Adaptive Randomized Trial Designs

Michael Rosenblum
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

April 9-May 3, 2010

Course Notes and Readings at:
http://people.csail.mit.edu/mrosenblum/Teaching/adaptive_designs_2010.html

Lecture 1: Introduction, and Skeptics' Points of View

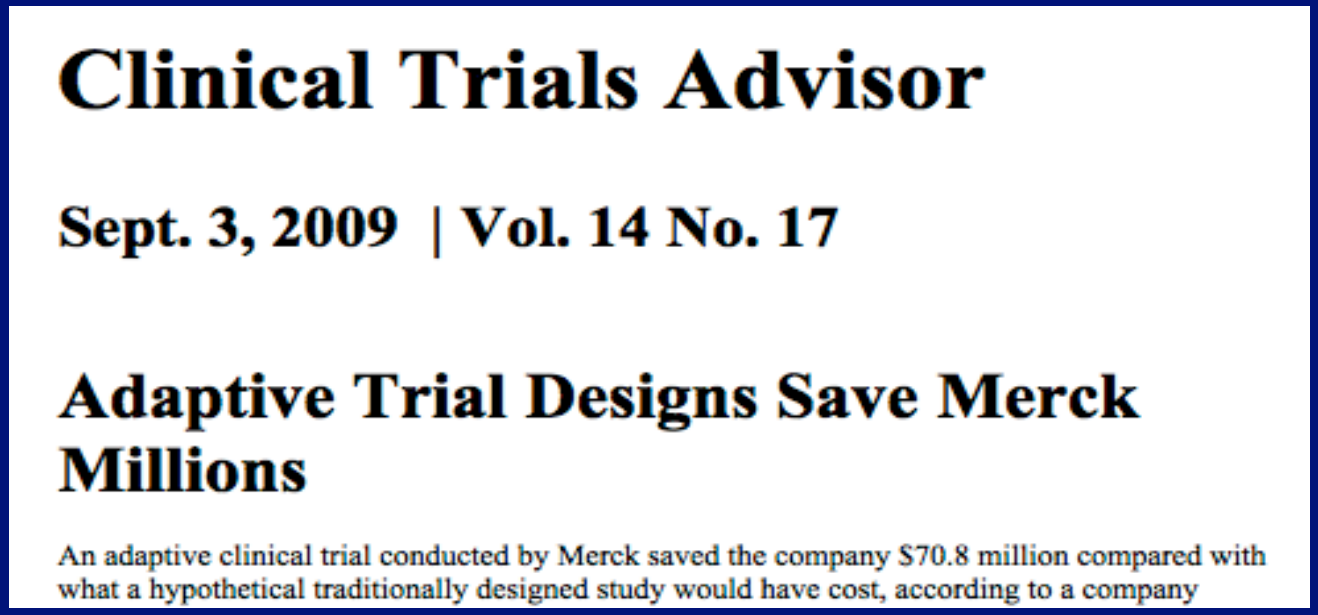
Adaptive Clinical Trial Designs

FDA is Interested:

"A large effort has been under way at FDA during the past several years to encourage the development and use of new trial designs, including enrichment designs."

Adaptive Clinical Trial Designs

- Pharmaceutical Companies are Interested:



Clinical Trials Advisor

Sept. 3, 2009 | Vol. 14 No. 17

Adaptive Trial Designs Save Merck Millions

An adaptive clinical trial conducted by Merck saved the company $70.8 million compared with what a hypothetical traditionally designed study would have cost, according to a company

"An adaptive clinical trial conducted by Merck saved the company $70.8 million compared with what a hypothetical traditionally designed study would have cost…"

Why Use Adaptive Designs?

Benefits:

- Can Give More Power to Confirm Effective Drugs and Determine Subpopulations who Benefit Most
- Can Reduce Cost, Duration, and Number of Subjects of Trials

Designs Must:

- Guarantee Correct Probability of False Positive Results (e.g. 0.05)
- Lead to Interpretable Results

Goals of Course

- Give an overview of adaptive randomized trial designs.
- Discuss the advantages, limitations, and open problems for various types of adaptation.

