Design and analysis of clinical trials

Lecture 3

1. Basic Design Considerations
2. Sample Size determination
3. Randomization

2020-01-29

Previous Lectures

- Definition of a clinical trial
- The drug development process

- How different aspects of the effects of a drug are studied in different phases
  - Phase I: Volunteer trials
  - Phase II: Explorative patient trials
  - Phase III: Confirmative patient trials

- Basic statistical concepts

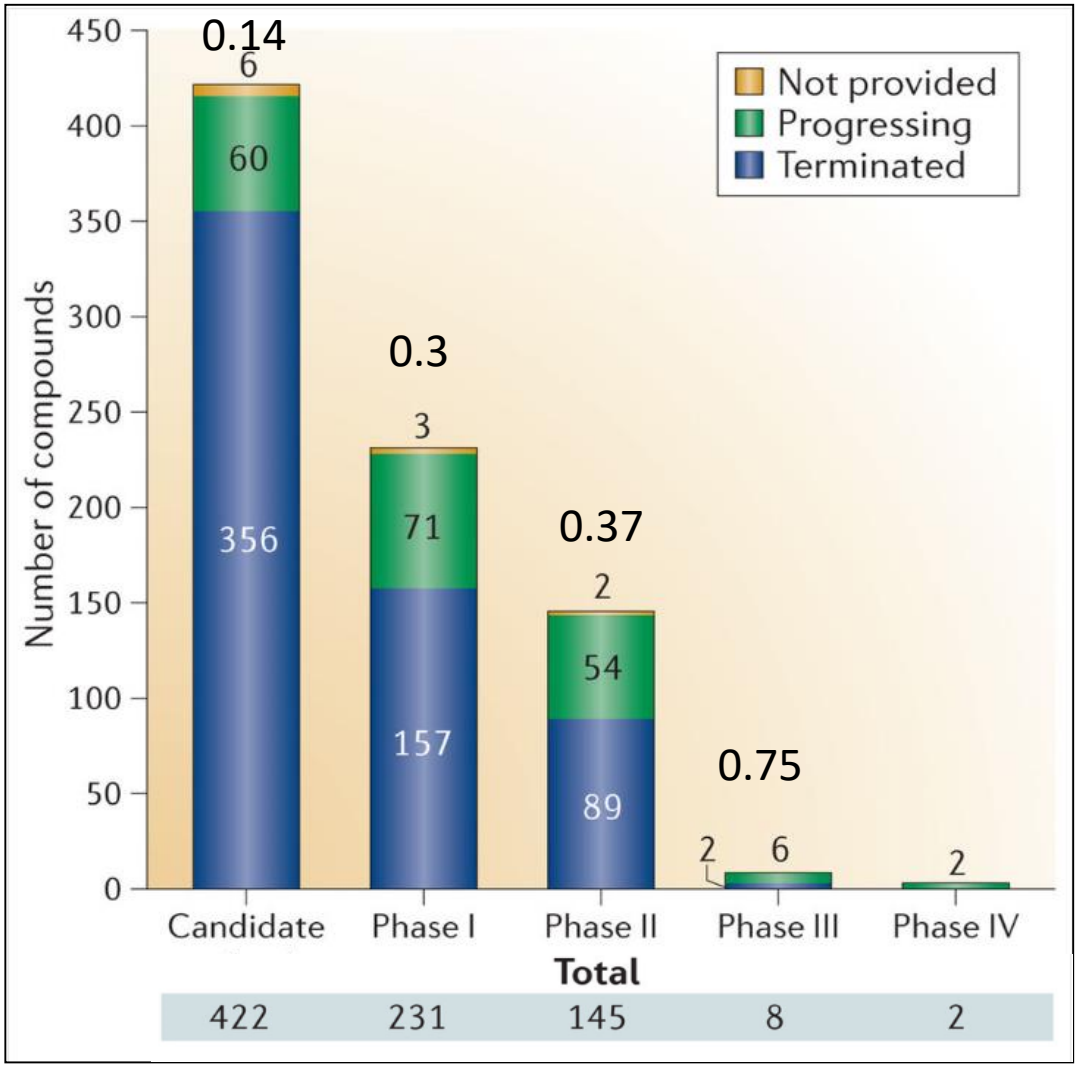
What is a clinical trial?

- A Clinical Trial (CT) is an experiment conducted on human subjects to evaluate some hypotheses related to a new treatment.
- CTs are risky to (i) the patients despite being highly regulated and to (ii) the sponsors (Pharma).
- A CT is usually part of a Clinical development Plan.

NATURE REVIEWS DRUG DISCOVERY | ANALYSIS

An analysis of the attrition of drug candidates from four major pharmaceutical companies

Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace & Alex Weir



- CT are risky to the Pharma Industry
- Clinical trials are difficult
- But: Probability of success increases with phase

Difficulties

•CT are difficult even under ideal circumstances. This due to:

- Bias
- Variability (heterogeneity)
- Not formulating the “right” scientific question
- Logistical complexity,
- Interdisciplinary complexity
- Uncertainty about Recruitment
- Patient dropout
- External changes mid-trial
- ….

Recent trend in clinical trials

- ”Innovative designs”
  - Adaptive
  - Bayesian
  - Sequential testing
  - Umbrella designs
  - Basket designs
  - …
- Modelling and simulation
- Real world evidence
- Personalised medicine/ diagnostics
- Biomarkers and surrogate endpoint
- Quantitative decision making (GNG)

Descriptive and inferential statistics

Inferential statistics forms a basis for a conclusion regarding a prespecified objective addressing the underlying population.

