Statistical Monitoring of Clinical Trials

Ken Cheung

Columbia University

@

NINDS CTMC 2017, Iowa City, IA

Outline

1. Overview
2. Stopping boundaries for efficacy
3. Stopping for futility
4. Internal pilot & sample size re-estimation
5. Discussion

Overview

- In clinical trials designed to assess the efficacy and safety of medical intervention, evolving data are typically reviewed on a periodic basis during the conduct of the trial.
  - To make sure that the study participants are protected.
  - It is the investigator's ethical responsibility to study participants to monitor the trial in terms of safety and clinical benefits
- Pre-specified statistical plan
  - Avoid bias
  - Scientific integrity
  - Efficiency

Reasons for "Early Stopping"

REASONS FOR STOPPING A TRIAL EARLY THAN PLANNED

1.Risk monitoring (Behrman et al., 2011, NEJM): If the intervention is harmful to the patients the study should be stopped early to minimize the number of patients exposed to the intervention

2.Stopping for efficacy: If the intervention proves to be much superior to the control the study should be stopped early to maximize the number of patients who may benefit from the intervention

3.Stopping for futility: If the intervention is not beneficial the study should be stopped early to minimize the number of patients exposed to a not efficacious intervention

4.External evidence: If another intervention is proven to be beneficial while the trial is in progress the trial may be stopped and participant treated with the new intervention

Other types for monitoring

- Sample size adaptation: To ensure the study is adequately powered as a result of interim data on nuisance parameters (e.g. variance, placebo rate, etc).

- Treatment adaptation: Adaptive randomization; play-the-winner; drop-the-loser; etc.

Beta Blocker Heart Attack Trial (BHAT)

- Propanolol vs Placebo
- Post MI patients
- Primary endpoint: Mortality
- Statistical analysis: Log-rank test (Z-value); two-sided 5% significance
- Monitoring tools:
  - O'Brien Fleming (OBF) stopping boundary
  - Conditional probability

Beta-Blocker Heart Attack Trial (BHAT)