Course Outline

1. Introduction: Skeptics' Perspectives
2. FDA Draft Guidance on Adaptive Designs
3. Adapting Randomization Probabilities
4. Adapting Sample Size (e.g. early stop)
5. Seamless Phase II/III Designs
6. Adapting Hypothesis Tested
7. Bayesian Designs

Themes

- Prespecify Decision Rules for Making Adaptations
- Tradeoff between Flexibility and Power
- Tradeoff between Power, Sample Size, Number of Patients in Inferior Arm
- Perspective of FDA, pharma company, subject in a trial

Group Sequential Randomized Trial Designs

- Participants Enrolled over Time
- At Interim Points, Can Change Sampling in Response to Accrued Data:
  - Can Stop Trial Early (e.g. for Efficacy, Futility, or Safety)
  - Can Change Probability of Assignment to Different Arms (e.g. to Maximize Number of Patients Assigned to Best Arm)
  - Can Recruit from Subpopulation in which Treatment Effect is Strongest ("Enrichment")

Example

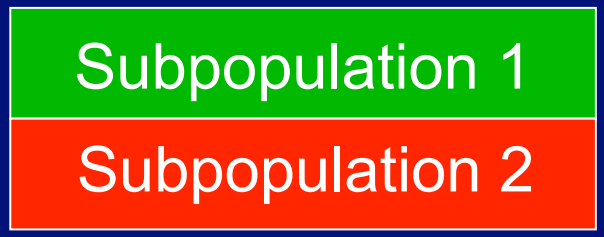
Population: Lung cancer patients with metastasis. Some are eligible for existing therapy, some are not.

Research Questions: Does addition of a new therapy improve mean outcome for total population? For those who are not eligible for existing therapy?

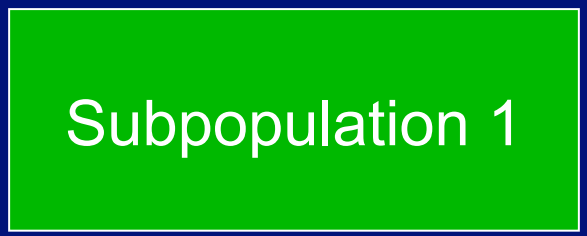
Prior Data Indicates: Treatment effect greatest for those not eligible for existing therapy.

Some Possible Fixed Designs

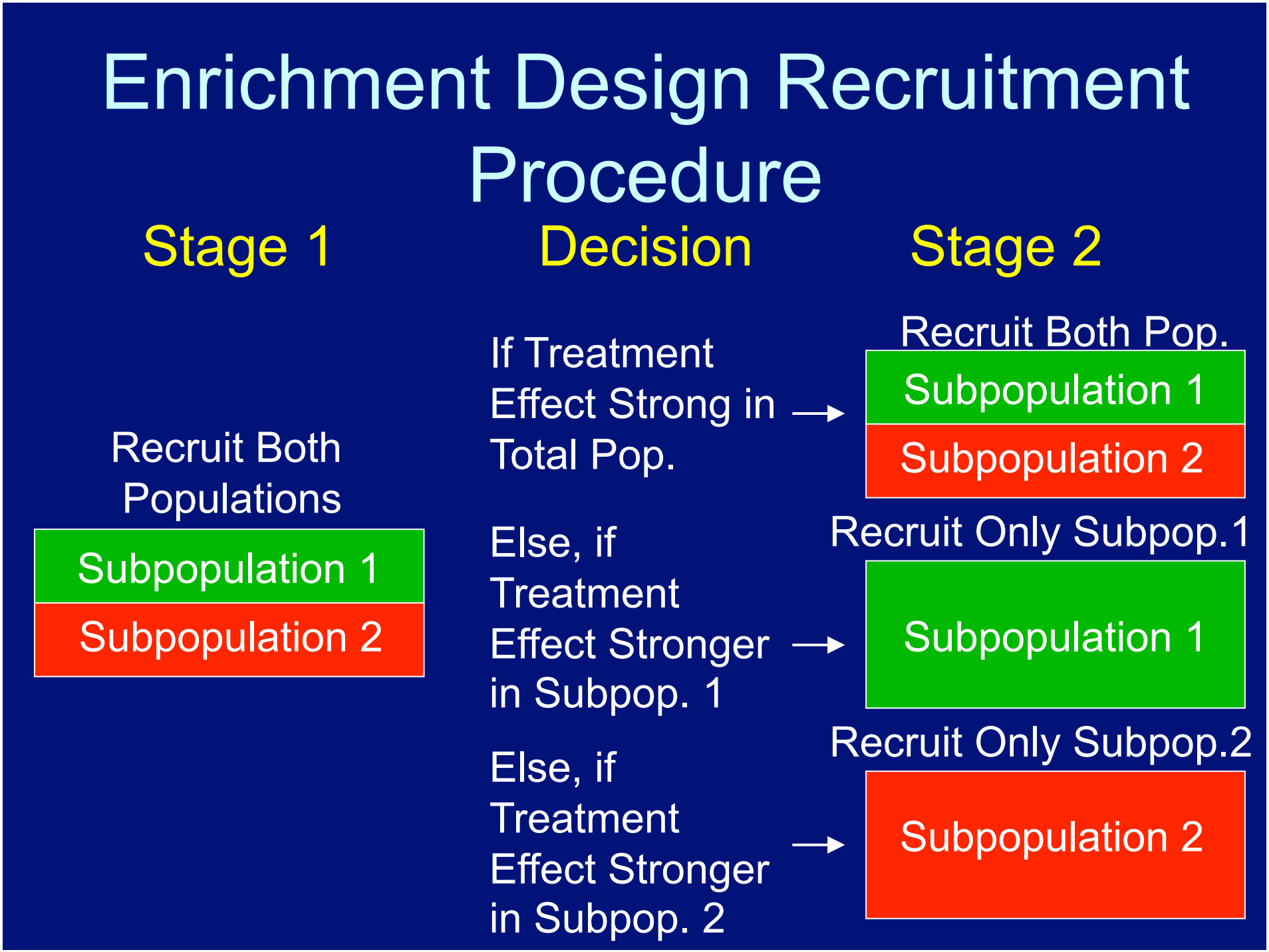
- Enroll from total population (both those eligible for existing treatment and those not)



