Module 5:

Statistical Design, Conduct, and Analysis of Clinical Trials

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

Summer Institute in Statistics for
Clinical & Epidemiological Research
July 23, 2019

Abstract

This module focuses on the statistical issues that must be addressed in the design, implementation, and reporting of randomized clinical trials. This module presumes the student has familiarity with the topics covered in Module 1. The premise of the module is that each clinical trial poses unique problems, and thus the best design for a particular clinical trial may not be appropriate for another. Hence, emphasis is placed on the iterative approach of considering alternative design structures, selecting candidate designs, evaluating the operating characteristics of those designs with respect to a variety of criteria, and comparing a number of designs to find an acceptable design for the scientific problem at hand. It will include discussion of

- alternative randomization strategies: randomization ratios, stratification, covariate adaptive randomization, response adaptive randomization
- choice of statistical analysis models: choice of statistic, covariate adjustment
- sequential adaptive designs: fixed sample, group sequential, blinded sample size re-estimation, unblinded sample size re-estimation, adaptive enrichment
- overview of approaches to missing data: sensitivity to missing not at random
- analysis,including covariate adaptive and response adaptive randomization

The module concludes with a discussion of the ways that the complete design and its operating characteristics might be summarized in the study protocol and fully documented in a statistical analysis plan (SAP) in order to satisfy study investigators and regulatory agencies.

2

Course Objectives

- Introduction to statistical design, monitoring, and analysis of randomized clinical trials (RCT)
  - Presumes clinical/scientific issues have been decided in collaboration with other researchers
- Topics generally relate to achieving statistical precision through appropriate choices of sample size as dependent on
  - Controls (parallel group vs matched, attenuated effects)
  - Randomization scheme (complete, blocked, stratified)
  - Analysis model (summary measure, covariate adjustment)
  - Sequential testing (group sequential, other adaptive)
  - Missing data analyses (overview)
- Discussion of plans for statistical analysis and content of a Statistical Analysis Plan (SAP)

3

The Problem of Clinical Trial Design

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy, *Anna Karenina, 1873-77*

4

The Problem of Clinical Trial Design

Unbiased clinical trials are all alike; every biased clinical trial is biased in its own way.

5

Take Home Message 0

- **Clinical / scientific vs statistical issues in RCT design**
  - Clinical and scientific issues are most important in RCT design
    - Define appropriate subjects, comparators, procedures, outcomes
    - Maintain scientific rigor: randomization and blinding when possible
  - Statistical expertise is crucial when addressing clinical and scientific issues
    - Interpreting results of prior studies
    - Understanding of conditional probabilities
    - Anticipating confounding and other biases
    - Looking ahead for statistical limitations that might preclude certain designs
  - Once clinical and scientific issues mostly resolved, statistician's role is to maximize and quantify RCT precision

6

Take Home Message 1

- **Statistical criteria for evidence (frequentist vs Bayesian)**
  - Frequentist inference is best starting place for RCT design
    - In designed experiment, frequentist inference has solid foundations
    - Regulatory and journal standards more often on frequentist basis
      - 95% confidence intervals, level 0.05 / .025 tests
  - Approximate Bayesian inference can be easily derived from frequentist inference (easier than vice versa)
    - Sensitivity of conclusions to a spectrum of priors can be effected using frequentist sampling distribution and conjugate normal priors
  - Emphasize interval and point estimates rather than p values
    - Facilitates clinical and scientific interpretation of results

7

Take Home Message 2

- **Probability models for RCT outcomes**
  - Avoid making strong (semi)parametric assumptions across treatment groups
    - (Semi)parametric assumptions across arms are not consistent with possible heterogeneity of effect across individuals
    - Fortunately many distribution-free approaches are very similar to some standard (semi)parametric model
  - Summarize outcome distribution according to clinical relevance
    - Plausibility of effect and statistical precision are secondary issues
  - Use distribution free estimates of summaries and their SE whenever possible
    - Such estimates most often have an approximately normal sampling distribution

8

Take Home Message 3

- **Precision of inference**
  - Precision of RCT is measured in width of interval estimates and statistical power to discriminate between important hypotheses
  - Statistical formulas to quantify precision are similar across a wide variety of commonly used statistical analysis models
  - Determinants of precision modifiable in RCT design include
    - Completeness of outcome collection (noninformative censoring)
    - Distributional summary measure used to assess treatment effect
    - Homogeneity of outcome within groups compared across arms
    - Ratio of sample sizes across groups compared across arms
    - Total sample size accrued to study
    - Avoiding potentially informative missing data

9

Take Home Message 4

- **Summarizing distribution of outcomes**
  - Choose summary measure according to (in order of importance):
    - Estimable within logistical constraints of RCT
    - Clinically / scientifically most important
    - Plausibly affected by treatment
    - Estimated with greatest statistical precision
  - Issues not related to precision
    - Univariate summaries allow better comparisons across studies
    - Proportion / odds above threshold often clinically most important
    - Means / geometric means may better quantify effects in population
    - With heavily censored data, some summaries may not be estimable
  - Means / geometric means most precise for many distributions
    - Means: normal, Poisson, binomial, exponential
    - Geometric means: lognormal

10

Take Home Message 5

- **Contrasting outcome distributions across treatment arms**
  - Differences or ratios of univariate summaries most interpretable
    - Allow estimation and testing on same measures across studies
    - Less desirable by this criterion: Wilcoxon rank sum test
  - With binary outcomes
    - Difference in proportions (attributable risk) gives population impact
    - Ratio of proportions (relative risk) is invariant to "contamination" with subjects insusceptible to treatment (either always 0 OR 1, not both)
    - Odds ratio may be better in adjusted analyses (less effect modification) and mimics relative risk in rare outcomes
  - Differences are expressed in units of original measurements
  - Ratios are dimensionless, hence need anchoring with baseline
    - Odds and geometric means tend to use ratios as contrasts

11

Take Home Message 6

- **Why control for baseline variables in RCT analyses?**
  - Precision of inference is the primary issue
    - Confounding is in some sense avoided by randomization
      - However, some critics will second guess results *post hoc*
    - Estimation of "within stratum" effect an issue with odds, hazards
- **How to control for baseline variables in RCT design?**
  - Restrict eligibility to subset of eventual target population
    - Only if certain no effect modification on efficacy or safety
  - Stratified randomization based on prognostic factors especially if
    - Want face validity of key prognostic factors in Table 1
    - Interested in prespecified analyses within subgroups
    - Covariate adaptive minimization can be usually be avoided
  - Obtain precision by prespecified adjustment for the most important prognostic factors known *a priori*
    - Do not allow data driven selection of adjusted models

12

Take Home Message 7

- **Randomization strategies**
  - 1:1 randomization unless really good reason
    - Most efficient randomization under strong null and homoscedasticity
  - Stratification
    - In large studies, no stratification necessary
    - In small studies:
      - Stratify on variables for which subgroup analyses are important
      - Stratify on clinical center if possible
      - No more than 4 important variables (including center)
      - No real advantage to dynamic minimization
    - Prespecify analysis including all fixed effects used in stratification
      - Less of an issue in logistic regression, but might as well
  - Use hidden blocks of varying sizes
    - Block sizes 4, 6, or 8 in 1:1 randomization

13

Take Home Message 8

- **Group sequential, adaptive studies**
  - RCT have always been sequential and adaptive across phases
    - Focus over past 50-70 years has been within individual studies
  - Sequential designs can help address efficiency / ethics of RCT
    - Greatest advantage is terminating “futile” studies earlier
    - For effective therapies, larger sample sizes may be needed
      - Safety and secondary endpoints
  - Group sequential approaches are generally sufficient
    - Fixed sample design as starting point
    - Explore alternative designs varying in schedule/timing of analyses and conservatism of stopping boundaries
    - Fully evaluate inference corresponding to early termination, statistical power, sample size requirements
      - Most boundary scales are nonlinear and difficult to predict
      - Conditional and predictive power have difficult interpretations

14

Take Home Message 9

- **Adaptive RCT designs**
  - Group sequential designs a special, well understood case
  - Methods based on unblinded adaptation
    - Statistical inference not based on minimal sufficient statistics
    - Unblinded adaptive modifications of statistical aspects of design generally not recommended by me due to inefficiency
      - Randomization ratio
      - Maximal statistical information (SSRE)
        - » Possible exception: altering accrual in time to event
  - Adaptive modifications of indication (phase 2 more than phase 3)
    - Adaptive enrichment of patient population to find greatest efficacy
    - Adaptive selection of doses or treatments
    - Adaptive selection of outcomes
      - Less useful because outcome should be based on clinical importance

15

Take Home Message 10

- **Statistical Analysis Plan**
  - Table 1: Compare distributions of baseline variables across arms
    - Materials and methods: range, quartiles, means, SD
    - Randomization imbalance: means (geometric means, proportions)
  - Formal sequential stopping rules
  - Complete pre-specification of analysis models for all primary and secondary endpoints
    - Adjustment for covariates and exact form used in modeling
      - Model misspecification not much of issue
    - Handling of missing data and sensitivity analyses
    - Handling of multiple comparisons: Hierarchy of testing
  - Confirmatory and supportive analyses re mechanism of action
    - Alternative summary measures and alternative clinical endpoints
  - Exploratory analyses of important subgroups

16

7

Scientific Setting: Clinical Research

Experimentation in Human Volunteers

"Treatment Discovery"

Where am I going?

Improvements in population health are the result of a long history of scientific research into diagnosis, treatment and (ideally) prevention of disease

A succession of scientific studies and experiments

- "The result of a scientific study is another scientific study"

Knowledge about treatment indications solidified through RCT

- Clinical goals
- Rigorous scientific standard
- Ethical treatment of individuals and populations

Medical 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)
- Medical science is about improving the health of people, but

"The physician must ... have two special objects in view with regard to disease, namely, to do good or to do no harm"

- Epidemics I, Hippocratic Corpus, c. 410 BC

"…it is only the second law of therapeutics to do good, its first law being this – not to do harm"

- Elisha Bartlett, 1844

18

Goals of Medical Research

- Identify methods to diagnose disease
- Identify risk factors for disease
- Identify treatments for disease
- Identify methods for disease prognosis
- Identify strategies for prevention of disease
- Basic science

19

Scientific Method

- Observation
- Form hypotheses
- Designed study
  - Overall goal
  - Specific aims
  - Materials and Methods
  - Data collection
  - Results (data analysis)
  - Interpretation and conclusions
- Repeat (new, refined hypotheses)

20

Bioethics in Clinical Research

- Classic principles
  - Autonomy
  - Beneficence
  - Non-maleficence
  - Justice
- Selected additional principles
  - Authority
  - Principle of double effect
  - Confidentiality
  - Equality
  - Equity
  - Mercy
  - Ownership
  - Privacy

21

Implementation in Clinical Research

- Governmental regulatory authorities
  - FDA (US), EMA (Europe), PMDA (Japan), MHRA (UK), etc.
  - International Council on Harmonisation (ICH)
- Institutional Review Boards (IRBs)
  - Institutional Ethics Commiteess, Research Ethics Boards, etc.
- Data Monitoring Committees
- Informed Consent Forms

22

U.S. Regulation of Drugs / Biologics

- 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)

23

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

24

Clinical Trials

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

25

The Enemy

"Let's start at the very beginning, a very good place to start…"

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

26

Overall Goal

- "Drug discovery"
- More generally
  - a therapy or preventive strategy
    - drug, biologic, device, behavior
  - for some disease
  - in some population of patients
- Medical diagnosis / prognosis
  - Evaluation of methods

27

Typical Chronology

- Observational epidemiology of disease, risk
- Preclinical experiments
  - Laboratory, animal studies of mechanisms, toxicology
- Clinical trials
  - Safety for further investigations / dose
  - Safety of therapy
  - Measures of efficacy
  - Confirmation of efficacy / effectivenesss
- Synthesis and quantification of evidence
- Adoption of new treatment indication

28

First: Where Do We Want To Be?

- Perform some innovative experiment?
- Find a use for some proprietary drug / biologic / device?
  - "Obtain a significant p value"
- Find an efficacious treatment?
  - "Efficacy" seems to benefit some people somehow
- Find a new treatment that improves health of individuals
  - "Personalized medicine" (fixed effect or random effect?)
- Find effective treatment that improves health of the population
  - Treatments administered to a community
  - Treatments tested on a population then administered correctly

29

Treatment "Indication"

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

30

Primary Endpoint: Clinical

- Consider (in order)
  - The most relevant clinical endpoint
    - Survival, quality of life ("feels, functions, or survives")
  - The endpoint the treatment is most likely to affect
  - The endpoint that can be assessed most accurately and precisely

31

Additional Endpoints

- Other outcomes are then relegated to a "secondary" status
  - Supportive and confirmatory
  - Safety
- Some outcomes are considered "exploratory"
  - Subgroup effects
  - Effect modification

32

Typical Regulatory Chronology

- Phase 1:
  - PK/PD; dose finding; dose limiting toxicities
- Phase 2:
  - Initial estimates of efficacy
    - Perhaps surrogate in enriched population over short timeframe
  - Initial assessment of safety
    - Adverse events (AEs), serious adverse events (SAEs)
    - Perhaps some protocolized safety measurements
- Phase 3: ("adequate and well-controlled investigations")
  - Confirmatory efficacy (closer to effectiveness)
  - Detailed safety: protocolized as indicated and AEs, SAEs
- Phase 4: (post marketing)
  - Additional safety and special populations as indicated

33

Some Regulatory Jargon

- Pivotal studies
  - Results so conclusive (extremely small p value) that a single study will suffice
- Noninferiority studies
  - Proof of efficacy by showing not unacceptably inferior to a previously proven treatment
- Bioequivalence studies
  - Neither markedly better than nor markedly worse than another (formulation of the same) treatment
- Bridging studies
  - Small studies to handle changes in formulation, environment, etc.

34

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

35

Statistical Planning

- Satisfy collaborators as much as possible
- Discriminate between relevant scientific hypotheses
  - Scientific and statistical credibility
- Protect economic interests of sponsor
  - Efficient designs
  - Economically important estimates
- Protect interests of patients on trial
  - Stop if unsafe or unethical
  - Stop when credible decision can be made
- Promote rapid discovery of new beneficial treatments

36

Take Home Message 0

- **Clinical / scientific vs statistical issues in RCT design**
  - Clinical and scientific issues are most important in RCT design
    - Define appropriate subjects, comparators, procedures, outcomes
    - Maintain scientific rigor: randomization and blinding when possible
  - Statistical expertise is crucial when addressing clinical and scientific issues
    - Interpreting results of prior studies
    - Understanding of conditional probabilities
    - Anticipating confounding and other biases
    - Looking ahead for statistical limitations that might preclude certain designs
  - Once clinical and scientific issues mostly resolved, statistician's role is to maximize and quantify RCT precision

37

8

Scientific Setting: Clinical Research

Statistical Criteria for Evidence

Where am I going?

Frequentist criteria are most often used in medical research, though Bayesian inference is also important

Quantifying precision by interval estimates is most important

Ideal Results

- Goals of "drug discovery" are similar to those of diagnostic testing in clinical medicine
- We want a "drug discovery" process in which there is
  - A low probability of adopting ineffective drugs
    - High specificity (low type I error)
  - A high probability of adopting truly effective drugs
    - High sensitivity (low type II error; high power)
  - A high probability that adopted drugs are truly effective
    - High positive predictive value
    - Will depend on prevalence of "good ideas" among our ideas

39

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)

40

Diagnostic Medicine: Evaluating a Test

- **We condition on diagnoses** (from gold standard)
  - Frequentist criteria: We condition on what is unknown in practice
- **Sensitivity: Do diseased people have positive test?**
  - Denominator: Diseased individuals
  - Numerator: Individuals with a positive test among denominator
- **Specificity: Do healthy people have negative test?**
  - Denominator: Healthy individuals
  - Numerator: Individuals with a negative test among denominator

41

Diagnostic Medicine: Using a Test

- **We condition on test results**
  - Bayesian criteria: We condition on what is known in practice
- **Pred Val Pos: Are positive people diseased?**
  - Denominator: Individuals with positive test result
  - Numerator: Individuals with disease among denominator
- **Pred Val Neg: Are negative people healthy?**
  - Denominator: Individuals with negative test result
  - Numerator: Individuals who are healthy among denominator

42

Points Meriting Special Emphasis

- Discover / evaluate tests using frequentist methods
  - Sensitivity, specificity
- Consider Bayesian methods when interpreting results for a given patient
  - Predictive value of positive, predictive value of negative
- Possible rationale for our practices
  - Ease of study: Efficiency of case-control sampling
  - Generalizability across patient populations
    - Belief that sensitivity and specificity might be
    - Knowledge that PPV and NPV are not
  - Ability to use sensitivity and specificity to get PPV and NPV
    - But not necessarily vice versa

43

Bayes' Rule

- Allows computation of "reversed" conditional probability
- Can compute PPV and NPV from sensitivity, specificity
  - **BUT: <u>Must</u> know prevalence of disease**

$$PPV = \frac{sensitivity \times prevalence}{sens \times prevalence \quad + \quad (1-spec)\times(1-prevalence)}$$

$$NPV = \frac{specificity \times (1-prevalence)}{spec \times (1-prevalence) \quad + \quad (1-sens)\times prevalence}$$

44

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

45

Slightly Different Setting

- Usually we are interested in some continuous parameter
  - E.g., proportion of infections cured is $0 < p < 1$
- "Prevalence" is replaced by a probability distribution
  - Prior (subjective) probability of selecting a drug to test that cures proportion p of the population
- Sum over two hypotheses replaced by weighted average (by some subjective prior) over all possibilities

$$\Pr(p \mid \hat{p}) = \frac{\Pr(\hat{p} \mid p) \times \Pr(p)}{\int \Pr(\hat{p} \mid p) \times \Pr(p)\, dp}$$

$$= \frac{freq\ samp\ distn\ \times\ prior\ prob}{weighted\ average\ freq\ samp\ distn}$$

46

Frequentist Inference

- Control type 1 error: False positive rate
  - Based on specificity of our methods
- Maximize statistical power: True positive rate
  - Sensitivity to detect specified effect
- Provide unbiased (or consistent) estimates of effect
- Standard errors: Estimate reproducibility of experiments
- Confidence intervals
- Criticism: Compute probability of data already observed
  - "A precise answer to the wrong question"

47

Bayesian Inference

- Hypothesize prior prevalence of "good" ideas
  - Subjective probability
- Using prior prevalence and frequentist sampling distribution
  - Condition on observed data
  - Compute probability that some hypothesis is true
    - "Posterior probability"
  - Estimates based on summaries of posterior distribution
- Criticism: Which presumed prior distribution is relevant?
  - "A vague answer to the right question"

48

Frequentist vs Bayesian

- 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:

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

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

$$posterior\ \ odds = \textbf{Bayes Factor} \times prior\ \ odds$$

49

Distribution-free Bayesian Models

- Regard estimate of summary measure as the data
  - Use asymptotic distributions under population model
  - Use conjugate normal priors to mimic other Bayesian analyses

Some joint distribution for $(\theta, \hat{\theta})$

Frequentist usually considers $\hat{\theta} \mid \theta \sim N(\theta, V(\theta)/n)$

Bayesian can consider $p\left(\vec{\theta} \mid \hat{\vec{\theta}}\right) = \dfrac{p\left(\hat{\vec{\theta}} \mid \vec{\theta}\right) \lambda(\vec{\theta})}{\int p\left(\hat{\vec{\theta}} \mid \vec{\theta}\right) \lambda(\vec{\theta}) d\vec{\theta}}$

where $\lambda(\vec{\theta})$ is a prior distribution for $\vec{\theta}$

50

Goals of Statistical Inference

- Traditional approach
  - Sample size to provide high power to "detect" a particular alternative
- Decision theoretic approach
  - Sample size to discriminate between hypotheses
    - "Discriminate" based on interval estimate
    - Standard for interval estimate: 95%
      - Equivalent to traditional approach with 97.5% power

51

Reporting Inference

- At the end of the study analyze the data
- Report three measures (four numbers)
  - Point estimate
  - Interval estimate
  - Quantification of confidence / belief in hypotheses

52

Reporting Frequentist Inference

- Three measures (four numbers)
- Consider whether the observed data might reasonably be expected to be obtained under particular hypotheses
  - Point estimate: minimal bias? MSE?
  - Confidence interval: all hypotheses for which the data might reasonably be observed
  - P value: probability such extreme data would have been obtained under the null hypothesis
    - Binary decision: Reject or do not reject the null according to whether the P value is low

53

Reporting Bayesian Inference

- Three measures (four numbers)
- Consider the probability distribution of the parameter conditional on the observed data
  - Point estimate: Posterior mean, median, mode
  - Credible interval: The “central” 95% of the posterior distribution
  - Posterior probability: probability of a particular hypothesis conditional on the data
    - Binary decision: Reject or do not reject the null according to whether the posterior probability is low

54

Parallels Between Tests, CIs

- If the null hypothesis not in CI, reject null
  - (Using same level of confidence)
- Relative advantages
  - Test only requires sampling distn under null
  - CI requires sampling distn under alternatives
  - CI provides interpretation when null is not rejected

55

Scientific Information

- "Rejection" uses a single level of significance
  - Different settings might demand different criteria
- P value communicates statistical evidence, not scientific importance
- Only confidence interval allows you to interpret failure to reject the null:
  - Distinguish between
    - Inadequate precision (sample size)
    - Strong evidence for null

56

Hypothetical Example

- Clinical trials of treatments for hypertension
  - Screening trials for four candidate drugs
    - Measure of treatment effect is the difference in average SBP at the end of six months treatment
    - Drugs may differ in
      - Treatment effect (goal is to find best)
      - Variability of blood pressure
    - Clinical trials may differ in conditions
      - Sample size, etc.

57

Reporting P values

Study	P value
A	0.1974
B	0.1974
C	0.0099
D	0.0099

58

Point Estimates

Study	SBP Diff
A	27.16
B	0.27
C	27.16
D	0.27

59

Point Estimates

Study	SBP Diff	P value
A	27.16	0.1974
B	0.27	0.1974
C	27.16	0.0099
D	0.27	0.0099

60

Confidence Intervals

Study	SBP Diff	95% CI	P value
A	27.16	-14.14, 68.46	0.1974
B	0.27	-0.14, 0.68	0.1974
C	27.16	6.51, 47.81	0.0099
D	0.27	0.06, 0.47	0.0099

61

Interpreting Nonsignificance

- Studies A and B are both “nonsignificant”
  - Only study B ruled out clinically important differences
  - The results of study A might reasonably have been obtained if the treatment truly lowered SBP by as much as 68 mm Hg

62

Interpreting Significance

- Studies C and D are both statistically significant results
  - Only study C demonstrated clinically important differences
  - The results of study D are only frequently obtained if the treatment truly lowered SBP by 0.47 mm Hg or less

63

Bottom Line

- If ink is not in short supply, there is no reason not to give point estimates, CI, and P value
- If ink is in short supply, the confidence interval provides most information
  - (but sometimes a confidence interval cannot be easily obtained, because the sampling distribution is unknown under the null)

64

Full Report of Analysis

Study	n	SBP Diff	95% CI	P value
A	20	27.16	-14.14, 68.46	0.1974
B	20	0.27	-0.14, 0.68	0.1974
C	80	27.16	6.51, 47.81	0.0099
D	80	0.27	0.06, 0.47	0.0099

65

Take Home Message 1

- **Statistical criteria for evidence (frequentist vs Bayesian)**
  - Frequentist inference is best starting place for RCT design
    - In designed experiment, frequentist inference has solid foundations
    - Regulatory and journal standards more often on frequentist basis
      - 95% confidence intervals, level 0.05 / .025 tests
  - Approximate Bayesian inference can be easily derived from frequentist inference (easier than vice versa)
    - Sensitivity of conclusions to a spectrum of priors can be effected using frequentist sampling distribution and conjugate normal priors
  - Emphasize interval and point estimates rather than p values
    - Facilitates clinical and scientific interpretation of results

66

7

Probability Models for RCT Outcomes

Where am I going?

Distribution-free models of outcome acknowledge that a treatment can affect many aspects of a distribution

Take Home Message 2

- **Probability models for RCT outcomes**
  - Avoid making strong (semi)parametric assumptions across treatment groups
    - (Semi)parametric assumptions across arms are not consistent with possible heterogeneity of effect across individuals
    - Fortunately many distribution-free approaches are very similar to some standard (semi)parametric model
  - Summarize outcome distribution according to clinical relevance
    - Plausibility of effect and statistical precision are secondary issues
  - Use distribution free estimates of summaries and their SE whenever possible
    - Such estimates most often have an approximately normal sampling distribution

68

9

Precision of Inference

Where am I going?

