Design and Analysis of Phase III Clinical Trials

Fei Ye, Ph.D.

Cancer Biostatistics Center, Biostatistics Shared Resource,
Vanderbilt University School of Medicine

June 19, 2008

Outline

Phase I Trials: Safety, Dosage Range, and Toxicity

Objective

To determine a safe dose, identify dose-limiting toxicities (DLTs) and the maximally tolerated dose (MTD), and understand drug metabolism.

Design

Usually single or multiple dose-escalation studies with a small number of patients. Approximately $20 < N < 80$.

Phase II Trials: Initial Clinical Investigation for Treatment Effect

Objective

Efficacy and Toxicity: to evaluate how well the treatment works and to continue safety (short-term side effects and risks) assessment.

Design

Often single arm; to be compared with historical controls or current treatment. Usually $N < 100$.

Phase III Trials: Full Scale Evaluation of Treatment

Objective

To compare efficacy of the new treatment with the standard treatment.

Design

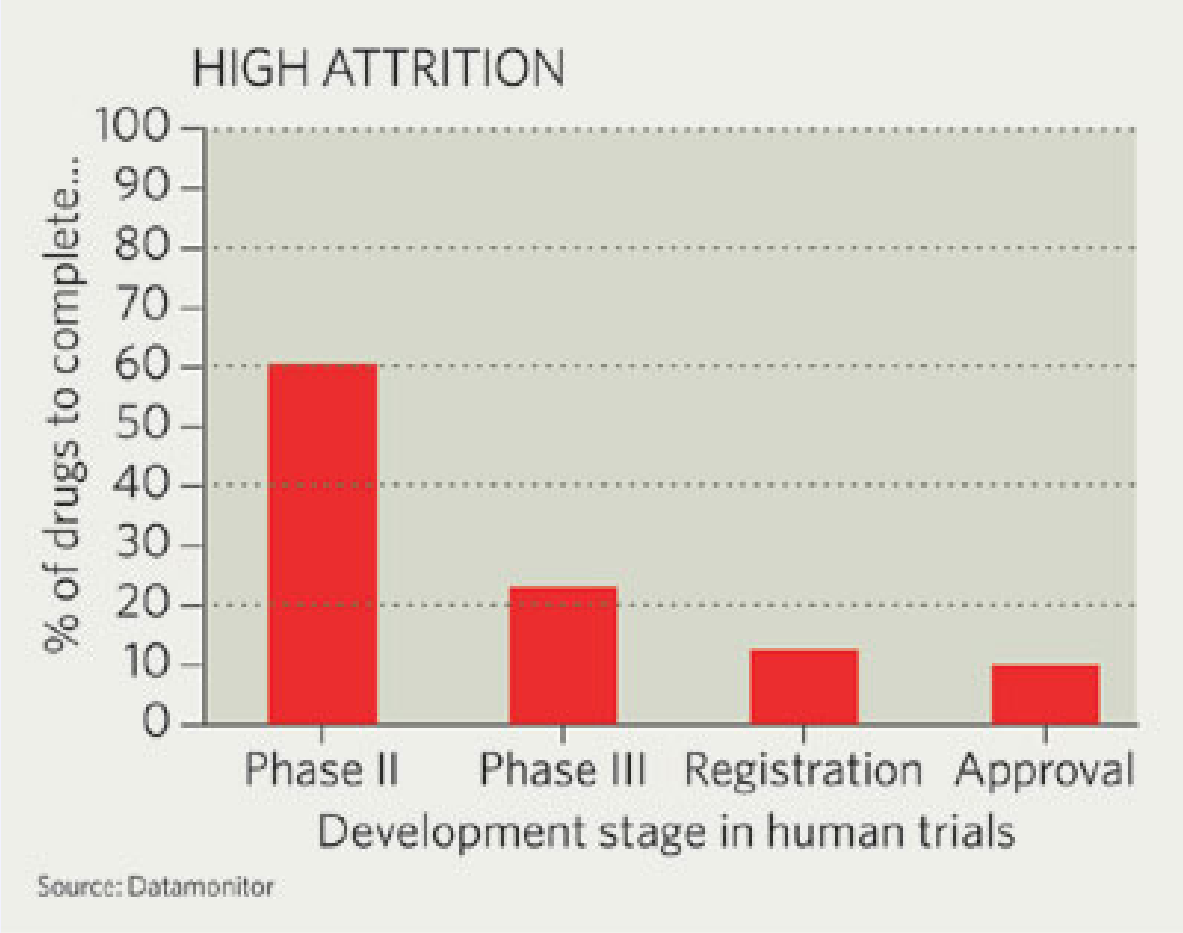
◇ Often randomized controlled multicenter trials and are aimed at being the definitive assessment of how effective the drug is, in comparison with current 'gold standard' treatment (**control**). Usually $100 < N < 3000$.

