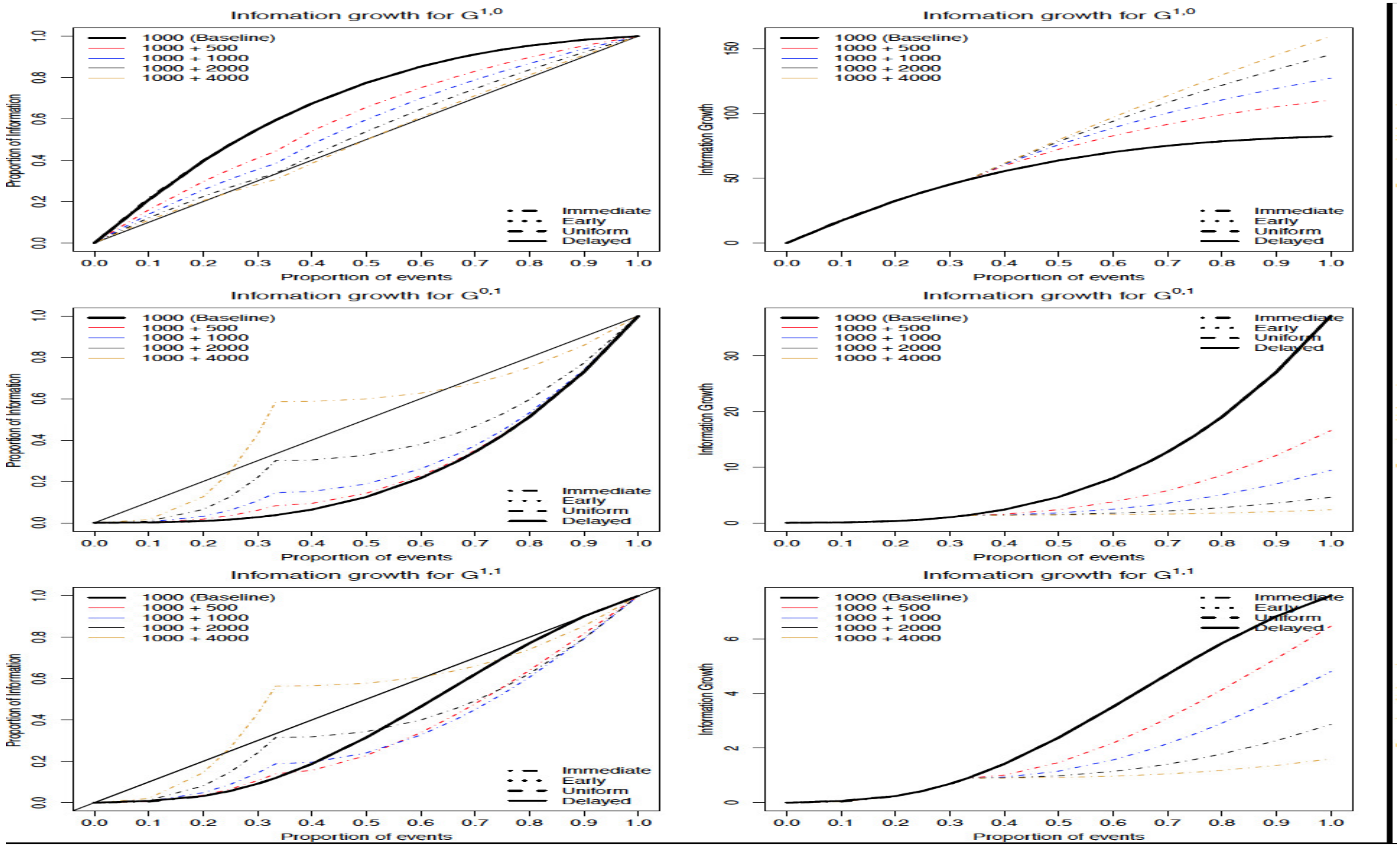
- At interim analysis, choose range of interim estimates that lead to increased accrual of patients

- How bad can we inflate type 1 error when holding number of events constant?