A major (but not the only) role of statistical design is to maximize the precision with which clinical hypothesis can be investigated

Determinants of precision include study design, statistical analysis models, and sample size

Issues

- Summary measure
  - Mean, geometric mean, median, proportion, hazard…
- Structure of trial
  - One arm, two arms, k arms
  - Independent groups vs cross over
  - Cluster vs individual randomization
  - Randomization ratio
- Statistic
  - Parametric, semi-parametric, nonparametric
  - Adjustment for covariates
- Attenuated treatment effects
  - "Contamination" with patients who cannot benefit from treatment

70

Refining Scientific Hypotheses

- Scientific hypotheses are typically refined into statistical hypotheses by identifying some parameter θ measuring difference in distribution of response
  - Difference/ratio of means
  - Ratio of geometric means
  - Difference/ratio of medians
  - Difference/ratio of proportions
  - Odds ratio
  - Hazard ratio

71

Inference

- Generalizations from sample to population
  - Estimation
    - Point estimates
    - Interval estimates
  - Decision analysis (testing)
    - Quantifying strength of evidence

72

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

73

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

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

(V related to average "statistical information")

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

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

74

Ex: One Sample Mean

$$iid\ Y_i \sim \left(\mu, \sigma^2\right), i = 1, \ldots, n$$

$$\theta = \mu \qquad \hat{\theta} = \bar{Y}$$

$$V = \sigma^2 \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{\sigma^2}{n}}$$

75

Ex: Difference of Indep Means

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

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

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

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

76

Ex: Difference of Paired Means

$$Y_{ij} \sim \left(\mu_i, \sigma_i^2\right), i = 1,2;\ j = 1,\ldots,n$$
$$corr\left(Y_{1j}, Y_{2j}\right) = \rho; \quad corr\left(Y_{ij}, Y_{mk}\right) = 0 \text{ if } j \neq k$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

$$V = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}}$$

77

Ex: Mean of Clustered Data

$$Y_{ij} \sim \left(\mu, \sigma^2\right), i = 1,\ldots,n;\ j = 1,\ldots,m$$
$$corr\left(Y_{ij}, Y_{ik}\right) = \rho \text{ if } j \neq k; \quad corr\left(Y_{ij}, Y_{mk}\right) = 0 \text{ if } i \neq m$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

$$V = \sigma^2\left(\frac{1 + (m-1)\rho}{m}\right) \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}}$$

78

Ex: Independent Odds Ratios

$$ind\ Y_{ij} \sim B(1, p_i), i = 1,2;\ j = 1,\ldots,n_i$$

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

$$\theta = \log\left(\frac{p_1 / (1-p_1)}{p_2 / (1-p_2)} \right) \qquad \hat{\theta} = \log\left(\frac{\hat{p}_1 / (1-\hat{p}_1)}{\hat{p}_2 / (1-\hat{p}_2)} \right)$$

$$\sigma_i^2 = \frac{1}{p_1(1-p_1)} = \frac{1}{p_1 q_1}$$

$$V = (r+1)\left[\sigma_1^2 / r + \sigma_2^2\right] \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{1}{n_1 p_1 q_1} + \frac{1}{n_2 p_2 q_2}}$$

79

Ex: Diff of Indep Proportions

$$ind\ Y_{ij} \sim B(1, p_i), i = 1,2;\ j = 1,\ldots,n_i$$

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

$$\theta = p_1 - p_2 \qquad \hat{\theta} = \hat{p}_1 - \hat{p}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

$$\sigma_i^2 = p_i(1-p_i)$$

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

80

Ex: Hazard Ratios

ind censored time to event (T_{ij}, δ_{ij}),

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

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

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

31

Ex: Linear Regression

$$ind\ Y_i \mid X_i \sim (\beta_0 + \beta_1 \times X_i, \sigma^2_{Y|X}),\ i = 1,\ldots,n$$

$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1$ from LS regression

$$V = \frac{\sigma^2_{Y|X}}{Var(X)} \qquad se(\hat{\theta}) = \sqrt{\frac{\sigma^2_{Y|X}}{nVar(X)}}$$

32

Controlling Variation

- In a two sample comparison of means, we might control some variable in order to decrease the within group variability
  - Restrict population sampled
  - Standardize ancillary treatments
  - Standardize measurement procedure

83

Adjusting for Covariates

- When comparing means using stratified analyses or linear regression, adjustment for precision variables decreases the within group standard deviation
  - Var (Y | X) vs Var (Y | X, W)

84

Ex: Linear Regression

$$ind\ Y_i \mid X_i, W_i \sim \left(\beta_0 + \beta_1 \times X_i + \beta_2 \times W_i, \sigma^2_{Y|X,W}\right), i = 1, \ldots, n$$

$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

$$V = \frac{\sigma^2_{Y|X,W}}{Var(X)(1 - r^2_{XW})} \qquad se(\hat{\theta}) = \sqrt{\frac{\sigma^2_{Y|X,W}}{nVar(X)(1 - r^2_{XW})}}$$

$$\sigma^2_{Y|X,W} = \sigma^2_{Y|X} - \beta_2^2 Var(W \mid X)$$

35

Precision with Proportions

- When analyzing proportions (means), the mean variance relationship is important
  - Precision is greatest when proportion is close to 0 or 1
  - Greater homogeneity of groups makes results more deterministic
    - (At least, I always hope for this)

36

Ex: Diff of Indep Proportions

$$ind\ Y_{ij} \sim B(1, p_i),\ i = 1,2;\ j = 1,\ldots,n_i$$

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

$$\theta = p_1 - p_2 \qquad \hat{\theta} = \hat{p}_1 - \hat{p}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

$$\sigma_i^2 = p_i(1 - p_i)$$

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

37

Precision with Odds

- When analyzing odds (a nonlinear function of the mean), adjusting for a precision variable results in more extreme estimates
  - odds = *p / (1-p)*
  - odds using average of stratum specific *p* is not the average of stratum specific odds
- Generally, little "precision" is gained due to the mean-variance relationship
  - Unless the precision variable is highly predictive

38

Precision with Hazards

- When analyzing hazards, adjusting for a precision variable results in more extreme estimates
- The standard error tends to still be related to the number of observed events
  - Higher hazard ratio with same standard error ➔ greater precision

89

Special Case: Baseline Adjustment

- Options
  - Final only (throw away baseline)
    - $V = 2\sigma^2$
  - Change (final – baseline)
    - $V = 4\sigma^2 (1 - \rho)$
  - ANCOVA (change or final adj for baseline
    - $V = 2\sigma^2 (1 - \rho^2)$
- Advantages (more precision) based on value of ρ

$-1 \leq \rho < 0$: "ANCOVA" better than "Final only" which is better than "Change"
$\rho = 0$: "ANCOVA" same as "Final only" which is better than "Change"
$0 < \rho < 0.5$: "ANCOVA" better than "Final only" which is better than "Change"
$\rho = 0.5$: "ANCOVA" better than "Final only" which is same as "Change"
$0.5 < \rho < 1$: "ANCOVA" better than "Change" which is better than "Final Only"
$\rho = 1$: "ANCOVA" same as "Change" which is better than "Final Only"

90

Ex: ANCOVA (Baseline Adjustment)

$$ind\ Y_{fi} \mid X_i \sim \left(\beta_0 + \beta_1 \times X_i + \beta_1 \times Y_{0i}, \sigma^2_{Y|X,Y_0}\right),$$
$$i = 1, \ldots, n \qquad \rho = corr\left(Y_{0i}, Y_{fi}\right)$$
$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

$$V = \frac{\sigma^2_{Y|X}\left(1-\rho^2\right)}{Var(X)} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{\sigma^2_{Y|X}}{nVar(X)}}$$

91

Heteroscedasticity

- Previous results presumed common variance and common correlation across treatment arms
- If variance or correlation varies across treatment arm, ideal is

$$a = \frac{n_0}{n_0+n_1}\rho_1\frac{\sigma_{1f}}{\sigma_{1b}} + \frac{n_1}{n_0+n_1}\rho_0\frac{\sigma_{0f}}{\sigma_{0b}}$$

- Note that if sample sizes are equal, this is same as ANCOVA model
  - Otherwise, need to estimate from each arm separately

92

Criteria for Precision

- Standard error
- Width of confidence interval
- Statistical power
  - Probability of rejecting the null hypothesis
    - Select “design alternative”
    - Select desired power

93

Statistics to Address Variability

- At the end of the study:
  - 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

94

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

95

Sample Size Computation

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

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

Design alternative $\theta = \theta_1$

Variability V within 1 sampling unit

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

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

96

When Sample Size Constrained

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

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

  - Compute alternative detected with high power

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

97

Increasing Precision

- Options
  - Increase sample size
  - Decrease V
  - (Decrease confidence level)

98

Comparison of Study Designs

– Single Arm: Mean; absolute reference	N= 25
– Single Arm: Mean; historical data	50
– Two Arms : Diff in Means	100
– Two Arms : Diff in Mean Change (r = 0.3)	140
– Two Arms : Diff in Mean Change (r = 0.8)	40
– Two Arms : ANCOVA (r = 0.3)	81
– Two Arms : ANCOVA (r = 0.8)	36
– Cross-over: Diff in Means (r = 0.3)	70
– Cross-over: Diff in Means (r = 0.8)	20

99

General Comments: Alternative

- What alternative to use?
  - Minimal clinically important difference (MCID)
    - To detect? (use in sample size formula)
    - To declare significant? (look at critical value)
  - Subterfuge: 80% or 90%

100

General Comments: Level

- What level of significance?
  - "Standard": one-sided 0.025, two-sided 0.05
  - "Pivotal": one-sided 0.005?
    - Do we want to be extremely confident of an effect, or confident of an extreme effect

101

General Comments: Power

- What power?
  - Science: 97.5%
    - Unless MCID for significance→ ~50%
  - Subterfuge: 80% or 90%

102

Role of Secondary Analyses

- We choose a primary outcome to avoid multiple comparison problems
  - That primary outcome may be a composite of several clinical outcomes, but there will only be one CI, test
- We select a few secondary outcomes to provide supporting evidence or confirmation of mechanisms
  - Those secondary outcomes may be
    - alternative clinical measures and/or
    - different summary measures of the primary clinical endpoint

103

Secondary Analysis Models

- Selection of statistical models for secondary analyses should generally adhere to same principles as for primary outcome, including intent to treat
- Some exceptions:
  - Exploratory analyses based on dose actually taken may be undertaken to generate hypotheses about dose response
  - Exploratory cause specific time to event analyses may be used to investigate hypothesized mechanisms

104

Subgroups

- Testing for effects in K subgroups
  - Does the treatment work in each subgroup?
  - Bonferroni correction: Test at α / K
    - No subgroups: N = 100
    - Two subgroups: N = 230
- Testing for interactions across subgroups
  - Does the treatment work differently in subgroups?
    - Two subgroups: N = 400

105

Safety Outcomes

- During the conduct of the trial, patients are monitored for adverse events (AEs) and serious adverse events (SAEs)
  - We do not typically demand statistical significance before we worry about the safety profile
    - We must consider the severity of the AE / SAE

106

Safety Outcomes: Conservatism

- If we perform statistical tests, it is imperative that we not use overly conservative procedures
  - When looking for rare events, Fisher's Exact Test is far too conservative
    - Safety criteria based on nonsignificance of FET is a license to kill
  - Unconditional exact tests provide much better power

107

Sample Size Considerations

- We can only choose one sample size
  - Secondary and safety outcomes may be under- or over-powered
- With safety outcomes in particular, we should consider our information about rare, devastating outcomes (e.g., fulminant liver failure in a generally healthy population)
  - The "three over N" rule pertains here
  - Ensure minimal number of treated individuals
    - Control groups are not as important here, if the event is truly rare

108

Sample Size Re-estimation

- Modify sample size to account for estimated information (variance or baseline rates)
- No effect on type I error IF
  - Estimated information independent of estimate of treatment effect
    - Proportional hazards,
    - Normal data, and/or
    - Carefully phrased alternatives
  - And willing to use conditional inference
    - Carefully phrased alternatives

109

Estimate Alternative

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

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

110

Estimate Sample Size

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

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

111

Attenuation Due to "Contamination"

- We sometimes anticipate that some accrued subjects will not be susceptible to the treatment
  - E.g., misdiagnosis of disease
- In that setting, both the mean and variance of the treatment effect can be affected
- We can, however use standard formulas to attenuate the expected treatment effect when proportion *p* is nonsuceptible
  - $\theta_T = (1-p)\theta_S + p\theta_{NS}$
  - $Var(\widehat{\theta_T}) = E[Var(\hat{\theta}|S)] + Var[E(\hat{\theta}|S)]$

112

Take Home Message 3

- **Precision of inference**
  - Precision of RCT is measured in width of interval estimates and statistical power to discriminate between important hypotheses
  - Statistical formulas to quantify precision are similar across a wide variety of commonly used statistical analysis models
  - Determinants of precision modifiable in RCT design include
    - Completeness of outcome collection (noninformative censoring)
    - Distributional summary measure used to assess treatment effect
    - Homogeneity of outcome within groups compared across arms
    - Ratio of sample sizes across groups compared across arms
    - Total sample size accrued to study
    - Avoiding potentially informative missing data

113

4

Summarizing Distributions of Outcomes

Where am I going?

RCT hypotheses are stated in terms of “tendencies” toward larger or smaller outcomes

Clinical and scientific issues should drive the choice of mean, geometric mean, proportions, hazards, etc.

But some statistical considerations can also aid in choice of outcome

Refining Scientific Hypotheses

- Scientific hypotheses are typically refined into statistical hypotheses by identifying some parameter θ measuring difference in distribution of response
  - Difference/ratio of means
  - Ratio of geometric means
  - Difference/ratio of medians
  - Difference/ratio of proportions
  - Odds ratio
  - Hazard ratio

115

Geometric Mean

- Geometric mean is related to mean of log transformed data
  - All comments about the arithmetic mean need to apply to the log transformed data
  - Good scientific foundations for many biomedical measurements
- Greater statistical precision than mean when standard deviations are proportional to mean
- When variables are measured on a exponential scale, geometric means tend to be more stable
  - E.g., Serial dilutions used for measuring titers
- Ratios of geometric means tend to be more stable than ratios of arithmetic means.
  - The log of a ratio is the difference of logs
    - » http://www.emersonstatistics.com/GeneralMaterials

116

Take Home Message 4

- **Summarizing distribution of outcomes**
  - Choose summary measure according to (in order of importance):
    - Estimable within logistical constraints of RCT
    - Clinically / scientifically most important
    - Plausibly affected by treatment
    - Estimated with greatest statistical precision
  - Issues not related to precision
    - Univariate summaries allow better comparisons across studies
    - Proportion / odds above threshold often clinically most important
    - Means / geometric means may better quantify effects in population
    - With heavily censored data, some summaries may not be estimable
  - Means / geometric means most precise for many distributions
    - Means: normal, Poisson, binomial, exponential
    - Geometric means: lognormal

117

8

Contrasting Treatments Across Treatment Arms

Where am I going?

Differences between or ratios of summary measures across treatment arms are most commonly used in the quantification of treatment effect

Take Home Message 5

- **Contrasting outcome distributions across treatment arms**
  - Differences or ratios of univariate summaries most interpretable
    - Allow estimation and testing on same measures across studies
    - Less desirable by this criterion: Wilcoxon rank sum test
  - With binary outcomes
    - Difference in proportions (attributable risk) gives population impact
    - Ratio of proportions (relative risk) is invariant to "contamination" with subjects insusceptible to treatment (either always 0 OR 1, not both)
    - Odds ratio may be better in adjusted analyses (less effect modification) and mimics relative risk in rare outcomes
  - Differences are expressed in units of original measurements
  - Ratios are dimensionless, hence need anchoring with baseline
    - Odds and geometric means tend to use ratios as contrasts

119

» http://www.emersonstatistics.com/GeneralMaterials

120

Adjusting for Baseline Covariates in Analyses of RCT Results

Where am I going?

Randomization precludes confounding on average.

However, precision of inference can be gained by making comparisons across groups that are more similar.

Adjusted analyses is one way of gaining such precision.

Regression Models

- According to the parameter compared across groups
  - Means → Linear regression
  - Geom Means → Linear regression on logs
  - Odds → Logistic regression
  - Rates → Poisson regression
  - Hazards → Proportional Hazards regr
  - Quantiles → Parametric (AFT) survival regr

121

Statistical Analysis Models in RCT

- Consider the distribution of response across groups defined by some "predictor of interest" (POI)
- We choose some summary measure of our response distribution
- We use regression to model our question
  - Simple regression with a binary POI often the standard test
- We estimate a "contrast"
  - Usually difference or ratio across treatment groups
- We operate as distribution-free as possible
  - Frequently: Methods originally derived under strong parametric models are found to be robust in a distribution-free sense

122

General Regression Notation

- General notation for variables and parameter

Y_i	Response measured on the ith subject
X_i	Value of the POI for the ith subject
$W_{1i}, W_{2i}, \ldots$	Value of adjustment variables for the ith subject
θ_i	Parameter of distribution of Y_i

- The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

123

Multiple Regression

- General notation for multiple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

$g(\ )$	"link" function used for modeling
β_0	"Intercept"
β_1	"Slope for Pred of Interest X"
β_j	"Slope for covariate W_{j-1}"

- The link function is usually either
  - none (difference of means), or
  - log (ratio of means, geom means, odds, hazards, quantiles)

124

Comparison of Models

- The major difference between regression models is interpretation of the parameters
  - Summary: Mean, geometric mean, odds, hazards
  - Comparison of groups: Difference, ratio
- Issues related to inclusion of covariates remain the same
  - Address the scientific question
    - Predictor of interest (sometimes modeled with multiple variables)
    - Effect modifiers
  - Address confounding
  - Increase precision

125

Adjustment for Covariates

- We “adjust” for other covariates
- Define groups according to
  - Predictor of interest, and
  - Other covariates
- Compare the distribution of response across groups which
  - differ with respect to the Predictor of Interest, but
  - are the same with respect to the other covariates
    - “holding other variables constant”

126

Unadjusted vs Adjusted Models

- Adjustment for covariates changes the scientific question
- Unadjusted models
  - Slope compares parameters across groups differing by 1 unit in the modeled predictor
    - Groups may also differ with respect to other variables
- Adjusted models
  - Slope compares parameters across groups differing by 1 unit in the modeled predictor but similar with respect to other modeled covariates
- (In RCT investigating a difference in means, the two questions may have similar numeric answers due to double expectation)

127

Interpretation of Slopes

- Difference in interpretation of slopes

$$\text{Unadj Model}: \quad g\left[\theta \,\middle|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

  - β_1 = Compares θ for groups differing by 1 unit in X
    - (The distribution of W might differ across groups being compared)

$$\text{Adj Model}: \; g\left[\theta \,\middle|\, X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

  - γ_1 = Compares θ for groups differing by 1 unit in X, but agreeing in their values of W

128

Comparing models

Unadjusted $g[\theta | X_i] = \beta_0 + \beta_1 \times X_i$

Adjusted $g[\theta | X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

Science: When is $\gamma_1 = \beta_1$?

When is $\hat{\gamma}_1 = \hat{\beta}_1$?

Statistics: When is $se(\hat{\gamma}_1) = se(\hat{\beta}_1)$?

When is $s\hat{e}(\hat{\gamma}_1) = s\hat{e}(\hat{\beta}_1)$?

129

General Results

- These questions can not be answered precisely in the general case
- However, in linear regression we can derive exact results
- These will serve as a basis for later examination of
  - Logistic regression
  - Poisson regression
  - Proportional hazards regression

130

Linear Regression

- Difference in interpretation of slopes

$$\text{Unadjusted Model}: \quad E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 \times X_i$$

  - β_1 = Diff in mean Y for groups differing by 1 unit in X
    - (The distribution of W might differ across groups being compared)

$$\text{Adjusted Model}: \quad E\left[Y_i \mid X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

  - γ_1 = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

131

Relationships: True Slopes

- The slope of the unadjusted model will tend to be

$$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

- Hence, true adjusted and unadjusted slopes for X are estimating the same quantity only if

  - $\gamma_2 \rho_{XW} = 0$ (X and W are truly uncorrelated), OR
  - (no association between W and Y after adjusting for X)

132

Relationships: Estimated Slopes

- The estimated slope of the unadjusted model will be

$$\hat{\beta}_1 = \hat{\gamma}_1\left(1+\hat{\gamma}_2 r_{XW}\left[\frac{s_W}{s_X\left(r_{YX}-r_{YW}r_{XW}\right)}\right]\right)$$

- Hence, estimated adjusted and unadjusted slopes for X are equal only if

$$\hat{\gamma}_2 = 0$$

  - $r_{XW} = 0$ (X and W are uncorrelated in the sample, which can be arranged by experimental design), OR
    (which cannot be predetermined, because Y is random)
  - $s_W = 0$ (W is controlled at a single value in which case $r_{XW} = 0$)

133

Relationships: True SE

Unadjusted Model $$\left[se\left(\hat{\beta}_1\right)\right]^2 = \frac{Var\left(Y\,|\,X\right)}{nVar\left(X\right)}$$

Adjusted Model $$\left[se\left(\hat{\gamma}_1\right)\right]^2 = \frac{Var\left(Y\,|\,X,W\right)}{nVar\left(X\right)\left(1-r_{XW}^2\right)}$$

$$Var\left(Y\,|\,X\right) = \gamma_2^2 Var\left(W\,|\,X\right) + Var\left(Y\,|\,X,W\right)$$

$$\sigma_{Y|X}^2 = \gamma_2^2\sigma_{W|X}^2 + \sigma_{Y|X,W}^2$$

134

Relationships: True SE

Unadjusted Model $\left[se\left(\hat{\beta}_1\right)\right]^2 = \dfrac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

Adjusted Model $\left[se\left(\hat{\gamma}_1\right)\right]^2 = \dfrac{Var\left(Y \mid X, W\right)}{nVar\left(X\right)\left(1-r_{XW}^2\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

Thus, $se\left(\hat{\beta}_1\right) = se\left(\hat{\gamma}_1\right)$ if

$$r_{XW} = 0$$

AND

$$\gamma_2 = 0 \quad \text{OR} \quad Var\left(W \mid X\right) = 0$$

135

Relationships: Estimated SE

Unadjusted Model $\left[s\hat{e}\left(\hat{\beta}_1\right)\right]^2 = \dfrac{SSE\left(Y \mid X\right)/\left(n-2\right)}{\left(n-1\right)s_X^2}$

Adjusted Model $\left[s\hat{e}\left(\hat{\gamma}_1\right)\right]^2 = \dfrac{SSE\left(Y \mid X, W\right)/\left(n-3\right)}{\left(n-1\right)s_X^2\left(1-r_{XW}^2\right)}$

Thus, $s\hat{e}\left(\hat{\beta}_1\right) = s\hat{e}\left(\hat{\gamma}_1\right)$ if

$$r_{XW} = 0$$

AND

$$SSE\left(Y \mid X\right)/\left(n-2\right) = SSE\left(Y \mid X, W\right)/\left(n-3\right)$$

136

Special Cases

- Behavior of unadjusted and adjusted models according to whether
  - *X* and *W* are uncorrelated (no association in means)
  - *W* is associated with *Y* after adjustment for *X*

	$r_{XW} = 0$	$r_{XW} \neq 0$
$\gamma_2 \neq 0$	Precision	Confounding
$\gamma_2 = 0$	Irrelevant	Var Inflation

137

Simulations

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

$$\text{Sampling} \quad : \quad E\left[W_i \,|\, X_i\right] = \alpha_0 + \alpha_1 \times X_i$$

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

$$\text{Unadjusted} : \quad g\left[\theta \,|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

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

138

Linear Regression

- Simulation results

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

139

Linear Regression

- Simulation results

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

140

Linear Regression

- Simulation results

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

141

Linear Regression

- Simulation results

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

142

Linear Regression

- Simulation results

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

143

Linear Regression

- Simulation results

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

144

Linear Regr Take Home: Estimate

- The magnitude of confounding is a product of
  - Magnitude of association between covariate and response AND
  - Difference of mean covariate value across POI groups
    - If adjustment would be linear, mean (not median, etc) matters
- If POI and covariate independent, adjusted and unadjusted estimates will tend to be equal
  - Randomization will lead to such independence
- If POI and covariate orthogonal, adjusted and unadjusted estimates will be exactly equal
  - Stratified, blocked randomization will lead to such orthogonality

145

Linear Regr Take Home: Std Err

- Adjusting for a confounder, precision of estimates for response-POI association can be larger or smaller than unadjusted analysis
- Association between covariate and response will tend to decrease SE in adjusted model
  - Thus some increased precision from modeling important prognostic variable in RCT
- Difference in mean covariate across treatment groups in sample will tend to increase SE in adjusted model
  - But in RCT, any such difference in means will tend to be small in large sample sizes or with stratified blocked randomization
  - And covariate adaptive minimization could attempt to minimize difference in means

146

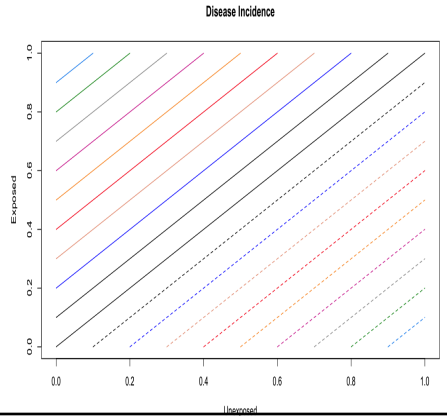
Noncollapsibility

- In logistic regression and proportional hazards regression, the differences between unadjusted and adjusted analyses is also affected by "noncollapsibility"
  - Double expectation formula does not protect us
- Even in the absence of confounding, the odds ratio computed when combining two strata can differ from the stratum specific odds ratios
  - And the hazard ratios behave similarly to the odds ratios
- We can consider l'Abbe plots

147

Contours: Equal Risk Difference (RD)

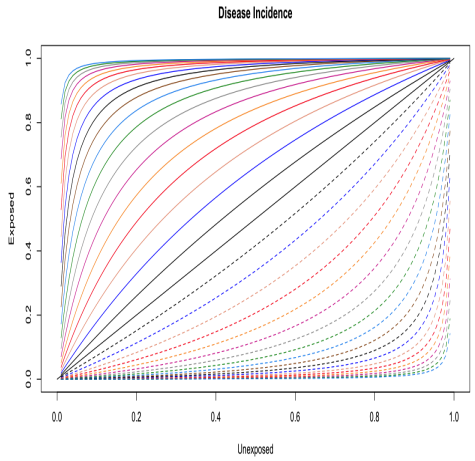
- Graph of all possible values for disease incidences
  - Two strata having the same RD will lie on the same line
  - Owing to the double expectation formula, with no confounding, the RD for the combined strata will also be on that same line



148

Contours: Equal Odds Ratio (OR)

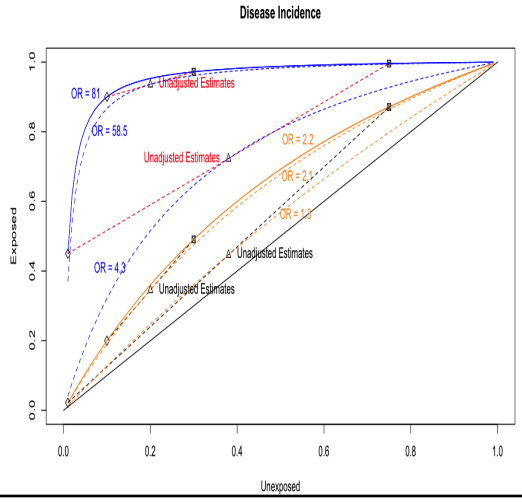
- Graph of all possible values for disease incidences
  - Two strata having the same OR will lie on the same line



149

Deattenuation of OR

- When "collapsing strata", the OR will lie on a different contour
  - Amount of attenuation will depend on prognostic strength



150

Logistic Regression

- Simulation results

	Truth					Avg Estimates (SE)	
	ΔMdn	α_1	r_{XW}	γ_2	γ_1	β_1	γ_1
Irrelevant	0.0	0.0	0.00	0.0	0.0	0.0 (0.42)	0.0 (0.42)
Precision	0.0	0.0	0.00	1.0	0.0	0.0 (0.40)	0.0 (0.42)
Precision	- 0.3	0.0	0.00	1.0	0.0	0.0 (0.42)	0.0 (0.43)
Precision	0.0	0.0	0.00	1.0	1.0	0.8 (0.43)	1.0 (0.49)
Confound	0.3	0.3	0.15	1.0	0.0	0.3 (0.43)	0.0 (0.48)
Confound	0.0	0.3	0.15	1.0	0.0	0.2 (0.41)	0.0 (0.47)
Var Inflatn	0.0	1.0	0.45	0.0	0.0	0.0 (0.41)	0.0 (0.47)

151

Logist Regr Take Home: Estimate

- The magnitude of confounding is primarily a function of
  - Magnitude of association between covariate and response AND
  - Difference of mean covariate value across POI groups
    - If adjustment would be linear, mean (not median, etc) matters most
- Even if no confounding, adjusted and unadjusted estimates will be different if new covariate (conditionally) associated with response
  - Adjusting for a precision variable will “deattenuate” OR (and estimate)
  - The amount of “deattenuation” depends on
    - The strength of association between the added covariate and the response, and
    - The adjusted strength of association between the POI and the response

152

Logist Regr Take Home: Std Err

- Standard errors of estimated OR measuring POI – response association are largely driven by the mean-variance relationship
  - SE of odds behaves like 1 / p(1-p)
  - Greater homogeneity of groups leads to p closer to 0 or 1
  - Usually statistical significance of the POI – response association is affected very little by adjustment for a "precision" variable unless it is very strongly associated with response

- Note that if we persist in estimating the population OR instead of the adjusted OR, we do gain precision by adjustment
  - We do have to take a weighted average, however
  - See Tangen & Koch, *Stat Med*, 2000

153

Proportional Hazards Regression

- Simulation results

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

154

Next

PH Regr Take Home: Estimate

- The magnitude of confounding is primarily a function of
  - Magnitude of association between covariate and response AND
  - Difference of mean covariate value across POI groups
    - If adjustment would be linear, mean (not median, etc) matters most
- Even if no confounding, adjusted and unadjusted estimates will be different if new covariate (conditionally) associated with response (PH regression is related to logistic regression)
  - Adjusting for a precision variable will "deattenuate" HR (and estimate)
  - The amount of "deattenuation" depends on
    - The strength of association between the added covariate and the response, and
    - The adjusted strength of association between the POI and the response

155

PH Regr Take Home: Std Err

- Standard errors of estimated HR measuring POI – response association are largely driven by the mean-variance relationship of the sample size ratios
  - Unless there is a very strong association, the SE tends to be fairly constant for a range of different HRs
- Holding the SE constant, we will obtain a lower p value with a deattenuated HR

156

Precision: Linear Regression

- E.g., X, W independent in population (or completely randomized experiment) AND W associated with Y independent of X

$$\rho_{XW} = 0 \qquad \gamma_2 \neq 0$$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 \approx \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) > se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

157

Precision: Logistic Regression

- Adjusting for a precision variable
  - Deattenuates slope away from the null
  - Standard errors reflect mean-variance relationship
    - Substantially increased power only in extreme cases
      - » (OR > 5 for equal samples sizes of binary W)

		True Value	Estimates
Slopes	$\beta_1 > 0$:	$0 < \beta_1 < \gamma_1$	$0 < \hat{\beta}_1 < \hat{\gamma}_1$
	$\beta_1 < 0$:	$0 > \beta_1 > \gamma_1$	$0 > \hat{\beta}_1 > \hat{\gamma}_1$
Std Errs		$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

158

Precision: Poisson Regression

- Adjusting for a precision variable (with robust SE)
  - No effect on the slope (similar to linear regression)
    - log ratios are linear in log means
  - Standard errors reflect mean-variance relationship
    - Virtually no effect on power

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 \approx \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) > se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

159

Precision: PH Regression

- Adjusting for a precision variable
  - Deattenuates slope away from the null
  - Standard errors stay fairly constant
    - (Complicated result of binomial mean-variance)

		True Value	Estimates
Slopes	$\beta_1 > 0$:	$0 < \beta_1 < \gamma_1$	$0 < \hat{\beta}_1 < \hat{\gamma}_1$
	$\beta_1 < 0$:	$0 > \beta_1 > \gamma_1$	$0 > \hat{\beta}_1 > \hat{\gamma}_1$
Std Errs		$se(\hat{\beta}_1) \approx se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) \approx s\hat{e}(\hat{\gamma}_1)$

160

Lin Reg: Stratified Randomization

- Stratified (orthogonal) randomization in a designed experiment
  - Also when designing observational study with matching
- Unless we adjust for the stratification variables, we estimate the true standard errors incorrectly

$r_{XW} = 0 \qquad \gamma_2 \neq 0$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) = se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

161

Take Home Message 6

- **Why control for baseline variables in RCT analyses?**
  - Precision of inference is the primary issue
    - Confounding is in some sense avoided by randomization
      - However, some critics will second guess results *post hoc*
    - Estimation of "within stratum" effect an issue with odds, hazards
- **How to control for baseline variables in RCT design?**
  - Restrict eligibility to subset of eventual target population
    - Only if certain no effect modification on efficacy or safety
  - Stratified randomization based on prognostic factors especially if
    - Want face validity of key prognostic factors in Table 1
    - Interested in prespecified analyses within subgroups
    - Covariate adaptive minimization can be usually be avoided
  - Obtain precision by prespecified adjustment for the most important prognostic factors known *a priori*
    - Do not allow data driven selection of adjusted models

162

3

Randomization Strategies

Where am I going?

