The Menu of Clinical Trial Design and Randomization Options

Jody D. Ciolino, PhD
Assistant Professor-Biostatistics
Department of Preventive Medicine
Biostatistics Collaboration Center (BCC)
• Introduction to Clinical Trials
• Defining the Question
• General Design and Analysis Considerations
  – Phase I
  – Phase II
  – Phase III
• General Analysis Considerations
• Randomization in Clinical Trials
• Handling Baseline Variables in Analysis
• Take-home Messages/Summary
Introduction:
Clinical Trials and Their Phases
Generate hypotheses

Establish causality

- Case reports
- Case series
- Ecologic studies
- Cross-sectional studies
- Case-control studies
- Cohort studies
- Randomized controlled trials
• A clinical trial has some sort of interventional/experimental component
• It is NOT necessarily randomized
Phases of Clinical Trials

• Pre-clinical/Phase 0
• Phase I: “First in man”, dose finding, safety
• Phase II: Safety, *hint* of efficacy
• **Phase III: “gold standard”**
  – Safety and Efficacy
  – Usually randomized, parallel control
• **Phase IV: post-market, often pragmatic, comparative effectiveness**
• The path through phases is NOT always a linear one
• Much grey area (i.e., variability)
Defining the Question and Determining Outcome(s)
All of these topics/ideas surrounding design are interrelated...

• Research question
• Study design
• Analysis plan
• Sample size
Defining the Research Question

• Iterative process
• Usually begins general, then becomes more specific
• E.g., “Does my new drug help lower blood pressure?”
  – In what population?
  – Compared to...?
  – Define “lower”?
  – After how long of therapy regimen?
  – → → → “In geriatric adults (aged ≥ 65 years), does daily dose of Drug X improve blood pressure by at least 5mmHg after 8 weeks of treatment and follow-up?”
Question Determines Outcome...

- Outcome and question combined determine analyses (statistical problem)
- Analyses determine sample size
- Analyses and sample size drive study design
- Example:
  - Average change in raw blood pressure (change from baseline) → Outcome = mean change (continuous)
  - Meeting “success” definition of a drop in SBP = 5mmHg → Outcome = binary
  - Average SBP at eight weeks of follow-up and intervention → Continuous
  - SBP trajectory over eight weeks → Continuous slope
a) Average change in raw blood pressure (change from baseline): $\Delta$

b) Meeting “success” definition of a drop in SBP = 5mmHg

c) Average SBP at eight weeks of follow-up and intervention ($\mu_2$)

d) SBP trajectory over eight weeks (look at slopes of dotted lines)
Summary of Considerations for Outcomes

• Be specific in defining question
• Primary question → primary outcome → analyses → power/sample size → study design
• How is outcome measured (what does a “success” mean)?
• What is population to be studied?
• When is outcome assessed?
  – Is it possible to assess outcome in reasonable time? Do you need to consider a surrogate and/or composite outcome?
  – Are you interested in difference from baseline? Value at certain time point (when controlling for baseline)?
• What are other relevant predictors of outcome?
• What is the comparator?
Early Phase Trial Designs
Early Phase Clinical Trials

- Following pre-clinical studies (in vitro, in silico) and animal experimentation, Phase “0” (PK/PD), and/or...
- Phase I → usually...
  - Dose-finding studies
  - First in human
  - SAFETY = primary goal
  - N ≈ 20
  - Healthy, paid volunteers
  - Sometimes, very ill volunteers with few options
- Assumptions: higher dose = more efficacious, higher dose = more toxic/dangerous
Phase I

• Goal: find the highest dose of investigational product (IP) that results in no more than a pre-defined proportion of (pre-defined) toxicities

• **Maximum tolerated dose** (MTD) = dose meeting above criterion

• Usually proportion is around 20% - 33%, but this depends on disease/available treatments

• **Dose-limiting toxicity** (DLT) = toxicity or adverse event (AE) that is of severity substantial enough to make us question the potential benefit of the IP at that dose
  – May again depend on disease and available treatments
  – E.g., Prolonged vomiting to treat skin rash vs. prolonged vomiting to treat cancer
Phase I Assumptions