◇ Phase III trial is most rigorous and extensive type of scientific clinical investigation of a new treatment.

Phase IV Trials: Post-Marketing Surveillance Studies

- Evaluation of efficacy and detection of rare or long-term adverse effects over a much larger patient population and longer time period.
- Evaluation of healthcare costs and outcomes.
- Pharmacogenetics (assess genetic responses).

How many drugs will be proven?

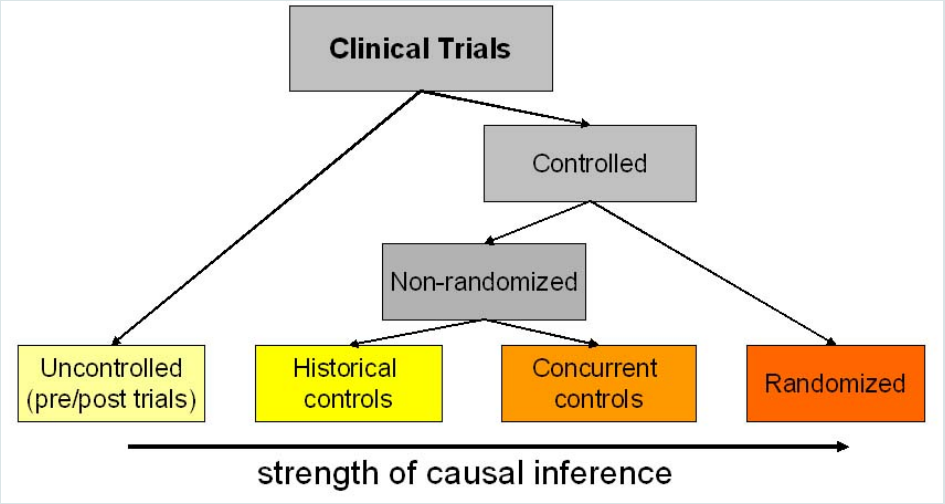


Experiment Design

Common Phase III Trial Designs

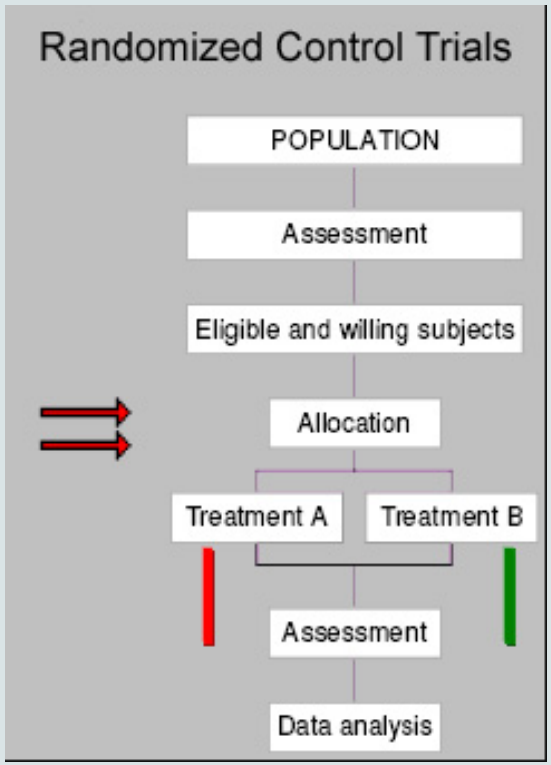
- **Randomized Control Trials (parallel design)**
- Uncontrolled Trials (single-treatment)
- Historical Controls
- Non-Randomized Concurrent Trials
- **Crossover Designs**
- **Factorial Designs**
- Group Sequential Design
- ...

