Putting the Novelty Back in Phase II Trials

Elizabeth Garrett-Mayer, PhD
Hollings Cancer Center
Medical University of South Carolina

AACR/ASCO Workshop
Methods in Clinical Cancer Research
July/Aug 2007
Outline of Talk

- Goals of Phase II study
- Single arm studies
  - Traditional (frequentist)
  - Bayesian designs
  - Multiple outcome designs
  - Time to event outcomes
- Two (or more) arm studies
  - Traditional randomized Phase II
  - Novel multi-arm Phase II
Goals of Phase II Trials

- Provide initial assessment of efficacy or ‘clinical activity’
  - Screen out ineffective drugs
  - Identify promising new drugs for further evaluation
- Further define safety and toxicity
  - Type
  - Frequency
Important Design Considerations in Phase II trials

- Minimize cost of the trial
  - Minimize number of patients exposed to an ineffective treatment
  - Enroll as few patients as “necessary” to show benefit or failure
Standard Single Arm Phase II Study

- Single arm:
- Comparison is “fixed” constant
- Binary endpoint (clinical response vs. no response)
- Often one-sided test

- Simple set-up: \( \alpha = 0.10 \)
  \[ \beta = 0.10 \text{ (power = 0.90)} \]
  \[ H_0 : p = 0.20 \text{ (null response rate)} \]
  \[ H_1 : p = 0.40 \text{ (target response rate)} \]

- Based on design parameters:
  - N=39
  - Conclude effective if 12 or more responses (i.e., observed response rate of \( \geq 0.31 \))
Two-Stage Designs

- *What if by the 15th patient you’ve seen no responses?*
- *Is it worth proceeding?*
- Maybe you should have considered a design with an early stopping rule
- Two-stage designs:

  Stage 1:
  - enroll $N_1$ patients

  $X_1$ or more respond
  - Fewer than $X_1$ respond

  Stage 2: Enroll an additional $N_2$ patients

  Stop trial
Revised Design

- **Stage 1**: enroll 19 patients
  - If 4 or more respond, proceed to stage 2
  - If 3 or fewer respond, stop
- **Stage 2**: enroll 20 more patients (total N=39)
  - If 12 or more of total respond, conclude effective
  - If 11 or fewer of total respond, conclude ineffective
- Design properties?
  \[ \alpha = 0.10 \]
  \[ H_0 : p = 0.20 \text{ (null response rate)} \]
  \[ H_1 : p = 0.40 \text{ (target response rate)} \]
- What about power?
Question 1:

- The power of this two stage design is:
  1. higher than in the single stage study
  2. lower than in the single stage study
  3. the same as in the single stage study
  4. I can’t remember what power is...

- Answer:
Two-stage Designs

- Simon two-stage (1989)
  - Used in example
  - MANY designs fit the criteria
  - “Optimal”
    - Minimum expected sample size under $H_0$
    - Minimum maximum sample size
  - Preserves alpha and power, and permits early look

- Gehan two-stage (1961)
  - At stage 1, stop if 0 responses
  - Choose $N_1$ such that early stopping has ‘good’ properties
  - “Special case” of Simon two-stage
Early Stopping

- FUTILITY stopping
- The designs discussed so far ONLY allow stopping if there is strong evidence that the treatment is not efficacious
- Can also have early stopping for efficacy
  - Generally not popular in single arm studies
  - Important to accumulate evidence to support claim of efficacy
  - But, not stopping prolongs time to launch phase III
Frequentist versus Bayesians

- So far, “frequentist” approaches
- Frequentists: $\alpha$ and $\beta$ errors
- Bayesians:
  - Quantify designs with other properties
  - General philosophy
    - Start with prior information ("prior distribution")
    - Observe data ("likelihood function")
    - Combine prior and data to get "posterior" distribution
    - Make inferences based on posterior
Bayesian inference

- No p-values and confidence intervals
- From the posterior distribution:
  - Posterior probabilities
  - Prediction intervals
  - Credible intervals
- Bayesian designs
  - Can look at data as often as you like (!)
  - Use information as it accumulates
  - Make “what if?” calculations
  - Helps decide to stop now or not
Bayesian Designs

- Requires ‘prior’
  - Reflects uncertainty about the response rate
  - Can be ‘vague’, ‘uninformative’
  - Can be controversial: inference may change

![Graph showing prior distributions for response rate](image)
Question 2: Which prior makes the most sense?
Bayesian design example

N = 0; 0 responses

Response Rate

Posterior Distribution
Posterior Probabilities

![Graph showing cumulative number of patients against probability]

- Blue line: Probability $p < 0.20$
- Red line: Probability $p > 0.40$
Other priors

- What if we had used a different prior?
- Assume informative “orange” prior
Likelihood Approach

