Interpreting Results from Clinical Research

Dr Suman Kumar, @sumankumarpram1

04 Jul 2017

Background

Around 80% of results of all published research are “FALSE” (NON REPLICABLE)

Huge problem of results not holding good on replication

- Wastage of resources: patients, time, money
- Wrong, sometimes fatal clinical decisions **(Ethical issues)**

Reasons for the “FALSE” results

- Human emotion

Reasons for the “FALSE” results

- Human emotion
- Bias by researcher: Hiding data, analyses and results

Reasons for the “FALSE” results

- Human emotion
- Bias by researcher: Hiding data, analyses and results
- Bias by editorial staff: Preferential publication of positive results

Reasons for the “FALSE” results

- Human emotion
- Bias by researcher: Hiding data, analyses and results
- Bias by editorial staff: Preferential publication of positive results
- Positive results are linked with professional growth

Reasons for the “FALSE” results

- Human emotion
- Bias by researcher: Hiding data, analyses and results
- Bias by editorial staff: Preferential publication of positive results
- Positive results are linked with professional growth

- Lack of replication of studies

Reasons for the “FALSE” results

- Human emotion
- Bias by researcher: Hiding data, analyses and results
- Bias by editorial staff: Preferential publication of positive results
- Positive results are linked with professional growth

- Lack of replication of studies

- **Mis-interpretation of statistical results**

Reasons for the “FALSE” results

- Human emotion
- Bias by researcher: Hiding data, analyses and results
- Bias by editorial staff: Preferential publication of positive results
- Positive results are linked with professional growth

- Lack of replication of studies

- **Mis-interpretation of statistical results**
- Commonest one: *P VALUE*

Actual meaning of “p value”: Complicated Concept

“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001”

- If we **assume** that there is no difference between reduction in blood sugar levels between A and B **(both are equal)**
- **Chances** that drug A and B are really equivalent, given the sample difference of $>$ $+$/- 2 mg/dl is 0.1%.

- Is the difference really significant? **Depends on us**

p value $<$ 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 (< 0.05)”**

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 (< 0.05)”**
- *Drug B is significantly better than drug A*

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 (< 0.05)”**
- *Drug B is significantly better than drug A*

- **“$p = 0.056$”**

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 (< 0.05)”**
- *Drug B is significantly better than drug A*

- **“$p = 0.056$”**
- *Drug B is not better than drug A and researcher does his best to reduce p value to less than 0.05*

p value $<$ 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 ($<$ 0.05)”**
- *Drug B is significantly better than drug A*

- **“p $=$ 0.056”**
- *Drug B is not better than drug A and researcher does his best to reduce p value to less than 0.05*

- **“p $=$ 0.01 is better discriminator than p $=$ 0.046”**

p value < 0.05: THE MAGICAL EXPRESSION

- DISCRIMINATOR FOR ASSESSING SIGNIFICANCE: The ill defined and often non relevant statistical significance.

- **“p value of difference in mean in reduction in fasting blood sugar levels between drug A (mean 34 mg/dl) and drug B (mean 36 mg/dl) is 0.001 (< 0.05)”**
- *Drug B is significantly better than drug A*

- **“p = 0.056”**
- *Drug B is not better than drug A and researcher does his best to reduce p value to less than 0.05*

- **“p = 0.01 is better discriminator than p = 0.046”**
- *Lesser the p value, better it is as discriminator*

Aims of this presentation

- Provide alternatives to p value for interpreting clinical research
- Provide more informative ways to interpret results from research

Trial characteristic to be discussed in presentation

- Comparative intervention trial
- Intervention A vs Intervention B
- Outcome of interest: **proportion of developing a given outcome** within a **period of time**
- Our aim is to **compare** Intervention A and Intervention B
  - Difference in proportion *(Risk difference)*
  - Ratio of proportion *(Risk ratio)*
  - Ratio of odds *(Odds ratio)*

Q 1: Comparability of populations

Is population being tested in trial comparable to our population?