Phase III Clinical Trial Designs



Randomized Controlled Trials (Parallel Designs)

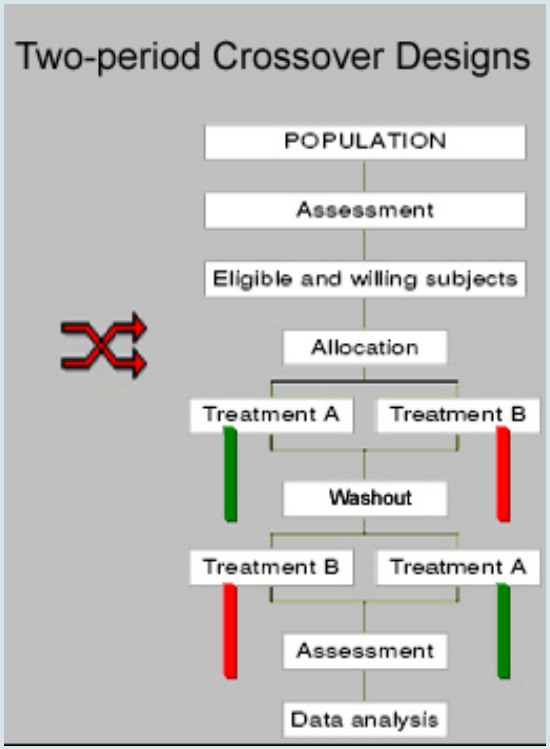
Parallel Design



- Each patient receives one and only one treatment in a random fashion.
- It is simple and easy to implement; applicable to acute conditions.
- The interpretation of the results is straightforward.
- Good if the interpatient variability relatively small compared to the intrapatient variability.

Cross-over Designs

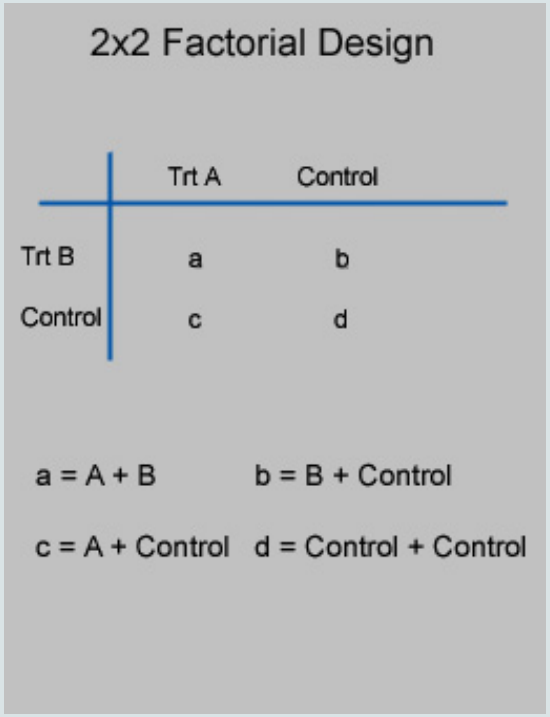
Two-period Cross-over Design



- Same patients receive different treatments at different dosing periods.
- Patients serve as their own controls: removes the interpatient variability from the comparison between treatments.
- Typically requires a much smaller sample size than a parallel design.
- Should only be used for chronic diseases.
- Need to be careful with "carryover" effect $\longrightarrow$ **washout** period.

Factorial Designs

2x2 Factorial Design



- Attempts to evaluate two or more treatments in a single experiment.
- Allows to evaluate the interaction effect between treatments.
- Many factorial trials lack sufficient power to look at the interaction effect.
- 2x2x2 or more complex design: require large sample size and the results are more difficult to interpret.

Randomized vs. Nonrandomized

Anticoagulant therapy studies - A

In **32** studies on the use of anticoagulant therapy in patients with acute myocardial infarctions, **18** used historical controls with 900 patients, **8** used nonrandomized concurrent controls with 3000 patients, **6** were randomized trials with 3800 patients.

15/18 (83%) studies with historical controls and **5/8 (63%)** with nonrandomized controls showed statistically significant results favoring anticoagulation therapy.

