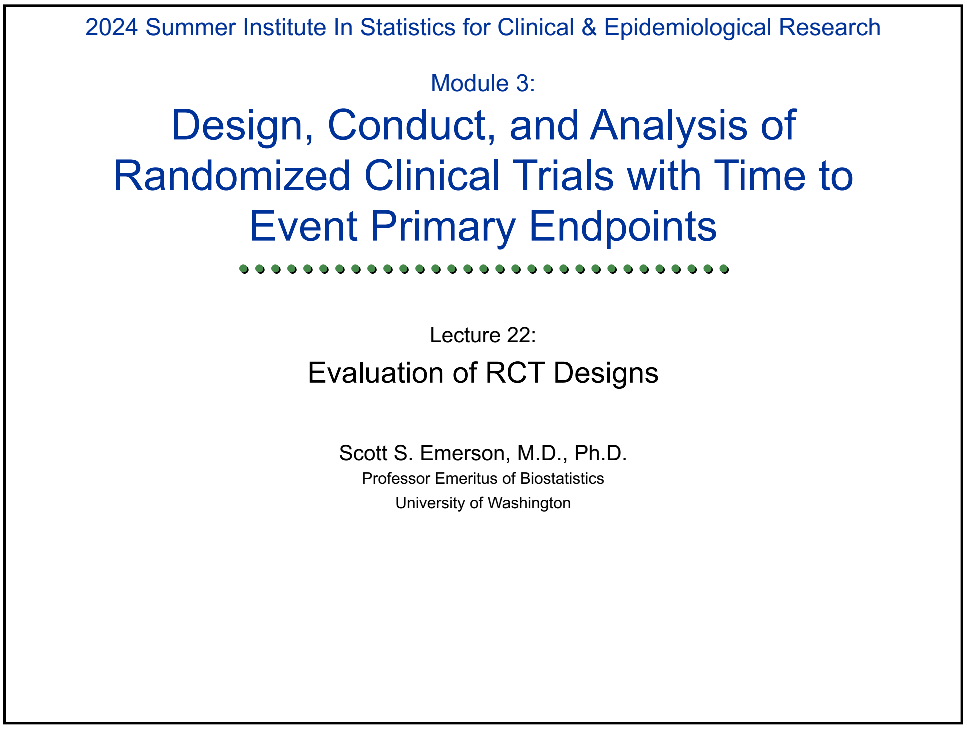


2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

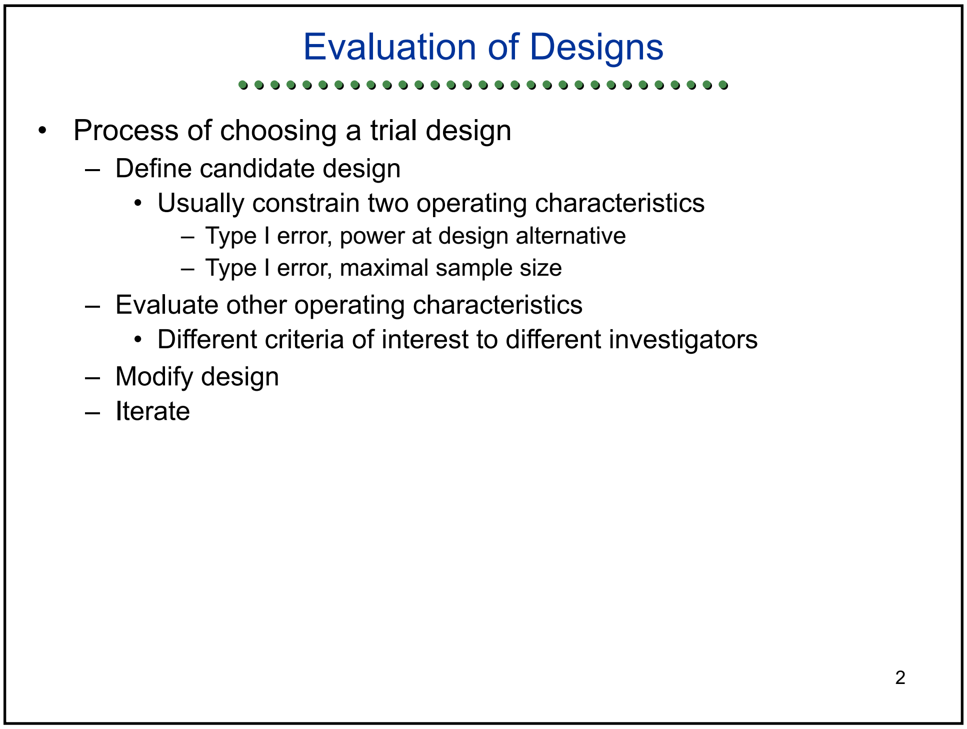
Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 22:

Evaluation of RCT Designs

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

1

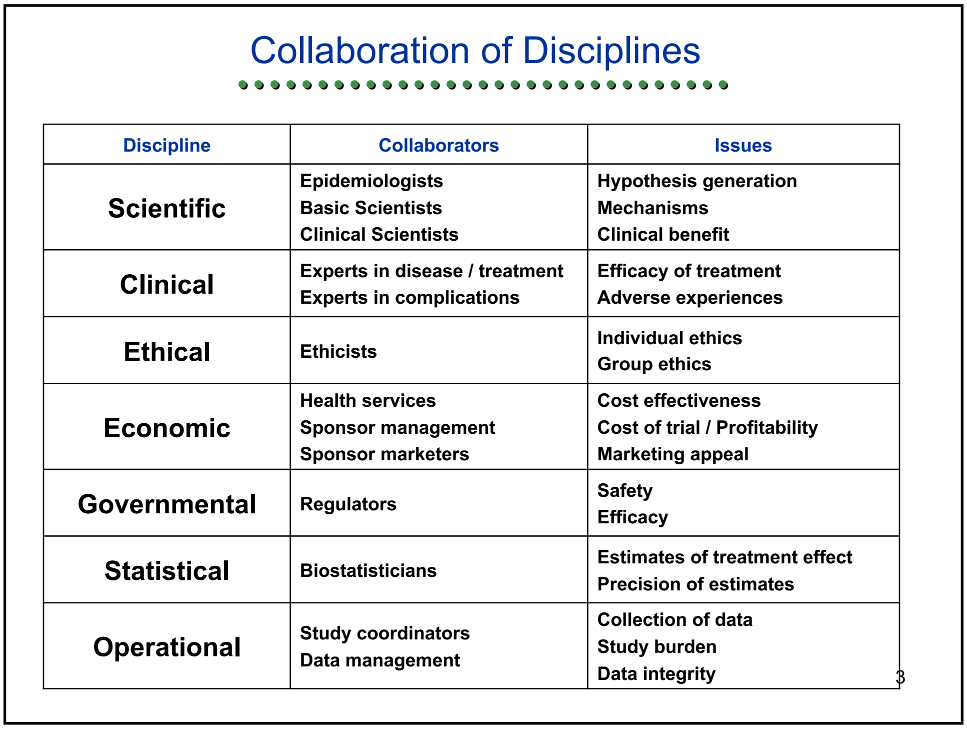


Evaluation of Designs

- 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

2

2

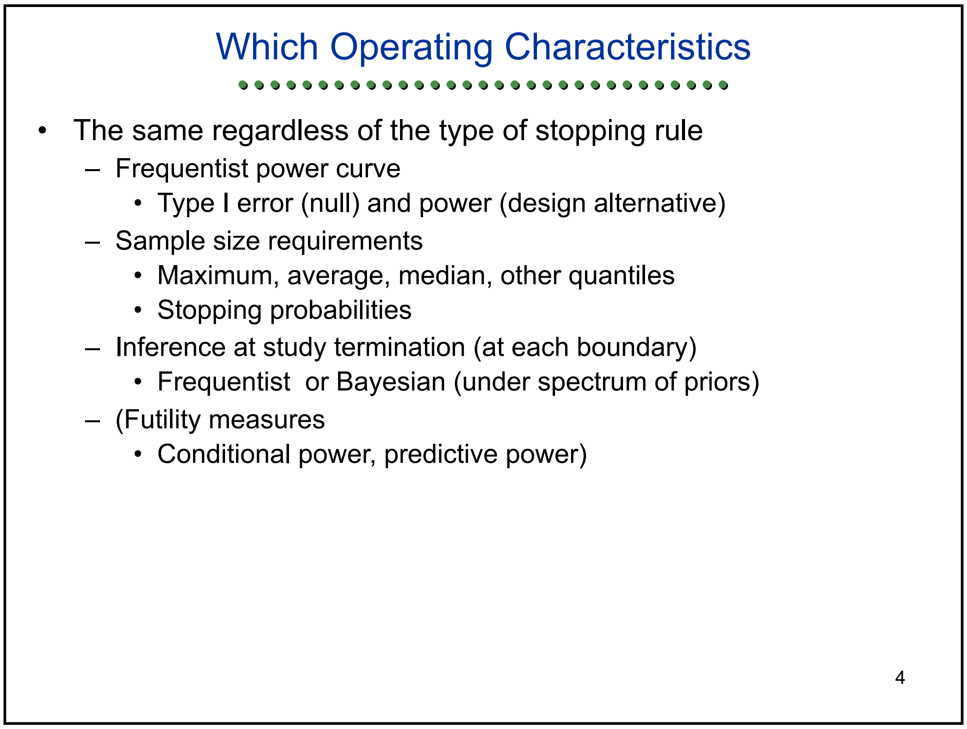


Collaboration of Disciplines

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

3

3

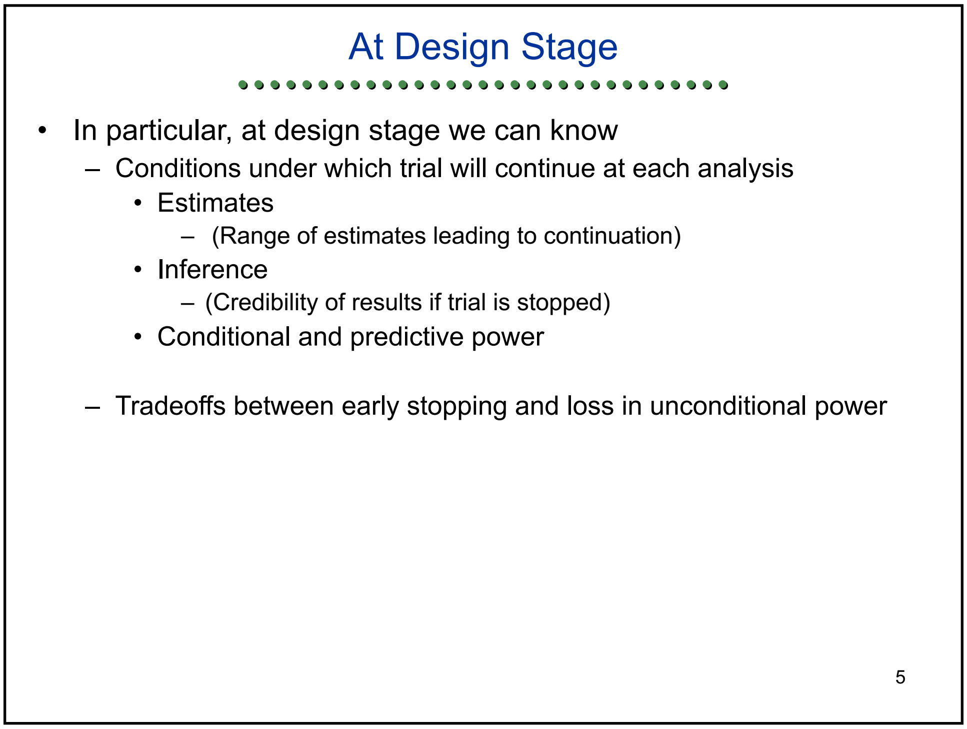


Which Operating Characteristics

- 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)