- Logrank test under strong null: Not at all

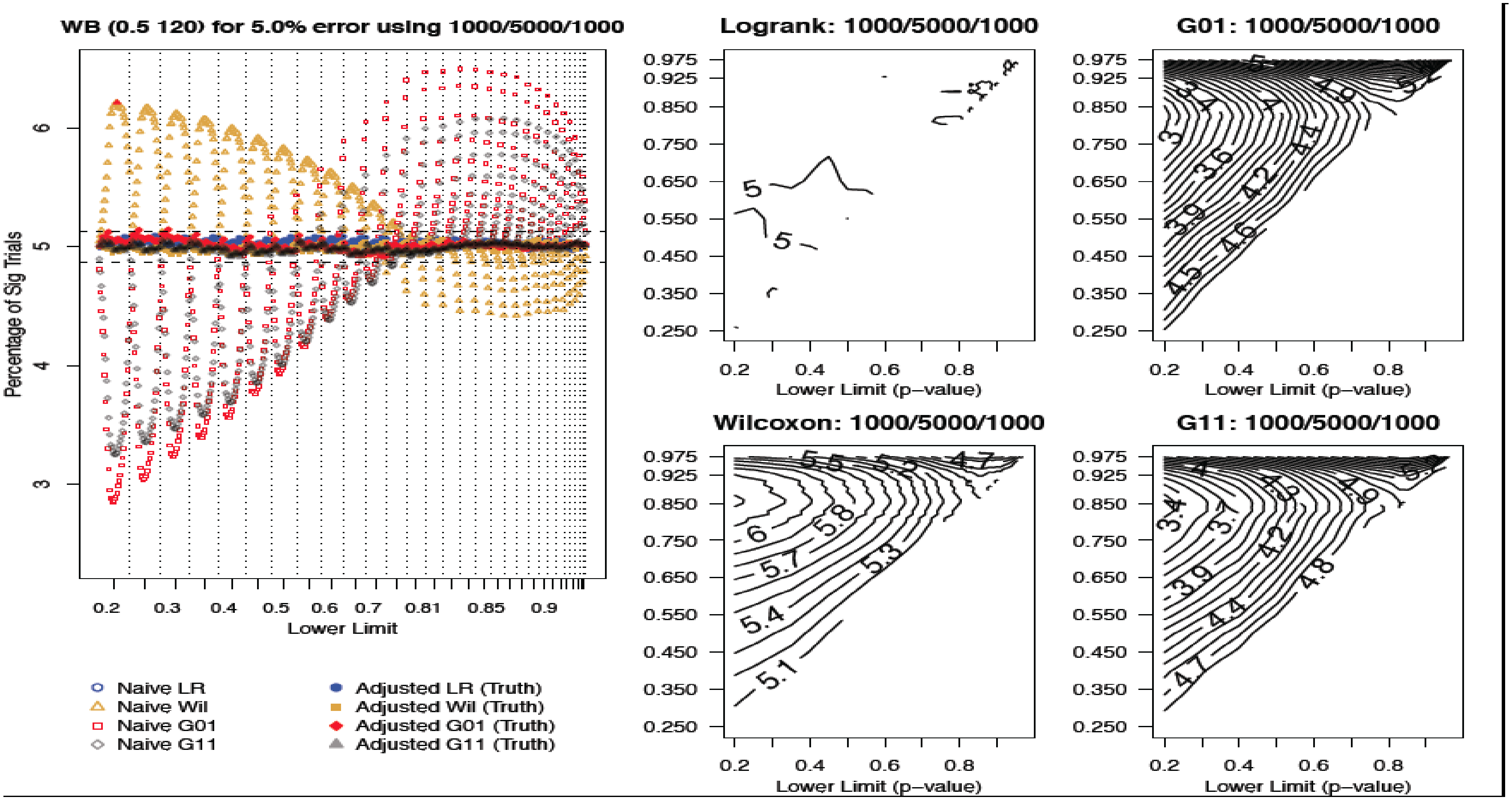
- Weighted logrank tests: Up to relative increase of 20%
  - Sequela of true information growth
    - **Information growth not linear in number of events**
  - Power largely unaffected, so PPV decreases