Randomized vs. Nonrandomized - *cont.*

Anticoagulant therapy studies - B

Only **1/6 (17%)** randomized control trials showed significant results in support of the therapy.

WHY?? - **Selection bias** in the nonrandomized trials being similar to the *presumed* true effect, could have yielded positive answers even if the treatment had no benefit.

The Purpose of Randomization

What is Randomization?

Randomization is a process that assigns research patients by chance, rather than by choice, to either the treatment group or the control group.

Why Randomize?

- To remove bias in patient allocation to treatments.
- To produce more comparable groups (w.r.t. risk factors).
- To allow statistical tests to have valid significance levels.

Simple Randomization: to overcome imbalance

Examples

◇ Toss a coin: H $\rightarrow$ treatment; T$\rightarrow$ control.

◇ Random digit: Even# = treatment; Odd# = control.

Pros&Cons

◇ Pro: Easy to implement.

◇ Con: Imbalance in the number of patients on each treatment. With n=20, the chance of a 12:8 split or worse is $\approx$ 50%.

◇ **Note**: most statistical tests are most powerful when the groups being compared have equal sizes, it is desirable for the randomization procedure to generate similarly-sized groups.

Permuted Block Randomization

PBR: to avoid serious imbalance

In this form of restricted randomization, 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 $\Rightarrow$ AABB, ABAB, ABBA, BAAB, BBAA, BABA.
◇ Block size=6 $\Rightarrow$ $C\binom{6}{3} = 20$ different arrangements.

Pros&Cons

◇ Pros: promotes *group balance* at the end of the trial, also *periodic balance* in the sense that sequential patients are distributed equally between groups.
◇ Cons: susceptible to selection bias: $AAB?$ $\Rightarrow$ **blinding!**

Stratified Randomization

Stratification: to control patient heterogeneity

◇ If a covariate (e.g., age, gender, race, or center) is known to be the cause of heterogeneity among patients, patients are stratified into several homogeneous strata w.r.t. the covariate. Randomization is then performed *within* each stratum (usually blocked).

◇ Stratified randomization guarantees treatment balance within risk factors. An extreme case of stratification: **matching**! (common in case-control studies)

Example

Age	Male	Female
< 40	ABBA, BAAB, ...	BABA, BAAB, ...
$40 - 60$	AABB, ABBA, ...	BAAB, ABAB, ...
> 60	BBAA, ABAB, ...	ABAB, BBAA, ...

Recommendations

Large Studies N: hundreds or more

◇ One center: Blocked Randomization.

◇ Multi-center: Stratified Randomization Blocked by Center.

Small Studies $N \approx 100$

◇ One center: Stratified Randomization Blocked by 1 or 2 Risk Factors.

◇ Multi-center: Stratified Randomization Blocked by Center+Risk Factors.

Blinding - To further reduce bias

The blinding/masking is used to prevent research outcomes from being influenced by either the placebo effect or the observer bias, by blocking identity of the treatments to eliminate such bias.

- Open label: phase I trials, trials for surgical procedures, premarketing & postmarketing surveillance studies.
- Single-blinded: patients
- Double-blinded: patients & investigators
- Triple-blinded: patients & investigators & sponsor (the project clinician, the CRA, the statistician, etc.)

◇ A classic *randomized*, *double-blinded* phase III trial.

Did blinding work?

At the completion of study, ask patients and investigators to guess which treatment/group the patient was on. If blinding worked, #correct guesses = 50%. $> 50\%?$ $< 50\%?$

Bias

◇ Let p= % who know their treatment assignment:
Expected proportion of correct guesses $= p + (1 - p)/2$.
◇ The *expected bias factor* is:
(#correct guesses - #incorrect guesses)/2.

Example

◇ If we observe 75% correct guesses, $0.75 = p + (1 - p)/2 \Rightarrow p = 0.50 \Rightarrow 50\%$ of people know their treatment assignment.
◇ If N=100 and p=75%, The *expected bias factor* is estimated to be $(75 - 25)/2 = 25$.

