Statistics for Clinical Trials: Basics of Phase III Trial Design

Gary M. Clark, Ph.D.

Vice President
Biostatistics & Data Management
Array BioPharma Inc.
Boulder, Colorado USA

NCIC Clinical Trials Group New Investigator Clinical Trials Course
August 12-14, 2015

Presenter’s Conflict of Interest

- Gary Clark is a full-time employee of Array BioPharma Inc.
- Gary Clark owns stock and has stock options in Array BioPharma Inc.
- However, the presentation today reflects the personal opinions of Gary Clark and not necessarily those of Array BioPharma Inc. or its partners

Outline of Presentation

- Historical vs. randomized controls
- Intent-to-treat principle
- Two-arm and multi-arm designs
- Superiority, equivalency, non-inferiority
- Interim analyses
- Time-to-event endpoints
- Sample size issues

Historical vs. Randomized Controls

Historical Controls

- Patients are unlikely to be comparable
  - Large patient heterogeneity
  - Unknown prognostic factors → Selection Bias
  - Cannot specify definitions for efficacy endpoints
- Choice of controls

Randomized Controls

- Patients are likely to be comparable
  - Can balance (stratify) on known prognostic factors
  - Unknown factors more likely to be balanced
  - Can specify definitions to be used in both arms
  - Can specify timing of assessment of efficacy endpoints

The Randomized Comparative Trial

Primary purpose/aim:

- Assess the efficacy of new treatment(s) relative to control treatment

Patients assigned at random to treatment(s) or control (considered the gold standard)

- Advantages
  - Eliminates assignment bias
  - Balance known and unknown factors
  - Basis for valid statistical tests
- Disadvantages
  - Generalizability of results
    - Selected patients based on inclusion/exclusion criteria
    - Volunteer effect
  - Acceptance of the randomization process
    - By patients and investigators

Randomization

Common Randomization Techniques

- Simple Randomization
- Block Randomization
- Stratified Randomization
- Dynamic Balance / Minimization

Simple Randomization

Examples:

- Toss a coin: H → arm A; T → arm B
- Random digit: Even #→ arm A; Odd # → arm B

Pros & Cons

- Pro: easy to implement
- Con: potential for imbalance in the number of patients on each treatment arm
  - With n=20, chance of a 12:8 split (or worse) ~50%
  - With n=100, chance of a 60:40 split (or worse) >5%
  - Chances decrease with larger n

Permuted Blocks

Blocks of k patients are created such that balance is enforced within each block. One of the blocks is then selected at random and the k patients are assigned accordingly.

Examples:

- Block size=4: AABB, ABAB, ABBA, BAAB, BBAA, BABA
- Block size=6: 20 different arrangements

Pros & Cons

- Pros: promotes group balance at end of study; also periodic balance in the sense that sequential patients are distributed equally between groups
- Cons: susceptible to selection bias: AAB? *(blinding!)*

Stratified Randomization

If a factors are known to affect outcome, stratify by those factors, then randomize within each stratum (simple or block randomization).

Example:

- Gender (male, female) and Age (<40, 40-60, >60) produce 6 strata
- Institution/site often included as a stratification factor

Pros & Cons

- Pros: insures balance within risk groups (most beneficial for small studies)
- Cons: over-stratification (too many factors) leads to sparse data which causes statistical problems.

Dynamic Balance / Minimization

- Balances treatments simultaneously over several factors
- Does not balance within strata; balances over the marginal totals of each stratum separately
- Is used when the number of strata is large relative to sample size
- Institution/site is usually one of the stratification factors

Pros & Cons

- Pros: achieve balance over a large number of covariates when the sample size is small to medium
- Cons: potential for overmatching; regulatory concerns about potential impact on subsequent analysis

See EMA Guideline on Adjustment for Baseline Covariates in Clinical Trials.
http://www.fdanews.com/ext/resources/files/03-15/03-30-15-covariates.pdf?1427736886

Phase III Studies: Key Points

- Traditionally, fixed sample size or multi-staged
- Involve large numbers of patients
- Frequently use resources from several institutions
- Commonly employ pre-defined interim analysis rules
- Require Data and Safety Monitoring Boards
- Primary analysis based on ‘intent-to-treat’ principle