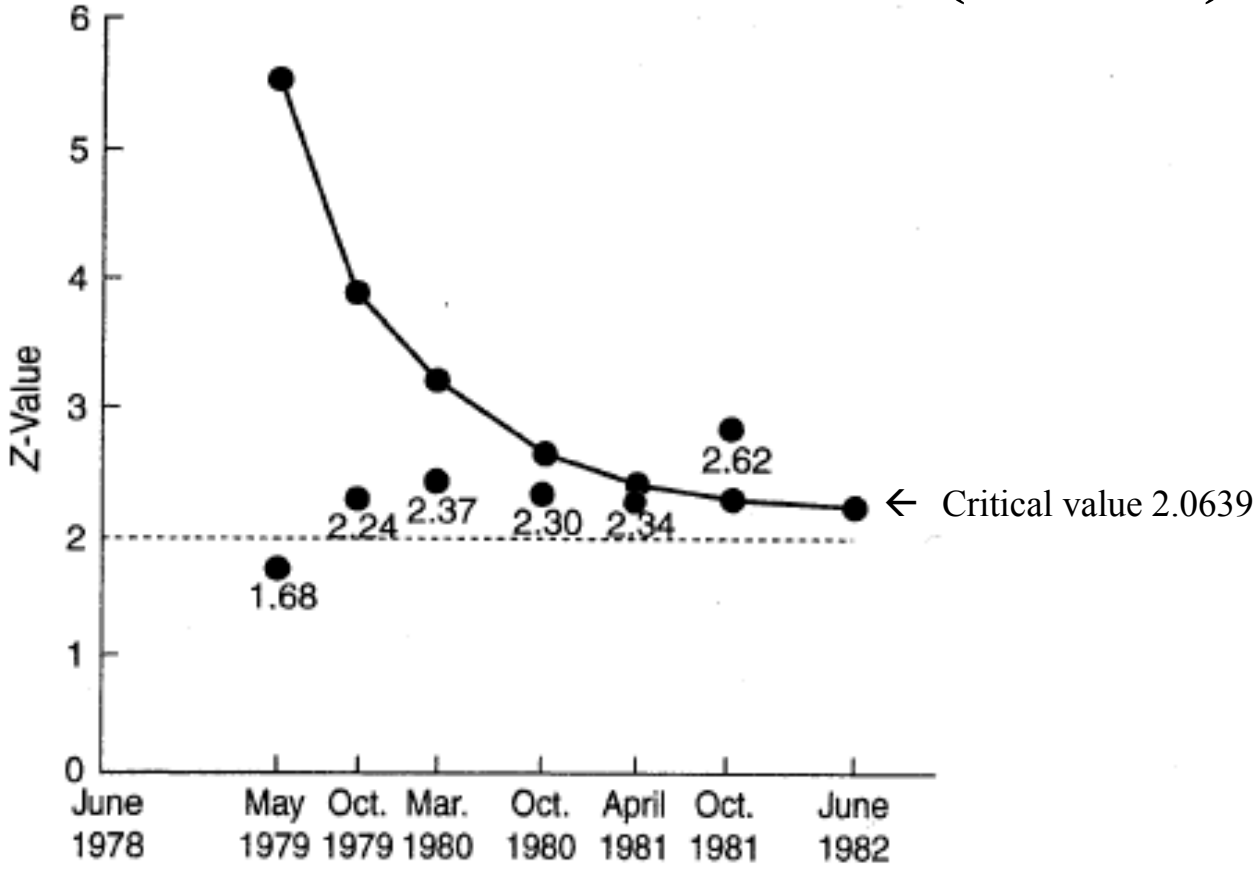


Figure 8.6 Interim mortality results for propranolol vs. placebo comparisons in the Beta-Blocker Heart Attack Trial (BHAT), using a one-sided 0.025 O'Brien–Fleming boundary. From DeMets *et al.*, 1984, copyright Elsevier Science.

- The DSMC made use of both the O'Brien-Fleming boundaries and conditional power.
- With one year FU remaining, the test statistic for comparing mortality crossed the boundary at the sixth of seven scheduled analyses
- Conditional power was used to determine the likelihood that the mortality difference would diminish or disappear, should the trial continue till its scheduled termination
- The probability was small and after thorough discussion the DSMC decided to terminate the trial.