4

4

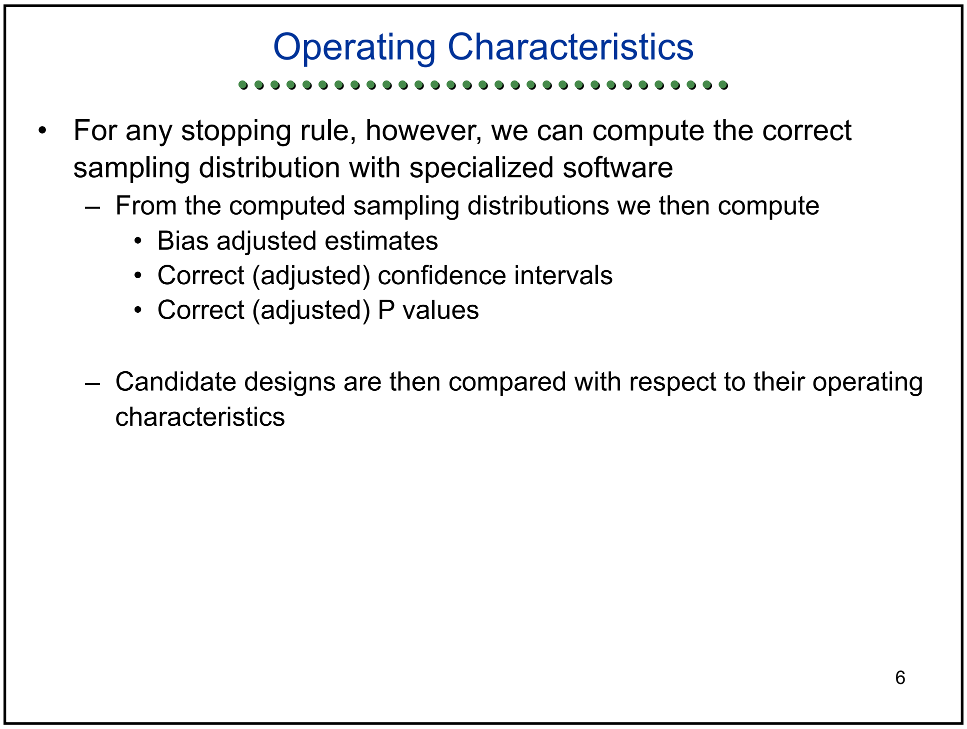


At Design Stage

- 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

5

5

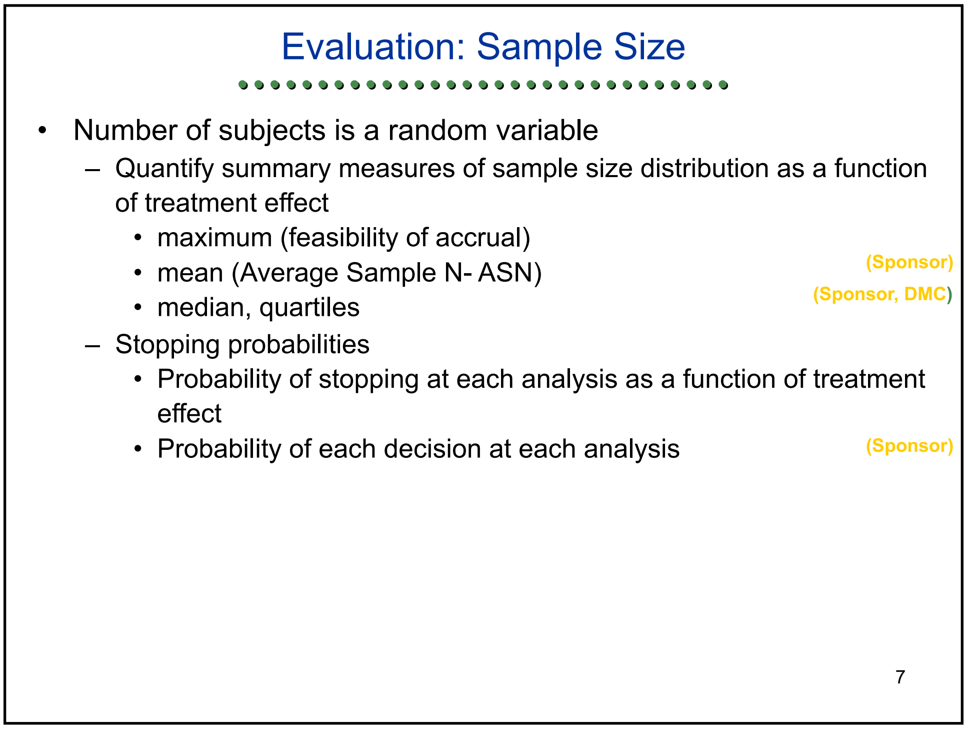


Operating Characteristics

- For any 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

6

6

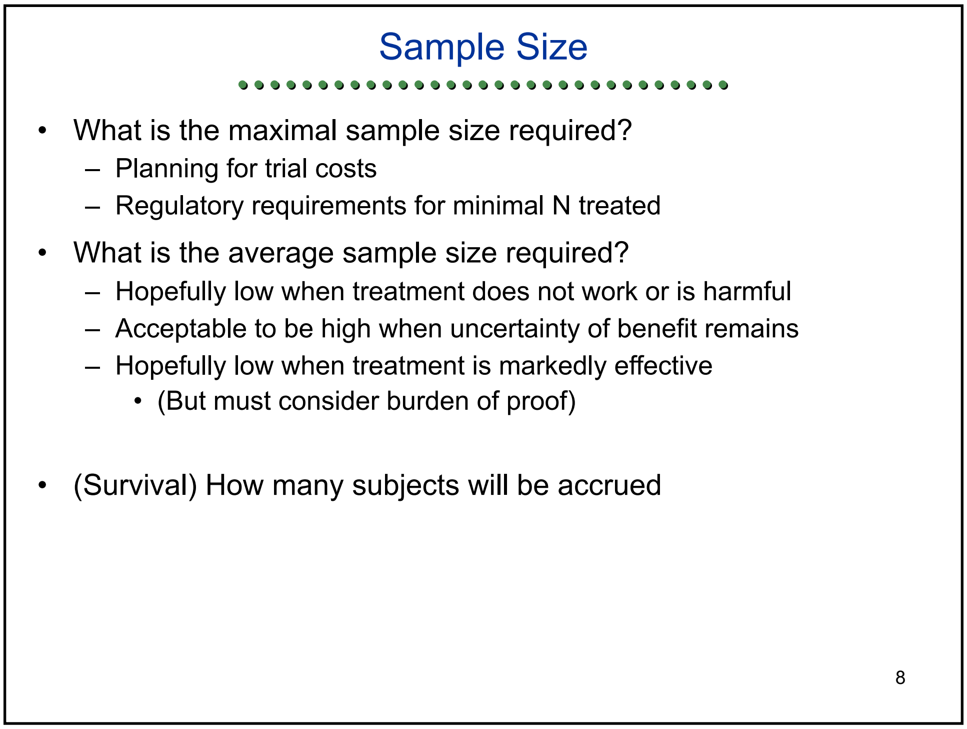


Evaluation: Sample Size

- Number of subjects is a random variable
  - Quantify summary measures of sample size distribution as a function of treatment effect
    - maximum (feasibility of accrual)
    - mean (Average Sample N- ASN) (Sponsor)
    - median, quartiles (Sponsor, DMC)
  - Stopping probabilities
    - Probability of stopping at each analysis as a function of treatment effect
    - Probability of each decision at each analysis (Sponsor)

7

7



Sample Size

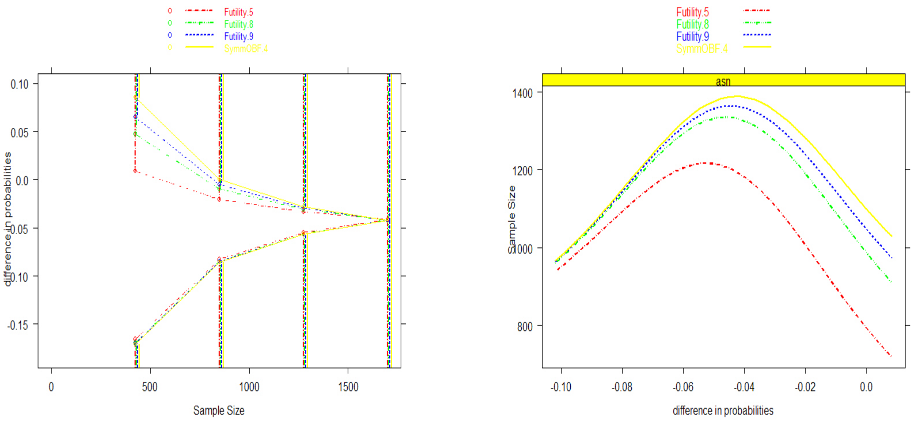
- What is the maximal sample size required?
  - Planning for trial costs
  - Regulatory requirements for minimal N treated
- What is the average sample size required?
  - Hopefully low when treatment does not work or is harmful
  - Acceptable to be high when uncertainty of benefit remains
  - Hopefully low when treatment is markedly effective
    - (But must consider burden of proof)
- (Survival) How many subjects will be accrued