- Enroll only from those not eligible for existing treatment



Enrichment Design Recruitment Procedure



FDA Critical Path Opportunities

"Advancing Innovative Trial Designs"

34. Design of Active Controlled Trials.

35. Enrichment Designs. If biomarkers can reliably identify individuals with a high probability of response to a therapy, trials could focus on such patients.

FDA Critical Path Opportunities

36. Use of Prior Experience or Accumulated Information in Trial Design.

"Consensus and clarification is needed on questions such as:

- When can extra trial arms be dropped?
- When can an early marker be used to choose which treatment to carry forward or to choose a subset for analysis?

FDA Critical Path Opportunities

"Consensus and clarification is needed on questions such as: (con't)

- When is it valid to modify randomization based on results, for example, in a combined phase 2/3 cancer trial?
- When is it valid and under what situations can one stage or phase of a study be combined with the second stage or phase?

A Skeptic

- Fleming (2006) Standard versus adaptive monitoring procedures: A commentary
- Issues:
  - Efficiency
  - Interpretability
  - Reliability of Interim Results
  - Leaking Information
  - Ethical Concerns

A Skeptic: Fleming

Issue of Efficiency:

Some adaptive sample size adjustment methods are inefficient, as they don't use sufficient statistics.

For example, Cui et al. (1999) method allows arbitrary change to sample size after interim analysis, but fixed weights on each stage's Z statistics.

E.g. final Z-statistic = $(Z_1+Z_2)/\sqrt{2}$.

A Skeptic: Fleming

Issue of Efficiency:

Some adaptive sample size adjustment methods are inefficient, as they don't use sufficient statistics.

However, some designs, e.g. response adaptive randomization that targets Neyman allocation, are more efficient than non-adaptive design.

A Skeptic: Fleming

Issue of Interpretability:

- Estimates of treatment effect will be biased if e.g. stop early.

Can make correction, by downweighting if stop early.

A Skeptic: Fleming

Issue of Reliability of Interim Results:

May be misled into making a poor adaptation decision by highly variable early results (due to low sample size at interim analysis).

True, but focus should be overall operating characteristics of trial.

A Skeptic: Fleming

Issue of Leaking Information:

Prejudgment of unreliable results based on limited data “could adversely impact patient accrual, continued adherence to trial regimens, and ability to obtain unbiased and complete assessment of trial outcome measures.”

A Skeptic: Fleming

Ethical Issues:

Will patients understand risks/benefits in complex design?

Semi-skeptics: Wittes and Lachenbruch

Wittes, J., and Lachenbruch, P. (2006) Discussion: Opening the Adaptive Toolbox.

Issues:

- Adaptive designs may be used as excuse to be lazy in planning a trial.
-Adapting based only on nuisance parameters.
-Internal vs. external information.

Semi-skeptics: Wittes and Lachenbruch

Wittes, J., and Lachenbruch, P. (2006) Issue that adaptive designs may be used as excuse to be lazy in planning a trial.

Companies may want to fund small trial, and then extend if it looks promising (since can argue for e.g. more venture capital money).

Could lead to sample size larger than a well-planned fixed trial.

Semi-skeptics: Wittes and Lachenbruch

Wittes, J., and Lachenbruch, P. (2006)

Issue of adapting based only on nuisance parameters.

Certain nuisance parameters, such as the variance for continuous outcomes, can be used to calibrate sample size without fear of inflated type I error.

Semi-skeptics: Wittes and Lachenbruch

Wittes, J., and Lachenbruch, P. (2006)

Issue of internal vs. external information.

Can make adaptation based on external information (e.g. results from a separate trial) without fear of increased Type I error.

Bias Due to Early Stopping

Lecture 2a: FDA Draft Guidance on Adaptive Designs

FDA Draft Guidance on Adaptive Designs

Focus is AW&C (adequate and well-controlled) trials.