Or some other percentage

% Toxicity

Dose

d₁ d₂ mtd
Phase I

• Usually test 3 – 10 different dose levels
• E.g., We have 5 different doses, and 20 subjects; we want DLT rate of no more than 25% (1 in 4)...
  – What might be a good way to go about finding the MTD?
  – Randomly assign 4 subjects to each dose and observe toxicity rate at each dose level?
  – Why might this not be a good idea?
Phase I Design Options

• Rule-based designs ("up-and-down")
  – Began in 1940s (Dixon, Mood, Storer)
  – 3+3
    • Treat 3 participants at Dose K
    • If 0 have DLT \(\rightarrow\) escalate to Dose K + 1
    • If 2 or 3 have DLT \(\rightarrow\) de-escalate to Dose K - 1
    • If 1 has DLT \(\rightarrow\) treat 3 more at same dose
      • If 1/6 has DLT \(\rightarrow\) Escalate to K + 1
      • If at least 2/6 have DLT \(\rightarrow\) De-escalate to K – 1
  – MTD = highest at which 0 or 1 out of 6 have DLT
  – Need to pre-specify doses
The problem with traditional design(s)...

- Imprecise/inaccurate estimate of MTD
- Does not use all data to determine MTD
- Algorithm based
- Potential implications
  - Choose dose too high → move on to Phase II or III and find too toxic
  - Choose dose too low → move on to Phase II or III and find not efficacious
Observed Data: with 90% CIs

Example 1: total N=21

<table>
<thead>
<tr>
<th>Dose</th>
<th>Cohort 1</th>
<th>Cohort 2</th>
<th>Cohort 3</th>
<th>Cohort 4</th>
<th>Cohort 5</th>
<th>Cohort 6</th>
<th>Cohort 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>DLTs</td>
<td>0/3</td>
<td>1/3</td>
<td>0/3</td>
<td>1/3</td>
<td>0/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>
Examples (Garrett-Mayer, 2010)

**Observed Data: with 90% CIs**

Example 2: total N=12

<table>
<thead>
<tr>
<th></th>
<th>Cohort 1</th>
<th>Cohort 2</th>
<th>Cohort 3</th>
<th>Cohort 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>DLTs</td>
<td>0/3</td>
<td>0/3</td>
<td>0/3</td>
<td>2/3</td>
</tr>
</tbody>
</table>
Another example

• Underlying toxicity rate (that we do not know) is actually 50% at dose level $k$
• We treat cohorts of size 3, and we plan to use 3+3 design
• We will escalate if no DLTs observed
  – (1) What is the probability of this happening?
  – Use binomial distribution
• We would also escalate if 1 out of 3 have DLT in first cohort AND 0 out of 3 have DLT in second cohort
  – (2) What is the probability of this happening?
  – (3) Thus, what is the probability of escalating to next dose under these conditions?
Phase I Design Options, cont’d

• Up-and-down approaches, cont’d
  – 1+5 design
  – 6+0 design
  – 2+4 design

• Are these a good idea? Bad idea?

• These designs are *popular*
  – Easy to understand
  – Easy to implement
  – Historically used/comfort aspect
  – Other approaches, less intuitive for non-statistical/mathematical minds
Phase I Design Options

• “Non-traditional” methods
  – Based on statistical models
  – **Adaptive**
    – Requires statistician from the beginning

• End goal =
  – Accurate selection of MTD with DLT close to specified rate
  – Few exposures to toxic doses

• “Novel” designs:
  – Continual Reassessment Method (CRM; O’Quigley et al., 1990)
  – Escalation with Overdose Control (EWOC; Babb et al., Rogatko et al., late 1990s-early 2000s)
  – Others: TITE-CRM, *many* modifications to CRM
Continual Reassessment Method (CRM)

• Dose for next participant depends on toxicity of all previous participants in trial

• **Bayesian** design
  – Choose likelihood [mathematical model of prob(toxicity | dose)]
  – Choose **prior distribution** (regarding parameter for toxicity)
  – Estimate **posterior** distribution = function of prior x likelihood, for parameters of interest
  – Find next dose consistent with DLT rate desired
Example:

$$p(\text{toxicity}|dose = d_i) = \frac{\exp(3 + \alpha d_i)}{1 + \exp(3 + \alpha d_i)} \quad (\text{where } d = \text{dose} - 7)$$

Garrett-Mayer, 2010
What are the goals?

