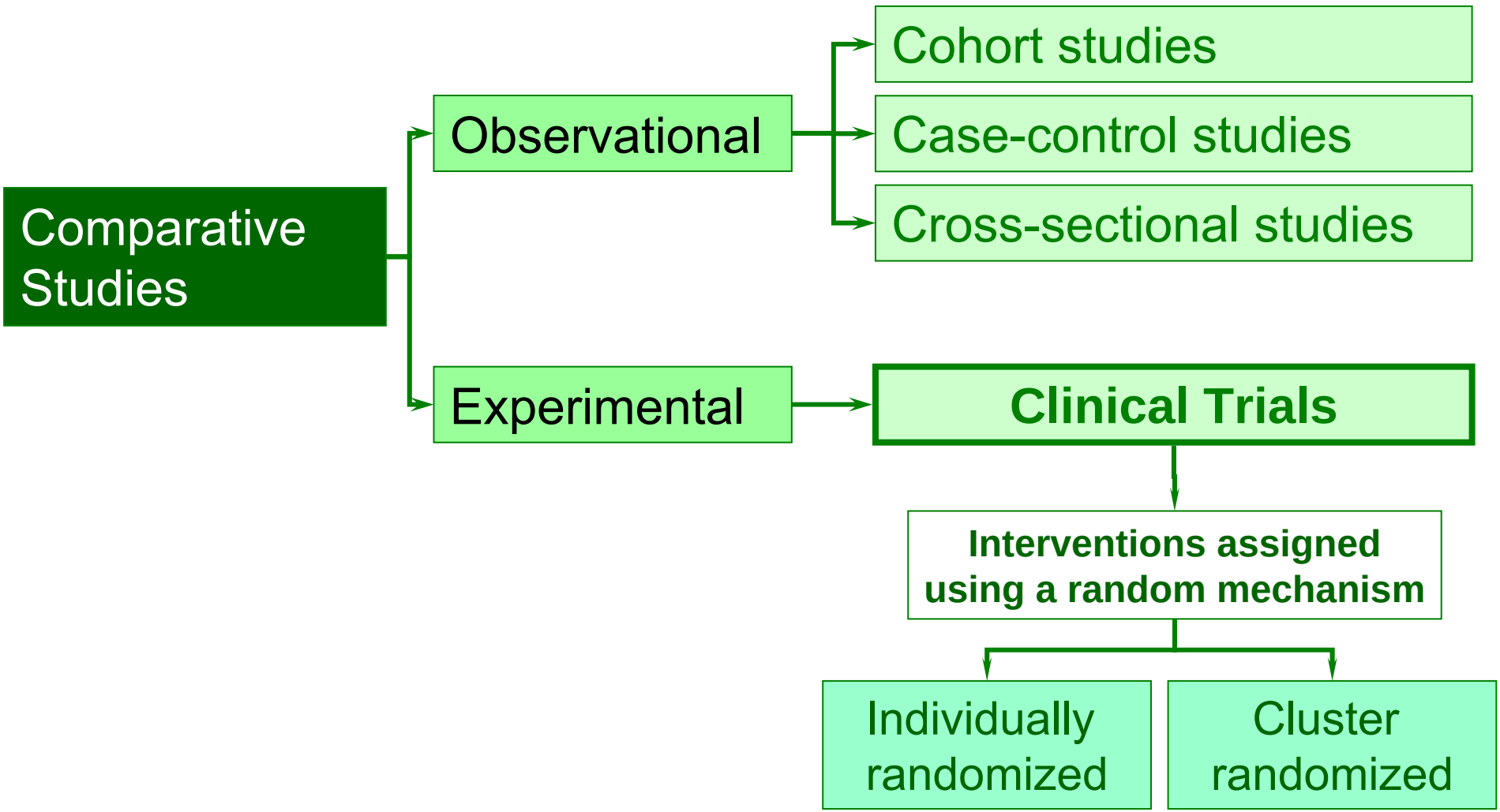
Cluster Randomized Trials and Equivalence Trials

Daniel Wojdyla

UNDP / UNFPA / WHO / World Bank Special Programme of Research, Development and Research Training in Human Reproduction
World Health Organization

2005

Cluster Randomized Trials: Introduction



Cluster Randomized Trials

Cluster randomized trials are experiments in which social units or clusters rather than individuals are randomly allocated to intervention groups

Examples:

Intervention	Social unit / Cluster
Mass education programs	Communities
Medical intervention	Clinics or hospitals
Smoking prevention programs	Schools
Dietary interventions	Families

Reasons for Adopting a Cluster Randomization

➔ Need to minimize or remove contamination

Example: In a trial for the prevention of coronary heart disease, factories were chosen as units of randomization to minimize the likelihood of subjects in different intervention groups sharing information concerning preventive advice on coronary risk factors.