Distinguishes well understood vs. less well understood adaptations.

Explains chief concerns: Type I error, bias, interpretability.

FDA Draft Guidance on Adaptive Designs

Well Understood Adaptations:

- **Adapt Study Eligibility Criteria Using Only Pre-randomization data.**
- **Adapt to Maintain Study Power Based on Blinded Interim Analyses of Aggregate Data (or Based on Data Unrelated to Outcome).**
- **Adaptations Not Dependent on Within Study, Between-Group Outcome Differences**

FDA Draft Guidance on Adaptive Designs

Well Understood Adaptations:

- **Group Sequential Methods (i.e. Early Stopping)**

FDA Draft Guidance on Adaptive Designs

Less-Well Understood Adaptations:

- **Adaptive Dose Selection**
- **Response-Adaptive Randomization**
- **Sample Size Adaptation Based on Interim-Effect Size Estimates**
- **Adaptation of Patient Population Based on Treatment-Effect Estimates**
- **Adaptive Endpoint Selection**

FDA Draft Guidance on Adaptive Designs

Adaptive Dose Selection

Dropping Doses (Arms).

Use of biomarker for dose selection.

[Need statistical adjustment.]

FDA Draft Guidance on Adaptive Designs

Response Adaptive Randomization

Population being enrolled may change over time (e.g. more events observed).

This could cause inflated Type I error and bias.

FDA Draft Guidance on Adaptive Designs

Adaptation of Patient Population Based on Treatment-Effect Estimates

"These designs are less well understood, pose challenges in avoiding introduction of bias, and generally call for statistical adjustment to avoid increasing the Type I error rate."

FDA Draft Guidance on Adaptive Designs

Guide to reporting simulations (pp. 38-39):

Investigate Type I error, power, bias, under variety of data generating distributions.

Compare to fixed designs.

Not sufficient to show Type I error controlled via simulations.

Interesting question: what is best time to do adaptations? Early vs. later?

Lecture 2b: Intro to Group Sequential Testing

Sequential Design, Adaptive Sample Size

Overview

Advantages: May be able to stop early if strong signal of treatment effect.

Can ensure adequate power by accruing enough data before doing hypothesis test.

Interim analysis times can be function of “information” accrued.

Disadvantage: If don’t stop early, need more subjects than in equivalent trial with no early stopping allowed. Biased estimates.

Sequential Testing (Early Stopping)

At prespecified interim analyses, do a test, and possibly stop the trial for efficacy or futility.

Advantage: May be able to stop early if strong signal of treatment effect. Interim analysis times can be function of "information" accrued.

Disadvantage: If don't stop early, need more subjects than in equivalent trial with no early stopping allowed. Biased estimates.

Simple Example: Static Design

[From Jennison and Turnbull (2000), Ch.2]

Two arm trial, ½, ½ randomization.

Responses are $N(\mu_T,\sigma^2)$, $N(\mu_C,\sigma^2)$.

Null Hypothesis: $\mu_T=\mu_C$.

Want Type I Error at most 0.05.

Want Power = 0.9 at alternative: $\mu_T-\mu_C=1$.

Assume $\sigma^2=4$. Then need in each arm:

$$n \approx 2 \times 4 \times \frac{[\Phi^{-1}(0.975)+\Phi^{-1}(0.9)]^2}{[1-0]^2} = 84.1$$

Simple Example: Seq. Design using Pocock Boundaries

At interim analyses, stop and reject null if Z-statistic exceeds Pocock cutoffs.

Consider 5 equally spaced interim analyses.

Cutoff is 2.41 at all interim analyses.

(Had it been 1.96, Type I error would be 0.14.)

What is max. sample size needed?

102 (> 84).

Pocock Stopping Boundaries

At alpha = 0.05, 2-sided, Z-statistic cutoffs:

Number Analyses	Pocock Boundary
1	1.96
2	2.18
3	2.29
5	2.41
10	2.56

Simple Example: Seq. Design, O'Brien-Fleming Boundaries

At interim analyses, stop and reject null if Z-statistic exceeds O'Brien-Fl. cutoffs.

Consider 5 equally spaced interim analyses.

Cutoffs are 4.56, 3.23, 2.63, 2.28, 2.04.

What is max. sample size needed?

86 (> 84).

O'Brien-Fleming Stopping Boundaries

At alpha = 0.05, 2-sided, Z-statistic cutoffs

Number Analyses	O'Brien Fleming Boundaries
1	1.96
2	2.80, 1.98
3	3.47, 2.45, 2.00
5	4.56, 3.23, 2.63, 2.28, 2.04

Max. Sample Size vs. Static Design

How much is max. sample size "inflated" in sequential testing vs. fixed design? R:

Number Interim Analyses	Pocock boundar.	O'Brien-Fleming
1	1	1
2	1.100	1.007
3	1.151	1.016
5	1.207	1.026

Expected Sample Size vs. Static Design

How does Expected Sample Size in sequential testing compare to fixed design, at alternat.?

