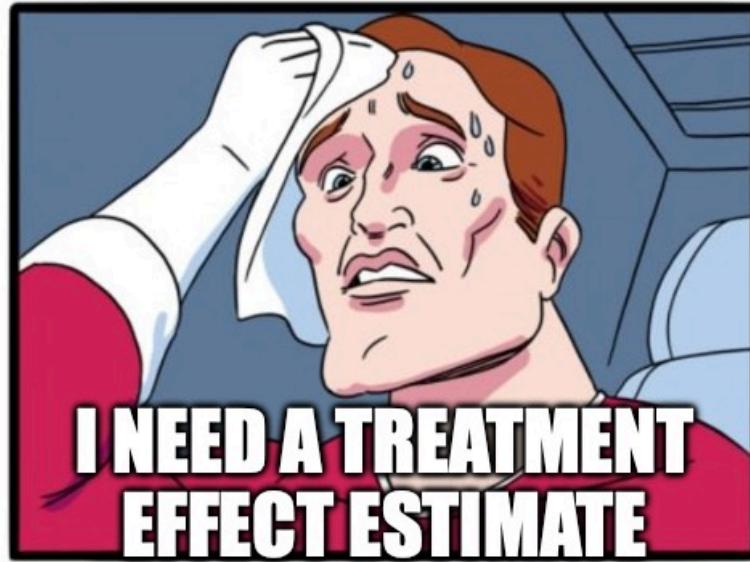


# Randomized Clinical Trials & Treatment Effects

Mabel Carabali

EBOH, McGill University

2024/10/01 (updated: 2024-10-28)



## **Expected competencies**

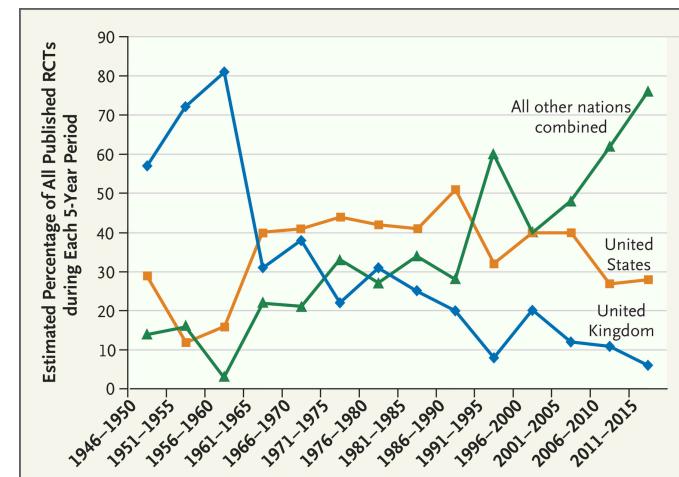
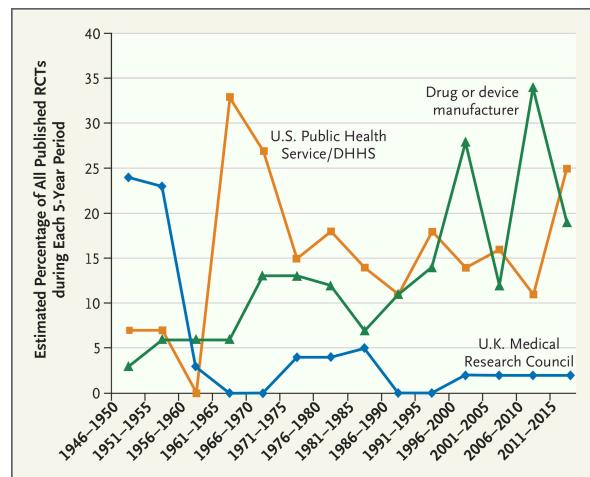
- Knows the rationale and understand different study phases in experiment designs
- Can recognize and describe RCTs according to their framework, objective, and design.

## **Objectives**

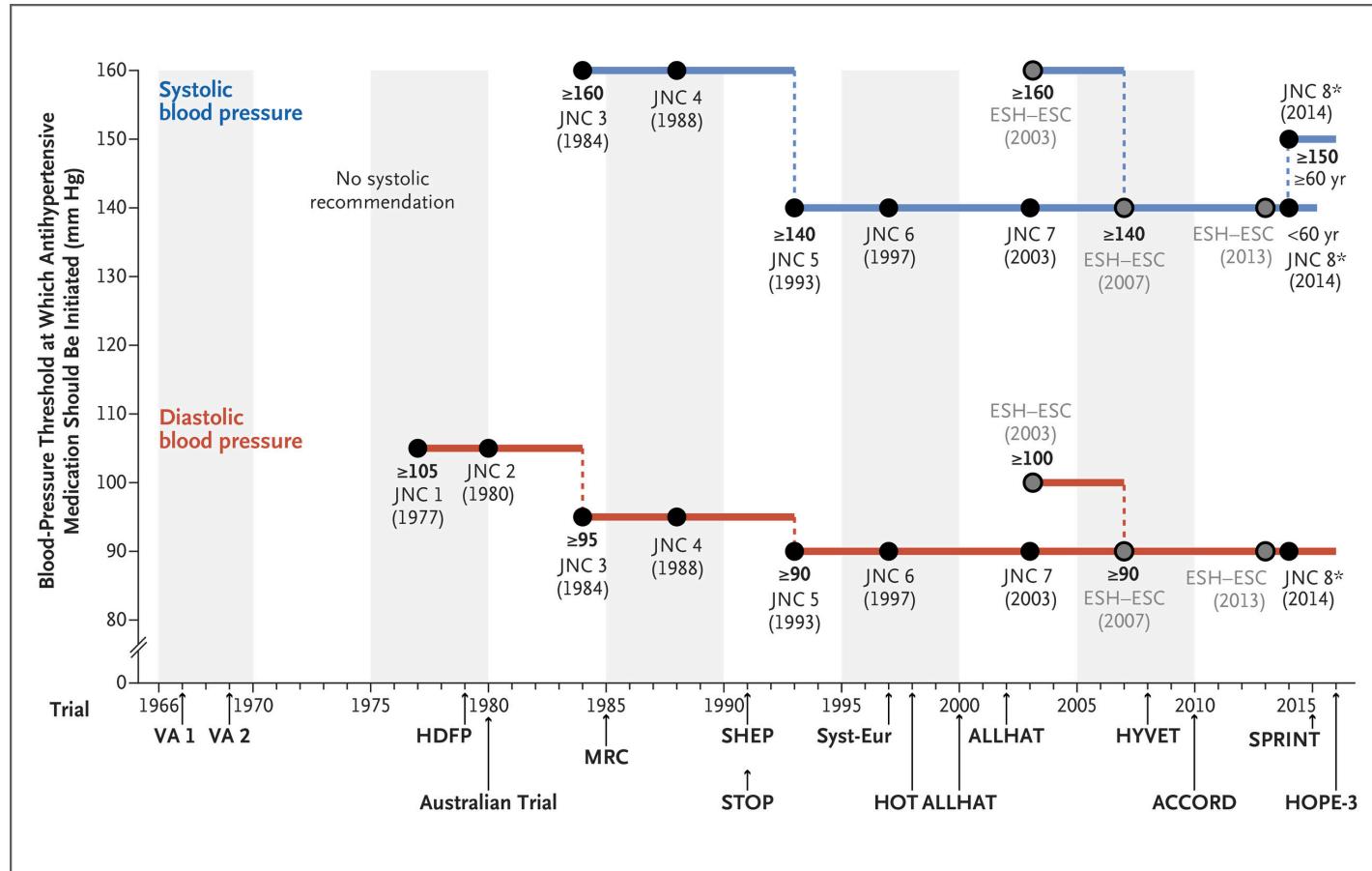
- Provide an overview of available designs for the realization of RCTs.
- Provide tools for the understanding of statistical analysis of RCTs.
- Provide an overview and revise the framework for Treatment Effects.

# RCT very brief history

- Hill formalized RCT methods (randomization, blinding, & statistical analysis) in 1940's
- Initially worries centered on ethics (today maybe more on \$)
- US Congress amendments to the FDA Act (1962), in response to thalidomide, new drugs must be proven efficacious in "adequate and well-controlled investigations"
- FDA (1970) interprets this to be RCTs
- Industry replaced governments and academic medicine as the primary producer of RCTs



# Lessons in Uncertainty and Humility — Clinical Trials Involving Hypertension



NOW 2024 Guidelines: OBPM >140/90 mmHg or H/ABPM >135/85 mmHg

Trials Influencing Blood-Pressure Thresholds at Which Antihypertensive Medications Should Be Used. N Engl J Med 2016;375:1756-1766

# Clinical trial protocols

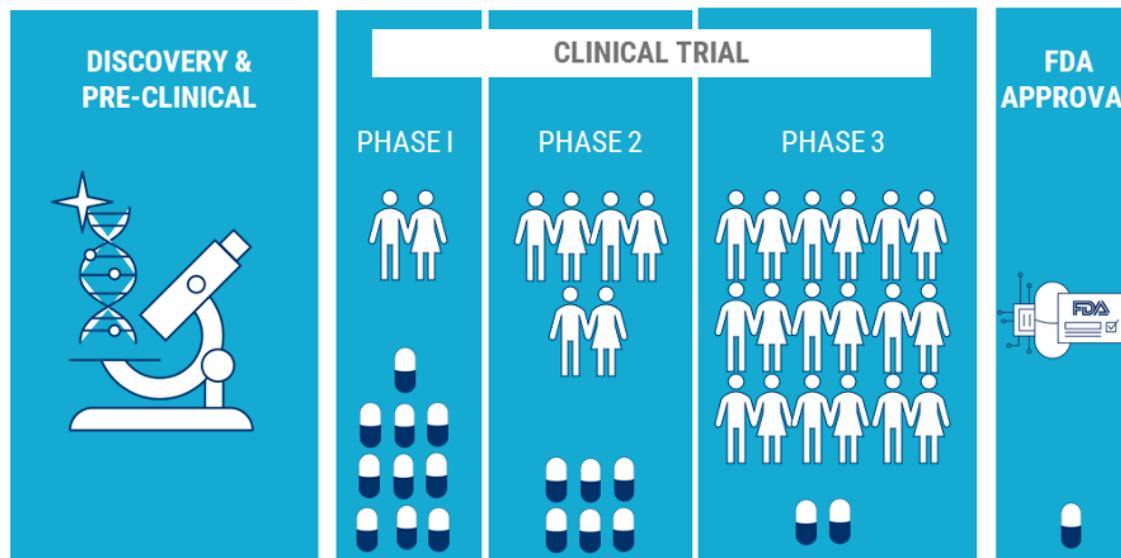
**SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials)** Intended to provide sufficient details about:

- Study rationale, intervention, trial design methods, study processes, outcomes, sample size, data collection procedures, proposed analyses and ethical considerations, with dissemination plans and administration of the trial.
- **Overarching goal:** to enable the research team to conduct high-quality, reproducible studies. Allowing external appraisal of the scientific, methodological and ethical rigour of the trial by relevant stakeholders.

The SPIRIT provides a checklist with **RECOMMENDATIONS** for the report of such protocols:  
<http://www.spirit-statement.org/>

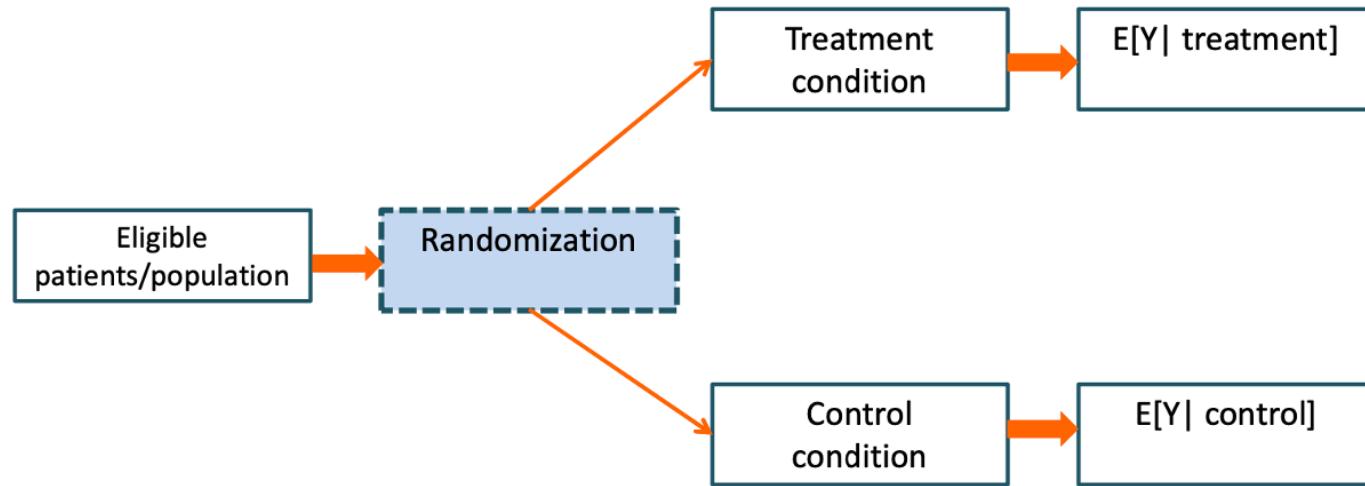
# What's a Clinical Trial

- Primary way to assess whether a treatment (molecule, posology, administration, devices, or technique) is safe and effective in people.
- To identify whether a treatment is more effective and/or has less harmful side effects than the standard treatment.



