Elizabeth Garrett-Mayer, PhD
Associate Professor of Biostatistics and Epidemiology
Director of Biostatistics, Hollings Cancer Center
Medical University of South Carolina

Phase II Trials: Are we asking too much?
*Incorporating Early Stopping Rules in Phase II Studies*
Disclosures

- I have no conflicts or relationships to disclose
- I will not discuss off label use and/or investigational use in my presentation
Are we asking too much?

- In this case, in Phase II trials, are we asking too much of the patients?

- Phase II trials are generally designed at academic institutions or industry, often small companies.

- Same care and consideration needs to be given to **study design** as larger studies.
Are we asking too much?

- Would a phase III study without early stopping considerations be acceptable?
- Why should the standard be less in Phase II?

- If the study is a bust, why enroll more patients?
  - FUTILITY STOPPING
- If there is early evidence of strong efficacy, shouldn’t we stop assigning patients to an inferior arm?
  - EFFICACY STOPPING
- If a treatment regimen is too toxic, why assign more patients?
  - SAFETY STOPPING
Focus is on treatment trials

Assumption that the primary aim regards evaluating the efficacy of one or more treatment regimens

“data dependent” stopping. There are lots of other reasons studies could stop.
Early stopping rules: why do we need them?

- They tell if we have “answered the question” earlier than expected.
- They make clinical research ETHICAL
  - They help to avoid exposure of patients to ineffective agents
  - Stopping early will serve to better utilize “resources” for other trials
    - money
    - time
    - patients!
Early stopping rules

- Futility vs. efficacy
- Misunderstanding?
- What do stopping rules “look” like?
Stopping a randomized comparative trial early

Efficacy Stopping

- Sample size accrued
- Critical Value
- Trt A > Trt B
- Trt B > Trt A
Stopping a randomized comparative trial early

Efficacy and Futility Stopping

- Trt A > Trt B
- Trt A = Trt B
- Trt B > Trt A

Sample size accrued vs Critical Value for looks 1 to 4.
Consider a single arm trial with clinical response (i.e., PR + CR) as the outcome.

- The standard of care has a 20% response rate
- You expect (i.e., hope) your new treatment has at least a 40% response rate.
What is futility stopping?

- Futility stopping allows you to stop the study early if there is strong evidence that the null hypothesis will NOT be rejected.
- Huh?
- Our example:
  - \( H_0: \ p = 0.20 \)
  - \( H_1: \ p = 0.40 \)
Basics of the design

- Single arm phase II: usually one-sided test
  - $H_0: \ p \leq 0.20$ (null)
  - $H_1: \ p > 0.40$ (alternative)

- Type I ($\alpha$) and Type II ($\beta$) errors rates? 5-10%
- With $\alpha = 0.05$ and power $= 0.90$, we need $N=44$.

- Rejection rule: if 14 or more patients respond ($\geq 32\%$), reject the null hypothesis.
After 1 year of accrual…

- 24 patients have been treated.
- 5 patients have responded
- Response rate = 21%.
- What are the chances that this trial will go on to be successful?
- Should the study be stopped?
- Well, was there consideration given to early stopping?
Dilemma

- What are the implications of
  a. unplanned early looks?
  b. statistical inferences from data based on early looks?
- Should you stop? You didn’t plan to, but it just doesn’t seem right to keep enrolling patients.
- Design properties (e.g., power, p-values) essentially go out the window
How can this be avoided?

- Include early stopping rules in your protocol.
- A simple solution to our current problem: Simon two-stage design.
  - look at the data once, about halfway through
  - stop if the response rate is ‘low’ relative to the alternative
How can this be avoided?

- Simon’s optimal design for our scenario:
  - $H_0: p \leq 0.20$; $H_1: p > 0.40$
  - one-sided $\alpha=0.05$; power = 90%

- In the 1\textsuperscript{st} stage, enroll 20 patients
  - if 5 or more respond, continue to stage 2
  - if 4 or fewer respond, stop the study
- In the 2\textsuperscript{nd} stage, enroll 29 more patients
- At the end of the 2\textsuperscript{nd} stage, reject $H_0$ if 15 or more (of 49 patients) respond
What is the cost?

