Some issues about design and statistical analysis of preclinical trials

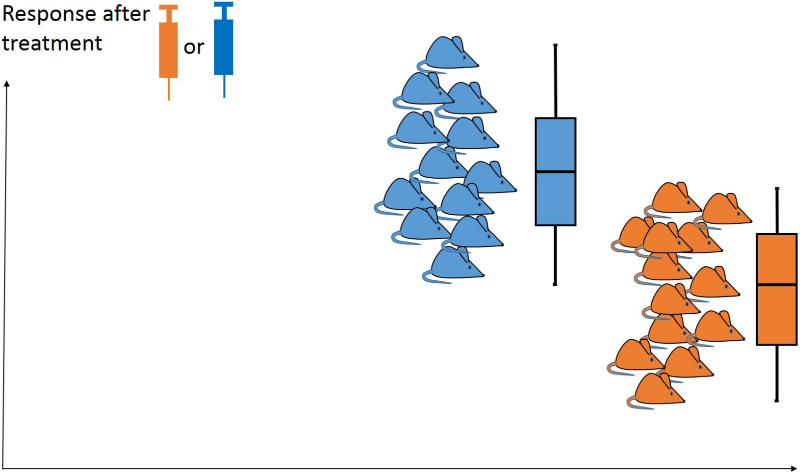
Marie Laure Delignette-Muller

15 janvier, 2021

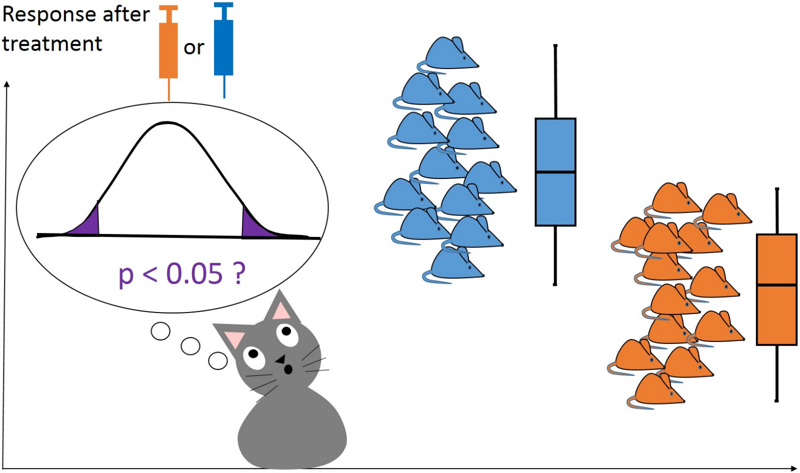
Context and definitions

The general objective of preclinical trials

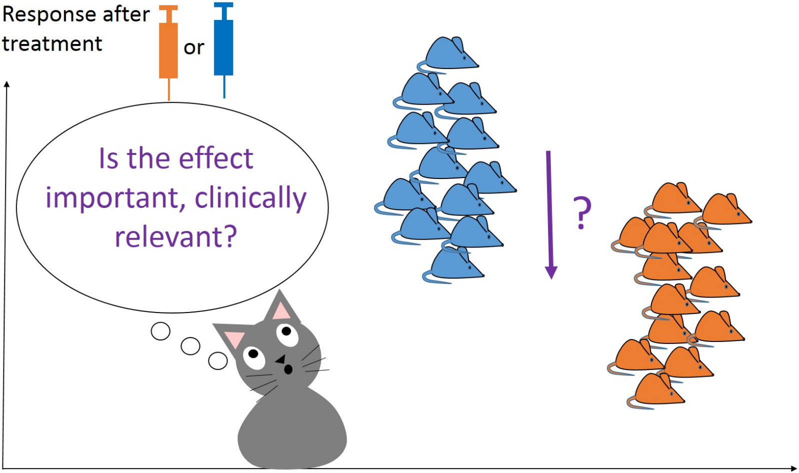
Evaluate the effet of a drug, a procedure, or another medical treatment in animals, in comparison to a placebo or a reference procedure or treatment.



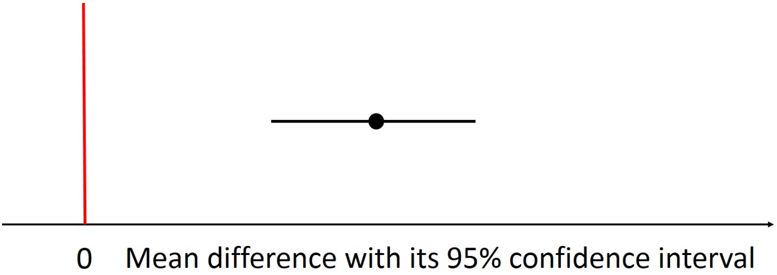
Is the effect significant ? (a preliminar question)



But the main issue is the characterization of this effect



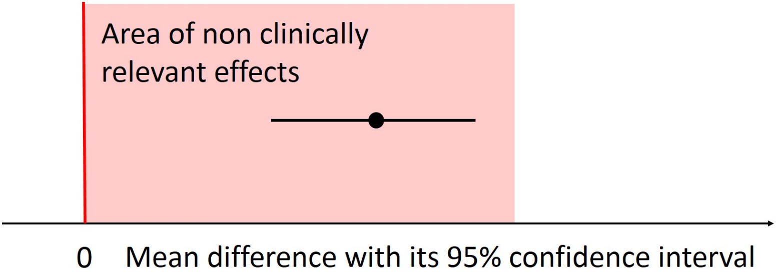
Estimation of the effect



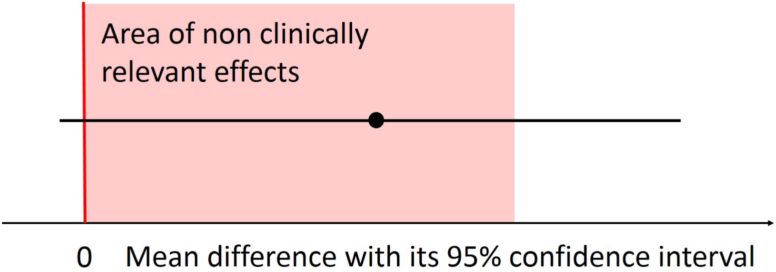
- Is the effect significant ?
- Is the effect clinically relevant (of biological interest)?

As soon as its 95% confidence interval does not contain 0 ($\Leftrightarrow$ p-value < 0.05) a difference is said significant.
But a significant difference is not necessarily clinically relevant.

