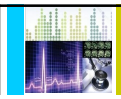



Topics in Clinical Trials

Yossi Levy

Topics to be covered




- What is a clinical trial
- Design considerations and sample size calculations
- Randomization
- Data monitoring – DMC and BDRM
- The ITT principle
- Analysis of incomplete data
- Adaptive clinical trial designs

What is a clinical trial

- A prospective study
- Comparing the effect and value of an intervention
- Against a control
- In human beings

Clinical Development

- Phase I
  - Safety
  - Pharmacology
- Phase II
  - Extended safety
  - Efficacy – proof of concept
- Phase III
  - Confirmation of clinical efficacy
  - Long term safety
- Phase IV
  - Post marketing

Clinical trials "must have"

- Everything should be pre-defined
  - Research questions
  - Study population
  - Study design
  - Assessment methods
  - Statistical analysis
- Randomization
- Blinding

The life cycle of a clinical trial

- Design
- Conduct
- Analysis

A good statistician can make significant contributions to each of these phases!

Most common study types

- Superiority study:

$$H_0 : \mu_T = \mu_C \qquad H_1 : \mu_T > \mu_C$$

- Non-inferiority study:

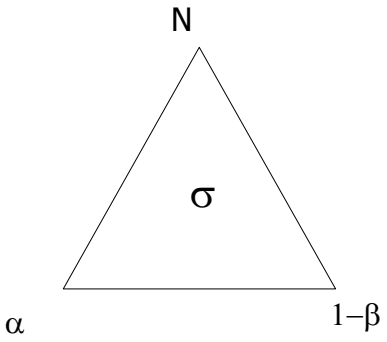
$$H_0 : \mu_C - \mu_T \geq \Delta \qquad H_1 : \mu_C - \mu_T < \Delta$$

- Bio-equivalence study:

$$H_0 : \mu_T \neq \mu_C \qquad H_1 : \mu_T = \mu_C$$

The "Introduction to Statistics story"

Hypothesis: **H_0: Δ=0**
H_1: Δ>0



The real life story

All of these consideration will affect the study's sample size and power

- Selection of clinical endpoint
- Study population - inclusion/exclusion criteria
- Assumptions on clinical conditions and therapeutic effect
- Drop out of subjects during the study
- How the data will be analyzed at the end of the study?

Sample size calculation in practice

- It is theoretically possible to develop formulae for the sample size needed, or to find them in the literature
- However, in many cases, it would be too complicated
- Moreover, it will probably not take into account population heterogeneity, correlation to baseline variable, drop-outs, etc.
- A practical way to evaluate sample size and power is to use simulation

Real life example

- Different patients may have different rate of clinical events – what is the rate distribution?
- You probably want to select “active patients” to the trial – for example, patients who had at least one or two clinical events during the year prior to enrollment. How do you account for that?

Simple example

- Endpoint – number of clinical events during a year
- Endpoint distribution – Poisson
- $H_0: \lambda_T = \lambda_P$ vs. $H_1: \lambda_T < \lambda_P$
- Normal approximation is not going to work
- Non-parametric test has low power
- There is probably a statistical test and a formula for the sample size out there for comparing the means of two independent Poisson Random variables
- Poisson regression will make more sense

Simulation of power/sample size

1. Generate a data set assuming H_1 is true for a given sample size N and given study parameters, and take into account all considerations
2. Analyze the data set using the statistical method you intend to use at the end of the trial. H_0 may be accepted or rejected
3. Repeat steps 1 and 2 a large number of times
4. The estimated power is the proportion of simulated data sets for which the null hypothesis is rejected

Randomization

- *Randomization* is an allocation of subjects to treatment regimens using a random element
- It is an essential component of clinical trials
- It promotes comparability of the treatment groups with respect to known as well as unknown covariates
- It reduces the chance for bias in the evaluation of the treatment effect
- It can also serve as a basis for a randomization approach to inference

Types of bias

- Selection bias
- Observer bias
- Imbalance in a covariate

Complete randomization

- “flip a coin” for every subject
- Totally unpredictable – eliminates selection bias
- Risk of undesirable imbalance in the number of subjects allocated to each arm => potential loss of power

Permuted block randomization

- A sequence of blocks that contain the treatment assignments in desired ratios
- The treatment assignments are randomly permuted within the blocks
- Provides a good balance in treatment assignments, when most of the blocks are filled
- Improves the efficiency of an interim analysis
- What would be the block size?