- Increased type I error? Not at all!
- Common misconception
- Early looks for futility DECREASE the power of the study (all else equal).
- To maintain the power, you generally need to increase the sample size a modest amount:
  - Our example:
    - N=45 without interim futility analysis
    - N=49 with interim futility analysis
Other approaches

- Simon’s two-stage designs allow one early look.
- There are other frequentist approaches with >2 (and even continuous) looks.
- Bayesian and likelihood approaches:
  - can implement multiple looks or continuous monitoring
  - if the evidence is ever ‘strong enough’ to conclude that either (a) the null will not be disproved, or (b) the alternative is unlikely to be true, then the study can be stopped
  - sometimes we hear about “conditional power”
Sequential futility stopping
Example: likelihood monitoring

![Graph showing likelihood ratios and patient numbers. The x-axis represents patient number from 0 to 50, and the y-axis represents likelihood ratios from 1/64 to 64. The graph shows two horizontal lines: one at 1/8 labeled "Favors Null" and another at 1/2 labeled "Favors Alternative."
Example: likelihood monitoring

![Graph showing likelihood monitoring with patient number on the x-axis and likelihood ratio on the y-axis. The graph indicates a line that crosses from 'Favors Null' to 'Favors Alternative' at certain points.](attachment:image.png)
What about error rates?

- Other paradigms (likelihood, Bayesian) do not evaluate evidence in the same way as the frequentist paradigm.
- But, they do have ways of controlling or bounding error rates.
Considered single arm binary

Number of arms?
- one vs. two or more
- comparative vs. non-comparative

Outcome of interest?
- binary
- continuous
- time to event (e.g. PFS)
What if you have two arms in your trial and one appears to be more efficacious at the interim look?

You should

a) continue the study with no changes
b) stop enrolling to the sub-optimal arm
c) stop enrolling to both arms.

Answer? it depends
Depends on what?

- Are the arms both experimental treatments?
- Is one arm the standard of care?
  - is the standard of care the superior or inferior arm?
- If both are experimental: drop the inferior arm.
- If one is experimental:
  - if the standard of care arm is superior: stop the study
  - if the experimental arm is superior: continue as is.
- If both are standards of care: stop the study.
Increasingly common to see progression-free survival (PFS) as a phase II endpoint in oncology (instead of response)

- For some cancers, we often use OS (e.g., brain cancer) even in Phase II

- Challenges: it may take longer to evaluate
Time to event endpoints

Example 1:
- randomized two-arm study of two agent combination vs. three agent combination
- first line treatment in metastatic pancreatic cancer
- Median PFS 6 months
- Accrual time of 18 months

Planned interim analysis at 12 months:
- 2/3 of patients accrued
- 50% of those on study have been followed for at least 6 months
- Early stopping for futility can avoid exposure to ineffective (additional) therapy for remaining patients
Time to event endpoints

- Example 2:
  - maintenance therapy [24 months] vs. standard of care (no treatment)
  - patients with relapsed chronic lymphocytic leukemia who have responded to induction therapy
  - Median PFS is about 30 months
  - Accrual time of 40 months

- Planned interim analysis at 44 months
  - all of the patients have been enrolled
  - What does interim analysis provide?
  - If there is evidence to stop:
    - for efficacy: early recommendation that maintenance therapy works
    - for futility: discontinuation of maintenance therapy for those on that arm
Many more possibilities to consider
Critically important to include safety stopping guidelines

Often, the phase II study dose may be based on a small phase I study with imprecise recommendation of the phase II dose.
Practical Guidelines

- Consider accrual rates and time of endpoint evaluation (per patient)
- Think about the **ethical implications** of the results of an interim look
  - for every current patient
  - for potential future patients
  - think about the consent form: does the stated purpose still apply?
Consider all possible interim look inferences:
- should the study stop?
- should one arm stop (and/or cross-over)?
- should the study continue?
- what would treatment options be if the study stopped vs. continued for those on or off study?

You may conclude that your design should have
- no early stopping rules?
- only futility stopping?
- only efficacy stopping?
- both?
Further Reading

- Lee (200) A predictive probability design for phase II cancer clinical trials. *Clinical Trials.*