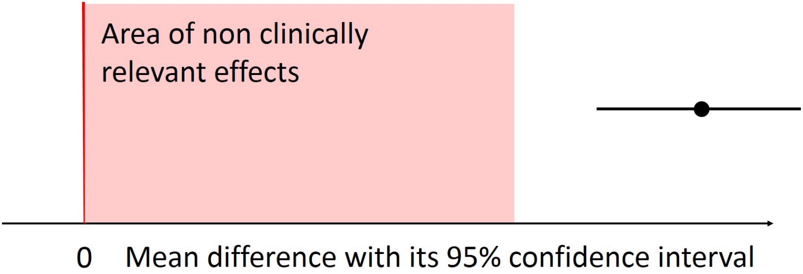
Case of a significant but not clinically relevant effect


Area of non clinically relevant effects
0
Mean difference with its 95% confidence interval

Case of a non significant difference but for which a clinically relevant effect cannot be excluded


Area of non clinically relevant effects
0
Mean difference with its 95% confidence interval

Expected case of a significant and clinically relevant effect

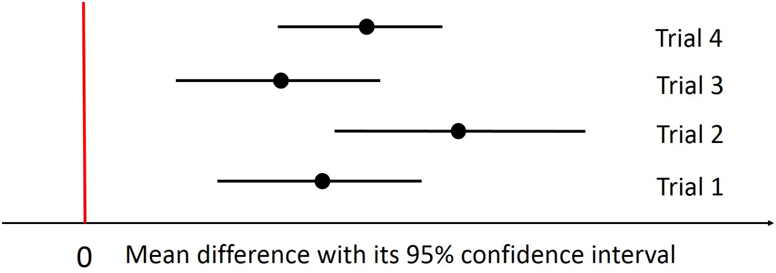

Area of non clinically relevant effects
0
Mean difference with its 95% confidence interval

What should we expect from preclinical trials in the $3R^s$ context (**R**eplacement, **R**eduction and **R**efinement) ?

An extract from the web site of the National center for the replacement refinement and reduction of animals in research (https://www.nc3rs.org.uk/the-3rs)

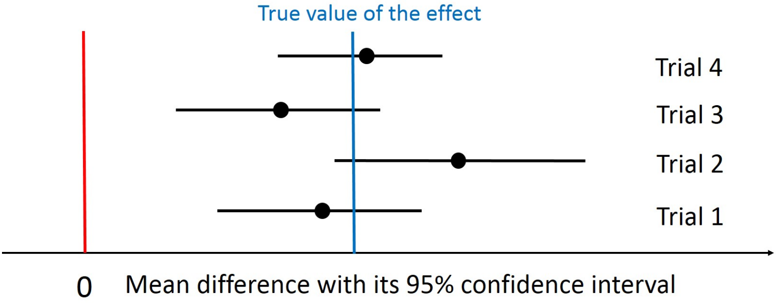
"**Reduction** refers to methods that minimise the number of animals used per experiment or study consistent with the scientific aims. It is essential for reduction that studies with animals are appropriately designed and analysed to ensure robust and **reproducible findings**."

Expected good reproducibility



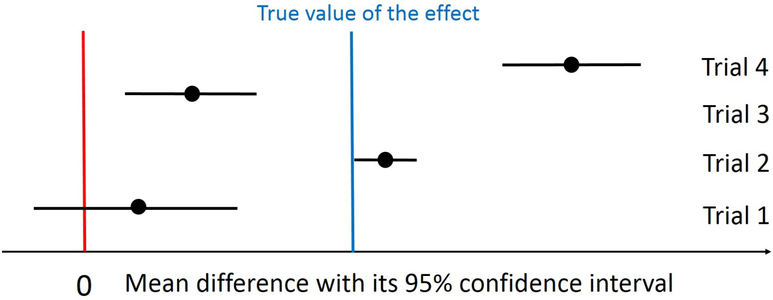
Good reproducibility: estimations from independant trials are consistent.

Good coverage of confidence intervals



Good reproducibility ⇒ good coverage of confidence intervals.
Coverage of a confidence interval = probability that it contains the true value.

Non expected bad reproducibility

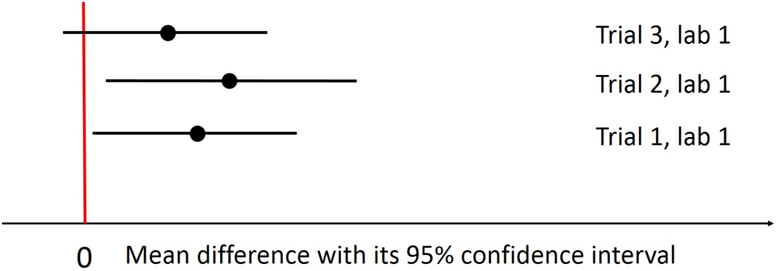


Bad reproducibility: estimations from independant trials are not consistent, confidence intervals have **low coverage**
⇒ **we cannot trust preclinical trial conclusions.**

1 - Choice of the design

1.1 Toward a better reproducibility

Do we need to repeat an experiment within a lab ?

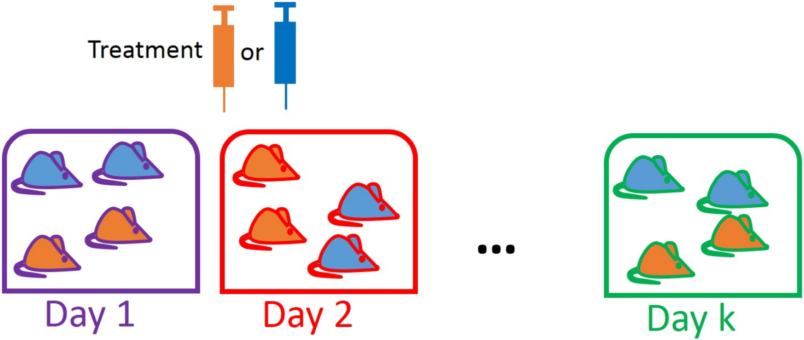


Three rather uncertain estimations.
What will be the conclusion ?

How to take into account various sources of variability ?