- Patient characteristics (Host, Disease, Co-morbidities, Demography)
- Environment around patients (in hospital and around the place of living)
- Equality of Supportive care
- Similarity in proficiency of measurement of variables and outcomes
- Similarity in proficiency of administering intervention

Q 2: Understanding Effect Size (Outcome measure)

Effect Size

- **Most important** number we should understand
- Population characteristic
  - Usually we can only estimate it from the sample
- **One population**
  - Mean/median of WBC, serum cholesterol, BP, HbA1C levels
  - Proportion surviving at the end of 1 year (OS)
  - Incidence rate (Hazard) of relapse over 1 year
  - Cumulative incidence of relapse over 1 year
- **Two populations** (comparision)
  - **Difference (Absolute and relative)**
  - Ratio (Hazard ratio, Odds ratio, Risk ratio)

Example (Difference in proportions)

Example 1

Intervention A (standard of care) and intervention B are given over a period of **1 month**. At the end of **1 year**, 50% of patients in intervention A and 60% of patients in intervention B arm are in remission.

Example 2

Intervention A (standard of care) and intervention B are given over a period of **1 year**. At the end of **5 years**, 2% of patients in intervention A and 1% of patients in intervention B arm relapse.

> *Is Intervention B better than intervention A (standard of care)?. We will use difference in proportion as our* ***Effect Size Measure****.*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 $= 0.2 =$ *20%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
- *Example 1:* 0.6 - 0.5 = 0.1 = *10%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
- *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
- *Example 2:* 0.02 - 0.01 = 0.01 = *1%*

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
- *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
- *Example 2:* 0.02 - 0.01 = 0.01 = *1%*

- Usually, **RRD is presented in literature** rather than ARD

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
- *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
- *Example 2:* 0.02 - 0.01 = 0.01 = *1%*

- Usually, **RRD is presented in literature** rather than ARD
- Inflates the effect size, especially when risks are nearer to zero (Example 2)

Relative vs absolute difference in proportion

- **Relative risk difference (RRD)**
- *Example 1:* (0.6 - 0.5)/0.5 = 0.2 = *20%*
- *Example 2:* (0.02 - 0.01)/0.02 = 0.5 = *50%*

- **Absolute risk difference (ARD)**
- *Example 1:* 0.6 - 0.5 = 0.1 = *10%*
- *Example 2:* 0.02 - 0.01 = 0.01 = *1%*

- Usually, **RRD is presented in literature** rather than ARD
- Inflates the effect size, especially when risks are nearer to zero (Example 2)
- For example 2, by using intervention B, there is only 1% decrease in relapse (in absolute term), but 50% reduction in relapse, when compared to intervention A

Clinically relevant effect size

- Needs to be defined by user
- Requires thorough knowledge of subject area and expertise
- Example 1: Say, the disease concern is an indolent and non life threatening disease. Improvement of remission rate by 10% may not be clinically relevant

ARD and NNT

- **Number Needed to Treat (NNT) = 1/ARD**
- Very useful effect size measure
- **Example 1:** $NNT = 10$
- **Example 2:** $NNT = 100$

- We need to treat *10 patients* to get 1 extra remission at the end of *1 year* (Example 1) and *100 patients* to prevent 1 extra relapse at the end of *5 years* (Example 2).

ARD and NNT

- **Number Needed to Treat (NNT) = 1/ARD**
- Very useful effect size measure
- **Example 1:** $NNT = 10$
- **Example 2:** $NNT = 100$

- We need to treat *10 patients* to get 1 extra remission at the end of *1 year* (Example 1) and *100 patients* to prevent 1 extra relapse at the end of *5 years* (Example 2).
- If I get 10 patients of the disease in Example 2 in my centre in a year, I will have to wait for 10 years to get one less relapse after waiting for 5 years (i.e., from 6th to 16th year)

ARD and NNT

- **Number Needed to Treat (NNT) = 1/ARD**
- Very useful effect size measure
- **Example 1:** $NNT = 10$
- **Example 2:** $NNT = 100$

- We need to treat *10 patients* to get 1 extra remission at the end of *1 year* (Example 1) and *100 patients* to prevent 1 extra relapse at the end of *5 years* (Example 2).
- If I get 10 patients of the disease in Example 2 in my centre in a year, I will have to wait for 10 years to get one less relapse after waiting for 5 years (i.e., from 6^{th} to 16^{th} year)
- Is intervention B really better for me at my centre??