Information Growth with Adaptation



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

Nonproportional Hazards

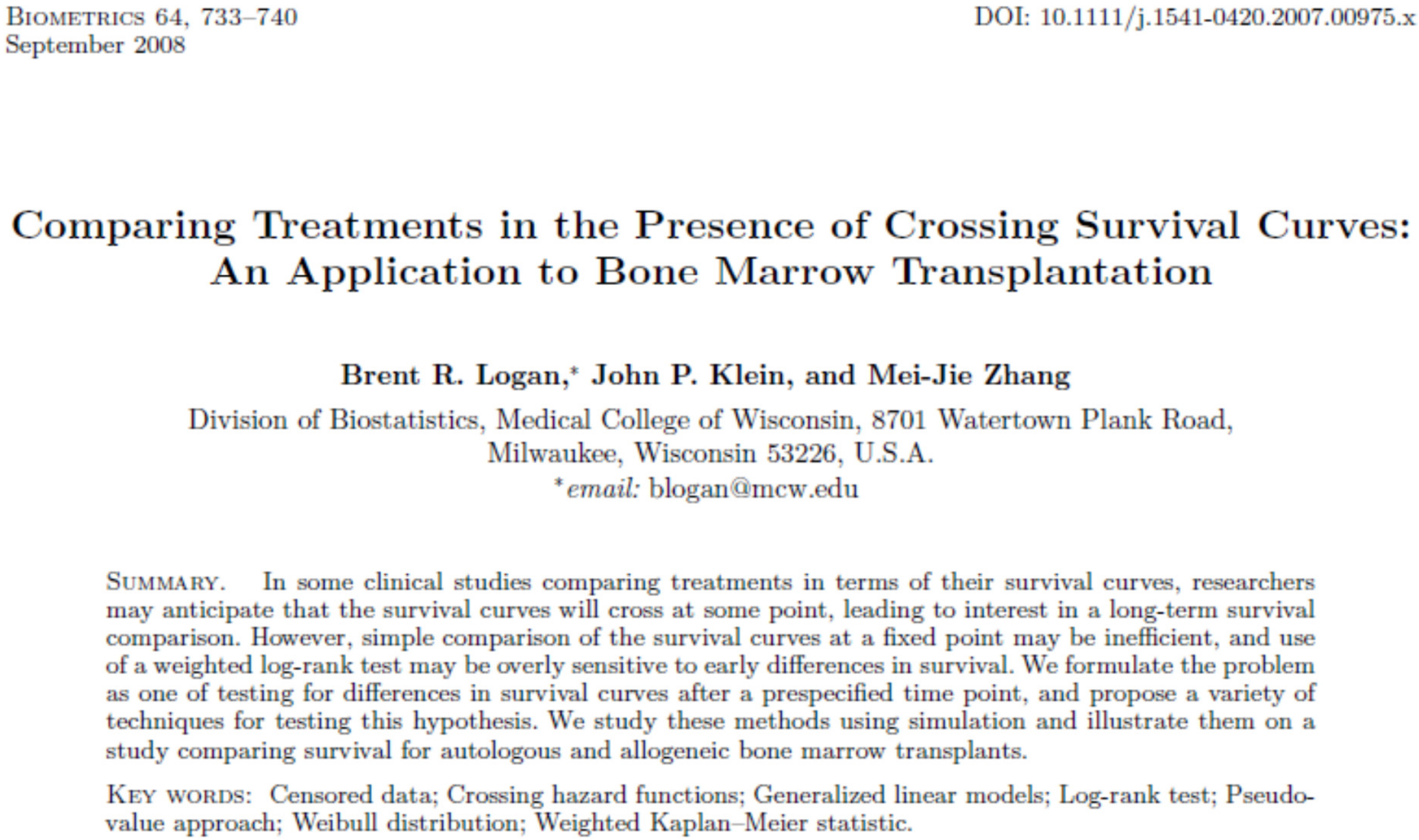
Crossing Survival Curves

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

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.