Confirmatory analysis:

Hypothesis → Results → Conclusion

Hypothesis testing is basic to the scientific method and statistical theory gives us a way of conducting tests of scientific hypotheses. Scientific philosophy today rests on the idea of falsification: For a theory to be a valid scientific theory it must be possible, at least in principle, to make observations that would prove the theory false. For example, here is a simple theory:

All swans are white

This is a valid scientific theory because there is a way to falsify it: I can observe one black swan and the theory would fall.

The most common approach to CT (Inferential)

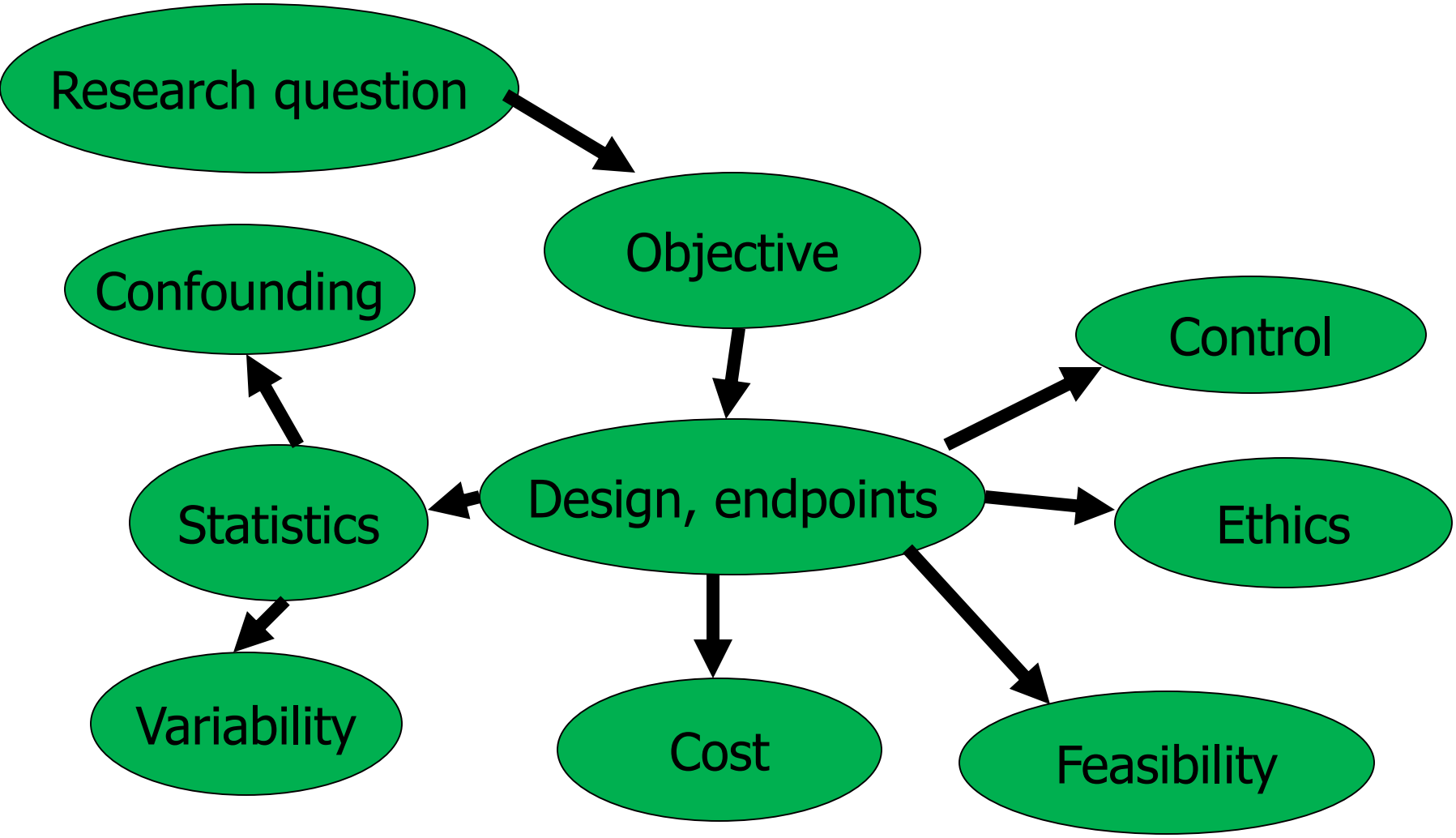
- We have some Theory (Hypothesis) we want to test
- **T** = The new treatment does not work
- **D** = The data from a clinical trial where we apply the treatment to a number of subjects
- We investigate the agreement between **T** and **D**

- **Frequentist: How likely is D given T?**
- **Bayesian: How likely is T given D?**

- If no agreement, we reject **T** (The treatment works)

Basic design considerations

Design issues



Trial design

Trial design should

- Avoid bias
- Generalize to the target population of interest
- Be efficient - avoid using more subjects than necessary
- Studies which are inadequately powered, or otherwise deficiently designed, are inefficient and ethically dubious

The Clinical Study Protocol

Specifies the research plan for a clinical investigation

Regulatory document/ethics approval

The most important document to ensure quality control of a clinical trial

The Clinical Study Protocol

- Important elements:
- Study objectives
- Study procedures
- Target patient population / Eligibility criteria
- Treatments / Blinding and randomisation
- Study design
- Data collection / Data management
- Statistical methods

What is the question?