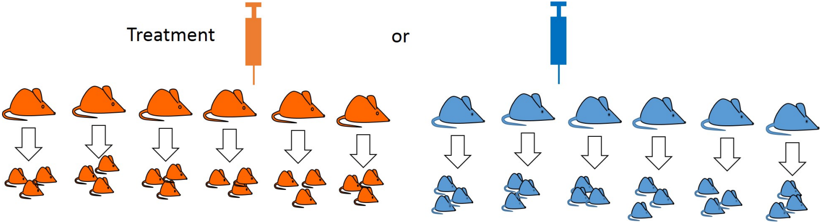
A randomized block design

Imagine you want (or need) to take into account variability linked to a block factor (e.g. day of the experiment).
You could design a unique trial using a **randomized block design** followed by a **unique analysis** taking into account the **studied factor** (the treatment) and the **nuisance factor** (the block).



A nested design

Imagine only one treatment can be administered per block / group (e.g. animals from various litters with the treatment administered to the mother).
You could use a **nested design** followed by a **unique analysis** taking into account the **studied factor** (the treatment) and the **nuisance factor** (the group).
In such a nested design **the number of groups is more important** than the number of animals in each group.



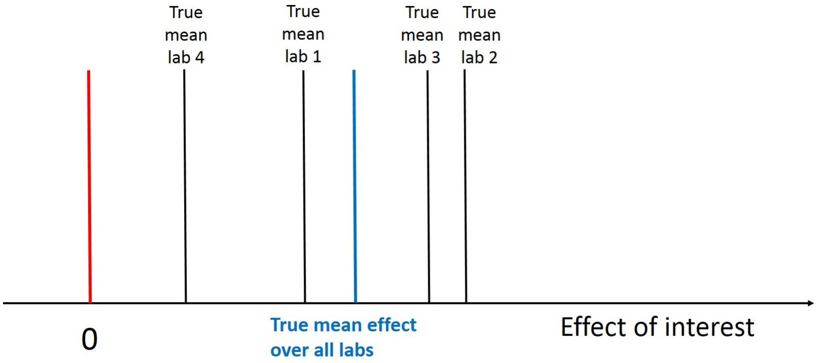
Standardization against reproducibility

From Voelkl *et al.* (2018). **Reproducibility of preclinical animal research improves with heterogeneity of study samples.** PLoS biology, 16(2), e2003693.

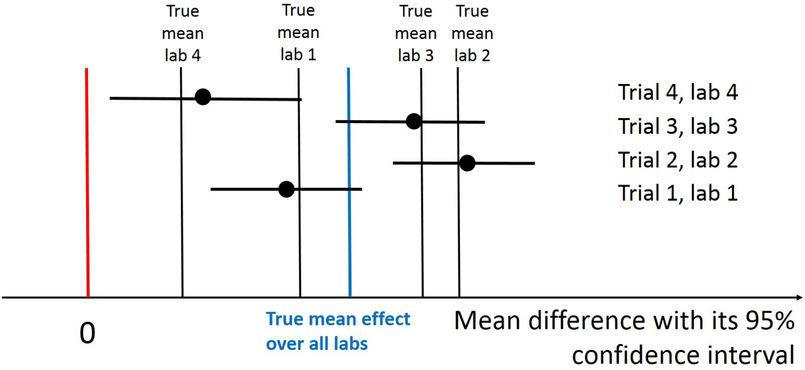
"Single-laboratory studies conducted under highly standardized conditions are the gold standard in preclinical animal research . . . Single-laboratory studies generally failed to predict effect size accurately . . . By contrast, multi-laboratory designs including as few as 2 to 4 laboratories increased coverage probability by up to 42 percentage points . . . **These findings demonstrate that within-study standardization is a major cause of poor reproducibility.**"

Standardization against reproducibility - WHY ?

Due to **inter-laboratory variability**

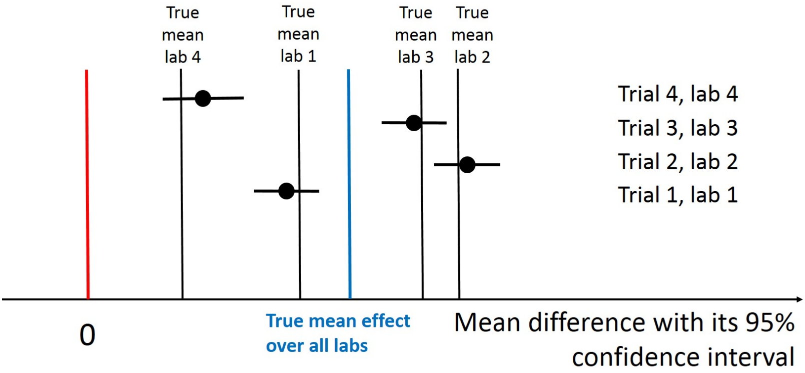


Single-lab trials



Single-lab trials give small confidence intervals that do not overlap, and that often **fail to contain the true mean effect** over all labs.

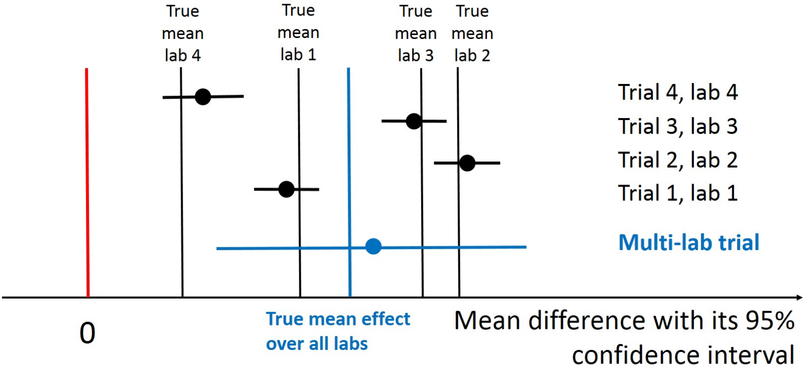
Single-lab trials



Single-lab trials give small confidence intervals that do not overlap, and that often **fail to contain the true mean effect** over all labs.

Increasing the sample size, so narrowing the confidence intervals, even exacerbates the problem !

Multi-lab trials



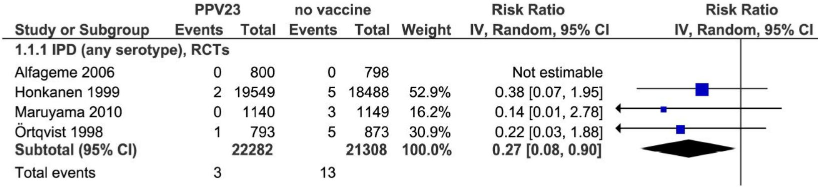
A multi-lab trial gives a larger confidence interval taking into account **inter-lab variability** (via mixed linear models), and far **more likely to contain the true mean effect**.