Monitoring of efficacy

- Why not 1.96?
- Stopping boundaries
- Alpha spending function

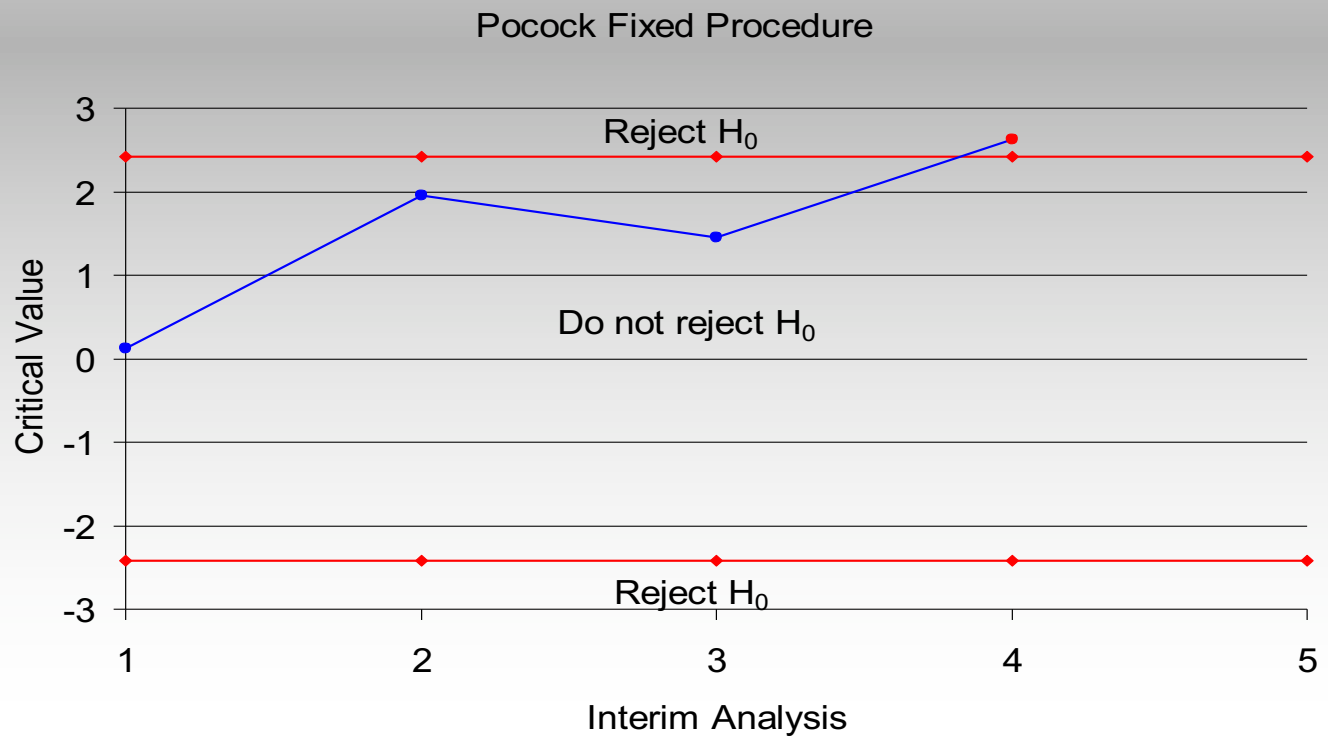
Stopping for efficacy

- Why not 1.96?
- Type I error: Probability of declaring a win under the null of no difference