Logan, et al.: Motivation

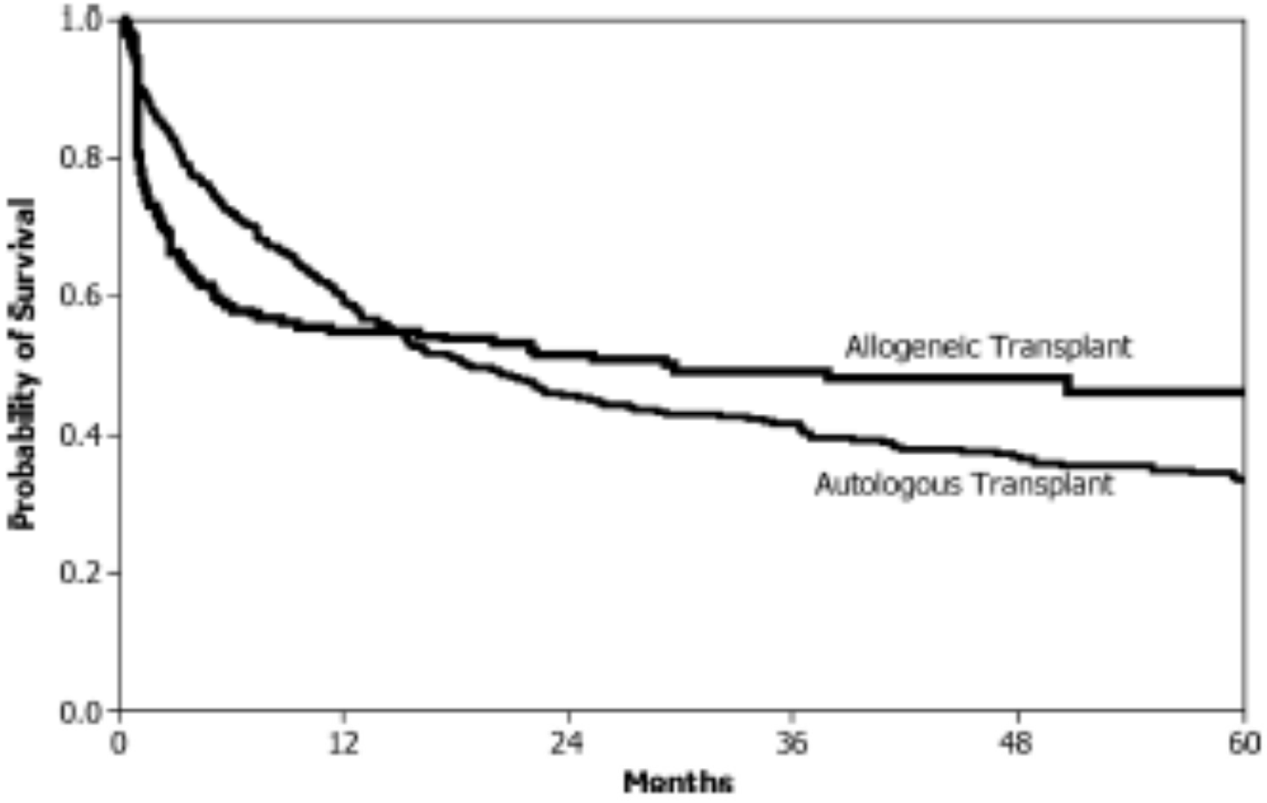


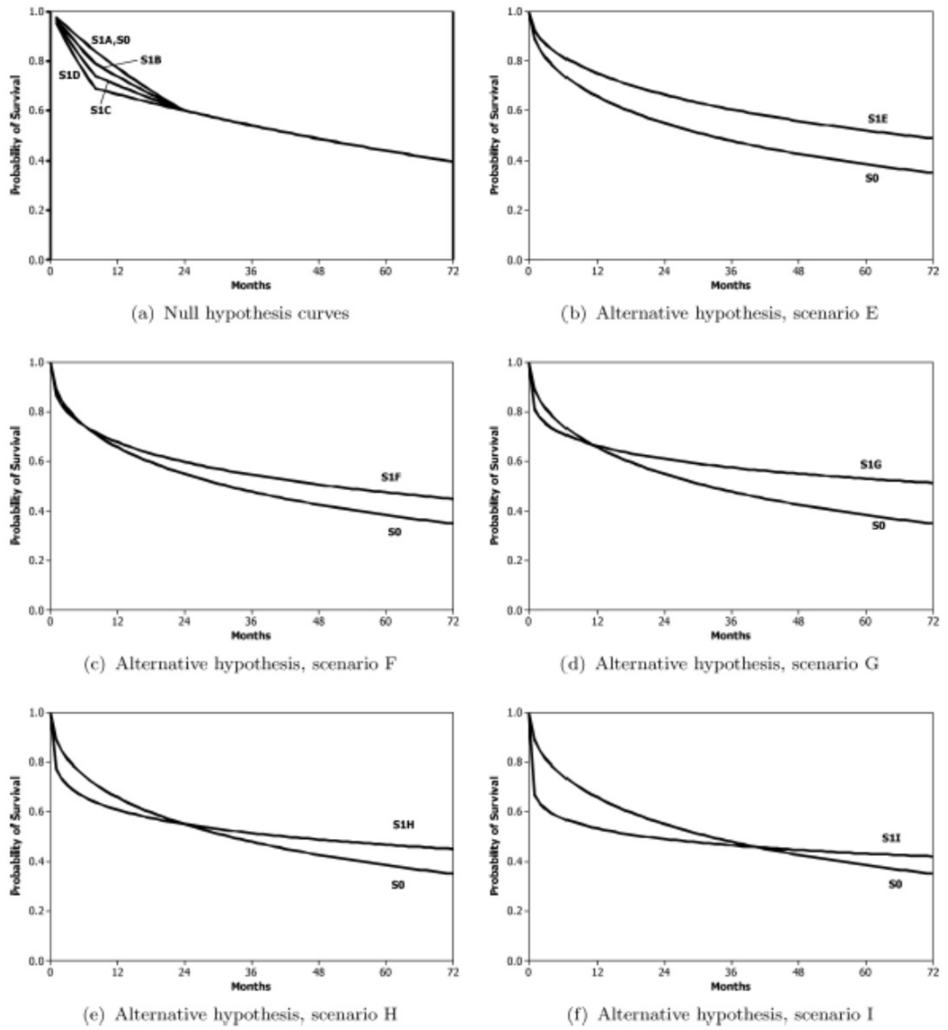
Figure 1. Kaplan–Meier estimate of DFS for follicular lymphoma example, by stem cell source.