Meta-analyses

A **meta-analysis** = a **quantitative analysis of all published data**, taking into account all sources of variability.

An example:

- Falkenhorst, G., Remschmidt, C., Harder, T., Hummers-Pradier, E., Wichmann, O., & Bogdan, C. (2017). Effectiveness of the 23-valent pneumococcal polysaccharide vaccine (PPV23) against pneumococcal disease in the elderly: systematic review and meta-analysis. PLoS One, 12(1), e0169368.



	PPV23		no vaccine			Risk Ratio	Risk Ratio
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**IV, Random, 95% CI**	**IV, Random, 95% CI**
1.1.1 IPD (any serotype), RCTs							
Alfageme 2006	0	800	0	798		Not estimable	
Honkanen 1999	2	19549	5	18488	52.9%	0.38 [0.07, 1.95]	
Maruyama 2010	0	1140	3	1149	16.2%	0.14 [0.01, 2.78]	
Örtqvist 1998	1	793	5	873	30.9%	0.22 [0.03, 1.88]	
Subtotal (95% CI)		**22282**		**21308**	**100.0%**	**0.27 [0.08, 0.90]**	
Total events	3		13				

Is it reasonable to intentionally reduce variability due to fixed factors ?

The example of the sex effect

The choice to work on one sex is often done

- to limit sources of variability (e.g. due to hormonal fluctuations with females)
- or to avoid technical problems (e.g. due to agressivity of males).

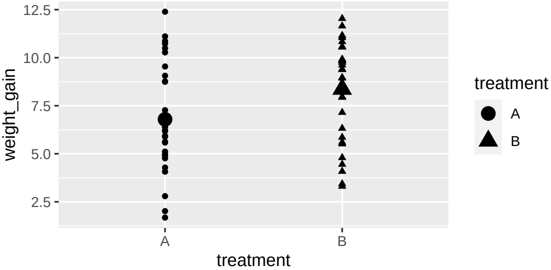
Is it a reasonable choice ?

Confusion bias

Extract of WHO recommendations for a specific vaccine test :

“The animals should be of the same sex or **both sexes in equal numbers** for each dose”

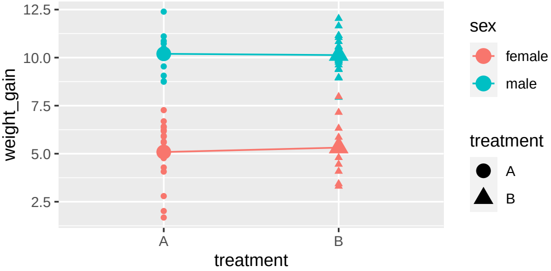
Why do they recommend equal numbers of both sexes ? (look at an example using males and females with sex uncontrolled).



The sex may be a confounding factor

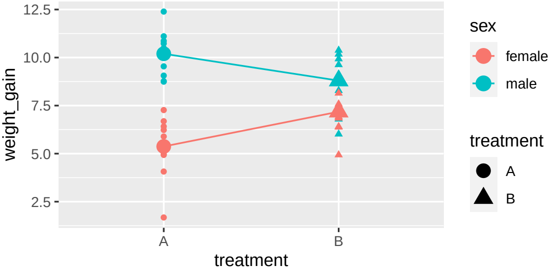
Why do they recommend to use equal numbers of both sexes ?

To avoid a confounding bias.



But is it sufficient ?

Potential interaction between sex and treatment



In case of such interaction, the conclusion on the treatment effect would depend on the chosen sex !

Is such a case possible ? Let see what is known in the area of vaccine tests.

Interaction effect between sex and vaccine

First observed and published a long time ago !

Few references from a quick Scholar Google research on "vaccine test mice sex".

- Fink, A. L., Engle, K., Ursin, R. L., Tang, W. Y., & Klein, S. L. (**2018**). Biological sex affects vaccine efficacy and protection against influenza in mice. Proceedings of the National Academy of Sciences, 115(49), 12477-12482.
- Hoyt, A., Moore, F. J., Knowles, R. G., & Smith, C. R. (**1957**). Sex differences of normal and immunized mice in resistance to experimental tuberculosis. American Review of Tuberculosis and Pulmonary Diseases, 75(4), 618-623.
- Pittman, M. (**1951**). Influence of sex of mice on histamine sensitivity and protection against Hemophilus pertussis. The Journal of infectious diseases, 296-299.

Higher order interactions should even be considered in some cases

- Potluri, T., Fink, A. L., Sylvia, K. E., Dhakal, S., Vermillion, M. S., Vom Steeg, L., . . . & Klein, S. L. 2019. **Age-associated changes in the impact of sex steroids on influenza vaccine responses in males and females**. NPJ vaccines, 4(1), 1-12.

"Sex differences in vaccine efficacy diminished with age in mice. . . Vaccine-induced antibody responses were increased in females by estradiol and decreased in males by testosterone. The benefit of elevated estradiol on antibody responses and protection against influenza in females is diminished with age in both mice and humans."

How to take into account the potential effect of sex and/or of other fixed factors

Trials should use animals of both sexes, preferentially in equal numbers and the analysis should take into account the sex effect using **multivariate linear models**, and its **potential interaction** with the main studied factor (treatment effect).

The analysis should also include other factors (such as age) that could have an effect on the response and in some cases take into account **higher order interactions**.

And take care of other potential confusion biases that could imply too early conclusions on a sex effect (e.g. the historical epidemiological example of correlation between sex and lung cancer due to smoking) !

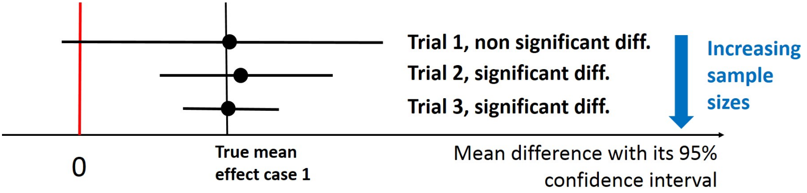
1.2 Choice of the sample sizes

Use the good number of animals !

An extract from the web site of the National center for the replacement refinement and reduction of animals in research (https://www.nc3rs.org.uk/the-3rs)

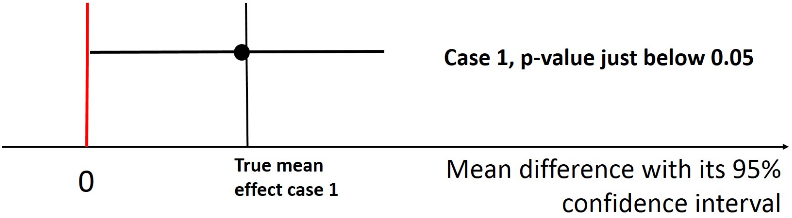
"**Reduction** refers to methods that **minimise the number of animals used per experiment** or study **consistent with the scientific aims**."