Number Interim Analyses	Pocock boundar.	O'Brien-Fleming
1	1	1
2	0.78	0.85
3	0.72	0.80
5	0.69	0.75

Expected Sample Size vs. Static Design

How does Expected Sample Size in sequential testing compare to fixed design, at null.?

Number Interim Analyses	Pocock boundar.	O'Brien-Fleming
1	1	1
2	1.08	1.01
3	1.13	1.01
5	1.18	1.02

Pocock vs. O'Brien-Fleming

Pocock more aggressive earlier, but larger max. sample size, and larger sample size variability. Better when true treatment effect relatively large, but worse otherwise.

Consider treatment of rare disease, subjects enter study at 40/year. Max duration is:

4.25 years for static design

4.5 years for O'Brien-Fleming

5.25 years for Pocock

Flexible Single Testing Time based on Information Accrued

Prespecify that trial will continue until a certain information level (I_{max}) is achieved, at which time a test will take place.

$$I_{\max} = \frac{[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2}{[\psi_{alt} - \psi_0]^2} = n/\sigma^2 .$$

Type I error (asymptotically) controlled.

Flexible Interim Analysis Times based on Information Accrued

Interim analysis times based on information accrued I(n).

E.g., if outcome binary:

$$I(n) = \frac{1}{Var_n(\hat{p}_A - \hat{p}_B)} \approx \frac{n}{\hat{p}_n(1-\hat{p}_n)}.$$

Interim analysis when information equals: e.g. ½ of

$$I_{\max} = R\frac{[\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)]^2}{[\psi_{alt}-\psi_0]^2}.$$

Lecture 3: Adapting Randomization Probabilities

Adapting Randomization Probabilities

Q: Why adapt the randomization probabilities?

A: To get more power and precision.

Adapting Randomization Probabilities

Q: How does adapting rand. Probabilities (potentially) give more power and precision?

A:

1. Improving balance on prognostic covariates (Covariate-adaptive designs)
2. Sampling more from population with greater variance in outcome (Response-adaptive designs)

Covariate Adaptive Designs

Methods to improve balance of prognostic covariates (compared to simple randomization):

1. Block randomization
2. Block randomization stratified by prognostic covariates
3. Biased-coin designs (bias randomization prob. of future subjects to correct observed imbalance)
4. Minimization (of a measure of imbalance)

Adapting Randomization Probabilities

Block randomization:

E.g. in blocks of 4 envelopes, with 2 “treatment” envelopes and 2 “control” envelopes. Overall balance can be off by at most 2!

Block randomization stratified by prognostic covariates

E.g. blocks of 4 envelopes for each stratum of prognostic covariates.

Balance in each stratum off by ≤ 2.

Adapting Randomization Probabilities

Biased coin:

Idea is to select randomization probability for each subject "biasing" toward balance.

E.g. Efron's biased coin: if more than ½ of subjects so far are in treatment group, then next subject gets prob. $p > ½$ of being in control group, and vice versa.

If $p = 1$, then this is example of minimization.

Adapting Randomization Probabilities

Biased coin designs for covariate adaptation:

1. Zelen's model: if imbalance in next subject's covariate stratum > 2, then deterministically assign to improve balance. Else assign with p = 1/2.
2. Pocock-Simon model: based on weighted combination of imbalances in each covariate stratum (with bigger weight for more important covariates), use p-biased coin to improve balance.

Adapting Rand. Probabilities

Friedman-Wei urn:

Wei's urn model: start with urn having

k red (treatment) and k white (control) balls. Draw one and assign to that arm, and replace it and also add b balls of opposite color. Repeat.

For covariate adaptation:

One urn for each covariate value. Draw from most unbalanced urn as above, and now add b opposite balls to each urn corresponding to that subject's covariate values.

Response Adaptive Randomization

Play the winner rules:

Deterministic version: if last patient outcome is “success,” assign that treatment again; else assign other treatment.

Randomized version:

Use an urn of course! Draw from urn for treatment assignment. If got treatment A and “success,” then add b Type A balls; else add b type B balls.

Response Adaptive Randomization

Play the winner rules:

Randomized version:

Use an urn of course! Draw from urn for treatment assignment. If got treatment A and "success," then add b Type A balls; else add b type B balls.

Properties: ratio of number assigned to

A vs. B converges to $(1-p_B) / (1-p_A)$, for

p_A, p_B the success probabilities.

Response Adaptive Randomization

Play the winner rules:

Randomized version:

Use an urn of course! Draw from urn for treatment assignment. If got treatment A and "success," then add b Type A balls; else add b type B balls.