- Similar to Bayesian
- Royall (1997), Blume (2002)
- No prior distribution required
- Quantified by intuitive properties
  - Instead of $\alpha$ and $\beta$
  - “Probability of misleading evidence”
  - (i.e. choosing the wrong hypothesis)
- Likelihood ratio used for making inferences
- Can look at data as it accumulates
Multiple Outcomes

- Phase II = “safety + efficacy” trial
- Then why are we only talking about efficacy?
- Bryant and Day (1995): extend Simon two-stage to incorporate both outcomes
Bryant and Day Design

Stage 1: Accrue $N_1$ Patients

- **$O_3$**
  - Response: Inadequate
  - Toxicity: Insufficient data
  - Continue to Second Stage

- **$O_1$**
  - Response: Inadequate
  - Toxicity: Excessive

- **$O_2$**
  - Response: Insufficient Data
  - Toxicity: Excessive

Stage 2: If $Y_{R1} > C_{R1}$ and $Y_{T1} > C_{T1}$, Accrue $N_2 - N_1$ Additional Patients

- **$O_5$**
  - Response: Inadequate
  - Toxicity: Acceptable

- **$O_6$**
  - Response: Adequate
  - Toxicity: Acceptable
  - Recommend Treatment For Further Investigation

- **$O_4$**
  - Response: Adequate
  - Toxicity: Excessive

- **$O_1$**
  - Response: Inadequate
  - Toxicity: Excessive
Examples of Bryant and Day Designs:

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Optimal Designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>Safety</td>
</tr>
<tr>
<td>$P_{R0}$</td>
<td>$P_{R1}$</td>
</tr>
<tr>
<td>$P_{T0}$</td>
<td>$P_{T1}$</td>
</tr>
<tr>
<td>$N_1$</td>
<td>$C_{R1}$</td>
</tr>
<tr>
<td>$C_{T1}$</td>
<td>$N_2$</td>
</tr>
<tr>
<td>$C_{R2}$</td>
<td>$C_{T2}$</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
</tr>
<tr>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>43</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>31</td>
</tr>
<tr>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>54</td>
</tr>
<tr>
<td>14</td>
<td>36</td>
</tr>
</tbody>
</table>

- Example (first row)
  - null rates: efficacy 5% and safety 60%
  - alternative rates: efficacy 25% and safety 80%
  - Stage 1: enroll 22 patients
    - stop if (1) one or fewer responses OR (2) 14 or fewer “safe” patients
  - Stage 2: enroll an additional 21 patients (total N=43)
    - conclude a negative study if (1) four or fewer responses OR (2) 29 or fewer “safe” patients
Thall and Cheng Design
A “novel” problem

- Time to event outcomes in Phase II
  - Response rate no longer the ‘outcome of choice’ in Phase II studies
    - targeted agents may not shrink cancer
    - we’re learning: tumor shrinkage ≠ increased survival
  - Time to event outcomes more common
    - time to progression
    - time to relapse
  - More than ever, need early stopping
    - Many agents
    - Many are targeted
    - Small ‘target’ population so we need to conserve resources
Time-to-event outcomes in Phase II studies

- High dose temozolomide, thiotepa and carboplatin with autologous stem cell rescue (ASCR) followed by continuation therapy with 13-cis-retinoic acid in patients with recurrent/refractory malignant brain tumors

- Primary outcome: 1 year progression-free survival

- Study design:
  - Stage 1: enroll 17 patients. If 5 or more have PFS > 1 year, continue to stage 2
  - Stage 2: enroll 19 patients. Conclude success if >14 patients total have PFS > 1 year
  - Power = 80%, alpha = 4%.
You have enrolled the 17th patient

Accrued data:
- 2 patients were progression-free at 1 year visit
- 9 patients progressed before 1 year
- 6 patients have been followed for less than 1 year and are currently progression-free (times = 1, 2, 5, 7, 7, 9)

Study design calls for ‘interim analysis’ at the 17th patient.
What do you do?

1. Halt enrollment (which may be for 6+ months) to wait to see if the stopping rule is met.

2. Continue enrolling while waiting to see if stopping rule is met.

3. Extrapolate what 1 year PFS would be for the 6 patients who haven’t reached 1 year based on what we’ve seen thus far.
Randomized phase II

- Why randomized?
  - Want to explore efficacy
  - Not willing to invest in phase III (yet)
  - Want some “control” or “prioritization”
  - Primarily two different kinds of randomized phase II studies
    - *Phase II selection design (prioritization)*
    - *Phase II designs with reference control arm (control)*
  - Also phase II/III studies
Phase II selection design (prioritization)

- Two parallel one arm studies (classic case)
- **Do not directly compare arms to each other.**
- Compare each to “null rate”
- Why? To compare to each other, you’ll need a study at least two times as large.