8

8

ASN Curve

- Expected sample size as function of true effect



9

Evaluation: Power Curve

- Probability of rejecting null for arbitrary alternatives
  - Level of significance (power under null)
  - Power for specified alternative

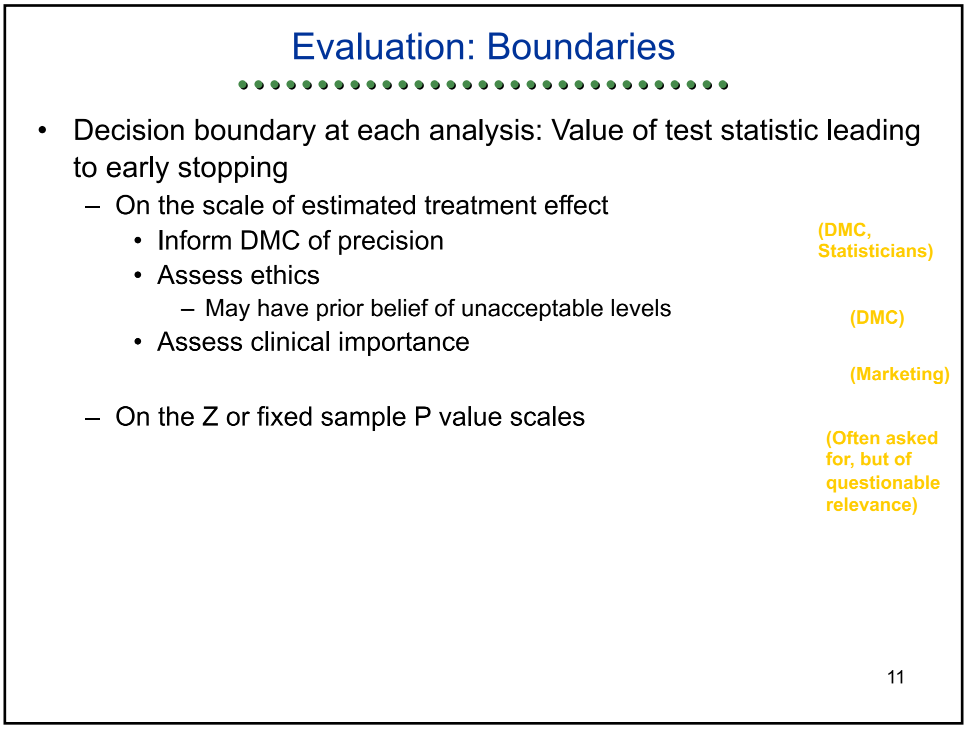
(Regulatory)

  - Alternative rejected by design
    - Alternative for which study has high power
      - Interpretation of negative studies

(Scientists)

10

10



Evaluation: Boundaries

- Decision boundary at each analysis: Value of test statistic leading to early stopping
  - On the scale of estimated treatment effect
    - Inform DMC of precision
    - Assess ethics
      - May have prior belief of unacceptable levels
    - Assess clinical importance
  - On the Z or fixed sample P value scales

(DMC, Statisticians)

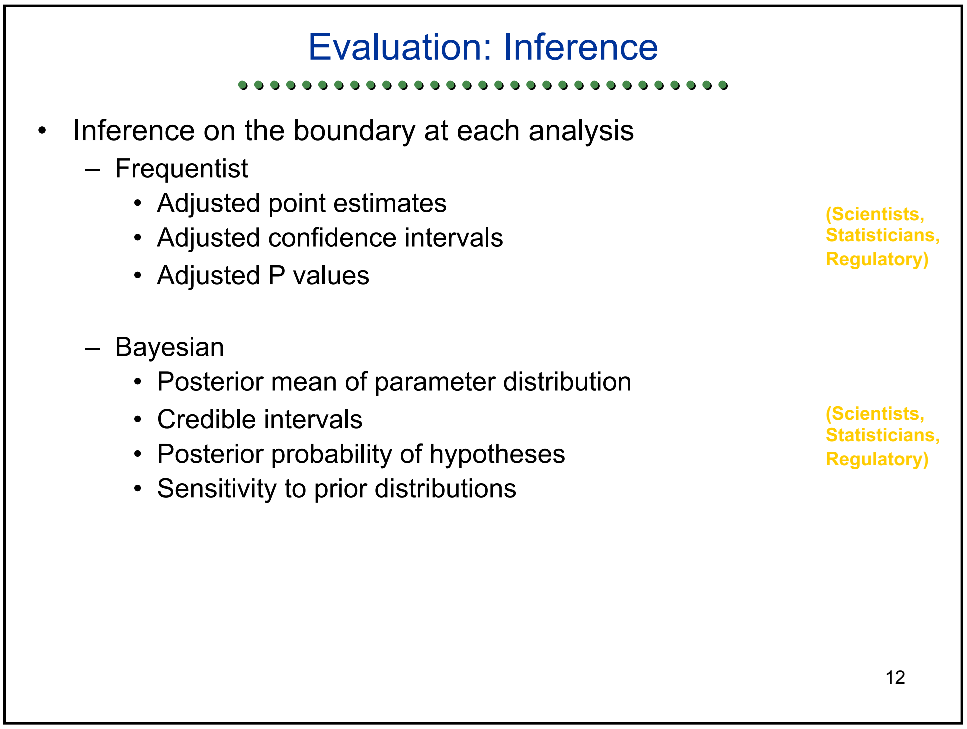
(DMC)

(Marketing)

(Often asked for, but of questionable relevance)

11

11



Evaluation: Inference

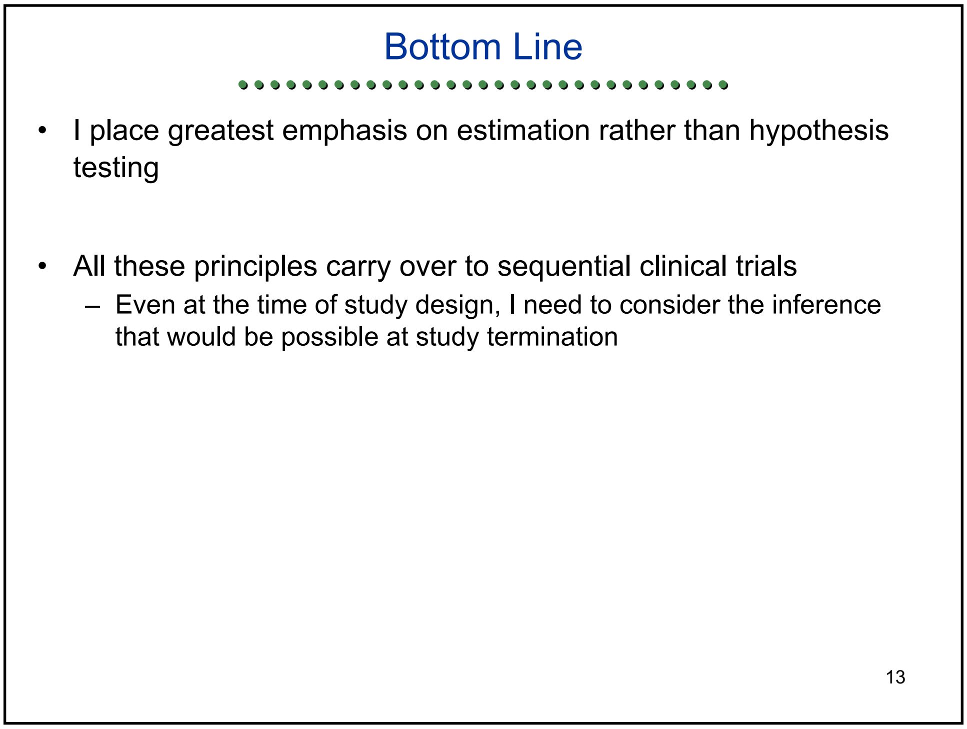
- Inference on the boundary at each analysis
  - Frequentist
    - Adjusted point estimates
    - Adjusted confidence intervals
    - Adjusted P values
  - Bayesian
    - Posterior mean of parameter distribution
    - Credible intervals
    - Posterior probability of hypotheses
    - Sensitivity to prior distributions

(Scientists, Statisticians, Regulatory)

(Scientists, Statisticians, Regulatory)

12

12

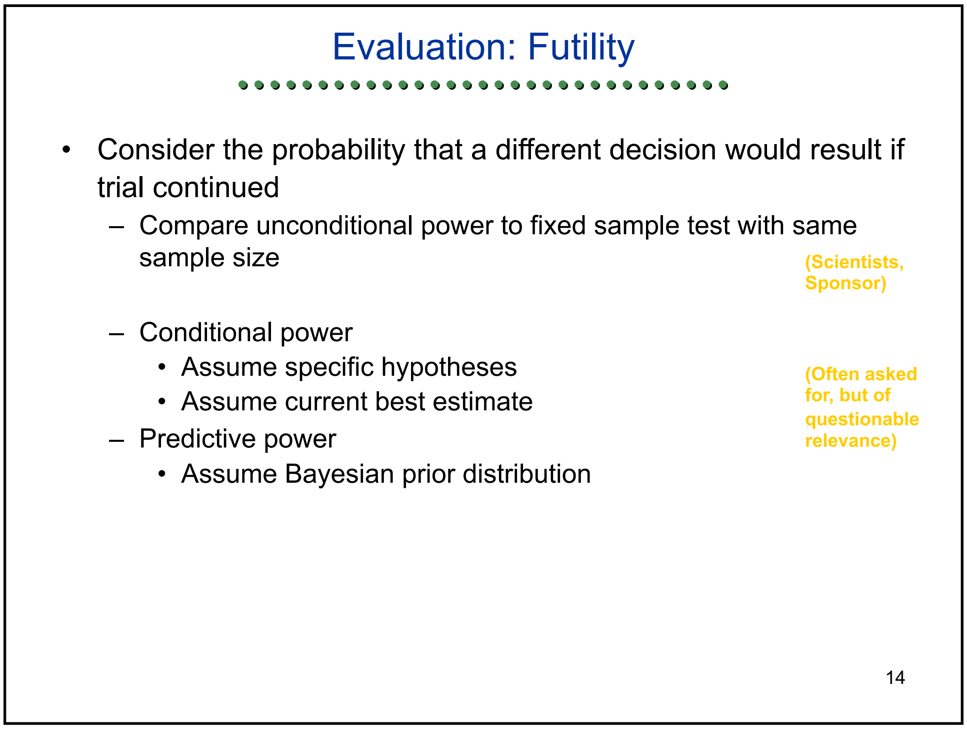


Bottom Line

- I place greatest emphasis on estimation rather than hypothesis testing
- All these principles carry over to sequential clinical trials
  - Even at the time of study design, I need to consider the inference that would possible at study termination