1. Find alpha
   - What is the alpha that is most consistent with the model?
   - Recall: Bayesian
     - Prior on alpha
     - Estimate likelihood
     - Find “best” alpha using posterior

2. Find the dose for the next patient
   - After alpha is estimated
   - Plug alpha “hat” in model
   - Find dose that is consistent with desired DLT rate
Prior Distribution on Toxicity Parameter

• Very important and may impact results largely
• MUST consider several scenarios/conduct sensitivity analyses
• Simulations generally involved here
• May use “vague” or non-informative prior distribution
  – Chi-squared distribution (modify df)
  – Normal distribution (modify variance, mean)
Need to spend time on the design

Try a normal prior with mean 1: tweak variance

No DLT

DLT
Example:

Let’s say we assume $\alpha = 1$.

$$p(\text{toxicity}|\text{dose} = d_i) = \frac{\exp(3 + \alpha d_i)}{1 + \exp(3 + \alpha d_i)} \quad (\text{where } d = \text{dose} - 7)$$
Scenarios for next patient

Dose-Toxicity Relationship

- Prior
- Next: no DLT
- Next: DLT

Dose
Many options for statistical model tied to CRM design...

Garrett-Mayer, 2006, Clinical Trials

Figure 2. Four different mathematical models that are commonly used for CRM studies. For each family of models, several pooling parameter $\beta$ (parameter $b$ is $\beta$) and $c$ are shown. A: one-parameter logistic: $P(dose) = \frac{1}{1 + e^{-(c-dose)}}$; B: two-parameter logistic: $P(dose) = \frac{1}{1 + e^{-(c-dose)^{1.5}}}$; C: one-parameter logistic model with pooling parameter $\beta$; D: one-parameter logistic model with pooling parameter $\beta$.
Modifications to CRM

**Box 2** Some suggestions for modifying CRM by Faries [8], Goodman *et al.* [6], and Möller [9].

1) Pre-define dose levels for escalation as if for a “3 + 3” design [6,8,9].
2) Always start at the lowest dose level under consideration [6,8,9].
3) Enroll two or three patients at each prescribed cohort (as opposed to only one per cohort) [6,8,9].
4) Proceed as a standard dose escalation design in the absence of dose-limiting toxicities [9].
5) Any given dose escalation cannot increase by more than one level, although dose de-escalations can be large [6,8,9].

**Box 3** Piantadosi *et al.*’s “practical” CRM.

1) Study preclinical information about toxicities.
2) Quantify clinical intuition about dose-toxicity by choosing dose level that would be guessed to incur low (eg, 10%) toxicity and level that would be guessed to incur high (eg, 90%) toxicity rate.
3) Estimate/draw the dose–toxicity curve assuming that the dose-toxicity curve passes through the dose–toxicity points described in step 2 above.
4) Use the dose–toxicity curve to find the target dose for the desired toxicity level.
5) Treat three (or more) patients at the target dose.
6) Re-estimate the dose–toxicity curve based on the toxicity outcomes of the treated patients.
7) Revise estimate of the dose associated with high toxicity (from step 2 above).
8) Repeat steps 3–6 until target dose changes by less than 10% or meets another appropriate criterion for stopping.
9) Use target dose for future trials.
CRM in R

- CRM package (download and install)
- Function: `crm(target, prior, ptdata, model=1, a0=1, b=3)`
  - Target = target toxicity probability
  - Prior = vector of prior probabilities for toxicity in each dose
  - Ptdata = n x 2 matrix of data
    - Column 1 = dose levels
    - Column 2 = indicator for toxicity (0 or 1)
  - Model = dose-toxicity model (default = 1: hyperbolic-tangential model; may be 2: one-parameter logistic)
  - a0 = initial value for model parameter
  - b = initial fixed intercept parameter for one-parameter logistic, default = 3
- OTHER resources: [https://biostatistics.mdanderson.org/SoftwareDownload/](https://biostatistics.mdanderson.org/SoftwareDownload/)
Analysis for Phase I

• Often limited by sample size
• Bayesian analyses
• Descriptive statistics are *always* a good idea
• Exact tests/nonparametrics (if any)
  – May want to relax type I error rate
  – Very seldom considerations for multiple hypothesis testing, etc.
  – Analyses seem “simple”, but data quality and precision = key
Phase II
Phase II...

