Experimental studies

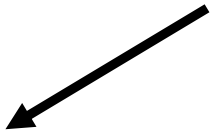
Oscar Bortolami
Senior Biostatistician

Giuseppe Verlato
Full Professor of Medical Statistics

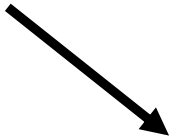
EXPERIMENT	PLANNED OBSERVATION
Researchers actively modify the course of events	Researchers just observe the course of events, without attempting to modify it
Only **positive perturbations** can be applied: 1) Preventive interventions, such as adding fluorine to tap water, or iodine to salt 2) Therapeutic measures (early thrombolysis in myocardial infarction, segmental vs total mastectomy) 3) Rehabilitation interventions	Also **etiologic factors** with **deleterious** health effects can be studied: 1) wrong lifestyle (smoking, excessive alcohol intake) 2) environmental situation (Chernobyl)
RANDOMIZATION	**SELF-SELECTION**
Participants are randomly assigned to different treatments ↓ Other risk factors (potential confounders) are balanced among groups	Potential confounders are not eliminated. For instance, it could be hypothesized that: Unknown genes → Craving for smoking Unknown genes → Increased risk of lung cancer

Observational studies

- If a diagnostic procedure (for example TC-scan), which does not modify the course of disease, is added to normal clinical practice



According to epidemiologists



According to the Italian law

(G.U. **31/03/2008** – AIFA - Determinazione 20 marzo 2008)

This is an observational study

The study is NO LONGER observational

Table of contents

1 Experimental design for a lay person (*uomo della strada*)

- We will start our lesson with a very rough and informal concept on what an experimental study means in the field of medicine:
- Briefly we would like to modify clinical practice for evaluating whether our modification "works".
- Let's go into deeper detail into this …..

1.1 … to modify clinical practice …

- Let's go into deeper detail on the first part of the sentence:

Why?	- You would improve wellbeing of patients (quantity or quality of life) - You have good guesses (e.g. from medical literature) your modification could work
How?	- This will be the topic of these two lessons. Briefly we should think about "smart" ways of doing it
May I do this modification straight away?	- No you can't. There are some ethics and legislation rules you have to fulfil before you are allowed to implement your modification.

1.2 … for evaluating…

Why?	- You should put in place some "indicators" that help you tracking the evaluation. Otherwise based on what will you make a decision?
How?	You should have an indicator. This indicator is collected for each subject (if necessary multiple times) either during or after the modification. If it is collected before the modification then it is illogical it can evaluate the modification (it can evaluate something else but surely not the modification!)
May I choose whatever I like for evaluation?	No. Choosing the right indicator requires expertise. The indicator should be "suitable" for the "intended use"

1.3 … whether our modification "works"

Why?	- When your experiment is finished you would like to know whether it is worth to proceed or not with the modification. Keep in mind that if you are "successful" it does not mean you can implement your modification: it could just be a "green light" for investigating more.
How?	Here comes the statistics. You can start with yes/no question like: "Do I have enough evidence for rejecting the null hypothesis?" "Does the 95% Confidence Interval for the treatment effect lies completely above 0 (or 1) ?"
Why statistics and not something else?	We are ignorant on many biological processes (including medicine). Hence we don't know exactly what is going to happen and what we observe following our modification (deterministic process: direct causation). We could have a good guess on what happened and we observed (statistical process: probabilities, risks, association).

2 Protocol in relation to experimental design: ICH guidelines

- As was telling before, for running an experimental design you have to fulfil some laws and ethics.
- Among them you have to produce in advance some documents.
- There are many documents to be produced showing how the experiment will look like, how the project will be explained to the patient and others.

- Keep in mind much of the methodology and the legislation has been mainly developed on pharmacological interventions. However many of the topics that will be covered also apply for non pharmacological interventions.

- If the experimental study involve a pharmacological intervention then the ICH guidelines are used. **ICH** stands for "**International Conference on Harmonisation** of Technical Requirements for Registration of Pharmaceuticals for Human Use". Some of ICH guidelines have been implemented into European (and Italian as well) legislation.

2.1 Good Clinical Practice

- The document E6 of ICH is called Good Clinical Practice (GCP). GCP are defined as : "A standard for the design, conduct, performance, monitoring, auditing, recording, analyses, and reporting of clinical trials that provides assurance that the data and reported results are credible and accurate, and that the rights, integrity, and confidentiality of trial subjects are protected".

- Briefly this for saying when we would like to run an experiment on drugs we have a specific legislation in place (which of course we must comply)

- Among the document to be produced according to the GCP there is the "study protocol".

- The study protocol is defined by GCP as "A document that describes the objective(s), design, methodology, statistical considerations, and organization of a trial. The protocol usually also gives the background and rationale for the trial, but these could be provided in other protocol referenced documents..."

2.2 SPIRIT Guidelines

- So before starting we need to have in mind how we are going to perform the experimental design as we need to write our plan on a study protocol!

- The GCP provides a template on what a study protocol should cover. This is dated 1995 however revision 2 released on 2017. For helping Investigators on writing study protocol on drugs on 2013 the SPIRIT guidelines have been released. The latter has not the status of law however SPIRIT implemented and expanded some recommendations from GCP.

- **SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials**) is an international initiative that aims to improve the quality of clinical trial protocols by defining an evidence-based set of items to address in a protocol.