A clinical study is conducted to adress a medical question regarding a drug substance in treatment of a specific patient population with a specific disease

In the clinical study protocol these questions are formulated in the objectives

A Research Question?

- **Is chocolate healthier than candy? How can we find out?**

Healthy Teeth: Chocolate Candy vs. Chewy Candy

According to Dr. Ana Paula Ferraz-Dougherty, DMD, ADA-certified dentist, and dentistry practitioner in San Antonio, TX, "chocolate is one of the better candies because it washes off your teeth easier than other types of candy." Additionally, research into dark chocolate (which has less sugar than milk chocolate), suggests that the cocoa flavonoids it contains may help with blood pressure and heart disease. So if you or your kids are really itching for a sweet treat, try to opt for dark chocolate. The higher the cocoa count, the better!

Objectives: Example 1

Primary objective:

- To evaluate the effect of gastrozole 20 mg versus placebo for the prevention of gastric ulcers in IBS patients

Secondary objectives:

- To evaluate the effect of gastrozole 20 mg versus placebo for the prevention of lesions in the oesophagus
- To evaluate symptoms of heartburn with gastrozole 20 mg and with placebo
- To evaluate the effect of gastrozole 20 mg on the Quality of Life of the patients.

Exploratory objectives:

- To evaluate the effect of gastrozole 20 mg on some Biomarkers

How choose primary variable?

The objective

To evaluate symptoms of heartburn with gastrozole 20 mg vs. placebo

can be evaluated using any of the measures:

- time to sustained absence of heartburn, defined as the time to the first of 7 consecutive days free of that symptom
- maximal intensity of heartburn during the first and fourth week of treatment
- proportion of subjects with absence of heartburn the 7 days preceding 4 weeks of treatment
- number of days free from heartburn during the entire 4-week treatment period
- …

Study objectives

- When possible, objectives determined primarily by norms for the given disease area and the target product profile
- Primary and secondary objectives should map to corresponding statistical hypotheses
- Safety objectives are given greater emphasis in Phases I and II; Phase III focuses on efficacy and safety
- Objectives should be as precisely as possible. At a minimum, include information on
  - What measure of efficacy/safety will be used?
  - Key features of the target patient population
  - Dosing regimen, i.e. amount, frequency, and route of dosing

Specify clear study objectives

- Instead of:
- To demonstrate the efficacy of rhIGF-I in improving glycemic control.
- Write the more precise statement:
- To investigate the effect of twice daily injection of 40µg/kg of rhIGF-I for 12 weeks on glycemic control, in subjects with moderate to severe Type II diabetes, as measured by the average change from baseline in HbA1c, compared to subjects in the placebo group.

Endpoints

- Ideally, one should use a well-established primary efficacy endpoint, accepted as a suitable measure of patient benefit. When such an endpoint exists, it cannot be ignored. (COPD FEV1 And Exacerbations)
- Often there may be consensus on the choice of primary efficacy variable, but secondary aspects, such as definition of “relapse” or “loss of control” may still be under debate
- It is not recommended to launch Phase I without a reasonably clear vision of what the primary efficacy variable will be in pivotal studies – postponing difficult discussions won’t necessarily make them any easier
- Agreement on conventions for endpoint specific handling of dropouts/missing data is important

Endpoints continued

- Generally speaking, endpoints which can be measured in a completely objective fashion are preferred
- This may not always be possible – some degree of subjectivity may be unavoidable (e.g. in endpoints such as physician's or patient's evaluation of improvement)
- In evaluating quality of life, use of a "validated" instrument is preferable. In many cases, a disease-specific QOL questionnaire exists
- Consultation with the Health Economics group is highly recommended, to ensure that collection of QOL data supports the target product profile (don't wait until Phase III to do this)

Endpoints: Multiple Endpoints

- Multiple primary endpoints (‘co-primary’) are sometimes used.

- In that case, there is an associated penalty, in terms of a higher bar to declare statistical significance at a given level α.

- A common simple approach is to require significance at level α/k, where k is the number of endpoints (Bonferroni).

- Bonferroni is in general inefficient; true attained significance will be $< \alpha$

Endpoints: a statistical taxonomy

- Continuous - e.g. reduction in cholesterol, HbA1c, visual acuity
- Categorical
  - Multiple categories with no natural ordering
  - Ordered categorical - e.g. different degrees of improvement
- Binary – e.g. response/non-response, dead/alive at a specific time post-treatment
- Time-to-event – e.g. survival, time to progression

- Different analysis methods are appropriate for each type of endpoint
- Sample size requirements differ as well

Endpoints

- Usually the focus is on efficacy endpoints but there can be many others:
  - **Pharmacokinetic** endpoints are generally standard parameters derived from the observed concentration-time profiles
  - **Safety** endpoints also tend to be fairly standard;
    - Incidence of adverse events
    - Changes in key laboratory parameters
  - **Pharmacodynamic** endpoints, in contrast, are measures of activity, and will vary from study to study.

