Design of Phase II Clinical Trials

Susan Stewart, Ph.D.
Division of Biostatistics

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).
Topics

- Objectives
- Types
  - Multi-stage
  - Randomized
  - Platform
  - Crossover
Phase II clinical trials

- Phase II (NIH definition): Study the biomedical or behavioral intervention in a larger group of people (several hundred) to determine efficacy and further evaluate safety.
  - Is there any biological activity?
  - May or may not have concurrent controls
  - May be shorter term with different outcome and more exclusion criteria than phase III trials
  - Phase IIA-evaluate dosing; phase IIB –determine effectiveness
Phase II: Multi-stage designs

➢ Purpose
  o Identify drugs that are promising for further testing in a Phase III trial
  o Preliminary efficacy assessment
  o Avoid exposing patients to sub-therapeutic dose levels
  o Terminate the study if the treatment is ineffective
Single arm trials

- Optimal two-stage designs
  - Permit early stopping if there is a moderately long sequence of initial failures
  - Enroll $n_1$ patients in stage 1
  - If $\leq r_1$ responses, stop the trial
  - Otherwise, enroll $n_2$ more patients
  - Decide whether or not treatment is promising based on the $n_1+n_2$ patients
Two-stage designs

- Null hypothesis: probability of response is unacceptably low
- Alternative hypothesis: probability of response is sufficiently high to warrant further study
- Simon’s optimal two-stage design minimizes the expected sample size under the null hypothesis for the given error constraints
- Simon’s minimax design minimizes the maximum sample size for the given error constraints
Example: Intravenous aflibercept in patients with ovarian cancer

- Drug is a vascular endothelial growth factor (VEGF) inhibitor
- 2 dose levels tested (2 mg/kg and 4 mg/kg), based on previous phase 1 & 2 studies
- Patients with advanced platinum-resistant ovarian cancer
- Simon minimax 2-stage design
- Primary outcome: objective response rate (ORR)
- Null hypothesis: ORR ≤ 5%
- Alternative hypothesis: ORR ≥ 15%
- Tested at the 0.025 level, 1-sided

Tew et al. Cancer 2014; 120:335-43
2-stage design

- Plan: enroll 42 patients in each group in stage 1
- If at least 3 responders in stage 1 in a group, go on to enroll 25 patients in stage 2
- Declare drug suitable for future study if at least 8 responders total (stages 1 & 2) in a group
- Allowed to enroll additional patients beyond the 2-stage design to reach a planned total sample size of 200
### Sample size calculation

http://cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx

---

**Anastasia Ivanova, Ph.D., University of North Carolina at Chapel Hill**

**Continuous monitoring for toxicity**  
Simon's two-stage design  
Fleming's two-stage design  
Simon's like design with delayed futility stopping  
Two-stage design for ordinal outcomes  
The Rapid Enrollment Design (RED) for Phase I trials  
Other programs

---

### Simon's Two-Stage Design

The program generates Simon's optimal two-stage designs (Simon, 1989) and adaptive designs from Jung et al. (2004) for Phase I single arm clinical trials.


**Type I error rate, α (one-sided):** 0.025

**Power:** 0.8

**Response probability of poor drug, p1:** 0.65

**Response probability of good drug, p2:** 0.15

### Calculations

<table>
<thead>
<tr>
<th>N</th>
<th>α1</th>
<th>α2</th>
<th>p1</th>
<th>p2</th>
<th>Type I Error</th>
<th>Power</th>
<th>CR0</th>
<th>Probability of early stopping</th>
<th>Interval for α</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>0.0106</td>
<td>0.0004</td>
<td>50.8</td>
<td>66</td>
<td>0.0001</td>
<td>0.8004</td>
<td>68.6</td>
<td>(0.8044, 1)</td>
<td>Minimax</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>0.0186</td>
<td>0.0002</td>
<td>45.7</td>
<td>68</td>
<td>0.0001</td>
<td>0.8002</td>
<td>67.8</td>
<td>(0.8078, 0.8369)</td>
<td>Minimax</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>0.0218</td>
<td>0.0004</td>
<td>43.5</td>
<td>68</td>
<td>0.0001</td>
<td>0.8004</td>
<td>66.6</td>
<td>(0.8093, 0.8388)</td>
<td>Minimax</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>0.0288</td>
<td>0.0006</td>
<td>39.6</td>
<td>69</td>
<td>0.0001</td>
<td>0.8006</td>
<td>64.4</td>
<td>(0.8094, 0.8382)</td>
<td>Optimal</td>
<td></td>
</tr>
</tbody>
</table>