- We will use some items from the SPIRIT guidelines to explain methodology on experimental studies.

- Again keep in mind examples will be mainly on drugs but can be extended to non pharmacological interventions.

3 General features: some items from SPIRIT guidelines

SPIRIT contains 33 items. Not all items will be explained. We will change a little bit the order for teaching purpose.

Item 6b: Explanation for choice of comparators (plus extension on items from 11a to 11d): why a comparator is needed and role of control group

Item 20b: Methods for any additional analyses (eg, subgroup and adjusted analyses): role of confounders/prognostic factors

Item 16a: sequence generation (randomisation)

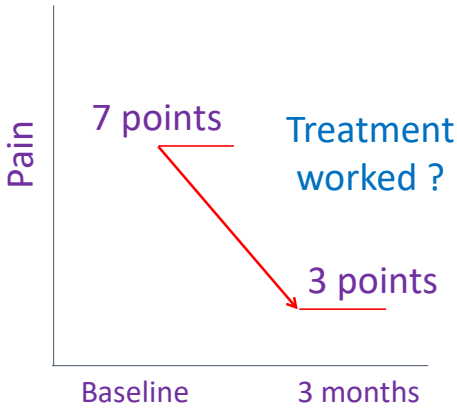
Items 16c, 17a and 17b: Implementation and blinding

Item 10: Eligibility criteria (in relation to internal/external validity of a trial)

Item 8: Description of trial design including type of trial (eg, parallel group, crossover, factorial, single group), allocation ratio, and framework (eg, superiority, equivalence, noninferiority, exploratory)

Item 21b: Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial

3.1 Choice of comparator and use of a control group (SPIRIT item 6b)



A simple experimental study:

i) A subject with a particular condition experiences pain. Hence I decide to administer a treatment hoping to reduce it
ii) I measure pain before and after administering a treatment
iii) For measuring pain I ask the subject to compile a scale ranging from 0 (no pain) to 10 (worst pain ever)
iv) The subject experienced a score of 7 at baseline and 3 after administering the treatment
v) Hence the pain decreased by 4 points

3.1.1 Did the treatment work?

i) Yes!

ii)No!

iii) Maybe …

3.1.2 Are we sure is treatment effect?

- The answer is … maybe!

- We observe a decrease in pain but the decrease could have been due:
- i) The treatment only
- Ii) The treatment and something else
- Iii) Something else only

- And it could be that the "treatment" was bad however "something else" was so good that overall we observe a decrease in pain

3.1.3 Evaluate the treatment effect

- Ideally for knowing the treatment effect on a single subject we should be able to separate the effect of the "treatment" from the effect of "something else"

- This idea is straightforward but how can we do it? How can I know exactly what "something else" contains?

- Some examples of "something else":

i) Background medication
ii) my genotype
iii) my current and past pathologies
iv) my exposure history (e.g.: food, smoke, sport, pollution, allergens)
vi) unknown biological factors on year 2017 (and still to be discovered)
vii) other unknown factors

- It's very hard if not impossible to exactly know the effect of "something else"! (There are special situations but we will not cover)

3.1.4 The counterfactual outcomes

- Given I am not able to determine the effect of "something else" I am not able to determine what was the treatment effect. Am I doomed?

- Well, I could change the point of view of the problem. Briefly I would see what happen in EXACTLY the same conditions if the treatment was not given. So:

i) Results that I see giving my treatment: Treatment effect + Something else effect
ii) Results that I see not giving my treatment: Something else effect

- Hence If I run exactly the same study in the same conditions and in the same time with and without the treatment I can know the treatment effect.

- For doing this I shall compute the difference in scores between the two studies: the "something else" effect cancels out and I have my treatment effect!!
- The counterfactual outcomes briefly is what would have happened with or without the modification of interest

N.B. Controfattuale = in logica proposizione ipotetica in cui l'antecedente è falso e quindi la conseguenza è ipoteticamente vera anche se non si è realizzata.

3.1.5 Problem solved?

- As was telling before we need to repeat the same experiment in the same conditions and the same time. It could be extremely hard to repeat the condition on same condition and impossible at the same time.

- This because if I'm giving a treatment to a subject I couldn't study the same subject on the same time on the condition "no treatment": I would need a time machine, go back in time, repeat the study without treatment and see what happened!

- So for a single subject it could be hard to know the treatment effect. (there are conditions in which we can study different treatments on same subject but we'll discuss later)

3.1.6 The control group

- We get the idea on how to get the treatment effect from a single subject. However we don't have a time machine. We should find other solutions.
- One extension to it could be to use a group of people instead of a single subject: some subjects having the condition studied (experimental) some other not having the condition (control).
- The idea is exactly the same: observe the results under experimental and control condition: having a "contrast" (e.g. difference) on the two conditions will give the treatment effect
- Given not a single subject is studied we'll have to use statistics for evaluating the treatment effect at group level. Hence we'll should be careful on what could happen.