Clinically relevant effect size (Surrogate Effect Size)

- Clinically relevant effect sizes are **Patient oriented**
  - Mortality, Morbidity, Quality of Life
  - **Adverse effects** attributable to the intervention
- **Surrogate markers** for Clinically relevant effect sizes
  - BP, Cholesterol $\Rightarrow$ CAD $\Rightarrow$ CAD associated deaths
  - Blood HbA1C levels $\Rightarrow$ Diabetic complications $\Rightarrow$ Diabetes associated deaths
  - Prevalence of CIN $\Rightarrow$ Prevalence of Cervical Cancer $\Rightarrow$ Cancer associated deaths
  - Major molecular response on CML $\Rightarrow$ CML associated deaths

Clinically relevant effect size (Surrogate Effect Size)

- **Questionable quality of surrogate markers** to extrapolate clinically relevant effect size
- Why surrogate markers are reported?
  - Assessing them takes less time and less resources
  - Researchers want to conceal the fact that the benefit of the drug is not clinically relevant

Q 3: Estimating Effect Size

Population vs Sample

- We donot know the real Effect Size as it is a population characteristic
- We can only estimate it from **Random Sample** chosen from the underlying population by **carrying out experiments**

Q 4: Quality of Effect Size Estimate

Three qualities

- **Validity** of estimate
  - Difference in average of sample estimates and actual effect size (Bias)
- **Magnitude** of estimate
  - Greater the magnitude in case of differences, we are surer of the real difference.
- **Precision** of estimate (denoted by Confidence Interval)
  - Greater the precision, we are surer of value of population effect size

Q 4a: Validity of effect size estimate (Problem of CONFOUNDERS)

What are confounders?

- Outcome is related to complex network of inter-related variables (known and unknown)
- Our job is to assess Exposure ⇒ Outcome effect size (SAMPLE EFFECT SIZE ESTIMATE)
- **CONFOUNDERS**
  - ASSOCIATED WITH OUTCOMES
  - UNEQUALLY DISTRIBUTED BETWEEN INTERVENTIONS
- Confounders change Exposure ⇒ Outcome effect size
- Creates BIAS

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
- If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
- If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?

- **Baseline unknown confounders** are equal between both groups

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
- If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?

- **Baseline unknown confounders** are equal between both groups
- Only way to take care of is by **appropriate randomisation** (randomised allocation of treatment ensures confounders to be distributed equally in both the groups)

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
- If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?

- **Baseline unknown confounders** are equal between both groups
- Only way to take care of is by **appropriate randomisation** (randomised allocation of treatment ensures confounders to be distributed equally in both the groups)

- **Blinding** of allocation of intervention arms, taking care of patients, measuring outcomes, performing statistical analyses

Are groups intervention A and intervention B equal in all respects other than the interventions?

- **Baseline known confounders** are equal between both groups
- If not, are they taken care of statistically (Multiple regression analysis, stratified analysis)?

- **Baseline unknown confounders** are equal between both groups
- Only way to take care of is by **appropriate randomisation** (randomised allocation of treatment ensures confounders to be distributed equally in both the groups)

- **Blinding** of allocation of intervention arms, taking care of patients, measuring outcomes, performing statistical analyses
- To maintain equality among both the groups till publishing the results

(contd . . .)

- Equality of **loss to follow up or cross over** between both groups: numbers and reasons

(contd . . .)