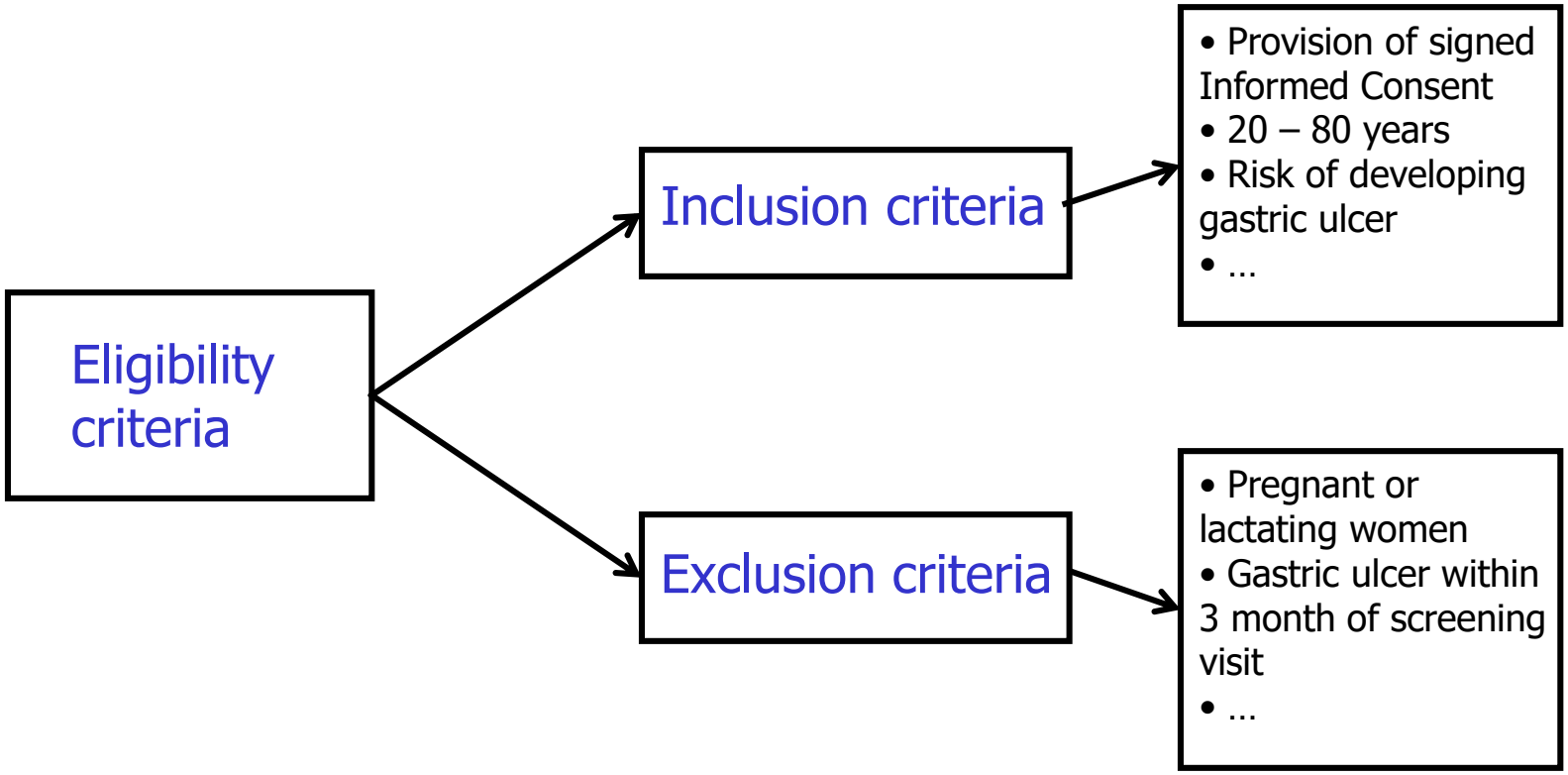
Endpoints in cancer trials

- Response rate (where response is based on change in tumor size)
- Duration of response (note that the resolution with which this can be determined will depend on the frequency of scheduled evaluations)
- Survival time: the holy Grail!
- Progression-free survival (PFS): is a treatment effect on response, in terms of reduction of tumor size, or PFS is predictive for treatment effect on survival. Unfortunately, this seems to vary by tumor and treatment class

Target population

- Subjects included in a trial should be a representative sample of the target population
- The target population should be should have the specific medical need
- A homogeneous population reduce bias and minimize variability
- Important to be able to generalize the results
- To be included in a trial a patient must be eligible

Target population



To be eligible a patient must meet *all* inclusion criteria

Patients meeting *any* of the exclusion criteria are excluded from the trial

Study design

Examples of common designs:

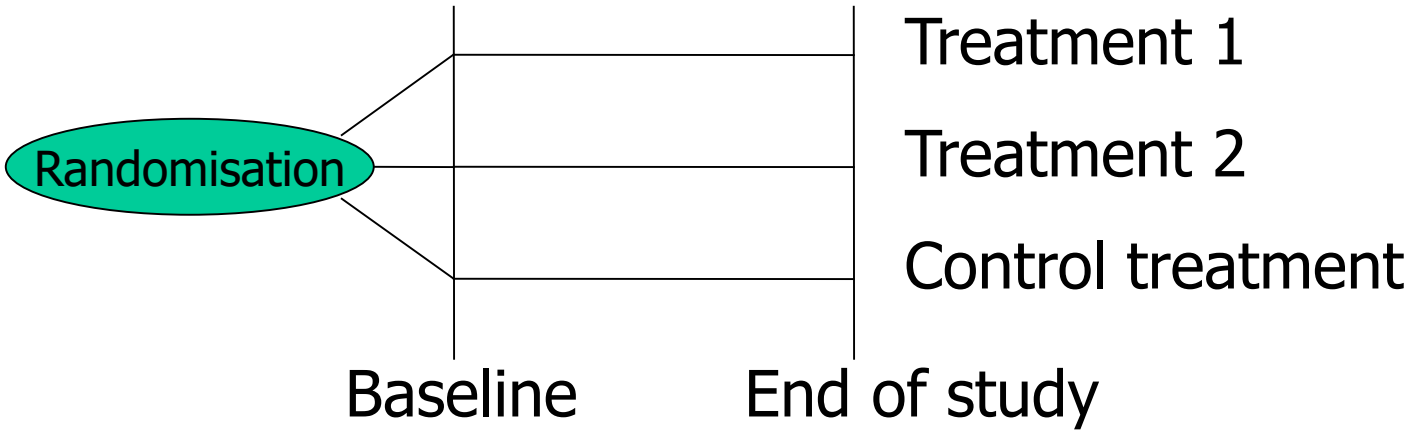
- Parallell group designs
- Crossover designs
- Group sequential designs
- Titration designs