➔ Basic feasibility considerations

Example: Evaluate a programme to enhance the effectiveness of hypertension screening and management in general practice. It was recognized that such a programme would not function effectively if some patients in a practice but no others were entered into it. Unit of randomization: physician practice.

➔ Only natural choice

Example: Intervention programmes that use mass education.
It is difficult to provide general recommendations concerning diet, smoking or exercise to some people and not to others in the same community

Impact on Design and Analysis

- Theory of experimental design assumes that experimental unit which is randomized is also the unit of analysis.

- Inferences are frequently intended to apply at the individual level, while randomization is applied at the cluster level.

- Problem with individual level analysis: lack of independence among members in a cluster (clustering effect).

  Application of standard sample size formulas will lead to underpowered studies. ➡ Larger sample size

  Application of standard statistical methods will tend to bias p-values downward risking a spurious claim of statistical significance. ➡ Sophisticated statistical methods

- Several other issues related to the conduct and interpretation of clinical trials are also affected.

Measuring Clustering Effect

Intraclass correlation (ρ): measure of the degree of similartity among responses within a cluster and may be interpreted as the standard Peason correlation coefficient between any two responses in the same cluster.

For sample size determination, "design effect" is defined as:

$$DE = 1 + (m - 1)\,\rho$$

where ρ is the intraclass correlation and m is the cluster size.

It gives a measure of how much the sample size in each group have to be increased to achieve the same statistical power as would be obtained by individual level randomization.

When ρ = 0, DE = 1 and the responses within clusters are independent.

Reasons for Between Cluster Variation ($\rho \neq 0$)

Subjects frequently select the clusters to which they belong.
Example: in a trial randomizing medical practices, the characteristics of the patients belonging to a practice could be related to age or sex differences among physicians. To the extent that these characteristics are also related to patient response, a clustering effect will be induced within practices.

Influence of characteristics at the cluster level, where all the individuals in a cluster are affected in similar manner (share common environment).
Example: Difference in temperature in nurseries may be related to infection rates.

The effect of personal interactions among cluster members who receive the same intervention.
Example: educational strategies or therapies provided in a group setting could lead to sharing information or predispositions that create a clustering effect.

Unit of Inference

In cluster randomized trials, unit of analysis and unit of randomization can be different, depending on the level of inference:

Inference at individual level: Antenatal Care Trial used clinics as the unit of randomiztion and women as unit of analysis

Inference at cluster level: Second Opinion Trial evaluated the effect of an intervention to lower the rate of caesarean section. The target of the intervention is defined explicitly as the hospital rate of caesarian section.

Common Designs in Cluster Randomized Trials

- Completely randomized: intervention allocated at random to clusters. Suitable when randomizing a fairly large number of clusters.

- Matched pairs: clusters are paired and the two clusters within each pair are allocated at random to the interventions. Advantage: provides very tight and explicit balancing of potentially important prognostic factors at baseline.

- Stratified: clusters are grouped in homogenous strata and then they are allocated at random to interventions.

General Issues in Sample Size Estimation

Issues common to sample size estimation that apply to any randomized trial:

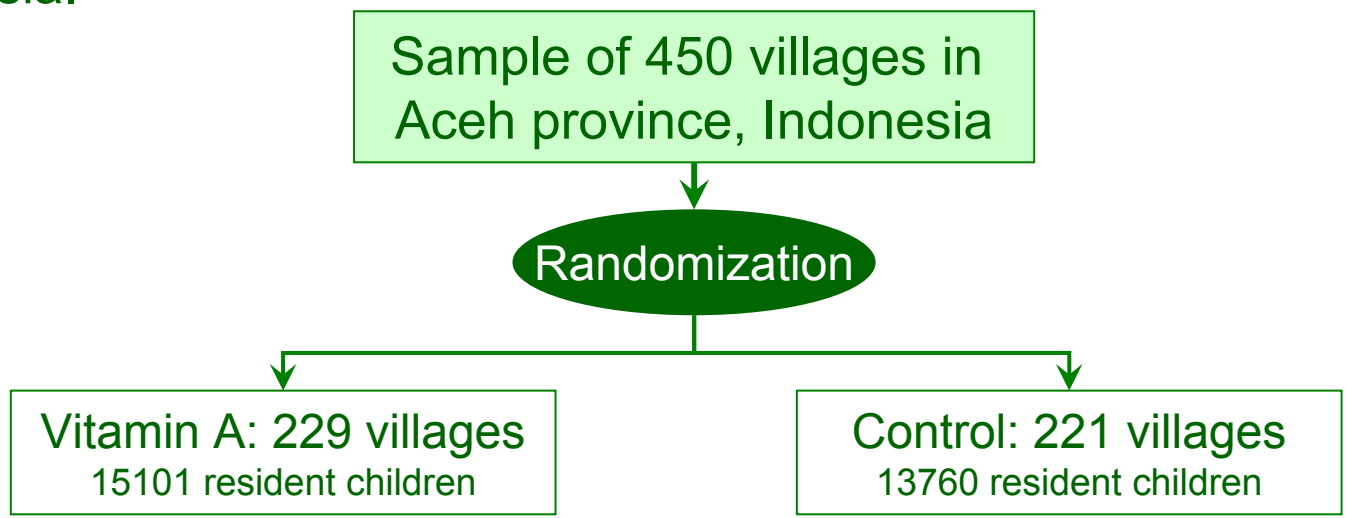
- Identification of the primary study outcome
- Determination of a minimally important effect of the intervention
- Specification of a statistical test or confidence interval method along with its directionality (one-sided / two-sided)

In addition, in cluster randomized trials:

- Determination of cluster size
- Prior assessment of intraclass correlation (ρ)

Example 1: Impact of Vitamin A on Morbidity

Examine the effect of vitamin A supplementation in reducing the frequency of symptoms of respiratory and enteric infections in preschool children in Indonesia.



Each village in the two groups was visited twice, first for the baseline survey and second for the follow-up survey 1 year later.

Capsules of vitamin A were distributed by a trained volunteer to preschool children in the treatment villages at the time of the first visit and 6 months later.

Example 1: Impact of Vitamin A on Morbidity

At baseline and 1 year follow up visits households having children under 60 months of age were identified and asked about history of their having cough and fever and diarrhea in the previous week.

Primary outcome: respiratory infection measured using the cough and fever information and enteric infection measured using the diarrhea information.

Example 2: COMMIT Trial

COMMIT: Community Intervention Trial for Smoking Cessation

Promote smoking cessation using a variety of community resources.

11 selected pairs of communities matched by geographic location, size, and general sociodemographic factors. (10 pairs in USA, 1 in Canada).

Within each pair, one community was randomly assigned to intervention and the other served as comparison.

Intervention: program to promote smoking cessation focused in heavy and light to moderate smokers.

Outcome: 5-year smoking cessation rate

	Heavy Smokers		Light to Moderate Smokers	
	Intervention	Comparison	Intervention	Comparison
Quit Rate	0.180	0.187	0.306	0.275

Example 3: The Antenatal Care Trial

Compare the standard model of antenatal care with a new model that emphasises actions known to be effective in improving maternal or neonatal outcomes and has fewer clinic visits.

Used a stratified cluster randomised design with strata based on countries and clinic characteristics.

Unit of randomization: clinics. (On average, 463 women recruited by clinic)

Clinics per intervention group: 27 new model clinics, 26 standard model clinics (12 clinics randomly assigned in each of three countries, 17 in one country).

Primary outcomes: low birthweight (< 2500 g),, preeclampsia/eclampsia, severe postpatim anaemia (< 90g/L haemoglobin) and treated urinary tract infection.

Example 4: The CATCH Trial

CATCH: Child and Adolescent Trial for Cardiovascular Health

Purpose: To assess the effect of health behaviour interventions, focusing on the elementary school environment.

Design: stratified cluster randomized (strata were four cities in the USA)

Unit of randomization: elementary school

Number of schools per intervention group: 56 intervention and 40 control elementary schools.