Intent-to-treat Principle

- Eligibility
  - Known at time of randomization
  - Sometimes confirmed (or not confirmed) after randomization
- Deviations
  - Based on events after randomization

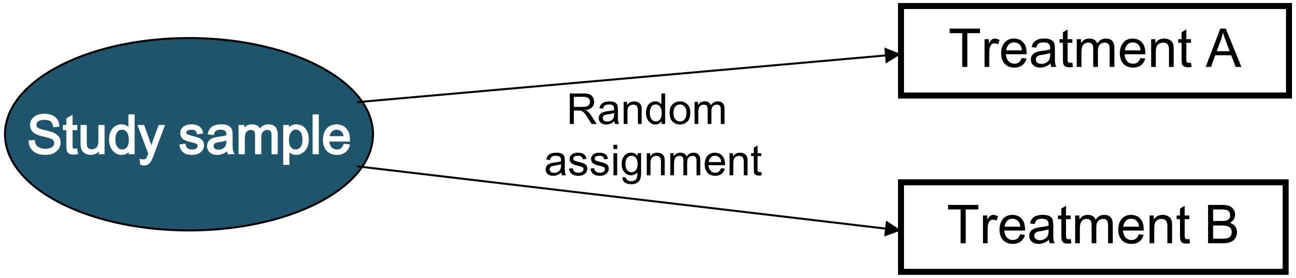
Intent-to-treat Principle

- Analyze all eligible patients on their randomized arm, without regard to treatment deviations
  - Clinical trials address practical questions
    - Deviations occur in practice
- Excluding patients with treatment deviations destroys comparability achieved by randomization

Two-Arm Parallel Design

- Simplest & most common
- Random allocation
- *Between* patient comparisons
  - each patient receives only 1 treatment or treatment regimen

Schema



Study sample

Random assignment

Treatment A

Treatment B

Two-Arm Parallel Design

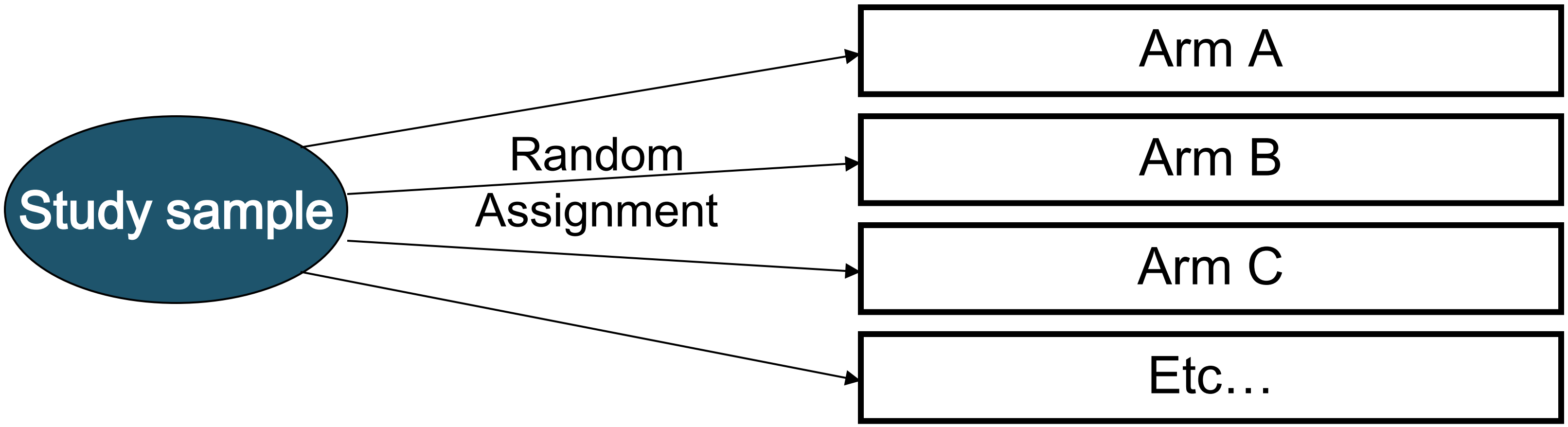
- Advantages
  - Simple
  - General use
  - Valid comparisons