Source: cbinsights.com

# Why do we conduct Clinical Trials?



Assuming successful randomization, no losses to follow-up, and complete adherence to treatment assignment, **RCTs provide the most credible method for constructing the counterfactual and measuring the causal effect of a particular treatment (or exposure)**; the effect estimate from a randomized experiment will approximate the true causal effect.

# Always back to → Research Questions!

A research Question should follow the FINER characteristics:

**Feasible:** Adequate number of subjects & Adequate technical expertise; Affordable in time and money; Manageable in scope

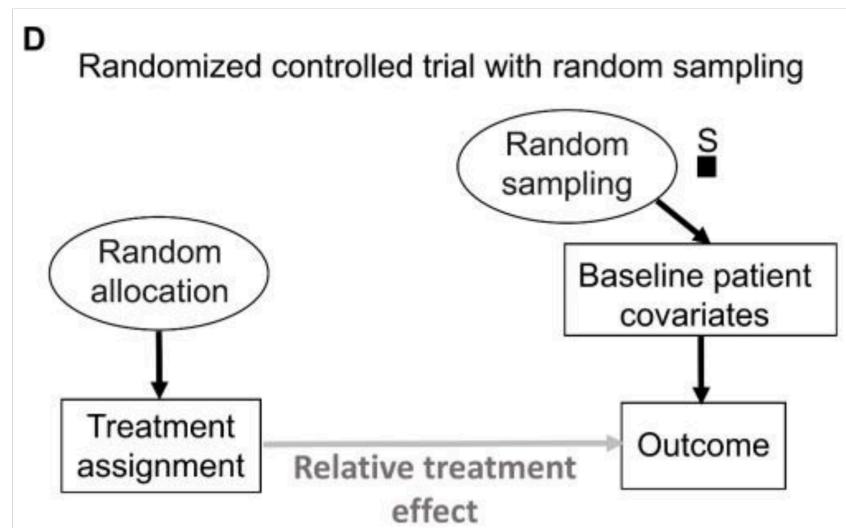
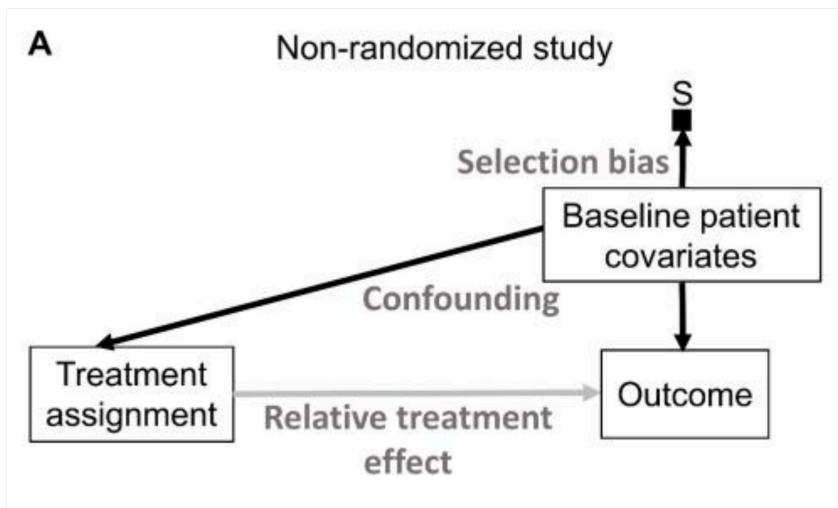
**Interesting:** Getting the answer intrigues the investigator and academic community

**Novel:** Confirms, refutes or extends previous findings; Provides new findings

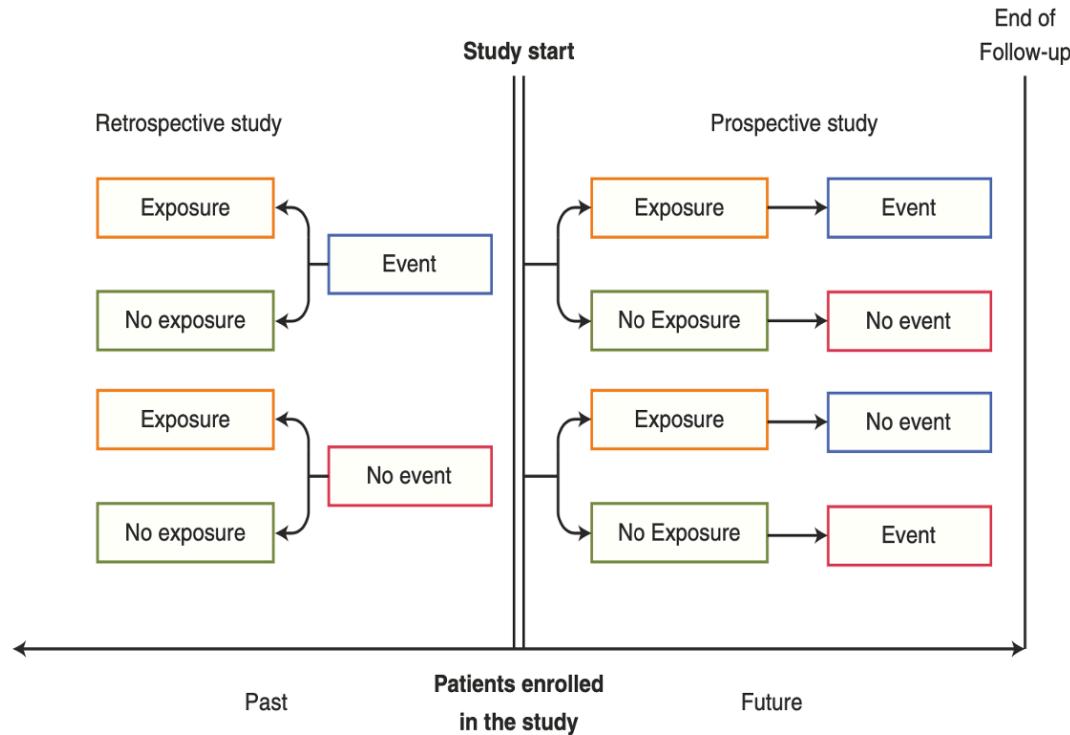
**Ethical:** Amenable to a study that institutional review board will approve

**Relevant:** To scientific knowledge; To clinical and health policy and To future research

# Trial Design, why prefer RCTs?



# Trial Design



Description of trial design including type of trial (eg, parallel group, crossover, factorial, single group)

# Type of Trial: Efficacy or Effectiveness

## Efficacy

- How well an intervention can work under ideal circumstances when administered by well-trained experts and perfectly compliant recipients

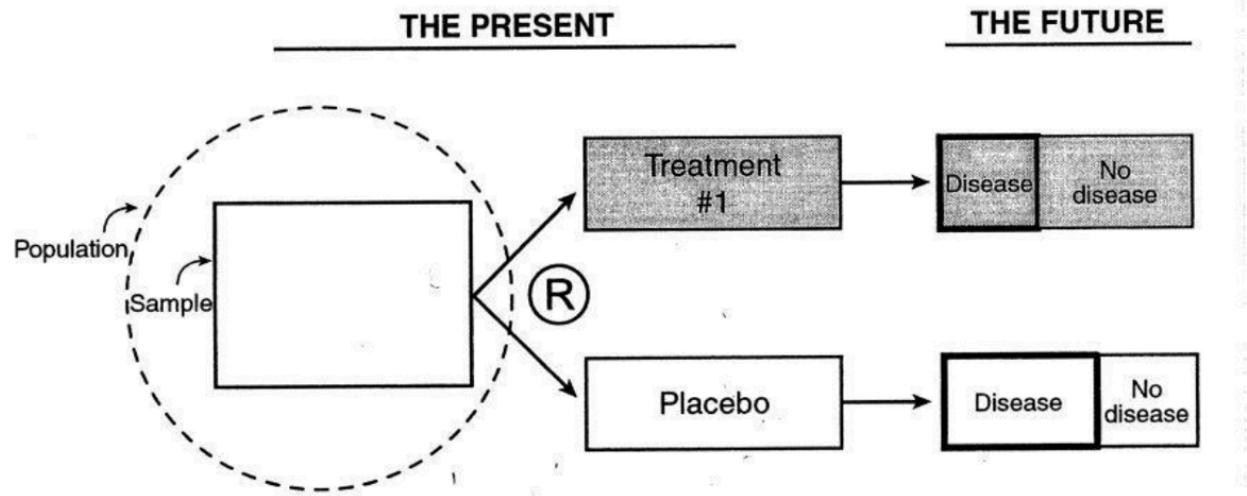
## Effectiveness

- How well an intervention does work under “field conditions”.
  - when administered by ordinary practitioners and offered to a relatively unselected (or ‘less’ selected) target population.

## RCT: In comparison to ...?

- **Absent (Nothing):** ← Can't make a causal inference (not and RCT)
- **Placebo:** ← Resembles the experimental intervention and main function of placebo is to help keep subjects unaware of their assignment status (intervention vs. control)
  - Useful for studying the rates of side effects or adverse reactions to the treatment.
  - Straightforward conclusion about.
- **Active alternative:** ← More ethical and relevant to compare the existent or new treatments.
  - Useful for prioritizing use of new treatment.
  - May present ambiguity in non-inferiority trials.
- **Usual care:** ← Suitable when current practice is a well-established treatment or when it's variable and hard to standardize

# Type of Trial – Basic RCT Design



**FIGURE 10.1**

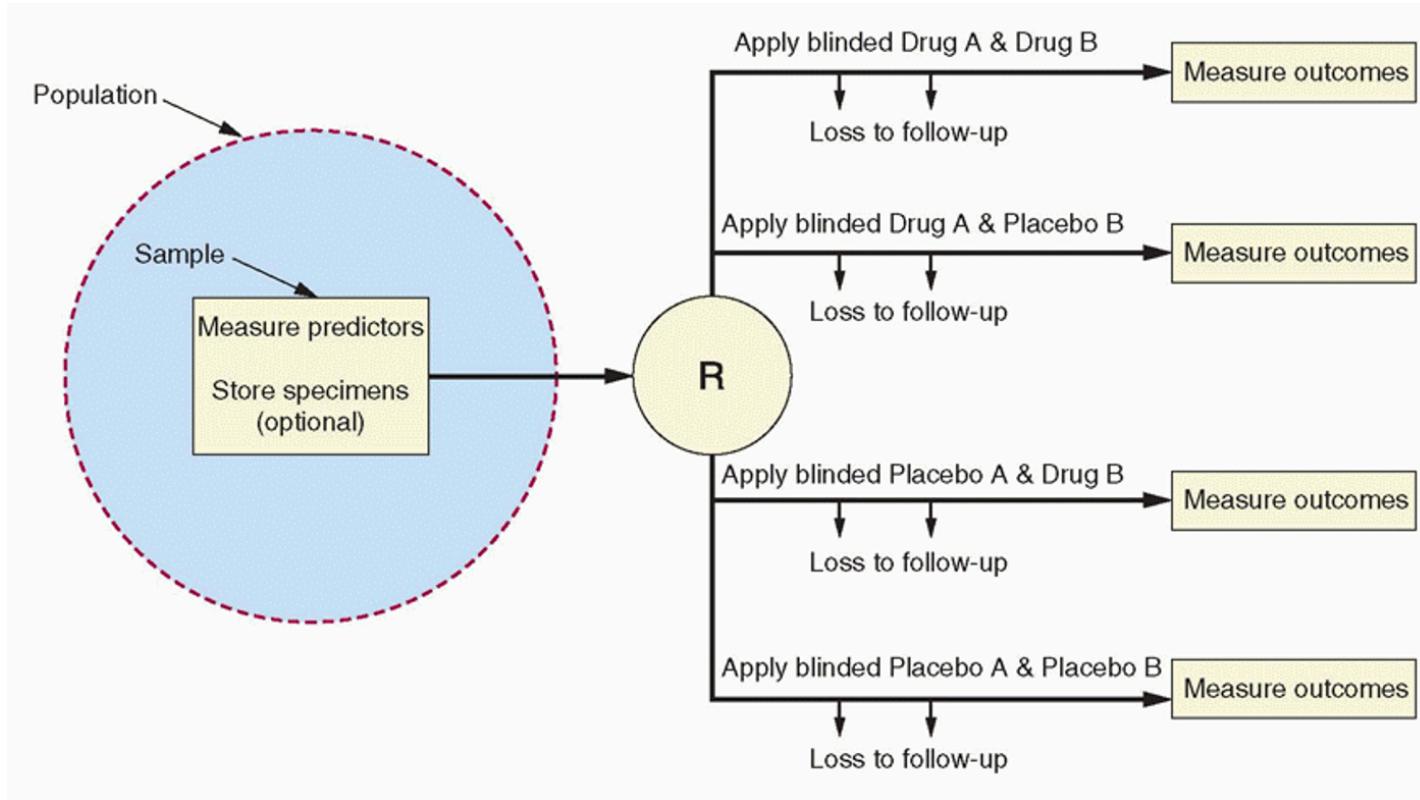
In a randomized trial, the investigator (a) selects a sample from the population, (b) measures baseline variables, (c) randomizes the participants, (d) applies interventions (one should be a blinded placebo, if possible), (e) follows up the cohort, (f) measures outcome variables (blindly, if possible) and analyzes the results.