13

13



Evaluation: Futility

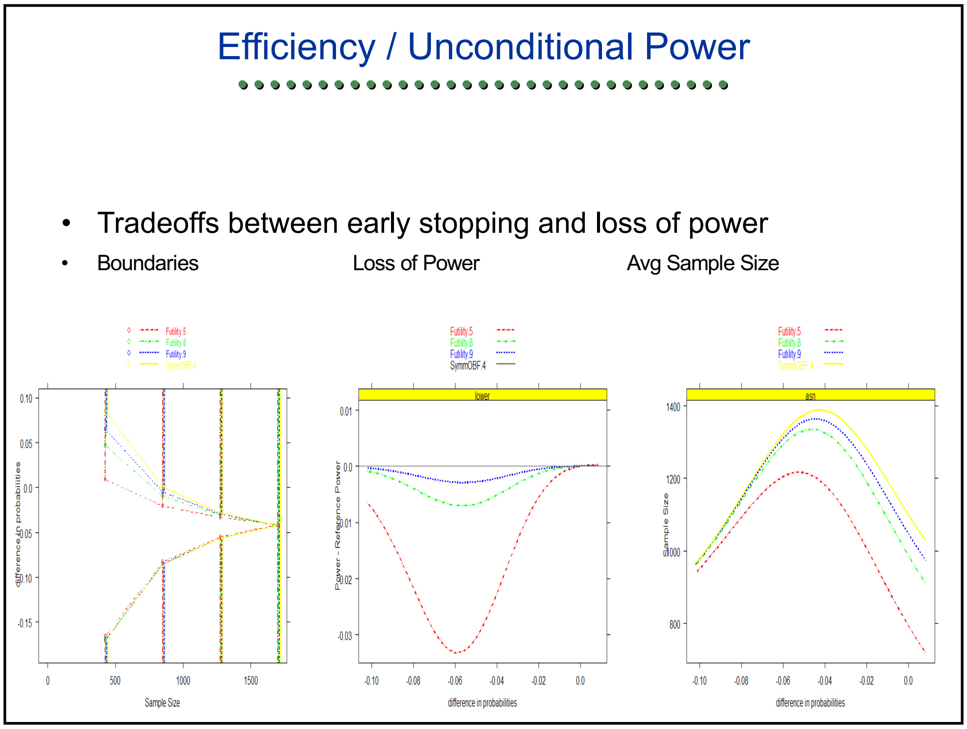
- Consider the probability that a different decision would result if trial continued
  - Compare unconditional power to fixed sample test with same sample size
  - Conditional power
    - Assume specific hypotheses
    - Assume current best estimate
  - Predictive power
    - Assume Bayesian prior distribution

(Scientists, Sponsor)

(Often asked for, but of questionable relevance)

14

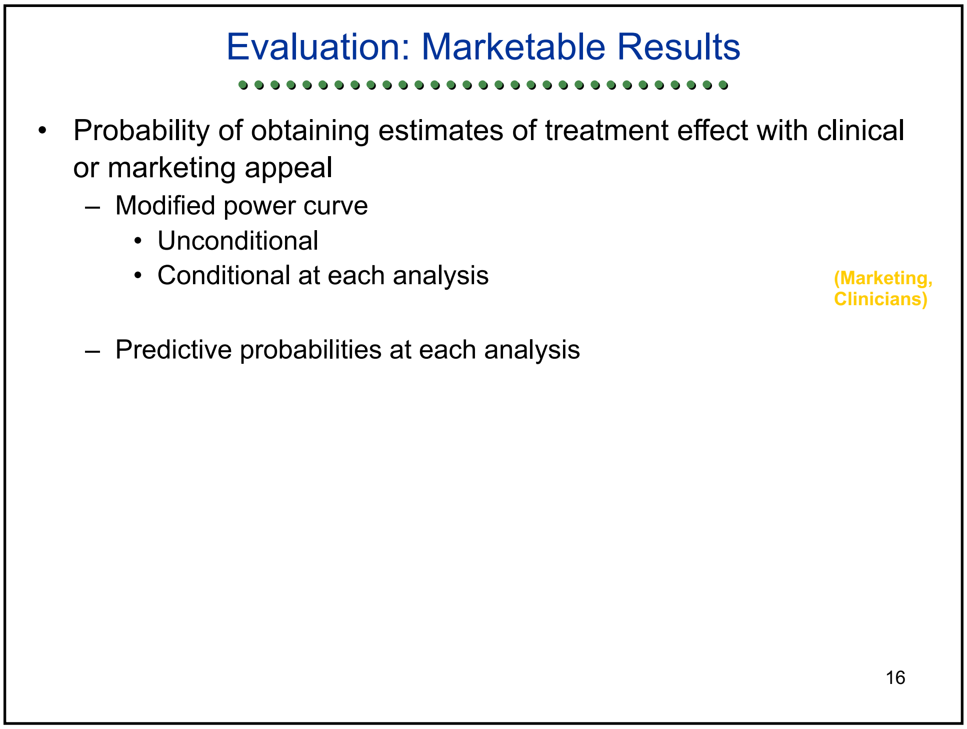
14



Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power
- Boundaries Loss of Power Avg Sample Size

15



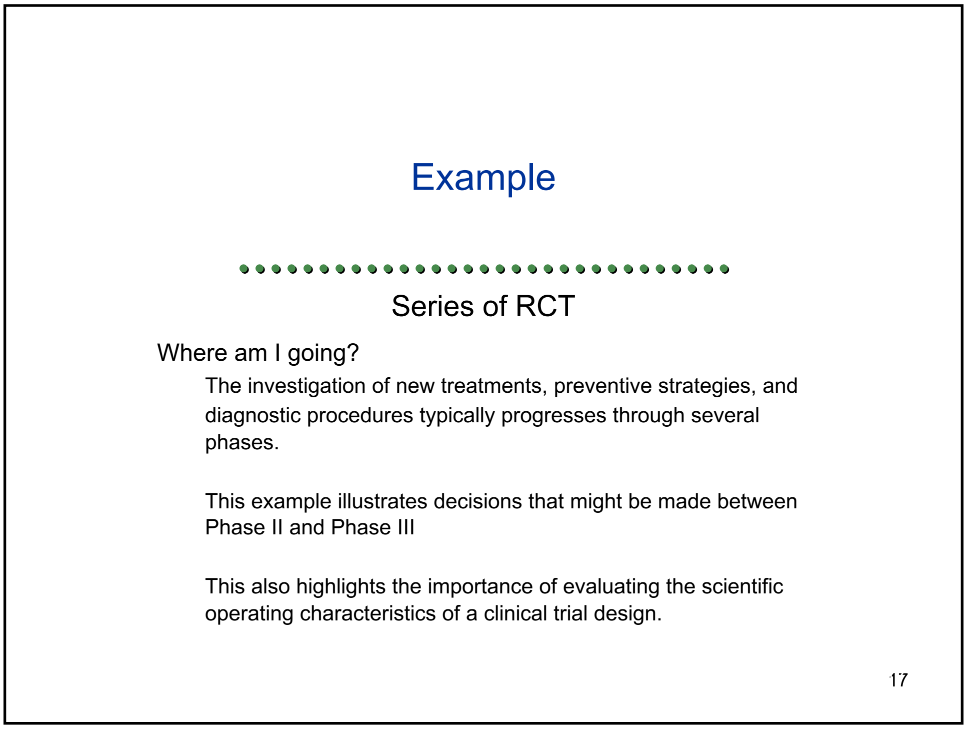
Evaluation: Marketable Results

- Probability of obtaining estimates of treatment effect with clinical or marketing appeal
  - Modified power curve
    - Unconditional
    - Conditional at each analysis
  - Predictive probabilities at each analysis

(Marketing, Clinicians)

16

16



Example

Series of RCT

Where am I going?

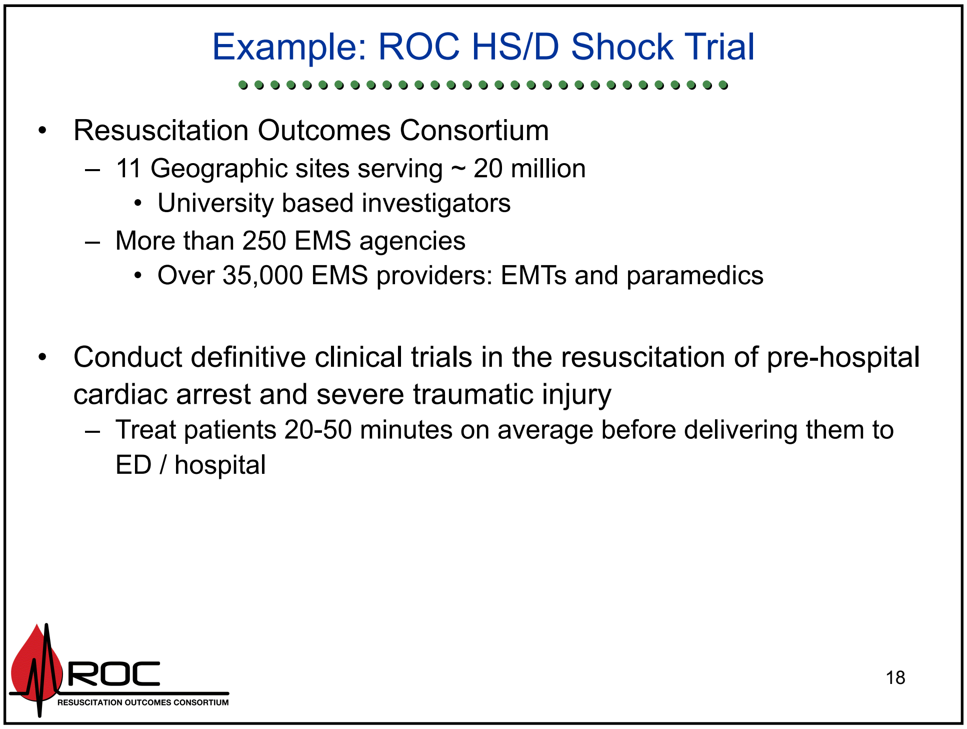
The investigation of new treatments, preventive strategies, and diagnostic procedures typically progresses through several phases.