- Disadvantage
  - Few study questions

Sample size is based on simple A vs. B comparison

Multi-Arm Parallel Design

Schema



Multi-Arm Parallel Designs

- Advantages
  - Can address more study questions
- Sample size
  - Depends on number of questions of interest
    - May have several competing standards
    - May have several experimental treatments vs. standard
- Problem of multiple comparisons
  - Probability of false positive conclusions is inflated
  - Do overall test before doing pairwise comparisons
  - Adjust each treatment comparison

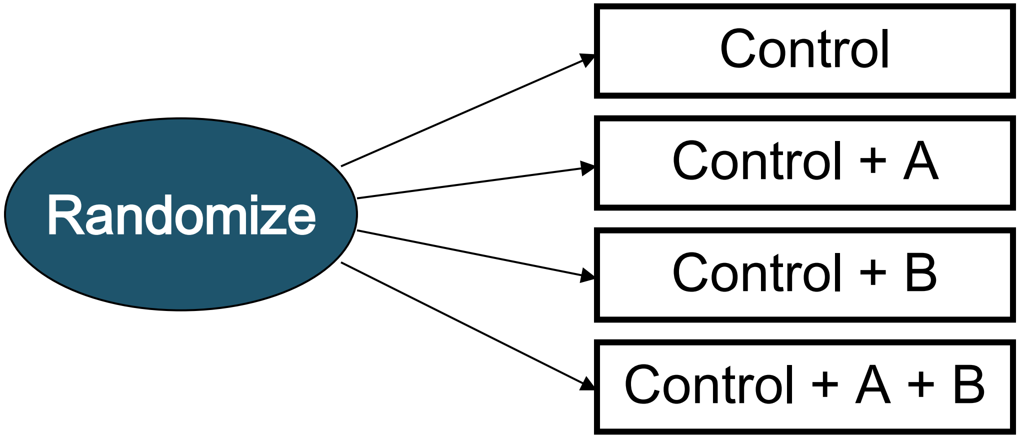
Multi-Arm Parallel Designs

- How many comparisons will we have?
  - Depends on number of questions of interest (also number of competing control or standard treatments)
  - All pair-wise comparisons?
      - 3 arms: (A vs. B, A vs. C, B vs. C)
      - 4 arms: (A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, C vs. D)
    - Experimental arms to control only?
      - 3 arms: (A vs. B, A vs. C)
      - 4 arms: (A vs. B, A vs. C, A vs. D)
    - An ordering?
      - 3 arms: (A < B < C)
      - 4 arms: (A < B < C < D) or (A < [B or C] < D), etc.
  - Number of possible comparisons increases as number of arms under study increases
  - Do not do pairwise tests unless overall test is significant at prespecified α; then adjust α for subsequent pairwise comparisons

Factorial Design

- Special case of parallel design
- Least complex factorial design has two new drugs (A and B) and four treatment regimens