3.2.1 Treatment effect - A simple experiment

Experimental	Control
Y_{11}	Y_{01}
Y_{12}	Y_{02}
...	...
Y_{1n}	Y_{0n}
$\overline{Y_1}$	$\overline{Y_0}$

Treatment effect: $\overline{Y1} - \overline{Y0}$

3.2.2 Treatment effect - How to form groups?

Treatment effect: $\overline{Y1} - \overline{Y0}$

- For the previous to be true the effect of "something else" should cancel out. Remember this is our ultimate goal.
- However how to deal with "something else" when you have many subjects could be more tricky.
- For example imagine you would study the effect on a drug on reducing mortality for a particular condition. The experimental group included 30 men aged 30-35 years while the control group included 30 men aged 80- 85 years. You observe a lower proportion of dead people on experimental.
- Do you believe the treatment was effective on reducing mortality?

3.3.1 Confounding (spirit item 20b)

What we wanted	Reduction in mortality due to treatment effect only	$\overline{Y1} - \overline{Y0}$
What we got	Reduction in mortality due to treatment effect and an unbalanced "something else" (a group is older; older people are more likely to die irrespective of what I'm doing)	$\overline{Y1} - \overline{Y0}$ + residual "something else"

- When we use groups instead of single subjects we should be careful on how we form groups.
- Remember our aim is cancelling out the "something else"
- If there are factors predictive (in this case age) of our response (mortality) we should be careful as can complicate our interpretation of treatment effect. If the predictive factors are "unbalanced" between groups they are called confounders.

3.3.2 How we deal with confounding?

- For cancelling out the confounders one idea is similarity among the groups on confounding factors (e.g. groups of similar age, similar proportion of males, similar distribution on comorbidities on the two groups).
- So if a confounding factors is balanced among groups its effect is cancelled out and it can't confound anymore: the factors are still predictive but are no more confounders.
- Remember the aim is estimate the treatment effect not the effect of confounding. The effect of confounding is something to "deal with" for having the "correct" treatment effect

So one idea could be:
i) let's have a list of confounders;
ii) balance groups based on this list (so that confounders are no longer a problem as they cancel out)
iii) Run the experiment.
iv) Estimate treatment effect.

End of the story?

3.3.3 "Unknown factors" acting as confounders

- Let's review for a moment the definition of "something else"
- This could contain known or observable factors (e.g. the gender or the age) as well as unknown factors (one day maybe we'll know or we'll never know)
- We have seen confounders are "problematic" if they are unbalanced.
- As a solution we would create "balanced group to "cancel out" the effect
- Let assume an unknown factor is a confounder. How can we "cancel out" its effect if we can't measure it and more over we even don't know it exists ?

3.4.1 Randomisation (SPIRIT item 16a)

- For answering that we assign the treatment between groups using a random process

- Let assume I would allocate subjects either to an intervention or to a control.

- I create a randomisation list. Something like: TTCTCCCCTCTCTTTC (where T is treatment while C is control). So the first subject is allocate to treatment, the second to treatment, the third to control and so on. The list should have been created under some criteria (e.g. computer algorithm for generating random sequences).

3.4.2 Some properties

- Randomization is considered the gold standard for quality when estimating the treatment effect as it can deal with both known and unknown factors. For quality it is meant to have an estimator with low bias* (and theoretically unbiased): briefly the treatment effect estimate is "cleaned" from concomitant effects.
- Remember that gold standard does not mean the truth.

- When thinking about repeating the same study many times each time randomising (Senn1994):
- 1. over all randomizations the groups are balanced;
- 2. for a particular randomization they are unbalanced.

*Tech part: under the frequentist framework (i.e. there is a true parameter for which we draw samples and calculate statistics like the point estimate), an estimator (such as sample mean) is unbiased if, over many repetitions, the difference between the value of the parameter (true population mean) minus the expectation of estimator (mean of sample means) is zero. A non zero difference quantifies the bias.

4 Not using randomisation

i) As we have seen so far, randomisation provides the best quality for having an estimation of treatment effect.

ii) In drug development it is mainly used on "late phases" of drug development to get approval from health authorities, i.e. when we would confirm effects found in earlier phase on volunteers or very few patients.

iii) On "early phases" of drug development, depending on context we can use or not randomisation. However we can do like this, because it is a "learning" phase. We are interested to have quick rather than strong evidence results. Later on we should confirm anyway.

iv) As a general rule, if randomisation can be used then it must be used. If we would not implement, then there should be very strong scientific and ethics motivations. Generally speaking, scarcity of resources or expertise are not good motivations for not implementing it (in this case don't run the study).

5.1 Is there something else to be taken in account?

- We have seen how a study for evaluating a treatment effect should work: ideally it should have both a treated group and a control group, and it should be randomised. Hence we called them Randomised Clinical Studies or Randomised Clinical Trial (RCT).

- We have seen how randomisation can provide an estimate of the treatment effect with "high quality". However a clinical trial is more than methodology and there could be other type of bias that can affect the "quality" of estimation of treatment effect.

- For example knowing which treatment a subject can be allocated could pose some challenges.

5.2 Blinding/Masking (SPIRIT item 17a and 17b)

Awareness of intervention can introduce (SPIRIT2013):

ascertainment bias	in the measurement of outcomes, particularly subjective ones (eg, quality of life)
performance bias	in the decision to discontinue or modify study interventions (eg, dosing changes) (Item 11b), concomitant interventions, or other aspects of care (Item 11d)
exclusion/attrition bias	in the decision to withdraw from the trial or to exclude a participant from the analysis

Blinding or masking is the process of keeping the study group assignment hidden after allocation. E.g.