Variations of permuted block randomization

- Permuted block design with a variable block size
- Constrained block randomization
- Stratified permuted block randomization

Example

- Suppose you want to randomize N patients into 2 arms (Drug / Placebo), in a 1:1 ratio
- There are 3 categorical stratification factors
  - Age: Under or over 40
  - Sex
  - Smoking status
- Goal: to balance the distribution of each of the stratification factors between the treatment group
- The 3 factors define 8 strata
- If N is relatively small, stratified permuted block randomization may fail to balance

Dynamic randomization

- May be applied when there are too many important prognostic factors for stratification to handle
- Used to provide a balance in selected covariates
- The treatment assignment of a subject depends on the subject's vector of covariates and thus is determined only when the subject arrives

Pocock and Simon Biased Coin Algorithm

- Draw in advance a random list of N numbers from U(0,1).
- When a new subject is to be randomized to the trial, calculate the total imbalance caused by assigning him to drug arm, or to placebo arm.
- The preferable arm for the current subject is the one that minimizes the total imbalance.
- If both arms are equally preferable, assign Drug / Placebo with probability 0.5. Otherwise, assign the preferable treatment to the subject with probability 0.5<p<1.

Pocock SJ, Simon R. Biometrics 1975;31(1) 103–115.

Total imbalance

- For each stratification factor:
  1. Look at the level of the new subject. Count the current number of subjects in the same level assigned to drug / placebo.
  2. Calculate the difference between the number of subjects assigned to drug and placebo, if the new subject is assigned to “drug”.
  3. Calculate the difference between the number of subjects assigned to drug and placebo, if the new subject is assigned to “placebo”.
- The total imbalance if the new subject is assigned to drug/placebo is the sum of differences for each stratification factor.

Total imbalance example

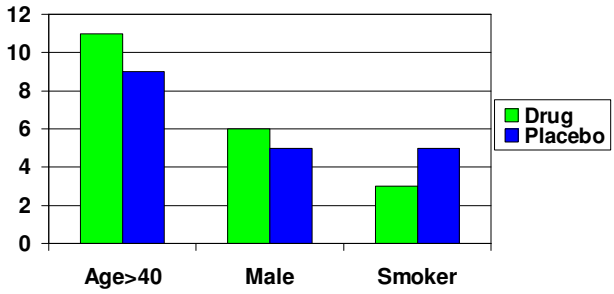
Current Count (N=22):