Properties: ratio of number assigned to

A vs. B converges to $(1-p_B) / (1-p_A)$, for

p_A, p_B the success probabilities.

Response Adaptive Randomization

Neyman Allocation:

How should allocation be done to get most power at a given sample size, when the final estimator/test based on estimated risk difference?

Intuitively, want to sample more from arm with larger variance. Neyman allocation:

$$n_A / n_B = \sqrt{\frac{p_A q_A}{p_B q_B}}.$$

Response Adaptive Randomization

Where does Neyman allocation come from?

Asymptotic variance of empirical risk difference: $\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}.$

To minimize it subject to total sample size = n: $n_A + n_B = n,$

use simple calculus to get: $n_A / n_B = \sqrt{\frac{p_A q_A}{p_B q_B}}.$

Response Adaptive Randomization

"Ethical" allocations:

How should allocation be done to minimize expected number of failures subject to power constraint?

Intuitively, want to sample more from arm with larger success probability.

"Ethical" allocation:

$$n_A / n_B = \sqrt{\frac{p_A}{p_B}}.$$

Lecture 4: Adapting the Hypothesis Tested

Testing Multiple Hypotheses

Designs that allow interim treatment selection, change of endpoint(s), or change of population sampled, all implicitly involve multiple testing.

We want designs to control the familywise Type I error, that is, the probability of rejecting one or more true null hypotheses.

Testing Multiple Hypotheses

For example, if two possible endpoints (e.g. death, MI or death), then implicitly testing null hypotheses corresponding to each.

Another example: in “seamless design,” if start with 5 treatments in Phase II and select one to continue in Phase III, then there are 5 hypotheses being tested (even though can reject at most 1).

Combination Tests

Given k null hypotheses $H_{01},\ldots, H_{0k}$, consider all possible intersection null hypotheses:

$$\bigcap_{i\in J} H_{0i}$$

For example, if H_{0i} is null that treatment has no effect in subpopulation i, then

is null that treatment has no effect in either of the subpopulations 1 or 2.

$$\bigcap_{i\in\{1,2\}} H_{0i}$$

Combination Tests

Interpreting rejection of combination tests:

If intersection null hypothesis $\bigcap_{i \in \{1,2\}} H_{0i}$ is false, this means at least one of null hypotheses 1 and 2 is false.

Hard to interpret—you really want to test each individual null hypothesis.

But combination tests important for controlling Type I error, as we'll see.

Note, intersection null hypothesis is "stronger" than the individual hypotheses.

Examples of Combination Tests

To test $\bigcap_{i\in\{1,2\}} H_{0i}$, can, for example:

1. Use Bonferroni: reject if $\min\{p_1,p_2\} < \alpha/2$.
2. Focus on just one of the hypotheses: reject if $p_1 < \alpha$.
3. If p-values independent, then can use weighted inverse normal method: reject if

$$\frac{Z_1}{\sqrt{2}} + \frac{Z_2}{\sqrt{2}} > 1.645$$

Method must be prespecified!!

Closure Principle

If H_{01} is null of no mean treatment effect in men, and H_{02} is null of no mean treatment effect in women, then $\bigcap_{i \in \{1,2\}} H_{0i}$ is null of no mean treatment effect in either of these two groups.

A "local test" is a level α test of an intersection null hypothesis. For example, a t-test including all men in the study is a local test of H_{01}.

Closure Principle

If H_{01} is null of no mean treatment effect in men, and H_{02} is null of no mean treatment effect in women, then $\bigcap_{i \in \{1,2\}} H_{0i}$ is null of no mean treatment effect in either of these two groups.

A "local test" of $\bigcap_{i \in \{1,2\}} H_{0i}$ could be, for example, a weighted combination of t-test within men, and t-test within women.

Closure Principle

Closure principle:

1. Run local test for each intersection null hypothesis $\bigcap_{i \in J} H_{0i}$.

2. For each original null hypothesis H_{0i}, reject it if for all sets J containing i, the local test rejected.

This guarantees familywise Type I error is correct (e.g. at most 0.05).

Closure Principle

Example: H_{01} is null of no mean treatment effect in men, and H_{02} is null of no mean treatment effect in women.

We prespecify local tests of H_{01}, H_{02}, and $\bigcap_{i \in \{1,2\}} H_{0i}$

E.g. individual nulls based on within group t-statistics; intersection based on weighted inverse normal combination test.

Reject individual null iff BOTH individual local test and combination test reject.

Closure Principle

Example: Thall, Simon, Ellenberg (1988) treatment selection design.

In Phase IIb, randomize subjects to k treatments and placebo. So we have

k null hypotheses.

In Phase III, randomize subjects to treatment that did best (largest t-stat.) in Phase IIb.

