

Stopping Rules

- ❑ Ethically responsible
- ❑ Often IRB mandated in phase II trials
- ❑ Phase II trials: DSMB not standard
- ❑ Phase III trials: DSMB almost always
- ❑ Due to lack of DSMB in phase II, stopping rules force monitoring

Phase II stopping rules

□ Safety

■ continuous monitoring

- Example: “if the lower limit of the 95% confidence interval for toxicity ever rises above 10%, then study will stop”
- Example: “if 2 or more deaths are observed during any point in the trial, the study will stop”

■ interim monitoring

- Example: “we will evaluate the study after the first 15 patients. if no more than 2 AEs have occurred, the study will continue to its full enrollment of 30”
- Example: “after the first 15 patients, safety data will be evaluated”

Phase II stopping rules

□ Safety (continued)

- most institutions still have DSMB and IRB that will be vigilant
- However, predefined stopping rules that are evaluated as part of the protocol review process help make the decisions seemingly more objective

Phase II stopping rules

□ Efficacy

- In phase II, worry about 'futility'
- no point in continuing if there is early evidence that treatment is not working
- different ethical issue than phase III
 - phase III: is one arm better than other arm?
 - phase II: is this single arm not working?
- very common approach to early stopping
phase II: Simon Two-Stage Design

Simon Two-Stage Design

- Requires BINARY outcome
- Examples:
 - complete response by RECIST
 - CR or PR by RECIST
 - PET response
 - PSA response
- NOT examples of binary outcome:
 - time to progression
 - time to death
 - time to relapse
- Major difference?
 - how quickly do we assess the outcome?

Simon Two-Stage

- Standard 'frequentist' approach
- Define:
 - alpha and power
 - null response rate and alternative response rate
- Note: naturally 'one-sided' test
- Result: a set of designs fulfilling your criteria
- Example:
 - choose $\alpha=0.05$, power=90%
 - null and alternative response rates: 20% and 45%
- 'Single' stage design sample size: **N=23**

Simon Two-Stage

ExpDesign

File Edit View Design Action Tools Window Help



Multiple-Stage Design

Multiple Stage Design with Early Stopping for Futility Only (One-sided Test)

Input

2-Stage Design

3-Stage Design

Alpha =

Proportion for Ho =

Power =

Proportion for Ha =

Sample size required for a standard design (1-stage) = 23

Example

Compute

Sort

Print

Cl

Utility

Rank the following with 1 to 10 scales
(A high score means important):

How important to have a small maximum sample size?

How important to have a small expected sample size under Ho?

Design Id	Total Sample Size	Expected Sample Size under Ho	Sample Size at Stage 1	Cutpoint r1 (Stop trial if <= r1 at stage 1)	Cutpoint r2 (Stop trial if <= r2 at stage 2)	Probability of Early Stopping Under Ho	Probability of Early Stopping Under Ha	Actual Type-I Error Rate, alpha	Actual Power, 1-beta	Utility
MaxUtility	24	18.2	15	3	7	0.648	0.042	0.084	0.9	1.094
MinMaxSize	24	18.2	15	3	7	0.648	0.042	0.084	0.9	1.094
MinExpSize	25	17.3	14	3	7	0.698	0.063	0.093	0.903	1.09
1	24	20.4	7	0	7	0.21	0.015	0.087	0.906	1.038
2	24	21.3	8	0	7	0.168	0.008	0.088	0.91	1.018
3	24	22	9	0	7	0.134	0.005	0.089	0.912	1.003
4	24	22.5	10	0	7	0.107	0.003	0.089	0.913	0.992

Phase III Early Stopping

- ❑ Interim analyses
- ❑ pre-planned analyses that specify times at which data will be analyzed
- ❑ usually no more than 1 or 2 interim analyses
- ❑ why not many?
 - problem with 'type I error'
 - interim analyses generally allow you to stop when significant p-value
 - do not suggest stopping for insignificant p-value
 - bias in the stopping
 - giving yourself additional opportunities to stop early based on 'chance' evidence