- “Pick the Winner” (Simon, 1985)
  - **Appropriate to use when:**
    - Selecting among NEW agents
    - Selecting among different schedules or doses
  - **NOT appropriate when**
    - Trying to directly compare treatment efficacies (not powered)
Phase II selection design (prioritization)

- “Pick the Winner” (continued)
  - 90% chance of choosing better arm so long as true difference in response rates is >15%.
  - Uses 2+ Simon two-stage designs
    - Each arm is compared to a null rate
    - Must satisfy efficacy criteria of Simon design
    - Move the “winner” to phase III
    - Only have to pick winner if more than one arm shows efficacy
  - Can be used when the goal is prioritizing which (if any) experimental regimen should move to phase III when no a priori information to favor one.
Randomized Phase II designs with reference arm (control)

- Includes reference arm to ensure that historical rate is “on target”
- Reference arm is not directly compared to experimental arm(s) (due to small N)
- Can see if failure (or success) is due to incomparability of patient populations
Other Randomized Phase II designs?

Lots of randomized studies are calling themselves randomized phase II studies these days:

- If outcome of interest is surrogate
  - Correlative (biomarker)
  - Clinical (response)
- If sample size is relatively small but direct comparison is made
- If study is comparative, but is not definitive for whatever reason (e.g. if $\alpha$ and $\beta$ are large)
Phase II/III studies

- Several versions {Schaid (1988), Storer (1990), Ellenberg and Eisenberger (1985), Scher and Heller (2002)}
- General idea
  - Begin with randomized phase II study
  - Randomize to control arm & experimental arm(s)
  - If some threshold of efficacy is met, continue to phase III sample size for direct comparison
- Benefits:
  - Allow use of phase II data in phase III inference
  - Minimize delay in starting up phase III study
  - Uses concurrent control
Phase II/III studies

- **Cons:**
  - The sample size for the phase II part is approximately twice as large as would be needed for standard phase II.
  - Need phase III infrastructure developed even if it stops early.
  - Phase II outcome is not always the same as the Phase III outcome.

- **Would be useful if MOST phase II studies showed efficacy (not the case!)**

- Really, these could be considered phase III designs with very aggressive early stopping rules.
Adaptive Randomization Designs

- Randomization is “adapted” based on accumulated information
- Adaptive on Outcome (Bayesian/Likelihood)
  - Assign treatments according to accumulated information about best treatment. (Berry and Eick, 1995)
  - Assign with higher probabilities to better therapies
- Example: Troxacitabine in AML (Giles et al. 2003)
Adaptive Designs

Adapt the randomization to learn while effectively treating patients on trial:

(1) Begin by randomizing with equal chance per arm
(2) Then, adjust probability of assignment to reflect the knowledge of the best treatment
Adaptive Designs

- Begin assuming equally effective (1/3, 1/3, 1/3)
- May wait until a minimum number have been treated per arm
- Based on currently available (accumulated) data, randomize next patient (i.e., “weighted” randomization)
- Stopping rules: drop an arm when there is “strong” evidence that
  - It has low efficacy OR
  - It has lower efficacy than competing treatments
Adaptive Designs

- **Summary of trial results:**
  - TI dropped after 24\textsuperscript{th} patient
  - Trial stopped after 34 patients (TA dropped)

  Complete responses by 50 days

<table>
<thead>
<tr>
<th>IA</th>
<th>10/18 = 56%</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>3/11 = 27%</td>
</tr>
<tr>
<td>TI</td>
<td>0/5 = 0%</td>
</tr>
</tbody>
</table>
Summary

- STRONGLY CONSIDER ALLOWING FOR EARLY STOPPING

- Bayesian and likelihood designs:
  - Allow early stopping as soon as strong evidence develops
  - More complicated to implement
    - High-maintenance: many analyses
    - Computationally intensive
  - For Bayesian: choice of prior can be tricky
  - Lack of objectivity and potential loss of “equipoise”

- Frequentist designs:
  - Usually just one interim analysis
  - Simple implementation
Summary

- Think about why/whether a multi-arm trial is needed
  - Very useful when there is lack of historical data for comparison
  - Phase II randomized is NOT a short-cut to avoid a larger more definitive trial
  - Adaptive designs can be very efficient for selection, but require more maintenance
Issues with innovative designs

- Statistically intensive
  - “buy your statistician a beer (or bourbon)”
  - Probably cannot be used “off-the-shelf”
  - Require specialized software

- Need to be validated
  - Do they behave as promised?
  - Are they ‘robust’ (i.e., do they work when incorrect assumptions are made)?
References (1)