	Age>40		Sex		Smoker	
	N	Y	M	F	N	Y
Drug	0	11	6	5	8	3
Placebo	2	9	5	6	6	5

New patient: Age=45, Male, Smoker

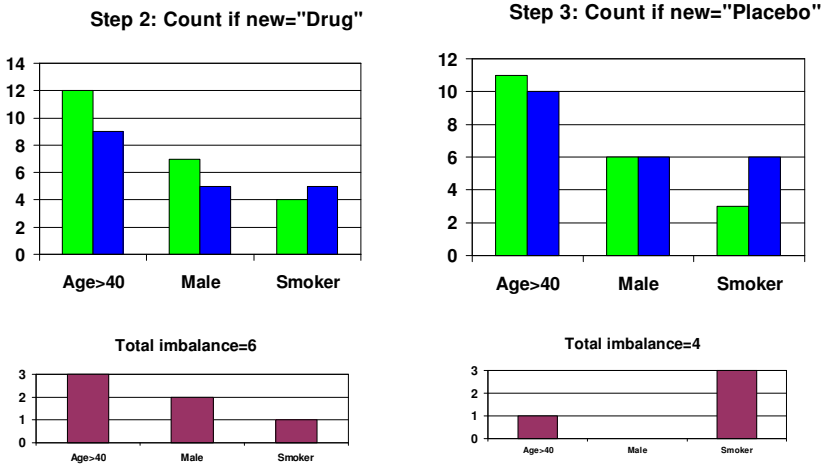
Total imbalance example

Step 1: Current Count



	Age>40	Male	Smoker
Drug	11	6	3
Placebo	9	5	5

Total imbalance example



Action: Randomize patient to “placebo” with probability 0.8

Pocock and Simon Biased Coin Algorithm

- The basic algorithm can be modified to deal with:
  - Different balance between arms (ratio of x:y)
  - More than 2 arms
  - Non-binary stratification factors (such as country)
  - Different weight to each stratification factor

Data monitoring during the trial

- BDRM – Blinded Data Review Meeting
- DMC – Data Monitoring Committee

The ITT principle

- Fisher developed the theory of randomization when experimenting with plants
- However, unlike rows of plants, people sometimes
  - Fail to comply with randomly assigned therapies
  - Fail to comply with study protocol
  - Do not complete the trial
- Any difference between groups that arises after randomization could be due to consequences of the randomized treatment assignment
- Adjusting the analysis of treatment effect by post-randomization group differences could introduce bias

Intention-to-Treat Analysis

Includes all randomized patients in the groups to which they were randomly assigned, regardless of their adherence with the entry criteria, regardless of the treatment they actually received, and regardless of subsequent withdrawal from treatment or deviation from the protocol

(Lloyd) Fisher et al., 1990

ITT key points

- Use every subject who was randomized according to randomized treatment assignment
- Ignore noncompliance, protocol deviations, withdrawal, and anything that happens after randomization
- As randomized, so analyzed
- The ITT analysis holds the randomization as of principal importance
- Deviation from the original randomized groups can contaminate the treatment comparison

Compliance with Treatment

- Even subjects who did not comply with their assigned treatment should be included in ITT analysis
- Statistical reasons
  - Compliance or noncompliance occurs **after randomization**
  - Attempting to account for noncompliance by excluding noncompliant subjects can bias the treatment evaluation
- Clinical reasons
  - In clinical practice, some patients are not fully compliant
  - Compliant subjects usually have better outcomes than noncompliant subjects, regardless of treatment

How to deal with missing data?

- Optimal study design and execution
- Statistical models that handle incomplete follow-up data
- Imputation
- Extreme case analysis
- Sensitivity analysis

Types of Missing Data

- MCAR - missing completely at random
  - Neither observed nor unobserved outcomes are related to (explain) missingness
- MAR - missing at random
  - Observed outcomes and/or baseline covariates are related to missingness, but unobserved outcomes are not
- MNAR - missing not at random
  - Unobserved outcomes are related to missingness
  - Also known as not missing at random (NMAR) and non-ignorable (NI)

Examples in Clinical Trials

From perspective of analysis of an efficacy outcome…

- Missingness caused by dropout due to patient moving to Hawaii is likely MCAR
- Missingness caused by dropout due to lack of efficacy is likely MAR if observe data up to the point of dropout but may be MNAR if don't, i.e. dropout is explained by unobserved but not observed data
- Missingness caused by dropout due to AE depends on relationship of AE to efficacy. If AE is related to PK levels that are also related to efficacy, then missingness may be MAR or MNAR

Implications for Analysis

- Maximum likelihood methods are valid (at least consistent) for MAR missingness
  - Important that all covariates and observed outcomes related to missingness are in the model, else analysis is not valid
  - Suggests repeated measures with ML or REML estimation as a good method for clinical trial analysis
- Methods that are not maximum likelihood require imputation

Implications for Analysis (cont)

- LOCF as an imputation method with ANCOVA is only valid under the assumption that patients outcomes would not have changed further had they remained in the trial (note that this is an stronger statement than assuming MCAR missingness).