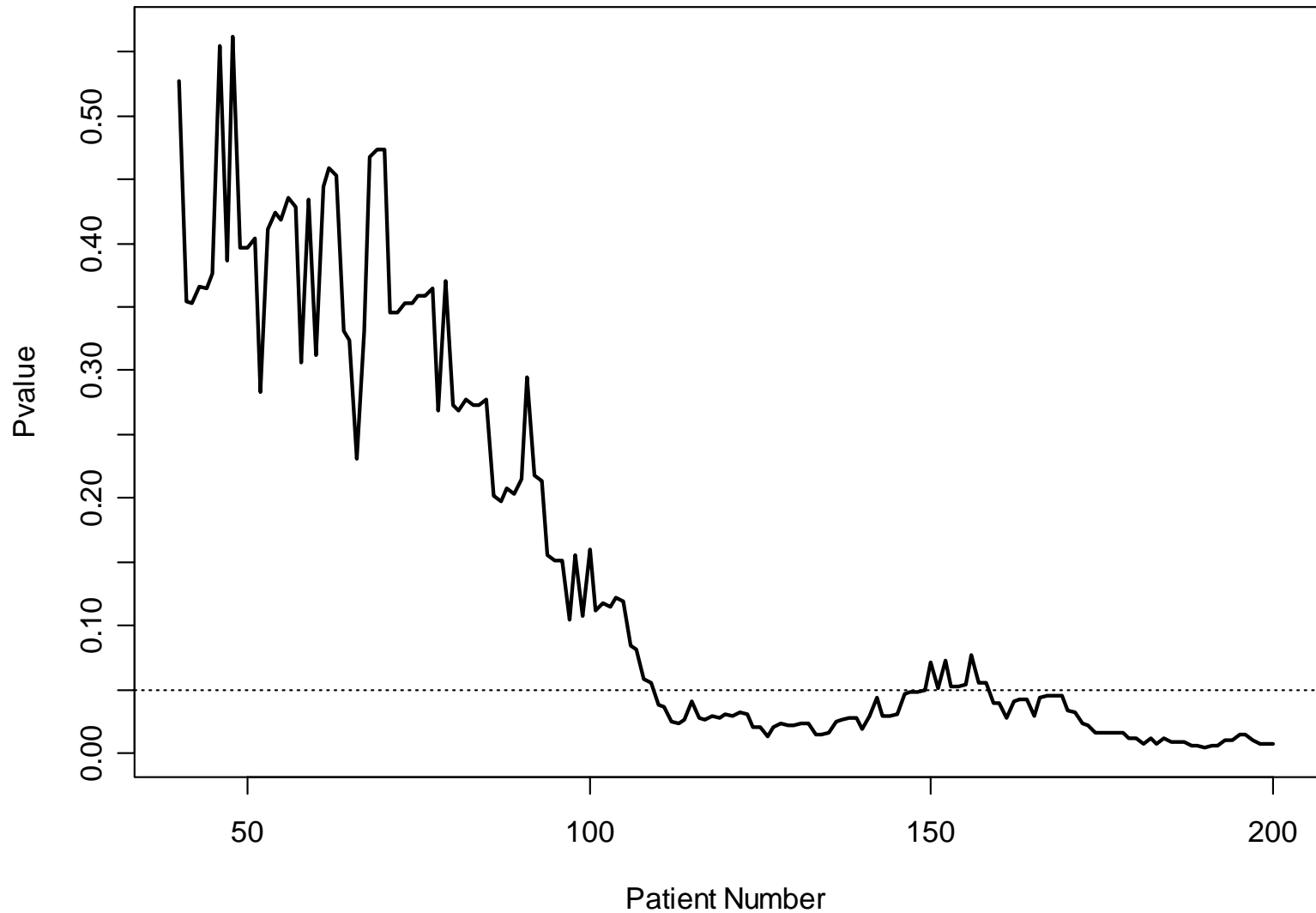
Phase III early stopping

- 'cumulative' type I error needs to be controlled
- what if we look at data after every patient?
- how likely are we to stop?
- Simulated examples

Example 1

- 20% difference in response rate
- $H_0: p_1 = p_2 = 0.50$
- $H_1: p_1 = 0.40; p_2 = 0.60$
- For power=80%, alpha = 0.05, randomize 200 patients
- Approach:
 - what if we analyze the data after every patient?
 - we would get 200 p-values
 - 'intuition' would dictate to stop when there was strong evidence to stop ($p < 0.05$)