Schema



Randomize

Control

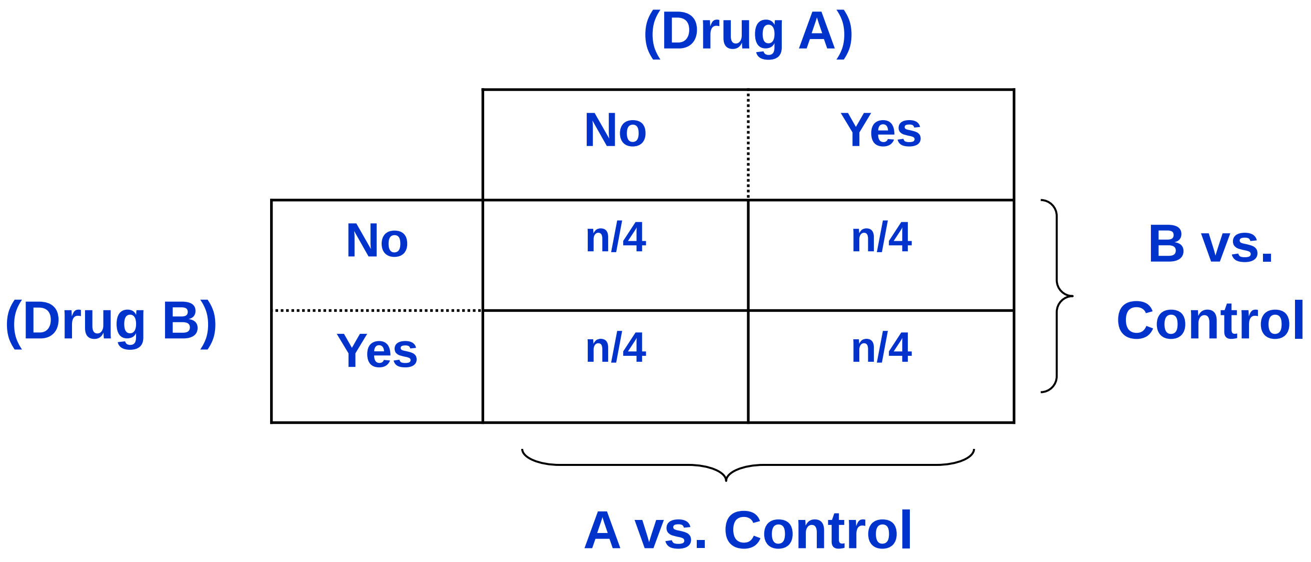
Control + A

Control + B

Control + A + B

Factorial Design

- Random allocation to all four groups
  - (Control, Control + A, Control + B, Control + A + B)
- Two main comparisons
  - A vs. Control, B vs. Control



		(Drug A)	
		No	Yes
(Drug B)	No	n/4	n/4
	Yes	n/4	n/4

B vs. Control

A vs. Control

Factorial Design

- Advantages
  - Two studies for one?
  - Discover interactions

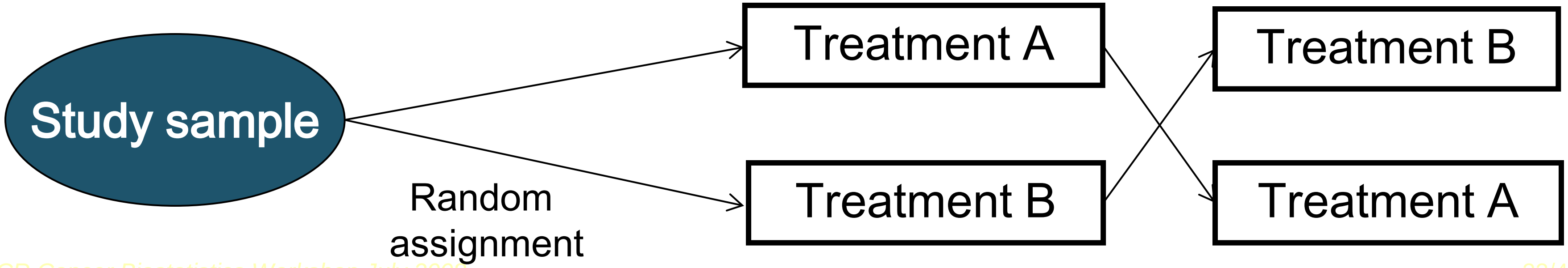
- Disadvantages
  - Test of main effects assumes no interaction
  - Often inadequate power to test for an interaction

    (effect of A differs depending on the presence or absence of B & vice versa)
  - Compliance

Crossover Design

- Initial randomization
- Crossover at a predefined event or point in time
  - Often only crossover from control to experimental treatment after documented disease progression
- If same endpoint, need to be careful about “carryover” effect (may need washout period)
- If different endpoint (eg, PFS, then OS), need to be careful about subsequent treatments

Schema



Types of Comparisons
(two groups)

New treatment versus Standard (or active control)

- Superiority trials
  - Hope that new treatment will prove ‘superior’ to standard
  - Use one or two-sided tests
- Equivalency trials
  - New treatment and standard are ‘similar’ (neither better nor worse)
  - Use two-sided tests
- Non-inferiority trials
  - New treatment is ‘not worse’ than standard
  - Use one-sided tests