# of looks	1	2	3	4	5	10	20
Type I error rate	0.050	0.083	0.107	0.126	0.142	0.193	0.248

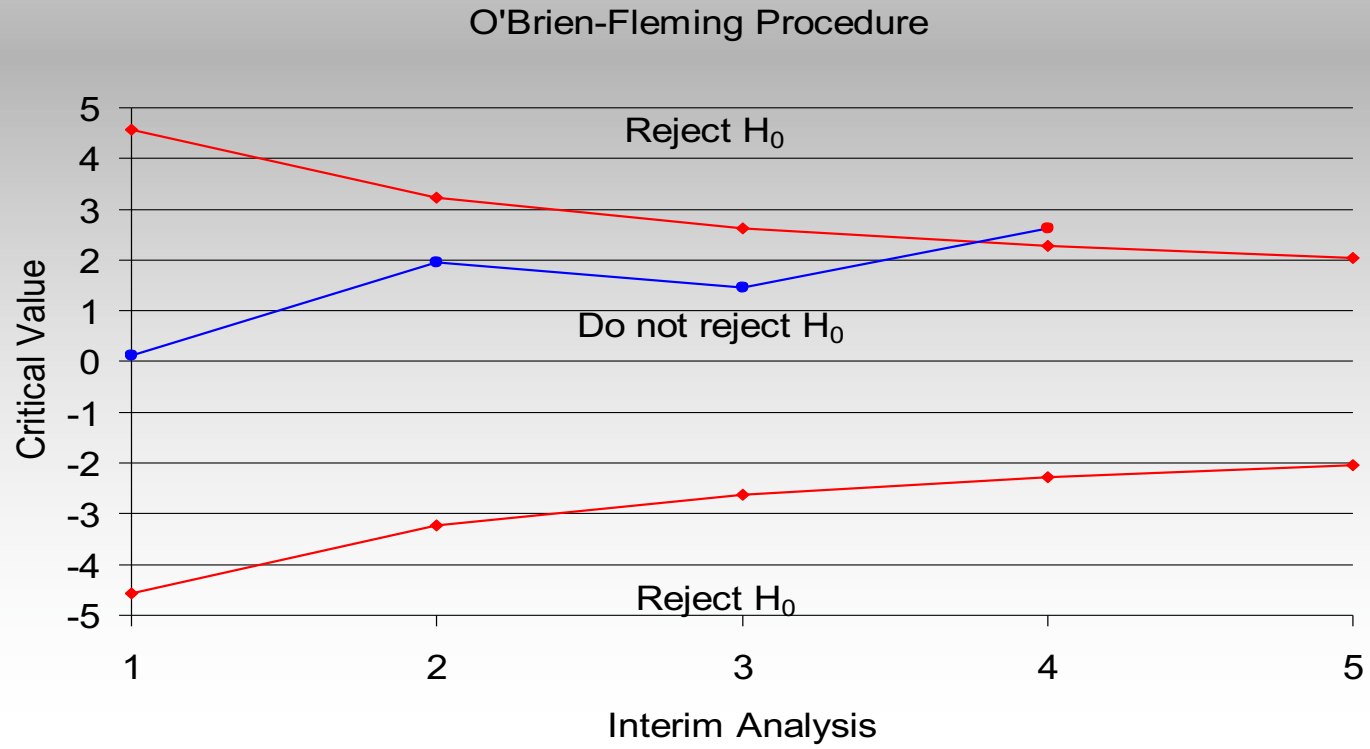
Stopping for efficacy

- Stopping boundaries
- Pocock: Same *adjusted* critical value at all analyses
  - E.g., K = 5 analyses, 5% significance: critical value = 2.413