- tablets of experimental and control product of similar shape, weight and packaging.
- Same formulation (e.g. both syrups or tablets)
- Same taste (e.g. add sweeteners if a neutral and bitter product is there)

The journey so far

Experimental study	I modify clinical practice for evaluating if the "modification" works
Protocol	Need a protocol and sorting ethics for running a study
Effect	What I observe is a mixture of treatment effect and "noise"
1 subject	If I had a time machine I could know treatment effect
Many subjects	Best to use a control and experimental group for estimating treatment effect
Confounding	If I don't balance predictive factors my treatment effect is "confounded" by these factors (if unbalanced)
Randomisation	I can balance what I know. I can't balance what I don't know. Randomisation provides a method for having the "best quality" treatment effect. Best does not mean the truth
Masking	Avoid to take subjective (intentional or unintentional) decisions

6 Validity (SPIRIT item 10)

- The previous slide shows an array of methodologies possibly used on a clinical trial
- Remember they come from Good Clinical Practice: following GCP gives a way to produce results of good quality. Hence it is said that a trial has validity.
- How about generalisability of results of the trial? Generally speaking a trial is said to have INTERNAL VALIDITY

- However we also have to think whether the results are generalizable or applicable to my setting. E.g.:
  - i) Other populations
  - Ii) Specialised/General Centres
  - iii) Type of procedures implemented within trial
  - iv) Measures used for having evaluations
  - v) Compliance (did the patient took the intervention?)
- Depending on "generalization" we can say whether the trial has EXTERNAL VALIDITY. Remember that having internal validity is a requisite for external validity!

7 Randomisation

Let's go back to randomisation. Spirit guidelines ask to specify whether we used :
i) simple *versus* restricted randomisation
ii) Fixed *versus* adaptive (eg, minimisation) randomisation

Method	What	Pro	Cons
Simple	Generate a list with simple randomisation	Best unpredictability	- risky in case of logistical problems. - Can be affected by confounding in case of "small" numbers
Restricted	What is not a simple randomisation is restricted: e.g. blocked, stratified randomisation	- Can deal better with logistical problems and confounding	- Can be more predictable

7.1 Blocked Randomisation

Logistical problem: at the beginning of trial an operator did a wrong procedure as it was not trained enough. With simple randomisation it could happen that, by chance, a high proportion of subjects in experimental arm experienced the wrong procedure. This could be an element of bias!

The trial stopped halfway: I made the sequence for the full trial (e.g. 100 subjects) however the trial stopped halfway (50 subjects). At that time I randomised 38 on control and 12 on experimental.

Permuted block randomisation. Instead of letting the algorithm randomise the full sequence, after some subjects the sequence reset. Within block the sequence is random. Then there is a new block. E.g. with a block size of 4: ABAB BBAA ABAB AABB ABBA.

Why blocked: If there is a problem at the beginning or end, the "problem" influences both experimental and control (and we can expect to cancel out).

On protocol do not disclose block size: state you will use "blocks of varying size"

7.2 Stratified Randomisation

If we there is a predictive factor, in case of imbalance it could be a confounder. We can use a trick.

TRICK: randomise within level of a predictive (hence potentially confounding) variable. This trick is stratified randomisation. E.g. produce a randomisation list for high and low severity of a pathology (if severity is predictive of response): within each severity level we allocate half to treatment and half to control.

Moreover if the stratification variable goes in the statistical model, it means it is not counted in the error term (bigger error terms increase variability and hence standard error).

Stratification could be a mean of killing two birds with a stone: you deal with confounding and have more precise estimates!

We can also stratify for logistical problem. E.g. if envelopes are used and patients go in different wards (reparti), we can think of stratifying by ward (for each ward we'll have a pile of envelopes containing randomisation codes).

7.3 Minimisation

What we have seen so far where examples of fixed randomisation.

A common choice is going on STRATIFIED BLOCKED randomisation (e.g. use blocked randomisation within strata).

However stratified randomisation can only deal with few stratification factors. For instance if we use 3 stratification factors (e.g. male vs female, <= vs > 45 years, smoker vs non smoker) we have at least 8 strata. If we have too many strata or "empty" strata there could be problems on analyses.

If we have many predictive factors an alternative to randomisation is MINIMISATION. Briefly allocation is not based on randomisation but on particular algorithms for having similarity on groups. There is a component of randomisation as well.

The validity of minimisation is controversial. As a general rule go with randomisation unless you have strong arguments and the regulator is happy with this.

7.4 An example of randomisation chapter on a protocol

"At Visit 2 (Randomisation Visit), after confirming the eligibility of the participant and performing all baseline assessments, the participant will be centrally randomised in the study using the XXXX online randomisation system (XYZ).

Participants will be assigned to either experimental or control (allocation 1:1) based on a predetermined randomisation schedule stratified by study site using permuted blocks. The block sizes will not be disclosed, to ensure concealment (occultamento). The study blind will not be broken except in an emergency or regulatory requirement."