Randomization is the major tool by which we maintain scientific rigor in our attempts to investigate cause and effect

Alternative randomization strategies can be used to increase precision

- Randomization ratio
- Blocking
- Stratification

Treatment of Variables

- Measure and compare distribution across groups
  - Response variable in regression
- Vary systematically (intervention)
- Control at a single level (fixed effects)
- Control at multiple levels (fixed or random effects)
  - Stratified (blocked) randomization
- Measure and adjust (fixed or random effects)
- Treat as “error”

16-4

Complete Randomization (CRD)

- With each accrued subject a (possibly biased) coin is tossed to determine which arm
  - Probability of treatment arm = *r / (r + 1)*
  - Independence of successive randomizations
- Issues
  - Bias
  - Face validity
  - Precision

165

CRD: Unbiased

- On average (across repeated experiments)
  - No correlation between treatment variable and other covariates
  - Individual type I errors come from samples in which other covariates are imbalanced

$$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

166

Face Validity: Table 1

	Methotrexate Arm		Placebo Arm	
	n	Mean (SD; Min – Max)	n	Mean (SD; Min – Max)
Age (yrs)	132	50.4 (8.5; 32 - 69)	133	52.2 (8.5; 26 - 67)
Female	132	92.4%	133	92.5%
Pruritus score	116	7.7 (3.8; 4 - 16)	124	6.9 (3.8; 4 - 20)
Splenomegaly	131	8.4%	133	10.5%
Telangiectasia	132	4.6%	133	11.3%
Edema	132	6.1%	133	3.0%
Alkaline phosphatase	132	242.6 (145.9; 53 - 933)	133	245.0 (187.6; 66 - 1130)
ALT	131	54.5 (41.7; 12 - 202)	132	50.6 (41.4; 12 - 311)
Total bilirubin	132	0.7 (0.4; 0.1 - 2.7)	133	0.7 (0.4; 0.1 - 2.4)
Albumin	132	4.0 (0.3; 3.1 - 6.0)	133	4.0 (0.3; 3.0 - 4.8)
Prothrombin time INR	124	1.0 (0.1; 0.7 - 1.3)	132	1.0 (0.1; 0.7 - 1.3)
Mayo score	128	3.8 (0.8; 1.6 - 6.3)	133	3.9 (0.8; 1.6 - 6.1)
Avg stage	128	2.2 (0.9; 1.0 - 4.0)	128	2.3 (0.9; 1.0 - 4.0)
Avg fibrosis	128	1.2 (0.8; 0.0 - 3.0)	128	1.3 (0.9; 0.0 - 3.0)

167

CRD: Face Validity

- Table 1: Potential for imbalance in covariates
  - Depends on number of covariates and correlations among them
  - Probability of at least one “significant” imbalance
-

Number Displayed	Worst Case	Correlation 0.00	0.30	0.50	0.75	0.90
1	.050	.050	.050	.050	.050	.050
2	.100	.098	.095	.090	.081	.070
3	.150	.143	.137	.126	.104	.084
5	.250	.226	.208	.184	.138	.101
10	.500	.401	.353	.284	.193	.127
20	1.000	.642	.540	.420	.258	.154
50	1.000	.923	.806	.624	.353	.193

168

CRD: Face Validity

- Of course, statistical significance is not the issue
  - "Conditional confounding"
    - How does unadjusted estimate compare to adjusted estimate?
    - Product of sample correlation between *X* and *W* and adjusted association between *Y* and *W*

$$\beta_1 = \gamma_1 + r_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

169

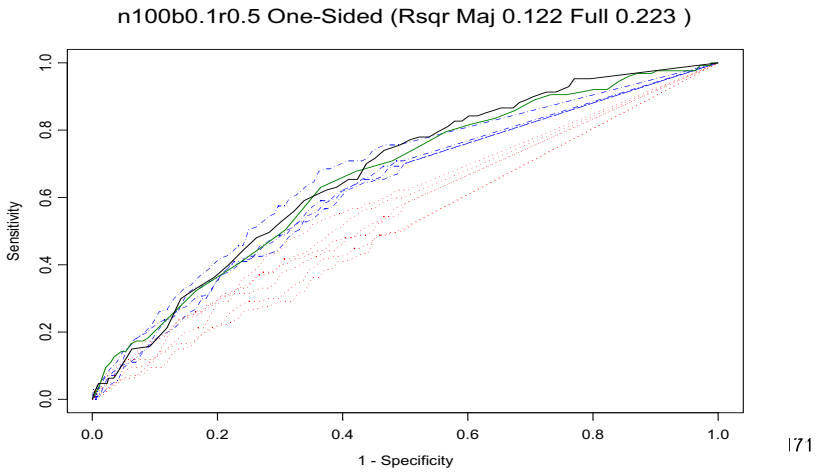
Demonstration of the Problem

- Consider a CRD in presence of
  - 4 highly correlated predictors with larger importance
  - 6 independent predictors with smaller importance
  - No treatment effect
- Questions about unadjusted analysis
  - What is type I error? ➔ 0.025
  - What does imbalance in predictors tell us about type I error?
    - Sensitivity, specificity of imbalance in predictors under null hypothesis
    - Dependence on R^2 of measured covariates

170

Low Association: 1-sided

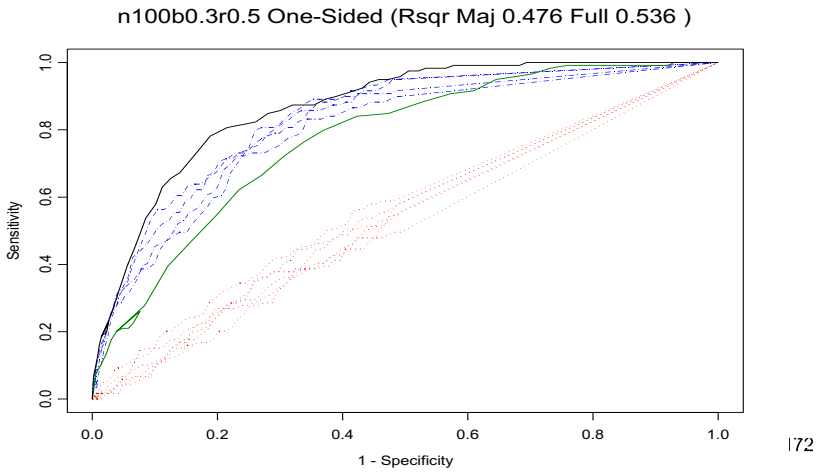
- ROC curve for covariate imbalance “explaining” statistical significance under the null



171

High Association: 1-sided

- ROC curve for covariate imbalance “explaining” statistical significance under the null



172

Face Validity

- Spurious results due to covariate imbalance
  - Unconditionally: Unbiased so no problem
  - Conditional on obtained randomization:
    - IF covariates are strongly predictive of outcome, then covariate imbalance is predictive of type I error
    - But need to consider that combined effect of other measured and unmeasured covariates may provide balance
- Ultimately, however, we need to have credible results
  - We do not always get to choose what others believe

173

Precision

- Impact of completely randomized design on precision of inference
  - Impact of imbalance in sample sizes
    - The number accrued to each arm is random
    - But maximum precision when $r = \frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$
  - Impact of imbalance in covariates
    - "One statistician's mean is another statistician's variance"

174

CRD: Improved Performance

- If we adjust for important covariates, we will often gain precision
  - Face validity in Table 1 if readers recognize that adjustment accounts for any observed imbalance
- Caveats:
  - If covariate imbalance by arm, model misspecification can be an issue re conditional bias
  - If covariate imbalance by arm, lack of effect can be an issue re variance inflation
  - If adjustment not TOTALLY prespecified, "intent to cheat" analysis can be an issue
    - Not too much loss of precision from imperfect model

175

CRD: Continuous vs Dichotomized

	CRD – Continuous Adjust				CRD – Dichotomized Adjust			
Tx Eff	SE Slope		Power		SE Slope		Power	
	Unadj	Adj	Unadj	Adj	Unadj	Adj	Unadj	Adj
0.0	.281	.211	.026	.024	.284	.231	.023	.026
0.1	.278	.209	.053	.062	.284	.229	.045	.062
0.3	.279	.209	.178	.285	.287	.231	.184	.243
0.5	.281	.209	.423	.655	.279	.225	.409	.581
0.7	.279	.209	.696	.909	.281	.229	.699	.858

176

Issues with CRD

- Imbalance across arms in covariate distribution
  - Loss of face validity
  - Conditional bias
  - Not much of an issue with large sample sizes
  - Could be problematic with sequential sampling
    - Interim analyses of data early in the study
  - Could be problematic with subgroup analyses
    - Possibility of very inefficient randomization ratio in small subgroups

177

Stratified Randomization

- Strata are defined based on values of important covariates
  - E.g., sex, age, disease severity, clinical site
- Within each stratum defined by a unique combination of stratification variables, CRD or blocked randomization
- Important caveats:
  - Number of strata is exponential in number of stratification variables
    - E.g., 4 two level stratification variables ➔ 16 strata

178

Statistical Inference

- Impact on statistical inference relative to CRD
  - Bias properties unchanged
  - Face validity improved for most important variables
  - Precision improved due to achieving closer to desired randomization ratio
  - Precision could be further improved if adjust for stratification variables in analysis
    - This should be done
      - Without adjustment for strata, may even lose power for some alternatives
    - Requires pre-specification of analysis model to avoid "intent to cheat" analysis

179

CRD vs Orthogonal Randomization

	CRD – Continuous Adjust				Orthogonal Randomization			
Tx Eff	SE Slope		Power		SE Slope		Power	
	Unadj	Adj	Unadj	Adj	Unadj	Adj	Unadj	Adj
0.0	.281	.211	.026	.024	.206	.206	.005	.026
0.1	.278	.209	.053	.062	.208	.208	.013	.069
0.3	.279	.209	.178	.285	.205	.205	.115	.313
0.5	.281	.209	.423	.655	.205	.205	.403	.684
0.7	.279	.209	.696	.909	.205	.205	.759	.924

180

Advantages

- Additional advantages of stratification
  - Balance within clinical center
    - Especially if quality control issues
  - Balance for interim analyses
  - Balance for subgroup analyses

131

Issues with Stratified Analyses

- The need to stratify on all combinations of variables
  - Good news:
    - Balances on interactions as well as main effects
  - Bad news:
    - Effect of interactions might be quite small
    - Really only need to adjust on “counterfactual” outcome based on linear combination of all covariates

132

Dynamic Randomization

- Subjects are assigned to the treatment arm that will achieve best balance
  - "Minimization": minimize the difference between the distribution of covariate effects between arms
    - Define a "distance" between arms for covariate vectors
    - Probability of assignment depends upon arm that would provide smallest difference

183

Distance Between Arms

- Two arms are "distant" based on one of:
  - Randomization ratio very different from $r : 1$ in some stratum
  - Summary measure of distribution of $(W_{i1},\ldots,W_{ip})$ differs
    - Mean, median, variance, …
  - Distribution of $(W_{i1},\ldots,W_{ip})$ differs
  - Contribution of covariates to the outcome differs

184

Conditional Confounding

$$\text{Unadjusted}: \quad g\left[\theta \,|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

$$\text{Adjusted} \quad : \; g\left[\theta \,|\, X_i, \vec{W}_i\right] = \gamma_0 + \gamma_1 \times X_i + \vec{W}_i^T \vec{\delta}$$

$$\text{Unadjusted}: \quad g\left[\theta \,|\, \mathbf{X}\right] = \mathbf{X}\vec{\beta}$$

$$\text{Adjusted} \quad : \; g\left[\theta \,|\, \mathbf{X}, \mathbf{W}\right] = \mathbf{X}\vec{\gamma} + \mathbf{W}\vec{\delta}$$

$$\vec{\beta} = \vec{\gamma} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\vec{\delta}$$

185

Conditional Confounding

$$g\left[\theta \,|\, \mathbf{X}\right] = \mathbf{X}\vec{\beta} \qquad g\left[\theta \,|\, \mathbf{X}, \mathbf{W}\right] = \mathbf{X}\vec{\gamma} + \mathbf{W}\vec{\delta}$$

$$\vec{\beta} = \vec{\gamma} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\vec{\delta}$$

$$\beta_1 = \gamma_1 + \sum_{j=1}^{p}\left(\overline{W}_{1j\bullet} - \overline{W}_{0j\bullet}\right)\delta_j$$

$$\overline{W}_{kj\bullet} = \frac{1}{n_k}\sum_{i=1}^{n} W_{ij} 1_{[X_i = k]}$$

186

Distance Metrics

Based on contribution to confounding:

$$d\left(\vec{X},\mathbf{W}\right)=\left|\sum_{j=1}^{p}\left(\overline{W}_{1j\bullet}-\overline{W}_{0j\bullet}\right)\delta_j\right|$$

Weighted distance between standardized means:

$$d\left(\vec{X},\mathbf{W}\right)=\sum_{j=1}^{p}c_j\left|\frac{\overline{W}_{1j\bullet}-\overline{W}_{0j\bullet}}{SD\left(W_j\right)}\right|^{\lambda}$$

Weighted imbalance in n across strata $\Omega_1,\ldots,\Omega_s$:

$$d\left(\vec{X},\mathbf{W}\right)=\sum_{s=1}^{S}c_s\left|\sum_{i=1}^{n}1_{[X_i=1]}1_{[\vec{W}_i\in\Omega_s]}-\sum_{i=1}^{n}1_{[X_i=0]}1_{[\vec{W}_i\in\Omega_s]}\right|^{\lambda}$$

137

Probability of Assignment

- Subjects are assigned to the treatment arm that will achieve best balance
  - When i-th patient accrued, compute a randomization probability

$$\Delta_i=d\left(\vec{X},\mathbf{W}\mid X_i=1\right)-d\left(\vec{X},\mathbf{W}\mid X_i=0\right)$$
$$\pi_i=\Pr\left(X_i=1\right)=f\left(\Delta_i\right)$$

  - $0 \leq \pi_i \leq 1$
  - Larger values of Δ_i ➔ smaller values of π_i
  - Probably best to avoid $\pi_i = 0$ and $\pi_i = 1$

138

Inference: Population Model

- Impact on statistical inference relative to CRD
  - Bias properties unchanged
  - Face validity improved for most important variables
  - Precision improved due to achieving closer to desired randomization ratio
  - Precision could be further improved if adjust for stratification variables in analysis for population model
    - This should be done
    - Requires pre-specification of analysis model to avoid "intent to cheat" analysis

189

Inference: Randomization Model

- Proschan, Brittain, Kammerman, 2010: Precision could be greatly hampered if you analyze under randomization hypothesis
- Alternative randomization schemes may be quite restrictive, especially under unequal randomization
  - Suppose sequential allocation
    - Randomization P value is identically 1 (or 0.5?)
  - If dynamic randomization has $\pi_i = 0$ or $\pi_i = 1$ too often, range of randomization P values is greatly restricted
- Also: Statistical analysis can be quite involved

190

Advantages / Disadvantages

- Advantages
  - Typically improved face validity
  - Can handle an arbitrary number of covariates
    - Depending on distance metric
- Disadvantages
  - Logistically more involved
  - Decreased credibility if too deterministic
    - Approaches sequential allocation
  - Some analytic strategies more complex
  - Does not necessarily facilitate subgroup analyses
    - Unless distance metric chosen carefully

191

Take Home Message 7

- **Randomization strategies**
  - 1:1 randomization unless really good reason
    - Most efficient randomization under strong null and homoscedasticity
  - Stratification
    - In large studies, no stratification necessary
    - In small studies:
      - Stratify on variables for which subgroup analyses are important
      - Stratify on clinical center if possible
      - No more than 4 important variables (including center)
      - No real advantage to dynamic minimization
    - Prespecify analysis including all fixed effects used in stratification
      - Less of an issue in logistic regression, but might as well
  - Use hidden blocks of varying sizes
    - Block sizes 4, 6, or 8 in 1:1 randomization

192

3

Group Sequential Adaptive Designs

Where am I going?

Issues related to efficiency and ethics are often addressed through period monitoring of accruing data.

Special methods must be used to protect the statistical inference from the multiple comparisons inherent in such sequential testing

To ensure that clinical, scientific, ethical, and statistical issues are adequately addressed, candidate designs must be carefully evaluated.

Classical Fixed Sample Designs

- Design stage:
  - Choose a sample size
- Conduct stage:
  - Recruit subjects, gather all the data
- Analysis stage:
  - When all data available, analyze and report

194

Statistical Sampling Plan

- Ethical and efficiency concerns are 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
    - Stop if scientifically relevant and statistically credible results have been obtained
  - If continuing, decide on any modifications to
    - sampling scheme
    - scientific / statistical hypotheses and/or

195

Ultimate Goal

- Modify the sample size accrued so that minimal number of subjects treated when
  - new treatment is harmful,
  - new treatment is minimally effective, or
  - new treatment is extremely effective
- Only proceed to maximal sample size when
  - not yet certain of treatment benefit, or
  - potential remains that results of clinical trial will eventually lead to modifying standard practice

196

Group Sequential Designs

- Design stage:
  - Choose an interim monitoring plan
  - Choose a maximal stopping time
    - Statistical information, sample size, calendar time
- Conduct stage:
  - Recruit subjects, gather data in groups
  - After each group, analyze for DMC
    - DMC recommends termination or continuation
- Analysis stage:
  - When study stops, analyze and report

197

Group Sequential Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined if independent of effect
- 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

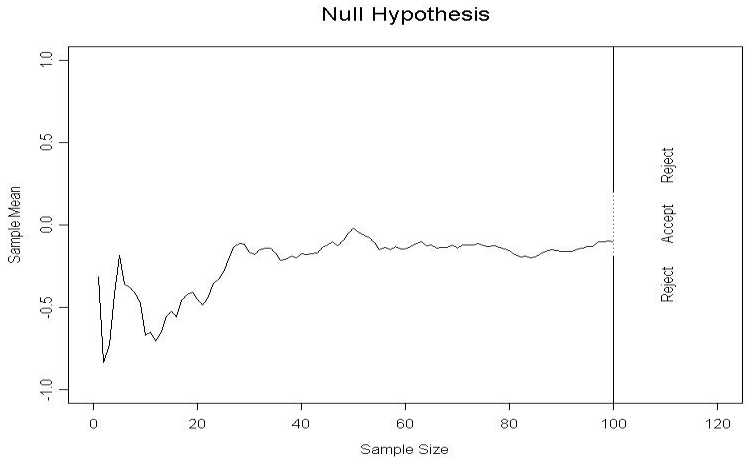
198

Statistical Design Issues

- Under what conditions should we use fewer subjects?
  - Ethical treatment of patients
  - Efficient use of resources (time, money, patients)
  - Scientifically meaningful results
  - Statistically credible results
  - Minimal number of subjects for regulatory agencies
- How do we control false positive rate?
  - Repeated analysis of accruing data involves multiple comparisons

199

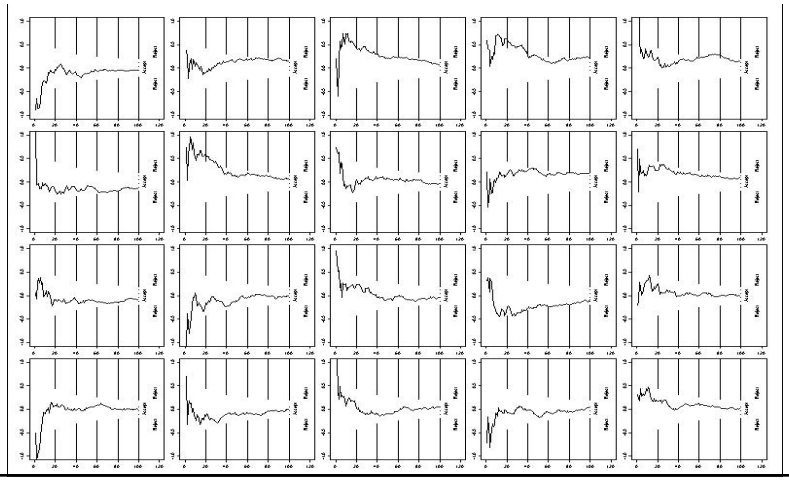
Sample Path for a Statistic



200

Fixed Sample Methods Wrong

- Simulated trials under null stop too often



201

Simulated Trials (Pocock)

- Three equally spaced level .05 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.03046			.03046
1st, 2nd	.00807	.00807		.00807
1st, 3rd	.00317		.00317	.00317
1st, 2nd, 3rd	.00868	.00868	.00868	.00868
2nd only		.01921		.01921
2nd, 3rd		.01426	.01426	.01426
3rd only			.02445	.02445
Any pattern	.05038	.05022	.05056	.10830

202

Pocock Level 0.05

- Three equally spaced level .022 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.01520			.01520
1st, 2nd	.00321	.00321		.00321
1st, 3rd	.00113		.00113	.00113
1st, 2nd, 3rd	.00280	.00280	.00280	.00280
2nd only		.01001		.01001
2nd, 3rd		.00614	.00614	.00614
3rd only			.01250	.01250
Any pattern	.02234	.02216	.02257	.05099

203

Unequally Spaced Analyses

- Level .022 analyses at 10%, 20%, 100% of data

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.01509			.01509
1st, 2nd	.00521	.00521		.00521
1st, 3rd	.00068		.00068	.00068
1st, 2nd, 3rd	.00069	.00069	.00069	.00069
2nd only		.01473		.01473
2nd, 3rd		.00165	.00165	.00165
3rd only			.01855	.01855
Any pattern	.02167	.02228	.02157	.05660

204

Varying Critical Values (O'Brien-Fleming)

- α=0.10; equally spaced using naïve .003, .036, .087

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.00082			.00082
1st, 2nd	.00036	.00036		.00036
1st, 3rd	.00037		.00037	.00037
1st, 2nd, 3rd	.00127	.00127	.00127	.00127
2nd only		.01164		.01164
2nd, 3rd		.02306	.02306	.02306
3rd only			.06223	.01855
Any pattern	.00282	.03633	.08693	.09975

205

Error Spending: Pocock 0.05

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.01520			.01520
1st, 2nd	.00321	.00321		.00321
1st, 3rd	.00113		.00113	.00113
1st, 2nd, 3rd	.00280	.00280	.00280	.00280
2nd only		.01001		.01001
2nd, 3rd		.00614	.00614	.00614
3rd only			.01250	.01250
Any pattern	.02234	.02216	.02257	.05099
Incremental error	.02234	.01615	.01250	
Cumulative error	.02234	.03849	.05099	

206

Distinctions without Differences

- Sequential sampling plans
  - Group sequential stopping rules
  - Error spending functions
  - Conditional / predictive power
  - Bayesian posterior probabilities
- Statistical treatment of hypotheses
  - Superiority / Inferiority / Futility
  - Two-sided tests / bioequivalence

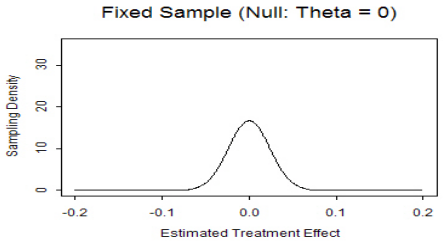
207

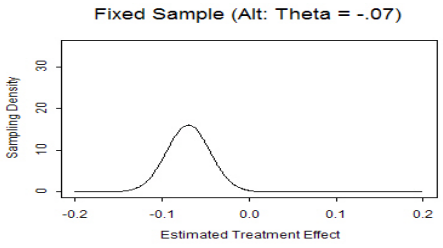
Major Statistical Issue

- Frequentist operating characteristics are based on the sampling distribution
- Stopping rules do affect the sampling distribution of the usual statistics
  - MLEs are not normally distributed
  - Z scores are not standard normal under the null
    - (1.96 is irrelevant)
  - The null distribution of fixed sample P values is not uniform
    - (They are not true P values)
- Bayesian operating characteristics are based on the sampling distribution and a prior distribution
  - Is stopping rule relevant?

208

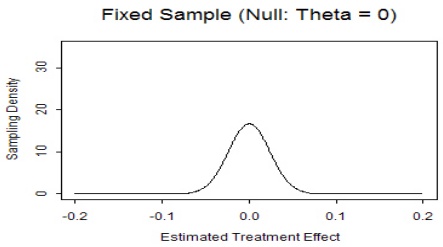
Sampling Distribution of MLE

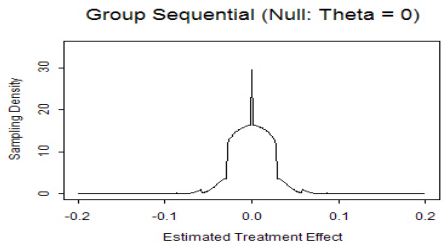


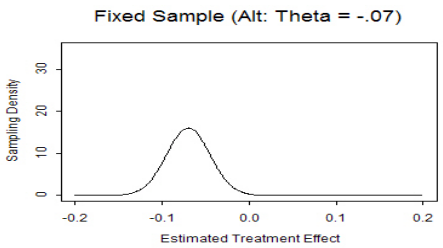


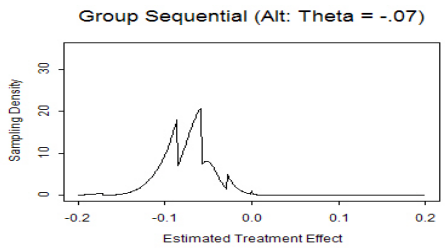
209

Sampling Distribution of MLE

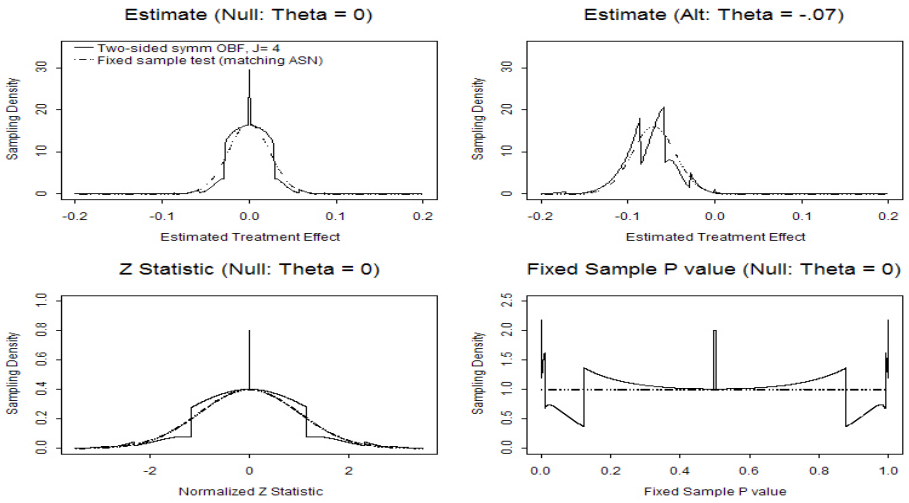








Sampling Distributions



So What?