Choice depends on:

- Objective(s) of the study
- Theraputic area
- Time and cost
- Regulatory requirements
- …

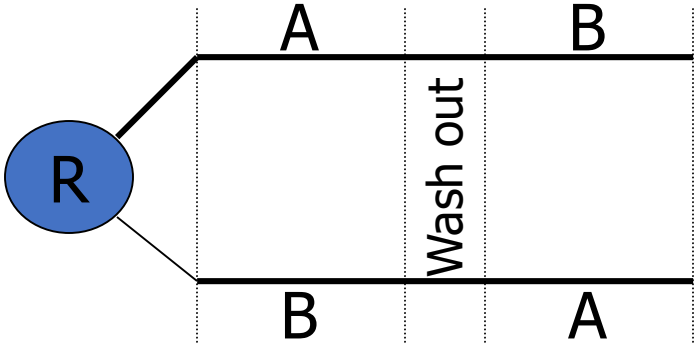
Study design, Example 1

- This study is a 26-week, multicenter, randomised, double blind, 2-arm, parallel group, placebo-controlled, efficacy and safety study, in patients with increased risk of developing gastric ulcers.



Crossover studies

- All subject get more that one treatment
- Comparisons within subject



Sequence 1

Sequence 2

A	B
B	A

- Within subject comparison
- Reduced sample size
- Good for chronic conditions
- Good for pharmaceutical studies

Statistical model, Example 2

Primary variable: Time to occurrence of gastric ulcer

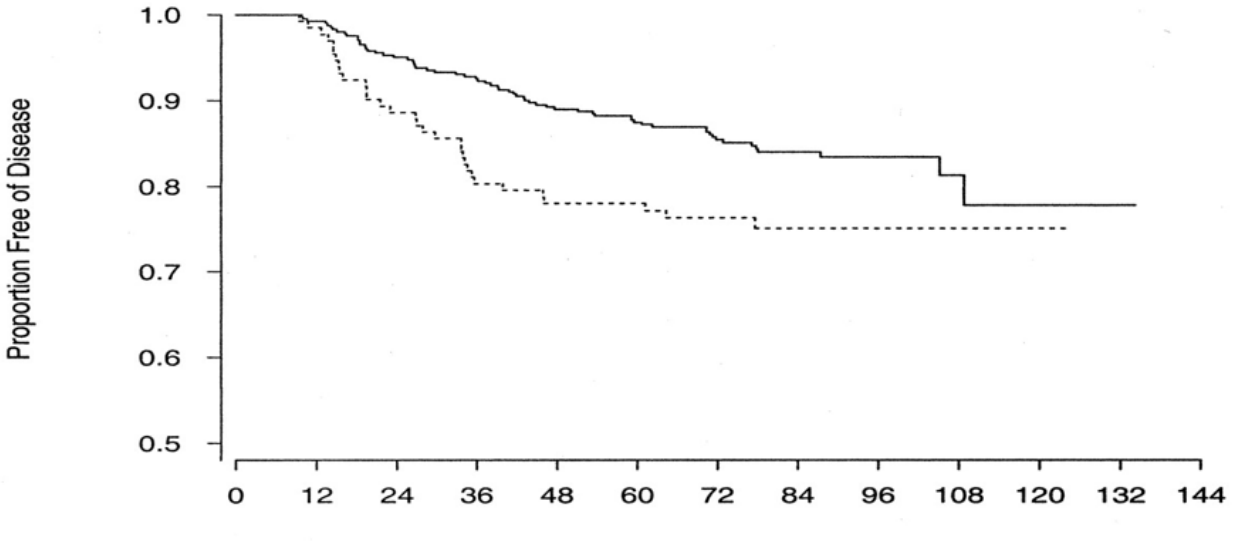
- To assess the difference between the active treatment group (T) and the placebo group (P) the null hypothesis will test if the two groups have the same survival function:

$$H_0: S_T(t) = S_P(t)$$

- The log rank test will be used

Kaplan-Meier time to occurrence of gastric ulcer

- Primary: Time-to event variable, to assess differences between the esomeprazole groups and the placebo group, the log rank test will be used.
- Kaplan-Meier life-table estimation will be used to graphically illustrate the primary variable, time to occurrence of gastric ulcer, for each treatment group.