- Equality of **loss to follow up or cross over** between both groups: numbers and reasons
- RCTs yield more valid estimate of Effect Size than observational studies (Cohort, Case Control studies)

Cross trial comparisons

- Trial 1: Drug A - remission rate 30%, Drug B - remission rate 40% (Drug B $>$ Drug A)
- Trial 2: Drug A - remission rate 30%, Drug C - remission rate 40% (Drug C $>$ Drug A)
- **Can we infer that Drug B $=$ Drug C?**

Cross trial comparisons

- **Dangerous to compare drugs across trials**
- Distribution of a poor prognostic factor

Trials	Drug A	Drug B	Drug C
Trial 1	30%	30%	-
Trial 2	70%	-	70%

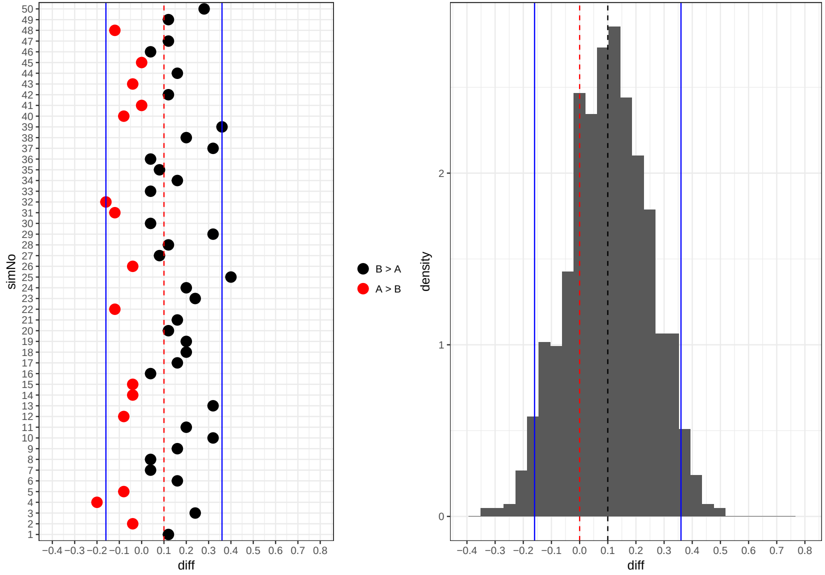
- **Intra-trial** poor prognostic factor (confounder) is equally distributed among treatment arms **(due to randomisation)**
- **Inter-trial:** Drug C (Trial 2) is given to patients with poorer prognosis than Drug B (Trial 1)
- **Bias is created when we are comparing drugs cross trials**

Q 4c: Understanding Precision

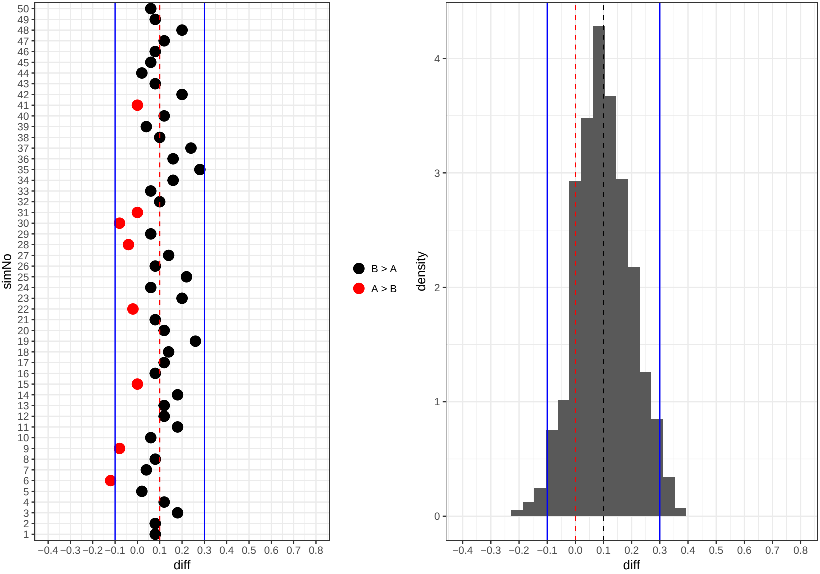
Simulation

We simulate example 1. Clinically relevant difference between both groups is 0.1. We will draw random samples from population treated with intervention A (prob of remission 0.5) and population treated with intervention B (prob of remission 0.6) 1000 times (equivalent as carrying out 1000 trials). We will compare intervention A and intervention B by difference of proportion.