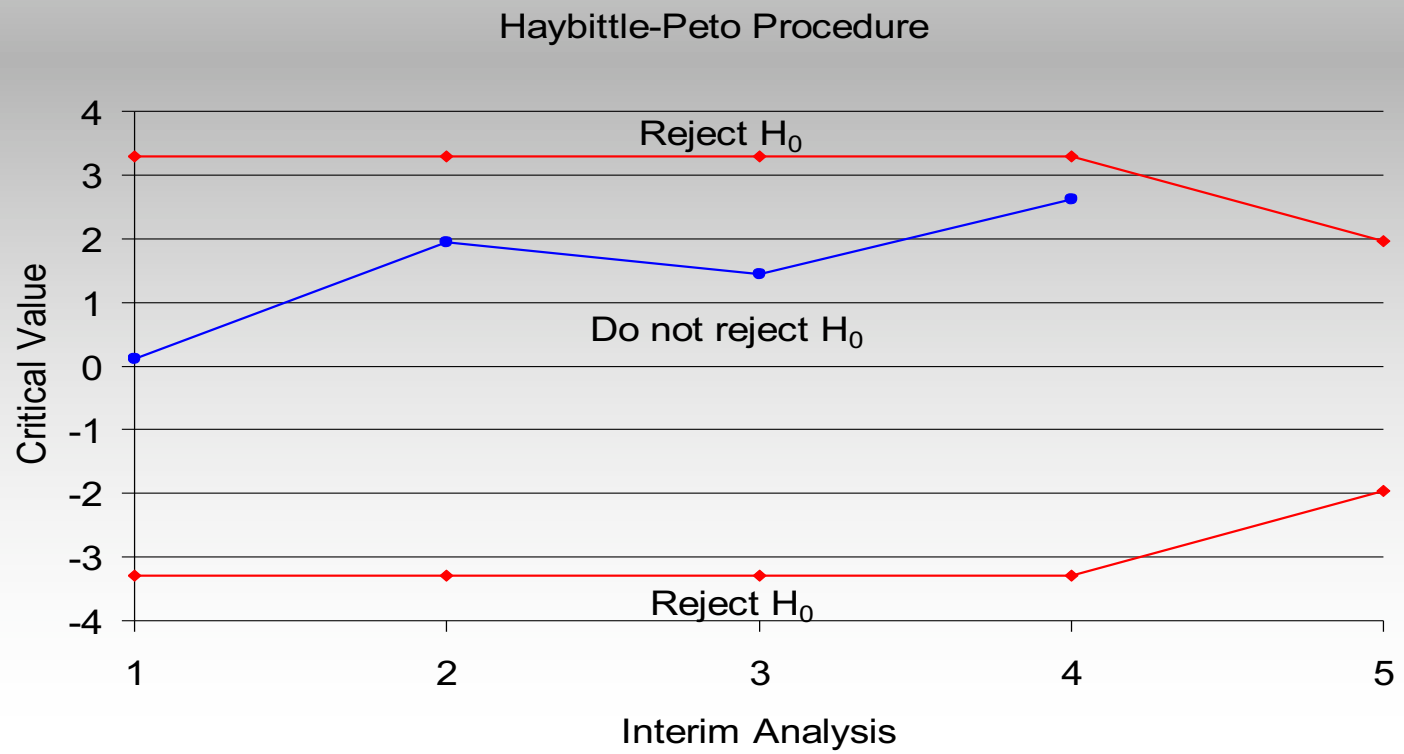
Stopping for efficacy

- Stopping boundaries
- OBF: Conservative at the beginning; e.g., K = 5 analyses, 5% significance: critical values = 4.562, 3.226, 2.634, 2.281, 2.040



Stopping for efficacy

- Stopping boundaries