8 Superiority vs non inferiority

So far we have spoken about treatment effect. Among the examples of contrasts:

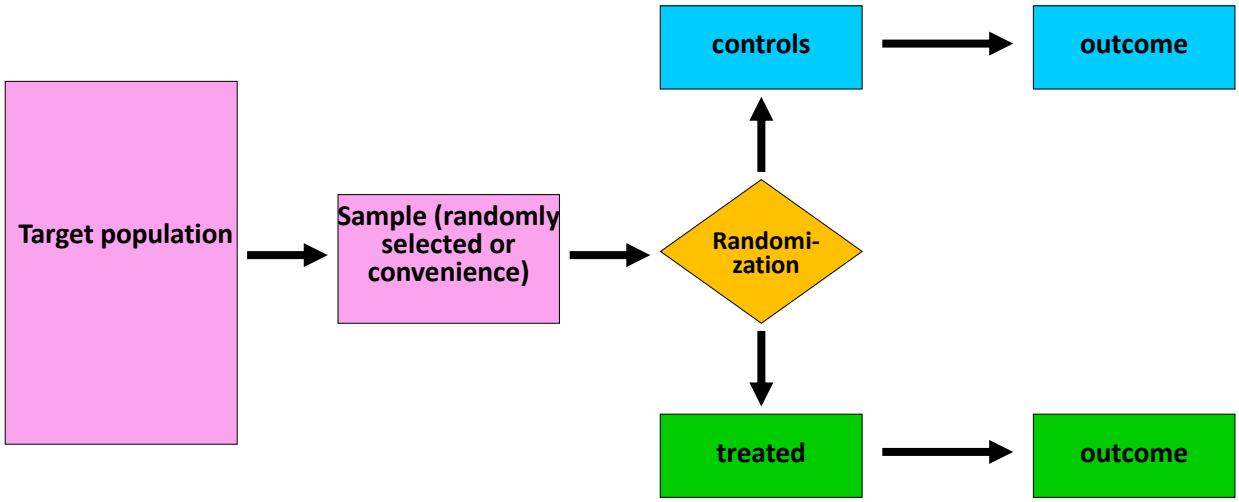
Experimental vs Standard
Experimental vs No Treatment
Experimental+Standard vs Standard

i) Choice of control depending on particular scenario
ii) Usually we think that a treatment should be better (superior) to control on an indicator
iii) In some conditions we can be interested that the experimental is "no worse" than the control that, in this case, is the standard therapy. This is NON INFERIORITY
iv) For instance we can be interested on a drug with similar efficacy (i.e. non inferior) but with less side effect or cheaper
v) Example: lower (Experimental) against full (standard) dose of a drug. Lower dose is expected to be cheaper and with less side effects

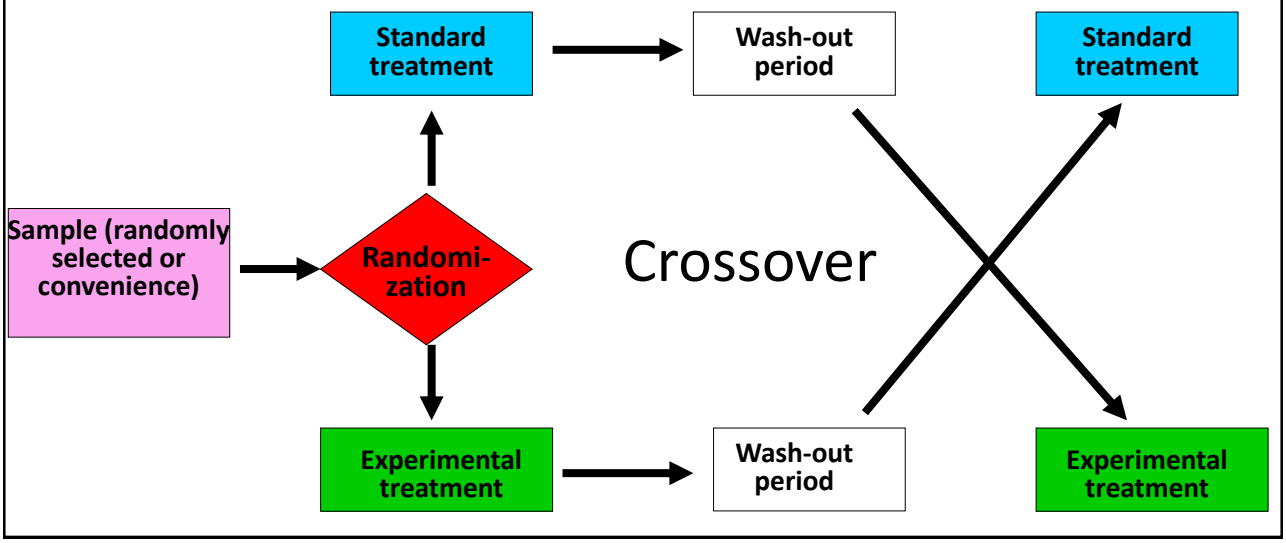
Randomized clinical trials (RTC), also named controlled clinical trials:

1) Parallel-group design
2) Cross-over design

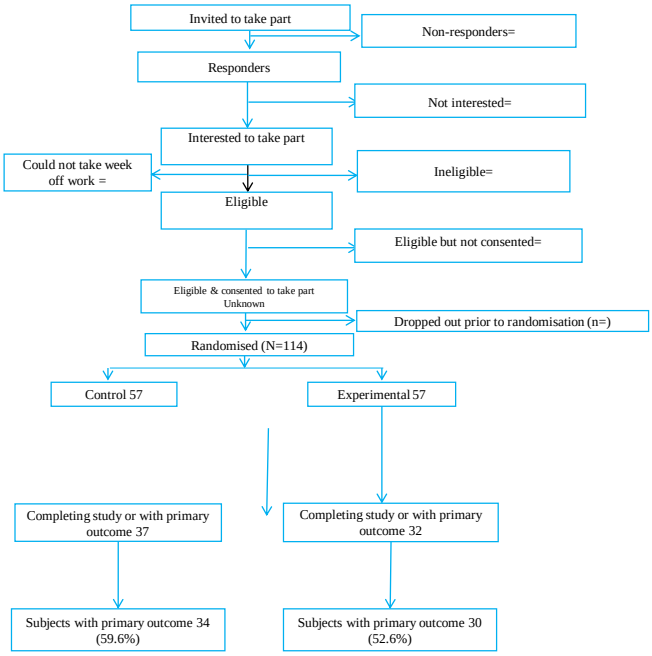
Parallel-group randomized clinical trial:
one group receives only the standard treatment (or placebo), while the other group receives only the experimental treatment.



Crossover randomized clinical trial.
Both groups receive both treatments in sequence. The sequence of administration is randomized.



9.1 **Parallel** controlled trials (SPIRIT item 8)



The "stereotype" of randomised clinical trial:

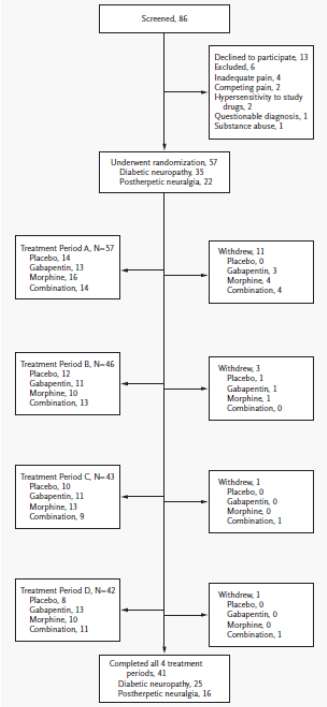
i) Subjects randomised to one of the intervention arm (2 is the simplest, however more arms are possible)

ii) Areas: late phase pharmacological interventions, non pharmacological interventions (e.g. physiotherapy, surgery, psychological interventions)

9.2 Default reporting of a parallel trial

So far we have seen some methodology top take in account. For reporting the results usually we use the CONSORT guidelines (Consolidated Standards of Reporting Trials). Among items to report for results:

Disposition data	Produce a CONSORT diagram (like the one seen on previous slide) and describe the flow of subjects
Baseline data	Report baseline data (i.e. just before randomisation) overall and by group.
Primary outcome	Report the results for the most important indicator on the trial (i.e. primary outcome) reporting the 95% Confidence Interval of the treatment effect and the p value
Sensitivity	Report ancillary analyses on the primary outcome for evaluating robustness of trial most important (primary) outcome
Secondary outcomes	Report results on the other indicators of the trial (secondary outcomes)



10.1 **Cross over** controlled trials

i) Subjects randomised to a sequence instead of a simple drug: a subject can take more than one drug and possibly all drugs tested.

ii) Subjects takes different treatments (or combination) in different time-points (usually called “periods”)

iii) Areas: pharmacological (early and late phases)

iv) Probably the most common design (Julious 2010), e.g. for pharmacokinetics and proof of concept (verifica parziale della fattibilità o fondatezza di un concetto)

10.2 Cross over features

Why crossover	Potentially they are more efficient (i.e. less subject to study for detecting the same effect) than parallel trials
Where to use crossover	On conditions where subjects return back to their baseline value prior to the start of each session and that there is equal carry-over between periods
What is carry over	The “treatment” effect does not end at the end of period but “carries over” on the next period
How to deal with carryover	Mainly by design (i.e. having a delay between periods called wash out; doing “proper” designs for which each treatment is preceded by a different treatment) and knowing the pharmacokinetics of drugs. After some time (e.g 4-5 half lives) the drug could not exert action hence there was a proper wash out

10.3 Multi period cross over

i) Simplest design is 2 treatments 2 periods or AB/BA design

ii) AB/BA design can be extended to more complicated designs; if more periods are there, then the design is called **multiperiod** cross over (e.g. 3 by 3)

iii) Potentially the number of sequences explodes (it's factorial for equal treatment equal periods). For example a 4*4 can potentially have 4! = 24 different ways while a 5*5 5! = 120. On the other hand in early phases of pharmacological development, there is a low number of subjects

iv) For even (pari)treatments, build a design called "William square" design: first sequence is 0, 1, t, 2, t-1, 3, t-2, …etc where t is the number of interventions minus 1. From the second sequence and so on, we add a "1" to the previous sequence with the exception of t which becomes zero . For a 4*4 we have 4 sequences instead of 24.

0	1	3	2
1	2	0	3
2	3	1	0
3	0	2	1

10.4 Multi period cross over

i) Odd (dispari) number: slightly different procedures. Total number of sequences is 2*n. For example for a 5*5 we have 10 sequences.

ii) Sequences are made with previous rule and the "mirror" i.e.: 0, 1, t, 2, t-1, 3, t-2… etc plus …t-2, 3, t-1,2, t, 1, 0.

iii) In this way we have a « small » number of sequences compatible with early phase drug development. Things could complicated however, this is an introduction.