- Demystification: All we are doing is statistics
  - Planning a study
  - Gathering data
  - Analyzing it

212

Sequential Studies

- Demystification: All we are doing is statistics
  - Planning a study
    - Added dimension of considering time required
  - Gathering data
    - Sequential sampling allows early termination
  - Analyzing it
    - The same old inferential techniques
    - The same old statistics
    - But new sampling distribution

213

Familiarity and Contempt

- For any known stopping rule, however, we can compute the correct sampling distribution with specialized software
  - Standalone programs
    - PEST (some integration with SAS)
    - EaSt
  - Within statistical packages
    - S+SeqTrial
    - SAS PROC SEQDESIGN
    - R: gsDesign and RCTdesign (SeqTrial)

214

Sampling Density

- Armitage, McPherson, and Rowe (1969)

Stopping time $M = \min\left\{j : \bar{X}_j \notin C_j\right\}$

Sequential statistic $\left(M, \bar{X}_M\right)$

$$p(m,x;\mu) = f(m,x;\mu)1_{[x \notin C_m]}$$

$$f(1,x;\mu) = \frac{1}{\sqrt{N_1}\sigma}\phi\left(\frac{N_1 x - N_1\mu}{\sqrt{N_1}\sigma}\right)$$

$$f(m,x;\mu) = \int_{C_{m-1}} \frac{1}{\sqrt{(N_m - N_{m-1})}\sigma}\phi\left(\frac{N_m x - u - N_m\mu}{\sqrt{(N_m - N_{m-1})}\sigma}\right) f(m-1,u;\mu)\,du$$

215

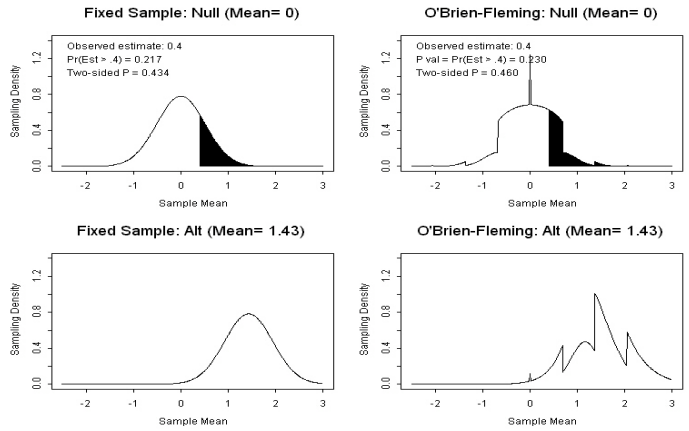
Familiarity and Contempt

- From the computed sampling distributions we then compute
  - Bias adjusted estimates
  - Correct (adjusted) confidence intervals
  - Correct (adjusted) P values
- Candidate designs can then be compared with respect to their operating characteristics

216

Example: P Value

- Null sampling density tail



217

Evaluation of Designs

- Define candidate design constraining two operating characteristics
  - Type I error, power at design alternative
  - Type I error, maximal sample size
- Evaluate other operating characteristics
  - Sample size requirements
  - Power curve
  - Inference to be reported upon termination
  - (Probability of a reversed decision)
- Modify design
- Iterate

218

Group Sequential Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined if independent of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
  - Often chosen according to some boundary shape function
    - O'Brien-Fleming, Pocock, Triangular, …
- 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

219

Stopping Boundary Scales

- Boundary scales (1:1 transformations among these)
  - Z statistic
  - P value
    - Fixed sample (so wrong)
    - Computed under sequential sampling rule (so correct)
  - Error spending function
  - Estimates
    - MLE (biased due to stopping rule)
    - Adjusted for stopping rule
  - Conditional power
    - Computed under design alternative
    - Computed under current MLE
  - Predictive power
    - Computed under flat prior (possibly improper)

220

Boundary Scales: Notation

- One sample inference about means
  - Generalizable to most other commonly used models

Probability model: $X_1, \ldots, X_N \quad iid \quad (\mu, \sigma^2)$

Null hypothesis: $H_0 : \mu = \mu_0$

Analyses after $N_1, \ldots, N_J = N$

Data at jth analysis: $x_1, \ldots, x_{N_j}$

Distributional assumptions:

in absence of a stopping rule $\bar{X}_j \sim N\left(\mu, \frac{\sigma^2}{N_j}\right)$

221

"S": Partial Sum Scale

- Uses:
  - Cumulative number of events
    - Boundary for 1 sample test of proportion
  - Convenient when computing density

$$S_j = \sum_{i=1}^{N_j} X_i$$

222

"X": MLE Scale

- Uses:
  - Natural (crude) estimate of treatment effect

$$\bar{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i = \frac{s_j}{N_j}$$

223

"Z": Normalized (Z) Statistic Scale

- Uses:
  - Commonly computed in analysis routines

$$z_j = \sqrt{N_j}\,\frac{\left[\bar{x}_j - \mu_0\right]}{\sigma}$$

224

"P": Fixed Sample P Value Scale

- Uses:
  - Commonly computed in analysis routine
  - Robust to use with other distributions for estimates of treatment effect

$$p_j = 1 - \Phi(z_j)$$

$$= 1 - \int_{-\infty}^{z_j} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$$

225

"E": Error Spending Scale

- Uses:
  - Implementation of stopping rules with flexible determination of number and timing of analyses

$$E_{dj} = \frac{1}{\alpha_u}\left[\sum_{i=1}^{j-1} \Pr\left(S_i \geq d_i, \bigcap_{k=1}^{i-1} \left(S_k \in (a_k, b_k) \cup (c_k, d_k)\right) ; \mu_d \right) + \Pr(S_j \geq s_j ; \mu_d)\right]$$

226

"B": Bayesian Posterior Scale

- Uses:
  - Bayesian inference (unaffected by stopping)
  - Posterior probability of hypotheses

Prior distribution $\mu \sim N\left(\zeta, \tau^2\right)$

$$B_j\left(\mu_*\right) = \Pr\left(\mu \geq \mu_* \mid \left(X_{1,\ldots,} X_{N_j}\right)\right)$$

$$= 1 - \Phi\left(\frac{\mu_*\left[N_j\tau^2 + \sigma^2\right] - N_j\tau^2\bar{x}_j - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right)$$

227

"C": Conditional Power Scale

- Uses:
  - Conditional power
    - Probability of significant result at final analysis conditional on data so far (and hypothesis)
  - Futility of continuing under specific hypothesis

Threshold at final analysis $t_{\bar{X}_J}$

Hypothesized value of mean μ_*

$$C_j\left(t_{\bar{X}_J}, \mu_*\right) = \Pr\left(\bar{X}_J \geq t_{\bar{X}_J} \mid \bar{X}_j; \mu = \mu_*\right)$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\bar{X}_J} - \mu_*\right] - N_j\left[\bar{x}_j - \mu_*\right]}{\sigma\sqrt{N_J - N_j}}\right)$$

228

"C": Conditional Power Scale (MLE)

- Uses:
  - Conditional power
  - Futility of continuing under specific hypothesis

Threshold at final analysis $t_{\bar{X}_J}$

Hypothesized value of mean $\mu_* = \bar{x}_j$

$$C_j\left(t_{\bar{X}_J}, \mu_*\right) = \Pr\left(\bar{X}_J \geq t_{\bar{X}_J} \mid \bar{X}_j; \mu = \bar{x}_j\right)$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\bar{X}_J} - \bar{x}_j\right]}{\sigma\sqrt{N_J - N_j}}\right)$$

229

"H": Predictive Power Scale

- Uses:
  - Futility of continuing study

Threshold at final analysis $t_{\bar{X}_J}$

Prior distribution $\mu \sim N\left(\zeta, \tau^2\right)$

$$H_j\left(t_{\bar{X}_J}\right) = \int \Pr\left(\bar{X}_j \geq t_{\bar{X}_J} \mid \bar{X}_j, \mu\right) \lambda\left(\mu \mid \bar{X}_j\right) d\mu$$

$$= 1 - \Phi\left(\frac{N_J\left[N_j\tau^2 + \sigma^2\right]\left[t_{\bar{X}_J} - \bar{x}_j\right] + \sigma^2\left[N_J - N_j\right]\left[\bar{x}_j - \zeta\right]}{\sigma\sqrt{\left[N_J - N_j\right]\left[N_J\tau^2 + \sigma^2\right]\left[N_j\tau^2 + \sigma^2\right]}}\right)$$

230

"H": Predictive Power (Flat Prior)

- Uses:
  - Futility of continuing study

Threshold at final analysis $t_{\bar{X}_J}$

Prior distribution $\mu \sim N\left(\zeta, \tau^2 \to \infty\right)$

$$H_j\left(t_{\bar{X}_J}\right) = \int \Pr\left(\bar{X}_j \geq t_{\bar{X}_J} \mid \bar{X}_j, \mu\right) \lambda\left(\mu \mid \bar{X}_j\right) d\mu$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\bar{X}_J} - \bar{x}_j\right]}{\sigma\sqrt{\frac{N_J}{N_j}\left[N_J - N_j\right]}} \right)$$

231

Specification of Designs

Error Spending Functions

Where am I going?

The default family in RCTdesign is the Unified Family that includes a wide variety of previously described families.

RCTdesign also includes a number of designs defined on the error spending scale.

It is generally unimportant which scale is used for definition of a stopping rule, so long as they are fully evaluated

In my experience, people do not understand the scientific impact of particular error spending functions very well

Error Spent at Each Analysis

- seqOC() returns stopping probabilities at each analysis
- When called under the null hypothesis, this is the error spent at each analysis

```
> seqOC(fut,theta=0)

### Asymptotic Operating Characteristics
Operating characteristics at theta= 0
ASN= 986.6778
Expected theta= 0.0092
Lower Power= 0.025

Stopping Probabilities:
                 Lower Null  Upper  Total
Analysis time 1 0.0000    0 0.1339 0.1339
Analysis time 2 0.0024    0 0.4956 0.4981
Analysis time 3 0.0092    0 0.2713 0.2805
Analysis time 4 0.0133    0 0.0742 0.0875
```

233

Error Spending Functions in RCTdesign

- No matter how a design is specified, RCTdesign returns the error spending function
  - "error.spend" component of a seqDesign object has the cumulative proportion of error spent

```
> fut$error.spend
STOPPING BOUNDARIES: Error Spending Function scale
                     Efficacy Futility
    Time 1 (N=  425)   0.0014   0.0341
    Time 2 (N=  850)   0.0993   0.2364
    Time 3 (N= 1275)   0.4684   0.5955
    Time 4 (N= 1700)   1.0000   1.0000
```

  - The efficacy boundary is the type 1 error spending function
  - The futility boundary is the type 2 error spending function for the alternative used to define that boundary
    - (Need to fully understand the stopping boundaries)

234

Error Spending Families in RCTdesign

- seqDesign() argument design.family allows specification of error spending families
  - design.family="E" uses Kim & DeMets power family
  - design.family="Hwang" uses Hwang, Shih, & DeCani family
- Authors have described approximate correspondences to OBF and Pocock boundaries within unified family

235

"O'Brien-Fleming" Design in Power Family

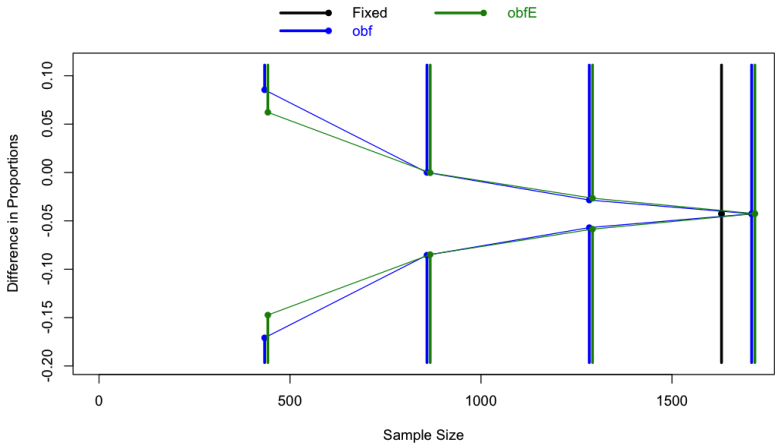
- I use P=-3.25 to approximate an OBF design
  - Others use P=-3

```
> obfE <- update(obf, design.family="E", P=-3.25)
> obfE
Call:
seqDesign(prob.model = "prop", null.hypothesis = 0.3, nbr.analyses = 4,
    sample.size = 1700, test.type = "less", power = 0.9, design.family = "E",
    P = -3.25)

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in probabilities (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >=  0.00000    (size  = 0.025)
    Alternative hypothesis : Theta <= -0.07038    (power = 0.900)

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N=  425)  -0.1474   0.0622
    Time 2 (N=  850)  -0.0848  -0.0003
    Time 3 (N= 1275)  -0.0586  -0.0266
    Time 4 (N= 1700)  -0.0426  -0.0426
```

236

Comparison of Designs

- Overlay of O'Brien-Fleming design and approximation based on power family error spending function
  - plot(obf, obfE)



237

Relative Advantages

- Which is the best scale to view a stopping rule?
  - Maximum likelihood estimate
  - Z score / fixed sample P value
  - Error spending scale
  - Stochastic curtailment
    - Conditional power
    - Predictive power

238

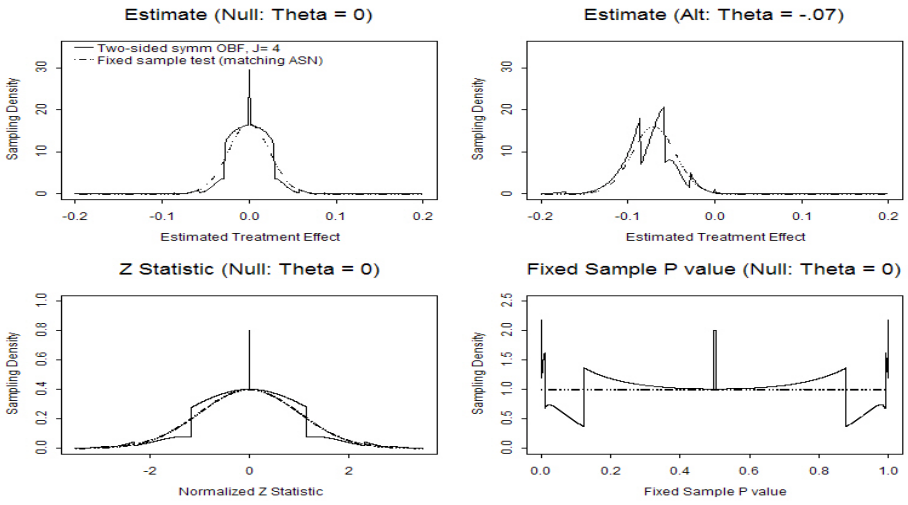
Current Relevance

- Many statisticians (unwisely?) focus on error spending scales, Z statistics, fixed sample P values when describing designs
- Some statisticians have (unwisely?) suggested the use of stochastic curtailment for
  - Defining prespecified sampling plans
  - Adaptive modification of sampling plans

239

Z scale, Fixed Sample P value

- Interpretation of Z scale and fixed sample P value is difficult

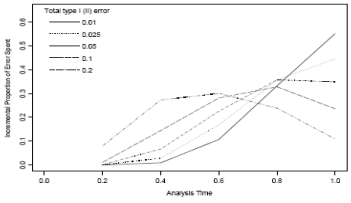


Error Spending Functions

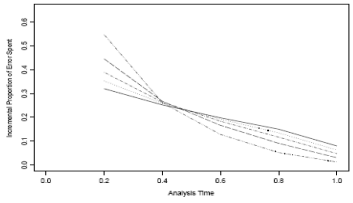
- My view: Poorly understood even by the researchers who advocate them
- There is no such thing as THE Pocock or O'Brien-Fleming error spending function
  - Depends on type I or type II error
  - Depends on number of analyses
  - Depends on spacing of analyses

241

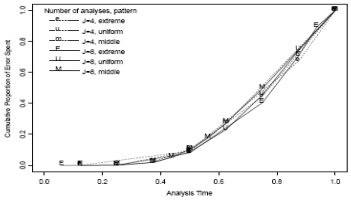
OBF, Pocock Error Spending

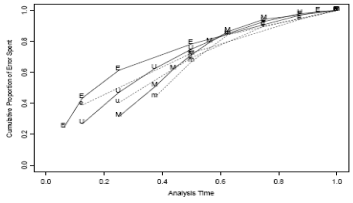


(a) O'Brien-Fleming Boundary Relationships



(b) Pocock Boundary Relationships





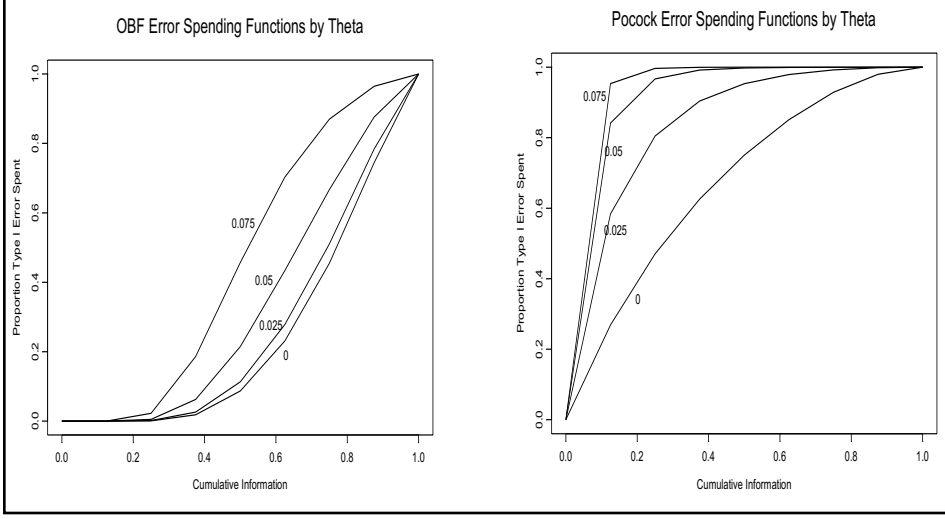
242

Function of Alternative

- Error spending functions depend on the alternative used to compute them
  - The same design has many error spending functions
- JSM 2009: Session on early stopping for harm in a noninferiority trial
  - Attempts to use error spending function approach
  - How to calibrate with functions used for lack of benefit?

243

Error Spent by Alternative



Stochastic Curtailment

- Stopping boundaries chosen based on predicting future data
- Probability of crossing final boundary
  - Frequentist: Conditional Power
    - A Bayesian prior with all mass on a single hypothesis
  - Bayesian: Predictive Power
- Users are typically poor at guessing good thresholds
  - More on this later

245

Well Understood Methods

Group Sequential Designs

Where am I going?

Borrowing terminology from the draft CDER/CBER Guidance, we first review the extent to which the "well understood" group sequential designs can be viewed as adaptive designs.

I show that almost all of the motivation for adaptation of the maximal sample size is easily addressed in the group sequential framework.

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

247

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

248

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)

249

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

250

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

251

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) (Sponsor)
    - mean (Average Sample N- ASN) (Sponsor, DMC)
    - median, quartiles
  - Stopping probabilities
    - Probability of stopping at each analysis as a function of treatment effect (Sponsor)
    - Probability of each decision at each analysis

252

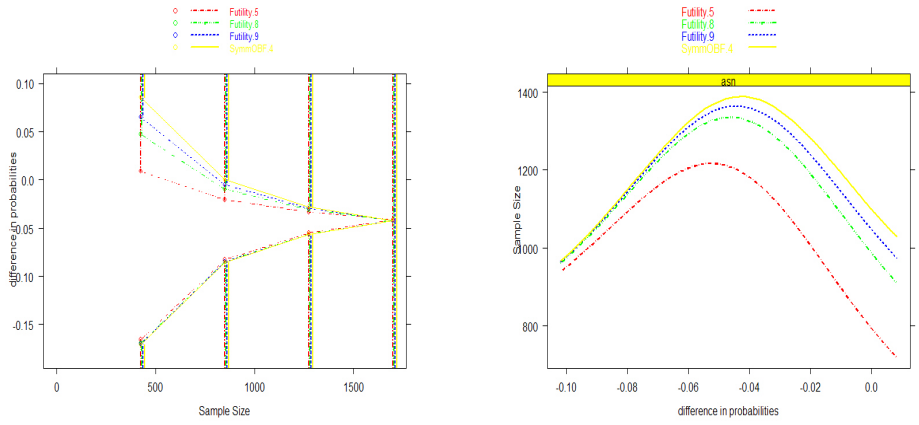
Sample Size

- 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)

253

ASN Curve

- Expected sample size as function of true effect



Evaluation: Power Curve

- Probability of rejecting null for arbitrary alternatives
  - Level of significance (power under null) (Regulatory)
  - Power for specified alternative
  - Alternative rejected by design (Scientists)
    - Alternative for which study has high power
      - Interpretation of negative studies

255

Evaluation: Boundaries

- Decision boundary at each analysis: Value of test statistic leading to early stopping
  - On the scale of estimated treatment effect (DMC, Statisticians)
    - Inform DMC of precision
    - Assess ethics (DMC)
      - May have prior belief of unacceptable levels
    - Assess clinical importance (Marketing)
  - On the Z or fixed sample P value scales (Often asked for, but of questionable relevance)

256

Evaluation: Inference

- Inference on the boundary at each analysis
  - Frequentist
    - Adjusted point estimates
    - Adjusted confidence intervals
    - Adjusted P values

    (Scientists, Statisticians, Regulatory)
  - Bayesian
    - Posterior mean of parameter distribution
    - Credible intervals
    - Posterior probability of hypotheses
    - Sensitivity to prior distributions

    (Scientists, Statisticians, Regulatory)

257

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

    (Scientists, Sponsor)
  - Conditional power
    - Assume specific hypotheses
    - Assume current best estimate

    (Often asked for, but of questionable relevance)
  - Predictive power
    - Assume Bayesian prior distribution

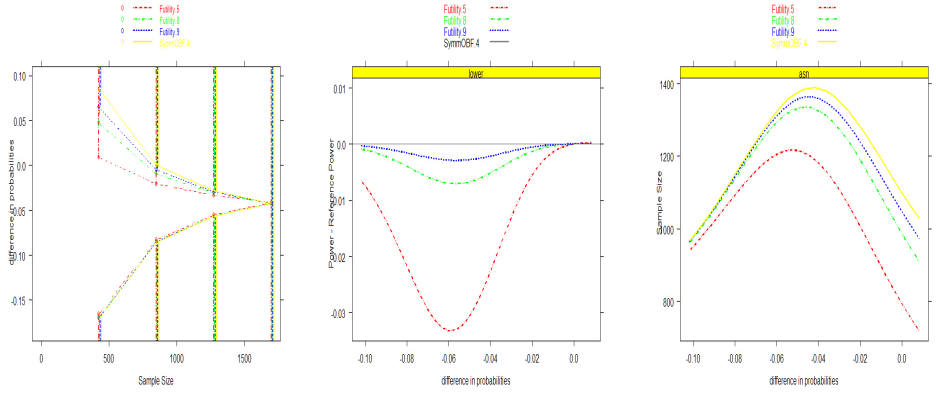
258

Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power
- Boundaries

Loss of Power

Avg Sample Size



Evaluation: Marketable Results

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

(Marketing, Clinicians)

  - Predictive probabilities at each analysis

250

Exploring Group Sequential Designs

- Candidate designs
  - J = 2 equal spaced analyses; O'Brien-Fleming efficacy boundary
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 equal spaced analyses; OBF efficacy, futility boundaries
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 equal spaced analyses; OBF efficacy, more efficient futility
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 4 equal spaced analyses; OBF efficacy, more efficient futlility
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 optimally spaced analyses; optimal symmetric boundaries
    - Maintain power under alternative (inflate N_J, but optimize ASN)

261

Exploring Group Sequential Designs

- Examining operating characteristics
  - Stopping boundaries
    - Z scale
    - Conditional power under hypothesized effects
    - Conditional power under current MLE
    - Predictive power under flat prior
  - Estimates and inference
    - MLE (Bias adjusted estimates suppressed for space)
    - 95% CI properly adjusted for stopping rule
    - P value properly adjusted for stopping rule
  - Power at specified alternatives
  - Sample size distribution (as function of true effect)
    - Maximal sample size
    - Average sample size

262

O'Brien-Fleming Efficacy: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	--	--	--	--	2.796	0.500	0.997	0.976
1.0	1.977	--	--	--	1.977	--	--	--

263

O'Brien-Fleming Efficacy: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	--	--	--	3.955	(1.16, 5.72)	0.003
1.0	1.977	(0.00, 3.93)	0.025	1.977	(0.00, 3.93)	0.025

264

O'Brien-Fleming Efficacy: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.999
1.96	0.496	0.960
2.80	0.797	0.896
3.24	0.898	0.847
3.92	0.974	0.755

265

O'Brien-Fleming Efficacy: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.50	--	--	--	3.943	(1.16, 5.70)	0.003
1.01	1.977	(0.00, 3.92)	0.025	1.977	(0.00, 3.92)	0.025

266

O'Brien-Fleming Efficacy: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	1.005
1.96	0.499	0.966
2.80	0.799	0.901
3.24	0.900	0.851
3.92	0.975	0.758

267

Take Home Messages GSD1

- Introduction of a very conservative efficacy boundary
  - Minimal effect on power even if do not increase max N
  - Minimal increase in max N needed to maintain power
- Ease and importance of evaluating a stopping rule
  - Even before we start the study, we can consider
    - Thresholds for early stopping in terms of estimated effects
    - Inference corresponding to stopping points
    - Conditional and predictive power under various hypotheses
  - We can judge a stopping rule by comparing it to a fixed sample test and look at the tradeoffs between
    - Increase in maximal sample size
    - Decrease in average sample size
    - Changes in unconditional power

268

O'Brien-Fleming Symmetric: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	0.000	0.500	0.003	0.024	2.796	0.500	0.997	0.976
1.0	1.977	--	--	--	1.977	--	--	--

269

O'Brien-Fleming Symmetric: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	0.000	(-1.76, 2.80)	0.375	3.945	(1.15, 5.71)	0.003
1.0	1.973	(0.00, 3.94)	0.025	1.973	(0.00, 3.94)	0.025

270

O'Brien-Fleming Symmetric: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.749
1.96	0.495	0.919
2.80	0.795	0.883
3.24	0.897	0.840
3.92	0.974	0.752

271

O'Brien-Fleming Symmetric: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.51	0.00	(-1.75, 2.78)	0.375	3.920	(1.14, 5.67)	0.003
1.01	1.960	(0.00, 3.92)	0.025	1.960	(0.00, 3.92)	0.025

272

O'Brien-Fleming Symmetric: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.758
1.96	0.500	0.930
2.80	0.800	0.893
3.24	0.900	0.848
3.92	0.975	0.758

273

Take Home Messages GSD2

- Introduction of a very conservative futility boundary
  - Again, minimal effects on power and/or max N
  - Dramatic improvement in ASN under the null
  - Conditional and predictive power thresholds are surprising
    - CP_{alt} = 0.50 for the extremely conservative OBF boundary
      - But the CI has already eliminated 3.92 with high confidence
    - CP_{est} = 0.003 and PP_{flat} = 0.024 are both very low thresholds

274

O'Brien-Fleming & Futility: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	0.331	0.644	0.017	0.068	2.776	0.500	0.997	0.975
1.0	1.963	--	--	--	1.963	--	--	--

275

O'Brien-Fleming & Futility: J = 2, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	0.468	(-1.30, 3.27)	0.228	3.925	(1.13, 5.69)	0.003
1.0	1.963	(0.00, 3.98)	0.025	1.963	(0.00, 3.98)	0.025

276

O'Brien-Fleming & Futility: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.684
1.96	0.492	0.886
2.80	0.791	0.869
3.24	0.893	0.830
3.92	0.972	0.747

277

O'Brien-Fleming & Futility: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.52	0.461	(-1.28, 3.22)	0.228	3.867	(1.11, 5.61)	0.003
1.03	1.934	(0.00, 3.92)	0.025	1.934	(0.00, 3.92)	0.025

278

O'Brien-Fleming & Futility: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.705
1.96	0.504	0.914
2.80	0.803	0.892
3.24	0.901	0.850
3.92	0.975	0.762

279

Take Home Messages GSD3

- More aggressive futility boundary better addresses ethical issues associated with ineffective drugs
  - I often find that sponsors are willing to accept this futility bound without increasing the sample size
  - But the minimal increase in maximal sample size would seem more appropriate to me

280

O'Brien-Fleming & Futility: J = 4

- Introduce four evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.25	-1.108	0.719	0.000	0.008	3.976	0.500	0.999	0.999
0.50	0.321	0.648	0.015	0.063	2.811	0.500	0.997	0.977
0.75	1.258	0.592	0.142	0.177	2.295	0.500	0.907	0.874
1.00	1.988	--	--	--	1.988	--	--	--

281

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.25	-2.216	(-4.71, 1.74)	0.846	7.951	(4.00, 10.5)	0.000
0.50	0.454	(-1.60, 3.31)	0.263	3.976	(1.14, 6.04)	0.003
0.75	1.452	(-0.36, 3.85)	0.053	2.650	(0.30, 4.48)	0.013
1.00	1.988	(0.00, 4.06)	0.025	1.988	(0.00, 4.06)	0.025

282

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.580
1.96	0.478	0.783
2.80	0.776	0.761
3.24	0.882	0.723
3.92	0.966	0.650