- Haybittle-Peto



Stopping for efficacy

- Alpha spending function

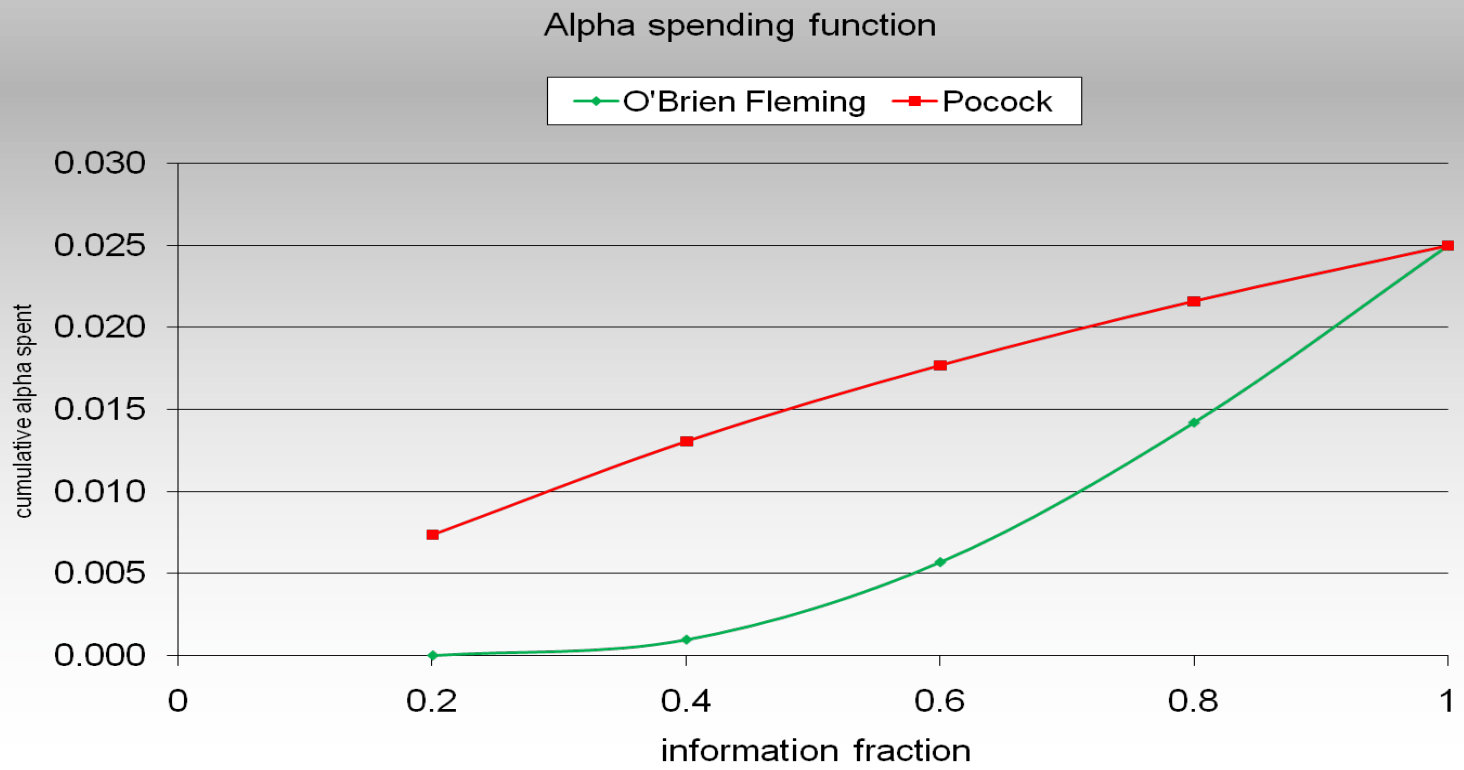
- Drawbacks of pre-specified stopping boundaries
  - Number of interim analyses are determined before the trial starts, and cannot be changed