Final test-statistic uses all data for the chosen treatment, with penalty calculated under the global null to control Type Ierror.

Closure Principle

Example: Thall, Simon, Ellenberg (1988) treatment selection design.

Consider 2 treatments in first stage, and pick the “winner” treatment for second stage.

If we simply combine all data and do t-test, we would inflate the Type I error.

Need to penalize with final cutoff that gives

P(Reject) = 0.05 under global null that both treatments do nothing.

Closure Principle

Example: Thall, Simon, Ellenberg (1988) treatment selection design.

But does this control familywise Type I error?

E.g. what if one treatment positive effect, and the other is 0 effect—what's the probability that we select the ineffective treatment and reject the null?

Closure principle gives way to formally prove control of familywise Type I error.

Closure Principle

Example: Thall, Simon, Ellenberg (1988) treatment selection design.

Define "local test" for any intersection null hypothesis $\bigcap_{i \in J} H_{0i}$ as follows:

If i* (selected treatment) not in J, then fail to reject. Else, p-value is that corresponding to t-test combining all data on treatment i*, with cutoff set so that local test controls Type I error under global null.

Closure Principle

Example: Thall, Simon, Ellenberg (1988) treatment selection design.

Define "local test" for any intersection null hypothesis $\bigcap_{i \in J} H_{0i}$ as follows:

I.e. let test statistic for this intersection null be

$$(S_{1,i^*} + S_{2,i^*})/n$$

if i* in J. This can be prespecified equivalently as $(\max_{i \in J} S_{1,i} + S_{2,i^*})/n$ if i* in J.

“Pedagogic” Example

Two stages, 4 treatments for asthma. In Phase IIb, 100 observations per treatment (and placebo). Phase IIb results are:

	Control	**Tx. 1**	**Tx. 2**	**Tx. 3**	**Tx. 4**
n	100	100	100	100	100
P-value		0.2	0.04	0.05	0.03
Z-score		0.84	1.75	1.64	1.88

“Pedagogic” Example

Choose treatment 4 for Phase III trial.

	Control	Tx. 1	Tx. 2	Tx. 3	Tx. 4
n	100	100	100	100	100
P-value		0.2	0.04	0.05	0.03
Z-score		0.84	1.75	1.64	1.88

“Pedagogic” Example

Phase III results:

	Control	Tx. 4
n	500	500
P-value		0.04
Z-score		1.75

Compare 3 approaches at 2-sided $\alpha=0.05$.

Conventional approach (ignore Phase IIb data in final test), TSE design, Bauer Kohne design.

“Pedagogic” Example

Phase III results:

	Control	Tx. 4
n	500	500
P-value		0.04
Z-score		1.75

Conventional Approach: fails to reject since p-value 0.04 more than 0.025.

“Pedagogic” Example

Phase III results:

	Control	Tx. 4
n	500	500
P-value		0.04
Z-score		1.75

TSE approach: combines data from both stages and uses sufficient Z-statistic, which equals 2.365. This exceeds “penalized” critical value 2.20, so reject.

“Pedagogic” Example

Bauer and Kohne combination test approach:

Compute p-value for each intersection null hypothesis J by combining both stages’ p-values:

Stage 1 p-value: $p_{1,J} = |J| \min_{i \in J} p_{1,i}$.