283

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.27	-2.141	(-4.55, 1.68)	0.846	7.682	(3.86, 10.1)	0.000
0.54	0.439	(-1.55, 3.20)	0.263	3.841	(1.10, 5.84)	0.003
0.80	1.403	(-0.34, 3.72)	0.053	2.561	(0.29, 4.33)	0.013
1.07	1.920	(0.00, 3.92)	0.025	1.920	(0.00, 3.92)	0.025

284

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.622
1.96	0.504	0.840
2.80	0.803	0.808
3.24	0.902	0.762
3.92	0.975	0.680

285

Take Home Messages GSD4

- Effect of adding more analyses
  - Greater loss of power if maximal sample size not increased
  - Greater increase in maximal sample size if power maintained
  - But, improvement in average efficiency
- Can also use this example for guidance in how to judge thresholds for conditional and predictive power
  - The same threshold should not be used at all analyses
  - It is not, however, clear what threshold should be used
    - I look at tradeoffs between average efficiency and power
  - We can look at optimal (on average) designs for more guidance

286

Efficient: J = 2

- Introduce two optimally spaced analyses to minimize ASN
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.43	0.573	0.818	0.049	0.141	2.776	0.182	0.951	0.859
1.00	2.129	--	--	--	2.129	--	--	--

287

Efficient: J = 2, Power

- Introduce two optimally spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.50	0.808	(-0.82, 3.58)	0.129	3.112	(0.34, 4.74)	0.014
1.18	1.960	(0.00, 3.92)	0.025	1.960	(0.00, 3.92)	0.025

288

Efficient: J = 2, Power

- Introduce two optimally spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.685
1.96	0.500	0.900
2.80	0.805	0.847
3.24	0.904	0.788
3.92	0.975	0.685

289

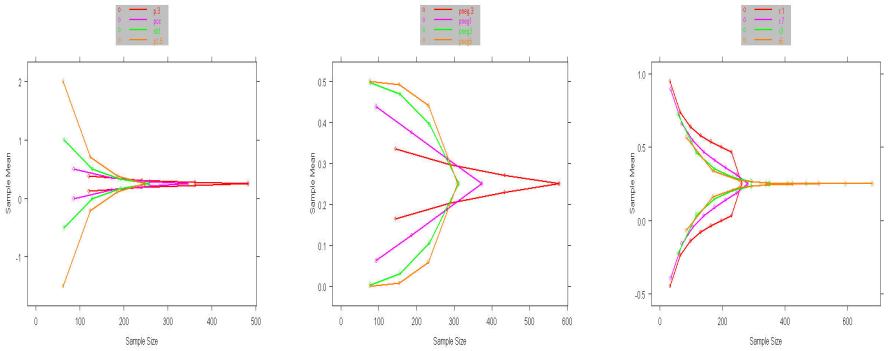
Take Home Messages GSD5

- Optimal spacing of analyses not quite equal information
- Optimal early conservatism close to a Pocock design
  - In unified family, OBF has P=1, Pocock has P= 0.5
  - Optimal P= .54
- With two analyses, increase maximal N by 18% over fixed sample
  - ASN decreases by about one third
- Again, the thresholds to use for conditional or predictive power are not at all clear
- Search for best designs should include many candidates
  - Examine many operating characteristics

290

Spectrum of Boundary Shapes

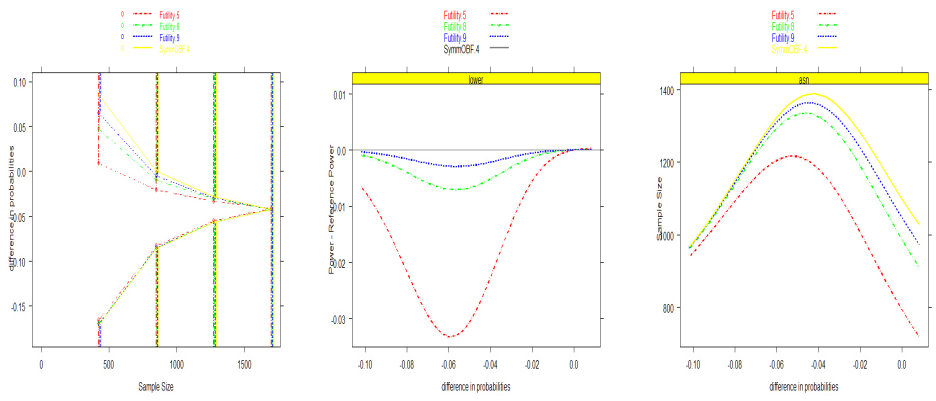
- All of the rules depicted have the same type I error and power to detect the design alternative



Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power

Boundaries | Loss of Power | Avg Sample Size



Delayed Measurement of Outcome

- Longitudinal studies
  - Measurement might be 6 months – 2 years after randomization
  - Interim analyses on variable lengths of follow-up
    - Use of partial data can improve efficiency (Kittelson, et al.)
- Time to event studies
  - Statistical information proportional to number of events
  - Calendar time requirements depend on number accrued and length of follow-up
- In either case: Interim analyses may occur after accrual completed
    - Group ethics of identifying beneficial treatments faster
    - Savings in calendar time costs, rather than per patient costs

293

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

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

295

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

296

21 CFR 50.24

- Exception to informed consent 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

297

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

298

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

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

299

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

300

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

301

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

302

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	

303

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

304

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

305

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)

306

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)

307

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

308

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

309

Bottom Line

You better think (think)

about what you're

trying to do…

-Aretha Franklin, "Think"

310

Case Study

A RCT of an Antibody to Endotoxin

in the Treatment of Gram Negative Sepsis

The Enemy

"Let's start at the very beginning, a very good place to start…"

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

312

First

- Where do we want to be?
  - Describe some innovative experiment?
  - Find a use for some proprietary drug / biologic / device?
    - "Obtain a significant p value"
  - Find a new treatment that improves health of some individuals
    - "Efficacy" seems to benefit some people somehow
  - Find a new treatment that improves health of the population
    - "Effectiveness" requires proper use of a treatment that modifies an important clinical endpoint

313

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

314

Evidence Based Medicine: PICO

- Population (patient)
  - Disease
  - Restrictions due to concomitant disease, etc.
- Intervention
  - Treatment or treatment strategy
- Comparator
  - Other alternatives being considered
- Outcome
  - Clinical vs surrogate; timeframe; method of measurement

315

Scientific Hypotheses: Background

- Critically ill patients often get overwhelming bacterial infection (sepsis), after which mortality is high
- Gram negative sepsis often characterized by production of endotoxin, which is thought to be the cause of much of the ill effects of gram negative sepsis
- Hypothesize that administering antibody to endotoxin may decrease morbidity and mortality
- Two previous randomized clinical trials showed slight benefit with suggestion of differences in benefit within subgroups
- No safety concerns

316

Indication

- Disease: Gram negative sepsis
  - Proven vs suspected
  - Gram stain, culture, abdominal injury
- Population: ICU patients
- Treatment
  - Single administration of antibody to endotoxin within 24 hours of diagnosis of sepsis
  - Reductions in dose not applicable
  - Ancillary treatments unrestricted
- Outcome?

317

Clinical Outcomes

- Goals
  - Primary: Increase survival
    - Long term (always best)
    - Short term (many other disease processes may intervene)
  - Secondary: Decrease morbidity
- Relevant logistical considerations
  - Multicenter clinical trial
    - Long term follow-up difficult in trauma centers
    - Data management is complicated

318

Candidate Clinical Outcomes

- Possible primary endpoints
  - Time to death
    - Heavy early censoring due to constraints of trauma centers
    - May emphasize clinically meaningless short term improvements
  - Mortality rate at fixed point in time
    - Allows choice of scientifically relevant time frame
      - single administration; short half life
    - Allows choice of clinically relevant time frame
      - avoid sensitivity to fleeting short term improvements
  - Time alive out of ICU during fixed period of time
    - Incorporates morbidity endpoints
    - May be sensitive to clinically meaningless improvements depending upon time frame chosen

319

Trial Design Strategy

- Selection of clinical trial design is iterative, involving scientists, statisticians, management, and regulators
  - Focus on measures with scientific meaning
  - Search through extensive space of designs
  - Compare designs with respect to variety of operating characteristics

320

Primary Endpoint: Statistical

- For a specific clinical endpoint, we still have to summarize its distribution
- Consider (in order or importance)
  - The most relevant summary measure of the distribution of the primary endpoint
    - Based on a loss function?
    - Mean, median, geometric mean, …
  - The summary measurement the treatment is most likely to affect
  - The summary measure that can be that can be assessed most accurately and precisely

321

Statistical Design

- Steps
  - Defining the probability model
    - Defining the comparison group, primary endpoint, analysis model
  - Defining the statistical hypotheses
    - Null, alternative
  - Defining the statistical criteria for evidence
    - Type I error, power
  - Determining the sample size
    - At each analysis, and maximal sample size
  - Evaluating the operating characteristics
  - Planning for monitoring
    - Updating stopping boundaries according to actual conditions
  - Plans for analysis and reporting results
    - Inference adjusted for sequential sampling plan

322

Chosen Endpoint / Study Structure

- Primary endpoint: 28 day all cause mortality
- Comparison group
  - Randomized, double blind, placebo controlled
  - 1:1 randomization
- Hypotheses
  - Treatment effect of 7% absolute increase in mortality
    - On placebo: 30% 28 day mortality
    - On antibody: 23% 28 day mortality
- Burden of proof: Registrational trial
  - Type I error: one-sided 0.025
  - Statistical power: high

323

Probability Models

- Common summary measures (contrasts) used in clinical trials
  - Means (difference)
    - 1, 2, k arms; regression
  - Geometric means (lognormal medians) (ratio)
    - 1, 2, k arms; regression
  - Proportions (difference)
    - 1, 2 arms
  - Odds (ratio)
    - 1, 2 arms; regression
  - Rates (ratio)
    - 1, 2 arms; regression
  - Hazard (ratio)
    - 2 arms

324

Creating a RCT Design : Model

```
• seqDesign (
   prob.model = "proportions",     (vs several others)
   arms = 2,                       (default)
   ratio = 1,                      (default)
   variance = "alternative",       (default vs "null","intermediate", number)
   (many others)
   )
```

325

Creating a RCT Design : Hypotheses

```
• seqDesign (
   null.hypothesis = 0.30,
   alt.hypothesis = 0.23,          (vs unspecified)
   variance = "alternative",       (vs "null","intermediate" or number)
   (many others)
   )
```

326

Creating a RCT Design : Fixed Sample

- seqDesign (
  nbr.analyses = 1, (default)
  test.type = "less", (vs "greater","two.sided")
  size = 0.025, (default)
  (many others)
)

327

Creating a RCT Design : Statistical Task

- seqDesign (
  alt.hypothesis = 0.23, (vs unspecified)
  power = "calculate", (vs a number, say, 0.90)
  sample.size = 1700, (vs unspecified)
  (many others)
)

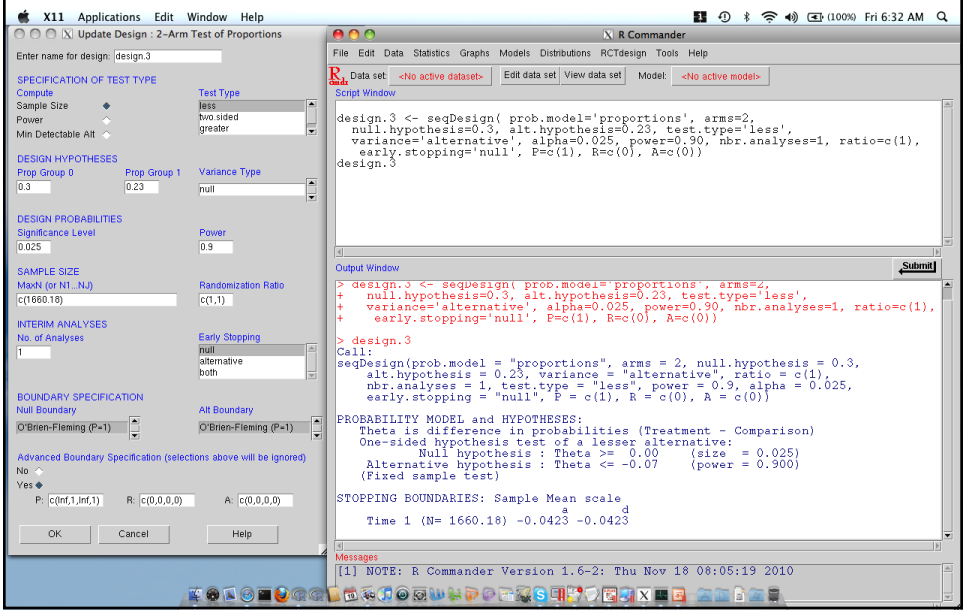
328

Creating a RCT Design : Output

- `seqDesign (`
  `display.scale = "X",` (default vs any boundary scale)
  `(many others)`
  `)`

329

Creating a RCT Design : GUI via Rcmdr (soon)



Evaluation of Fixed Sample Designs

- Clinical trial design is most often iterative
  - Specify an initial design
  - Evaluate operating characteristics
  - Modify the design
  - Iterate

331

Fixed Sample Operating Characteristics

- Level of Significance (often pre-specified)
- Sample size requirements
  - Scientific / regulatory credibility; Feasibility
- Power Curve
  - Under null (type I error); design alternative
- Decision Boundary
  - Clinical significance
- Frequentist / Bayesian inference on the Boundary
  - Clinical significance of hypotheses discriminated
  - Sensitivity to Bayesian priors

332

Evaluation in RCTdesign

- Printing
  - Design: `changeSeqScale (x, scale)`
  - Operating characteristics: `seqOC (x, theta, power)`
  - Inference at boundary: `seqInference (x, theta, power)`
  - Everything: `seqEvaluate (x, theta, power)`
- Plotting
  - `plot (x); seqPlotBoundary (x)`
  - `seqPlotASN (x)`
  - `seqPlotPower (x)`
  - `seqPlotStopProb (x)`
  - `seqPlotInference (x)`

333

RCTdesign: Creation of Design

```
> fxd90 <- seqDesign("proportions",null=0.30,alt=0.23,power=0.90)
> fxd90
Call:
seqDesign(prob.model = "proportions", null.hypothesis = 0.3,
    alt.hypothesis = 0.23, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in probabilities (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >=  0.00    (size  = 0.025)
    Alternative hypothesis : Theta <= -0.07    (power = 0.900)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                       Efficacy Futility
    Time 1 (N= 1660.17)  -0.0423  -0.0423
```

334

RCTdesign: evalGST

```
> seqEvaluate(fxd90)

Stopping Boundaries:
    Anlys SampSize CrudeEst     Z  FxdP Hnoninf
Eff     1 1660.173  -0.0423 -1.96 0.025      NA
Fut     1 1660.173  -0.0423 -1.96 0.025      NA

ASN and Cumulative Stopping Probability at Each Analysis
 Power TrueEff AvgSampSiz CumStpPrb 1
 0.975 -0.0847   1660.173           1
 0.950 -0.0778   1660.173           1
 0.900 -0.0700   1660.173           1
 0.800 -0.0605   1660.173           1

Inference at the Stopping Boundaries
    Anlys SampSize     BAM  CIlo.m CIhi.m Pval.m
Eff     1 1660.173 -0.0423 -0.0847      0  0.025
Fut     1 1660.173 -0.0423 -0.0847      0  0.025
```

335

Alternative Fixed Sample Designs

- Alternative sample sizes
- Sensitivity to assumptions about variability
  - Comparison of means, geometric means
    - Need to estimate variability of observations
  - Comparison of proportions, odds, rates
    - Need to estimate event rate
  - Comparison of hazards
    - Need to estimate number of subjects and time required to observe required number of events

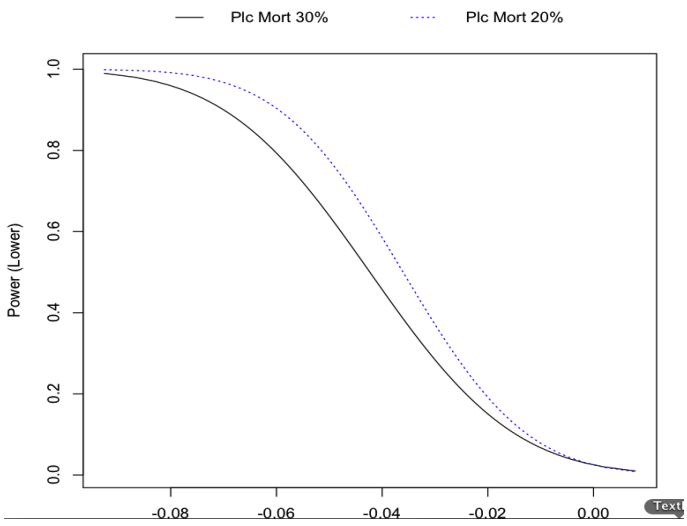
336

What if Lower Event Rate?

```
> afxd90 <- update(fxd90,null=0.20,alt=,test.type="less",
+   sample.size=sampleSize(fxd90))
> seqEvaluate(afxd90)

Stopping Boundaries:
    Anlys SampSize CrudeEst     Z  FxdP Hnoninf
Eff     1 1660.173   -0.036 -1.96 0.025      NA
Fut     1 1660.173   -0.036 -1.96 0.025      NA

ASN and Cumulative Stopping Probability at Each Analysis u
 Power TrueEff AvgSampSiz CumStpPrb 1
 0.975 -0.0721   1660.173           1
 0.950 -0.0663   1660.173           1
 0.900 -0.0596   1660.173           1
 0.800 -0.0515   1660.173           1

Inference at the Stopping Boundaries
    Anlys SampSize    BAM  CIlo.m CIhi.m Pval.m
Eff     1 1660.173 -0.036 -0.0721      0  0.025
Fut     1 1660.173 -0.036 -0.0721      0  0.025
```

337

What if Lower Event Rate?



338

Group Sequential Stopping Rules

Stopping Rules

- Basic Strategy
  - Find stopping boundaries at each analysis such that desired operating characteristics (e.g., type I and type II statistical errors) are attained
- Issues
  - Conditions under which the trial might be stopped early
  - When to perform analyses
  - Test statistic to use
  - Relative position of boundaries at successive analyses
  - Desired operating characteristics

340

Boundary Scales

- Stopping boundaries can be defined on a wide variety of scales
  - Sum of observations
  - Point estimate of treatment effect
  - Normalized (Z) statistic
  - Fixed sample P value
  - Error spending function
  - Conditional probability
  - Predictive probability
  - Bayesian posterior probability

341

Boundary Shape Function

- Boundary shape functions
  - Π_j measures the proportion of information accrued at the j-th analysis
    - often $\Pi_j = N_j / N_J$
  - Boundary shape function $f(\Pi_j)$ is a monotonic function used to relate the dependence of boundaries at successive analyses on the information accrued to the study at that analysis

342

Unified Design Family

- Stopping Boundaries for Sample Mean Statistic:
  - $a_j = \mu_a - f_a(\Pi_j)$
  - $b_j = \mu_b + f_b(\Pi_j)$
  - $c_j = \mu_c - f_c(\Pi_j)$
  - $d_j = \mu_d + f_d(\Pi_j)$

3-43

Unified Design Family

- Parameterization of boundary shape functions

$$f_*(\Pi_j) = [A_* + \Pi_j^{-P_*}(1-\Pi_j)^{R_*}] \times G_*$$

  - Distinct parameters possible for each boundary
  - Parameters A*, P*, R* typically chosen by user
  - Critical value G* usually calculated from search

3-44

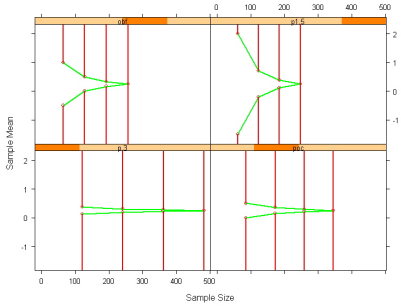
Unified Design Family

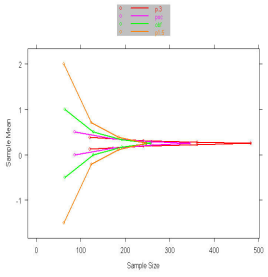
- Choice of P parameter
  - $P \geq 0$:
    - Larger positive values of P make early stopping more difficult (impossible when P infinite)
    - When A=R=0, $0.5 < P < 1$ corresponds to power family parameter (Δ) in Wang & Tsiatis (1987): $P = 1 - \Delta$
    - Reasonable range of values: $0 < P < 2.5$
    - P=0 with A=R=0 possible for some (not all) boundaries, but not particularly useful

345

Unified Design Family

- Effect of varying P>0 (when A=0, R=0)
  - Higher P leads to early conservatism
  - P > 0 has infinite boundaries when N=0





346

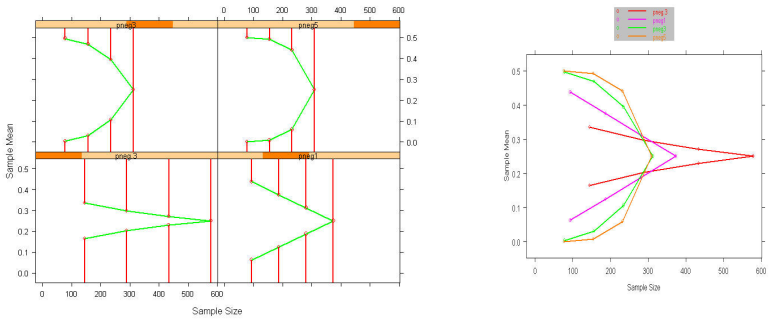
Unified Design Family

- Choice of P parameter
  - P < 0:
    - Must have R = 0 and (typically) A < 0
    - More negative values of P make early stopping more difficult

347

Unified Design Family

- Effect of varying P<0 (when A=2, R=0)
  - More negative P leads to early conservatism
  - P < 0 has finite boundaries when N=0



348

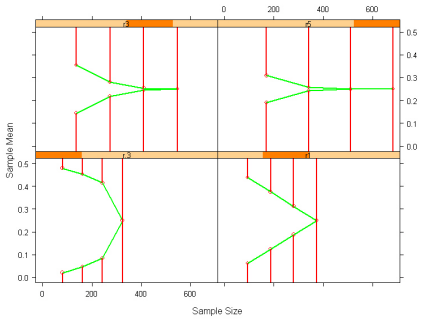
Unified Design Family

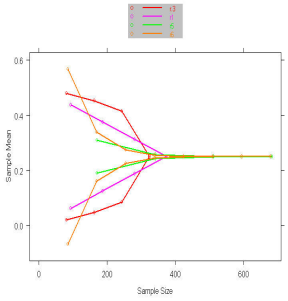
- Choice of R parameter
  - R > 0:
    - Larger positive values of R make early stopping easier
    - When R>0 and P=0, typically need A>0
    - Reasonable range of values: 0.1 < R < 20
    - R < 1 is convex outward
    - R > 1 is convex inward
    - When R>0 and P>0, can get change in convexity of boundaries

349

Unified Design Family

- Effect of varying R (when A=1, P=0)
  - R < 1 leads to convex outward
  - R > 1 leads to convex inward

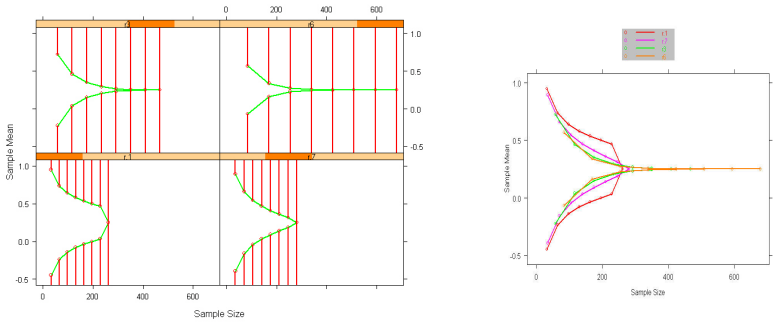




350

Unified Design Family

- Effect of varying R (when A=1, P=0.5)
  - With P > 0, boundaries infinite when N=0
  - R < 1 and P > 0 has change in convexity



351

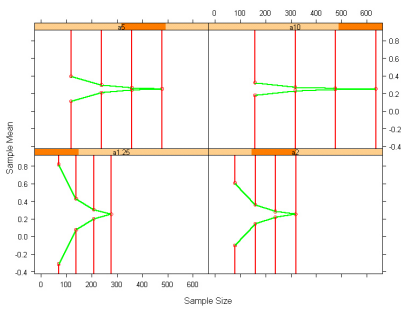
Unified Design Family

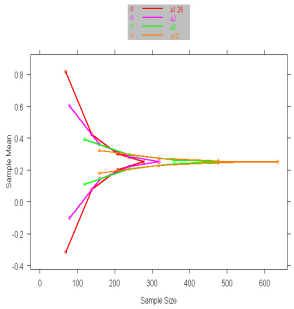
- Choice of A parameter
  - Lower absolute values of A makes it harder to stop at early analyses
  - Valid choices of A depend upon choices of P and R
  - Useful ranges for A
    - $P \geq 0, R \geq 0$: $0.2 \leq A \leq 15$
    - $P \leq 0, R = 0$: $-15 \leq A \leq -1.25$

352

Unified Design Family

- Effect of varying A (when P=0, R=1.2)
  - Values of A close to 0 make it harder to stop early
  - Higher absolute value of A -> flatter boundaries





353

Unified Design Family

- Parameterization of boundary shape function includes many previously described approaches
  - Wang & Tsiatis Boundary Shape Functions:
    - $A_* = 0$, $R_* = 0$, $P_* > 0$
    - P_* measures early conservatism
      $P_* = 0.5$ Pocock (1977)
      $P_* = 1.0$ O'Brien-Fleming (1979)
    - ($P_* = \infty$ precludes early stopping)

354

Unified Design Family

- Parameterization of boundary shape function includes many previously described approaches
  - Triangular Test Boundary Shape Functions (Whitehead)
    - $A^* = 1$, $R^* = 0$, $P^* = 1$
  - Sequential Conditional Probability Ratio Test (Xiong):
    - $R^* = 0.5$, $P^* = 0.5$

355

Creating a RCT Design : Sequential

```
seqDesign (
  nbr.analyses = 4,                 (default is 1)
  sample.size = (1:4)/4*1700,       (default spacing)
  design.family = "X",              (default)
  early.stopping = "both",          (default vs "null", "alternative")
  P = c(1,1),                       (default corresponds to OBF)
  A = c(0,0),                       (default)
  R = c(0,0),                       (default)
  minimum.constraint = ,            (default)
  maximum.constraint = ,            (default)
  exact.constraint = ,              (default)
  (many others)
  )
```

356

Evaluation of Designs

- Process of choosing a trial design
  - Define candidate design
  - Evaluate operating characteristics
  - Modify design
  - Iterate

357

Evaluation of Designs: Fixed Sample

- Operating characteristics for fixed sample studies
  - Level of Significance (often pre-specified)
  - Sample size requirements
  - Power Curve
  - Decision Boundary
  - Frequentist inference on the Boundary
  - Bayesian posterior probabilities

358

Evaluation of Designs: Sequential

- Additional operating characteristics for group sequential studies
  - Probability distribution for sample size
  - Stopping probabilities
  - Boundaries at each analysis
  - Frequentist inference at each analysis
  - Bayesian inference at each analysis
  - Futility measures at each analysis

359

Evaluation of Designs

- Sample size requirements
  - Number of subjects needed is a random variable
  - Quantify summary measures of sample size distribution
    - maximum (feasibility of accrual)
    - mean (Average Sample N- ASN)
    - median, quartiles
  - (Particularly consider tradeoffs between power and sample size distribution)

360

Evaluation of Designs

- Stopping probabilities
  - Consider probability of stopping at each analysis for arbitrary alternatives
  - Consider probability of each decision (for null or alternative) at each analysis

361

Evaluation of Designs

- Power curve
  - Probability of rejecting null for arbitrary alternatives
    - Power under null: level of significance
    - Power for specified alternative
  - Alternative rejected by design
    - Alternative for which study has high power
  - RCTdesign defines
    - Power curves for upper and lower boundaries
    - Alternatives having specified power for each boundary

362

Evaluation of Designs

- Decision boundary at each analysis
  - Value of test statistic leading to rejection of null
    - Variety of boundary scales possible
  - Often has meaning for applied researchers (especially on scale of estimated treatment effect)
    - Estimated treatment effects may be viewed as unacceptable for ethical reasons based on prior notions
    - Estimated treatment effect may be of little interest due to lack of clinical importance or futility of marketing

363

Evaluation of Designs

- Frequentist inference on the boundary at each analysis
  - Consider P values, confidence intervals when observation corresponds to decision boundary at each analysis
  - Ensure desirable precision for negative studies
    - Confidence interval identifies hypotheses not rejected by analysis
    - Have all scientifically meaningful hypotheses been rejected?