  - The times of interim analyses are determined before the trial starts, and cannot be changed. Previous boundaries assume equal intervals (number of events/n) between interim analyses

Stopping for efficacy

- Alpha spending function $\alpha(t)$

Calculate the critical value at an interim analysis based on a pre-specified function how much alpha can be spent



Stopping for efficacy

- Alpha spending function $\alpha(t)$

Calculate the critical value at an interim analysis based on a pre-specified function how much alpha can be spent

At the k^{th} interim analysis, with *information fraction t_k:

The critical value c_k is calculated based on previous critical values and the alpha spending function

$$\Pr\{|S_k| > c_k, |S_l| < c_l, l = 1, \ldots, k-1\} = \alpha(t_k) - \alpha(t_{k-1}).$$

*Information fraction is the relative amount of information accumulated in the trial at the analysis. Information fraction = 1 at the end of the trial.

Stopping for efficacy

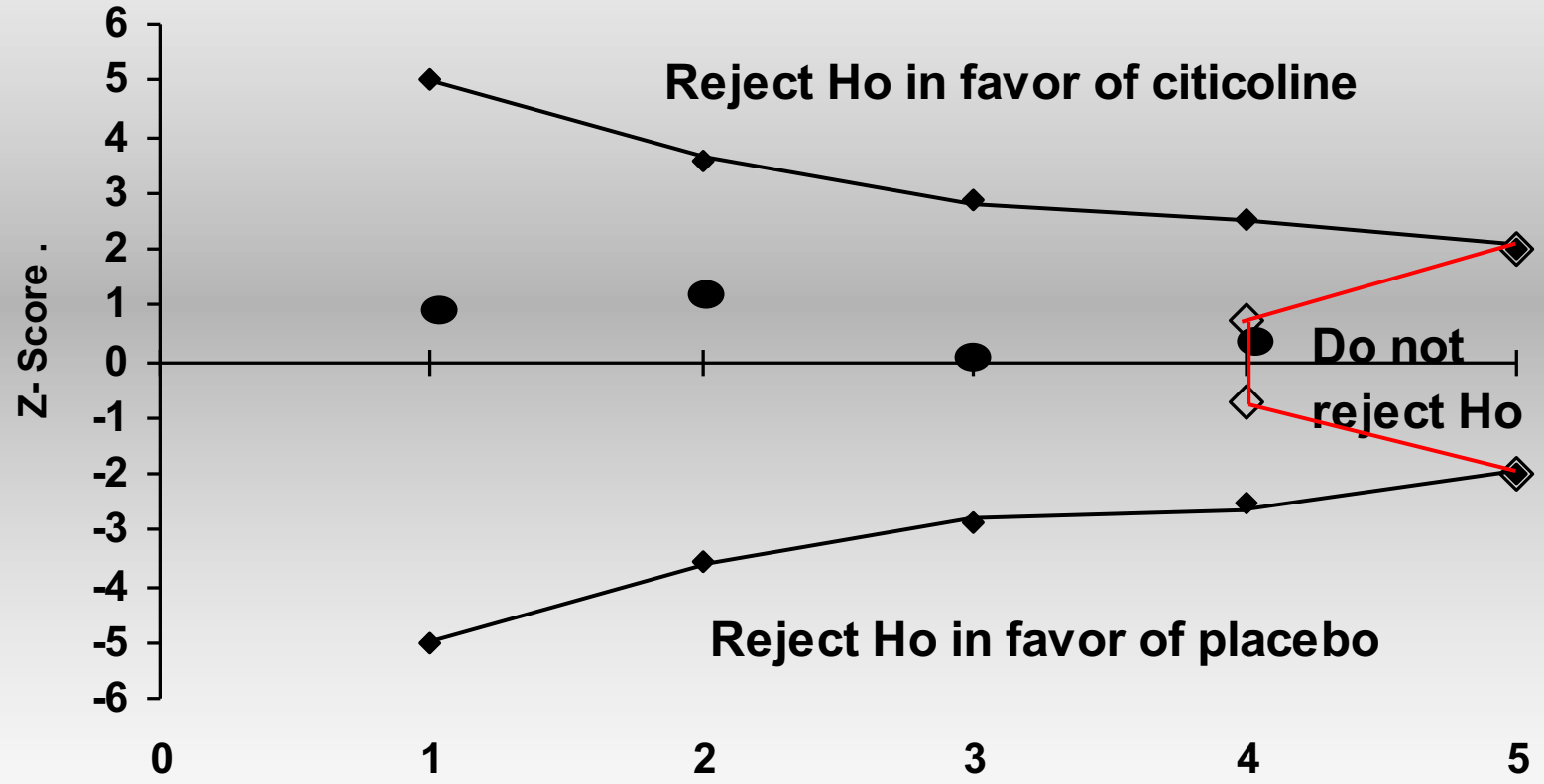
- Alpha spending function $\alpha(t)$

While alpha spending function allows unscheduled interim analyses, it is useful to plan a trial with a pre-specified number of interim analyses and schedule

Use alpha spending to calculate the actual critical values at the time of the analysis

Larger number of interim analyses generally results in larger critical values, and does not necessary stop a trial earlier

Stopping for futility



Stopping for futility

- Stopping for futility is primarily motivated by economic considerations
- In phase 3 trial, the futility boundaries are non-binding:
  - It does not affect how the efficacy boundaries are chosen
  - It does not increase type I error rate
  - The power will be slightly lower than without a futility boundary
  - Beta-spending function
- Alternative method: Conditional power = Probability of reaching statistical significance at the end given the current observations
- Different from "futility design", which is typically used in early phase trials to "prove" a treatment is not promising. Stopping for futility implies a high likelihood of inconclusiveness.

Futility interim analysis in early phase trial

- Example (Simon's two stage):
- *Stage 1:* Enroll $n_1 = 20$ subjects
- ***Futility interim analysis:***
  - Stop the trial and conclude *futility* if number of responses $S_{20} < 6$
- *Stage* 2 (if $S_{20} \geq 6$):
  - Enroll another $n_2 = 51$ subjects (i.e. **total n = 71**)
  - Conclude the treatment is good if $S_{71} \geq 24$; conclude futility otherwise
- Numerical values of the design were determined for 5% type I error under 25% response rate, and 80% power under 40% response.