Primary outcome: serum cholesterol change after 3 years of follow-up

Equivalence Trials

Equivalence Trials: Introduction

The aim of an equivalence trial is to show the therapeutic equivalence of two treatments, usually a new drug under development and an existing drug for the same disease used as standard active comparator.

Problems:

- Often include too few patients
- Have intrinsic design biases which tend towards the conclusion of no difference
- The application of hypothesis testing in analysing and interpreting data from such trials sometimes lead to inappropiate conclusions
- Inclusion and exclusion of patients from analysis may be poorly managed

Equivalence Trials: Introduction

Randomized placebo controlled double blind clinical trial is the gold standard in clinical research.

Sometimes is not ethical to use a placebo group as comparison group and an active comparator or standard treatment is used instead.

New treatment better than the standard treatment: no special methodological problems.

New treatment expected to match the efficacy of the standard treatment but have advantages in safety, convenience or cost: the objective of the trial is to show equivalent efficacy.

Equivalence Trials: Introduction

Equivalence trials:

- Generally need to be larger than "comparative" trials (or "superiority trials")
- Standard of conduct needs to be especially high
- Handling of withdrawals, losses and protocol deviations needs more care than usual
- Different approaches to analysis and interpretations are appropiate

Statistical Methods and Sample Size

In comparative trials the standard analysis uses:

> Statistical significance tests to determine whether the null hypotesis of "no difference" may be rejected,
>
> Confidence intervals limits to place bounds on the possible size of the difference between treatments.

In equivalence trials the conventional significance test has little relevance: failure to detect a difference does not imply equivalence.

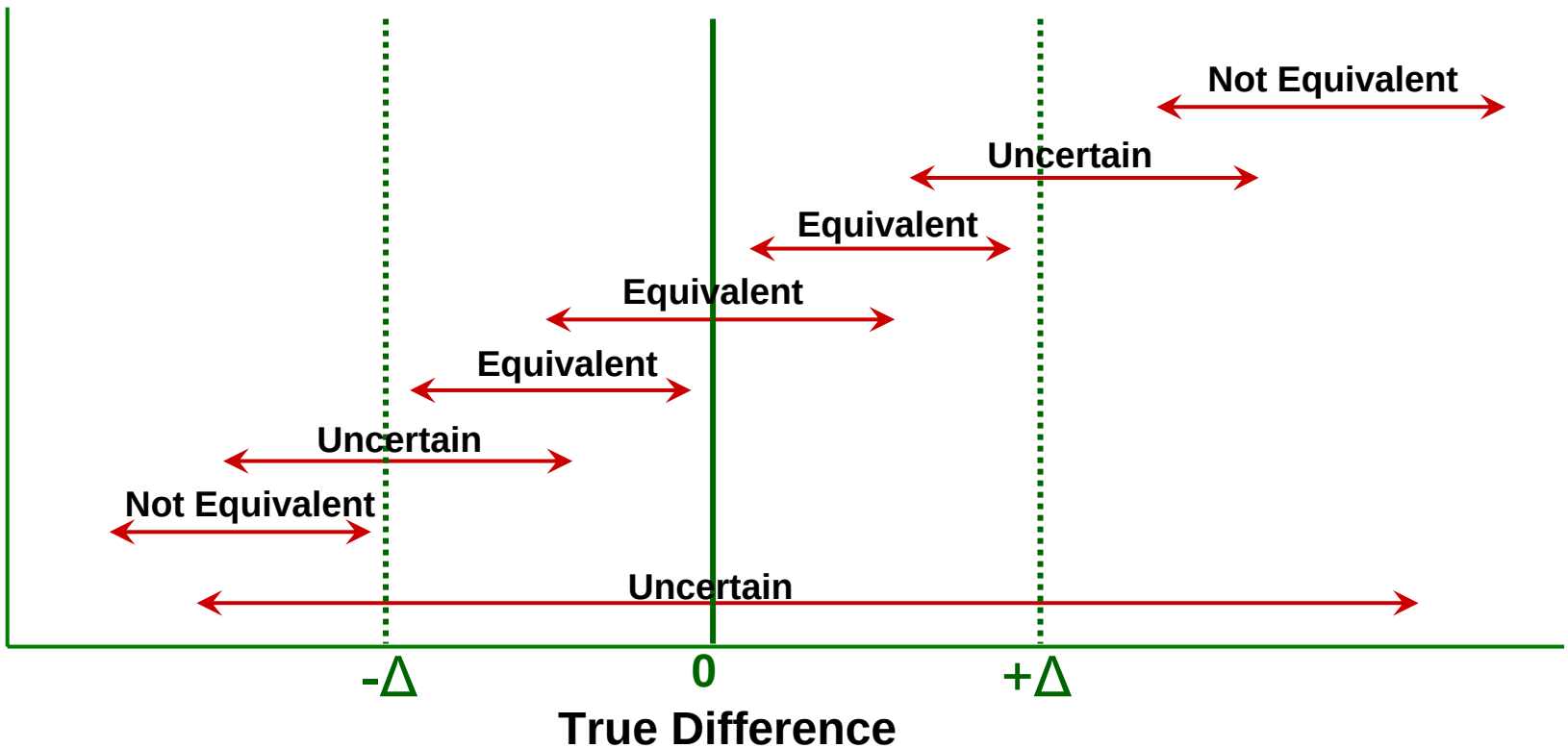
A confidence interval defines a range for the possible true difference between treatments, any point of which is compatible with the observed data.