364

Evaluation of Designs

- Bayesian posterior probabilities at each analysis
  - Examine the degree to which the frequentist inference leads to sensible decisions under a range of prior distributions for the treatment effect
    - Posterior probability of hypotheses
  - Bayesian estimates of treatment effect
    - Median (mode) of posterior distribution
    - Credible interval (quantiles of posterior distribution

365

Evaluation of Designs

- Futility measures
  - Consider the probability that a different decision would result if trial continued
  - Can be based on particular hypotheses, current best estimate, or predictive probabilities
  - (Perhaps best measure of futility is whether the stopping rule has changed the power curve substantially)

366

O'Brien-Fleming Symmetric

```
> obf <- seqDesign("prop",null=0.30,test.type="less",
+   sample.size=1700,nbr.analyses=4,P=1,power=0.9)
> seqEvaluate(obf)

Stopping Boundaries:
    Anlys SampSize CrudeEst       Z   FxdP Hnoninf
Eff     1      425  -0.1709 -4.0065 0.0000  0.9997
Eff     2      850  -0.0855 -2.8330 0.0023  0.9774
Eff     3     1275  -0.0570 -2.3131 0.0104  0.8763
Eff     4     1700  -0.0427 -2.0032 0.0226      NA
Fut     1      425   0.0855  2.0032 0.9774  0.0003
Fut     2      850   0.0000  0.0000 0.5000  0.0226
Fut     3     1275  -0.0285 -1.1566 0.1237  0.1237
Fut     4     1700  -0.0427 -2.0032 0.0226      NA

ASN and Cumulative Stopping Probability at Each Analysis under Alternatives
 Power TrueEff AvgSampSiz CumStpPrb 1 CumStpPrb 2 CumStpPrb 3 CumStpPrb 4
 0.975 -0.0855   1098.676      0.0226      0.5026      0.8897           1
 0.950 -0.0786   1162.491      0.0153      0.4144      0.8351           1
 0.900 -0.0706   1236.314      0.0095      0.3213      0.7603           1
 0.800 -0.0610   1315.958      0.0053      0.2309      0.6675           1

Inference at the Stopping Boundaries
    Anlys SampSize     BAM  CIlo.m  CIhi.m Pval.m
Eff     1      425 -0.1624 -0.2242 -0.0866 0.0000
Eff     2      850 -0.0795 -0.1296 -0.0250 0.0024
Eff     3     1275 -0.0543 -0.0957 -0.0068 0.0123
Eff     4     1700 -0.0427 -0.0855  0.0000 0.0250
Fut     1      425  0.0770  0.0011  0.1387 0.9765
Fut     2      850 -0.0060 -0.0605  0.0442 0.4011
Fut     3     1275 -0.0312 -0.0786  0.0102 0.0672
```

357

O'Brien-Fleming Efficacy with Futility

```
> fut <- update(obf,P=c(1,0.8))
> seqEvaluate(fut)

Stopping Boundaries:
    Anlys SampSize CrudeEst       Z   FxdP Hnoninf
Eff     1      425  -0.1695 -3.9756 0.0000  0.9997
Eff     2      850  -0.0848 -2.8112 0.0025  0.9766
Eff     3     1275  -0.0565 -2.2953 0.0109  0.8744
Eff     4     1700  -0.0424 -1.9878 0.0234      NA
Fut     1      425   0.0473  1.1082 0.8661  0.0076
Fut     2      850  -0.0097 -0.3211 0.3741  0.0625
Fut     3     1275  -0.0310 -1.2577 0.1043  0.1768
Fut     4     1700  -0.0424 -1.9878 0.0234      NA

ASN and Cumulative Stopping Probability at Each Analysis under Alternatives
 Power TrueEff AvgSampSiz CumStpPrb 1 CumStpPrb 2 CumStpPrb 3 CumStpPrb 4
 0.975 -0.0865   1079.055      0.0266      0.5292      0.9052           1
 0.950 -0.0794   1140.971      0.0188      0.4409      0.8556           1
 0.900 -0.0713   1211.075      0.0133      0.3495      0.7875           1
 0.800 -0.0615   1283.396      0.0110      0.2654      0.7038           1

Inference at the Stopping Boundaries
    Anlys SampSize     BAM  CIlo.m  CIhi.m Pval.m
Eff     1      425 -0.1610 -0.2228 -0.0852 0.0000
Eff     2      850 -0.0791 -0.1289 -0.0243 0.0026
Eff     3     1275 -0.0548 -0.0955 -0.0064 0.0129
Eff     4     1700 -0.0437 -0.0865  0.0000 0.0250
Fut     1      425  0.0378 -0.0371  0.1005 0.8458
Fut     2      850 -0.0173 -0.0707  0.0341 0.2628
Fut     3     1275 -0.0348 -0.0821  0.0076 0.0530
Fut     4     1700 -0.0437 -0.0865  0.0000 0.0250
```

358

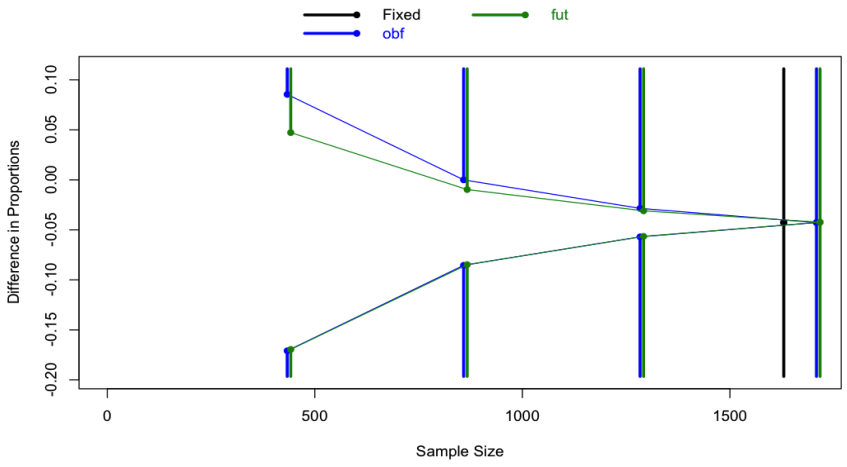
Plotting Stopping Boundaries

- plot() or seqPlotBoundary()
  - Arbitrary number of designs
  - By default, include fixed design with same power as first specified design
- Axes
  - Y-axis: Critical values for stopping on arbitrary boundary scale
    - MLE scale as default
  - X-axis: Statistical information
    - Sample size (number of events for hazards probability models)
- Display of boundaries
  - Vertical lines are true stopping regions
  - Critical values connected to allow better visualization of boundary shape

369

Display of Boundaries

- plot(obf, fut)



370

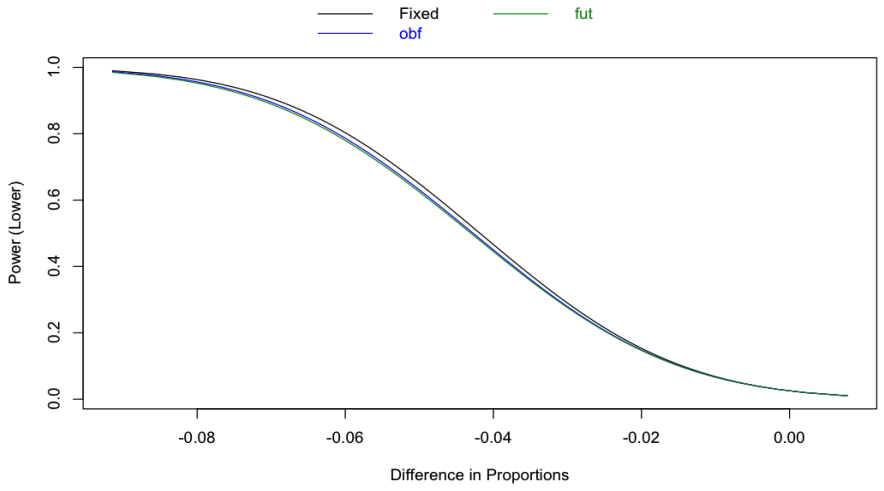
Plotting Power Functions

- seqPlotPower()
  - Arbitrary number of designs
  - By default, include fixed design with same maximal sample size as first specified design
- Axes
  - Y-axis: Power or difference in power from reference design
  - X-axis: Treatment effect (θ)
    - Interpretation based on probability model

371

Display of Power Curves

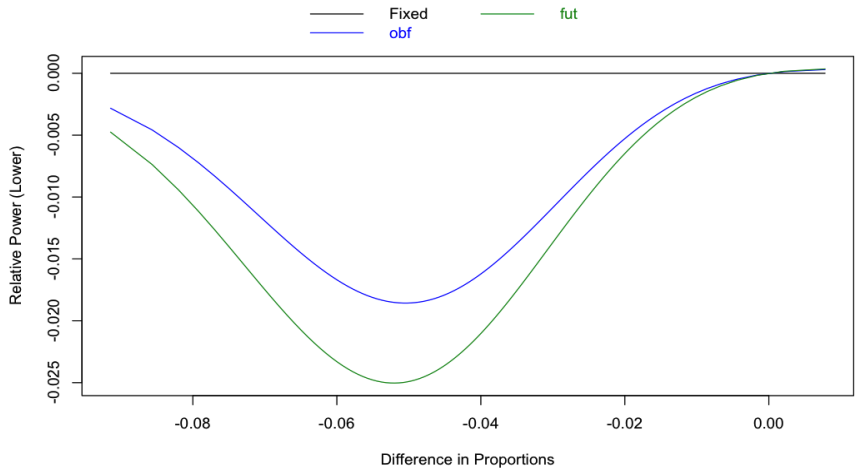
- seqPlotPower(obf, fut)



372

Display of Power Curves (Reference)

- seqPlotPower(obf, fut, reference= T)



373

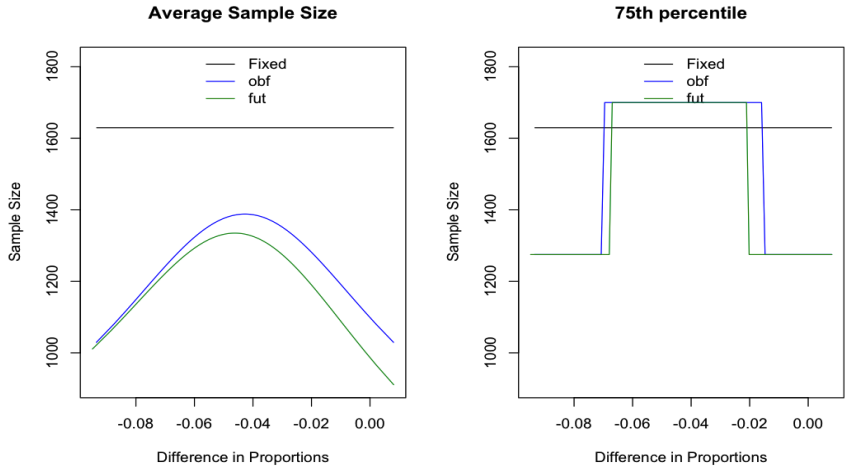
Plotting Sample Size Distribution

- seqPlotASN()
  - Arbitrary number of designs
  - Average sample N (ASN) and arbitrary quantiles
    - Default is 75th percentile
  - By default, include fixed design with same power as first specified design
- Axes
  - Y-axis: Statistical Information (average or quantile)
  - X-axis: Treatment effect (θ)
    - Interpretation based on probability model

374

Display of ASN Curves

- seqPlotASN(obf, fut)



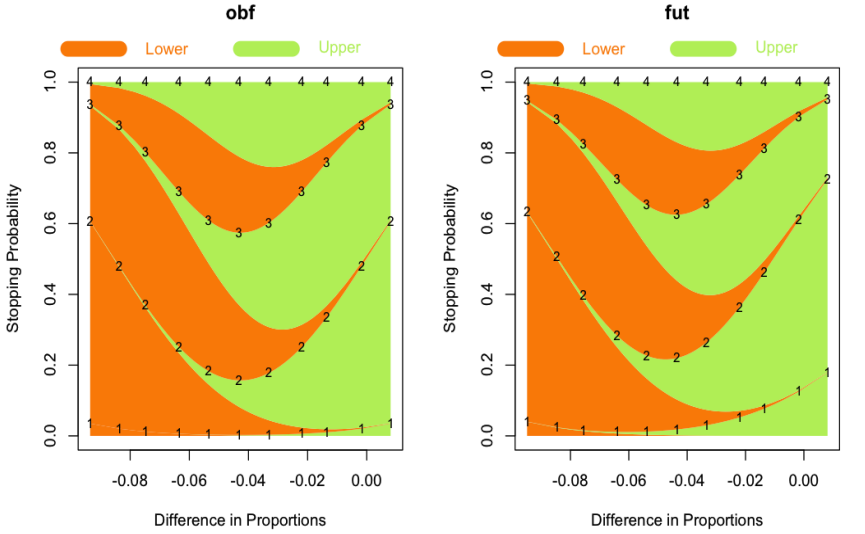
375

Plotting Stopping Probabilities

- seqPlotStopProb()
  - Arbitrary number of designs
    - By default, separate plots for each design
    - Different designs may have different sample sizes
  - Essentially plots of error spending functions for each alternative
- Axes
  - Y-axis: Cumulative stopping probability
  - X-axis: Treatment effect (θ)
    - Interpretation based on probability model
  - Contours labeled by analysis number
    - May not correspond to same sample size
  - Color coding by decision corresponding to stopping rule

376

Display of Stopping Probabilities



377

Take Home Message 8

- **Group sequential, adaptive studies**
  - RCT have always been sequential and adaptive across phases
    - Focus over past 50-70 years has been within individual studies
  - Sequential designs can help address efficiency / ethics of RCT
    - Greatest advantage is terminating "futile" studies earlier
    - For effective therapies, larger sample sizes may be needed
      - Safety and secondary endpoints
  - Group sequential approaches are generally sufficient
    - Fixed sample design as starting point
    - Explore alternative designs varying in schedule/timing of analyses and conservatism of stopping boundaries
    - Fully evaluate inference corresponding to early termination, statistical power, sample size requirements
      - Most boundary scales are nonlinear and difficult to predict
      - Conditional and predictive power have difficult interpretations

378

9

Additional Adaptive Designs

Where am I going?

Other approaches have been described for the adaptive modification of RCT designs.

Distinctions should be made between adaptations that are primarily statistical in nature and those that substantially change the scientific hypotheses being addressed

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

380

Adaptive Sampling Plans

- 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
  - 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

381

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)

382

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

383

Topic for Today: Optimizing the Process

- How do we maximize the number of drugs adopted while
  - Ensuring effectiveness of adopted drugs
  - Ensuring availability of information needed to use drugs wisely
  - Minimizing the use of resources
    - Patient volunteers
    - Sponsor finances
    - Calendar time
- The primary tool at our disposal: Sequential sampling
  - Decrease average sample size used for each drug
  - Maximize number of new drugs using limited resources

384

Distinctions without Differences

- 1:1 correspondence between sequential sampling plans
  - Group sequential stopping rules
  - Error spending functions
  - Conditional / predictive power
  - Bayesian posterior probabilities
- Statistical treatment of hypotheses
  - Superiority / Inferiority / Futility
  - Two-sided tests / bioequivalence

385

Phases of Investigation

- Series of studies support adoption of new treatment
- Preclinical
  - Epidemiology including risk factors
  - Basic science:
    - Biochemistry, physiologic mechanisms, physics / engineering
  - Animal experiments: Toxicology / safety
- Clinical
  - Phase I: Initial safety / dose finding
  - Phase II: Preliminary efficacy / further safety
  - Phase III: Confirmatory efficacy / effectiveness
- Approval of indication
  - (Phase IV: Post-marketing surveillance, REMS)

386

Phase III Confirmatory Trials

- The major goal of a "registrational trial" is to confirm a result observed in some early phase study
- Rigorous science: Well defined confirmatory studies
  - Eligibility criteria
  - Comparability of groups through randomization
  - Clearly defined treatment strategy
  - Clearly defined clinical outcomes (methods, timing, etc.)
  - Unbiased ascertainment of outcomes (blinding)
  - Prespecified primary analysis
    - Population analyzed as randomized
    - Summary measure of distribution (mean, proportion, etc.)
    - Adjustment for covariates

387

Why Confirmation: Real-life Examples

- Effects of arrhythmias post MI on survival
  - Observational studies: high risk for death
  - CAST: Specific anti-arrhythmics have higher mortality
- Effects of beta-carotene on lung CA and survival
  - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
  - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival
- Effects of hormone therapy on cardiac events
  - Observational studies: HT has lower cardiac morbidity and mortality
  - WHI: HT in post menopausal women leads to higher cardiac mortality

388

Multiple Comparisons in Biomedicine

- Observational studies
  - Observe many outcomes
  - Observe many exposures
  - Perform many alternative analyses
    - Summary of outcome distribution, adjustment for covariates
  - Consequently: Many apparent associations
    - May be type I errors
    - But even when valid, may be poorly understood due to confounding
- Interventional experiments
  - Exploratory analyses ("Drug discovery")
    - Modification of analysis methods
    - Multiple endpoints
    - Restriction to subgroups

389

Statistics and Game Theory

- Multiple comparison issues
  - Type I error for each endpoint – subgroup combination
    - In absence of treatment effect, will still decide a benefit exists with probability, say, .025 in each such combination
- Multiple endpoints and subgroups increase the chance of deciding an ineffective treatment should be adopted
  - This problem exists with either frequentist or Bayesian criteria for evidence
  - The actual inflation of the type I error depends
    - the number of multiple comparisons, and
    - the correlation between the endpoints

390

U. S. Regulation of Drugs / Biologics

- 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)

391

U.S. Regulation of 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

392

Phase III Clinical Trials

- Confirmation of efficacy / effectiveness
  - Goals:
    - Obtain measure of treatment's efficacy on disease process
    - Incidence of major adverse effects
    - Therapeutic index
    - Modify clinical practice (obtain regulatory approval)
  - Methods
    - Relatively large number of participants from true target population (almost)
    - Clinically relevant outcome

393

Need for Exploratory Science

- Before we can do a large scale, confirmatory Phase III trial, we must have
  - A hypothesized treatment indication to confirm
    - Disease
    - Patient population
    - Treatment strategy
    - Outcome
  - Comfort with the safety / ethics of human experimentation
- In "drug discovery", in particular, we will not have much experience with the intervention

394

Phase II Clinical Trials

- Preliminary evidence of efficacy
  - Goals:
    - Screening for any evidence of treatment efficacy
    - Incidence of major adverse effects
    - Decide if worth studying in larger samples
      - Gain information about best chance to establish efficacy
        - » Choose population, treatment, outcomes
  - Methods
    - Relatively small number of participants
    - Participants closer to true target population
    - Outcome often a surrogate
    - Sometimes no comparison group (especially in cancer)

395

Screening Studies as Diagnostic Tests

- Clinical testing of a new treatment, preventive agent, or diagnostic method is analogous to using laboratory or clinical tests to diagnose a disease
  - Goal is to find a procedure that identifies truly beneficial interventions
- Not surprisingly, the issues that arise when screening for disease apply to clinical trials
  - Predictive value of a positive test is best when prevalence is high
  - Use screening trials to increase prevalence of beneficial treatments

396

Preliminary Studies in Screening

- In cancer less than 5% of treatments studied in clinical trials are adopted
- NCI drug development program 1970 - 1985
  - 350,000 unique chemical structures studied
  - 83 pass preclinical and phase I testing
  - 24 pass phase II tests for biological activity

397

A Simple (Simplistic?) Setting

- We consider best use of 1,000,000 patients who consent to RCT
  - We consider at most one phase 2 RCT and
  - One phase 3 RCT only among drugs that “pass” phase 2
- We presume a binary prior at the start of phase 2
  - 10% of candidate drugs truly work with a treatment effect of 0.125
  - 90% of candidate drugs have a treatment effect of 0.000
- We presume outcome variability such that 500 subjects provide 80% statistical power in a one-sided level 0.025 test

398

Preliminary Studies in Screening

- Two general approaches to studying new treatments

- Scenario 1:
  - Study every treatment in a large definitive experiment
    - Only do Phase III studies
      - Level of significance 0.025, high power
    - (Ignore, for now, the safety / ethics of this)

- Scenario 2:
  - Perform small screening trials, with confirmatory trials of promising treatments passing early tests
    - Phase II studies
      - Level of significance, power (sample size) to be determined

399

Scenario 1: Only Phase III

- Only large trials using 1,000,000 subjects
  - 10% of drugs being investigated truly work
  - Level of significance .025, .025, or 0.05

  - Sample size / power
    - 979 subjects, α=0.025, 97.5% power ➔ 1,021 RCT
    - 500 subjects, α=0.025, 80.0% power ➔ 2,000 RCT
    - 394 subjects, α=0.050, 80.0% power ➔ 2,538 RCT

  - Results
    - N= 979: 99 effective / 23 ineffective (PV+ = .81)
    - N= 500: 160 effective / 45 ineffective (PV+ = .78)
    - N= 394: 202 effective / 114 ineffective (PV+ = .64)

400

Scenario 2a: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 100 subjects provide 24% power ➔ 7,000 RCT
  - Results
    - N= 100: 168 effective / 158 ineffective (PV+ = .52)
- Use 300,000 subjects in confirmatory Phase III studies
  - 52% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 921 subjects provide 96.7% power ➔ 326 RCT
  - Results

401

Scenario 2b: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .10
  - Sample size / power
    - 342 subjects provide 85% power ➔ 2,047 RCT
  - Results
    - N= 342: 173 effective / 184 ineffective (PV+ = .49)
- Use 300,000 subjects in confirmatory Phase III studies
  - 49% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 839 subjects provide 95% power ➔ 357 RCT
  - Results

402

Summary

		Scenario 1	Scenario 2a	Scenario 2b
Phase 2	Number RCT	2,000 (10% eff)	7,000 (10% eff)	2,047 (10% eff)
	N per RCT	0	100	342
	Type 1 err; Pwr		0.025; 24%	0.100; 85%
	"Positive" RCT		168 eff; 158 not	173 eff; 184 not
Confirmtaory Phase 3	Number RCT	2,000 (10% eff)	326 (52% eff)	357 (49% eff)
	N per RCT	500	921	839
	Type 1 err, Pwr	0.025; 80%	0.025; 97%	0.025; 95%
	# Effctve Adopt	160	162	165
	# Ineff Adopt	45	4	5
	Pred Val Pos	78%	98%	97%
	N per Adopt	500	1,021	1,181

Summary: Phase 2

		Scenario 1	Scenario 2a	Scenario 2b
Phase 2	Number RCT	**2,000 (10% eff)**	**7,000 (10% eff)**	**2,047 (10% eff)**
	N per RCT	**0**	**100**	**342**
	Type 1 err; Pwr		**0.025; 24%**	**0.100; 85%**
	"Positive" RCT		**168 eff; 158 not**	**173 eff; 184 not**
Confirmtaory Phase 3	Number RCT	2,000 (10% eff)	326 (52% eff)	357 (49% eff)
	N per RCT	500	921	839
	Type 1 err, Pwr	0.025; 80%	0.025; 97%	0.025; 95%
	# Effctve Adopt	160	162	165
	# Ineff Adopt	45	4	5
	Pred Val Pos	78%	98%	97%
	N per Adopt	500	1,021	1,181

Summary: Phase 3

		Scenario 1	Scenario 2a	Scenario 2b
Phase 2	Number RCT	2,000 (10% eff)	7,000 (10% eff)	2,047 (10% eff)
	N per RCT	0	100	342
	Type 1 err; Pwr		0.025; 24%	0.100; 85%
	"Positive" RCT		168 eff; 158 not	173 eff; 184 not
Confirmtaory Phase 3	Number RCT	**2,000 (10% eff)**	**326 (52% eff)**	**357 (49% eff)**
	N per RCT	**500**	**921**	**839**
	Type 1 err, Pwr	**0.025; 80%**	**0.025; 97%**	**0.025; 95%**
	# Effctve Adopt	**160**	**162**	**165**
	# Ineff Adopt	**45**	**4**	**5**
	Pred Val Pos	78%	98%	97%
	N per Adopt	500	1,021	1,181

Screening Phase II: Bottom Line

- Pilot studies increase the predictive value of a positive study while using the same number of subjects.
  - Screening parameters can be optimized
    - Proportion of subjects in Phase II vs Phase III
    - Type I error at Phase II
    - Power at Phase II
- Additional considerations when choosing among screening parameters
  - Will we have same prevalence of "good" ideas when we screen 2,000 drugs vs 7,000 drugs?
  - Holding predictive value of positive constant, which strategy provides more information about safety and secondary endpoints for the treatments eventually adopted?

406

Summary: "Drug Discovery"

		Scenario 1	Scenario 2a	Scenario 2b
Phase 2	Number RCT	**2,000 (10% eff)**	**7,000 (10% eff)**	**2,047 (10% eff)**
	N per RCT	0	100	342
	Type 1 err; Pwr		0.025; 24%	0.100; 85%
	"Positive" RCT		168 eff; 158 not	173 eff; 184 not
Confirmtaory Phase 3	Number RCT	2,000 (10% eff)	326 (52% eff)	357 (49% eff)
	N per RCT	500	921	839
	Type 1 err, Pwr	0.025; 80%	0.025; 97%	0.025; 95%
	# Effctve Adopt	**160**	**162**	**165**
	# Ineff Adopt	**45**	**4**	**5**
	Pred Val Pos	**78%**	**98%**	**97%**
	N per Adopt	**500**	**1,021**	**1,181**

Burden of Larger Phase II Studies?

- It appears to be advantageous to use larger Phase 2 studies than is typical currently in cancer research
- BUT: 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

408

Ultimate Goal

- Modify the sample size accrued so that minimal number of subjects treated when
  - new treatment is harmful,
  - new treatment is minimally effective, or
  - new treatment is extremely effective
- Only proceed to maximal sample size when
  - not yet certain of treatment benefit, or
  - potential remains that results of clinical trial will eventually lead to modifying standard practice

409

General Classification of Approaches

- What aspects of the RCT are modified?
  - *Statistical*: Modify only the sample size to be accrued
  - *Scientific*: Possibly modify the hypotheses related to patient population, treatment, outcomes
- Are all planned modifications described at design?
  - *"Prespecified adaptive rules":* Investigators describe
    - Conditions under which trial will be modified and
    - What those modification will consist of
  - *"Fully adaptive":* At each analysis, investigators are free to use current data to modify future conduct of the study

410

Statistical Design Issues

- Under what conditions should we use fewer subjects?
  - Ethical treatment of patients
  - Efficient use of resources (time, money, patients)
  - Scientifically meaningful results
  - Statistically credible results
  - Minimal number of subjects for regulatory agencies
- How do we control false positive rate?
  - Repeated analysis of accruing data involves multiple comparisons

411

Potential Benefits of Stopping Rules

- Group sequential sampling
  - Aggressive early stopping for futility: Pocock boundaries
    - Greatest efficiency (or nearly so)
  - Conservative early stopping for efficacy: O'Brien-Fleming
    - Burden of proof, other endpoints
- Type I error, power maintained exactly at each phase
  - Worst case maximum sample size increases
- Average sample size requirements assuming 10% truly effective drugs at start of Phase II
  - Only large studies : 58.5% of fixed sample

412

Relevance to Adaptive SSRE

- Rather than using group sequential designs, we could have considered unblinded sample size re-estimation
- At the penultimate analysis, unblinded estimates of treatment effect are used to determine the final sample size
- Our experience: Most people "adapt" using an inefficient rule
  - They most often try to use conditional power without fully understanding its behavior
- Furthermore, if the adaptive rule is not fully pre-specified, the test statistic must be based on an inefficient weighting of the

413

Average Efficiency of Adaptive SSRE

- When using minimal sufficient statistics, the "optimal" adaptive design is negligibly more efficient on average than the "optimal" group sequential design (Levin, et al., 2011)
  - When using the reweighted statistic, virtually all of those slight efficiency gains disappear
  - The maximal sample size for adaptive SSRE is greater

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

414

Furthermore

- Additional advantages of screening trials
  - Gathering more detailed preliminary safety data before embarking on expensive, large scale Phase 3 trials
  - Gathering preliminary efficacy data that allows fine tuning
  - Fine tune eligibility criteria
    - Include only susceptible patient populations
    - Exclude patients at high risk for AEs
  - Optimal treatment strategies
    - Fine tune formulation, dose, administration, frequency, duration
    - Develop dose modification strategies
    - Prophylactic treatments, rescue treatments for AEs
  - Optimal clinical endpoints
- Major disadvantage
  - "White space" (time delay) between phase 2 and phase 3
  - (Truly an issue for sponsors, rather than public health)

415

What Can Go Wrong?