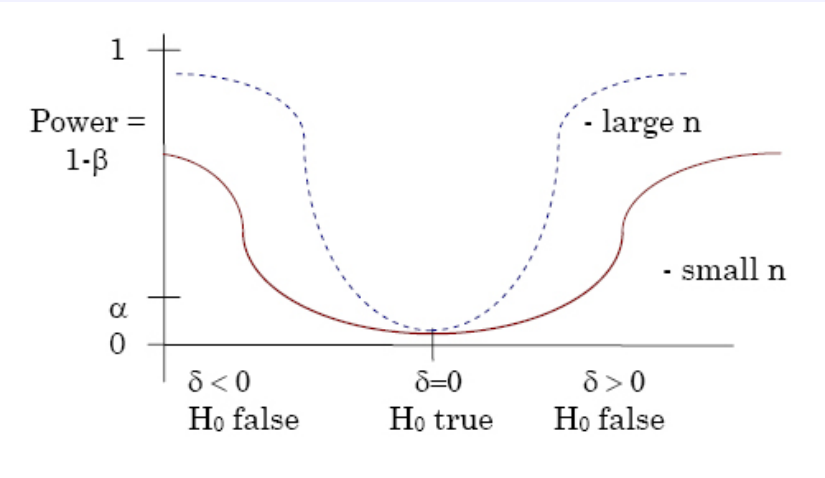
Preparing to Calculate Sample Size

Sample Size = Fn. (α, Power, Clinical significant level, Variation).

- What is the main purpose of the trial? (the question on which sample size is based).
- What is the principal measure of the patient outcome (endpoint) continuous or discrete? Censoring?
- What result is anticipated with the standard treatment? (e.g., average value or rate).
- What statistical test will be used to test treatment difference (e.g., t-test, logrank, chi-square)? One-tailed or two-tailed?
- How big a risk can be taken if the drug is concluded effective when in fact it is NOT? (This is the level of significance, type I error rate α).
- What is the smallest difference between treatments that is important to detect (δ), and with what degree of certainty? (power, 1-power=type II error rate)
- What is the size of the variance (σ^2)?

Power and Size

- Calculation of a proper sample size is necessary to assure adequate levels of significance and power to detect differences of clinical interest. The size of a trial should be considered early in the **planning phase**.
- Biggest danger: Sample size too small $\Rightarrow$ No significant difference found $\Rightarrow$ Treatment is discarded which may be useful.
- Note: $\alpha \downarrow$ as $\beta \uparrow$ and $\alpha \uparrow$ as $\beta \downarrow$.
  How to decrease both error rates?? **Increase the Sample Size!**



Sample Size Determination - Two Samples (A)

Dichotomous

$$H_0 : P_A = P_B \text{ vs. } H_1 : P_A \neq P_B$$

EX: A trial is planned to evaluate an anti-infective drug compared to active control. The response rate of the active control drug is assumed to be **80%** (P_A) based on previous studies; a difference of **10%** between the two drugs is interested (δ); $\alpha = \mathbf{5\%}$ and *power* $= \mathbf{80\%}$.

$$\begin{aligned} n &= \frac{[Z(\alpha/2)\sqrt{2\overline{P}(1-\overline{P})} + Z(\beta)]\sqrt{P_A(1-P_A)+P_B(1-P_B)}}{(P_A - P_B)^2} \\ &= \frac{[(1.96)\sqrt{2(0.85)(0.15)} + 0.842\sqrt{(0.9)(0.1)+(0.8)(0.2)}]^2}{(0.1)^2} \\ &= 199.02 \approx 200 \quad (\textit{per arm}). \end{aligned}$$

Sample Size Determination - Two Samples (B)

Continuous

$$H_0 : \mu_A = \mu_B \text{ vs. } H_1 : \mu_A \neq \mu_B$$

EX: A trial is designed to compare two cholesterol lowering drugs: the clinically meaningful difference in LDL-C is **8%** (δ); the standard deviation is assumed to be **15%** (σ); $\alpha = \mathbf{5\%}$ and *power* $= \mathbf{80\%}$.

$$\begin{aligned} n &= \frac{2\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\delta^2} \\ &= \frac{2(15)^2[1.96 + 0.842]^2}{(8)^2} \\ &= 55.2 \approx 56 \quad (\textit{per arm}). \end{aligned}$$

Sample Size Determination - Multiple Samples

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs. H_1 : At least one of μ_i's are not equal to the others.