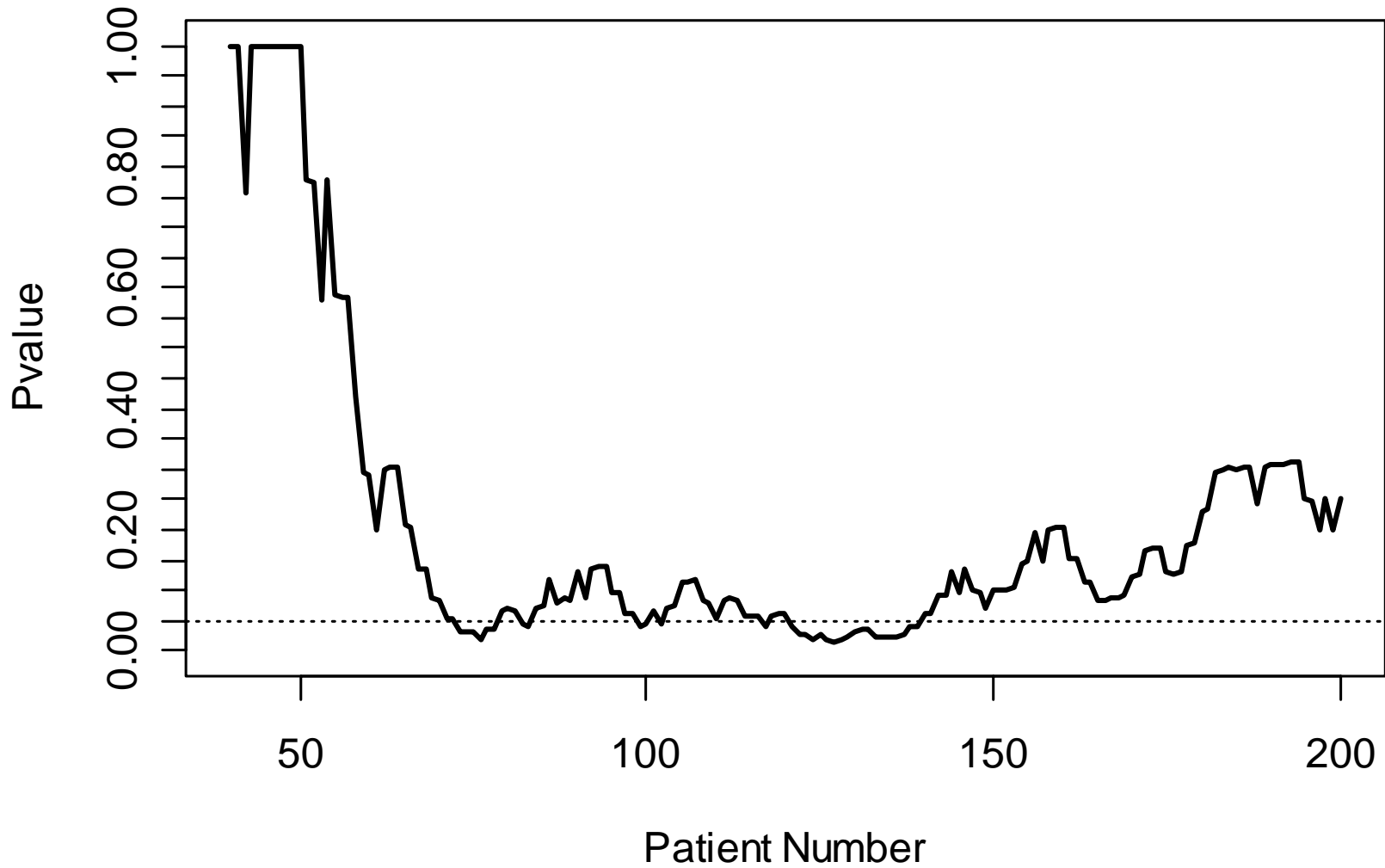
P-values



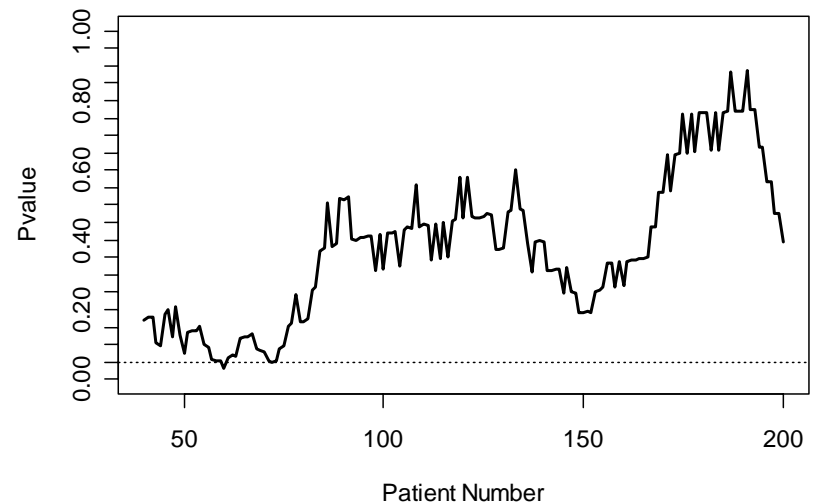
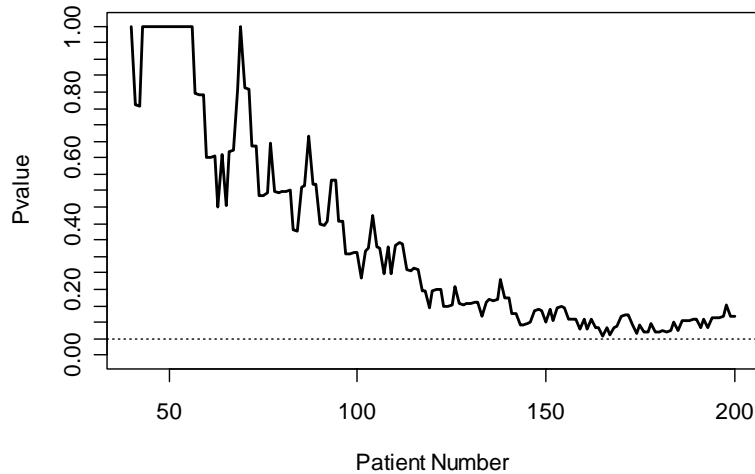
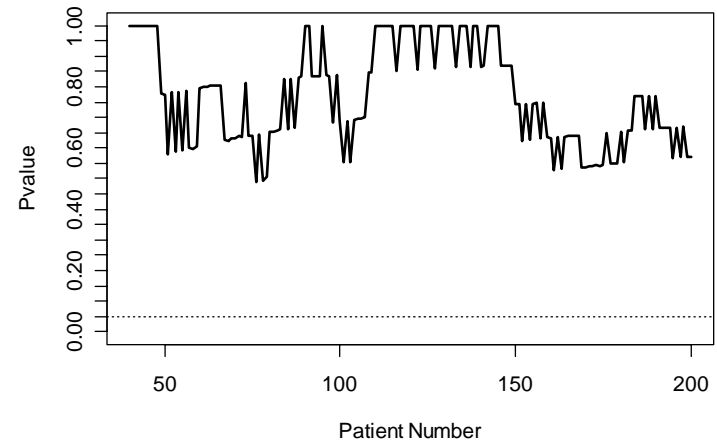
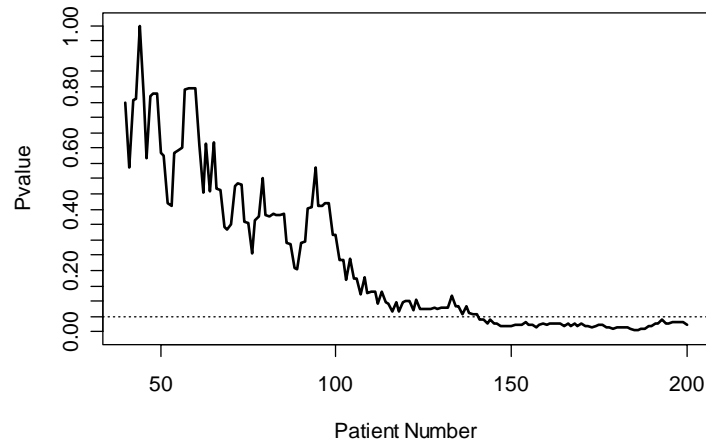
What if there is NO 'significant' difference?

- Assume same study design as previously
- $H_0: p_1 = p_2 = 0.50$
- $H_1: p_1 = 0.40; p_2 = 0.60$
- But, the 'truth' is that $p_1 = 0.40$ and $p_2 = 0.50$
- power to detect 0.40 vs. 0.50 is 0.26 with $N = 200$
- **What if we look at the p-values sequentially after each patient?**
- (Note: to find difference between $p_1 = 0.40$, $p_2 = 0.50$ you would need a total of $N = 780$ for power of 80%)

One possible outcome



Very many possibilities



Sequential Designs

- Designs have been developed to allow early looks
- usually just a few early looks (1 to 4)
- 'alpha-spending function'
- use up a little bit of alpha each time
- over all 'looks', the total alpha level is 'preserved'

Other designs?

- There are designs that let you look as often as you like without alpha penalties
- these do not use p-values
- Bayesian and likelihood approaches
- measures of evidence are different
 - likelihood ratios
 - posterior probabilities
- Becoming more popular
- But, FDA is a tough sell.