What to do when have MNAR?

- Need to make an assumption about the effect of treatment in the unobserved data
  - pattern mixture models, selection models, and pattern set models are examples of well known frameworks to implement this by relating observed to unobserved values
- Since in general we do not know the relationship, the suggested method is sensitivity analysis where the analysis result is examined under a range of assumptions about the effect of treatment in the unobserved data

How to perform sensitivity analysis? One sensible solution:

Concept based on (unpublished? JSM 2006) proposal of Jie Zhang and Liansheng Zhu for binary data:

1. Subset the missing data according to MCAR, MAR, and MNAR. This helps avoid the problem of analysis results depending too much on assumptions and the sensitivity analysis being too broad.
2. Choose a model for imputing the missing data for each case. MAR and MCAR can be based on the observed data. MNAR missing data should be imputed using a distribution where you can vary the parameter(s) that specify the treatment effect.

sensitivity analysis - One sensible solution

3. For each value of the parameters examined, perform multiple imputation. Combine the analysis result from each imputation set as follows:

$$\overline{\theta_D} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d \qquad \overline{V_D} = \frac{1}{D}\sum_{d=1}^{D}\hat{V}_d$$

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \overline{\theta_D}\right)^2 \qquad T_D = \overline{V_D} + \frac{D+1}{D}B_D$$

where T_D is the total variability associated with $\overline{\theta}_D$.

Form a test statistic to evaluate θ using T_D and $\overline{\theta}_D$ and then evaluate using the t distribution with degree of freedom estimated as outlined in Little and Rubin 2002.

sensitivity analysis - One sensible solution

4. Create a grid of the results across values of the parameter(s) that specifies the treatment effect in the MNAR group indicating where the p-value is <0.05.

5. If the p-value is <0.05 across a reasonable range of alternatives, then the primary analysis result can be considered robust

Example: Linear Model with Repeated Measurements

Step 1

- Reasons for discontinuation/missing data need to be classified as MCAR or MAR versus MNAR

MCAR or MAR

- Adverse event
- Meeting pre-defined withdrawal criteria

MNAR

- Lost to follow up
- Disease exacerbation
- Unwillingness to continue

Example: Linear Model with Repeated Measurements (2)

Step 2

- Build models for each visit using only observed data

$$y_j = \beta_0 + \beta_1 y_1 + \ldots + \beta_{j-1} y_{j-1}$$

where y_j is the change from baseline in FEV_1 at the j^{th} visit.

Step 3

- Impute the MCAR and MAR missing data using these models as follows

$$y_{j\,imputed} = \hat{y}_j + z_i \hat{\sigma}_e$$

where z_i is a randomly drawn standard normal deviate.

- Any intermittent missing values will also be imputed using this regression method.

Example: Linear Model with Repeated Measurements (3)

Step 4

- Using only the MCAR, MAR, and completers data including the imputed values, perform repeated measures analysis to estimate β's for the primary analysis model written here for each subject.

Step 5 $\underline{y}_i = \underline{x}_{i1}\beta_{trt} + X_i\underline{\beta}$

- Impute MNAR missing data using the model from Step 4 as follows:

$$\underline{y}_{i\,imputed} = \delta\,\underline{x}_{i1}\hat{\beta}_{trt} + X_i\underline{\hat{\beta}} + \underline{z}_i\hat{\sigma}_\varepsilon$$

where $\delta \in [0,1]$ and represents the amount of treatment benefit preserved in the unobserved MNAR data, 0 being none and 1 being all.