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

Logan, et al.: Simulations

Comparing Treatments with Crossing Survival Curves



(a) Null hypothesis curves

(b) Alternative hypothesis, scenario E

(c) Alternative hypothesis, scenario F

(d) Alternative hypothesis, scenario G

(e) Alternative hypothesis, scenario H

(f) Alternative hypothesis, scenario I

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

Logan, et al.: Results

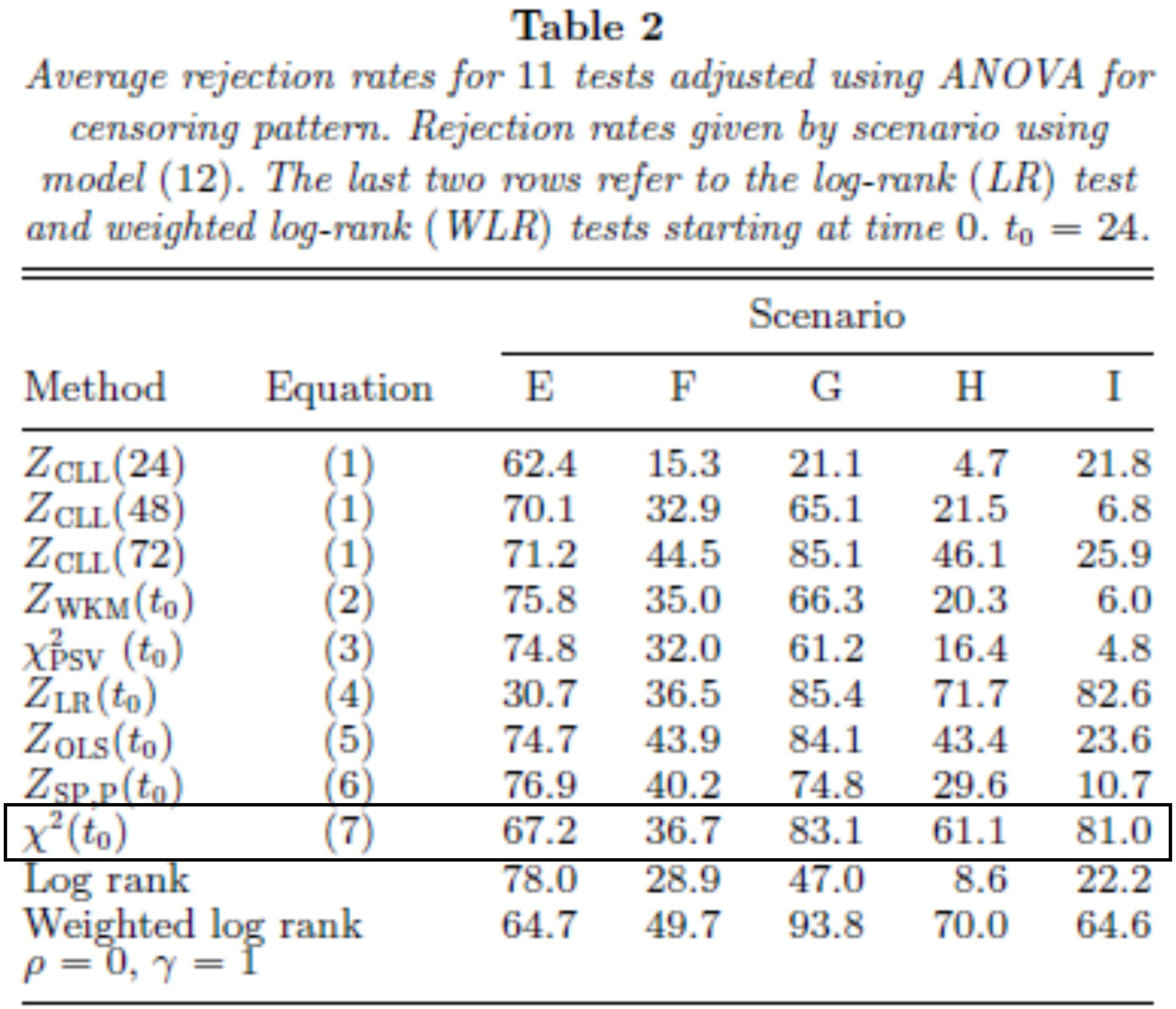


Table 2

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. $t_0 = 24$.