Analysis of Variance Models (ANOVA)

- The power can be expressed as a function of cumulative noncentral F distribution.
- No explicit form exists for calculating the required sample size.
- Normal approximation can be used to determine the sample size (Laubscher 1960).
- Not easy to calculate "by hand".

Generalized Linear Models (GLM)

- No explicit form exists for calculating the required sample size.
- Power and sample size calculations based on noncentral χ^2 approximation.
- Trials with repeated measurement (correlated observations) (Liu & Liang 1995).
- Not easy to calculate "by hand".

Sample Size Determination - Censored Data

$$H_0 : M_A = M_B \text{ vs. } H_1 : M_A > M_B$$

- Often sample size calculations are based on the assumption that time-to-event/censoring has an exponential distribution.
- If further assume uniform censoring time, then the hazard rate $\hat{\lambda}_i = \frac{\#Events}{Total\ \ follow-up\ \ time}$ approximately follows a normal distribution with mean λ_i and variance $\phi(\lambda_i)/n$:

$$n = \frac{[Z(\alpha)\sqrt{2\phi(\overline{\lambda})}+Z(\beta)\sqrt{\phi(\lambda_1)+\phi(\lambda_2)}]^2}{(\lambda_2-\lambda_1)^2}.$$

- Accrual throughout the study requires more patients than if all start at beginning of the study: choose appropriate $\phi(\lambda_i)$.

Sample Size Calculation Software - Too much formula?

- **SAS**: PROC POWER; PROC GLMPOWER
- **SamplePower** (SPSS)
- **PS** (Power and Sample Size Calculation), by William D. Dupont and Walton D. Plummer, Jr. `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize`
- **S-PLUS/R**: pfnc, pchisqnc, glhpwr, etc.
- ...

Statistical Analysis - Continuous Data

Two-Sample Comparison of Means

- **Paired t-test**: pre- and post- treatment comparison (usually positively correlated) e.g., noncomparative open-label trials.
- **Two-Sample t-test**: comparative clinical trials. e.g., treatment vs control/placebo.

Multiple Treatment Groups

Analysis of Variance (ANOVA)

- One-Way Classification: simultaneous comparison of k groups.
- Two-Way Classification: to test two categorical factors and their interaction. e.g., multicenter trials or 2x2 factorial designs.

Statistical Analysis - Continuous Data *cont.*

Nonparametrics

When the normality assumption of random errors is not met:

- **Wilcoxon Signed Rank Test**: analogate to paired t-test.
- **Wilcoxon Rank Sum Test**: analogate to two-sample t-test.
- **Kruskal-Wallis Test**: analogate to ANOVA.

Repeated Measures

Generalized Estimating Equations (GEE): longitudinal data (multiple assessments of the endpoint variable are performed at various time points from each patient).
◇ GEE can assess overall average treatment effect across time, time effect, and treatment-by-time interaction effect.

Statistical Analysis - Categorical Data

One Sample

- **Z test (normal approximation)**: large sample size.
- **Exact Binomial confidence interval**: small sample size or the number of events is close to 0 or to n.
- χ^2 **test**: pre- and post- treatment comparison.

Multiple Independent Samples

- **Z test or** χ^2 **test**: large sample size.
- **Fisher's Exact Test**: small sample size or small number of events.

Model-Based Method

Logistic Regression To describe the relationship between the endpoint variable and the covariates. $->$ Odds Ratio (OR).
e.g., parallel group trials with *binary* endpoint.

Statistical Analysis - Censored Data

Kaplan-Meier Estimator

- To compare survival distributions of two samples: **Logrank Test** (nonparametric).

Statistical Analysis - Censored Data - *cont.*

Cox Proportional Hazard Model - Cox 1972

- Widely used in survival analysis to explain the effect of covariates on survival times.
- *Semi-parametric* model: the baseline hazard can take any form (no distribution assumed for the hazard function), but the covariates enter the model linearly.
- Assume **proportional hazards**: the covariate effects multiple hazard. e.g., if taking drug X halves your hazard at time 0, it also halves your hazard at time 1, or time 0.5, or time t for any value of t.
- **Hazard Ratio** is a function of covariates and does not vary with time if all covariates are NOT time-dependent.
- **Likelihood ratio test**, **Wald test**, and **score (logrank test)**.