- Small studies of efficacy
- Meant to inform decision whether to conduct larger, phase III trial
- “small” may have many meanings
  - Sample size usually not more than 50-60, but can be in 100s range
  - Follow-up usually relatively short (less than 1 year)
- Often a single arm, but not always the case
- “Distinction is not necessarily sharp” (Cook and DeMets, 81)
Gehan Design (1961)

- Use binary outcome ("success" or "response")
- Two stages: "Screening" for futility and then follow-up
- **Stage 1:** Historically, focus = 20% response rate, but methods apply to other rates
- Assume $H_0$: response = 20%, $H_1$: response > 20%
  - We want to control type II error rate = $p(\text{fail to reject } H_0 \mid H_1 \text{ true})$
  - What if we conduct a trial and see **zero** successes?
    - We would probably fail to reject $H_0$
    - What is type II error rate if $y = 0$ for given underlying proportion?
      - Again, use the binomial distribution...
Gehan, cont’d

• If we use \( n = 14 \) subjects in a small study for efficacy...
  – Assume 20% response rate
  – Probability of committing type II error if no subjects observe response = \( 0.80^{14} = 0.044 \)
  – Using Gehan design, treat \( n = 14 \) subjects and if > 0 responses \( \rightarrow \) move onto “stage 2” of trial

• Simon and Fleming (and others) have adapted this idea, and developed a formal two-stage design: 

• Be careful with these: make sure they make logical sense
Phase II Design Options

• Flexibility: consider restricting type II error rate, but relaxing type I error rate
• You don’t want to dismiss an IP as ineffective
• But be transparent about limitations
• Can be very similar to phase III designs, just on a smaller scale
• Analysis considerations will be similar to phase III considerations
Phase III Design Options

• Parallel group, randomized, controlled superiority trial ("usual" RCT)
  – Screen → consent → enroll → randomize → intervene → follow-up
  – May add a “run-in” period and drop/exclude participants prior to randomization
    • E.g., maybe medication dose is not stable or calibrated
    • May appear that these participants will not adhere to protocol
      • Medication schedule
      • Visit schedule
      • Too much within-participant variability

• Withdrawal study: same ideas, simply pick a pre-determined time point at which you withdraw a random subset of participants from therapy
Phase III Design Options (Cook and DeMets)

- Non-inferiority
- Bioequivalence

Figure 3.10 Non-inferiority design (absolute difference) (modified from Antman (2001)).
Visual of Non-Inferiority (Example: Mothers and Babies Study)

μ = Model-estimated adjusted 24-week QIDS score

\[ \Delta = \mu_{HV} - \mu_{MH} \]
Crossover Studies

• Randomize to intervention arm
• Follow for period of time
• Allow “wash out” period: stop therapy, and allow residual effects to wear off
• Cross over to other treatment
  – Each participant may now serve as own control
  – This may save sample size
  – It can also become very complex very quickly
• Time or period effect?
• Carry-over effect?
• Interaction? Time-by-treatment?
Factorial Designs

- Drug X and Drug Y
- Can test each simultaneously
  - Efficiency: interest is in marginal effects of each drug independently
  - Effect modification: interaction effects are of interest
- If interaction cannot be ruled out: must test for it!
Factorial Designs

- Advantages:
  - Small or no interaction $\rightarrow$ efficiency in design (two trials in one)
  - Adequately powered study $\rightarrow$ ability to test for interaction

- Special considerations:
  - Interpretation
  - Complexity and potential polypharmacy
  - Subgroup analyses: sample size/power