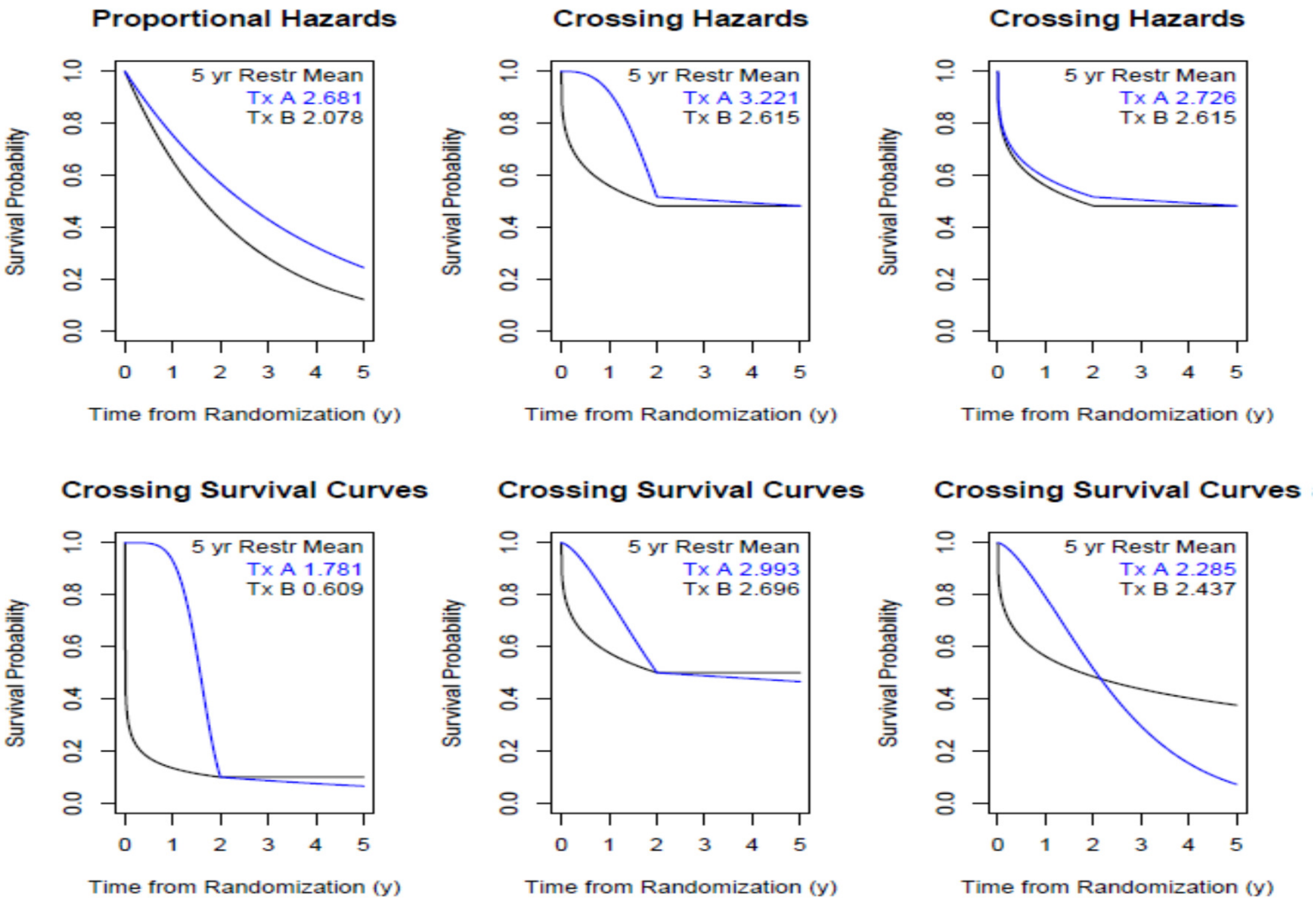
		Scenario				
Method	Equation	E	F	G	H	I
$Z_{\text{CLL}}(24)$	(1)	62.4	15.3	21.1	4.7	21.8
$Z_{\text{CLL}}(48)$	(1)	70.1	32.9	65.1	21.5	6.8
$Z_{\text{CLL}}(72)$	(1)	71.2	44.5	85.1	46.1	25.9
$Z_{\text{WKM}}(t_0)$	(2)	75.8	35.0	66.3	20.3	6.0
$\chi^2_{\text{PSV}}(t_0)$	(3)	74.8	32.0	61.2	16.4	4.8
$Z_{\text{LR}}(t_0)$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