Futility interim analysis in early phase trial

- When the treatment is inefficacious (25% response), there is over 60% probability the trial will end with 20 subjects.
- A fixed sample size design with no interim analysis will require n=62 subjects for the same statistical accuracy
- Useful when screening multiple new treatments with a concurrent control in a randomized fashion:
  - Randomize n_1 subjects to K treatments and control
  - Interim analysis: Stop the trial early if none of the treatments shows promise. Otherwise, continue with the most promising treatment and control to stage 2 with additional n_2 subjects

INTERNAL PILOT

Setting

- **Phase: 3**
- **Study endpoint:** Binary response (yes/no), e.g. 4-point change on NIHSS, mRS ≤ 2
- **Standard plan for design and analysis:**
  - Two-armed, placebo-controlled, double-blinded
  - 1:1 randomization
  - Z-test test at 5% significance and 80% power
- **Assumptions required for N calculation:**
  - Placebo rate ("nuisance" parameter)
  - Effect size (parameter of interest)
- **Design question:** N per arm

Power lowest when placebo rate near 50%

1:1 randomization
Two-sided test at 5% significance
Effect size = 10 percentage points

Placebo rate (assumed)	Treatment rate	N required per arm
10	20	219
20	30	314
30	40	376
40	50	408
50	60	408
60	70	376
70	80	314
80	90	219

Better-than-expected placebo rate

- If we assume 20% placebo rate at planning, and the actual rate is 40%, the study is underpowered

- If we assume 60% placebo rate at planning, but the actual rate is 70%, the study is overpowered (unnecessary $$$)

Internal Pilot

- Idea: Conduct an internal pilot to estimate the placebo rate in a blinded fashion.
- Example:
  - Effect size: 10 percentage points
  - Initial assume 10% placebo rate → N = 219 per arm
  - At N = 50 per arm, observe 35 responders; 35% of 100 subjects
  - Revised estimate of placebo rate with an effect size of 10 percentage point: 30% (drug response: 40%)
  - Revised N: 376 per arm (an additional 326)

Internal Pilot

- Not a tool to re-calculate N based on observed effect size
- Possible to account for variability in the interim placebo rate estimate
- Using the pilot data twice!
  - Data in the internal pilot will be used in the final analysis
  - Simulation: Type I error rate of unadjusted Z-test okay unless pilot n is small (Friede and Kieser, 2006)
  - Weighted analysis: View data from pilot and from the second stage as two independent strata. Slight loss in power

Internal pilot: weighted z-score

- The basic idea of weighted analysis is to *pre-specify* w_1 and w_2 so that $w_1 + w_2 = 1$
- *The weighted* Z-score is standard normal under null, i.e., type I error rate is preserved

$$Z = w_1^{1/2} Z_1 + w_2^{1/2} Z_2$$

Sample size re-adjustment

- Idea first appear in 1940s (Stein's two-stage)
- Sample size re-adjustment can in theory be extended to multiple stages using weighted analyses (Fisher, 1998) or P-value combination (Bauer and Kohne, 1994)
- Adjustments based on observed effect size have been proposed – not a great idea in general.
- Adjustments based on blinded estimate of nuisance parameters (e.g., pooled variance for continuous outcomes)

Discussion

- Number of interim analyses (adaptations):
  - Gain in early stopping greatest from one analysis (no interim) to two; diminishing benefits afterwards
  - Delayed outcomes
  - Practical risks: Unblinding; administrative burden
- Group sequential (alpha spending): Design with a maximum N, and stop "early" for efficacy/futility based on observed treatment effects
- Internal pilot: Design with a small pilot n, and expand (based on estimate of nuisance parameters)
- All stopping rules and adaptations need to be pre-specified at the planning stage.

References

Proschan, Lan, Wittes. Statistical Monitoring of Clinical Trials. 2006. Springer

Early phase trial monitoring

•Simon. Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials 1989; 10:1-10.

•Thall, Simon, Ellenberg. Two-stage selection and testing designs for comparative clinical trials. Biometrika 1988; 75:303-310.

•Cheung. Simple sequential boundaries for treatment selection in multi-armed randomized clinical trials with a control. Biometrics 2008; 64: 940-949.

Internal pilot/sample size re-calculation

•Bauer, Kohne: Evaluation of experiments with adaptive analyses. *Biometrics* 1994; 50, 1029—1041.

•Wittes, Brittain: The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; 65—72.