<table>
<thead>
<tr>
<th></th>
<th>Drug Y</th>
<th>Control</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug X</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Control</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>a+c</td>
<td></td>
<td>b+d</td>
<td></td>
</tr>
</tbody>
</table>
Cluster-Randomized Design

• Randomization at the clinician or clinic level
  – E.g., sanitation practices at a hospital
  – Prescription or quality-improvement protocols
  – Group-based therapy
• Often “pragmatic” (large, simple)
• Must consider intra-cluster or intra-class correlation
  – Are individuals within a clinic or within a clinician likely to be more similar for some reason?
  – What are the analysis units? What is the outcome of interest?
  – How many individuals per cluster? How many clusters?
Screening and Prevention Trials

- Screening trials usually very large (assuming prevalence of disease is low)
- False positive rate and false negative rate are crucial in design of these trials
- Prevention trials require identification of population at risk
  - Primary prevention (e.g., prevent individuals with arterial disease from having stroke/heart attack)
  - Secondary prevention (e.g., prevent recurrent of stroke or thrombolytic event)
  - May need special considerations for adherence
Adaptive Designs

• Many examples of this
• Randomization can be adaptive
  – With respect to arm imbalance
  – With respect to covariate imbalance
  – With respect to response
• May re-calculate sample size at an interim point
  – Is variance larger than anticipated?
  – Is event rate smaller than anticipated?
  – NOT generally done for effect size
• May stop for overwhelming efficacy, futility, or harm
Some Analysis Considerations
General Ideas

• Starts with the research question:
  – # arms?, Outcome?
  – When? (single time point [in relation to baseline?], over time [trajectory/repeated measures?])

• Clinical trials have special considerations re: analyses
• Majority of issues start with analysis population
• Others include missing data/outliers
• Multiple testing procedures
• Special topics: non-inferiority/equivalence, cross-over designs, interim analyses
Analysis Population

• In general, hope that analyzed population = randomized population, but there may be exceptions (and this is often the case):
  – Ineligibility discovered after the fact
  – Nonadherence to drug/dosing regimen
  – Special cases/outliers, poor data quality
  – Competing events

• First decision: Intent-to-treat principle vs. as treated (also called “on treatment” or “per protocol”)

• This should be pre-specified
Intent-to-Treat (ITT)

- One of most fundamental principles underlying analysis considerations for randomized clinical trials
- Two major elements:
  - Once randomized, always analyzed (regardless of adherence/dropout)
  - Analyzed according to group to which randomized (regardless of adherence)
- In an RCT: any differences observed after randomization are thought to be result of (a) chance or (b) treatment
- \(\rightarrow\) ITT allows for unbiased hypothesis testing
**Objection**: if subjects discontinue active treatment, their data will not illustrate potential benefits of treatment → bias toward no difference

− But alternatives are subject to bias as well (may be subjective assessment on whether to include/exclude participant data in analyses)
− ITT mirrors real-world scenario more closely (effectiveness): once treatment approved, adherence = just as poor (if not worse) in a real clinical setting

**Objection**: ITT can result in decreased power (dilute observed treatment effect)

− Inflate sample size at design to account for dropout and drop-ins
− If non-adherence is so great to result in large loss of power → issue with procedures

• What about non-inferiority and equivalence studies?
ITT in Non-inferiority/Equivalence Setting Setting

• ITT may dilute any observed differences across arms:
  – Superiority → potential bias toward null (conservative)
  – NI/equivalence → potential bias toward alternative-

• As treated in NI/equivalence: bias in other, difficult to determine ways (may be subjective again, may only be making inferences based on specific subgroups)

• Friedman, Furberg, DeMets: ideally, somewhere between ITT and as treated population

• Suggestion: perform analyses under both principles (same suggestion for superiority case too)
  – Compare
  – Interpret with caution if results are inconsistent
Randomization in Clinical Trials
Randomization in Clinical Trials