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

Logan, et al.: Critique

- Additional scenarios that are of interest



Logan, et al.: Critique

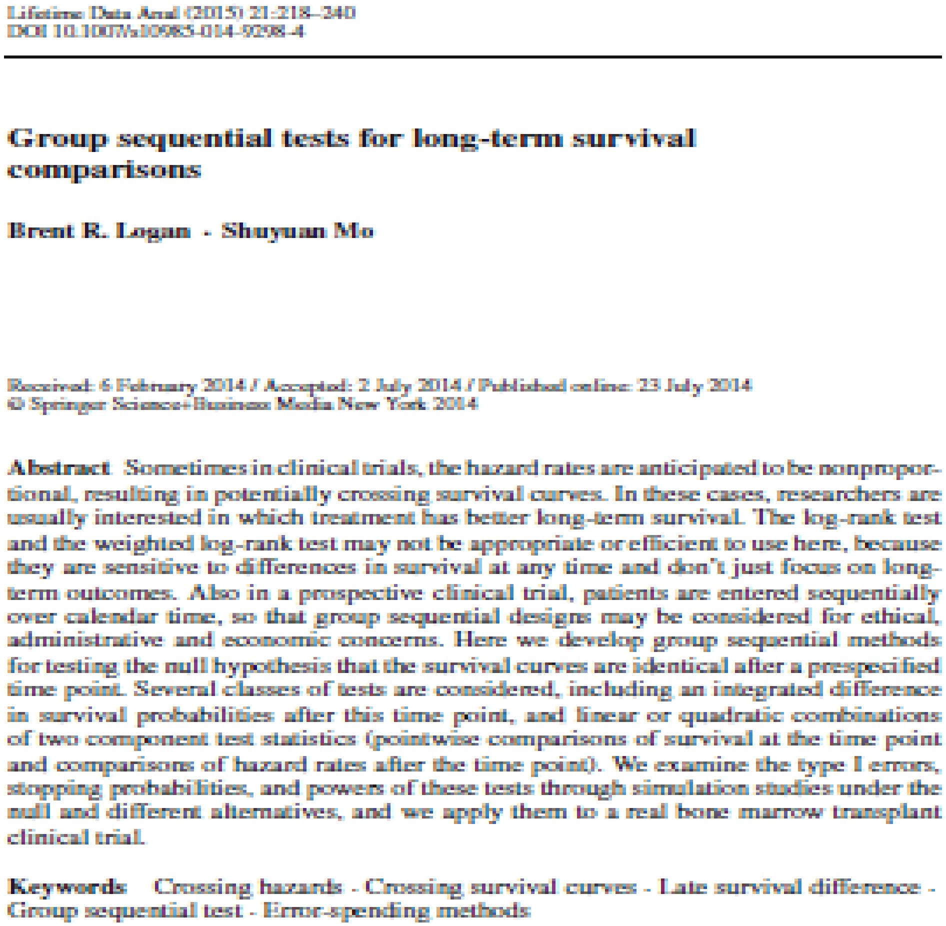
- How might a naïve investigator use this test?
  - If the observed survival curves cross and the hazard is significantly higher after that point, the presumption might be that we have significant evidence that the group with higher hazard at later times has worse survival at those times

- "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%

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

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?)

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.

Really Bottom Line

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

-Aretha Franklin, "Think"