## Some other RCT designs

To try and overcome some of the limitations of classic **parallel-group fixed sample size RCTs**, some other designs include

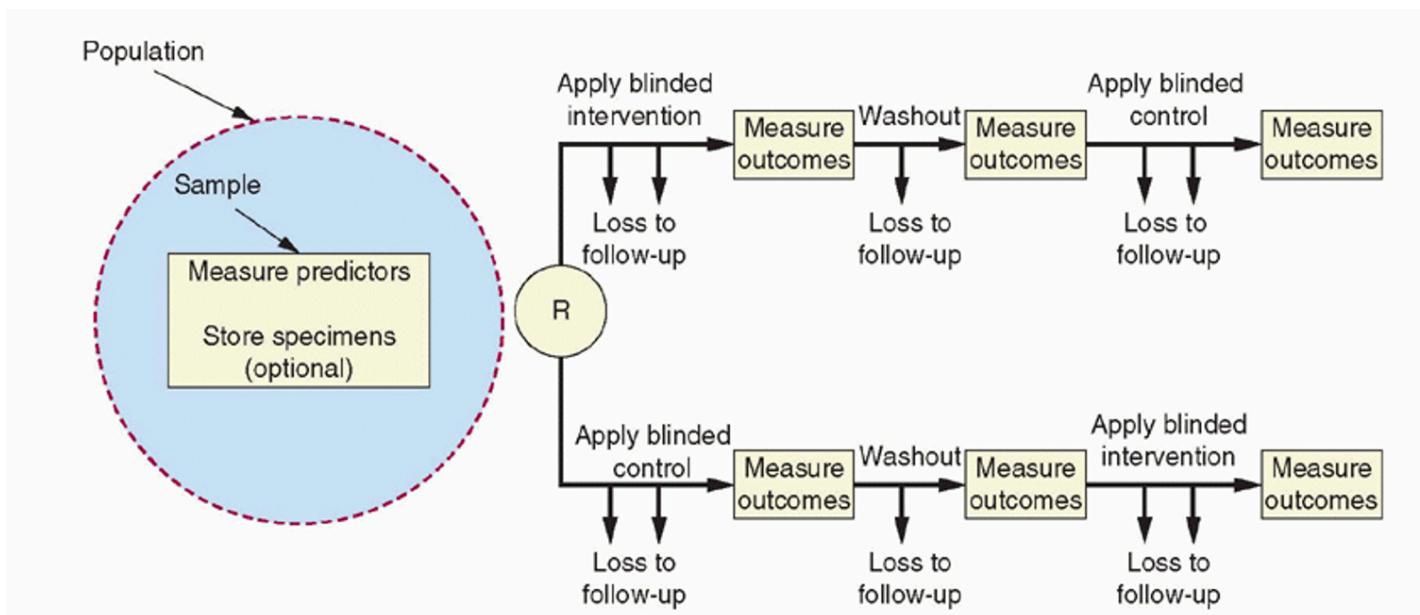
- **Factorial** trials (evaluate 2 or more treatments simultaneously)
- **Cluster randomized** trials
- **Crossover** trials
- **Stepped wedge** trials
- **Pragmatic** trials (to overcome generalizability concerns)
- **Randomized registries** (overcome generalizability and cost issues)

# Factorial Design

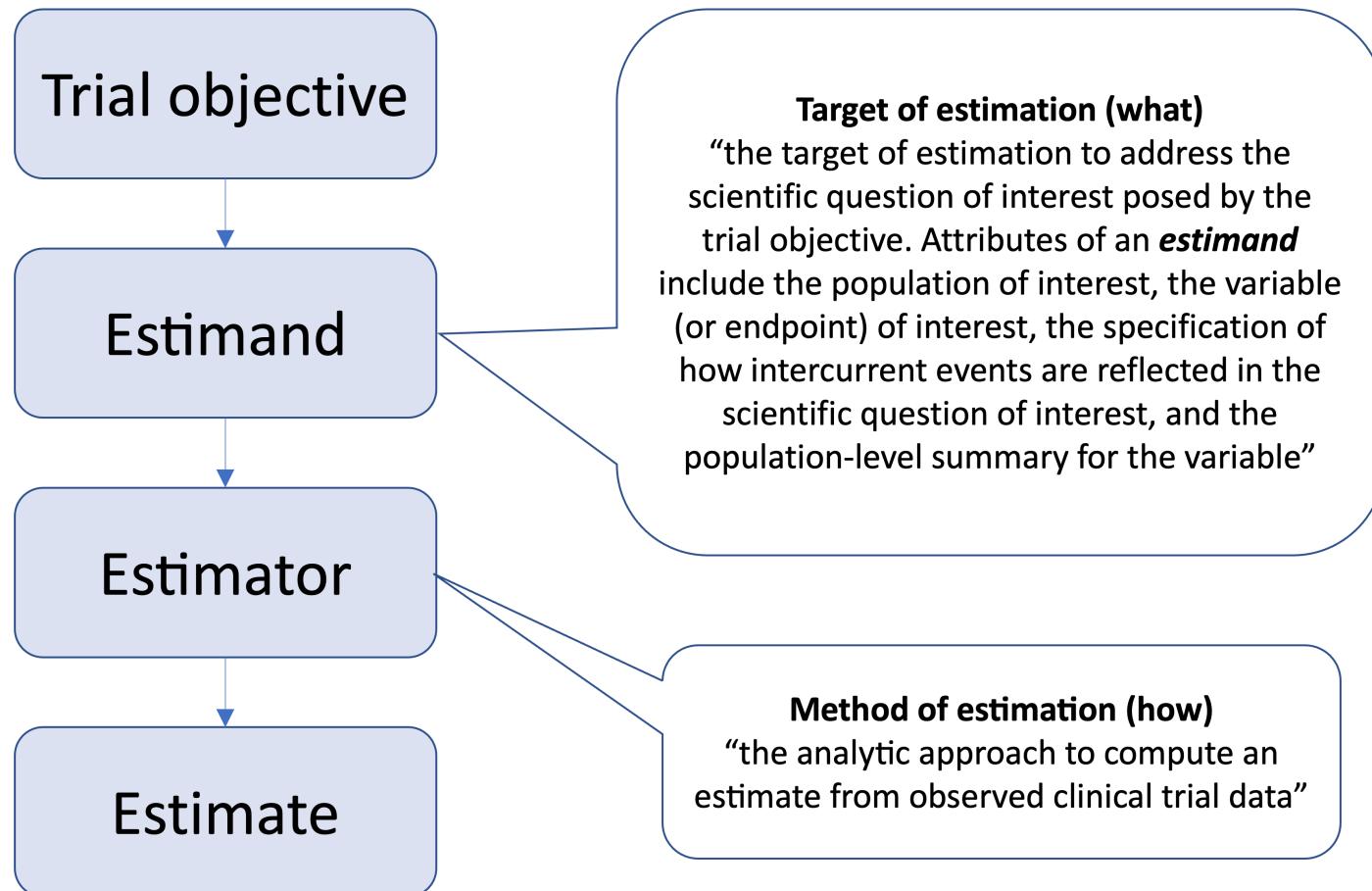


Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing Clinical Research.  
Lippincott Williams & Wilkins; 2013

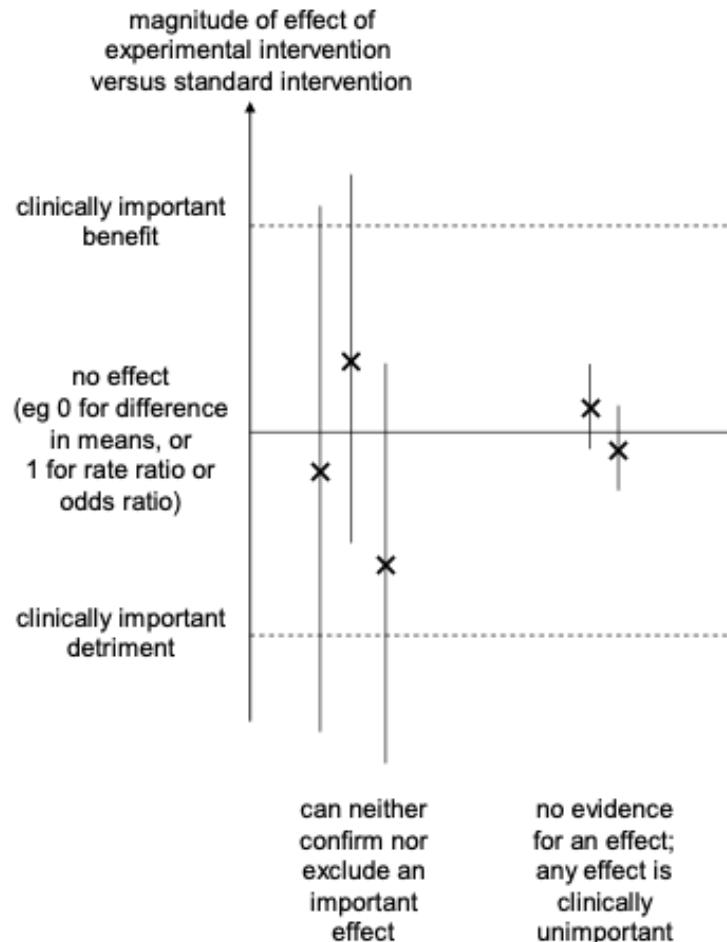
# Planned Cross-over RCT design



# Estimation: What and How?

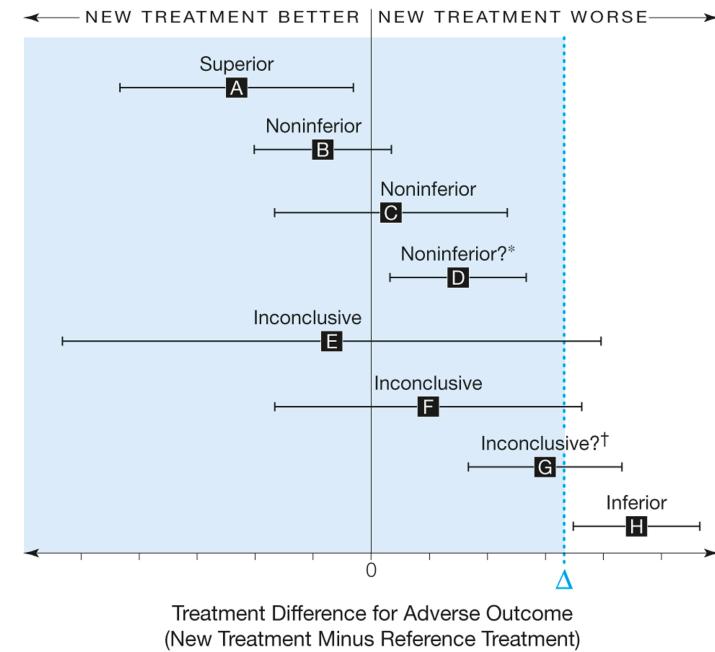
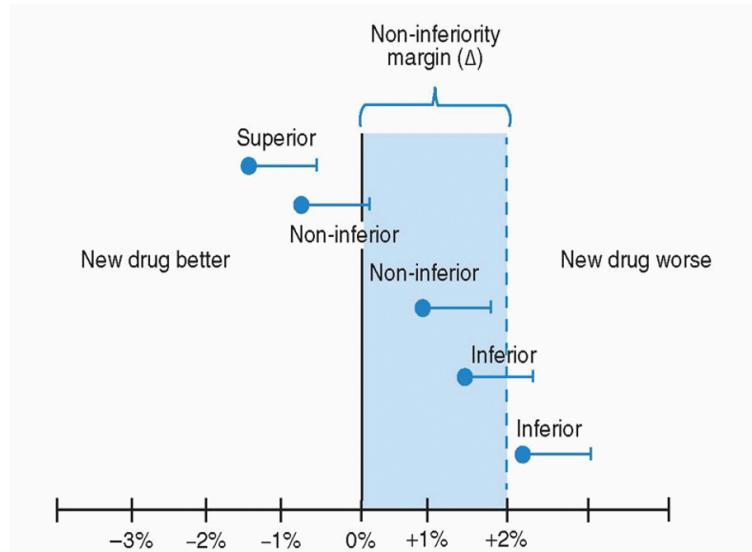


**Estimation vs Significance:** Suppose a trial compares the **efficacy of two interventions**. Does absence of *statistically significant difference* means equivalence of efficacy?