- We ignore bias / lack of precision and believe our phase 2 results
  - Phase 3 study will be poorly designed
- We fish through phase 2 data to change outcome
  - We might inflate type 1 error without sufficient increase in power
- We use a surrogate at phase 2 that is not sensitive/specific for true clinical outcome used at phase 3
  - We inflate "effective" type 1 error without increasing power

416

Phase 3 Sample Size from Phase 2

- Phase 2: type 1 error 0.10 with N= 342 provides 85% power
  - Estimated treatment effects: mn (sd; min – max)
    - Ineffective drugs 0.095 (0.022; 0.069 - 0.313)
    - Effective drugs 0.140 (0.043; 0.069 - 0.396)
- Phase 3 using Phase 2 results
  - Estimated N for 95% power:
    - Ineffective drugs 1665 (610; 134 – 2745)
    - Effective drugs 893 (571; 84 – 2745)
  - Type 1 error 0.025, avg power 86%
- Screen 1,759 patients with 1,000,000 patients

417

Relevance to Adaptive SSRE

- The above results highlight a major problem with adaptive sample size re-estimation
- Partway through a RCT, estimates of treatment effect are markedly imprecise

Table 2: Average and Maximal Sample Sizes of Adaptive Designs in Setting 2

	R (Proportion of n_{min} at which adaptation occurs)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ASN	0.99	0.97	0.94	0.91	0.88	0.86	0.84	0.82	0.80	0.78
Maximal N	1.07	1.12	1.16	1.18	1.20	1.21	1.21	1.20	1.18	1.17
$P(N > 1.0)$	0.61	0.68	0.55	0.45	0.36	0.29	0.23	0.18	0.14	0.09
$P(N > 1.1)$	0	0.38	0.36	0.28	0.23	0.18	0.14	0.11	0.09	0.06
$P(N > 1.2)$	0	0	0	0	0	0.11	0.09	0.07	0	0

Coherent vs Incoherent Bayes

- The previous results were based on "staying the course"
- *A priori* we presumed a certain treatment effect
  - Phase 2 studies powered for that treatment effect
  - When progress to phase 3, still power for that treatment effect
- The problem: Phase 2 results are used to decide to go to phase 3
  - Results from "promising phase 2 trials" are biased

419

The Problem of Small Studies

- Using 700,000 patients
  - Small sample size ➔ Big bias of "positive" studies

			Null: Δ =0			Alt: Δ= .125		
N per RCT	**RCTs**	**Crit Value**	**Prob Sig**	**N Sig RCT**	**Expected Estimate**	**Prob Sig**	**N Sig RCT**	**Expected Estimate**
7000	**100**	**0.0234**	0.025	2	0.028	1.000	100	0.125
3500	**200**	**0.0331**	0.025	5	0.039	1.000	200	0.125
700	**1000**	**0.0741**	0.025	25	0.089	0.912	912	0.132
350	**2000**	**0.1048**	0.025	50	0.125	0.649	1,298	0.156
70	**10000**	**0.2343**	0.025	250	0.280	0.180	1,801	0.299
35	**20000**	**0.3313**	0.025	500	0.390	0.114	2,271	0.407

420

Coherent vs Incoherent Bayes

- An alternative optimistic strategy
  - Phase 2 studies powered for presumed treatment effect
  - Phase 3 studies powered for observed phase 2 estimate
- An alternative Bayesian strategy
  - Phase 2 studies powered for presumed treatment effect
  - Phase 3 studies powered for posterior mean treatment effect
    - (up to some maximum)

421

Summary

		Scenario 2b	Optimistic	Mod. Bayes
Phase 2	Number RCT	2,047 (10% eff)	1,759 (10% eff)	1,959 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 85%	0.100; 85%	0.100; 85%
	"Positive" RCT	173 eff; 184 not	150 eff; 159 not	163 eff; 176 not
Confirmtaory Phase 3	Number RCT	357 (49% eff)	309 (49% eff)	339(48% eff)
	N per RCT	839	894 vs 1665	941 vs 998
	Type 1 err, Pwr	0.025; 95%	0.025; 95 vs 86%	0.025; 95%
	# Effctve Adopt	165	129	156
	# Ineff Adopt	5	4	4
	Pred Val Pos	97%	97%	97%
	N per Adopt	1,181	1,259	1,285

Inflation of the Type I Error

- Recall that in order to avoid inflation of type I error, we require confirmatory studies using prespecified
  - Patient population
  - Treatment
  - Primary clinical outcome
  - Statistical analysis
- Hence, we must be concerned about data dredging ("data mining") of the phase 2 data, because it may lead to differences between phase 2 and phase 3 due to
  - Revising outcomes to reflect the most promising results
  - Revising eligibility criteria based on subgroup analyses
  - Changing from surrogate efficacy to effectiveness endpoints
    - "Treating the symptom not the disease"

423

Data Dredging Examples: Endpoints

- We might look for the endpoint for which the treatment has the largest estimated effect
- Examples
  - Overall survival
    - Logrank test vs Wilcoxon logrank vs survival at fixed time …
  - Progression free survival
  - Major adverse cardiovascular events (MACE)
  - MACE plus hospitalization
  - …

424

Data Dredging Examples: Dose / Arms

- We might look for the dose group or treatment arm with largest effect
  - Treatment effect
  - Risk / benefit ratio
  - P value

425

Data Dredging Examples: Subgroups

- In phase 2 trials that are not significant, we search for subgroups that might show significant differences
  - If the results were significant overall, we use the overall results
- In phase 2 trials that are significant, we look for cases in which all the effect seems to be in a subgroup
  - Statistical significance in, say, males
  - Point estimate in wrong direction in females
- We look for the smallest p value among the overall comparison and several subgroups
  - We choose the indication with the smallest p value

426

Examples

- We can explore the impact of adaptive changes in RCT in several examples
  - Consideration of multiple summary measures
    - Mean, geometric mean, Wilcoxon, median, two proportions
  - Consideration of subgroups
    - Overall sample
    - Plus equal sized subgroups defined by three variables
  - Consideration of change of endpoint between phase 2 and 3
    - Phase 2: potential surrogate
    - Phase 3: clinical outcome
- We consider
  - Adaptations that do or do not control type I error
  - Treatment effect in all groups or only in one subgroup
  - Surrogates that do or do not always predict outcome

427

Homogeneous Effects, No Error Control

- First we consider treatments that are equally effective in all subjects
- Prevalence of beneficial treatments: 10%
- Possible adaptations
  - Adaptive choice of statistical summary measure
    - Mean, geometric mean, median, Wilcoxon, two thresholds
  - Look for subgroups having effects
    - Sex, Age (young vs old), BMI (normal vs obese)
    - Strategies
      - If significant overall, proceed with all, otherwise choose most significant subgroup
      - Choose subgroup if it is highly significant and opposite subgroup has estimated nil effect
      - Choose analysis with smallest p value

428

Summary (Homogeneous Effects)

		Scenario 2b	Alt Smry Meas	Subgroups
Phase 2	Number RCT	2,047 (10% eff)	1,695 (10% eff)	1,485 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 85%	0.227; 92%	0.334; 95%
	"Positive" RCT	173 eff; 184 not	155 eff; 346 not	141 eff; 446 not
Confirmtaory Phase 3	Number RCT	357 (49% eff)	501 (31% eff)	587 (24% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 95%	0.025; 94%	0.025; 95%
	# Effctve Adopt	165	147	134
	# Ineff Adopt	5	9	11
	Pred Val Pos	97%	94%	92%
	N per Adopt	1,181	1,181	1,181

Inhomogeneous Effects, No Error Control

- Consider treatments effective only in females
- Prevalence of beneficial treatments: 10%
- Look for subgroups having effects
  - Sex, Age (young vs old), BMI (normal vs obese)
  - Strategies
    - If significant overall, proceed with all, otherwise choose most significant subgroup
    - Choose subgroup if it is highly significant and opposite subgroup has estimated nil effect

430

Impact of Strategies for Subgroups

Analysis	Sig	Pref All	Choice	Min P
All	.64	.64	.40	.07
Females	.85	.20	.40	.60
Males	.10	.00	.00	.00
Young	.45	.02	.03	.06
Old	.45	.02	.03	.06
Norm Wt	.45	.02	.03	.06
Obese	.45	.02	.03	.06

431

Summary (Inhomogeneous Effects)

		Scenario 2b	Prefer All	Choose Subgrp
Phase 2	Number RCT	2,123 (10% eff)	1,490 (10% eff)	1,490 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 64%	0.334; 92%	0.334; 92%
	"Positive" RCT	136 eff; 191 not	137 eff; 448 not	137 eff; 448 not
Confirmtaory Phase 3	Number RCT	327 (42% eff)	584 (23% eff)	584 (23% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 73%	0.025; 75%	0.025; 80%
	# Effctve Adopt	99	103	109
	# Ineff Adopt	5	11	11
	Pred Val Pos	95%	90%	91%
	N per Adopt	1,181	1,181	1,181

Adaptive Sampling Plans

- At each interim analysis, possibly modify
  - Conditions for early stopping
  - Schedule of analyses
  - Randomization ratios
  - Maximal statistical information
  - Statistical criteria for credible evidence
  - Scientific and statistical hypotheses of interest
    - Summary measures used to quantify treatment effect
      - Mean, median, etc.
    - Clinical endpoint
      - Objective response rate, progression, survival, etc.
    - Eligibility criteria
      - Restrict to a subgroup
    - Definition of treatment
      - Drop dose groups, change ancillary treatments, etc.

433

When Stopping Rules Not Pre-specified

- Methods to control the type I error have been described for fully adaptive designs
  - Most popular: Preserve conditional error function from some fixed sample or group sequential design
  - Can have loss of efficiency relative to prespecified plan
- Can choose revised sample size to maintain power
- Methods to compute bias adjusted estimates and confidence intervals not yet well-developed

434

"Partitioning Type 1 Error"

- When designing an adaptive design to look for alternative endpoints, subgroups, doses, we have to decide how to prioritize the different decisions
- This is akin to "spending type 1 error" in sequential trials
- We have to consider our relative beliefs in the treatment effect
  - Is it likely to be homogeneous across all subgroups examined?
  - Is it likely to be concentrated in some pre-specified subgroup?

435

Strategies for Subgroups: Type 1 Error

Example: Assuming independent covariates with 50-50 split

Analysis	Sig	Pref All	Choice	Min P
All	.023	.022	.021	.007
Females	.023	.013	.013	.015
Males	.023	.013	.013	.015
Young	.023	.013	.013	.015
Old	.023	.013	.013	.015
Norm Wt	.023	.013	.013	.015

436

Strategies for Subgroups: Alternatives

- Need to consider how we thing the overall treatment effect might differ from effects within subgroups

- Cases we have considered
  - Hypothesized treatment effect actually occurs only in subpopulation
    - Overall test is extremely underpowered: 45% instead of 85%
  - Slightly stronger hypothesized treatment effect only in subpopulation
  - Overall population's treatment effect as hypothesized
    - But one subgroup has double that effect and opposite subgroup has no effect

437

Generalizability

- We need to consider type 1 and type 2 errors relative to the ultimate result of the "drug discovery" process

- The previous results are dependent on
  - A mixture of 10% effective drugs and 90% ineffective drugs, *where "effectiveness" is defined based on the clinical outcome used in the phase 3 trial*
  - Phase 2 and phase 3 type I errors being controlled at the specified level *based on the phase 3 outcome*
  - Phase 2 and phase 3 power being controlled at the specified level *based on the phase 3 outcome*

438

- Many early phase RCT use alternative outcomes

Inhomogeneous Effects, Control Errors

- Consider treatments effective only in females
- Prevalence of beneficial treatments: 10%
- Look for subgroups having effects
  - Sex, Age (young vs old), BMI (normal vs obese)
  - Strategies as before
- Perform all tests using type I error of 0.023
  - Yields experimentwise type I error of 0.100

439

Control Error (Inhomogeneous Effects)

		Scenario 2b	Inflate Error	Control Error
Phase 2	Number RCT	2,123 (10% eff)	1,490 (10% eff)	1,720 (10% eff)
	N per RCT	342	342	438
	Type 1 err; Pwr	0.100; 64%	0.334; 92%	0.100; 80%
	"Positive" RCT	136 eff; 191 not	137 eff; 448 not	138 eff; 156 not
Confirmtaory Phase 3	Number RCT	327 (42% eff)	584 (23% eff)	294 (47% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 73%	0.025; 80%	0.025; 76%
	# Effctve Adopt	99	109	105
	# Ineff Adopt	5	11	4
	Pred Val Pos	95%	91%	96%
	N per Adopt	1,181	1,181	1,277

Control Error (Inhomogeneous Effects)

- With inhomogeneous effects, we also need to consider additional errors
- A "True Positive" would be adoption of a new treatment in exactly the population that benefits
- "False Positives" might include drugs with too broad an indication
  - It does not work in part of the population
- "False Negatives" might include a drug that has omitted part

441

Homogeneous Effects, Surrogates

- We consider treatments that are equally effective in all subjects
- Prevalence of beneficial treatments: 10%
  - "Beneficial" defined based on phase 3 endpoint
- Prevalence of misleading treatments: 0%, 10%, or 20%
  - "Misleading" = efficacious on surrogate but not effective
  - 85% power to detect efficacy on surrogate
- No adaptations

442

Surrogates (Homogeneous Effects)

		0% Misleading	10% Misleading	20% Misleading
Phase 2	Number RCT	2,046 (10% eff)	1,812 (10% eff)	1,627 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 85%	0.100; 85%	0.100; 85%
	"Positive" RCT	174 eff; 184 not	154 eff; 337 not	138 eff; 494 not
Confirmtaory Phase 3	Number RCT	358 (49% eff)	491 (31% eff)	632 (22% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 95%	0.025; 95%	0.025; 95%
	# Effctve Adopt	166	147	132
	# Ineff Adopt	5	8	12
	Pred Val Pos	97%	95%	91%
	N per Adopt	1,181	1,181	1,181

Comparisons

	RCT	Eff (TP)	Not(FP)	n
Nonadaptive				
• Homogeneous effect	2,040	165(165)	5	1,181
• Homogeneous,10% misleading	1,812	147(147)	8	1,181
• Homogeneous,20% misleading	1,627	132(132)	12	1,181
• Inhomogeneous effect	2,123	99(0)	5	1,181

444

Seamless Phase 2 / 3

- In cases that no changes will be made between Phase 2 and Phase 3, can try to use same trial
  - Need to ensure that same level of evidence is provided as would be in two independent trials
    - Pivotal 0.005 vs 0.000625 in two independent trials?
    - One RCT setting vs two RCT settings (random effects)
- Such would eliiminate “white space”
  - But note that white space is truly an issue for those whose focus is on a particular agent
  - During “white space” other agents in the pipeline can be investigated
  - Eliminating “white space” limits scientific, regulatory, and ethical review of phase 2 results

445

Comments

- Screening Phase II trials provide great protection
  - Ensure that overwhelming majority of adopted therapies are truly effective
- However, control of type I and II errors are of great importance even at phase 2
  - But note that type 1 error of 0.025 not necessarily indicated
- Adaptive designs can help provide that control
  - But need to re-power the study to get greatest benefit
  - The added benefit over nonadaptive designs is not huge, but there are advantages
    - Higher power and predictive value of the positive
    - More beneficial drugs identified
    - More patient exposure for adopted drugs

446

- Adaptation cannot protect against false surrogates

Adaptive Sample Size Determination

- Design stage:
  - Choose an interim monitoring plan
  - Choose an adaptive rule for maximal sample size
- Conduct stage:
  - Recruit subjects, gather data in groups
  - After each group, analyze for DMC
    - DMC recommends termination or continuation
  - After penultimate group, determine final N
- Analysis stage:
  - When study stops, analyze and report

447

Adaptive Sample Size Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - $N_1 \ldots N_{J-1}$ can be randomly determined indep of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- At $N_1 \ldots N_{J-1}$ compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
    - At N_{J-1} determine N_J according to value of T_j

448

When Stopping Rules Not Pre-specified

- Methods to control the type I error have been described for fully adaptive designs
  - Most popular: Preserve conditional error function from some fixed sample or group sequential design
  - Can have loss of efficiency relative to pre-specified designs
- Recently methods to compute bias adjusted estimates and confidence intervals have been developed
  - Loss of precision relative to pre-specified designs

449

Scientific Design Issues

- Under what conditions should we alter our hypotheses?
  - Definition of a “Treatment Indication”
    - Disease
    - Patient population
    - Definition of treatment (dose, frequency, ancillary treatments, …)
    - Desired outcome
- How do we control false positive rate?
  - Multiple comparisons across subgroups, endpoints
- How do we provide inference for Evidence Based Medicine?

450

Adaptive Sampling: Issues

"Make new friends, but don't forget the old.
One is silver and the other is gold"

- Children's song

451

CDER/CBER Guidance on Adaptive Designs

- Recommendations for use of adaptive designs in confirmatory studies
  - Fully pre-specified sampling plans
  - Use well understood designs
    - Fixed sample
    - Group sequential plans
    - Blinded adaptation
  - For the time-being, avoid less understood designs
    - Adaptation based on unblinded data
- Adaptive designs may be of use in early phase clinical trials

452

Perpetual Motion Machines

- A major goal of this course is to discern the hyperbole in much of the recent statistical literature
  - The need for adaptive clinical trial designs
    - What can they do that more traditional designs do not?
  - The efficiency of adaptive clinical trial designs
    - Has this been demonstrated?
    - Are the statistical approaches based on sound statistical foundations?
  - "Self-designing" clinical trials
    - When are they in keeping with our scientific goals?

453

Important Distinctions in this Course

- What aspects of the RCT are modified?
  - *Statistical*: Modify only the sample size to be accrued
  - *Scientific*: Possibly modify the hypotheses related to patient population, treatment, outcomes
- Are all planned modifications described at design?
  - *"Prespecified adaptive rules":* Investigators describe
    - Conditions under which trial will be modified and
    - What those modification will consist of
  - *"Fully adaptive":* At each analysis, investigators are free to use current data to modify future conduct of the study
- When do the modifications take place?
  - *Immediately:* Potential to stop study at current analysis
  - *Future stages of the RCT:* Change plans for next analysis

454

Response Adaptive Modifications of an RCT Design

Overview of General Methods

Where am I going?

Various authors have advocated the use of unblinded interim estimates of the treatment effect to modify a broad spectrum of RCT design parameters.

In this section of this course, I review the most commonly espoused statistical methods that would be used to maintain an experimentwise type I error.

Blinded Adaptive Sampling

- Modify sample size to account for estimated information (variance or baseline rates)
- No effect on type I error IF
  - Estimated information independent of estimate of treatment effect
    - Proportional hazards,
    - Normal data, and/or
    - Carefully phrased alternatives
  - And willing to use conditional inference
    - Carefully phrased alternatives

456

Estimation of Statistical Information

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

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

457

Estimation of Statistical Information

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

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

458

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

459

Two Stage Design

- Proschan & Hunsberger consider worst case
  - At first stage, choose sample size of second stage
    - $N_2 = N_2(Z_1)$ to maximize type I error
  - At second stage, reject if $Z_2 > a_2$

460

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)

461

Flexible Adaptive Designs

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

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

  - Often choose function from some specified test

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

  - Find critical value to maintain type I error

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

462

Other Approaches

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

463

Incremental Statistics

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

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

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

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

464

Conditional Distribution

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

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

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

465

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)

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

466

Unconditional Distribution

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

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

467

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

468

Response Adaptive Modifications of an RCT Design

Modification of the Maximal Sample Size

Where am I going?

The majority of the statistical literature on adaptive clinical trial design is directed toward using interim estimates of the treatment effect to re-power a study.

In this section of the course I review the basis for those methods and demonstrate the settings in which such adaptive designs can improve substantially on more traditional methods.

Adaptive Sample Size Determination

- Design stage:
  - Choose an interim monitoring plan
  - Choose an adaptive rule for maximal sample size
- Conduct stage:
  - Recruit subjects, gather data in groups
  - After each group, analyze for DMC
    - DMC recommends termination or continuation
  - After penultimate group, determine final N
- Analysis stage:
  - When study stops, analyze and report

470

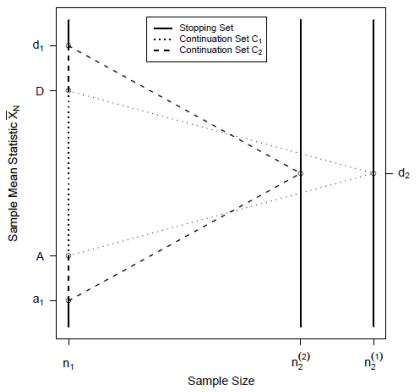
Adaptive Sample Size Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - $N_1 \ldots N_{J-1}$ can be randomly determined indep of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- At $N_1 \ldots N_{J-1}$ compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
    - At N_{J-1} determine N_J according to value of T_j

471

Operating Characteristics

- Given an adaptive rule that modifies the final sample size at the penultimate stage, this can be viewed as adaptively switching between two group sequential tests
  - Standard group sequential software can be used



472

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

473

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)

474

Prespecified Modification Rules

- Adaptive sampling plans exact a price in statistical efficiency
- Tsiatis & Mehta (2002); Jennison & Turnbull (2003)
  - A classic prespecified group sequential stopping rule can be found that is more efficient than a given adaptive design
- Shi (2003, unpublished thesis)
  - Fisher's test statistic in the self-designing trial provides markedly less precise inference than that based on the MLE
    - To compute the sampling distribution of the latter, the sampling plan must be known

475

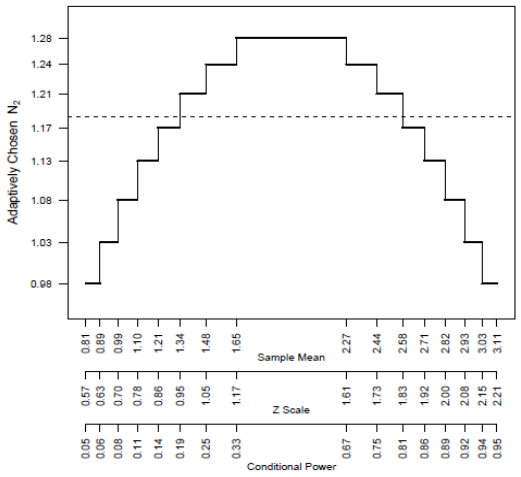
Comments re Conditional Power

- Many authors propose adaptations based on conditional or predictive power
- Neither conditional power nor predictive power 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

476

Optimal Adaptions by Conditional Power

- Difficulty understanding conditional / predictive power scales can lead to bad choices for designs



477

Optimal Timing of Adaptation

- When should adaptive analysis take place relative to minimal acceptable sample size?

Table 2: Average and Maximal Sample Sizes of Adaptive Designs in Setting 2

	R (Proportion of n_{min} at which adaptation occurs)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ASN	0.99	0.97	0.94	0.91	0.88	0.86	0.84	0.82	0.80	0.78
Maximal N	1.07	1.12	1.16	1.18	1.20	1.21	1.21	1.20	1.18	1.17
$P(N > 1.0)$	0.61	0.68	0.55	0.45	0.36	0.29	0.23	0.18	0.14	0.09
$P(N > 1.1)$	0	0.38	0.36	0.28	0.23	0.18	0.14	0.11	0.09	0.06
$P(N > 1.2)$	0	0	0	0	0	0.11	0.09	0.07	0	0

478

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*

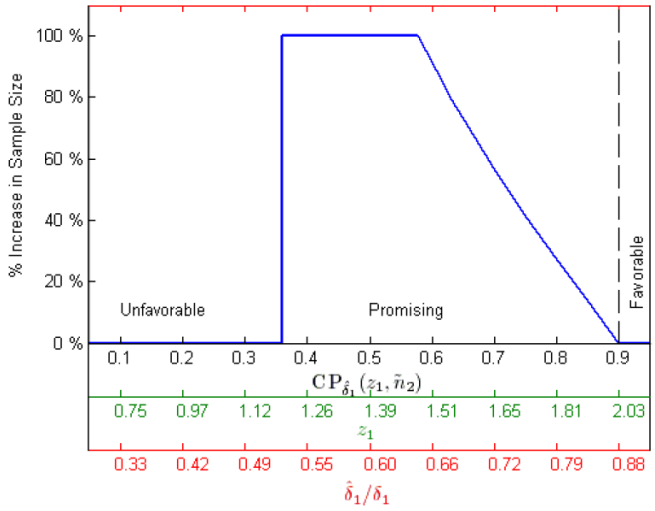
SUMMARY

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.

- http://www.cytel.com/pdfs/Mehta_Pocock_PromisingZone_StatsinMed_9.11.10.pdf

479

Example Modification Plan



480

Alternative Approaches

Table 1: Comparison of RCT Designs for Example 1

Design	Hypothesized Treatment Effect $\delta = 0$	$\delta = 1.5$	$\delta = 1.6$	$\delta = 1.7$	$\delta = 1.8$	$\delta = 1.9$	$\delta = 2.0$
				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

481

Alternative Approaches

Table 1: Comparison of RCT Designs for Example 1

Design	Hypothesized Treatment Effect $\delta = 0$	$\delta = 1.5$	$\delta = 1.6$	$\delta = 1.7$	$\delta = 1.8$	$\delta = 1.9$	$\delta = 2.0$
				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

482

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

483

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
- Even experienced clinical trialists get it wrong
  - The authors also seemed somewhat chagrined that efforts to boost conditional power did not result in substantial gains in unconditional power

484

Delayed Measurement of Response

- This example had delayed measurement of response
- It is in that situation that we have found the greatest advantage of an adaptive design
  - Early evidence of an effect may allow curtailing accrual
  - If constraining schedule of analyses, then adaptive design sometimes outperform GST depending on relative magnitude of
    - Per patient costs
    - Calendar time costs
- With relaxation of constraints, we easily found a better GST
  - (Emerson, Rudser, Emerson, 2011)

485

The Cost of Planning Not to Plan

- In order to provide frequentist estimation, we must know the rule used to modify the clinical trial
  - We can then evaluate candidate designs
- Hypothesis testing of a null is possible with fully adaptive trials
  - Statistics: type I error is controlled
  - Game theory: chance of "winning" with completely ineffective therapy is controlled
  - Science:
    - At best: ability to discriminate clinically relevant hypothesis may be impaired
    - At worst: uncertainty as to what the treatment has effect on
- With prespecified adaptation, we know how to do inference, but standard orderings may lead to surprising results

486

Take Home Message 9

- **Adaptive RCT designs**
  - Group sequential designs a special, well understood case
  - Methods based on unblinded adaptation
    - Statistical inference not based on minimal sufficient statistics
    - Unblinded adaptive modifications of statistical aspects of design generally not recommended by me due to inefficiency
      - Randomization ratio
      - Maximal statistical information (SSRE)
        - » Possible exception: altering accrual in time to event
  - Adaptive modifications of indication (phase 2 more than phase 3)
    - Adaptive enrichment of patient population to find greatest efficacy
    - Adaptive selection of doses or treatments
    - Adaptive selection of outcomes
      - Less useful because outcome should be based on clinical importance

487

8

Statistical Analysis Plan

Where am I going?

A formal Statistical Analysis Plan (SAP) is developed and finalized prior to unblinding treatments.

This serves to ensure that all analysis methods are totally pre-specified, thereby avoiding data-driven hypothesis.