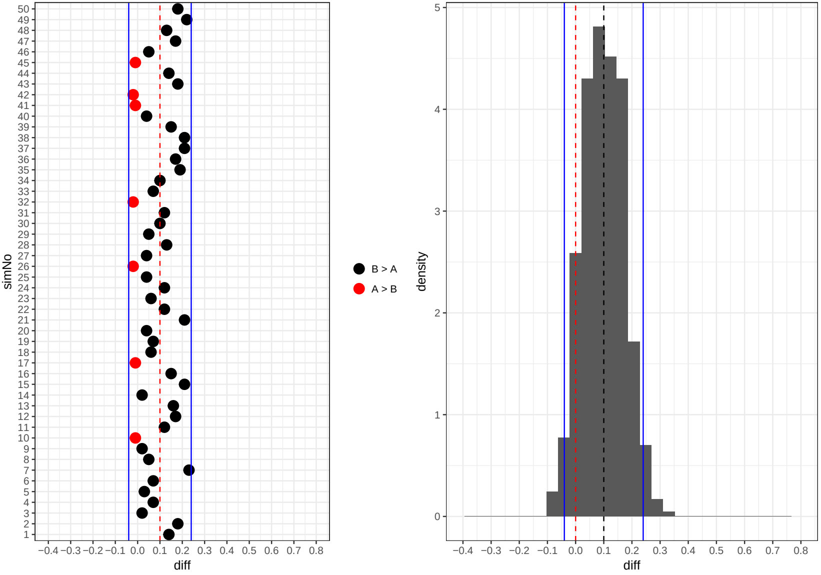
Simulation: probA: 50%, probB: 60%, sample size: 25



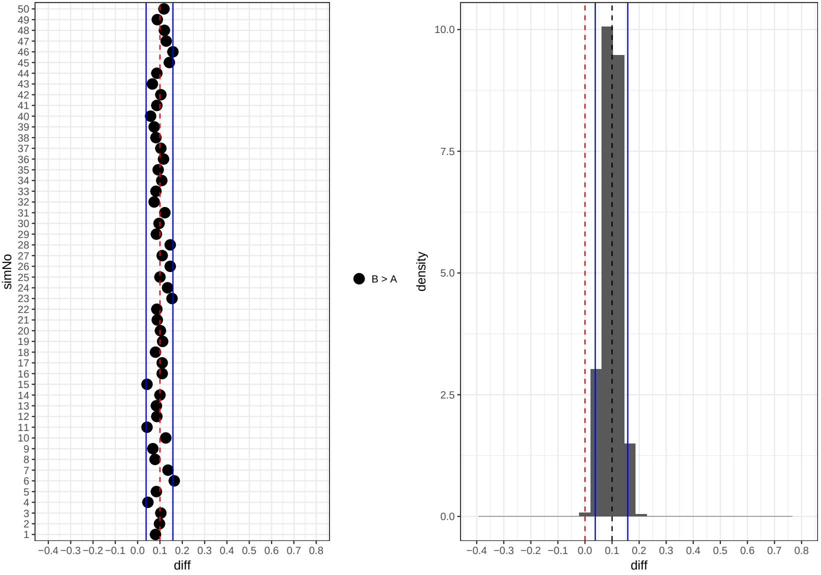
Simulation: probA: 50%, probB: 60%, sample size: 50



Simulation: probA: 50%, probB: 60%, sample size: 100



Simulation: probA: 50%, probB: 60%, sample size: 500



Explanation of simulation

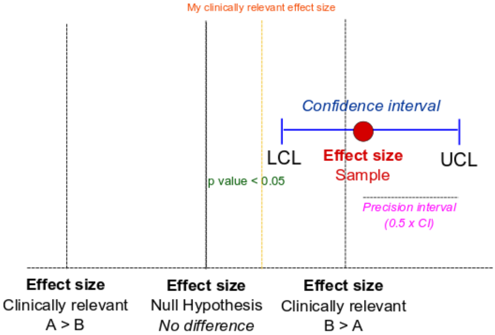
- The **blue** lines, which denote the **bound for mid 95% of all the estimates** is the measure of **Precision**, the width of which is the width of corresponding confidence interval