If every point within this range corresponds to a difference of no clinical importance then the treatments may be considered to be equivalent.

Statistical Methods and Sample Size

Usually a range of equivalence is predefined as the interval from -Δ to +Δ.

If the confidence interval centered on the observed difference lies entirely between -Δ and +Δ equivalence is demonstrated.



Statistical Methods and Sample Size

In comparative trials:

The null hypothesis is that there is no difference between treatments.

The alternative hypotesis is that a difference exists.

In equivalcne trials:

The null hypothesis is that a difference of at least Δ exists.

The trial is targeted at disproving this in favoir of the alternative that no difference exists.

Statistical Methods and Sample Size

To compute the sample size the following is needed:

Range of equivalence (Δ)

Probabilities of type I and type II error (α and β)

The choice of Δ is difficult and requires extensive debate with clinical experts.

The chosen Δ should be generally smaller than in a comparative trial.

Internal Validity of Trials

The second special feature affecting the equivalence trial is the lack of any natural internal check on its validity.

In a comparative trial there is a strong incentive to remove any sloppiness in in design, conduct and analysis because such sloppiness is likely to obscure any differences between treatments.

Therefore, the detection of a treatment difference not only implies that a difference exists but also that the trial was of sufficient quality to detect it.

In equivalence trials the finding of equivalence may arise either from true equivalence or from a trial with poor discriminative power –a trial which was too small, for example.

Internal Validity of Trials

The finding in a trial that two treatments are equivalent does not require that both treatment were effective.

It is equally compatible with the alternative hypothesis that neither was.

In a equivalence trial is important to have means of confirming that both treatments were indeed effective (3rd placebo arm, for example)

The degree of certainty can be increased only by paying careful attention to the design of the equivalence trial, by being strict about matters of conduct and by making additional checks during analysis.

The equivalence trial should mirrors as closely as possible the methods used in earlier placebo controlled trials (comparative trials which found that the active comparator was effective)

Internal Validity of Trials

Important design features:

Inclusion and exclusion criteria: carefully chosen on the basis of previous experience of the active comparator to ensure that the trial contains patients likely to respond to the active comparator.

Dosing schedule of the standard treatment: should reflect the standard manner of use known to be effective on the basis of earlier clinical trials

Use of concomitant medication and other interventions: the use in all patients of a standard dose of concomitant medication with known beneficial effect can result in patients reaching their upper threshold of response and lead to the masking of treatment differences.

Primary response variable and its schedule of measurements

During analysis:

Show similarities between the equivalence trial and the earlier comparative trials in terms of patient compliance, the response during any run period and the scale of patient losses and the reasons for them

Analysis

The most difficult issue relating to the analysis of an equivalence trial concerns which patients and which data from these patients to include.

Most common approaches for the analysis of RCT:

intention to treat analysis

per protocol analysis

Intention to treat: patients are analysied according to their randomized treatment, irrespective of wheter they actually received the treatment.

Patients may fail to take a treatment altogether, may be given the wrong treatment, or may violate the protocol in some other way, but under intention to treat analysis this does not affect matters.

Analysis: Intention to Treat

The strength claimed for such an analysis is that it is pragmatic (it mirrors what will happen when treatment is applied in practice).