Superiority Trials

- Motivation
  - New treatment will prove ‘superior’ to standard therapy
- Benefit of new treatment
  - More effective
- Must specify a superior difference (denoted as Δ)
- Test new treatment versus standard
  - New better by pre-specified Δ

    ± Δ → two-sided alternative

    + Δ → one-sided alternative

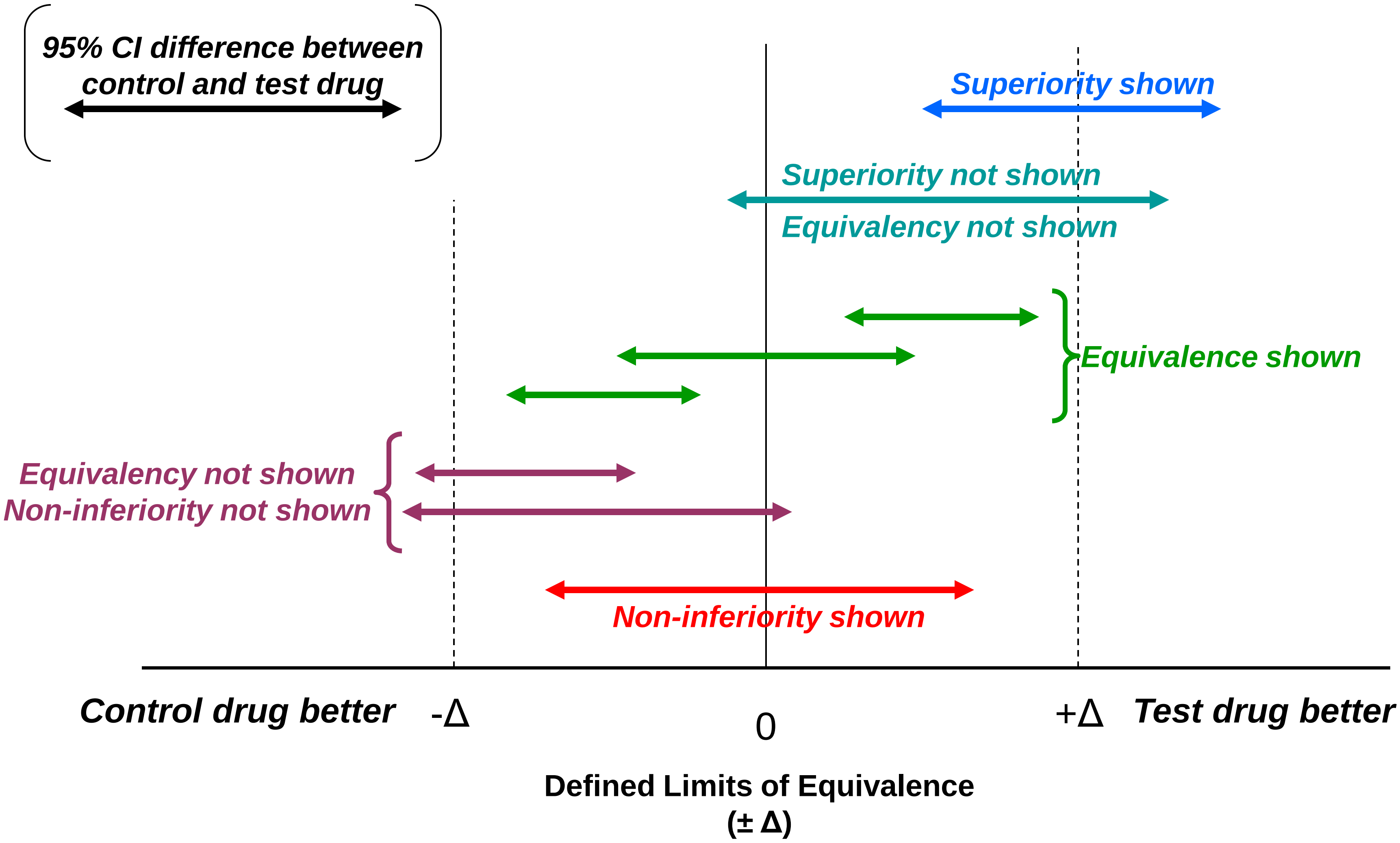
Equivalency Trials

- Motivation
  - New treatment is ‘as effective’ as standard therapy
- Benefit of new treatment
  - Less adverse events
  - Less expensive
  - Easier to administer
  - Profit (‘me too’)
- Proving ‘equal’ effectiveness is not possible
  - Must specify range of ‘equivalence’, denoted as Δ
- Test new treatment versus standard
  - New does not differ by ± Δ → two-sided alternative
- Sample size is much larger than for superiority trial