Explanation of simulation

- The **blue** lines, which denote the **bound for mid 95% of all the estimates** is the measure of **Precision**, the width of which is the width of corresponding confidence interval

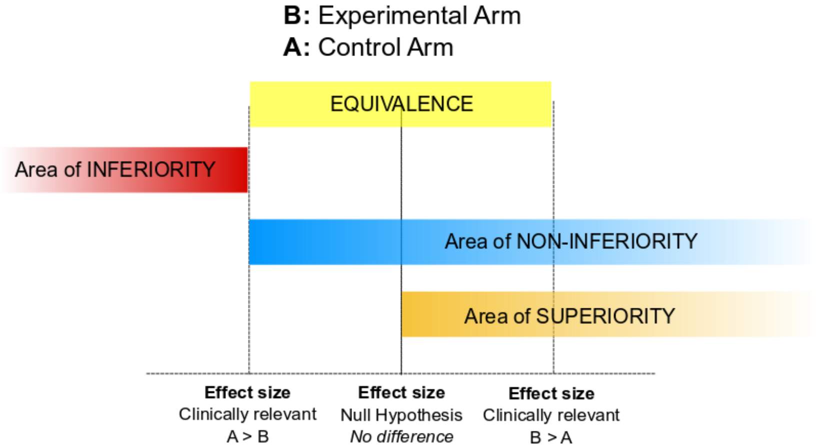
- The precision increases (width of the distribution decreases) with increasing the sample size

Understanding Confidence Interval



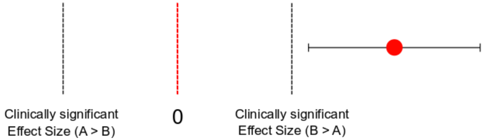
- Width of LCL - UCL, dependent on **variability** of sample and **sample size**
- Any of the points bounded by LCL and UCL can be the **Population Effect Size** (with 95% certainty)

Understanding clinically relevant regions



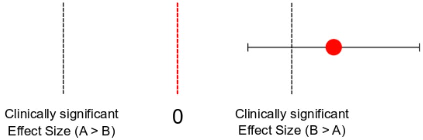
Interpreting effect size estimate and CI

Scenario 1



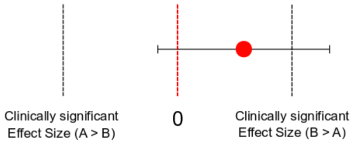
- The population effect size is more than 0 and clinically significant effect size (Intervention B is definitely clinically better to Intervention A)

Scenario 2



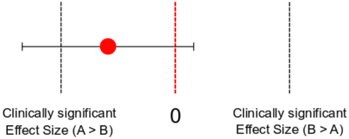
- The population effect size is more than 0 and but may not be more than clinically significant effect size (Intervention B is better than intervention A but may not be clinically relevant).

Scenario 3



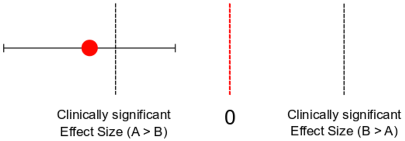
- The population effect size crosses 0, **we cannot say that B is better than A**. We can say that B is **not inferior** to A. We should not say that B is not better than A, we need to be **more precise.**
- Absence of evidence that a fact is true does not mean that fact is not true.

Scenario 4



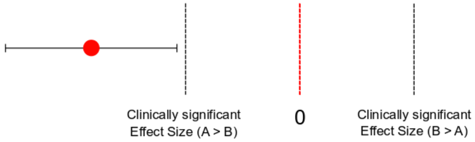
- We are not sure that B is not inferior to A. We are sure that B is not better than A in a clinically relevant manner.