Remember the **Range of Practical Equivalence ROPE**

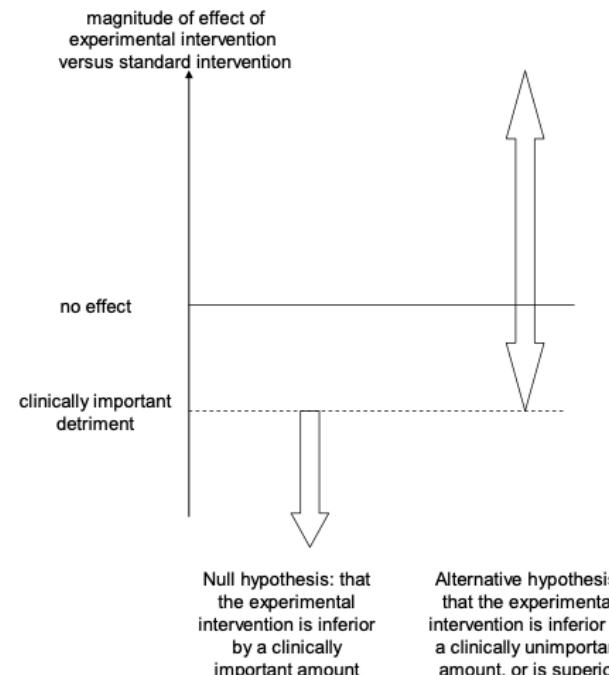
# Framework: Superior vs. Inferior?



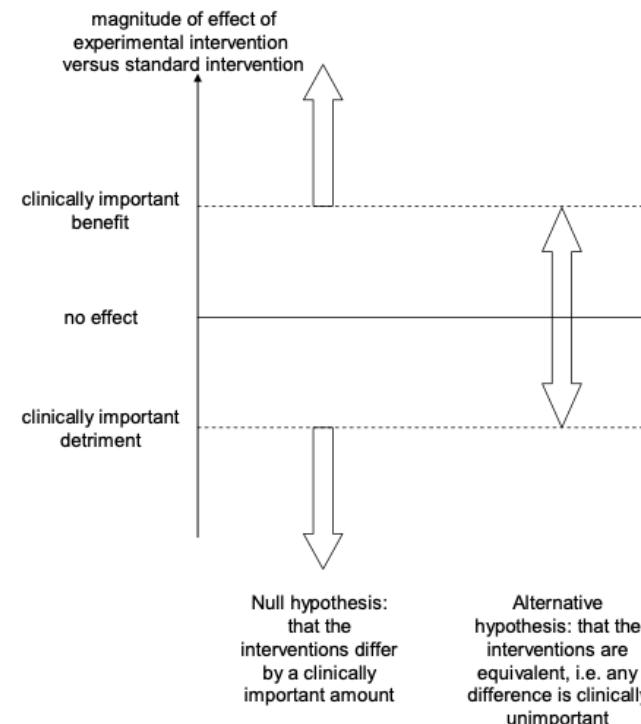
- Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement. *JAMA*. 2006;295(10):1152–1160. doi:10.1001/jama.295.10.1152.
- Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing Clinical Research. Lippincott Williams & Wilkins; 2013

# Framework: Superior vs. Inferior?

## Null & alternative hypotheses for non-inferiority



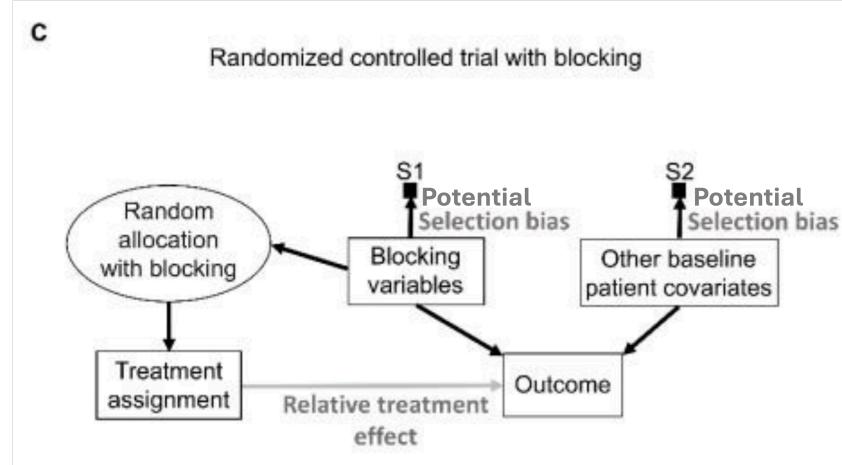
## Null & alternative hypotheses for equivalence



# Type of Randomization.

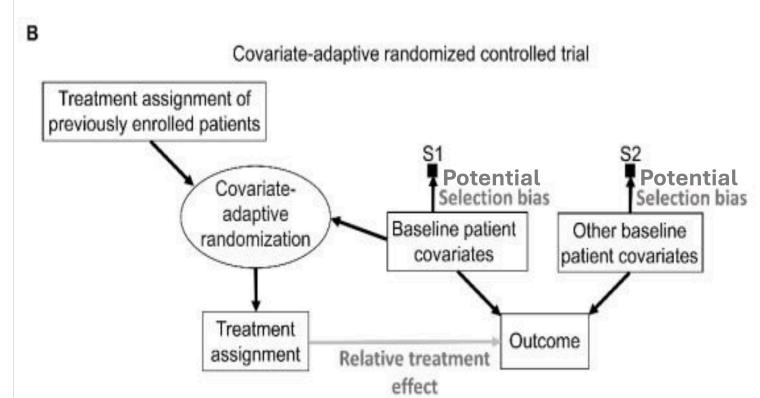
## Randomization with Same Probability

- Simple Randomization
- Blocked Randomization
- Stratified Randomization

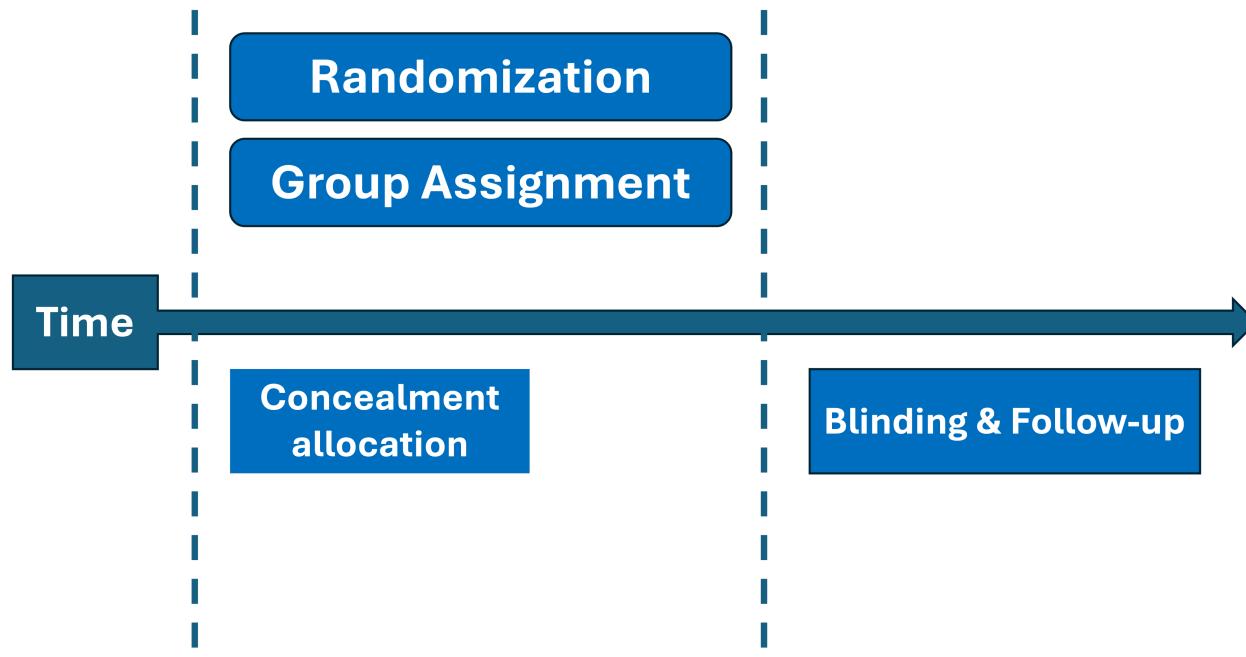


## Randomization with Varied Probability

- Treatment-Adaptive Randomization
- Adaptive Randomization
- Response-Adaptive Randomization

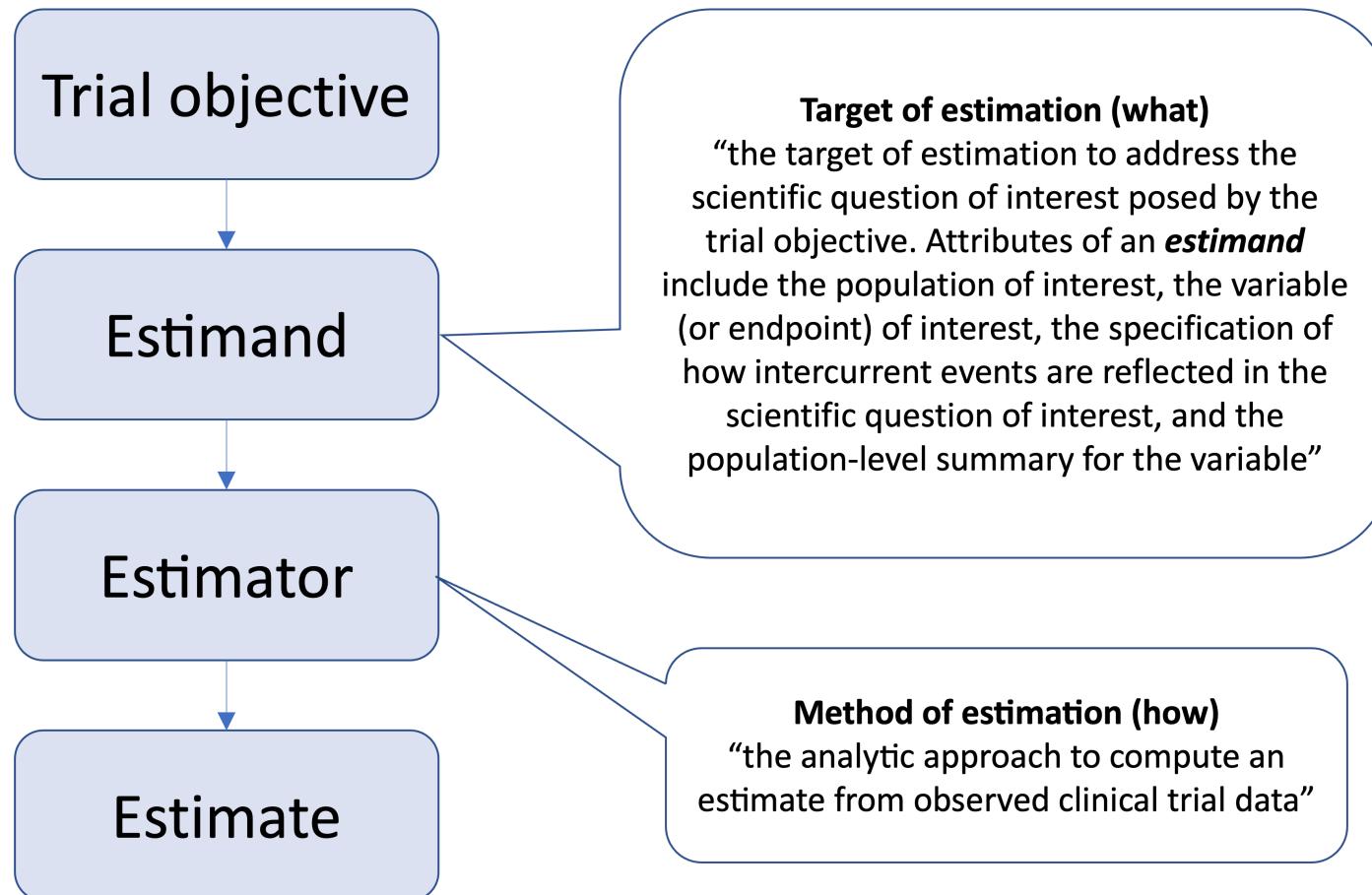


# Considerations for successful RCTs



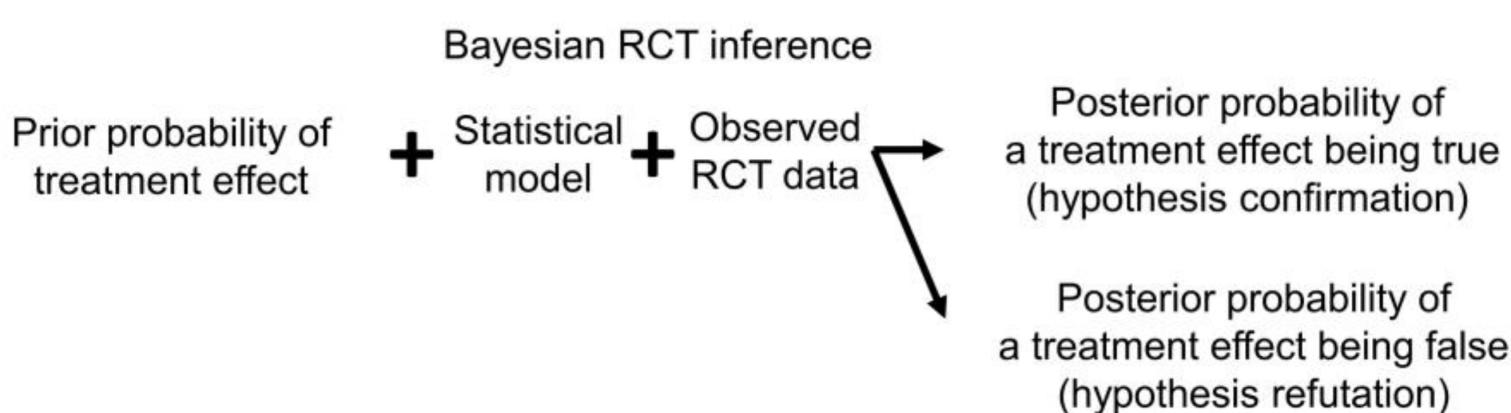
Allocation concealment and randomization