Statistical model, Example 2

$j = 1,\ldots,J$: Distinct times of observed events (occurrence of gastric ulcer) in either group

N_{Tj} = Number of patients at risk in group T at time j
N_{Pj} = Number of patients at risk in group P at time j
O_{Tj} = Observed number of events in group T at time j
O_{Pj} = Observed number of events in group P at time j

O_{Tj} is hypergeometric distributed under H_0

Expectation: $E_j = O_j \cdot N_{Tj} \,/\, N_j$
Variance: $V_j = \{\, O_j \,(N_{Tj} \,/\, N_j)(1 - N_{Tj} \,/\, N_j)(N_j - O_j) \,\} \,/\, (N_j - 1)$

Statistical model, Example 2

The log rank statistic

$Z = \Sigma_j (O_{Tj} - E_j) \; / \; (\Sigma_j V_j)^{½}$

Is approximately standard normal under H_0

A one-sided test at significance level α will reject H_0 if $Z > z_a$

z_a : upper α quantile of the standard normal distribution

Objectives: Example 3

Testing a drug that lowers the blood pressure on patients with hypertension

Primary objectives:

- To compare sitting blood pressure (BP) lowering effect of hypersartan 16 mg, 8 mg and 4 mg

Secondary objectives:

- To compare the proportions of responders on hypersartan 16 mg, 8 mg and 4 mg wrt sitting BP

Tertiary objectives:

- To compare standing BP lowering effect of hypersartan 16 mg with that of hypersartan 8 mg and 4 m

[a] Responders have a decrease in sitting DBP ≥ 10 mmHg from baseline to end of study
[b] Patients with controlled sitting DBP have sitting DBP < 90 mmHg at end of study

Objectives: Example 3

Primary objectives:

- To compare sitting diastolic blood pressure (DBP) lowering effect of hypersartan 16 mg with that of hypersartan 8 mg
- To compare sitting systolic blood pressure (SBP) lowering effect of hypersartan 16 mg with that of hypersartan 8 mg
- To compare sitting DBP lowering effect of hypersartan 8 mg with that of hypersartan 4 mg
- To compare sitting SBP lowering effect of hypersartan 8 mg with that of hypersartan 4 mg

Objectives: Example 3 cntd.

Secondary objectives:

- To compare the proportion of responders[a] on hypersartan 16 mg and on hypersartan 8 mg
- To compare the proportion of responders[a] on hypersartan 8 mg and on hypersartan 4 mg
- To compare the proportion of patients with controlled DBP[b] on hypersartan 16 mg and on hypersartan 8 mg
- To compare the proportion of patients with controlled DBP[b] on hypersartan 8 mg and on hypersartan 4 mg

[a] Responders have a decrease in sitting DBP ≥ 10 mmHg from baseline to end of study
[b] Patients with controlled sitting DBP have sitting DBP < 90 mmHg at end of study

Objectives: Example 3 cntd.

Tertiary objectives:

- To compare standing DBP lowering effect of hypersartan 16 mg with that of hypersartan 8 mg
- To compare standing SBP lowering effect of hypersartan 16 mg with that of hypersartan 8 mg
- To compare standing DBP lowering effect of hypersartan 8 mg with that of hypersartan 4 mg
- To compare standing SBP lowering effect of hypersartan 8 mg with that of hypersartan 4 mg

Variables, Example 3

Primary objective:

To compare sitting diastolic blood pressure (DBP) lowering effect of hypersartan 16 mg with that of hypersartan 8 mg

Primary variable:

Change in sitting DBP from baseline to the end of study

Study design, Example 3

This study is a 8-week, multicenter, randomised, double blind, 3-arm, parallel group, efficacy and safety study, in patients with moderate to severe hypertension.

Statistical model, Example 3

The change from baseline to end of study in sitting DBP (sitting SBP) will be described with an ANCOVA model, with treatment as a factor and baseline blood pressure as a covariate:

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim NID(0,\sigma^2)$

blood pressure change from baseline to end of study

overall mean

treatment effect
i = 1,2,3
{16 mg, 8 mg, 4 mg}

baseline blood pressure

Statistical model, Example 3

Differences between two treatments will be tested using the residual variance from the ANCOVA table, p-values and confidence intervals will be calculated from Student's t-distribution.

- H_{01}: $\tau_1 = \tau_2$ (DBP)
- H_{02}: $\tau_1 = \tau_2$ (SBP)
- H_{03}: $\tau_2 = \tau_3$ (DBP)
- H_{04}: $\tau_2 = \tau_3$ (SBP)

Statistical model, Example 3

- The family-wise type I error for the 4 primary objectives will be controlled at the 5% level using a step-wise testing procedure.

- All 4 comparisons will be conducted at a significance level of 5%, but a comparison will only be confirmed as statistically significant if it is significant at a 5% level and all preceding comparisons were statistically significant at a 5% level.

Results, Example 2

Objective	Treatment difference	Variable	LS Mean	Confidence interval (95%)	p-value	Statistically significant
Primary objective 1	**Hyp. 16 mg – Hyp. 8 mg**	**Sitting DBP**	**-3.7 mmHg**	**[-4.6, -2.8]**	**<0.001**	**Yes**
Primary objective 2	**Hyp. 16 mg – Hyp. 8 mg**	**Sitting SBP**	**-7.6 mmHg**	**[-9.2, -6.1]**	**<0.001**	**Yes**
Primary objective 3	**Hyp. 8 mg – Hyp. 4 mg**	**Sitting DBP**	**-0.9 mmHg**	**[-1.8, 0.0]**	**0.055**	**No**
Primary objective 4	**Hyp. 8 mg – Hyp. 4 mg**	**Sitting SBP**	**-2.1 mmHg**	**[-3.6, -0.6]**	**0.005**	**No**

Are the results statistically significant?

Are the results clinically relevant?