Sample size and statistical power


Trial 1, non significant diff.
Trial 2, significant diff.
Trial 3, significant diff.
Increasing sample sizes
0
True mean effect case 1
Mean difference with its 95% confidence interval

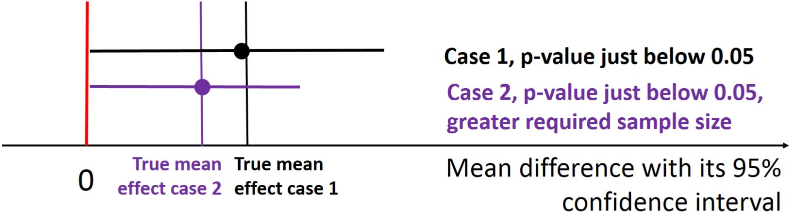
The power of a test is the probability that it will detect a difference in case there is one.
The greater the sample size, the smaller the confidence interval (and the smaller the p-value), so the greater the power !

Prior determination of sample sizes


Case 1, p-value just below 0.05
0
True mean
effect case 1
Mean difference with its 95% confidence interval

For a **given true effect**, what is the **minimal sample size** that will ensure the test to get a **certain probability** (named the **power** of the test) **to detect the effect** ?

Need to define a minimal effect we want to be able to detect


Case 1, p-value just below 0.05
Case 2, p-value just below 0.05, greater required sample size
0
True mean effect case 2
True mean effect case 1
Mean difference with its 95% confidence interval

For a **given true effect**, what is the **minimal sample size** that will ensure the test to get a **certain probability** (named the **power** of the test) **to detect the effect** ?

The smaller the effect, the higher the required sample size !

As the true effect is generally not known in advance, **prior calculation of sample sizes requires the definition of the minimal effect we want to be able to detect** (the minimal effect of biological interest).

What do we need for a prior determination of sample sizes ?

For a t-test of comparison of two means we need:

- the minimal difference (δ) we want to have the probability $1 - \beta$ to detect,
- the chosen power ($1 - \beta$),
- and the standard deviation (σ) supposed common in both groups.

```r
power.t.test(n = NULL, delta, sd, power)
```

For a χ^2-test of comparison of two proportions we need:

- the true proportion in each group, resp. p_1 and p_2,
- the chosen power ($1 - \beta$),

```r
power.prop.test(n = NULL, p1, p2, power)
```

Example for a t-test ?

In a given trial, we want to have a probability of 90% to detect a difference of mean weight gain between two groups as soon as it reaches a threshold of biological interest of 5g. We expect the standard deviation within each group to be near 10g.

```r
power.t.test(n = NULL, delta = 5, sd = 10, power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##               n = 85
##           delta = 5
##              sd = 10
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Another type of power calculation for the same test ?

In a given trial, we would like to detect a difference of mean weight gain between two groups as soon as it reaches a threshold of biological interest of 5g. We expect the standard deviation within each group to be near 10g. We have funding to work on 50 animals per group. What will be the power ?

```r
power.t.test(n = 50, delta = 5, sd = 10, power = NULL)
```

```
##
##      Two-sample t test power calculation
##
##               n = 50
##           delta = 5
##              sd = 10
##       sig.level = 0.05
##           power = 0.697
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Possible bias due to small samples

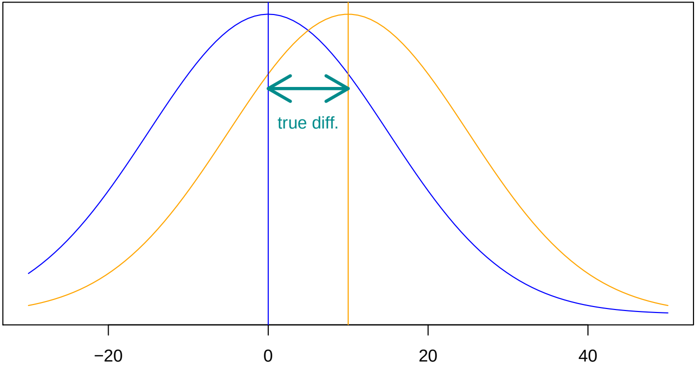
From Button *et al.* (2013). **Power failure: why small sample size undermines the reliability of neuroscience.** Nature Reviews Neuroscience, 14(5), 365.

"A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. . . "

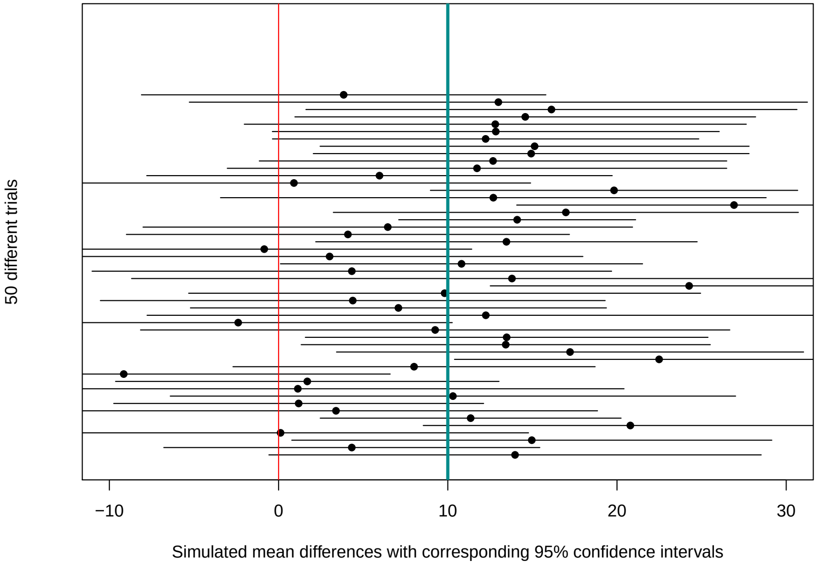
Why working on small samples could induce a bias ?

Let us simulate 50 trials with two samples of size 10 (resp. 40) assuming two Gaussian distributions of respective means 0 and 10 for each treatment and a common standard deviation of 15.