In a comparative trial where the aim is to decide if two treatments are different, an intention to treat analysis is generally conservative: the inclusion of protocol violators and withdrawals will usually tend to make the results from the two treatment groups more similar.

For an equivalence trial this effect is no longer conservative: any blurring of the differences between treatment groups will increase the chance of declaring equivalence.

Analysis: Per Protocol

A per protocol analysis compares patients according to the treatment actually received and includes only those patients who satisfied the entry criteria and properly followed the protocol.

This approach might be expected to enhance any difference between treatments rather than diminishing it, because of the removal of uninformative "noise". (Unfortunately, in some circumstances, per protocol analysis might bias the results towards a conclusion of no difference)

In an equivalence trial it is probably best to carry out both types of analysis and hope to show equivalence in either case.

In preparation for this policy it is important to collect complete follow up data on all randomized patients.

Analysis

The result of the analysis of the primary endpoint should be one the following:

that the confidence interval for the difference between two treatments lies entirely within the equivalence range so that equivalence may be concluded with only a small probability of error.

that the confidence interval covers at least some points which lie outside the equivalence range, so that differences of potential clinical importance remain a real possibility and equivalence cannot safely concluded.

that the confidence interval is wholly outside te equivalence range (though this is likely to be rare)

Equivalence Trials: Examples

ASSENT-2: Assessment of the Safety of a New Thrombolytic Study
The Lancet, Vol 354, (1999), 716-722.

Objective: Comparison of efficacy and safety of two fibrinolytic therapies: Tenecteplase vs. Alteplase

Design: double-blind, randomized, controlled trial in 1021 hospitals with 16949 patients with AMI of less than 6 hours duration.
All patients received aspirin and heparin.

Definition of equivalence: the null hypothesis was that 30 day mortality after tenecteplase would exceed 30-day mortality after alteplase by more than 1% or that the relative risk in 30-day mortality with tenecteplase compared with alteplase would exceed 14%, whichever difference proved smallest.

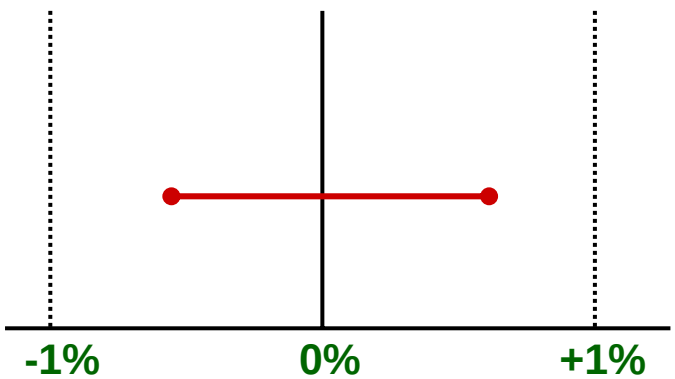
Equivalence Trials: Examples

Primary Outcome: equivalence of all-cause mortality at 30 days

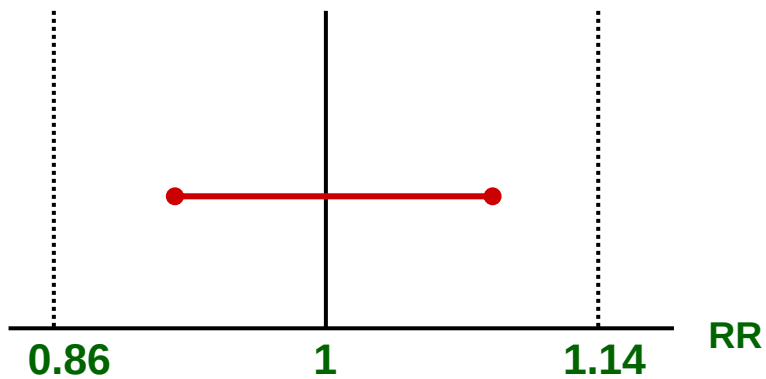
Result:

Tenecteplase (%)	Alteplase (%)	Absoulte Difference (90% CI)	Relative Risk
6.179	6.151	0.028 (-0.554 to 0.609)	1.004 (0.914 to 1.104)

Absolute Difference



Relative Difference



Equivalence Trials: Examples

Prostaglandins for prevention of postpartum haemorrhage
The Lancet, Vol 358, (2001), 689-695.