Example: Linear Model with Repeated Measurements (4)

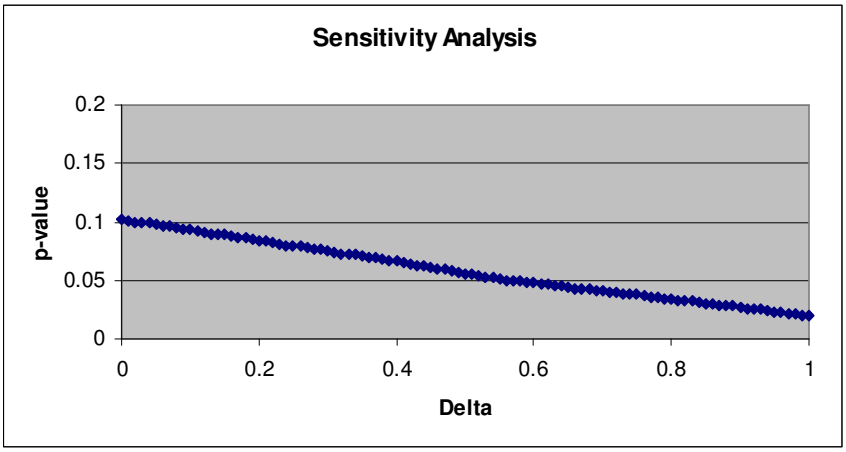
Step 6

- Estimate β_{trt} using all data.

Step 7

- Iterate Steps 3-6 for each δ 100 times. Then calculate a single estimate and p-value for each δ using standard multiple imputation techniques (available in SAS Proc MIANALYZE). A plot of p-value by δ can then be drawn to help visualize the result.

Example: Linear Model with Repeated Measurements (5)



Example: Poisson Regression

Step 1 – same as previous example

- Reasons for discontinuation/missing data need to be classified as MCAR or MAR versus MNAR

MCAR or MAR

- Adverse event
- Meeting pre-defined withdrawal criteria

MNAR

- Lost to follow up
- Disease exacerbation
- Unwillingness to continue

Example: Poisson Regression (2)

Step 2

- Build "overdispersed" Poisson model using only observed data

$$\lambda(\underline{x}_i) = e^{\beta_0 + x_{1i}\beta_1 + \underline{x}_i \underline{\beta}}$$

where β_1 represents the treatment effect and is written separately simply for explanatory reasons.

Example: Poisson Regression (3)

Step 3

- Impute the MCAR and MAR missing data using this model as follows:

1. For each patient who drops out from the study, use their baseline covariates and treatment to estimate their annual event rate, λ_i. Also, calculate τ'=percent of study unobserved. For example, in a one-year study, a patient dropping out after 6 months has a value of 0.5. Calculate their rate for the unobserved portion of the study, $\lambda_{i\ un} = \lambda_i \tau_i$'.
2. We use the negative binomial distribution to simulate a random "overdispersed" Poisson value since we can set the variance to be the mean x overdispersion for the negative binomial distribution and get integer values. In other words, using the negative binomial distribution, we get values that look like overdispersed Poisson. In order to generate a random negative binomial value since you cannot get directly from SAS, you can use the following procedure.

Example: Poisson Regression (4)

3. First generate a random gamma distributed value, *g*, with from a gamma distribution with parameters

$$\alpha = \frac{\lambda_{i\,un}}{\sigma^2 - 1} \qquad \beta = \sigma^2 - 1$$

where σ^2 is the overdispersion parameter from the Poisson regression and for the following parameterization of the gamma distribution:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

4. Then, generate the imputed number of events for the unobserved period by randomly drawing a Poisson with mean *g*. Then add this value to the number of events experienced by the patient during the observed period to get $y_{i\ imputed}$.

Example: Poisson Regression (5)

Step 4

- Using only the MCAR, MAR, and completers data including the imputed values, perform Poisson regression analysis to estimate β's for the primary analysis model.