• Goals of randomization:
  – Control selection bias (when process is predictable)—Blinding also helps this!
  – Prevent accidental bias: control imbalance across treatment groups → precise estimates of treatment effect
  – Provide valid estimate of treatment effect (independence assumption)
• Purely random assignment does not ensure balance with respect to covariate distributions (or allocation number)
Reasons for Achieving Balance

• Efficiency
• Power in detection of treatment effect
• Face validity
• Important for interim analyses
• Secondary outcome analyses that may be unadjusted for covariates
Randomization Options

• Many options with regard to randomization or treatment allocation scheme
  – Simple random allocation, random or permuted block, urn designs, etc.
  – **Stratified or stratified block**
  – **Adaptive techniques (e.g., minimization, minimal sufficient balance, etc.)**
• Which one is ‘best’ depends on scenario of the trial...
  – In general, the most flexible designs tend to be the adaptive designs
  – A brief review of the issues follows...
Simple Randomization

- \( P(\text{assignment to one group}) \) is equal throughout
- If 1:1 allocation, \( p(\text{assignment to arm 1}) = 0.50 \)
- Easy to implement
- Easy to understand
- May result in large imbalance by chance \(\rightarrow\) loss of efficiency/power

```plaintext
data new;
  do sequence = 1 to 100;
    rand_num = ranuni(213);
    if rand_num < 0.50 then tx = 1;
    else tx = 0;
  output;
end;
run;
```

```
  tx Frequency Percent
  0    55      56.00
  1    45      45.00
```

```
proc freq data=new;
  tables tx/nocum;
run;
```
Restricted Randomization Methods

- Blocked design = most well known
  - Choose block sizes = multiple of number of arms
    - Should be > 2 for 2 arms
    - Block size of 4 for 2 arms: AABB, BBAA, ABAB, BABA, BAAB, ABBA
  - Ensures equal number after every block completed
  - Drawbacks: pre-generated, many assignments = deterministic \(\rightarrow\) selection bias?

- Others
  - Efron’s biased coin design
  - Wei’s urn design
  - Random allocation rule/truncated binomial

\[
\text{prob}(T_{n+1} = 1) = \begin{cases} 
    0.50 & \text{if } D = 0 \\ 
    p & \text{if } D < 0 \\ 
    1 - p & \text{if } D > 0 
\end{cases}
\]
Stratified Block Design

- Most commonly used method for attempting to balance covariates
- Uses blocking within strata of influential covariates
- Example: Gender (M/F) and Age (older/younger) = important predictors
- We have four strata:
  - Older males
  - Older females
  - Younger males
  - Younger females
- Within each stratum, apply the blocked design
Drawbacks of Stratified Block Design

• What if we stop the trial now?
  – Unfilled blocks: Male/ Older and Female/ Older have unfilled blocks
  – How do we really know that we are balancing age? Must categorize continuous variables

• As number of strata increases, performance = similar to simple randomization
• Example: Clinical center (assume 5), Gender (2 categories), age (4 categories: 21-30, 30-35, 36-40, >40 years), baseline disease status (mild, moderate, severe)
  – Each center has 2 x 4 x 3 = 24 strata that need to be balanced!
  – Thus, 5 x 24 = 120 strata total!
  – Requires pre-generated lists: may be electronic, sealed envelopes, pharmacy houses list, etc. opportunity for error
• Issues re: unfilled blocks and categorization are magnified
Covariate-Adaptive Methods

• AKA ‘minimization’ (Taves, Simon, Pocock [1970s])
• Choose imbalance function to minimize (range, variance) for each variable ($D_i, i=1,...,# variables$)
• Weight each variable wish to balance ($w_i$)
• Let overall imbalance = $D = \sum w_i D_i$
• For incoming subject, calculate $D$ under assignment to each possible arm
• Assign subject to arm with smallest $D$ with higher probability (0.5,1]
• Well known, less commonly implemented than stratified block
• More recent methods can handle both categorical and continuous variables (e.g., Minimal sufficient balance [Zhao et al., 2012])
Minimization: Example