Scenario 5



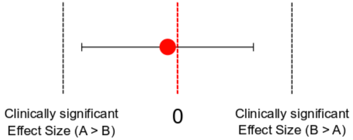
- We are sure that B is not better than A

Scenario 6



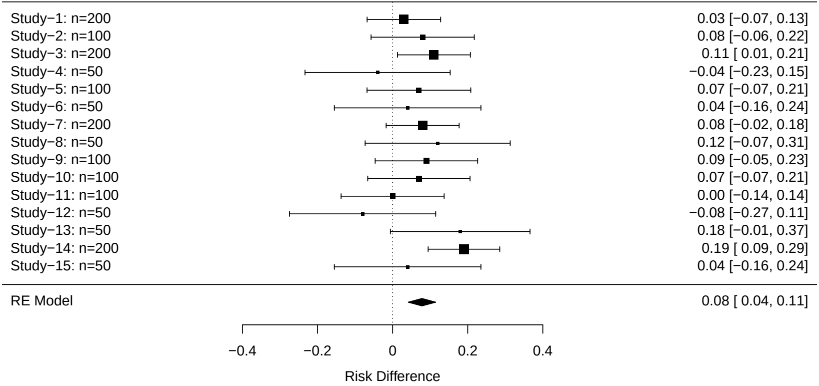
- We are sure that B is inferior to A

Scenario 7



- B is equivalent in effect to A

Finally, Importance of Replication of Experiments (Meta-analysis)



- We are **more certain about the population effect size.** Miniscule confidence interval
- Interpretation of effect size depends on us.

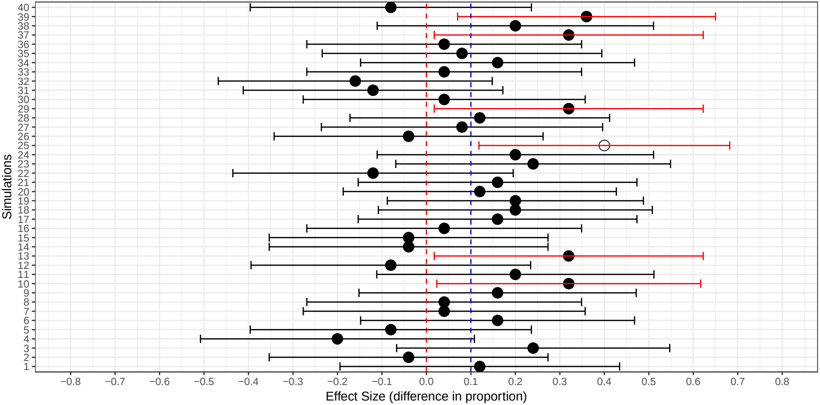
To summarise, interpretation of study results means

- Assessing similarity of population depicted in study with ours
- Understanding relevant effect size
- Be careful of surrogate outcome measures and cross trial comparisons
- Ascertaining equality of groups A and B (Tackling Bias)
- Assessing position and precision of effect size estimate