10.5 Parallel versus Cross over

	Cross over	Parallel
Condition	Have to "reset" on new period	No constrain
Subjects to study	Lower	Higher
Sample size	Formula uses variance within: $\sigma^2_{\text{within}}=\sigma^2_{between}$(1-correlation). For the most inefficient case (correlation zero), we have half of the sample size of a parallel trial	Formula uses variance between
Stratification/ confounding	Less effect compared to a parallel trial as information mainly "within"	Useful to stratify and increase in precision
Within	Each subject is his/her own control: more informative on safety and efficacy (better approximation of counterfactuals)	Not possible
Default statistical analysis	Mixed model with subject (random effect), treatment, period (fixed effects)	Regression analysis with treatment, baseline, stratification factors

10.6 shall we always do cross overs?

i) From previous table it looks like that a cross over trial is much better than a parallel trial from the statistical point of view.
ii) On drugs we can "potentially" deal with carry over effect, but on non drug interventions it could be harder
iii) At analysis stage we can see that there is a different treatment effect between periods. This is called treatment by period interaction. If it is there, it could be very hard to interpret results. Some investigators consider only the first period. However doing like this means we have run an underpowered parallel trial!
iv) Going with a crossover trial needs more planning and thinking about what can go bad, while a parallel trial, even if more inefficient, is more robust. If not sure a solution could be: "don't do a crossover trial"

11.1 Variability in Clinical Trials

Senn (2001) identifies four sources of variation in clinical trials

TABLE 1
Sources of Variation in Clinical Trials

Label	Source	Description
A	Between treatments	The average difference between treatments over all randomizations (and hence over all patients). The 'true' mean difference between treatments
B	Between patients	The average difference between patients. (Averaged over both experimental and control treatments.)
C	Patient-by-treatment interaction	The extent to which the difference between treatment differs from one patient to another. (Equivalently, the extent to which the difference between patients being given the same treatment depends on treatment given.)
D	Within-patient error	The variability shown from treatment period to treatment period when the same patient is given the same treatment

11.2 Which variability can I ascertain?

TABLE 2
Identifiability and Clinical Trials

Type of Trial	Description	Identifiable Effects	Error Term
Parallel	Each patient receives one treatment	A	B + C+ D
Cross-over	Each patient receives each treatment in one period only	A and B	C + D
Repeated period cross-overs (sets of n-of-1 trials)	Each patient receives each treatment in at least two periods	A and B and C	D

In a parallel trial we can only ascertain variance between, in a crossover trial variance both between and within. Potentially with a new design (N of 1) we can ascertain a third source, i.e. subject by treatment interaction.

11.3 N of 1 trial

i) A subject is given at least two treatments for at least two periods: it is an extension of cross over trial on a single patient

ii) In the context of making decisions about an individual patient's care, N-of-1 trials have been considered to be among the most relevant and rigorous study designs for assessing treatment efficacy (Consort N of 1)

iii) Can have the same challenges as cross over trials. Not so widely used

12.1 Pilot Trials

i) So far we have spoken about the methodology. However trials pose lot of feasibility challenges (sfide di fattibilità). For example, we can compute a sample size and discover later on we are not able to include in the study the required numbers.

ii) There could be uncertainty of variability of indicators: when computing sample size we have to plug in (inserire) a measure of the expected variance. However we can go completely wrong.

iii) How to mitigate the risk of going completely wrong on recruitment (i.e. the patient to study) or estimating the variability?

iv) The idea is to run a small trial (or do an internal check). Small is related to time (i.e. should be quick to have results) as well on subjects to study. We are not going to put emphasis on efficacy indicators (i.e. for seeing if the treatment work). Instead we go on feasibility indicators (e.g. we see patients recruited or finishing the trial).

v) Pilot trials (Thabane 2010) are useful for evaluating whether we would be capable to "finish" the study hence not wasting resources (time, money). If the pilot is successful we go into the main trial, otherwise we stop.

12.2 Pilot Trials

	Pilot trials
Main motivation	Not wasting time and money if uncertain about logistics
Areas	Any
Indicators	Process (e.g. recruitment/retention rates), resources (e.g. how much time it takes running a process; do centres do what they should do ?), management (challenge by study personnel), scientific (estimation of SD)
Sample size	Precision of a feasibility indicator (e.g. Recruitment rate) or estimation of standard deviation (remember variance is distributed as a chi squared with n minus one degrees of freedom)
Reporting	Descriptive statistics and Confidence Intervals. Don't report p value
What is not a pilot trial (from Thabane 2010)	- A small single-centre study/ student project - I do a pilot as before I did like this and I got this published
Main focus (from Thabane 2010)	A study should only be conducted if the results will be informative; otherwise waste of the researchers' and participants' efforts

13.1 Cluster randomised trials

i) As we have seen, statistics and methodology are not the only driver for implementing a trial. There are decisions to be made depending on the setting, pathology and interventions.

ii) For example, if the interventions is not a drug but a surgery procedure, we need to train the surgical team. And it could happen that groups of patients, treated by the same health professional, are more likely to receive the same intervention. This latter phenomenon is called CONTAMINATION.

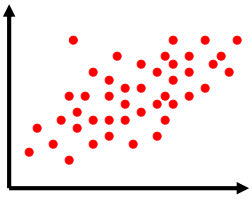
iii) Idea: instead of randomising subjects, I randomise the centres or the health professionals. This is called cluster randomisation. In this way "contamination" is dealt by design.

iv) Cluster trials can be a solution to some challenges like administrative reason, having investigator cooperation, enhancing subject compliance. Remember that cluster trials are statistically more inefficient compared to individually randomised trials. Reporting strategy is explained on CONSORT for cluster trials (Campbell 2012).