Chapter 3 Reading instructions

3.1 Introduction: Read

3.2 Goals of clinical trials: Read

3.3 Target Population and Patient Selection: Read through

3.4 Selection of controls: Read through

3.5 Statistical considerations: Read

3.6 Other issues: Read through

3.7 Discussion: Read through

Sample size determination and randomization

Sample Size Determination

Why do we need to compute the sample size for a clinical study?

–We must have enough patients to draw conclusions that are "certain enough"

–We do not want *too* many patients (time, cost, ethics)

Sample Size Determination

To perform a sample size calculation we need to:

–Choose primary variable with distribution (variability)
–State the expected effect
–Decide on statistical model
–Set the type I error
–Decide power for a certain effect size

Sample Size Determination

Type I error:

– *Prob* (H_0 rejected given H_0 true)

– The risk to state that we do have an effect even though we have not

Power:

– *Prob* (H_0 rejected given H_1 true)

– The probability to find an effect given that there actually is one

Sample Size Determination

We cannot influence the real effect size!

For which effect size do we want to power the study?

–The smallest clinically relevant effect
–The smallest commercially viable effect
–The effect seen for a competitior substance
–The effect seen in previous studies

Sample Size Determination

Example of protocol text for sample size determination:

"With 260 evaluable subjects
the power is 80% to detect a 7 mmHg
change in DBP from baseline to week 8
at the significance level 5%,
assuming a standard deviation of 20 mmHg."

Is this the best we can do?

Standard deviation assumption is based on historical data, competitor results or regulatory requirements
'Detect' means that we get a p-value below 5% given the true effect is 7 mmHg.

Sample Size Determination

$$n \geq \frac{2(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\Delta^2}$$

nQuery Advisor - [MTT0-1]

File Edit View Options Assistants Randomize Plot Window Help

Two group t-test of equal means (equal n's)

	1	2	3
Test significance level, α	0,050		
1 or 2 sided test?	2		
Group 1 mean, μ_1			
Group 2 mean, μ_2			
Difference in means, $\mu_1 - \mu_2$	7,000		
Common standard deviation, σ	20,000		
Effect size, $\delta = \lvert\mu_1 - \mu_2\rvert / \sigma$	0,350		
Power (%)	80		
n per group	130		

Randomisation

- Example: When performing a survey in a large group of teenagers it is found that teenagers who play computer games more that 3 hours/day have a lower verbal ability as compared to those who computer games less than 3 hours/day.

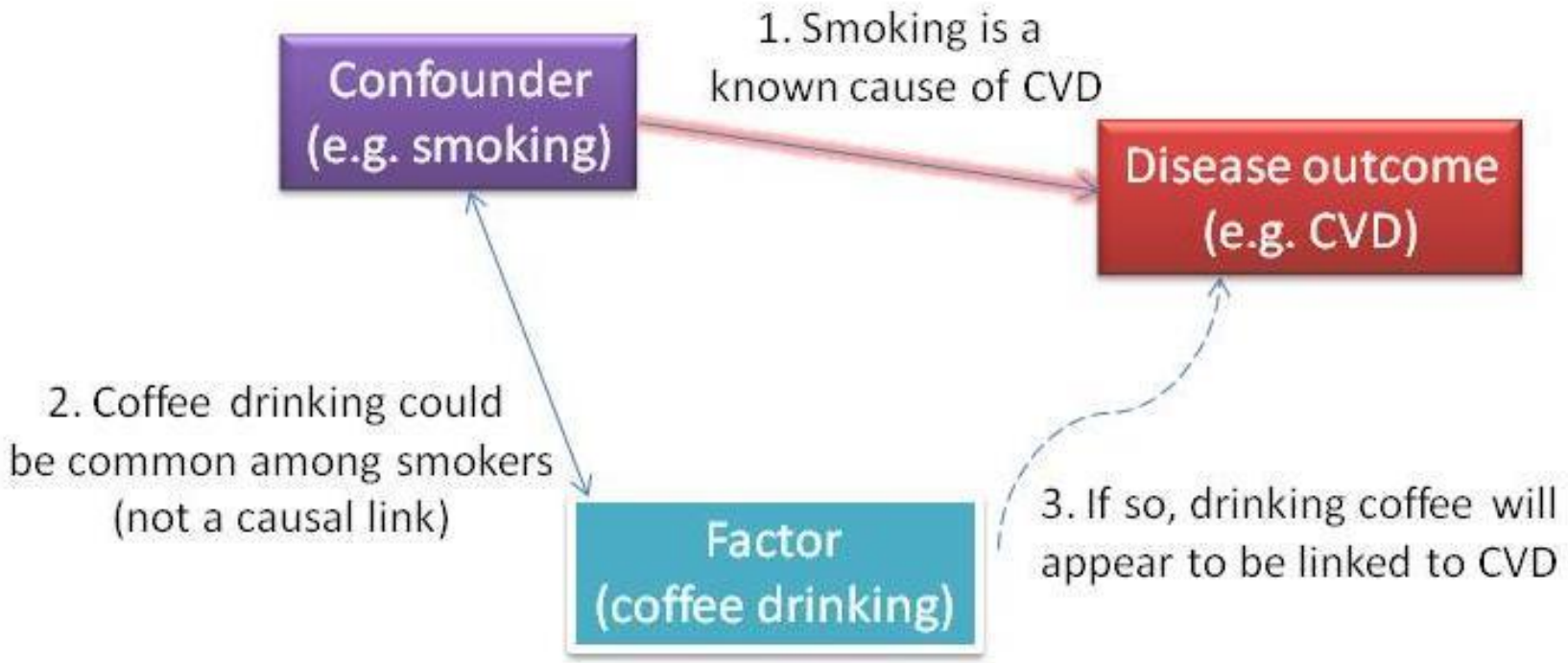
- Can we draw the conclusion that excess of playing computer games causes low verbal ability?