This example illustrates decisions that might be made between Phase II and Phase III

This also highlights the importance of evaluating the scientific operating characteristics of a clinical trial design.

17

17



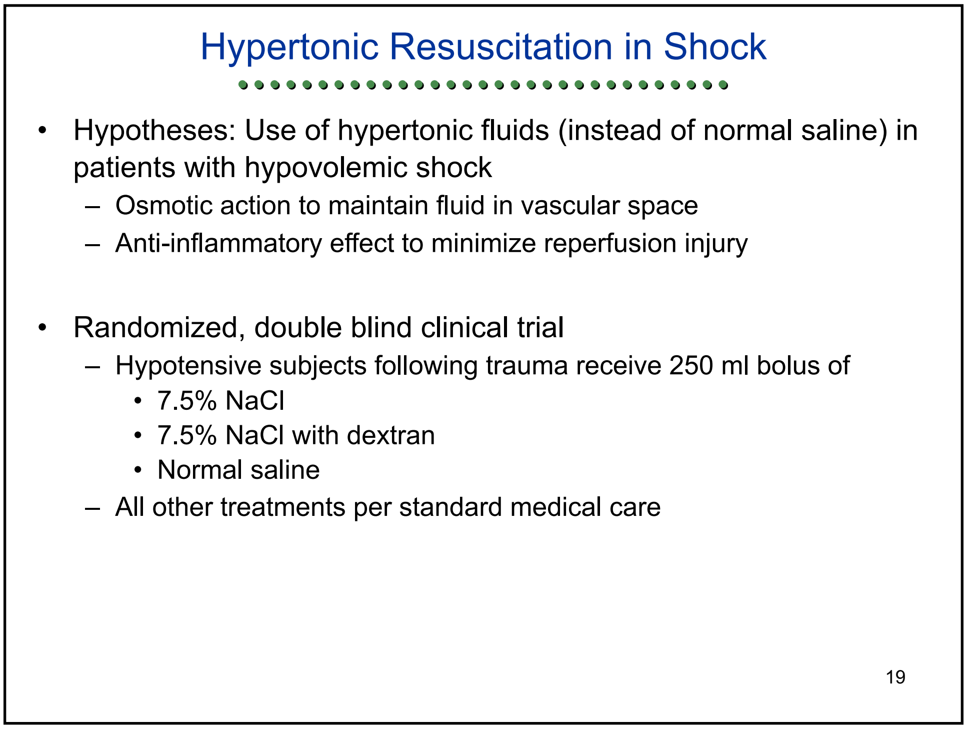
Example: ROC HS/D Shock Trial

- Resuscitation Outcomes Consortium
  - 11 Geographic sites serving ~ 20 million
    - University based investigators
  - More than 250 EMS agencies
    - Over 35,000 EMS providers: EMTs and paramedics
- Conduct definitive clinical trials in the resuscitation of pre-hospital cardiac arrest and severe traumatic injury
  - Treat patients 20-50 minutes on average before delivering them to ED / hospital

ROC
RESUSCITATION OUTCOMES CONSORTIUM

18

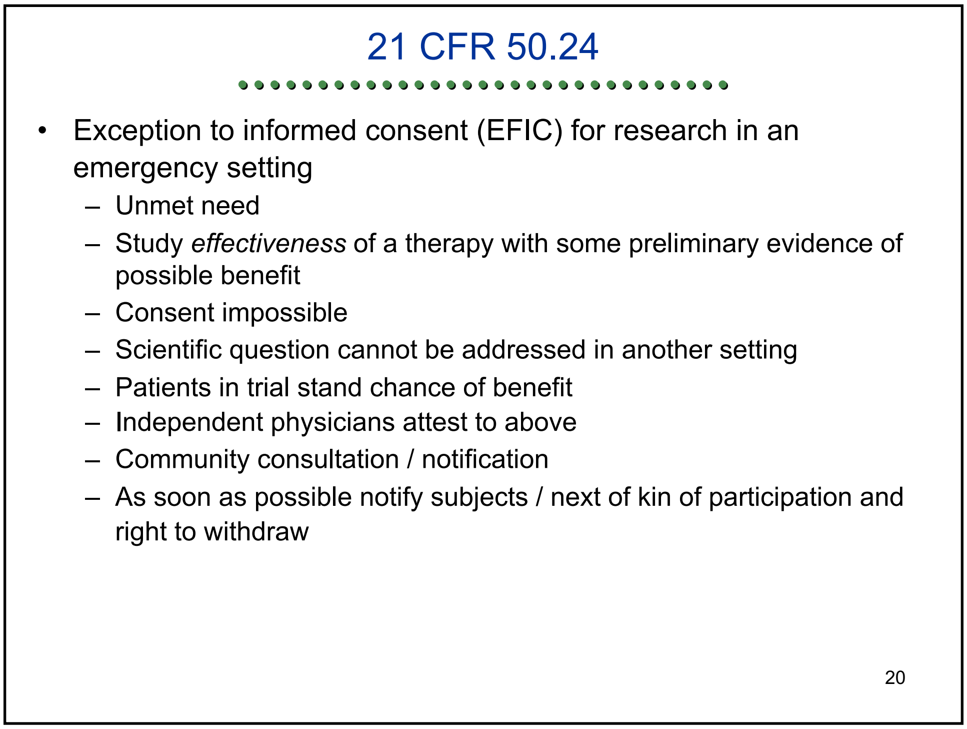
18



Hypertonic Resuscitation in Shock

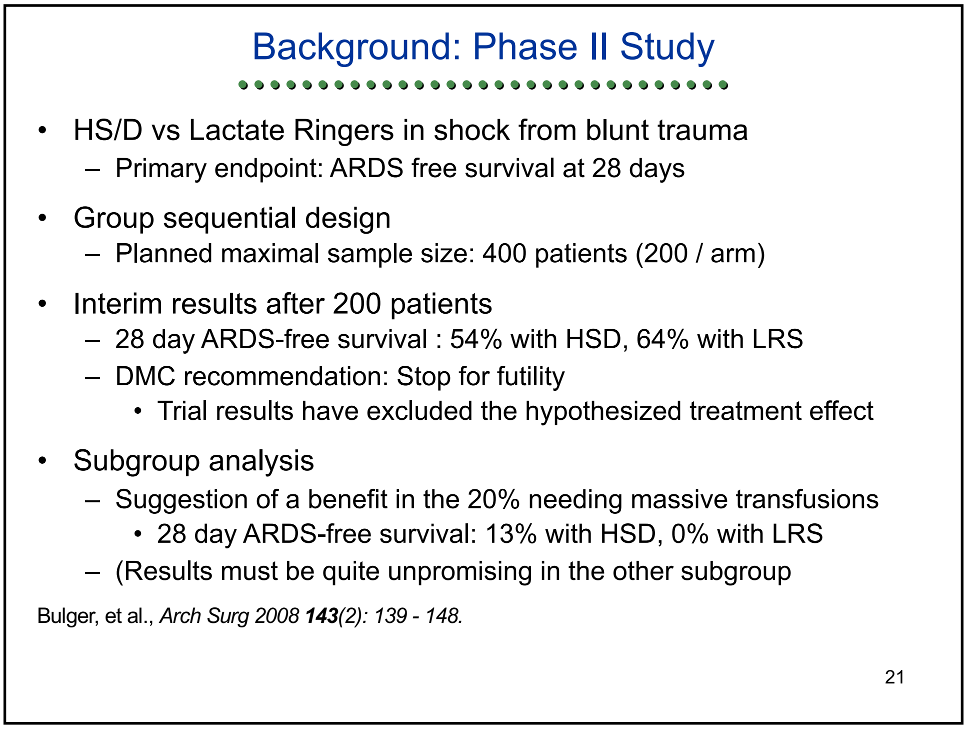
- Hypotheses: Use of hypertonic fluids (instead of normal saline) in patients with hypovolemic shock
  - Osmotic action to maintain fluid in vascular space
  - Anti-inflammatory effect to minimize reperfusion injury

- Randomized, double blind clinical trial
  - Hypotensive subjects following trauma receive 250 ml bolus of
    - 7.5% NaCl
    - 7.5% NaCl with dextran
    - Normal saline
  - All other treatments per standard medical care



21 CFR 50.24

- Exception to informed consent (EFIC) for research in an emergency setting
  - Unmet need
  - Study *effectiveness* of a therapy with some preliminary evidence of possible benefit
  - Consent impossible
  - Scientific question cannot be addressed in another setting
  - Patients in trial stand chance of benefit
  - Independent physicians attest to above
  - Community consultation / notification
  - As soon as possible notify subjects / next of kin of participation and right to withdraw



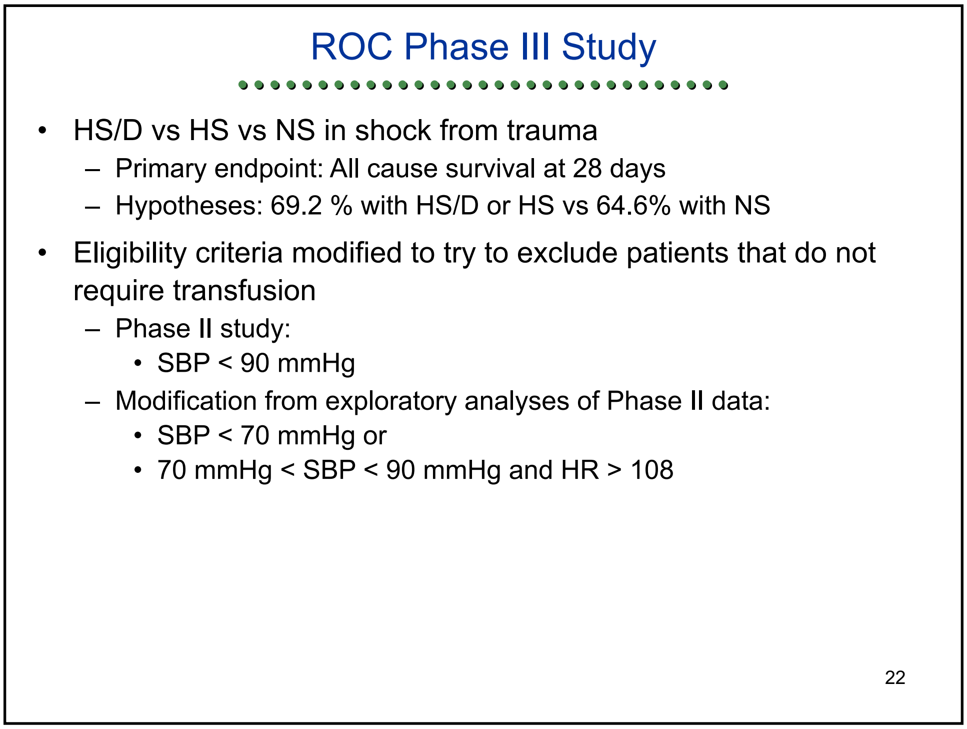
Background: Phase II Study

- HS/D vs Lactate Ringers in shock from blunt trauma
  - Primary endpoint: ARDS free survival at 28 days
- Group sequential design
  - Planned maximal sample size: 400 patients (200 / arm)
- Interim results after 200 patients
  - 28 day ARDS-free survival : 54% with HSD, 64% with LRS
  - DMC recommendation: Stop for futility
    - Trial results have excluded the hypothesized treatment effect
- Subgroup analysis
  - Suggestion of a benefit in the 20% needing massive transfusions
    - 28 day ARDS-free survival: 13% with HSD, 0% with LRS
  - (Results must be quite unpromising in the other subgroup