Step 5 $y_i = e^{\beta_0 + x_{1i}\beta_1 + \underline{x}_i \underline{\beta}} + \varepsilon_i$

- Impute MNAR missing data using the β's from Step 4 and generate a random "overdispersed" Poisson value for the unobserved portion as in Step 3 with the exception that now for the missing portion

$$\hat{\lambda}_{i\,un} = \tau_i' e^{\beta_0 + \delta\, x_{1i}\beta_1 + \underline{x}_i \underline{\beta}}$$

where $\delta \in [0,1]$ and represents the amount of treatment benefit preserved in the unobserved MNAR data, 0 being none and 1 being all. Finally, add the imputed count to the observed count for the new total count as in Step 3.

Example: Poisson Regression (6)

Step 6 – same as previous example

- Estimate β_{trt} using all data.

Step 7 – same as previous example

- Iterate Steps 3-6 for each δ 100 times. Then calculate a single estimate and p-value for each δ using standard multiple imputation techniques (available in SAS Proc MIANALYZE). A plot of p-value by δ can then be drawn to help visualize the result.

Recommendation

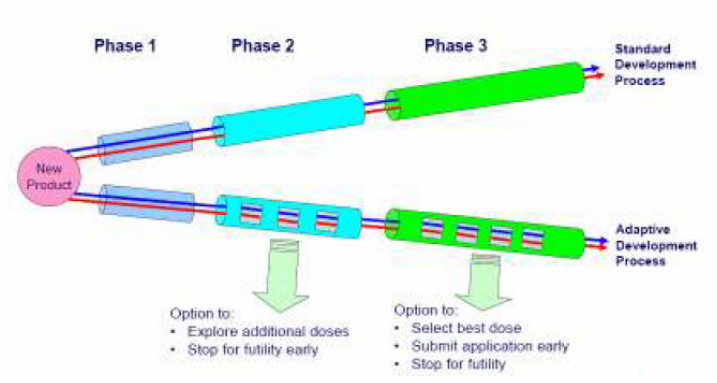
- At a minimum, reports of confirmatory clinical trials should contain discussion of the impact of missing data
- If there is concern about the validity of the primary analysis result in particular, a sensitivity analysis should be performed to evaluate the robustness of the result

Adaptive clinical trial design

Adaptive design is defined as

- A multistage study design
- That uses accumulating data
- To decide how to modify aspects of the study
- Without undermining the validity and integrity of the trial

The adaptive design principle



Types of adaptations

- Early stopping for efficacy or futility
- Blinded sample size re-estimation
- Response adaptive randomization

- Seamless phase II/III design
- Unblinded sample size re-estimation

- Enrich subpopulation
- Change in choice of test statistic
- Change in primary hypothesis/endpoint

Adaptive does not mean "flexible"

Rules for adaptation can be completely specified in the protocol and implemented as such

We are reluctant to support the view that more and more decisions regarding design issues and statistical methodology can be deferred to later phases ... of the trial. This may not be beneficial for trial quality in general.

The high credibility of the results from randomized clinical trials ... is not exclusively a direct consequence of randomization, but stems from the need to carefully pre-plan the scientific investigation.

Armin Koch, Federal Institute for Drugs and Medical Devices, Bonn, Germany
Biometrical Journal, 4 8 (2006) p.11.

Advantages

- Potential modification are approved up front by regulatory agencies
- There is no need to file protocol amendments
- Logistics can be planned in advance
- Credibility of results is maintained, especially with the DMS as a firewall