# Estimation: What and How?

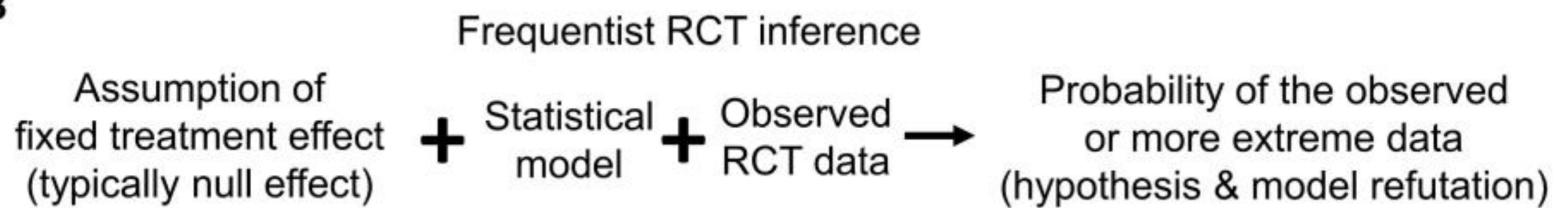


# Frequentist and Bayesian Inference for RCTs

A



B



## RCTs: it's still about the research question

Without a precise description of the trial objective and the treatment effect that is targeted for testing and estimation there is a risk that:

- The study will not be designed appropriately to address its objective;
- The **statistical analyses will be misaligned** to the trial objective and the target of estimation;
- The treatment effect that is reported will be **incorrectly interpreted**, which risks misleading decision makers.

# What do we want to know?

We really care about the difference between  $Y^0$  and  $Y^1$ .

Let  $\delta_i = y_i^1 - y_i^0$  (observed values)

$$E[\delta] = E[Y^1 - Y^0]$$

$$E[\delta] = E[Y^1] - E[Y^0]$$

This is the definition of a **Treatment Effect (TE)**.

# Why do experiments work?

Since

$$T \perp Y^0$$

$$T \perp Y^1$$

Then

$$E[Y^0|T = 0] = E[Y^0|T = 1]$$

$$E[Y^1|T = 1] = E[Y^1|T = 0]$$

In other words, strong assumption that actual observations = unobserved counterfactuals

## Or in other words...

**In a properly executed experiment, no association between potential outcome variables and treatment assignment**

$$E[Y^0|T = 0] \simeq E[Y^0]$$

$$E[Y^1|T = 1] \simeq E[Y^1]$$

So...

$$E[\delta] = E[Y^1] - E[Y^0] = E[Y|T = 1] - E[Y|T = 0]$$

Treatment effect =  $\Delta$  between **observed** treatment and control averages

# What is this treatment effect?

$E[\delta]$  is the expected value (mean) of the difference between each unit's value of  $Y^1$  and  $Y^0$ , denoted the **average treatment effect (ATE)**

- In a sample, this is the **sample average treatment effect (SATE)**.

Even though the **individual differences are unobservable** (because either  $Y^0$  or  $Y^1$  will be counterfactual for each unit), we can estimate the mean difference via experiment.

$$\text{SATE} = \frac{1}{n} \sum_{i=1}^n (y_i^1 - y_i^0)$$

## What's the role of randomization again?

- Well performed experiments identify causal effects, **ATE**, because cases are randomly assigned to the treatment and control group and are, therefore, identical *on average*, on all pre-treatment characteristics.
- Experiments of interest for causal interpretation are **randomized controlled trials** (or RCTs)
- **(S)ATE** = the average treatment effect for **switching** everyone's treatment.

# Causal estimands and RCTs

The clinical question defines the estimand and not the inverse.

- The statistical analysis must then align with the choice of estimand.
- Classically, the estimand for RCTs has been **ATE** (in reality a SATE) and the statistical approach is based on **intention to treat (ITT)**  
**ITT** measures the effect of randomization on the outcome of interest, in all patients, as randomized, **irrespective of compliance to the planned course of treatment**.
- ITT ignores all intercurrent events (i.e., intermediate changes from randomization to outcome).

Why would one be interested in this estimand?

- Maximally exploits the advantages of randomization.
- Its preservation provides a secure foundation for statistical tests.

# How to Analyze RCTs?

**Super Easy (statistically)!**

ITT assumes that all confounding have been removed with randomization, and assuming successful allocation and adequate measurement, there should not be any selection bias or measurement error.

**ITT requires only 2 steps:**

1. Check balance of covariates.
2. Contrast results (i.e., estimate the difference (Absolute or Relative) in average outcome between groups).

# Causal inference assumptions (yes, they still apply to RCTs)

**Identical to causal assumptions in non-experimental designs**

- Consistency
- Positivity
- Exchangeability (“ignorability of the treatment assignment and measurement of the outcome”)
  - Requires “no unmeasured confounders and no informative censoring”

Randomization takes care of 1st element (if sample large enough) but 2nd requires good trial conduct with no or minimal lost to follow-up

**In general, only no informative loss to follow-up is likely to be violated in any sufficiently large randomized trial.**

## How to analyze RCTs?

**Randomized treatment assignment only protects against confounding at the time of randomization**

Therefore, estimation procedures that involve post-randomization confounding for treatment adherence and/or loss to follow-up, needs to be very thoughtfully considered, otherwise bias may occur despite randomization.

## How to analyze RCTs? ...It depends !

If you **trust** the RCT (all assumptions are met) then follow the two steps:

1. Check balance, and
2. Contrast outcomes (no need to adjust for anything else).

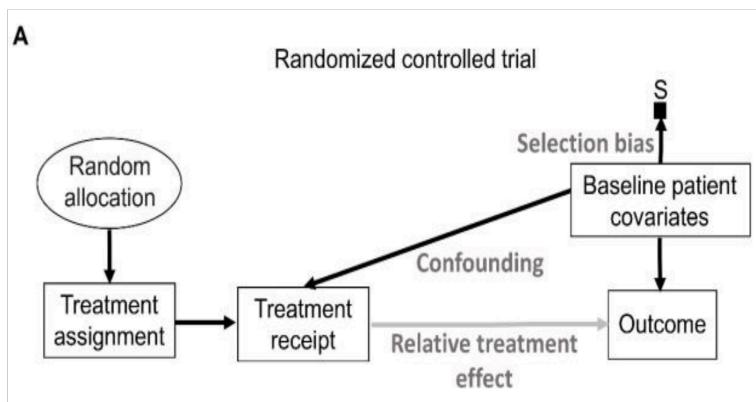
If you consider that there will be **deviations** to the RCT that can alter

$E[\delta] = E[Y^1] - E[Y^0]$  and hence  $= E[Y|T = 1] - E[Y|T = 0]$ , then you may need to:

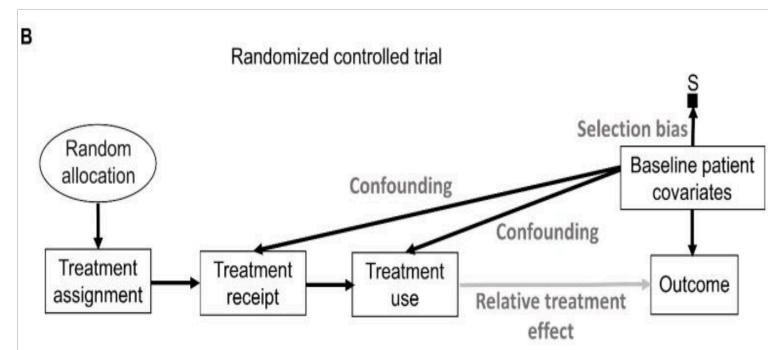
1. Identify the type/mechanism of the variation/deviation of the RCT.
2. Identify if/how to make valid causal inferences with the data.

## *Deviations from the planned RCTs:*

### Treatment



Despite Randomization, **Treatment was not always received**.



Despite Randomization, treatment was **not always received or if received not always used**.

# What are the causal estimands for RCTs in presence of such deviations?

If RCT is as planned, then ITT is used to estimate the ATE.

But, depending on the **deviations** in the trial, one can also be interested in the **per-protocol (PP)** effect: The effect of **adherence** to assigned treatment strategy; **Average Treatment Effect among the Treated = ATT**)

**If perfect treatment adherence by 100% of participants, PP = ITT ; ATE = ATT**)

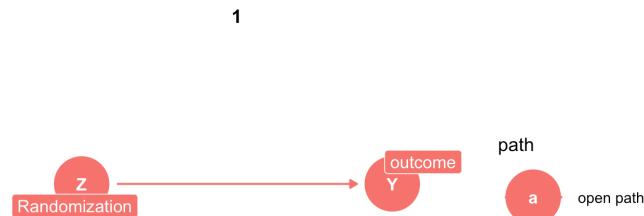
PP effects most useful in pragmatic trials (which inform a clinical or policy decision by providing evidence for adoption of the intervention into real-world clinical practice) because patients and providers want a measure of effectiveness that is not influenced by adherence.

PP effect is trial-specific, and more than one per-protocol effect definition is possible for a given trial (e.g. Tx adherence of 80%, 90%,... or adherent to a chosen time).

# RCT DAGs

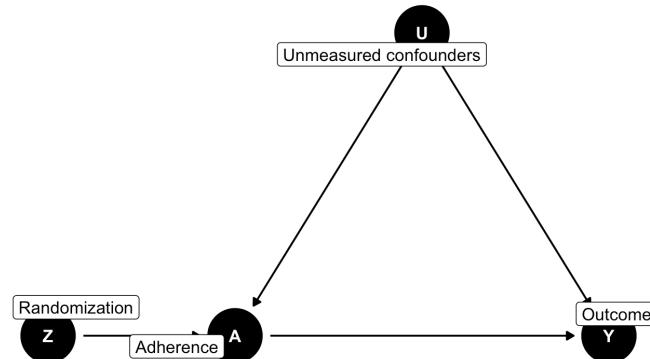
## ITT DAG

RCT:  $H_0 = \text{FALSE}$   
ITT analysis



## ITT DAG with mechanism

RCT:  $H_0 = \text{FALSE}$   
ITT with adherence



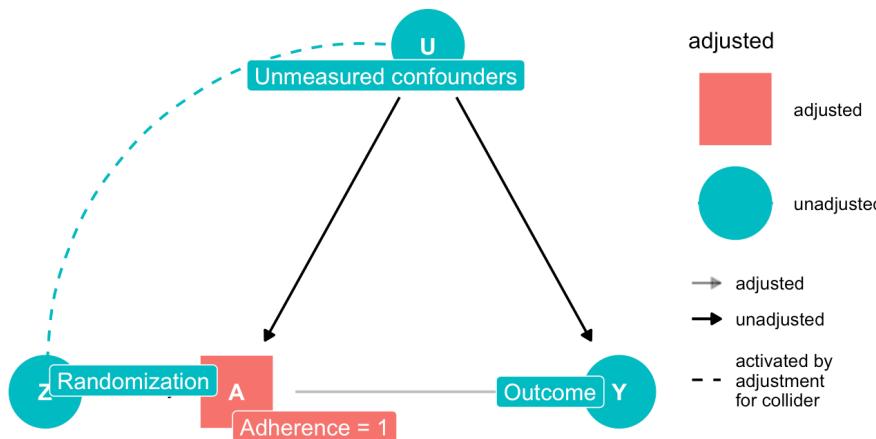
```
## $paths
## [1] "Z -> A -> Y"           "Z -> A <- U -> Y"
##
## $open
## [1] TRUE FALSE
```