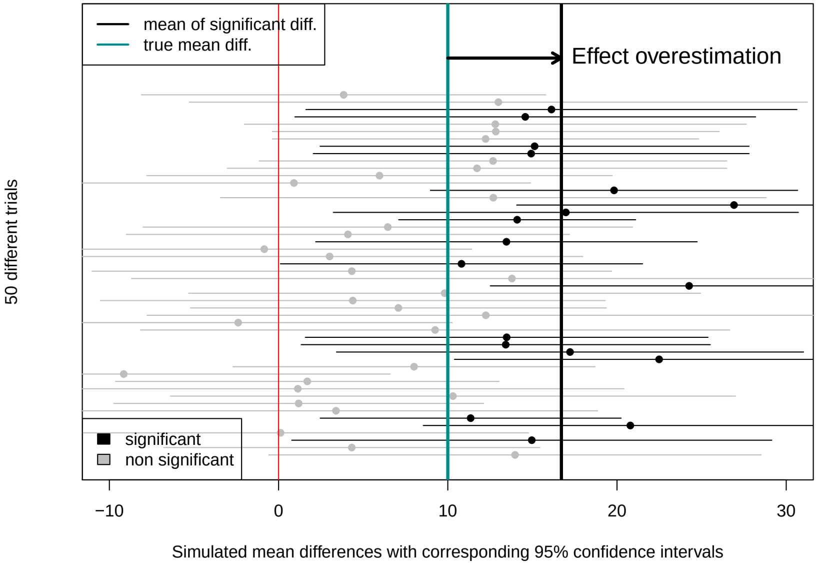
And suppose that only significant differences are reported (**common publication bias**).



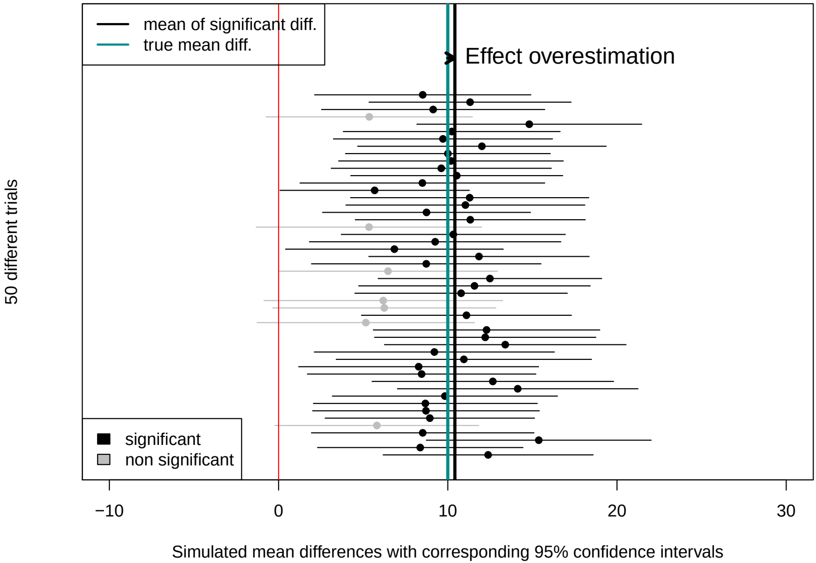
Simulation of 50 trials each with a sample size = 10



Mean of significant differences



Simulation of 50 trials each with a sample size = 40



What solution ?

- To work on bigger samples ? Not always possible.
- To collaborate and design multi-lab trials.
- To publish every result, whatever it is significant or not, as soon as its design is valuable. It is the spirit of registered reports.

PUBLISHED RESULTS
PUBLISHED REGISTERED REPORT
FINAL REVIEW
STAGE 2: RESULTS
REVIEW
STAGE 1: PLAN
RESEARCH PLAN

From Chambers 2019, What's next for registered reports, Nature, vol 573, 187

Why not adopt and promote the registered report format of publication ?

More and more journals are accepting submissions in this format.

- Chambers et al. (2014). **Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond.** AIMS Neuroscience, 1(1), 4-17.
- Hardwicke & Ioannidis (2018). **Mapping the universe of registered reports**. Nature Human Behaviour, 2(11), 793.
- Parker et al. (2019). **Making conservation science more reliable with preregistration and registered reports.**Conservation Biology, 33 (4) Editorial section.

" *The in-principle acceptance is a guarantee of publication, regardless of results, given that the researchers follow the methods that were approved during the original review process. The editor and reviewers evaluate conformity to the original methods at stage 2 when the final manuscript is submitted to the journal.*"

Sharing all data is necessary

Another extract from the web site of the National center for the replacement refinement and reduction of animals in research (https://www.nc3rs.org.uk/the-3rs)

"**Sharing data** and resources (e.g. animals, tissues and equipment) between research groups and organisations can also contribute to **reduction**."

Sharing all data is especially necessary **to enable unbiased meta-analyses.**

1.3 Other classic issues

Why a control group is imperative ?

Illustration of a placebo effect from Glaus et al. (2019). Efficacy of long-term oral telmisartan treatment in cats with hypertension: Results of a prospective European clinical trial. Journal of veterinary internal medicine, 33(2), 413-422.

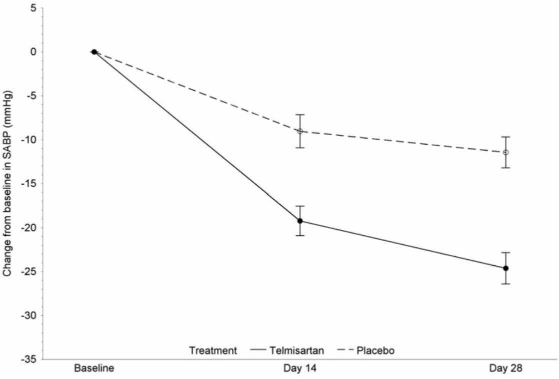


FIGURE 2 Mean (95% confidence interval of the mean) changes from baseline in systolic arterial blood pressure during the blinded efficacy phase (per protocol set population)

A placebo effect could be due to regression to the mean

Imagine a group of animals selected at random in a population,
among which a **second selection is performed** to keep only
animals with hypertension,
hypertension being defined by a **tension above a fixed threshold**.