THANK YOU

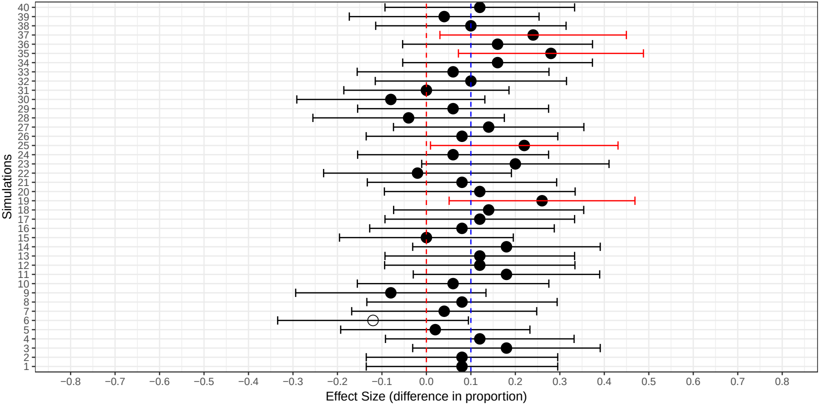
Manipulating CI

Simulation with CI: probA: 50%, probB: 60%, sample size: 25



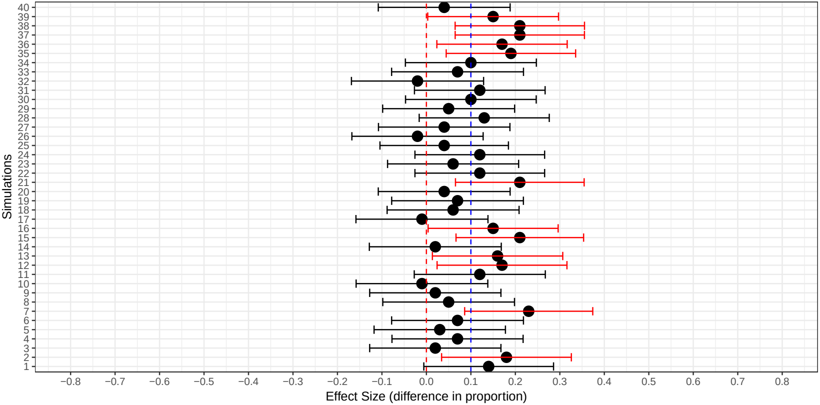
- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.023
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.918

Simulation with CI: probA: 50%, probB: 60%, sample size: 50



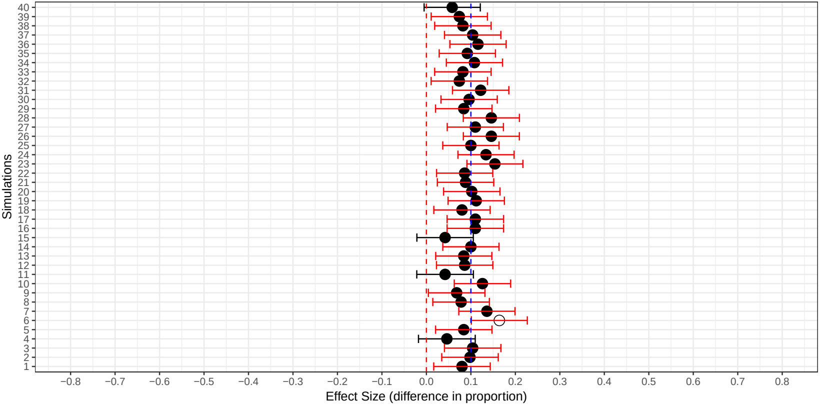
- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.034
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.857

Simulation with CI: probA: 50%, probB: 60%, sample size: 100



- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.043
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.713

Simulation with CI: probA: 50%, probB: 60%, sample size: 500

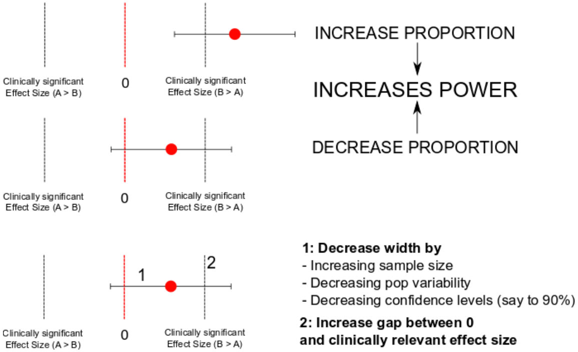


- Proportion of experiments failing to include population effect size in CI (Alpha Error): 0.034
- Proportion of experiments failing to show difference between both groups (Beta Error): 0.128

We learnt about . . .

- **Alpha error:** Proportion of times when CI **fail** to include the **population effect size**
    - Usual value: 0.05
- **Beta error:** Proportion of times when CI include **effect size of null hypothesis (0)**
    - Usual value: 0.20
- **Power** of study *(1 - Beta error)*: Proportion of times when CI **do not** include **effect size of null hypothesis (0)**

Steps to increase power of study


Clinically significant Effect Size (A > B)
0
Clinically significant Effect Size (B > A)
Clinically significant Effect Size (A > B)
0
Clinically significant Effect Size (B > A)
1
2
Clinically significant Effect Size (A > B)
0
Clinically significant Effect Size (B > A)

INCREASE PROPORTION

INCREASES POWER

DECREASE PROPORTION

1: Decrease width by
- Increasing sample size
- Decreasing pop variability
- Decreasing confidence levels (say to 90%)

2: Increase gap between 0 and clinically relevant effect size