Purpose of Descriptive Statistics

- Identify errors in measurement, data collection
- Characterize materials and methods
- Assess validity of assumptions needed for analysis
- Straightforward estimates to address scientific question
- Hypothesis generation

489

Purpose #3: Validity of Assumptions

- Sampling scheme (of crucial importance)
  - Is our sample relevant to our question?
  - Missing data
  - Confounding
    - A function of means (arithmetic, geometric, proportions)
  - Influential cases
- Modeling of dose response (importance depends on question)
  - Not so important with modeling of baseline covariates for precision
  - Effect modification (scientific question or exploration)
  - Linearity of association (estimation in groups or exploration)
- Distributional assumptions
  - Technical assumptions with strong (semi)parametric models
  - Avoided with use of distribution-free approaches

490

Take Home Message 10

- **Statistical Analysis Plan**
  - Table 1: Compare distributions of baseline variables across arms
    - Materials and methods: range, quartiles, means, SD
    - Randomization imbalance: means (geometric means, proportions)
  - Formal sequential stopping rules
  - Complete pre-specification of analysis models for all primary and secondary endpoints
    - Adjustment for covariates and exact form used in modeling
      - Model misspecification not much of issue
    - Handling of missing data and sensitivity analyses
    - Handling of multiple comparisons: Hierarchy of testing
  - Confirmatory and supportive analyses re mechanism of action
    - Alternative summary measures and alternative clinical endpoints
  - Exploratory analyses of important subgroups

491

Implementation for Monitoring

Schedule of Analyses

- Design of clinical trial : Selection of stopping rule to provide desired operating characteristics
  - Type I error
  - Statistical power to detect design alternative
  - Efficiency
  - Bayesian properties
  - Futility considerations

493

Schedule of Analyses

- At time of study design : Sample size (power, alternative) calculations based on
  - Specifying a maximum of J analyses
  - Specifying sample sizes at which analyses will be performed
- During conduct of study: Number and timing of analyses may be different
  - Monitoring scheduled by calendar time
  - Slow (or fast) accrual
  - Estimation of available information at time of locking database
  - External causes
  - (should not be influenced by study results)

494

Impact of Altered Schedule

- Summary for Pocock boundary relationships

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

495

Impact of Altered Schedule

- Summary for O'Brien-Fleming boundary relationships

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

496

Implementation

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

497

Estimation of Statistical Information

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

498

Estimation of Statistical Information

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

499

Blinded Adaptive Sampling

- Modify sample size to account for estimated information (variance or baseline rates)
- No effect on type I error IF
  - Estimated information independent of estimate of treatment effect
    - Proportional hazards,
    - Normal data, and/or
    - Carefully phrased alternatives
  - And willing to use conditional inference
    - Carefully phrased alternatives

500

Estimation of Statistical Information

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

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

501

Estimation of Statistical Information

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

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

502

Schedule of Analyses

- Use of constrained families is necessary because critical values are dependent upon exact schedule
  - In Unified Family, boundary at first analysis is affected by timing of later analyses
  - Compare boundary at first analysis when timing of second analysis differs:
    - `a' boundary: -0.0743 versus -0.0740
    - 'd' boundary: 0.5744 versus 0.5713
  - Must constrain first boundaries at the levels actually used, and then use parametric form for future analyses

503

Flexible Determination of Boundaries

- When implementing stopping rules, must be able to accommodate changes
- Previously described methods for implementing stopping rules
  - (Adhere exactly to monitoring plan)
  - (Approximations based on design parameters: Emerson and Fleming, 1989)
  - Christmas tree approximation for triangular tests: Whitehead and Stratton, 1983
  - Error spending functions: Lan and DeMets, 1983; Pampallona, Tsiatis, and Kim, 1995
  - Constrained boundaries in unified design family: Emerson, 2000; Burington and Emerson, 2003

504

Common Features

- Stopping rule specified at design parameterizes the boundary for some statistic (boundary scale)
- At the first interim analysis, parametric form is used to compute the boundary for actual time on study
- At successive analyses, the boundaries are recomputed accounting for the exact boundaries used at previously conducted analyses
- Maximal sample size estimates may be updated

505

Constrained Boundaries Approach

- Use of constrained families in flexible implementation of stopping rules
  - At the first analysis, compute stopping boundary from parametric family
  - At successive analyses, use parametric family with constraints (on some scale) for the previously conducted interim analyses
  - When the error spending scale is used, this is just the error spending approach of Lan & DeMets or Pampallona, Tsiatis, & Kim

506

Case Study: Summary of Study

- Target population
  - Patients in ICU with newly diagnosed gram negative sepsis
- Primary endpoint
  - 28 day mortality (binary)
- Summary measure (contrast)
  - Difference in probability of mortality within 28 days
- Test type
  - One-sided level .025 test of binomial proportions
- Baseline event rate and clinically relevant difference
  - Assumed to be 30%
  - 5% absolute difference in 28 day mortality

507

Case Study: Monitoring the Trial

- Maintaining operating characteristics
  - Depends upon statistical information
  - In the binomial case, this is determined by the event rates
  - If event rates are different from those assumed at design stage, need to choose whether to
    - Re-estimate maximum sample size to maintain power
    - Maintain maximum sample size and re-compute power
  - Re-compute these by using the POOLED event rate on observed data

508

Case Study: Simulated Monitoring

- Simulated example
  - Event rate in control arm is 0.35
  - Event rate in treatment arm is 0.25
- First interim analysis takes place after 400 patients observed (200 per arm)

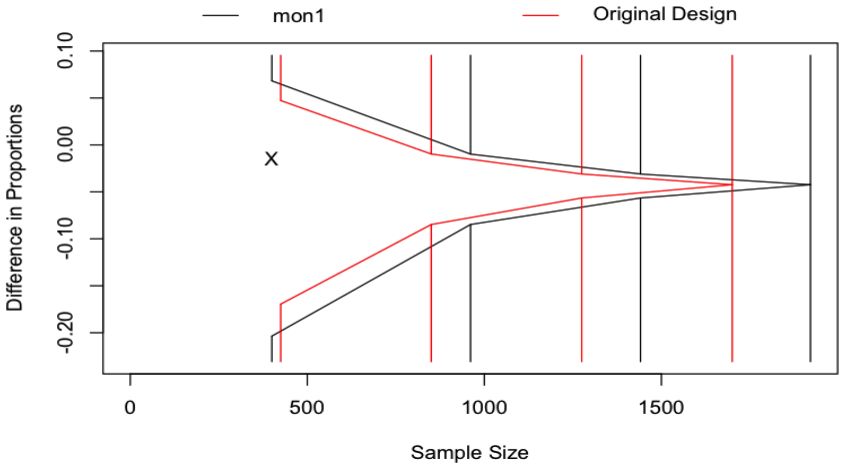
509

First Monitoring Analysis: Re-estimate N

```
> obs <- c(rbinom(200,1,0.25),rbinom(200,1,0.35))
> tx <- rep(1:0,each=200)
> mon1 <- seqMonitor(fut,obs,tx)
> mon1
Call:
seqMonitor(x = fut, response = obs, treatment = tx)

RECOMMENDATION:
     Continue

OBSERVED STATISTICS:
 Sample Size Crude Estimate Z Statistic
         400         -0.015      -0.3209

MONITORING BOUNDS:
Call:
"[for original design call, see $seqDesignCall in seqMonitor object]"

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in probabilities (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=  0.0000    (size  = 0.025)
    Alternative hypothesis : Theta <= -0.0713    (power = 0.900)

STOPPING BOUNDARIES: Sample Mean scale
                           a       d
    Time 1 (N=  400) -0.2036  0.0682
    Time 2 (N=  961) -0.0848 -0.0098
    Time 3 (N= 1441) -0.0565 -0.0310
    Time 4 (N= 1921) -0.0424 -0.0424
```

510

First Monitoring Analysis



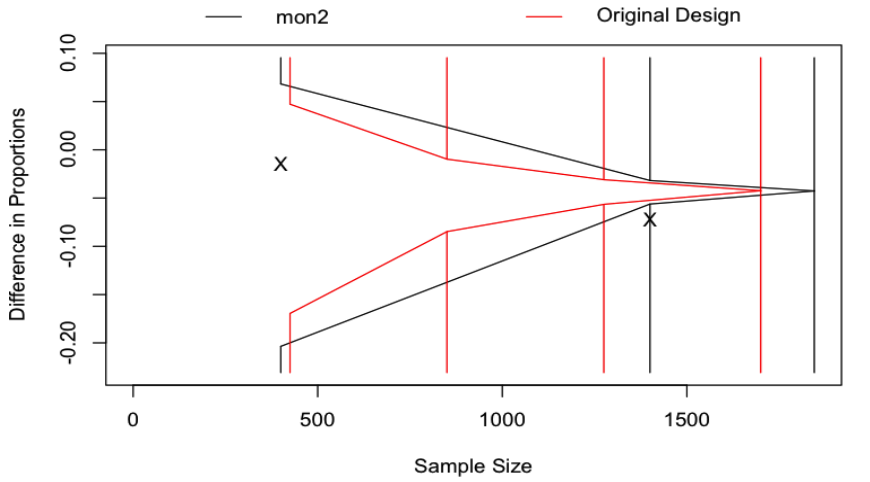
511

Second Monitoring Analysis

```
> obs <- c(obs,c(rbinom(500,1,0.25),rbinom(500,1,0.35)))
> tx <- c(tx,rep(1:0,each=500))
> mon2 <- seqMonitor(mon1,obs,tx)
Warning message:
In seqMonitor(mon1, obs, tx) :
  1 specified future analysis time(s) were within min.increment of the current time or earlier, and are deleted
> mon2
Call:
seqMonitor(x = mon1, response = obs, treatment = tx)

RECOMMENDATION:
     Stop with decision for Lower Alternative Hypothesis

OBSERVED STATISTICS:
 Sample Size Crude Estimate Z Statistic
         400       -0.01500     -0.3209
        1400       -0.07286     -2.9544

INFERENCE:

Adjusted estimates based on observed data:
  analysis.index observed      MLE      BAM    RBadj
1              2 -0.07286 -0.07286 -0.07147 -0.07282

Inferences based on Analysis Time Ordering:
       MUE  P-value       **** CI ****
1 -0.07285 0.001570 (-0.1212, -0.0245)

Inferences based on Mean Ordering:
       MUE  P-value       **** CI ****
1 -0.07235 0.001585 (-0.1207, -0.0243)
```

512

Second Monitoring Analysis



513

Time to Event Endpoints

Estimating Subject Accrual

Where am I going?

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

Additional consideration must be given to accrual of subjects.

Time to Event Endpoints

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

515

Case Study: Stopping Rule

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

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

PROBABILITY MODEL and HYPOTHESES:
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1.00    (size  = 0.025)
    Alternative hypothesis : Theta <= 0.77    (power = 0.900)

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

516

Specification of Accrual

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

517

Case Study: Accrual

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

518

Case Study: Accrual

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

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

Timing of analyses:

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

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

519

Case Study: Accrual (months)

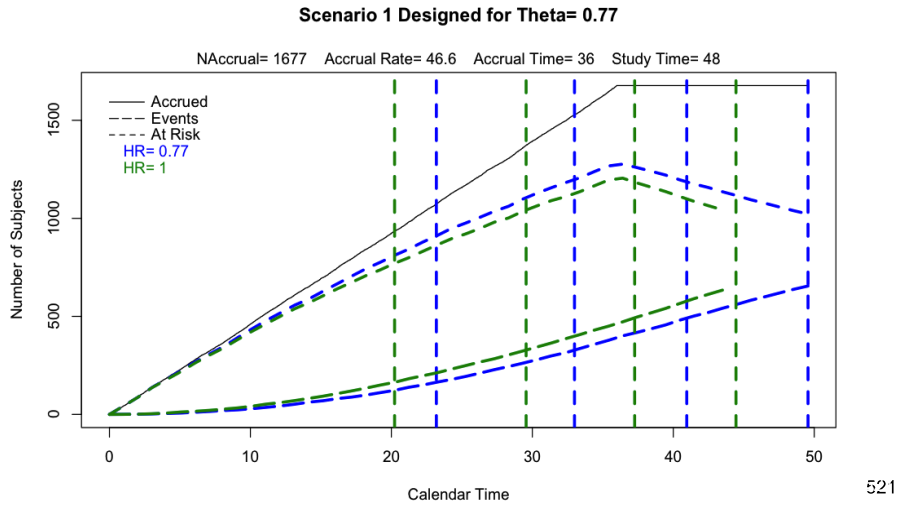
- tte <- update(tte, accrualTime= 36, studyTime= 48, eventQuantiles= 36)

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

Timing of analyses:

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

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

520

Case Study: Accrual Plots

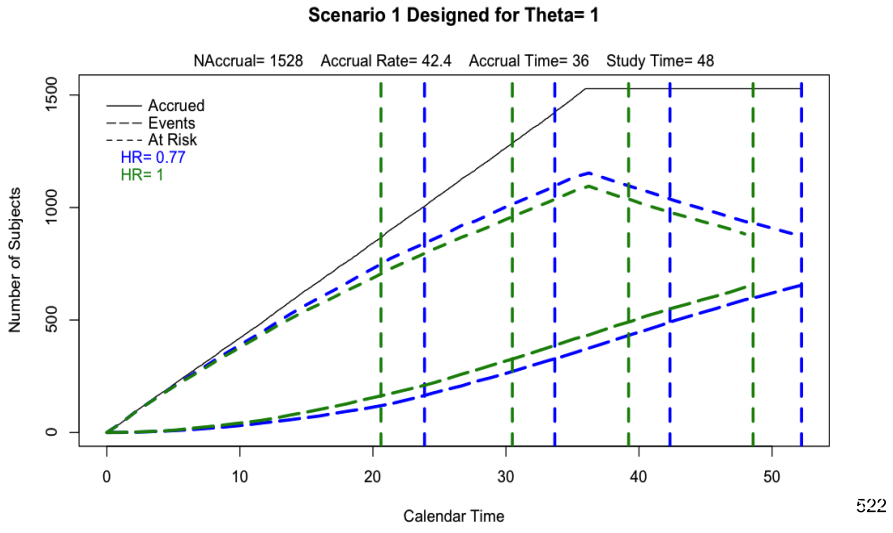
- seqPlotPHNSubjects(tte)



521

Case Study: Accrual Plots

- seqPlotPHNSubjects(tte)



522

Frequentist Inference Following GSD

Where am I going?

Group sequential and adaptive testing modifies the sampling distribution of the estimates.

Frequentist inference must account for the proper sampling distribution.

Methods have been well-developed for these situations.

- There is no uniformly optimal method of computing CI and p values, so particular orderings of the sample space must be selected.

Frequentist Inference

Observed data : $\vec{Y} = \vec{y}$

Probability model : $\vec{Y} \mid \vec{\theta}_{\bar{X}} \sim F\left(\vec{y}, \vec{\theta}_{\bar{X}}\right)$

P value : $\Pr\left(\vec{Y} > \vec{y} \mid \vec{\theta}_0\right)$

CI for $\vec{\theta}$: $\left\{\vec{\theta} : \frac{\alpha}{2} \le \Pr\left(\vec{Y} > \vec{y} \mid \vec{\theta}\right) \le 1 - \frac{\alpha}{2}\right\}$

Mdn Unbiased estimates : $\Pr\left(\vec{Y} > \vec{y} \mid \hat{\vec{\theta}}\right) = 0.5$

524

Implementation

- Define "extreme data" for each hypothesis
  - Choose test statistic to define ordering

$$\vec{y}_1 > \vec{y}_2 \quad \Leftrightarrow \quad T(\vec{y}_1) > T(\vec{y}_2)$$

525

Statistical Theory

- Parametric probability models
  - Neyman – Pearson Lemma
    - LR is MP for simple hypotheses
  - Karlin – Rubin
    - LR is UMP if monotone likelihood ratio
  - Cramer – Rao
    - Efficient estimates in "regular" problems
  - Asymptotic likelihood theory
    - CAN estimators in "regular" problems

526

Clinical Trials

- Nonparametric probability models
  - Regulatory agencies do not allow model checking
- Nuisance parameters
  - Continuous data: Within group variance
  - Binary data: Placebo event rate in comparative trials
- Finite sample sizes
  - Especially when assessing severe toxicity

527

Sequential Sampling

- Curved sample space
  - With one dimensional parameter, the sufficient statistic is two-dimensional
  - Impact on completeness
- No uniformly most powerful tests
- No uniformly most accurate CI

528

Ordering the Outcome Space

- Issue similar to decisions commonly faced with 2 x 2 contingency tables
  - Chi squared vs Fisher's exact vs LR vs Wald ordering
- Ordering of the sequential outcome space
  - Orderings of outcomes within an analysis time intuitive
    - Based on the value of T_j at that analysis
  - Need to define ordering between outcomes at successive analyses
    - How does sample mean of 3.5 at second analysis compare to sample mean of 3 at first analysis (when estimate more variable)?

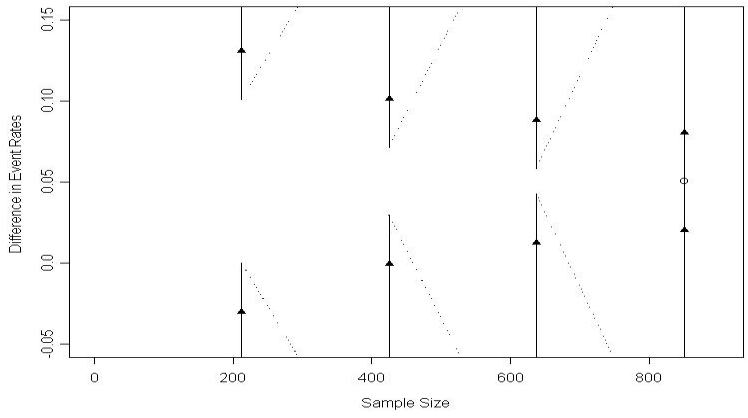
529

Stagewise Ordering

- Jennison & Turnbull, 1983; Tsiatis, Rosner, & Mehta, 1984
- Results leading to earlier stopping are more extreme
  - Linearizes the outcome space
  - Depends on the sequential stopping rule
  - Not defined for two-sided tests with early stopping for both null and alternative
  - Independent of alternative

530

Stagewise Ordering



531

MLE ordering

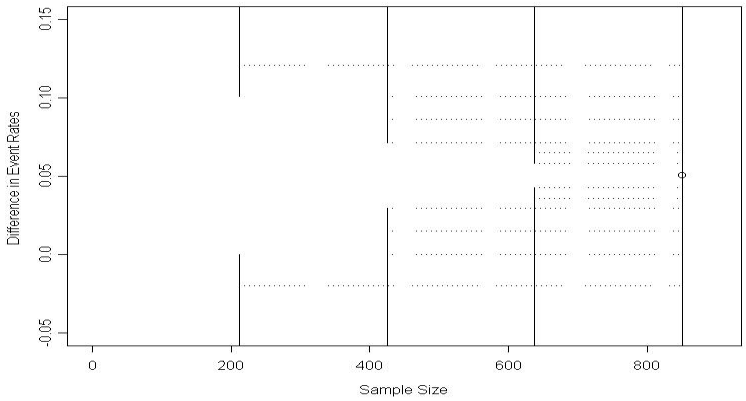
- Duffy and Santner, 1987; Emerson and Fleming, 1990
- Consider only magnitude of sample mean
  - Independent of stopping rule

$$(m_1, \bar{x}_1) > (m_2, \bar{x}_2) \quad \Leftrightarrow \quad \bar{x}_1 > \bar{x}_2$$

532

MLE Ordering

- Contours of equality



533

LR ordering

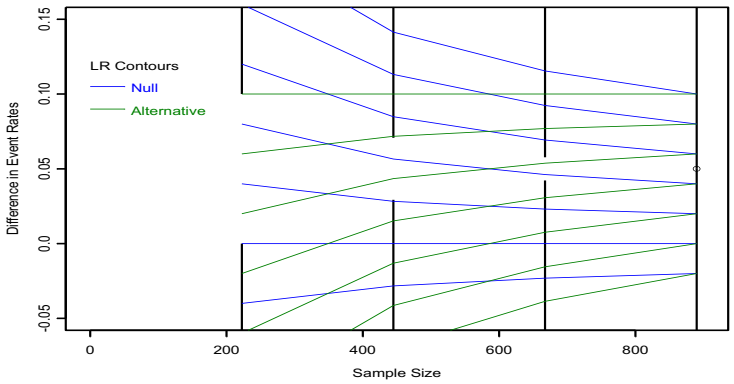
- Chang and O'Brien, 1986; Chang, 1988
- Consider only Likelihood Ratio
  - Independent of stopping rule
  - Depends on alternative

$$(m_1, \bar{x}_1) >_\theta (m_2, \bar{x}_2) \quad \Leftrightarrow \quad \frac{p(m_1, \bar{x}_1 \mid \hat{\theta}_1)}{p(m_1, \bar{x}_1 \mid \theta)} > \frac{p(m_2, \bar{x}_2 \mid \hat{\theta}_2)}{p(m_2, \bar{x}_2 \mid \theta)}$$

534

LR Ordering

- Contours of equality
  - Under null and alternative



535

Optimality Criteria

- Dependence on sampling plans
  - AT: depends on sampling plan, but not future analysis times
- Shorter CI (on average)
  - MLE < AT
  - MLE < LR for O'Brien-Fleming; LR < MLE for Pocock
- Probability of obtaining a low P value (Pivotal studies)
  - LR > MLE > AT
- Agreement with stopping boundaries
  - MLE: if boundaries are appropriately monotonic
  - AT: unless different thresholds at similar sample sizes

536

Adaptive Sample Size Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - $N_1 \ldots N_{J-1}$ can be randomly determined indep of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- At $N_1 \ldots N_{J-1}$ compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
    - At N_{J-1} determine N_J according to value of T_j

537

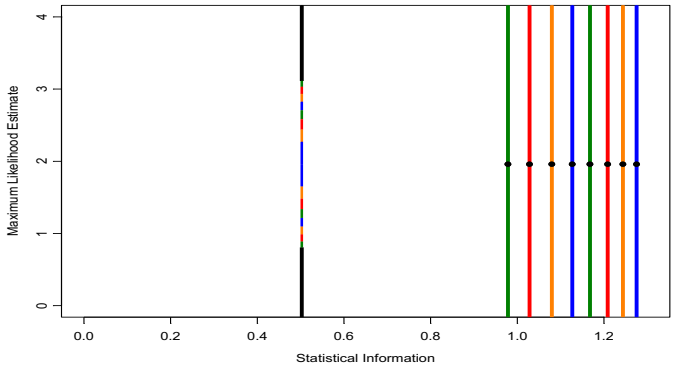
Inference: Major Issues

- Prespecification of adaptive plan is necessary to know sampling distribution
  - FDA CDER/CBER Guidance suggest prespecification is necessary for adequate and well-controlled studies
- How to generalize orderings used in GST to adaptive designs
  - MLE, LR straightforward
  - SW needs to consider multiple possible analysis times with empty continuation region
    - Use stage (creates many ties similar to MLE)
    - Use sample size instead of stage
    - Use weighted statistics instead of sufficient statistic
- Ordering specific to adaptive designs
  - Branath, Mehta, Posch (2009)

538

Prespecified Two-stage Adaptive Design

- Approximately optimal ASN under null and alternative
  - 0.44% more efficient than optimal two-stage GST (N_2 = 1.18)
  - ASN: GST 0.6854; Adaptive 0.6824



539

Inference on the Boundary by Ordering

- 95% CI for results on boundary at second analysis
  - Using sufficient statistic

N	MLE,SWm	SWz	AT	LR
0.98	0, 3.92	-0.11, 3.83	-1.87, 3.62	-0.11, 4.03
1.03	0, 3.92	-0.08, 3.86	-1.77, 3.66	-0.08, 4.00
1.08	0, 3.92	-0.05, 3.88	-1.66, 3.71	-0.05, 3.97
1.13	0, 3.92	-0.03, 3.90	-1.53, 3.75	-0.03, 3.95
1.17	0, 3.92	-0.01, 3.92	-1.39, 3.79	-0.01, 3.93
1.21	0, 3.92	0.01, 3.94	-1.20, 3.83	0.01, 3.91
1.24	0, 3.92	0.03, 3.96	-0.92, 3.86	0.03, 3.89
1.28	0, 3.92	0.04, 3.98	0.00, 3.92	0.04, 3.88

540

Control of Type 1 Error

- Methods to control the type 1 error for adaptive designs
  - Ignore the data and generate a random uniform
  - Use only the data gathered prior to adaptation
  - Use only the data gathered after adaptation
  - Most popular: Preserve conditional error function from some fixed sample or group sequential design
  - (Prespecification and proper sampling distribution)

541

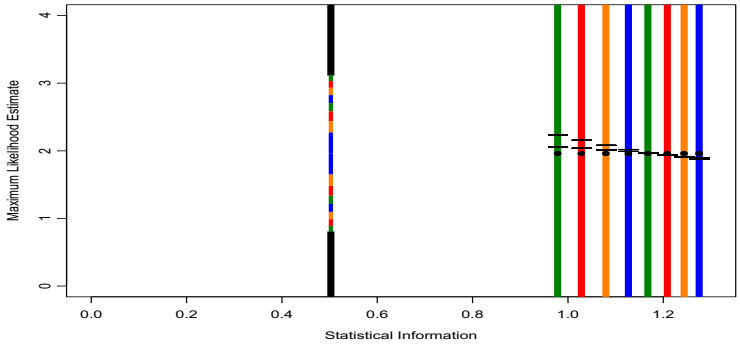
When Stopping Rules Not Pre-specified

- Methods to control the type 1 error for fully adaptive designs
  - Most popular: Preserve conditional error function from some fixed sample or group sequential design
    - Note the impact this would have on pre-specified inference
  - Alternative: Reweighting of the information accrued at different stages
- Methods to compute bias adjusted estimates and confidence intervals not yet well-developed
  - Have to anticipate what adaptations performed under other results

542

Two-stage Fully Adaptive Design

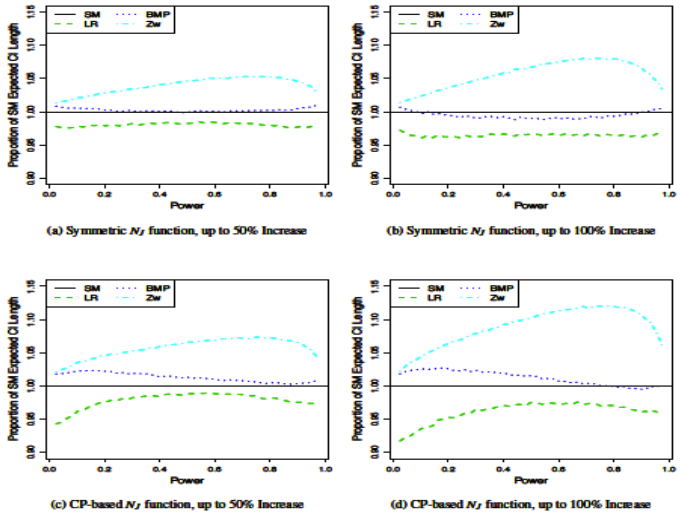
- Critical values determined using conditional error function
  - Loss of power relative to use of sufficient statistic:
  - Maintain N: Power 0.9744 vs 0.975
  - Maintain power: Inflate N by 0.902% to achieve ASN 0.6852
    - cf: Group sequential design had ASN of 0.6854



543

Comparison of CI Precision

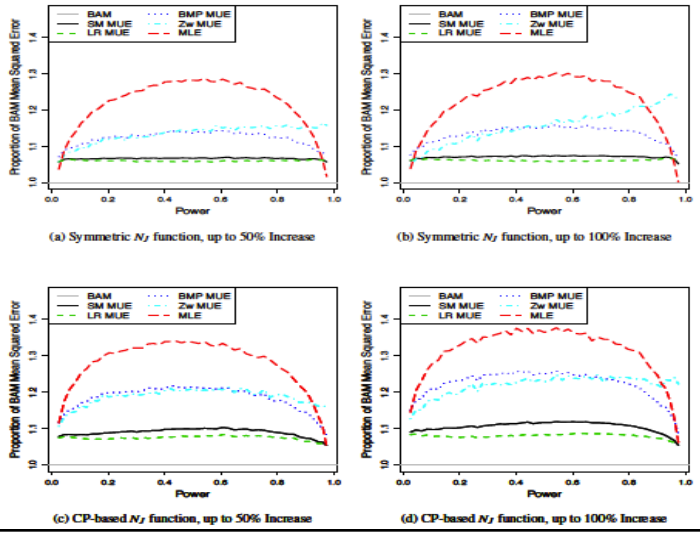
- O'Brien-Fleming conditional error function



544

Comparison of MSE of Estimates

- O'Brien-Fleming conditional error function



(a) Symmetric N_2 function, up to 50% Increase

(b) Symmetric N_2 function, up to 100% Increase

(c) CP-based N_2 function, up to 50% Increase

(d) CP-based N_2 function, up to 100% Increase

545

Enrichment, Dropping Arms

- When using adaptive designs to drop subgroups or to drop doses or treatments, the indication that is being investigated is a random variable
- CI can be computed conditionally on the "winning" indication
  - Sampling distribution depends on full dose response
  - Conservative coverage
    - (CI for dropped groups is infinite)

546

CDER/CBER Guidance on Adaptive Designs

- Recommendations for use of adaptive designs in confirmatory studies
  - Fully pre-specified sampling plans
  - Use well understood designs
    - Fixed sample
    - Group sequential plans
    - Blinded adaptation
  - For the time-being, avoid less understood designs
    - Adaptation based on unblinded data
- Adaptive designs may be of use in early phase clinical trials

547

Major Conclusions

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

548

Additional Resources

- www.RCTdesign.org
  - SS Emerson, DL Gillen, JM Kittelson, GP Levin, SC Emerson
- Software
  - Documentation
  - Tutorials
  - Extensions (Bayesian evaluation; adaptive design evaluation)
- Learning
  - Short courses
  - Research talks
  - Case studies
- Methodology
  - Technical reports on a variety of RCT-related topics

549