Illustration of the regression to the mean using simulations

Global distribution of the variable in the whole population

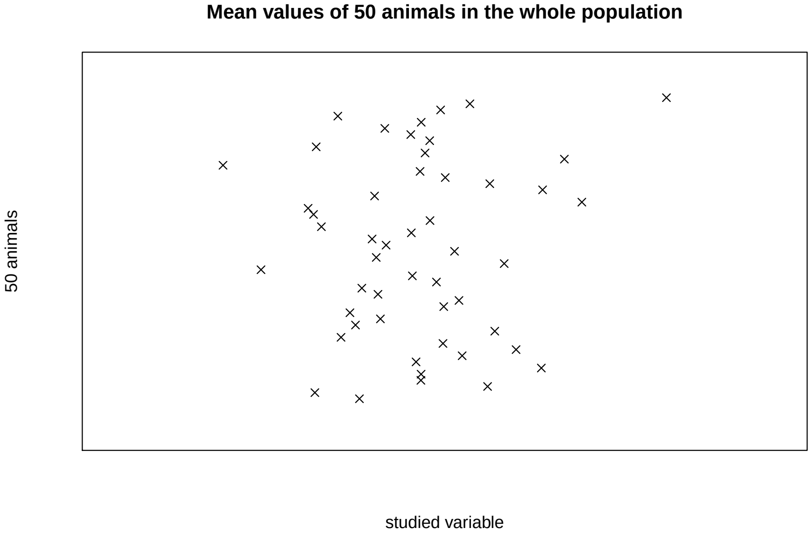


Illustration of the regression to the mean using simulations

Taking into account fluctuations for each animal

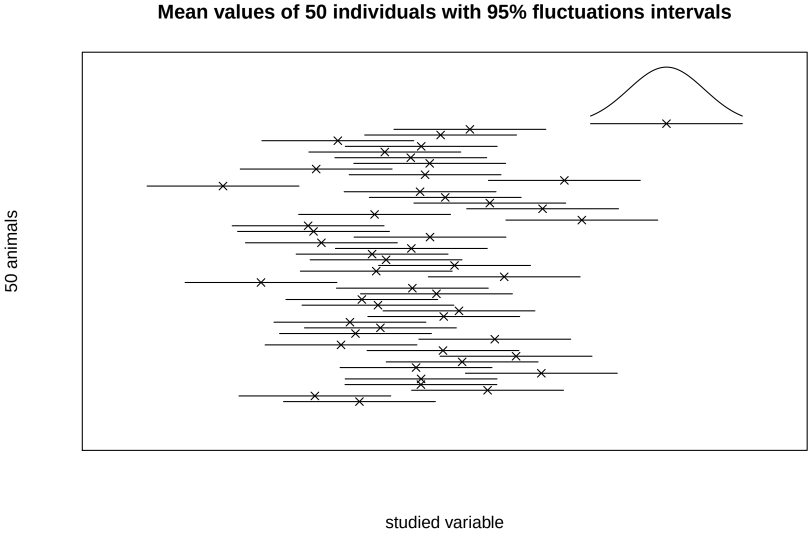


Illustration of the regression to the mean using simulations

Simulation of a first observation of the variable on animals at the beginning of the study

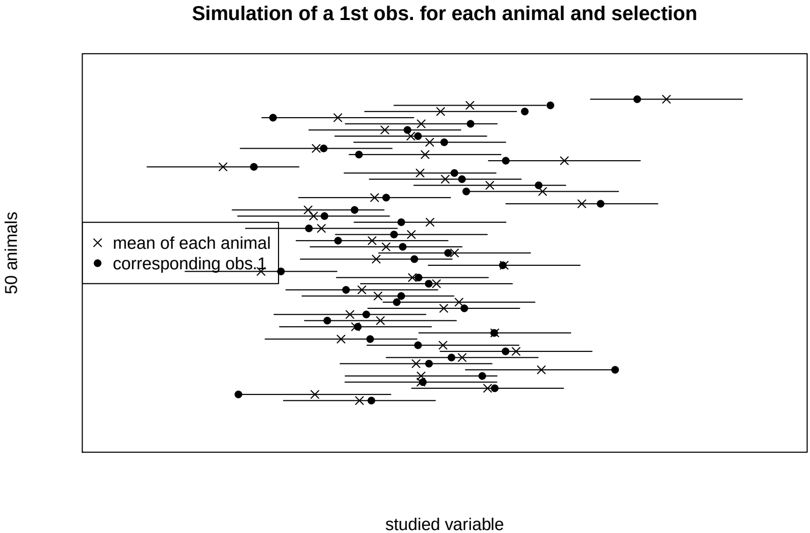


Illustration of the regression to the mean using simulations

Selection of animals reaching the defined threshold

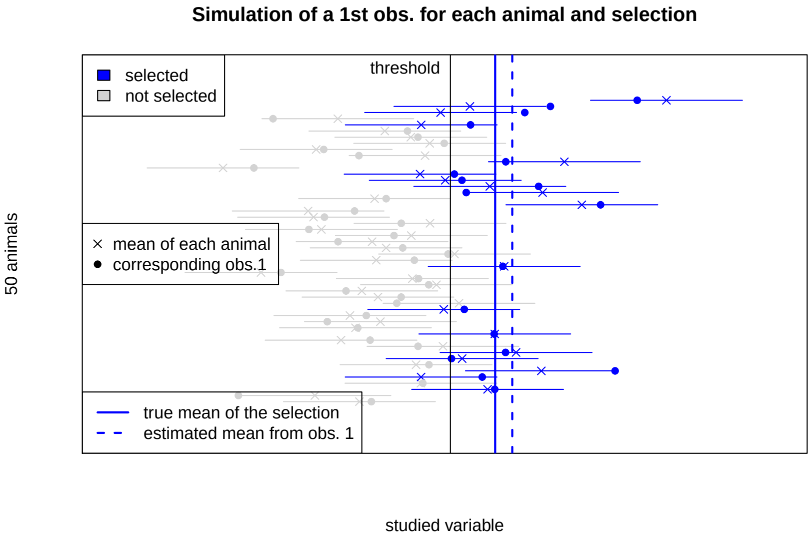
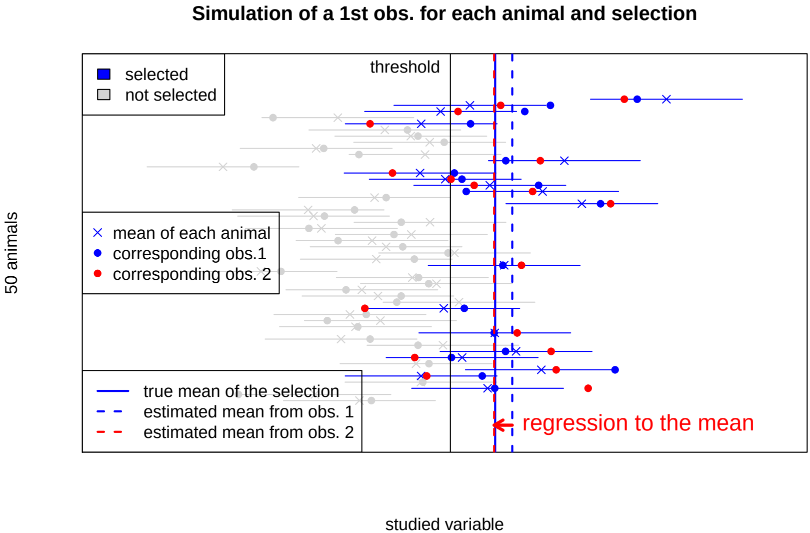


Illustration of the regression to the mean using simulations

Simulation of a second observation on each selected animal at the end of the study



Why randomization is imperative ?