# RCT DAGs

## Per-Protocol DAG conditioning on adherence

RCT:  $H_0 = \text{FALSE}$

PP analysis



```
## $paths
## [1] "A -> Y"      "A <- U -> Y"
##
## $open
## [1] TRUE TRUE
```

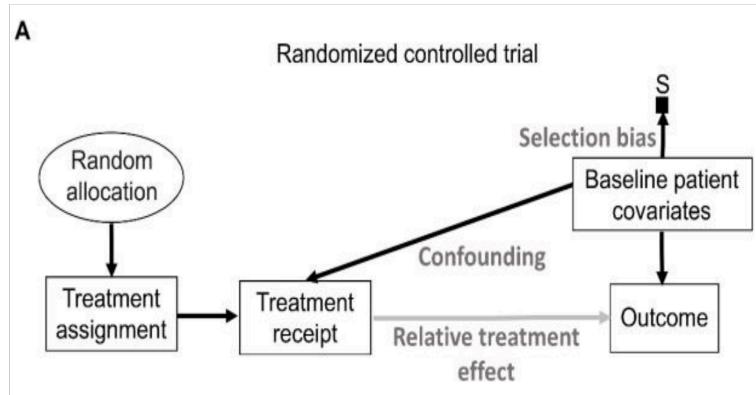
# Per-Protocol Effects (PP) approaches

Two common approaches to estimate the per-protocol effect are:

- i) Comparing the outcomes of those who took treatment A=1 and treatment A=0 (regardless of the treatment they were assigned to), e.g.,  $\Pr[Y=1 | A=1] - \Pr[Y=1 | A=0]$ , referred to the **as treated analysis**.
- ii) Comparing the outcomes of those who took treatment A=1 among those assigned to Z=1 and treatment A=0 among those assigned to Z=0, e.g.,  $\Pr[Y=1 | A=1, Z=1] - \Pr[Y=1 | A=0, Z=0]$ , referred to the **per protocol analysis**. **The ATT**

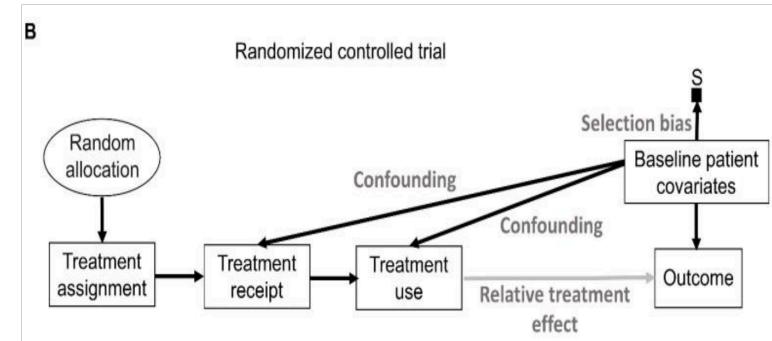
# Per-Protocol Effects (PP) approaches

## Per-Protocol



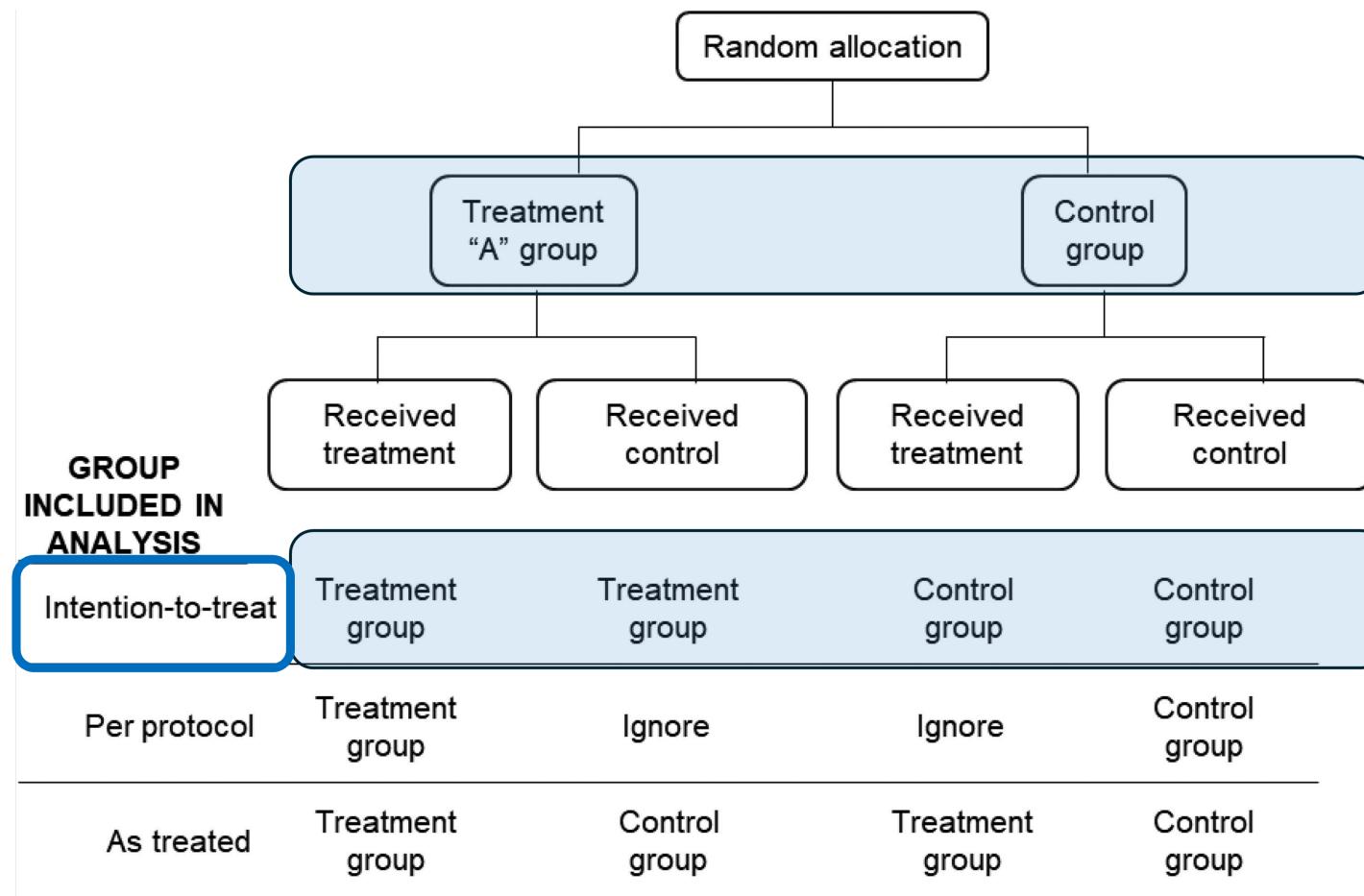
Despite Randomization, treatment was not always received.

## As treated.



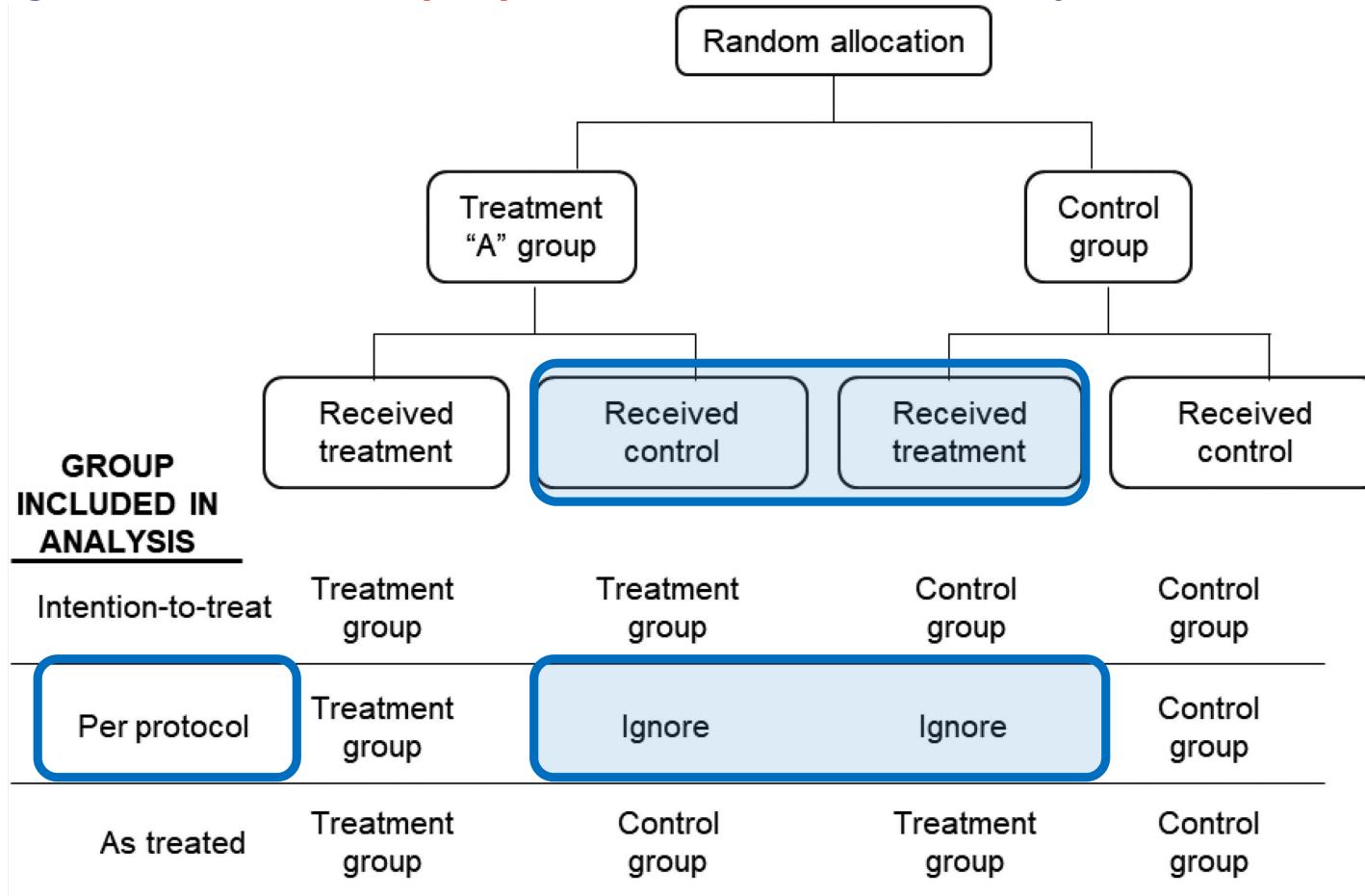
Despite Randomization, treatment was not always received or if received not always used.