Bulger, et al., *Arch Surg 2008* ***143****(2): 139 - 148.*

21

21

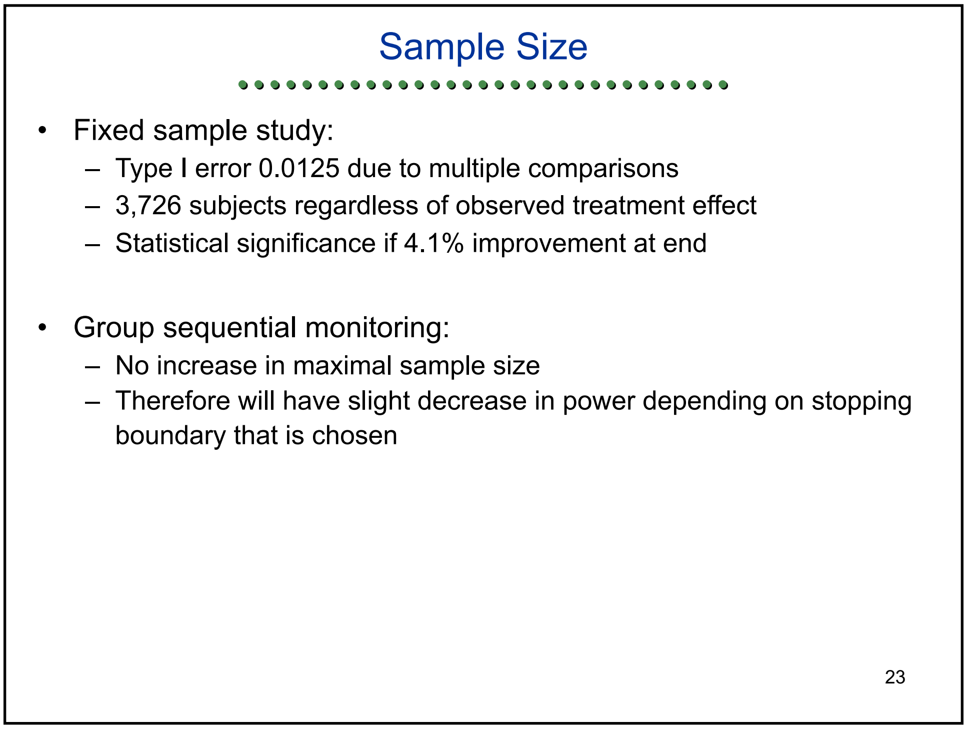


ROC Phase III Study

- HS/D vs HS vs NS in shock from trauma
  - Primary endpoint: All cause survival at 28 days
  - Hypotheses: 69.2 % with HS/D or HS vs 64.6% with NS
- Eligibility criteria modified to try to exclude patients that do not require transfusion
  - Phase II study:
    - SBP < 90 mmHg
  - Modification from exploratory analyses of Phase II data:
    - SBP < 70 mmHg or
    - 70 mmHg < SBP < 90 mmHg and HR > 108

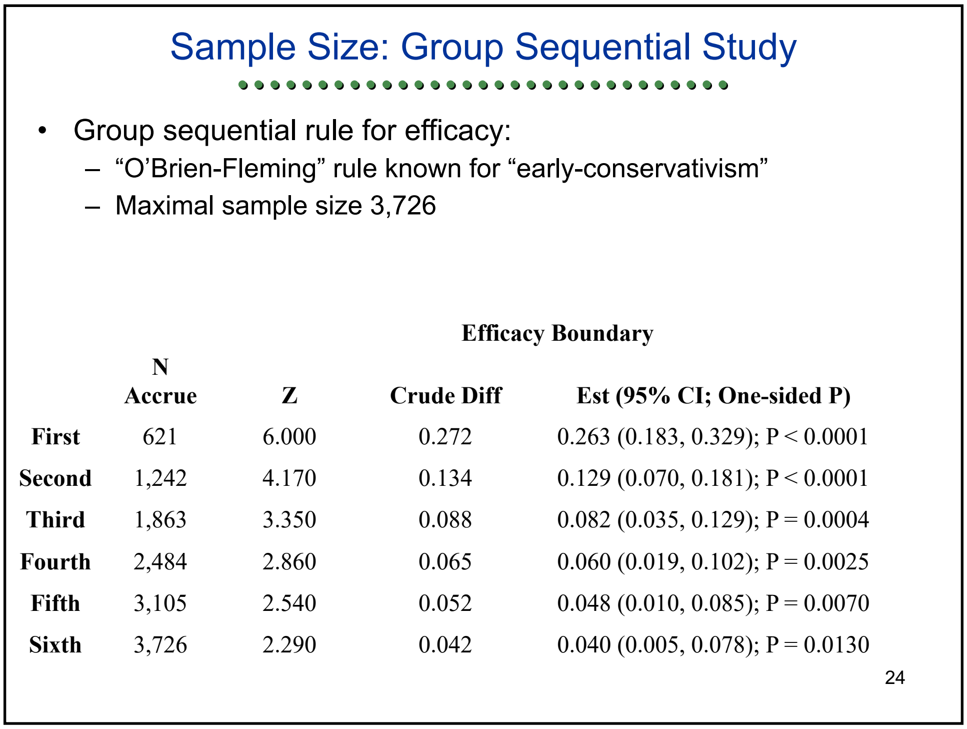
22

22



Sample Size

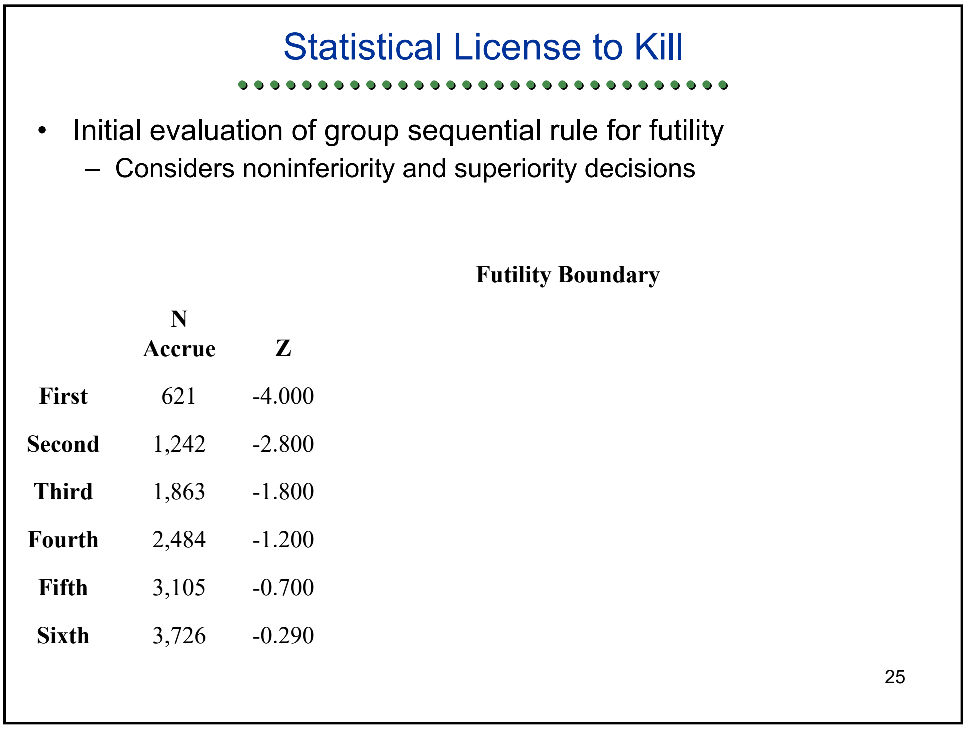
- Fixed sample study:
  - Type I error 0.0125 due to multiple comparisons
  - 3,726 subjects regardless of observed treatment effect
  - Statistical significance if 4.1% improvement at end
- Group sequential monitoring:
  - No increase in maximal sample size
  - Therefore will have slight decrease in power depending on stopping boundary that is chosen



Sample Size: Group Sequential Study

- Group sequential rule for efficacy:
  - “O’Brien-Fleming” rule known for “early-conservativism”
  - Maximal sample size 3,726

		Efficacy Boundary		
	N Accrue	Z	Crude Diff	Est (95% CI; One-sided P)
First	621	6.000	0.272	0.263 (0.183, 0.329); P < 0.0001
Second	1,242	4.170	0.134	0.129 (0.070, 0.181); P < 0.0001
Third	1,863	3.350	0.088	0.082 (0.035, 0.129); P = 0.0004
Fourth	2,484	2.860	0.065	0.060 (0.019, 0.102); P = 0.0025
Fifth	3,105	2.540	0.052	0.048 (0.010, 0.085); P = 0.0070
Sixth	3,726	2.290	0.042	0.040 (0.005, 0.078); P = 0.0130