---

**Notes:**

- n is the total number of subjects
- n1 is the number of subjects accrued during stage 1
- n2 is the number of patients whose responses are observed during stage 1, the trial is stopped early for futility
- p1 or p2 are the response rates observed by the end of stage 1, then no further investigation of the drug is warranted
- CR0 is the expected sample size for the trial when response rate is p2
- Interval for α is the set of values α such that the design minimizes α = e^-1 + α2

---

*The development of this software was supported by funds from the National Institutes of Health 58051C0502015A1. 
For comments, questions, and suggestions, email: biostatistics@ivanova@n.unc.edu.*

---

**UNC Lineberger Comprehensive Cancer Center**

---

**Clinical and Translational Science Center**
Multiple stage designs

- Can extend to 3 (or even 4 stages)
- May require at least one response at first stage to go on to the second stage

Considerations for any multi-stage design
- How long will it take to determine whether there are enough responses to proceed to the next stage?
- Will we stop the study or keep on enrolling while waiting for the results from the previous stage?
Randomized phase II designs

- May randomize patients to different drugs or dose levels of the same drug
- Can estimate differences between treatments
- Can pick the treatment with best response
- Randomization produces balanced groups
Example: Phase II trial—Oncken (2006)

- **Background:** Evaluated 4 varenicline dose regimens for promoting smoking cessation.

- **Methods:** Multicenter, double-blind, placebo-controlled. Randomized healthy smokers aged 18-65 to varenicline tartrate or placebo twice daily for 12 weeks
  - 0.5 mg non-titrated (n=129); 0.5 mg titrated (n=130)
  - 1.0 mg non-titrated (n=129); 1.0 mg titrated (n=130)
  - placebo (n=129)

with 40-week follow-up to assess long-term efficacy.

Primary efficacy outcomes: carbon-monoxide confirmed 4-week continuous quit rates; continuous abstinence

*Arch Intern Med.* 2006 166(15):1571-7
Data Analysis

- Quit rates: binary
  - Compared each treatment group separately vs. placebo
  - Compared pooled dosage groups vs. placebo
  - Step-down procedure to account for multiple comparisons
  - Logistic regression
    - Independent variables: treatment and center
    - Computed odds ratios with 95% confidence intervals

- MNWS (withdrawal), mCEQ (cigarette evaluation): numeric
  - Analysis of covariance (ANCOVA)
    - Covariate: baseline level of outcome variable
    - Independent variables: treatment and center
Results

- Weeks 9-12 continuous quit rates greater in 1.0 mg group and 0.5 mg group than placebo
- Weeks 9-52 abstinence rates greater in 1.0 mg group and 0.5 mg group than placebo
- Generally well tolerated
  - Nausea in 16%-42% of varenicline treated subjects
  - Less nausea with titrated dosing
Continuous quit rates. P<.001 for each treatment group vs placebo. BID indicates twice daily. The odds ratios (ORs) and 95% confidence intervals (CIs) for the weeks 4 through 7 evaluation were 4.96 (95% CI, 2.66-9.22) for the 0.5-mg group and 5.86 (95% CI, 3.16-10.90) for the 1.0-mg group; for the weeks 9 through 12 evaluation, 6.32 (95% CI, 3.47-11.50) and 8.07 (95% CI, 4.42-14.70), respectively.

Figure Legend:
Figure Legend:
Carbon monoxide–confirmed weekly point prevalence abstinence rates. BID indicates twice daily. *P<.001 vs placebo.
Figure Legend:
Mean changes in Minnesota Nicotine Withdrawal Scale “urge to smoke” scores from week 1 to week 12 for all subjects. BID indicates twice daily. In comparison with placebo, asterisk indicates $P<.001$; dagger, $P<.01$; and double dagger, $P<.05$. 
Conclusion

- Varenicline tartrate, 0.5 mg and 1.0 mg twice daily, is efficacious for smoking cessation.
Platform Trials