## Comparing intention-to-treat, per-protocol, and as-treated analyses.



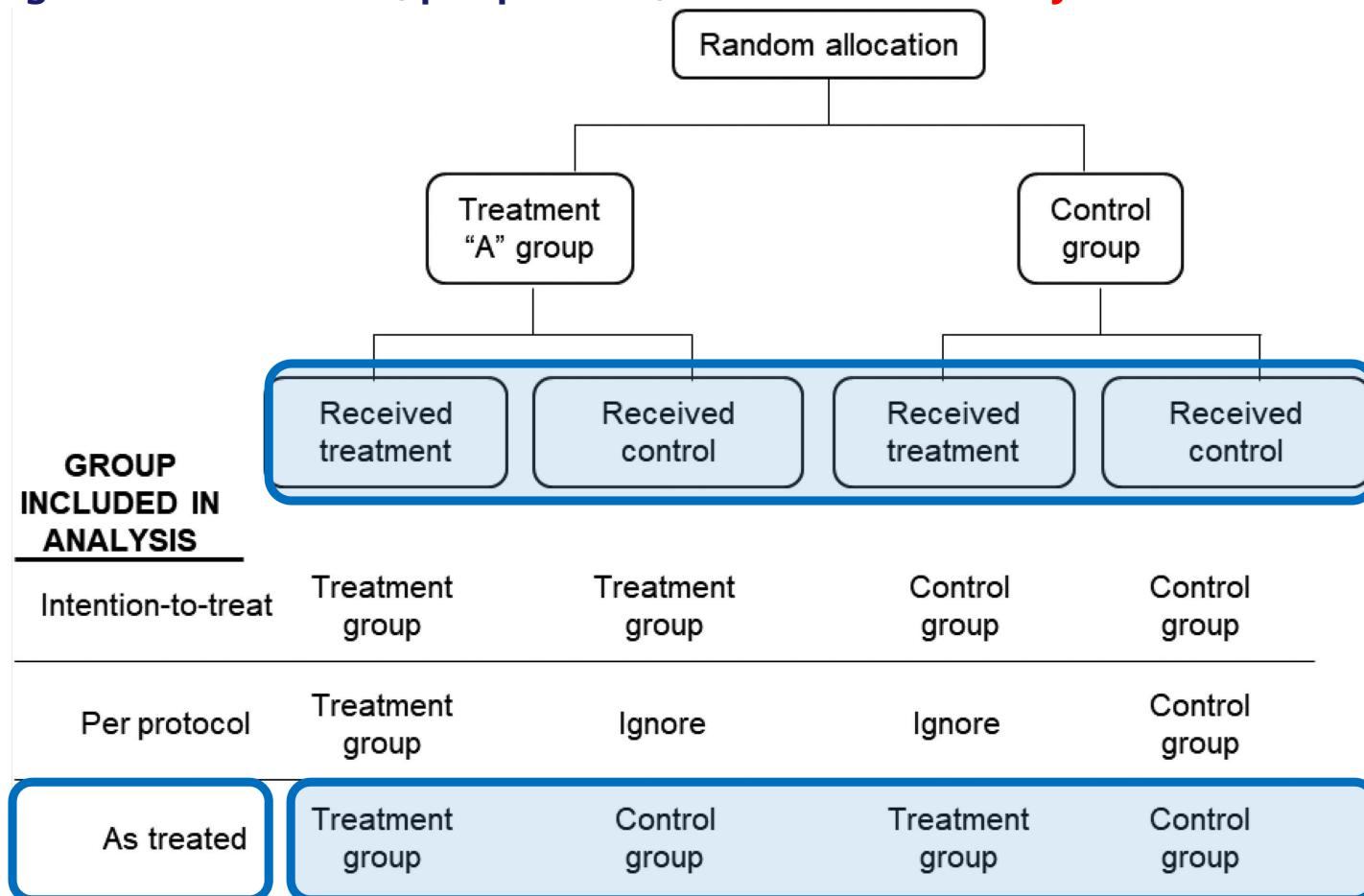
Modified from: Dettori JR, Norvell DC. Intention-to-Treat: Is That Fair? Global Spine Journal. 2020;10(3):361-363.  
doi:10.1177/2192568220903001

## Comparing intention-to-treat, per-protocol, and as-treated analyses.



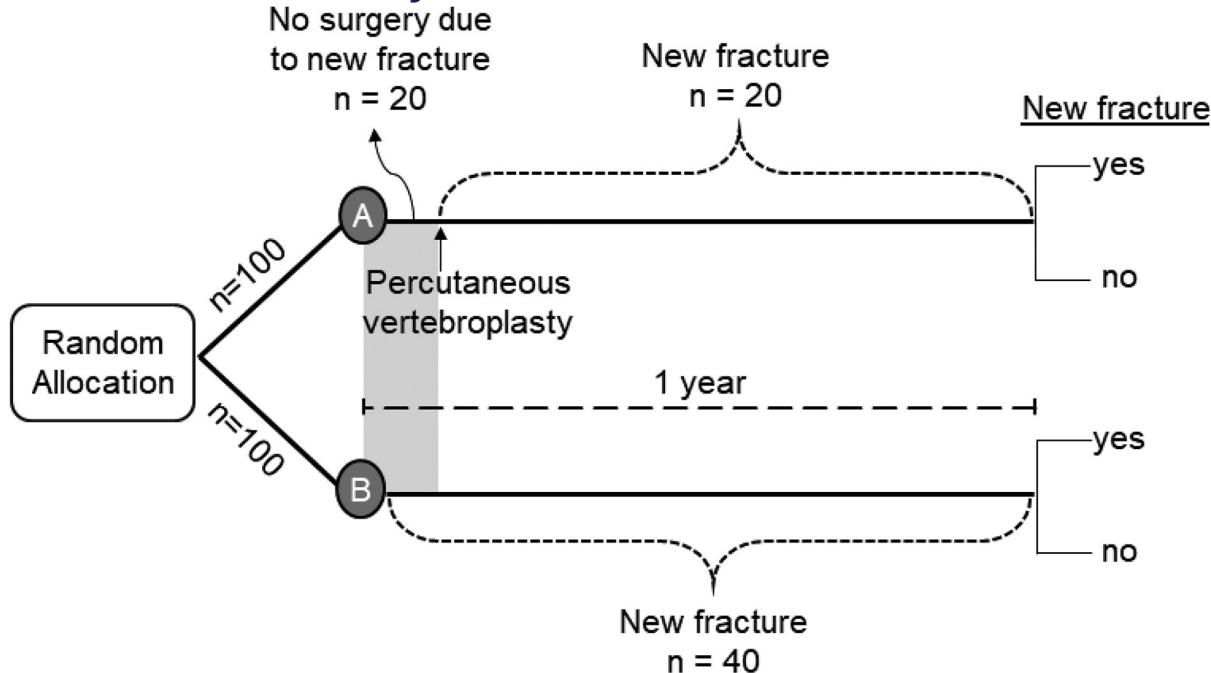
Modified from: Dettori JR, Norvell DC. Intention-to-Treat: Is That Fair? Global Spine Journal. 2020;10(3):361-363.  
doi:10.1177/2192568220903001

## Comparing intention-to-treat, per-protocol, and as-treated analyses.



Modified from: Dettori JR, Norvell DC. Intention-to-Treat: Is That Fair? Global Spine Journal. 2020;10(3):361-363.  
doi:10.1177/2192568220903001

## Schematic of an ITT and a PP analysis for RCTs.



Analysis	PV	Watchful Waiting	RR (95% CI)
<b>Per-protocol</b>	$20/80 = 25\%$	$40/100 = 40\%$	$0.63 (0.40, 0.98)$
<b>Intention-to-treat</b>	$40/100 = 40\%$	$40/100 = 40\%$	$1.00 (0.71, 1.40)$

Modified from: Dettori JR, Norvell DC. Intention-to-Treat: Is That Fair? Global Spine Journal. 2020;10(3):361-363.  
doi:10.1177/2192568220903001

# Other Types of Treatment Effects

- Average treatment effect (**ATE**)
- Average treatment effect on the treated (**ATT** or **ATET**)
- Average treatment effect on the controls (or untreated) (**ATC** or **ATU**)
- Average treatment effect among the evenly matchable (**ATM**), nearly equivalent to cohort formed by one-to-one pair matching
- Average treatment effect among the overlap population (**ATO**), estimates the treatment effect among those likely to have received either treatment or control

Sometimes you will see these prefixed with **P** for "population" (e.g., **PATT** = population ATT) or **S** for "sample" (e.g., **SATT** = sample ATT)

Sample estimates are interpreted conditional on the sample data

We will largely concentrate on **ATE** and **ATT**

# What are the differences between these TEs?

- **ATE** is  $E(Y^1 - Y^0)$  for *all* units (effect of *switching*) - **makes both groups look like the total sample**
- **ATT** is  $E(Y^1 - Y^0)$  for *treated* units (effect of *taking away* treatment) - **makes the controls look like the treatment**
- **ATC** is  $E(Y^1 - Y^0)$  for *untreated* units (effect of *adding* treatment) - **makes the treatment look like the controls**
- **ATM** estimates equivalent to cohort formed by one-to-one pair matching
- **ATO** estimates the treatment effect among those likely to have received either treatment or control

## Balancing weights for different TEs

Weights for participant  $i$  is defined here as  $e_i$  and the treatment assignment is  $T_i$ , where  $T=1$  indicates the participant received the treatment and  $T=0$  indicates they received the control → comparable pseudo populations for each treatment effect

$$w_{ATE} = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

$$w_{ATT} = \frac{e_i T_i}{e_i} + \frac{e_i (1 - T_i)}{1 - e_i}$$

$$w_{ATC} = \frac{(1 - e_i) T_i}{e_i} + \frac{(1 - e_i)(1 - T_i)}{1 - e_i}$$

$$w_{ATM} = \frac{\min\{e_i, 1 - e_i\}}{T_i e_i + (1 - T_i)(1 - e_i)}$$

$$w_{AT0} = (1 - e_i) T_i + e_i (1 - T_i)$$

# ATEs and ATTs

## Recall

**ATE:** Expected causal effect of the treatment for the whole population

$$E[\delta] = \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\ - \{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\}$$

**ATT:** Expected causal effect of the treatment for treated individuals

$$E[\delta|D=1] = E[Y^1 - Y^0|D=1] \\ = E[Y^1|D=1] - E[Y^0|D=1]$$

**For a RCT, ATE = ATT** because we assume

$$E[Y^1|D=1] = E[Y^1|D=0] \text{ and}$$

$$E[Y^0|D=1] = E[Y^0|D=0]$$

**In summary, a good RCT is a great way to estimate an ATE and an ATT**

**RCT primary outcomes measure** should have the following attributes:

- Clinically meaningful, capturing the main aspects of feeling/function/survival for the outcome.
- Penalize a treatment causing serious adverse outcomes (net clinical benefit)
- Avoids ties, continuous measures improve statistical power and lower sample size (i.e. is sensitive for detecting treatment effects)
- Measured over the relevant clinical time course
- Does not have its interpretation clouded by inter-current therapies or events
- Is easily interpretable to clinicians, regulators and patients
- Allows for simple and complete data capture while handling partially available data with minimal hidden assumptions

**However, there are few common statistical pitfalls in traditional RCTs**

**1. p-value > 0.05**

Commonly researchers will conclude that the treatment is ineffective -> usually wrong

Remember absence of evidence is not evidence of absence -> reality is the study is inconclusive (especially if the confidence interval for the treatment difference is wide).

**2. Powering study for a miracle**

In that scenario a clinically important, but not miraculous, effect is unfortunately often then interpreted as a conclusion of no effect.

**3. Inflexibility**

Design cannot be modified after randomization begins, otherwise one would not know how to compute a p-value as this requires repeated identical sampling.

## Questions to be asked if the RCT results are positive

- Does a P value of <0.05 provide strong enough evidence? (Lecture 1 slide 17 toss-up exp. with prior &  $p = 0.05 \rightarrow 71\% \text{ posterior}$ )
- What is the magnitude of the treatment benefit?
- Is the primary outcome clinically important?
- Are subgroups & secondary outcomes supportive?
- Is the trial large enough to be convincing?
- Was the trial stopped early?
- Do concerns about safety counterbalance positive efficacy?
- Are there flaws in trial design and conduct?
- Do the findings apply to my patients?

**Reference:** The Primary Outcome Is Positive — Is That Good Enough?- NEJM

## Questions to be asked if the RCT results are negative

- Is there some indication of potential benefit?
- Was the trial underpowered?
- Were the primary population, treatment, outcome appropriate
- Were there deficiencies in trial conduct or analyses?
- Is a claim of noninferiority of value?
- Do subgroup, 2<sup>o</sup> findings elicit positive signals?
- Does more positive external evidence or strong biologic rationale exist?

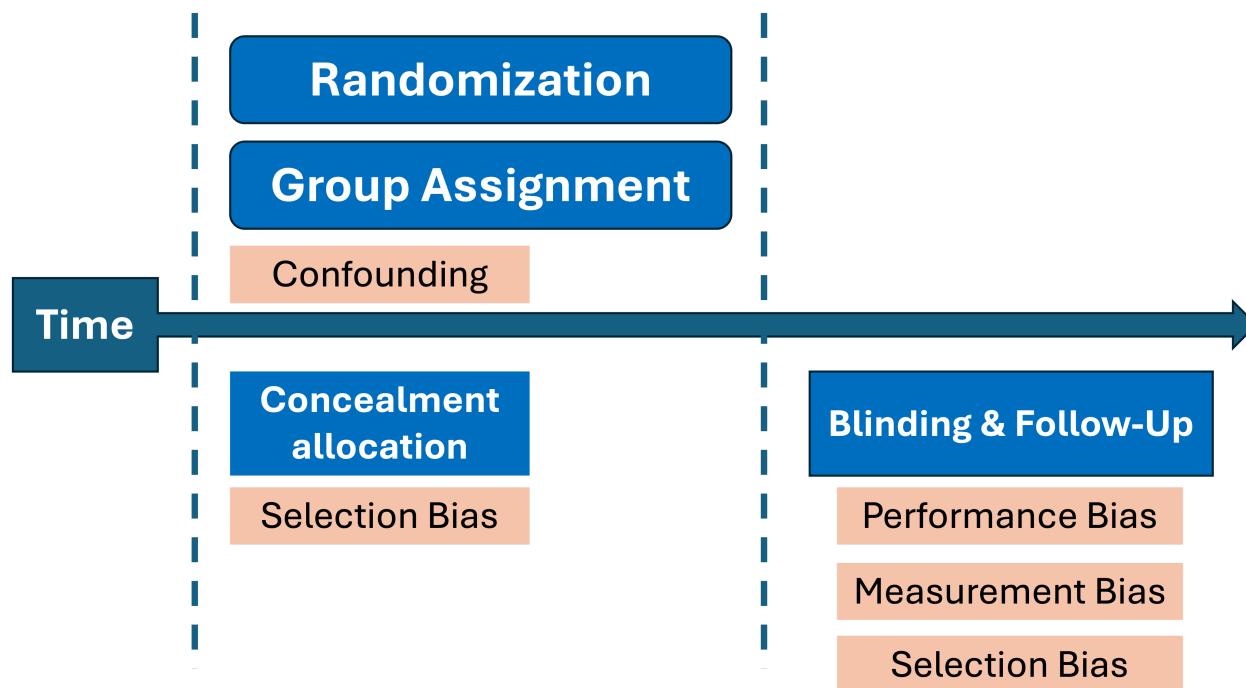
**Reference:** The Primary Outcome Fails — What Next?- NEJM

## Other *Deviations* from the planned RCTs:

The modified RCT can exclude participants who:

- Withdrew their consent
- Failed to receive any study drug
- Met the definition of a certain subgroup
- Dropped out because of toxicity of the study drug
- Were given the wrong treatment by the healthcare provider
- Failed to receive study drug long enough to have a measurable effect
- Violated certain aspects of the Clinical Study Protocol, e.g., taking prohibited drugs
- Determination of not meeting inclusion or exclusion criteria, after enrollment
- Had other reasons for **differential adherence or loss to follow-up.**

# Issues with *deviations* of planned RCTs



We *may* encounter bias despite the initial randomization

## What are those biases? Don't be lost in Translation:

Translation of Cochrane bias in randomized trials domains into common epidemiologic terms.