Oxytocin in the management of third stage of labour is administered by injection, requires refrigeration and protection from light.

Misoprostol, if equally effective, could be an alternative.

Effectiveness was defined in terms of the outcomes

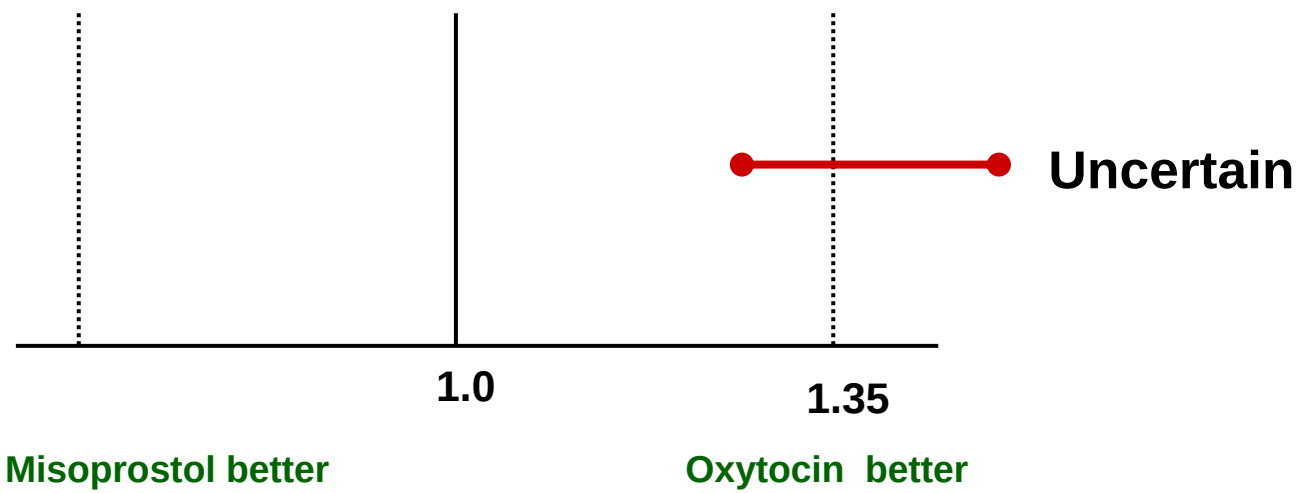
measured postpartum vaginal blood loss of 1000 ml or more need for additional uterotonics

"The sample-size calculation was based on the occurrence of measured blood loss of 1000 mL or more. An increase in relative risk of up to 35% with misoprostol was regarded as acceptable. 20,246 women were needed to provide 90% power for a two-sided, 5% level test to detect a proportional change of 35% or more if the rate of blood loss of 1000 mL or more with oxytocin was 2%"

Equivalence Trials: Examples

Outcome	Misoprostol	Oxytocin	RR (95% CI)
Blood Loss ≥ 1000 mL	366 / 9214 (4 %)	263 / 9228 (3 %)	1.39 (1.19 to 1.63)
Additional Uterotonics	1398 / 9225 (15 %)	1002 / 9228 (11 %)	1.40 (1.29 to 1.51

Clinical Equivalence Range



Sample Size

Test for the Difference of 2 Means

Most epidemiologic studies are comparative.

The problem:
A sample will be extracted. How many subjects should be selected?

How to split the sample in two groups?

Often sample size for each group is equal: 50%..

It will be assumed that the two groups has the same standard deviation and that it is known (σ)

Test for the Difference of 2 Means

Let n_1 and n_2 the sample sizes (to be determined) in the 2 groups.

Define the assignment ratio r as n_1/n_2, and let n be the total sample size.

That is,

$$n = n_1 + n_2 = (r + 1)n_2$$

A test for the null hypothesis of no difference between the means of both groups will be considered against the alternative that the difference is δ.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 - \mu_2 = \delta$$

$\delta < 0$ or $\delta > 0$ for one-side tests and $\delta \neq 0$ for two-side tests.

Test for the Difference of 2 Means

Sample size is determined in such way that a power (or probability) of 1 - β of detecting the true difference when this is δ is achieved.