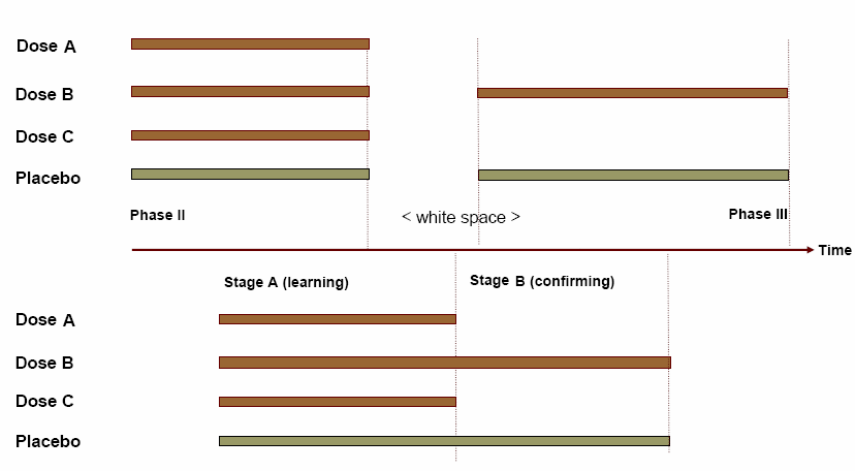
Disadvantages

- Rigid design lack flexibility to respond to unexpected developments during long term trial
- Sponsors must be blinded. Responsibility to implement changes falls to the DMC, who may be “uncomfortable” implementing the rules as written
- institutional review boards (“Helsinki committee”) may find it more difficult to agree to amore complex protocol. Some protocol adaptations may need to be included in patient informed consent form

Seamless adaptive designs

- A seamless design is a clinical trial design which combines into a single trial objectives which are traditionally addressed in separate trials
- A seamless adaptive design is a seamless trial in which the final analysis will use data from patients enrolled before and after the adaptation

Traditional vs. Seamless



Seamless development pros

- Up-front planning of "phase III"
- Real time move from "phase II" to "phase III"
- Saving operational issues of phase III (selection of sites/investigators, IRB submissions and approvals)
- Other operational issues

Seamless development cons

- Logistical problems, marketing tactics
- Maintenance of the blind
- Who will instigate and direct adaptations between phases?
- Complicated protocol may be problematic for acceptance by IRBs, DMC, informed consent forms.
- Interpretability/credibility of the results?

Statistical advantages

- Type I error better controlled
- Higher overall power = smaller overall sample size

Traditional development viewed as "super- experiment"

α_{II} = Phase II type I error $\quad 1-\beta_{II}$ = Phase II power

α_{III} = Phase III type I error $\quad 1-\beta_{III}$ = Phase II power

$\Rightarrow$

Overall development program type I error = $\alpha_{II} \cdot \alpha_{III}$

Overall development program power = $(1-\beta_{II}) \cdot (1-\beta_{III})$

Seamless phase II/III development

α_{II} = Seamless trial type I error

$1-\beta_{II}$ = Seamless trial type II error

$\Rightarrow$

Overall development program type I error = α_{II}

Overall development program power = $1-\beta_{II}$

Statistical difficulties

- How to actually control the type I error level?
- How to combine results from the two stages of the trial?

* The two questions are closely related

Key to answer the two questions is the

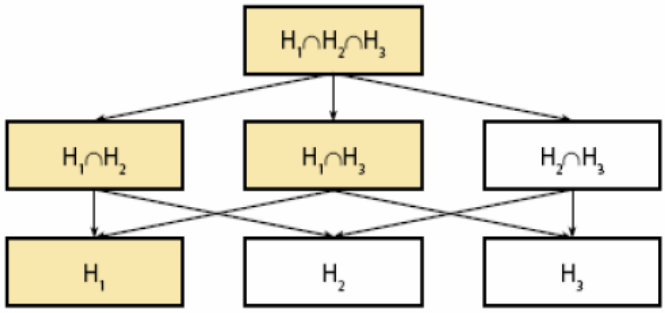
Closed Multiple Testing Principle

Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655—660

The closed testing principle

- Suppose there are k hypotheses $H_1,\ldots, H_k$ to be tested and the overall type I error rate is α.
- The closed testing principle allows the rejection of any one of these elementary hypotheses, say H_i, if all possible intersection hypotheses involving H_i can be rejected by using valid local level α tests.
- It controls the family-wise error rate for all the k hypotheses at level α in the strong sense.