Statistical License to Kill

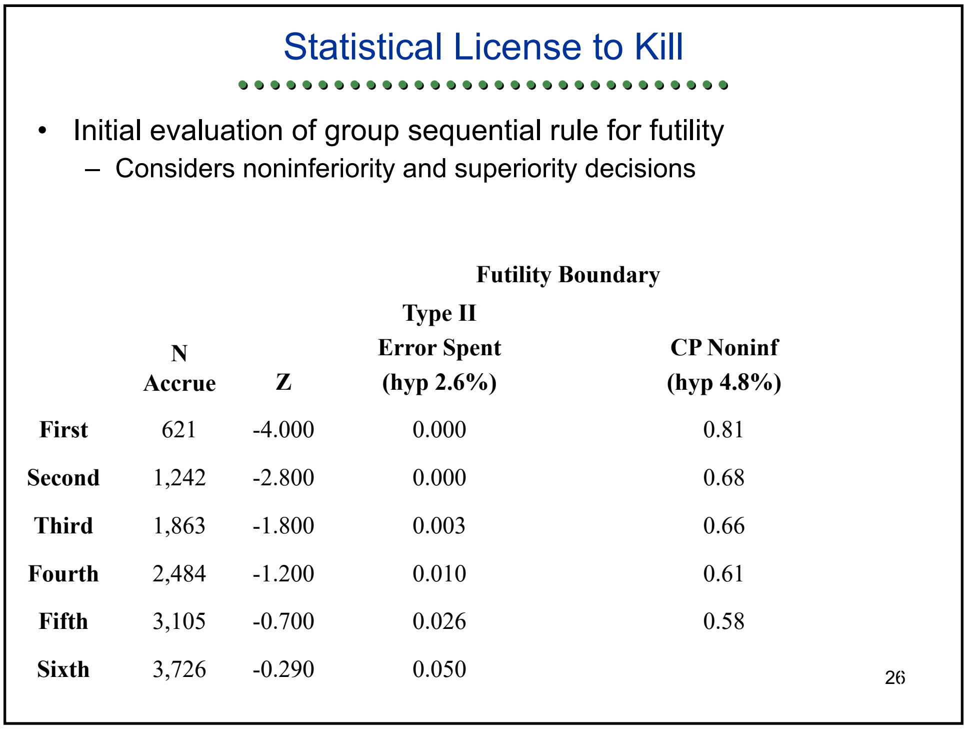
- Initial evaluation of group sequential rule for futility
  - Considers noninferiority and superiority decisions

Futility Boundary

	N Accrue	Z
First	621	-4.000
Second	1,242	-2.800
Third	1,863	-1.800
Fourth	2,484	-1.200
Fifth	3,105	-0.700
Sixth	3,726	-0.290

25

25



Statistical License to Kill

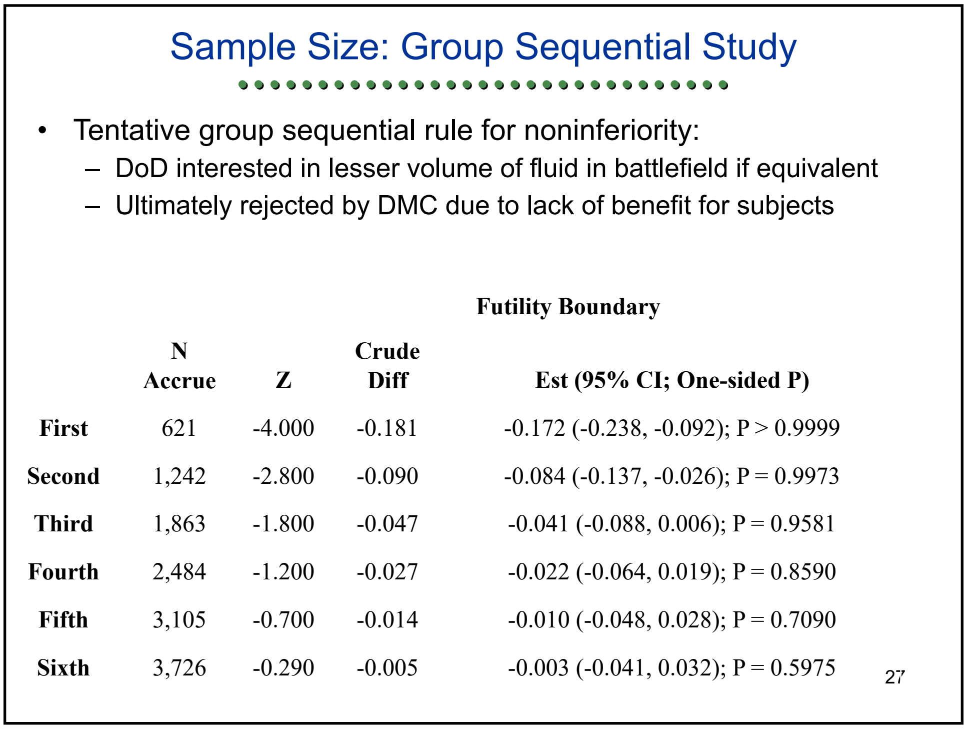
- Initial evaluation of group sequential rule for futility
  - Considers noninferiority and superiority decisions

Futility Boundary

	N Accrue	Z	Type II Error Spent (hyp 2.6%)	CP Noninf (hyp 4.8%)
First	621	-4.000	0.000	0.81
Second	1,242	-2.800	0.000	0.68
Third	1,863	-1.800	0.003	0.66
Fourth	2,484	-1.200	0.010	0.61
Fifth	3,105	-0.700	0.026	0.58
Sixth	3,726	-0.290	0.050	

26

26



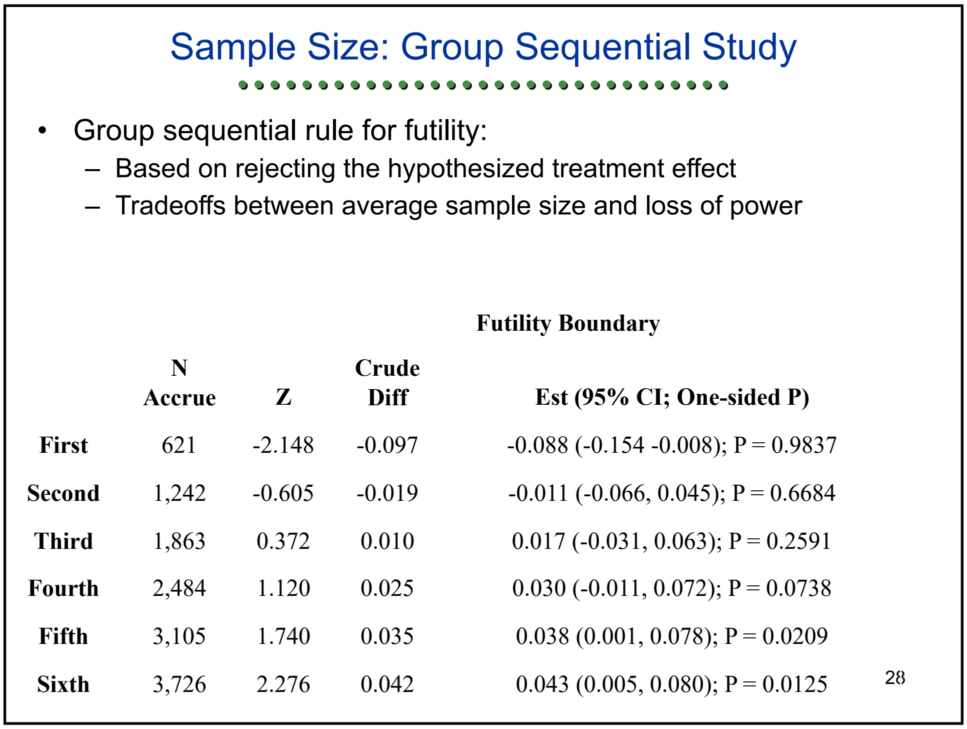
Sample Size: Group Sequential Study

- Tentative group sequential rule for noninferiority:
  - DoD interested in lesser volume of fluid in battlefield if equivalent
  - Ultimately rejected by DMC due to lack of benefit for subjects

				Futility Boundary
	N Accrue	Z	Crude Diff	Est (95% CI; One-sided P)
First	621	-4.000	-0.181	-0.172 (-0.238, -0.092); P > 0.9999
Second	1,242	-2.800	-0.090	-0.084 (-0.137, -0.026); P = 0.9973
Third	1,863	-1.800	-0.047	-0.041 (-0.088, 0.006); P = 0.9581
Fourth	2,484	-1.200	-0.027	-0.022 (-0.064, 0.019); P = 0.8590
Fifth	3,105	-0.700	-0.014	-0.010 (-0.048, 0.028); P = 0.7090
Sixth	3,726	-0.290	-0.005	-0.003 (-0.041, 0.032); P = 0.5975

27

27



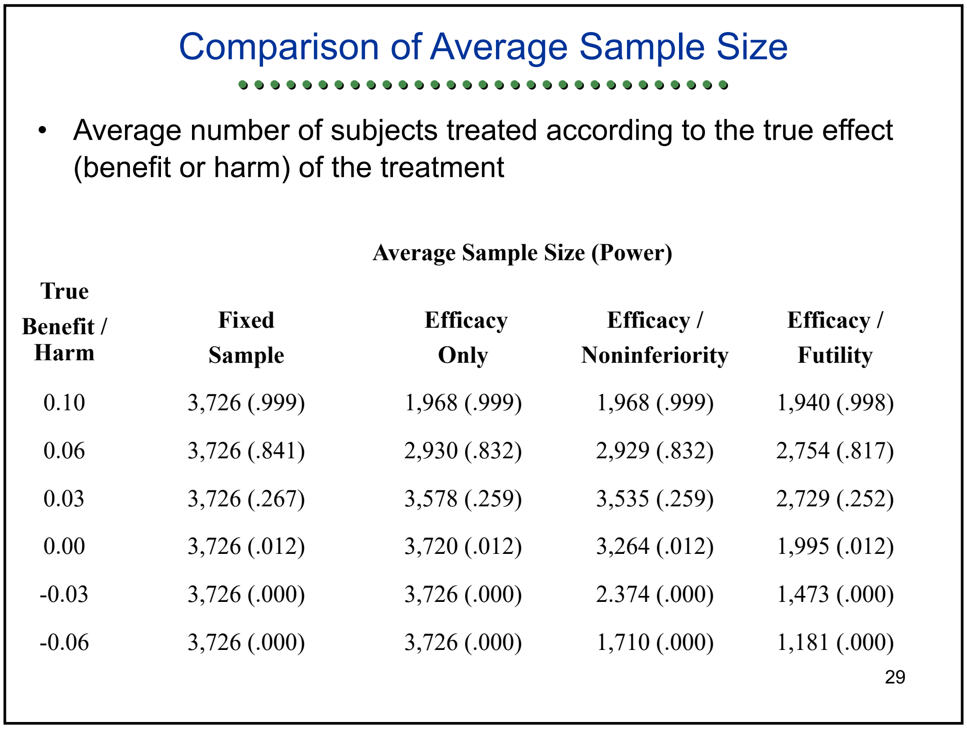
Sample Size: Group Sequential Study

- Group sequential rule for futility:
  - Based on rejecting the hypothesized treatment effect
  - Tradeoffs between average sample size and loss of power