- Incoming subject = Male, BMI <30 kg/m², Cholesterol >6.0 mmol/l
- Use ‘range’ as measurement of imbalance
- Use equal weight for each of these variables
- Assign to treatment A:
  \[ \text{Imbalance} = |5-5| + |5-3| + |4-2| = 4 \]
- Assign to treatment B:
  \[ \text{Imbalance} = |4-6| + |4-4| + |3-3| = 2 \]
- Minimize imbalance by assigning to treatment B
- Use probability of assignment to B = (0.50, 1]

### Features of 17 Subjects Entered Into a Trial of Obesity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male¹</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>&lt;30¹</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>≥30</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Fasting cholesterol (mmol/l)</td>
<td>≤6.0</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>&gt;6.0¹</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Total number of subjects</td>
<td></td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>already allocated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹ Values for next subject to be allocated.

McEntegart 2005; Drug Information Journal
Minimization/Covariate-Adaptive Methods

• More flexible: adaptive, weighting, more covariates, differing variable types (categorical, continuous, etc.)
• More difficult to guess treatment assignment when balancing several covariates
• Does not handle imbalance as well as stratified block in presence of interactions
• Complex: requires algorithmic feedback on ongoing basis
  – Interactive voice response
  – Web-based
  – Need to consider: back-up, speed of process, 24-hour availability
• Taves (2010) reports <2% of published randomized clinical trials use minimization
Things to Consider when Choosing an Allocation Scheme

• Study workflow and setting
  – Time lag between consent, enrollment, randomization
  – Blinding capabilities
• Secure data transfer
• Control over system/randomization dashboard access
• Control of ability to randomize
• 24-hour availability
• Back-up, speed of process
• Testing!!!
Handling Baseline Variables in Analysis
Analysis Techniques to Account for Randomization

• There is inconsistency in literature about appropriate methods of analysis.

• Most analyses do not take randomization scheme in its true form into account in analyses (e.g., accounting for blocks, inducing correlation among observations).

• General consensus: minimal impact if failure to account for blocks in analysis (tends to be slightly conservative).

• **Permutation tests** are one method (not common, but a good option).

• My suggestion re: covariates: if try to balance covariate at allocation → plan/try to **account for covariate in analyses** (Ciolino et al., 2011; 2013; Friedman, Furberg, DeMets text).
Why are confounders still a problem in analysis of clinical trials?

Chance imbalances can affect:

- **Power**
- Type I error rate
- Bias in treatment effect estimates (over/underestimation is possible)

\[
\gamma(d_z) = \text{prob} \left[ Z \geq \frac{Z_{\alpha}}{\sqrt{1 - \rho^2}} - \frac{d_z\rho}{\sqrt{1 - \rho^2}} \frac{\Delta}{\sqrt{(1 - \rho^2)\sigma_1^2(\frac{1}{n_1} + \frac{1}{n_2})}} \right]
\]

Senn, 1989; Ciolino et al., 2011

**Imbalance across two arms favoring control arm**

\[\rho = \text{cor(baseline variable, outcome)}\]
Why are confounders still a problem in analysis of clinical trials?

Chance imbalances can affect:
- Power
- **Type I error rate**
- Bias in treatment effect estimates (over/underestimation is possible)

\[
\alpha(d_x) = \text{prob} \left[ Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho \bar{d}_x}{\sqrt{(1-\rho^2)\sigma_x^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right] \\
= \text{prob} \left[ Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho \bar{d}_x}{\sqrt{1-\rho^2}} \right]
\]

Senn, 1989; Ciolino et al., 2011

Imbalance across two arms favoring active arm \( \rightarrow \) [\( \rho = \text{cor(baseline variable, outcome)} \)]
Analysis at the End of the Study

• Good news!
  • Appropriate adjustment *often* solves many of the statistically-related problems (Ciolino et al. 2011, 2014; Raab and Day 2000; Ford and Norrie 2002)
    – Increases precision on treatment effect estimate
    – Decreases bias in treatment effect estimate
    – → tends to preserve type I error rate and power

• Bad news?
  – We can’t adjust for everything
    – *Sometimes the benefit of adjusted analyses depends heavily on nature of outcome and magnitude/direction of imbalance* (Gail et al. 1984; Greenland 1999; Hauck et al. 1998; Ciolino et al. 2013)
      • Binary outcome/nonlinear relationships
      • Precision may decrease and unadjusted estimates ≠ adjusted estimates
At the End (Analysis)