- Why not?

Example, cntd

- We might draw the conclusion that playing computer games causes low verbal ability.
- We might also draw the conclusion that teenagers with a low verbal ability tend to playing computer games.
- But perhaps lack of childhood reading experience causes both playing computer games and low verbal ability.
- How could we correctly assess a possible causal relationship between playing computer games and having low verbal ability?

Confounding



Randomization and Causality

We could randomize subjects to treatments

The aim of randomisation is to ensure that ONE and only one factor is different between the different groups

The consequences of this specific factor can be observed

We can attribute a causal relationship between the factor and the effect

Can we always use randomization?

Observational vs. Randomized

Observational studies

- Can only show association
- We will never know all possible confounders

Randomized studies

- Can show association and causality
- Appropropriate randomisation should eliminate effects of unknown confounders

Randomization

- Randomization is the basis for statistical inference
- Without randomization a “statistically significant difference” may be the result of non random differences in the distribution of unknown prognostic factors
- Randomization does not ensure that groups are medically equivalent, but it distributes randomly the unknown biasing factors
- Randomization plays an important role for the generalization of observed clinical trial data

Randomization – Practical Tips

- If prognostic factors are known use randomization methods that can account for it
  - Stratification / blocking
  - Adaptive randomization
- If possible randomize patients within a site
- Patients enrolled early may differ from patients enrolled later
- Protocol amendments that affect inclusion/exclusion criteria may be tricky
- Even in open label studies randomization codes should be locked

Randomisation methods

The goal is to obtain a representative sample of the target population,

with homogeneous groups,

treatment being the one and only factor differing between the groups.

Randomisation methods

Assume a trial with N patients comparing a test drug (T) and placebo (P) with equally sized treatment groups

Assign either test drug or placebo with 50% probability independently for each patient

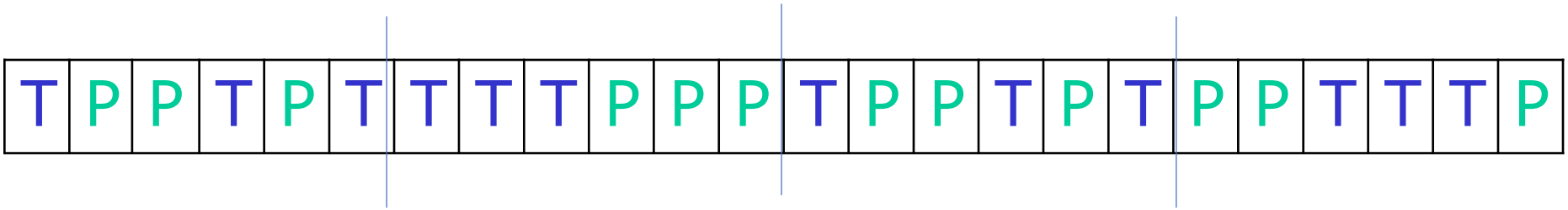
T	P	T	P	T	T	P	P	P	P	T	P	T	T	P	P	T	T	T	T	T	T	T	T

Complete randomisation

Randomisation methods

Assume in the same trial that randomisation codes are generated within blocks of size n=6

Each block is a random permutation of the two treatments in equal proportions



Permuted-block randomisation

Stratification

- Covariates with possible impact on the statistical inference:
  - Age
  - Gender
  - Race
  - Geographical location
  - Disease severity

Stratification

Assume a trial with permuted-block randomisation with equally sized groups treated with test drug and placebo.

Note the gender of each of the randomised patients.

T	P	P	T	P	T	T	T	T	P	P	P	T	P	P	T	P	T	P	P	T	T	T	P

M	F	F	M	F	M	M	M	M	F	F	F	M	F	F	M	F	M	F	F	M	M	M	F

Confounding!

Stratification

- A method to achieve balance between groups for a prognostic factor/covariate
- Each subgroup is randomised separately
- Stratification may be extended to two or more factors
- Rarely feasible to go beyond two factors

Adaptive randomisation

Assume that randomisation codes for each patient are generated based on information on previously randomised patients

Adaptive randomisation

- Treatment adaptive (Biased coin)
- Response adaptive (Play-the-winner)

Blinding

- Randomization does not guarantee that there will be no bias by subjective judgment in evaluating and reporting the treatment effect
- Such bias can be minimized by blocking the identity of treatment (blinding)
- Even in open label studies randomization codes should be locked

Blinding

Parties that can be blinded

- The patient
- The investigator
- The sponsor

Types of blinding

- Open label
- Single blinding
- Double blinding
- Triple blinding

- The investigator: Also includes other personel at study site, such as study nurses etc.
- The sponsor: Includes monitors, statisticians, programmers, etc.

Blinding

Why should we blind

–The patient

–The investigator

–The sponsor

?

Is it always possible to blind

–The patient

–The investigator

–The sponsor

?

Chapter 4 Reading instructions

- 4.1 Introduction: Read
- 4.2 Randomisation Models: Read
- 4.3 Randomisation Methods: Read
- 4.4 Implementation of Randomisation: Less important
- 4.5 Generalization of Controlled Randomised Trials: Less important
- 4.6 Blinding: Read
- 4.7 Discussion: Less important