Closed testing example



Decision rule for H_1: All intersection hypotheses containing H_1 must be rejected, e.g., $H_1 \cap H_2 \cap H_3$, $H_1 \cap H_2$, $H_1 \cap H_3$ and H_1

Closed test based methods for p-value combination

- Bauer and Kohne (BK) - Biometrics, 1994
- Thall, Simon and Ellenberg (TSE) -Biometrika, 1988
- Schaid, Wieand and Therneau - Biometrika, 1990
- Stallard and Todd - Statistics in Medicine, 2003

BK method

- BK method combines the adjusted p-values from first ("Phase IIb") and second ("Phase III") stages of the seamless trial.
- The p-values can be combined using either Fisher's method or the Inverse Normal method

P-value combination

- Under H_0 the (adjusted) p-values are uniformly distributed on [0,1]
- Let p_1 and p_2 be the (adjusted) p-values from the first and second stage of the trial
- Note that p_1 and p_2 are independent under H_0
- Fisher's method (R.A. Fisher 1932):

$$-2\log(p_1 \cdot p_2) \sim \chi^2_{(4)}$$

- Inverse normal method (Mosteller & Bush 1954): test statistic is

$$w_1 \cdot \Phi^{-1}(1-p_1) + w_2 \cdot \Phi^{-1}(1-p_2) \quad \text{where} \quad w_1^2 + w_2^2 = 1$$

TSE method

- At step 1 test the global hypothesis of no effect. If rejected, select one treatment and proceed to stage 2
- Weight the treatment effect from both stages by sample size at each stage
- Reject H_0 if weighted effect is sufficiently large
- But… critical value for weighted effect must be adjusted to appropriately control the overall type I error

Example (Jennison &Turnbull, 2006)

- Endpoint is normally distributed with σ=5
- At stage 1: 5 treatment groups (including placebo), arm size=100
- The treatment group with the highest mean (provided it is highest than the placebo group mean) is selected to continue to stage 2.
- At stage 2: 2 treatment groups, arm size=500

The results

DATA: Phase IIB results

	Control	Trt1	Trt2	Trt3	Trt4
n	100	100	100	100	100
P (1-sided)		0.20	0.04	0.05	0.03
Z		0.84	1.75	1.64	1.88

Trt $i^* = 4$ was selected to go to Phase III.

Phase III results

	Control	Trt4
n	500	500
P (1-sided)		0.04
Z		1.75

Application of BK method

- In stage 1, the adjusted p-value, according to Simes method, is p_1=0.075
- In stage 2, no adjustment is needed, so P_2=0.04
- According to Fisher's method: χ^2=11.6, p=0.0204
- According to the inverse normal method: z=2.19, p=0.0144
- The result of the trial is positive

Application of TSE method

- Effect sizes can be back calculated:

$$\text{Stage 1}: \quad \hat{\theta}_1 = \sigma \cdot Z_1 / \sqrt{n_1} = 5 \cdot 1.88 / \sqrt{100} = 0.940$$
$$\text{Stage 2}: \quad \hat{\theta}_2 = \sigma \cdot Z_2 / \sqrt{n_2} = 5 \cdot 1.75 / \sqrt{500} = 0.391$$

- Combined TSE statistic is

$$\hat{\theta}_{\text{weighted}} = \frac{100 \cdot \hat{\theta}_1 + 500 \cdot \hat{\theta}_2}{600} = 0.483 = \frac{\sigma \cdot Z}{\sqrt{n}} = \frac{5 \cdot Z}{\sqrt{600}}$$
$$\Rightarrow \quad Z = 2.365 \quad \Rightarrow \quad \text{p} - \text{value} = 0.0090$$

Design issues

- The power of the methods presented are similar (although TSE is uniformly more powerful, nut not by much)
- Nominal type I errors in this example are 0.0212 for the Fisher method, 0.0206 for the inverse normal method, because of the futility boundary at stage 1
- Therefore we could change the significance test levels of the combined p-values to ~0.03, maintain the overall type I error at 2.5%, and gain more power