• When in doubt, adjust
• CONSORT (2009):
  – Adjustment may be ‘sensible, especially if one or more variables is thought to be prognostic’ (Journal of Clinical Epidemiology, 2010)
  – Ideally...pre-specified in the protocol or analysis plan

Continuous Outcome:
Adjusted vs. Unadjusted p-value
ZERO correlation w/baseline variable and outcome

Binary Outcome:
(Simulated data; Ciolino 2013)
At the End (Analysis)

- What we should not be doing:
  - Allow baseline test for significant differences to dictate adjustment (Senn, Ciolino et al., CONSORT)
  - Failing to pre-specify or transparently explain post hoc decisions to adjust
  - CONSORT (J Clinical Epidemiology, 2010)

- “Unfortunately significance tests of baseline differences are still common…”
- “[these tests]…assess the probability the observed baseline differences could have occurred by chance; however, we already know that any differences are caused by chance.”
- “illogical”, “superfluous”, and misleading
- “...comparisons at baseline should be based on consideration of prognostic strength and the size of any chance imbalances.”

Table 1: Baseline Characteristics of the Study Participants

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.84</td>
</tr>
<tr>
<td>Sex</td>
<td>0.77</td>
</tr>
<tr>
<td>Race or Ethnic</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Some Anecdotes

Common questions/comments from collaborators:

- Shouldn’t the randomization take care of it?

- There are no ‘significant differences’ at baseline, so we don’t need to worry (our randomization ‘worked’)

- We stratified, so these variables should be balanced

- On average, yes; there is no guarantee (every trial will exhibit some baseline variable imbalance)

- Not necessarily (even ‘insignificant’ imbalances have an impact [if we fail to adjust] on analyses)

- See above + stratification may not always help the cause
Some Anecdotes, cont’d

Common questions/comments from collaborators:

• Can’t we just adjust for these in analyses?

• Yes, but...
  – What about face validity?
  – What if we have too many variables for which we’d like to adjust?
  – We can’t adjust for everything nor do we know all influential variables ahead of time
  – Unadjusted effect ≠ adjusted effect
Areas of potential gap between ideal/theory/guidelines and practice/real

• Dominant use (69%) of stratified block despite shortcomings
• 11% employ covariate-adaptive methods, with less prevalence over time
• “substantial and confusing variation...in handling baseline covariates” (Austin et al. 2010)
  – 10% of the time unable to determine allocation technique
  – ‘unclear’ as high as 23% of the time (may be related to number of arms/trial complexity)
  – Superfluous test of baseline differences in 43% of trials (similar to 38% in review by Austin et al. in 2010)
Why the gap?

- Lack of education/understanding...refer to anecdotes
  - Over-simplification of design (‘it’s just a simple/small trial’)
  - Poor planning/time commitment to design and a pre-specified analysis plan
  - Sometimes a ‘black box issue’
- Programming/software requirements and expense
- Lack of statistician or programmer involvement from beginning to end
- Individual trial logistical complexities overpower design and analysis considerations
My Take-home Messages

• We should be thinking about baseline variables in design and analysis phase even though RCTs are ‘randomized’
• Increased education and collaborative efforts can help mitigate the gap we are seeing
• But sometimes logistical or practical constraints simply cannot be avoided
  – Something can (and will) always come up
  – We cannot predict everything with 100% certainty when designing a study (think of a simple sample size calculation – based on assumptions)
  – In these situations: critical thinking (‘trickle down effects’); involvement of a statistician throughout; compromises between ideal and real; transparency in reporting
Overall Summary

• Clinical trials have much flexibility with design/implementation
• Design should always be driven by question/phase of IP, which will in turn drive analyses/sample size
• Always an iterative process
• “There is no such thing as a perfect study” (FFD)
• Pre-specify as much as you can and be transparent about reporting
End. Thank you! Questions?

jody.ciolino@northwestern.edu