13.2 Cluster randomised trials

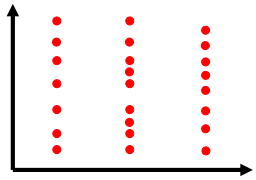
	Cluster trials
Main motivation	Contamination could be an issue
Areas	Mainly non pharmacological interventions
Sample size	Have to inflate sample size compared to conventional formula. This factor is called Design Effect (DE). The formula is : DE= 1+(m-1)*ICC
ICC	Intra cluster correlation. Usually a number between 0.01 and 0.05. The higher the number, the higher the correlation (results more similar). Correlation of zero means no design effect.
m	size of each cluster (e.g. each surgeon treats 10 patients)
Practical advice	Larger ICC and large clusters are detrimental to sample size. There is a diminishing return on increasing cluster size. Better to have new clusters.
Analysis	Clustering should be taken into account in analysis (e.g. mixed models with cluster as random term). If not, you have wrong standard errors and p values.

Pearson's correlation coefficient

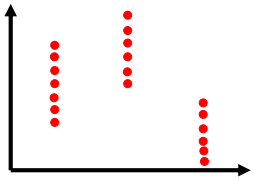


Intraclass correlation coefficient
Data are group in cluster

ICC closed to null (zero)



ICC closed to unity (one)



14.1 Adaptive trials

i) Methodology, presented so far, behaves in an "automatic pilot" way: when we start the trial, we wait until the end for seeing the results. Adaptive Design involves an array of methodologies to make decisions on a trial before this is ended.

ii) Among the decisions:

Do nothing and continue
Make an adaptation and continue
Stop for efficacy/futility/harm

iii) Among the methodologies (examples):

- Sequential Designs (some authors don't consider these adaptive)
- Calculate conditional power
- Sample Size re-estimation (SSR)
- Change on eligibility criteria
- Change allocation ratio
- Drop an arm

14.2 Adaptive trials

	Adaptive designs
May I used them directly?	No, you have to show some statistical properties like nominal type 1 error preserved, unbiasedness of the estimates.
What does it mean "to maintain type 1 error"	On adaptive design conventional statistics could not work. For example , if a z distribution is used, it could not be true anymore that between 1.96 and infinity you got a probability of 2.5% (under null hypothesis). This could be due to multiple comparison, conditional testing and so on. Indeed you can stop a trial earlier, so *p* values could not be correct.
How can I deal with this?	Check whether a published method is there (e.g. sequential designs deal correctly with nominal type 1 error). If not, you have to show (e.g. by simulations) that you got the level intended. For example you simulate 100.000 trials on computers. If overall you have 2500 trials considered "success" (irrespective on when they were declared success) under the null hypothesis, then you have evidence that your method could maintain type 1 error. Some tricks involve multiplication factors or changing a little bit statistics.

14.3 Adaptive Design

Some methodologies could be complex (time to spend and resources to invest).

Are there "straightforward" (semplici) methods (low hanging fruits)?

Low hanging fruit 1: Blinded Sample Size Re-estimation (SSR) (Kieser 2000, Kieser 2003) on continuous outcome:

Estimate interim variance with its LCI and UCI (lower and upper limit of Confidence Interval).

Compute new sample size.

Final statistics shall take in account the SSR (if unblinded).

For instance, if one-sided unadjusted alpha = 0.025, adjusted alpha accounting for SSR (performed at 50 patients) = 0.0233 [Tables on Kieser 2000].

14.4 Adaptive Design

Low hanging fruit 2: futility with conditional power:
if we "recalculate" the power on the trial this is no more "unconditional" (as at the beginning) but "conditional" on results given.

In classic hypothesis testing we have two hypotheses, null and alternative. When computing the conditional power we have (at least) three:

- Calculate the conditional power assuming the new data behave like the null (e.g. no difference)
- Calculate the conditional power assuming the new data behave like the alternative (e.g. pre specified effect)
- Calculate the conditional power assuming the new data behave like the observed effect so far

The trial could be stopped for futility if not "enough" conditional power is observed (Sully 2013, Lachin 2005)

14.5 Adaptive trials

i) Think about internal checks. In pragmatic trials it would be more useful to check to stop for futility rather than efficacy
ii) Think about whether the statistics is compatible with the logistics: fast recruitment and long term outcomes make hard implementing adaptation.
iii) if you go outside published methods, evaluate (e.g. simulations) that nominal type I error is maintained and unbiased estimates provided
iv) Beware that design and Quality Control process could be more time consuming compared to traditional methods, for instance: 5 (traditional) vs 300-500 (simulations) lines of STATA code (a statistical program); internal validation; quality control with external statistician

15 DMC

A Data Monitoring Committee (DMC) is an independent committee that monitor accumulating data of the trial [Damocles 2005].

The key role of a DMC is to monitor patient safety and advise the trial steering committee (TSC) where it believes the study protocol should be altered or the trial should be discontinued.

It is the only body involved in the trial that could have access to unblinded data.

It is the DMC that can suggest to stop/continue or adapt the trial on an adaptive design.

Clinical and statistical judgment are used to take decisions.