- Multiple treatments evaluated simultaneously
- Single master protocol
- Adaptive platform designs
  - Drop treatments for futility
  - Declare one or more treatments superior
  - Add new treatments
- Multi-arm, multi-stage
- More efficient than traditional RCT designs


Example: ACCORD Seamless Phase 2 Platform Study to Assess Multiple COVID-19 Treatments

- **Objectives:**
  - Stage 1 (screening stage): Evaluate safety and efficacy of candidate agents as add-on therapy to standard of care (SoC) in hospitalized patients
  - Stage 2 (expansion stage): Confirm efficacy of agents selected based on evidence from Stage 1

- **Participants:**
  - Hospitalized patients age ≥18 with Grade 3-5 COVID-19 in UK

- **Main outcomes:**
  - Time to sustained clinical improvement ≥2 points on WHO 9 point ordinal scale
  - Live discharge or fit for discharge (0-2 on WHO scale) by Day 29

ACCORD trial (cont’d)

 Comparator and candidate interventions
  o Current SoC for COVID-19
  o Bemcentinib
    • Could reduce viral infection; blocks spike protein
  o MEDI3506
    • Anti-IL-33 monoclonal antibody; could treat respiratory failure
  o Acalabrutinib
    • BTK inhibitor; anti-viral and anti-inflammatory
  o Zilucoplan
    • Complement C5 inhibitor; may block severe inflammatory response
  o Nebulized heparin
    • Binds with spike protein
  o Others TBD
ACCORD trial (cont’d)

- Randomization
  - Stratified by baseline severity grade
  - Equal allocation to each experimental arm and contemporaneous SoC arm
  - May be changed to 2:1 in favor of experimental arms

- Sample size per agent
  - Stage 1: 60
  - Stage 2: 126
  - Total: up to 1800
Crossover Trial

- Definition (Chow & Liu): Modified randomized block design in which each block receives more than one treatment at different dosing periods.
- Simplest case: each participant is randomized to receive 2 treatments, A and B, in the order AB or BA.
- Between the 2 treatments, there is a washout period.

Design and Analysis of Clinical Trials (3rd Ed.) Chow & Liu, Wiley, 2014
Crossover Trial

➢ Advantages
  o Each participant serves as his or her own control
  o Removes inter-patient variability from the comparison of treatments
  o Therefore, requires a smaller sample size than a parallel groups design

➢ Disadvantage
  o Have to worry about carryover between treatments
    • Carryover effects may not be equal
  o Vulnerable to dropouts
Higher Order Crossover Designs

- Definition (Chow & Liu):
  - Number of periods > number of treatments
    - Two-sequence dual (extra period) design: ABB, BAA
    - Doubled (replicated) design: AABB, BBAA
  - Number of sequences > number of treatments
    - Balaam’s design: AA, BB, AB, BA
  - Both
    - Four-sequence design: AABB, BBAA, ABBA, BAAB
- These designs allow estimation of carryover effects and intra-patient variability
Crossover Trial

- Example: Randomized double blind trial of dark chocolate/cocoa snack vs. control snack in overweight people aged 40-64 (n=30)
- 2 periods, 4 weeks each, with 2-week washout period
- Outcomes: large & small blood vessel dilatation, peripheral blood flow, arterial stiffness
- Comparison: Active vs. control & baseline

West et al., British Journal of Nutrition 2014; 111:653-61
Data Analysis

- Initial model
  - Fixed effects: treatment (baseline, active, control), period, treatment X period interaction
  - Random effect: participant