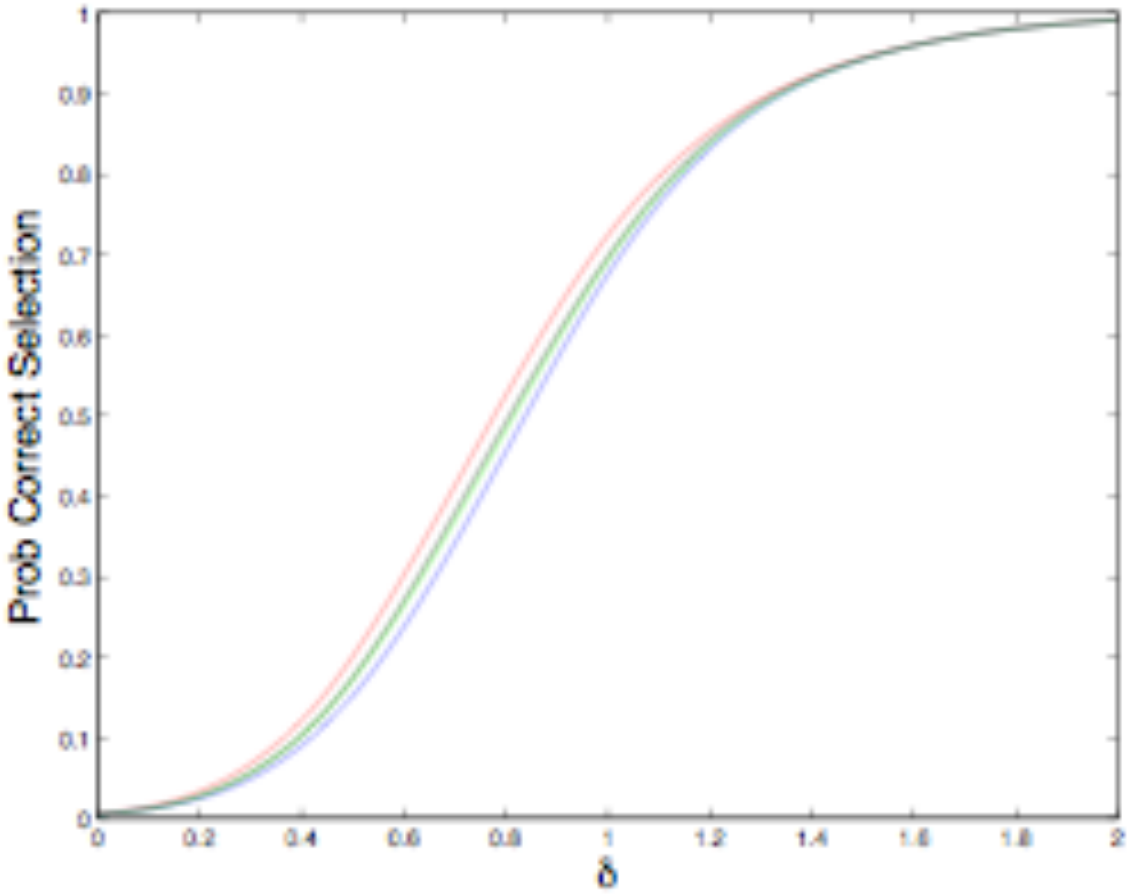
Stage 2 p-value: if i* in J, then p_{2,i^*}

(else fail to reject). Combine to get local test:

$$\sqrt{\frac{100}{600}}\Phi^{-1}(1-p_{1,J}) + \sqrt{\frac{500}{600}}\Phi^{-1}(1-p_{2,i^*}) > 1.96$$

“Pedagogic” Example

Power Curves (almost identical):



Example 2: Enrichment Design

- Consider total population treatment effect (θ_1), and treatment effects in L-1 subpopulations ($\theta_2,\theta_3,\ldots,\theta_L$).
- At end of stage 1, pick subpopulation with large estimated treatment effect: θ_{i^*} (and possibly using other criteria).
- In stage 2, enroll from this subpopulation only.
- Want to combine data from both stages to test H_{0i^*}.

Example 2: Enrichment Design

Example: Subpopulations:

1. Entire population
2. Men only
3. Men over 50
4. Men who are smokers

Each intersection null hyp. tested by combination test:

$$\sqrt{\frac{1}{2}}\Phi^{-1}(1-p_{1,J})+\sqrt{\frac{1}{2}}\Phi^{-1}(1-p_{2,J})>1.96$$

Example 2: Enrichment Design

Example: Subpopulations:

1. Entire population
2. Men only
3. Men over 50
4. Men who are smokers

At stage 1, intersection null hypothesis tested by $p_{1,J} = |J| \min_{i \in J} p_{1,i}$.

At stage 2, same but on reduced set J' for which data is collected in stage 2.

Stage 1 results

Full population:	$P_{1,1} = 0.20$
All men:	$P_{1,2} = 0.10$
Men over 50 years:	$P_{1,3} = 0.03$
Men who smoke:	$P_{1,4} = 0.03$

Stage 2 results

All men:	$P_{2,2} = 0.11$
Men over 50 years:	$P_{2,3} = 0.08$
Men who smoke:	$P_{2,4} = 0.03$

To test null hypothesis for "all men", we have to reject intersection nulls: J=

{2},{1,2},{2,3},
{2,4},{1,2,3},
{1,2,4},{2,3,4},
{1,2,3,4}

Stage 1 results

Full population:	$P_{1,1} = 0.20$
All men:	$P_{1,2} = 0.10$
Men over 50 years:	$P_{1,3} = 0.03$
Men who smoke:	$P_{1,4} = 0.03$

Stage 2 results

All men:	$P_{2,2} = 0.11$
Men over 50 years:	$P_{2,3} = 0.08$
Men who smoke:	$P_{2,4} = 0.03$

E.g. to test J={2,3}, we compute

$$P_{1,\{2,3\}} = 2\min(0.1, 0.03)$$

$$P_{2,\{2,3\}} = 2\min(0.11, 0.08)$$

$$\sqrt{\frac{1}{2}}\Phi^{-1}(1 - p_{1,J}) + \sqrt{\frac{1}{2}}\Phi^{-1}(1 - p_{2,J}) = 1.15$$