				Futility Boundary
	N Accrue	Z	Crude Diff	Est (95% CI; One-sided P)
First	621	-2.148	-0.097	-0.088 (-0.154 -0.008); P = 0.9837
Second	1,242	-0.605	-0.019	-0.011 (-0.066, 0.045); P = 0.6684
Third	1,863	0.372	0.010	0.017 (-0.031, 0.063); P = 0.2591
Fourth	2,484	1.120	0.025	0.030 (-0.011, 0.072); P = 0.0738
Fifth	3,105	1.740	0.035	0.038 (0.001, 0.078); P = 0.0209
Sixth	3,726	2.276	0.042	0.043 (0.005, 0.080); P = 0.0125

28

28



Comparison of Average Sample Size

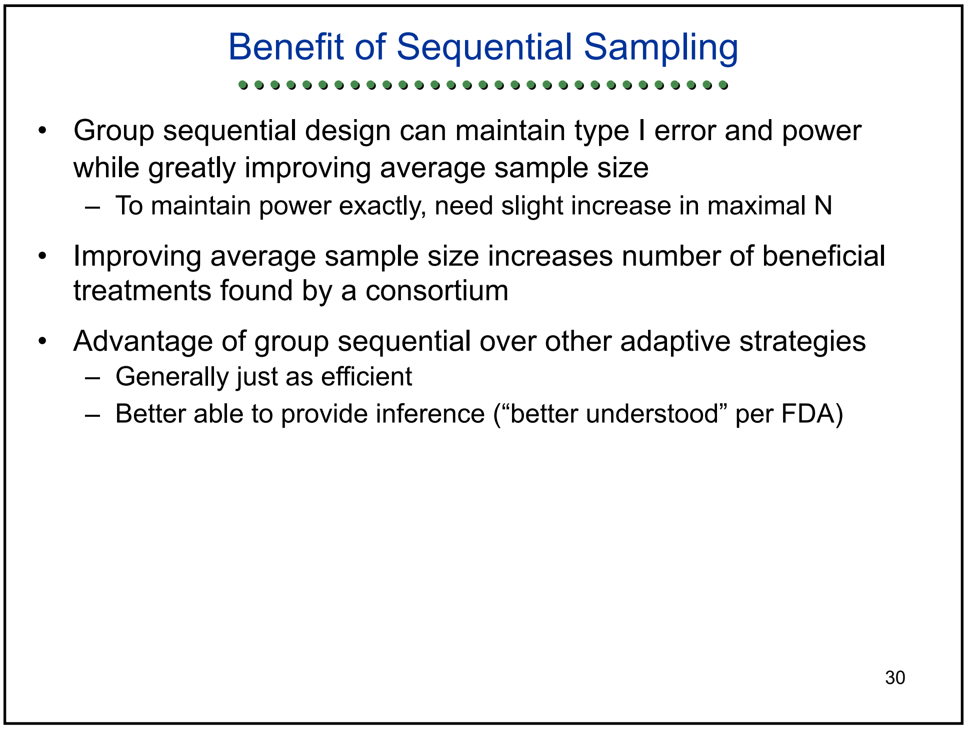
- Average number of subjects treated according to the true effect (benefit or harm) of the treatment

Average Sample Size (Power)

True Benefit / Harm	Fixed Sample	Efficacy Only	Efficacy / Noninferiority	Efficacy / Futility
0.10	3,726 (.999)	1,968 (.999)	1,968 (.999)	1,940 (.998)
0.06	3,726 (.841)	2,930 (.832)	2,929 (.832)	2,754 (.817)
0.03	3,726 (.267)	3,578 (.259)	3,535 (.259)	2,729 (.252)
0.00	3,726 (.012)	3,720 (.012)	3,264 (.012)	1,995 (.012)
-0.03	3,726 (.000)	3,726 (.000)	2.374 (.000)	1,473 (.000)
-0.06	3,726 (.000)	3,726 (.000)	1,710 (.000)	1,181 (.000)

29

29

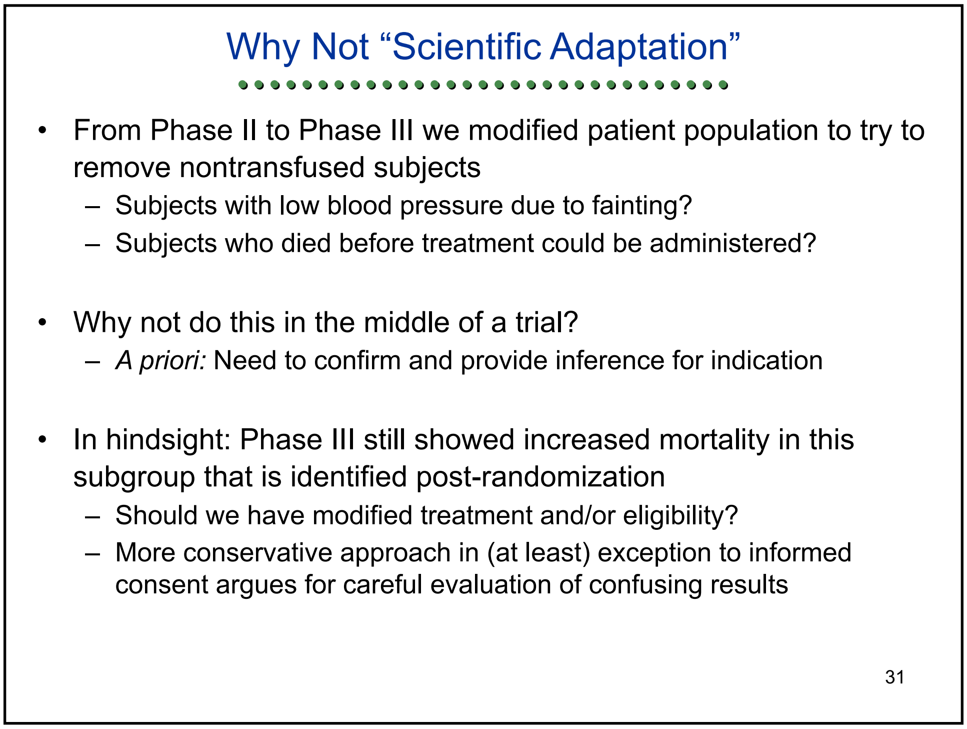


Benefit of Sequential Sampling

- Group sequential design can maintain type I error and power while greatly improving average sample size
  - To maintain power exactly, need slight increase in maximal N
- Improving average sample size increases number of beneficial treatments found by a consortium
- Advantage of group sequential over other adaptive strategies
  - Generally just as efficient
  - Better able to provide inference (“better understood” per FDA)

30

30

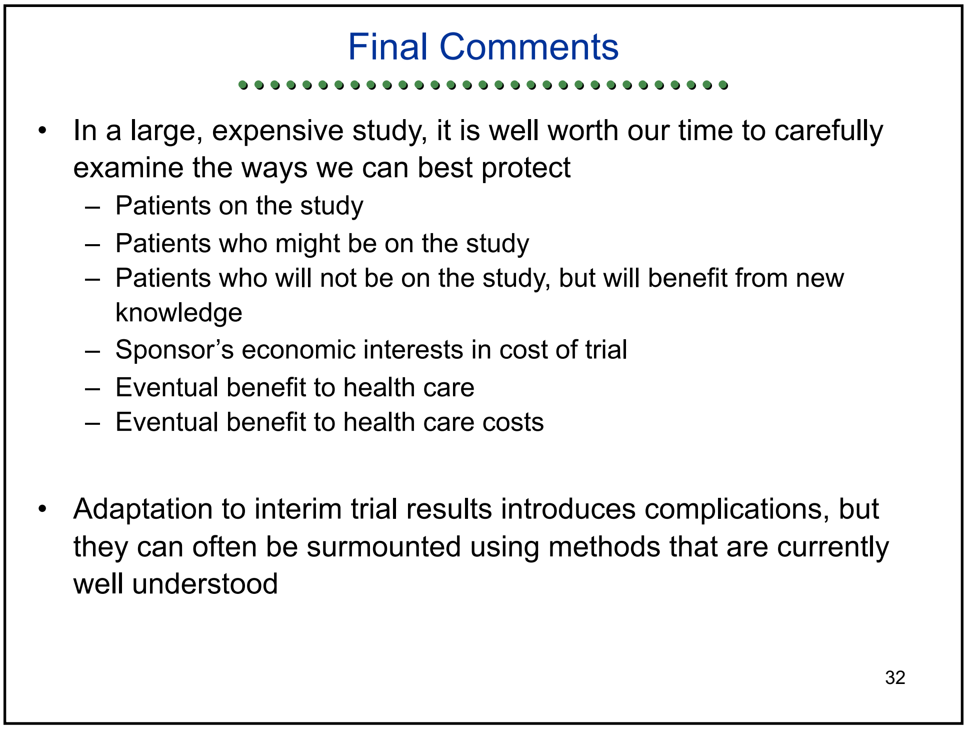


Why Not "Scientific Adaptation"

- From Phase II to Phase III we modified patient population to try to remove nontransfused subjects
  - Subjects with low blood pressure due to fainting?
  - Subjects who died before treatment could be administered?
- Why not do this in the middle of a trial?
  - *A priori:* Need to confirm and provide inference for indication
- In hindsight: Phase III still showed increased mortality in this subgroup that is identified post-randomization
  - Should we have modified treatment and/or eligibility?
  - More conservative approach in (at least) exception to informed consent argues for careful evaluation of confusing results

31

31



Final Comments

- In a large, expensive study, it is well worth our time to carefully examine the ways we can best protect
  - Patients on the study
  - Patients who might be on the study
  - Patients who will not be on the study, but will benefit from new knowledge
  - Sponsor's economic interests in cost of trial
  - Eventual benefit to health care
  - Eventual benefit to health care costs
- Adaptation to interim trial results introduces complications, but they can often be surmounted using methods that are currently well understood

32

32