- Treatment X period was not statistically significant

- Some models included treatment X sex interaction

- Tukey’s post-hoc tests for multiple comparisons
### Table 4: Results

<table>
<thead>
<tr>
<th></th>
<th>Pre-treatment‡</th>
<th>Control§</th>
<th>Active§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
</tr>
<tr>
<td><strong>Ultrasound measurements</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basal arterial diameter (mm)</td>
<td></td>
<td>4.20***</td>
<td>0.17</td>
</tr>
<tr>
<td>Peak arterial diameter (mm)</td>
<td></td>
<td>4.39***</td>
<td>0.18</td>
</tr>
<tr>
<td>FMD (% change)</td>
<td>4.73</td>
<td>0.41</td>
<td>5.12</td>
</tr>
<tr>
<td><strong>Doppler-derived measures</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basal flow volume (ml/s)</td>
<td></td>
<td>166**</td>
<td>18</td>
</tr>
<tr>
<td>Peak flow volume (ml/s)†</td>
<td>1059*</td>
<td>76</td>
<td>1032*</td>
</tr>
<tr>
<td>Reactive hyperaemia (% change)††</td>
<td>612*</td>
<td>37</td>
<td>587</td>
</tr>
<tr>
<td><strong>EndoPAT variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RHI</td>
<td>2.26</td>
<td>0.14</td>
<td>2.19</td>
</tr>
<tr>
<td>fRHI</td>
<td>0.60</td>
<td>0.09</td>
<td>0.55</td>
</tr>
<tr>
<td>AI‡‡</td>
<td>9.92**</td>
<td>3.9</td>
<td>5.90**</td>
</tr>
<tr>
<td>AI at 75 bpm§§</td>
<td>2.75**</td>
<td>3.9</td>
<td>-2.72**</td>
</tr>
<tr>
<td><strong>Anthropometrics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>80.9</td>
<td>2.3</td>
<td>80.7</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>27.4</td>
<td>0.5</td>
<td>27.5</td>
</tr>
<tr>
<td>Waist circumference (cm)</td>
<td>94.6</td>
<td>1.2</td>
<td>94.7</td>
</tr>
<tr>
<td>Hip circumference (cm)</td>
<td>106.8</td>
<td>0.9</td>
<td>106.9</td>
</tr>
<tr>
<td>Waist:hip ratio</td>
<td>0.89</td>
<td>0.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Mean values were significantly different from those of the active group: * P ≤ 0.05, ** P ≤ 0.01, *** P ≤ 0.001.
Fig. 1 Sex difference in vascular response to the cocoa+dark chocolate treatment. Women () exhibited significant reductions in the augmentation index, whereas men () did not (sex × treatment interaction, P= 0.01).
2-Period 2-Treatment Crossover Trial: Outcome by Sequence & Period

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Period 1</th>
<th>Period 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>$Y_A$</td>
<td>$Y_B$</td>
</tr>
<tr>
<td>BA</td>
<td>$Y_B$</td>
<td>$Y_A$</td>
</tr>
</tbody>
</table>
Simplifying Assumptions

- $H_0: \mu_B = \mu_A; H_a: \mu_B \neq \mu_A$
- Specify $\mu_B - \mu_A = \delta$
  (difference in treatment effects)
- No sequence or period effect: paired t-test comparing treatment B with treatment A over the entire sample
  - Specify SD=$\sqrt{2} \times$ (within-person SD)=$SD(Y_B-Y_A)$
  - Or specify $SD(Y_B)$, $SD(Y_A)$, and $corr(Y_A,Y_B)$
One Arm Normal

One Arm Normal is a program to calculate either estimates of sample size or power for one sample normal problem.

<table>
<thead>
<tr>
<th>User Input</th>
<th>Program Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1 Sided</td>
</tr>
<tr>
<td>Power</td>
<td>2 Sided</td>
</tr>
</tbody>
</table>

Select Calculation and Test Type

Select Hypothesis Test Parameters

<table>
<thead>
<tr>
<th>Null Mean</th>
<th>Alternative Mean</th>
<th>Standard Deviation</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1.414</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>22</td>
</tr>
</tbody>
</table>

Calculate

Help Document
Crossover Trial vs. Parallel Group Sample Size

- For a given
  - difference in treatment mean responses $\mu_B - \mu_A = \delta$
  - treatment response variance $\text{Var}(Y)$
    - (between-person plus within-person)
  - levels of type I & II error

$$\frac{n_{\text{crossover}}}{n_{\text{parallel}}} = 0.5\times[1-\text{corr}(Y_B,Y_A)]$$

- Even if there is no within-person correlation, the crossover trial requires half the sample size
- The greater the correlation, the greater the reduction in sample size
Considerations

- If intra-patient variability ≥ inter-patient variability, parallel groups preferred to crossover
- If inter-patient variability is large and the number of treatments is small, consider a cross-over design
  - However, disease state must be stable
Selecting a design

- Need to consider (Chow & Liu)
  - Number of treatments to be compared
  - Characteristics of the treatment
  - Study objectives
  - Availability of participants
  - Inter- and intra-person variability
  - Duration of the study
  - Dropout rates