The sample size can be computed as:

$$n = \frac{(r + 1)^2 (z_\alpha + z_\beta)^2 \sigma^2}{\delta^2 r}$$

for one side tests. For two side tests, replace z_α by $z_{\alpha/2}$. Usually r = 1, and the previous formula simplify to:

$$n = \frac{4 (z_\alpha + z_\beta)^2 \sigma^2}{\delta^2}$$

Power and Minimum Detectable Difference

$$z_\beta = \frac{\delta\sqrt{nr}}{(r+1)\sigma} - z_\alpha$$

$$\delta = \frac{(r+1)(z_\alpha + z_\beta)\sigma}{\sqrt{nr}}$$

The expression for z_β can be converted in an expression for the power (1 - β) using the equation:

$$p\left(Z < z_\beta\right) = 1 - \beta$$

Tests for a Relative Risk

A problem of two samples where propotions are compared:

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 - \pi_2 = \delta$$

for a previously specified δ where π_1 and π_2 are the population proportions.

Alternatively, the ratio between π_1 and π_2 (relative risk) instead of the difference can be tested.

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 / \pi_2 = \lambda$$

Tests for a Relative Risk

Group 2 will be the reference group.

For one-sided tests: $\lambda > 1$ or $\lambda < 1$,

For two sided tests: $\lambda \neq 1$

λ can be computed from δ as: $\lambda = 1 + (\delta / \pi_2)$

Problems of comparison of proportions can arise in cohort studies, cross-sectional studies or clinical trials.

Although case-control studies also compare two proportions, the equations given are not appropiate due to the special sampling design used in those studies.

Tests for a Relative Risk

The sample size is:

$$n = \frac{r+1}{r(\lambda-1)^2\pi^2}\left[z_\alpha\sqrt{(r+1)p_c(1-p_c)} + z_\beta\sqrt{\lambda\pi(1-\lambda\pi)+r\pi\,(1-\pi)}\right]^2$$

where $\pi = \pi_2$ is the proportion in the reference group and p_c is the common proportion in both groups, which can be estimated as:

$$p_c = \frac{\pi(r\lambda+1)}{r+1}$$

When r =1,

$$p_c = \frac{\pi(\lambda+1)}{2} = \frac{\pi_1+\pi_2}{2}$$

For two-sided tests, replace z_α by $z_{\alpha/2}$.

Example

A cohort study investigating smoking habit and coronary heart disease (CHD) is planned in middle-aged men.

A random sample from a population of men will be selected and and questionaire will be completed.

Men will be folllowed recording events such associated with coronary heart disease such as deaths .

After 5 years of follow-up the investigators want to be 90% sure that they will be able to detect a relative risk equal to 1.4 using a one-sided test.

Previous evidence suggest that **non smokers** have a annual mortality rate due to CHD of 413 per 100,000.

Assuming that the same number of smokers and non-smokers are sampled, what should be the sample size?

During the 5 years period, the probability of death is 5 x 413/100.000 = 0.02065. That's the value of π.

The relative risk to be detected is $\lambda = 1.4$

$z_\alpha = 1.6449$ $z_\beta = 1.2816$

Assuming r = 1,

$$p_c = \frac{0.02065 \times 2.4}{2} = 0.02478$$

Applying the equation for n, we obtain

n = 12130.16

Rounding up the next even integer, n = 12,132, 6066 smokers and 6066 non-smokers should be sampled.

Intraclass Correlation Coefficient

ρ may be interpreted as the usual pairwise correlation coefficient between any two members of the same cluster. If the additional assumption that the ICC cannot be negative is added, ρ can be interpreted as the proportion of overall variation in response that can be accounted for by the between cluster variation.

With this interpretation,

$$\rho = \frac{\sigma_A^2}{\sigma^2} = \frac{\sigma_A^2}{(\sigma_A^2 + \sigma_W^2)}$$

where σ_A^2 is the between cluster component of variance and σ_W^2 the within cluster component.

Estimating the Intraclass Correlation Coefficient

Consider a sample of k clusters, each of size m, and denote the mean square error among and within cluster by MSC and MSW respectively. The "analysis of variance" estimator of ρ is given by

$$\rho = \frac{MSC - MSW}{MSC + (m-1)MSW} = \frac{S_A^2}{(S_A^2 + S_W^2)}$$

where S_A^2 = (MSC – MSW)/m and S_W^2 = MSW are sample estimates of σ_A^2 and σ_W^2 respectively.