Cochrane bias domain	Epidemiologic term	Bias in intention-to-treat effect	Bias in per-protocol effect
Selection bias	Confounding or selection bias	Yes	Yes
Performance bias	Biased direct effect or confounding	No	Yes
Detection bias	Measurement bias	Yes	Yes
Attrition bias	Selection bias	Yes	Yes
Reporting bias	Non-structural bias that cannot be represented in our causal diagrams	Yes	Yes

Modified from: Mansournia, M. A., et al. (2017). Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. *Epidemiology* (Cambridge, Mass.), 28(1), 54–59. <https://doi.org/10.1097/EDE.0000000000000564>

## Other potential RCT limitations

1. Expensive to perform.
2. May be ethically challenging.
3. Difficult to recruit (both investigators & patients).
4. Findings are too broad (average treatment effect not representative or benefit for any given individual).
5. Findings may lack external generalizability (trial population and setting not representative of general practice).
6. Often reported and interpreted in isolation from other pertinent studies.
7. (Often long delays before RCT results diffuse into practice).

## Key messages

- No study design is flawless.
- Elevating non feasible RCTs at the expense of other research designs can be counter productive, especially in public health
- Recognize the strengths and limitations in all data sources and designs to obtain the most useful and valid data
- Health decision making often optimized by considering all data sources from well performed experimental and non-experimental sources
- If non-experimental designs are chosen, try to emulate <sup>1</sup> a RCT to minimize bias

<sup>1</sup> Lecture on Target Trials on Nov 14, 2024.

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

## Extra resources

- Twisk J, et al. Different ways to estimate treatment effects in randomised controlled trials. Contemporary clinical trials communications, 10, 80–85.  
<https://doi.org/10.1016/j.conctc.2018.03.008>
- Cole SR, Edwards JK, Zivich PN, Shook-Sa BE, Hudgens MG, Stringer JS. Reducing Bias in Estimates of Per Protocol Treatment Effects: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Netw Open*. 2023;6(7):e2325907. doi:10.1001/jamanetworkopen.2023.25907
- Morris, T.P., Walker, A.S., Williamson, E.J. et al. Planning a method for covariate adjustment in individually randomised trials: a practical guide. *Trials* 23, 328 (2022).  
<https://doi.org/10.1186/s13063-022-06097-z>
- Harrer, M., Cuijpers, P., Schuurmans, L.K.J., Kaiser, T., Buntrock, C., van Straten, A. & Ebert, D. Evaluation of randomized controlled trials: a primer and tutorial for mental health researchers. *Trials* 24, 562 (2023). doi: 10.1186/s13063-023-07596-3.

## Designed versus actual power

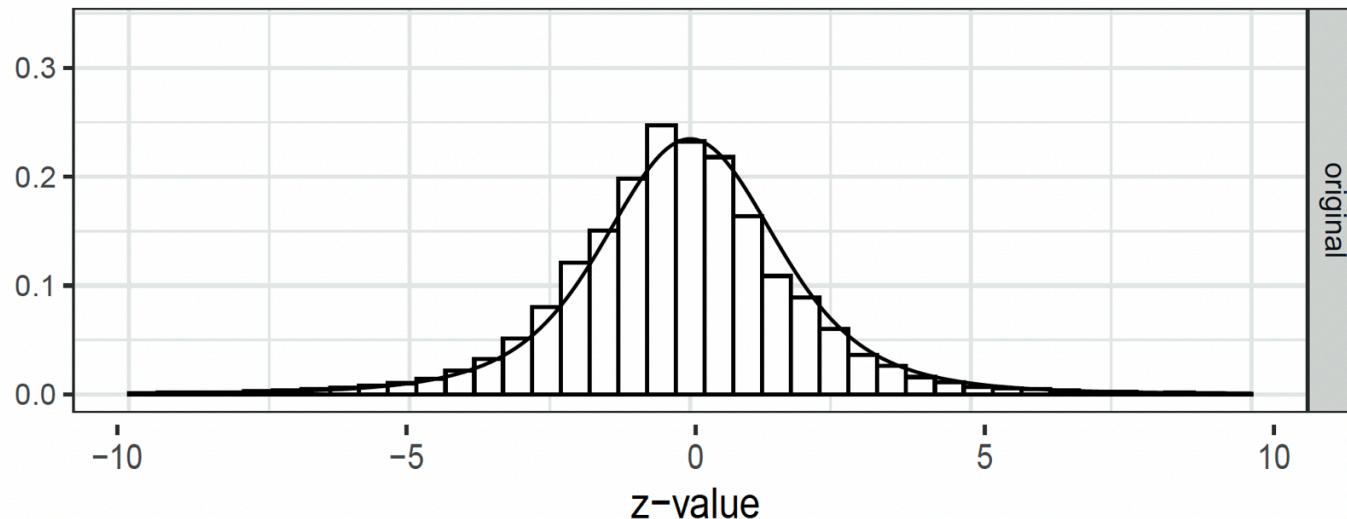
Ensure that there is at least 80% or 90% probability or “power” ( $1 - \beta$ ) for “detecting” an effect at the  $\alpha = 0.05$  level under the assumption that the effect of the treatment is a particular size.

While a trial may be designed to have 80% power to detect a particular effect of clinical relevance, that does not mean it has 80% probability of yielding  $p < 0.05$ : the latter probability depends on the **true** effect of the treatment, so we refer to it as the **actual power**

# Actual power

Actual power can't be observed, but can estimate its standardized distribution

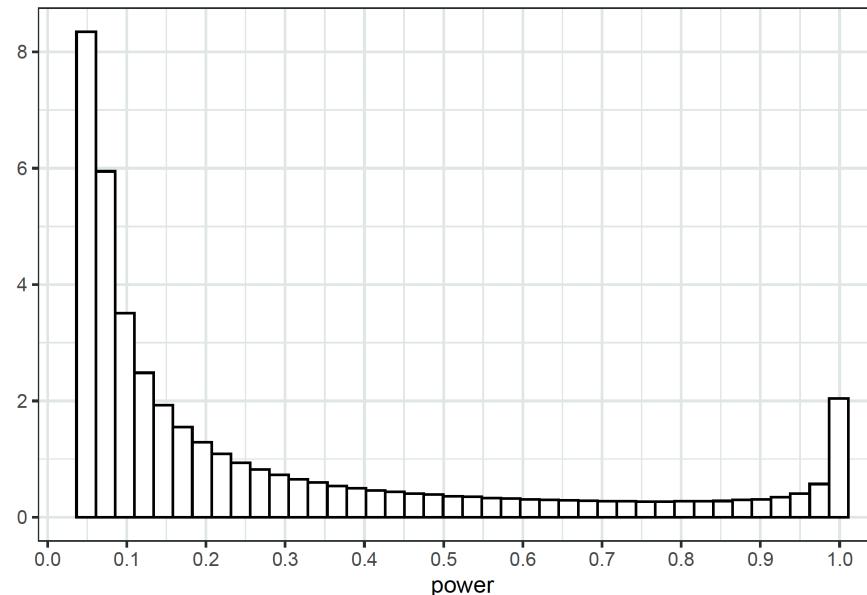
Based on results of 23,551 RCTs of treatment efficacy that were extracted from the Cochrane database



Notice this is **not** Gaussian as have heavy tails ( $7\% > z=4$ ), raising some concerns about possible bias in these studies with extreme results

# Actual power

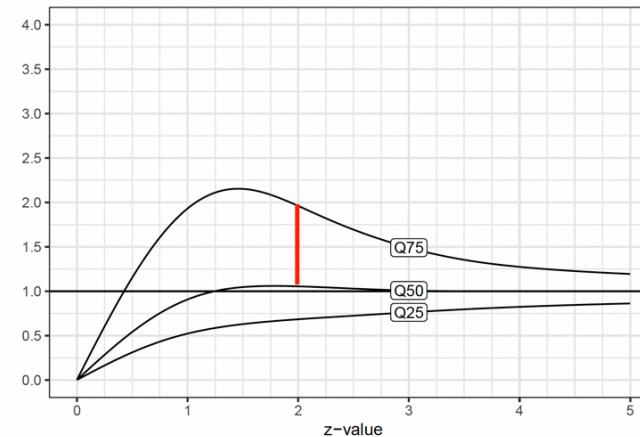
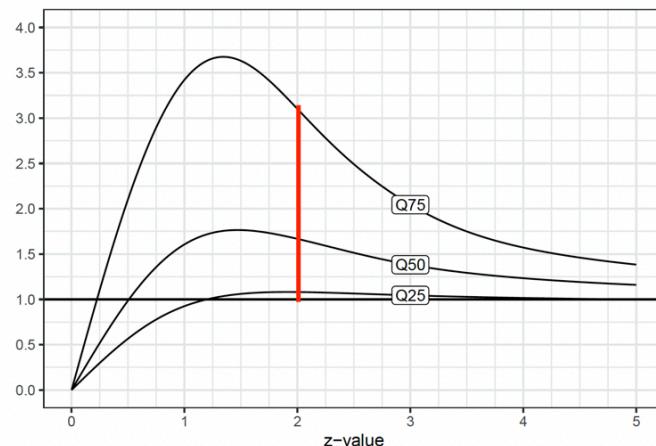
From this observed distribution can estimate the actual power in these 23,000 RCTs



- Around 9 out of 10 RCTs have actual power less than 80%
- The mean actual power is just 28% , median actual power just 13%
- Flip side if despite low power RCT +, **estimated effect size must be exaggerated** – unlikely to be replicated in later studies - -> winner's curse
  - IOW, most RCTs are radically underpowered to detect the true effect

# Dealing with exaggerated RCT effect sizes

van Zwet proposed to counter the resulting exaggeration of effect estimates by using shrinkage estimators, essentially a regularizing prior using **exchangeability with RCTs in Cochrane database**



Full Bayesian combines single trial with study specific "prior" capturing unique insights & experience, rather than assuming the RCT resembles the "average" Cochrane study

# Standard power analysis in R

In R, power analysis can use

- `library(pwr)` (vignettes can be found [here](#))
- `library(epiDisplay)`
- `power.prop.test` for proportions in base R

If we have any of the three parameters given below, we can calculate the fourth one

- Sample size
- Effect size
- Significance level
- Power of the test

# Power analysis in R - example

Suppose you have a trial with 500 patients in each group & mortality proportions = 0.3 & 0.4  
What is the power for a Type 1 error ( $\alpha$ ) = 0.05?

```
# method 1
library(pwr)
pwr.2p.test(h=ES.h(p1 = 0.4, p2 = 0.3), sig.level = 0.05)

## 
## Difference of proportion power calculation for two-sample proportions
## 
## h = 0.2101589
## n = 500
## sig.level = 0.05
## power = 0.9135494
## alternative = two.sided
## 
## NOTE: same sample sizes
```

```
# method 2
epiDisplay::power.for.2p(p1=.4, p2=.3, n1=500, n2=500, alpha=0.05)

## 
## Power for comparison of 2 proportions.
## for binomial distribution (arcsine transformation)
## 
## p2 = 0.3
## n1 = 500
## n2 = 500
## alpha = 0.05
## power = 0.902
```

Yet a third way is with `?power.prop.test`

# Choice of outcome and power

Outcome Type	Statistical Efficiency
binary	Minimum (assumes time is not important)
time to first binary outcome	high if event is very frequent
continuous response (e.g. blood pressure)	maximum power among univariate outcomes
ordinal response measured at a single time from randomization	high if at least 4 well-populated categories
longitudinal ordinal responses measured e.g. daily or weekly	very high if at least 4 well-populated categories
longitudinal continuous response	highest