Non-Inferiority Trials

- Motivation
  - New treatment is ‘not worse’ than standard therapy
- Test new treatment versus standard
  - New is at least $-\Delta$ → one-sided alternative
    - Δ must be pre-specified
  - Would have beaten placebo if a placebo arm had been included (regulatory requirement)
- Challenges
  - Requires high quality control & assay sensitivity
    - The ability of a study to distinguish between active and inactive treatment
  - Specifying Δ
    - Must include an assessment of difference between standard and placebo
    - Sample size is much larger than for superiority trial

Types of Comparisons



Statistical Methods for Interim Analyses

- Most large comparative trials provide for interim analyses of efficacy and/or safety
- Purposes include determining if the trial should be closed early for:
  - Issues with patient safety
    - Adverse events are too severe
    - Treatment compliance is too low
  - Treatments under study are convincingly different (or similar)
  - Demonstration of target difference of the experimental regimen(s) is unlikely (futility)
  - To provide some direction for the planning of the next study

Guidelines for Interim Analyses

- Basic approach should be included in the protocol during the design phase of the study
- To avoid the ‘repeated testing problem’ common design approaches include:
  - Group sequential methods
    - Specify number of interim looks and probability of stopping
    - O’Brien-Fleming Boundary (or others)
    - Lan & DeMets alpha spending function
  - Triangular Test
  - Conditional power or stochastic curtailment
  - Adaptive monitoring
  - Futility
- The choice of which to use *varies* greatly!
- The method employed should be viewed as a monitoring guideline and *not* a rigid rule to be followed

Time-to-Event Endpoints

From a statistical perspective

- Any time-to-failure or time-to-event endpoint, provided that the “failure” or “event” is unambiguously defined

Examples of Time-to-Event Endpoints

- Overall Survival
- Disease-specific survival
- Progression-free survival (PFS)
- Disease-free survival (DFS)
- Time to progression (TTP)
- Time to treatment failure (TTF)
- Duration of response
- Time to deterioration of QoL/symptoms
- Time to tumor doubling (animal studies)

Sample Size Issues for Comparative Trials

How many patients?

- Estimates are approximations
  - Uncertain assumptions
  - Over optimism about treatment effect
- Need a series of estimates
  - Vary assumptions, pick most reasonable
- Be conservative yet reasonable

Factors that Affect Sample Size

- Number of study arms
- Allocation ratio
- Effect size to be detected (clinically important difference and expected variability)
- The test statistics used to analyze the data:

  Type I and Type II errors

  α = P(type I error) = P(false positive)

  → exposure to ineffective treatment

  β = P(type II error) = P(false negative)

  → active agent may be missed

  Power = $1 - \beta$ = P(true positive)

Additional Factors that Affect Sample Size for Time-to-Event Endpoints

- Sample size refers to number of events, not number of patients
- Need to specify:
  - Accrual rate or accrual duration
  - Minimum (or maximum) length of follow-up for each patient

Complicating factors

- Lost to follow-up
  - Patient lost before final outcome observed
- Drop out
  - Patient stops taking protocol therapy
- Drop in
  - Patient starts taking other protocol therapy

➢ All of these dilute effective sample size and impact the observed treatment effect

➢ Therefore, need to adjust sample size to compensate for dilution effects

Sample Size Calculations

Will be covered by Chris O'Callaghan in:

Workshop 1: Sample Size Determination, Methodology, Analysis and Philosophy

Take-Home Message

- Design of a Phase III trial requires a multidisciplinary team
- Many decisions need to be made before a successful clinical trial protocol can be written
- Biostatisticians should be included early in these discussions and should be a collaborator throughout the study