To ensure comparability of treatment groups,
achieve valid comparison of the effects of the studied treatment(s),
and so limit the risk of confounding bias.

When blinding is imperative ?

Blinding is a procedure asuring that people involved in a trial do not know which treatment has been administered to each animal. It is used to limit biased interpretation of a trial results.

In trials on animals, **blinding is imperative at the outcome assessment step, as soon as the assessment of this outcome may be subjective.**

2 - Analysis of data

Data analysis should never be left for last !

Another extract from the web site of the National center for the replacement refinement and reduction of animals in research (https://www.nc3rs.org.uk/the-3rs)

"It is essential for reduction that studies with animals are **appropriately designed and analysed** to ensure robust and reproducible findings.

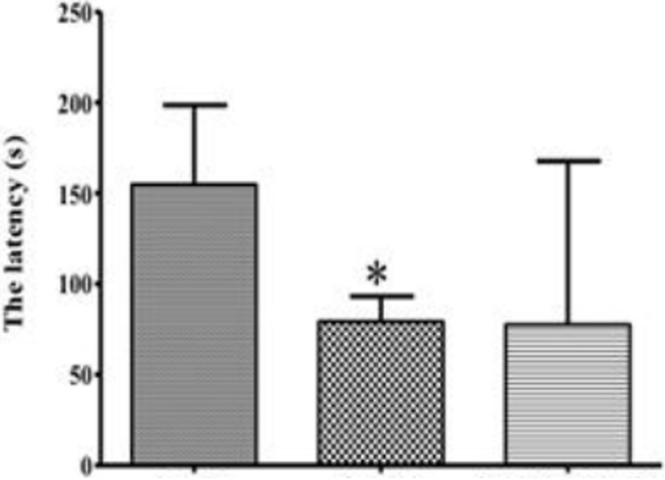
Reduction also includes methods which allow the **information gathered per animal in an experiment to be maximised** in order to reduce the use of additional animals."

One must think to the data analysis while designing the trial, not after the data collection!

2.1 Graphical representation

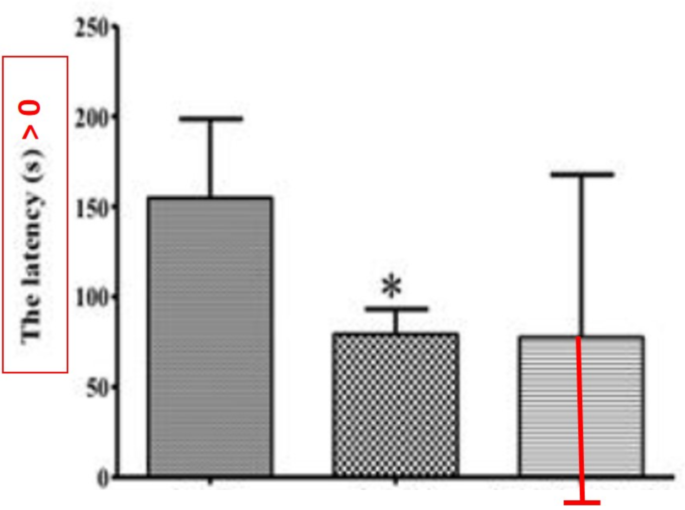
A unfortunately so common plot !

Do you see any problem on the example below ?



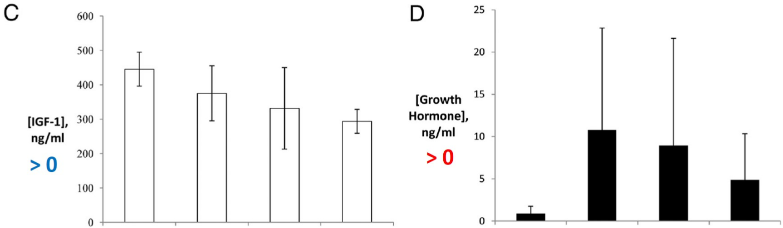
Do not tell me you never saw that !

Do you see the problem now ?



Why such a plot ?

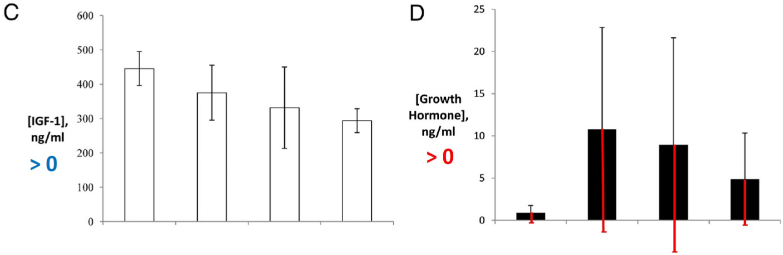
Why the margin of error (or margin of fluctuation) is sometimes represented only above the estimation ?



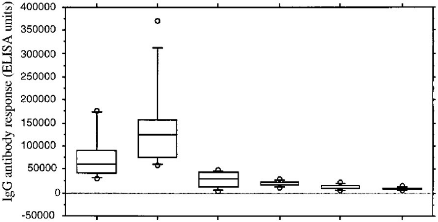
Why such a plot ?

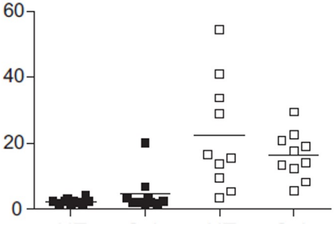
Why the margin of error (or margin of fluctuation) is sometimes represented only above the estimation ?

I am afraid the answer is clear from this extract of a published figure.



Don’t you think there are some more relevant and more informative alternatives ?

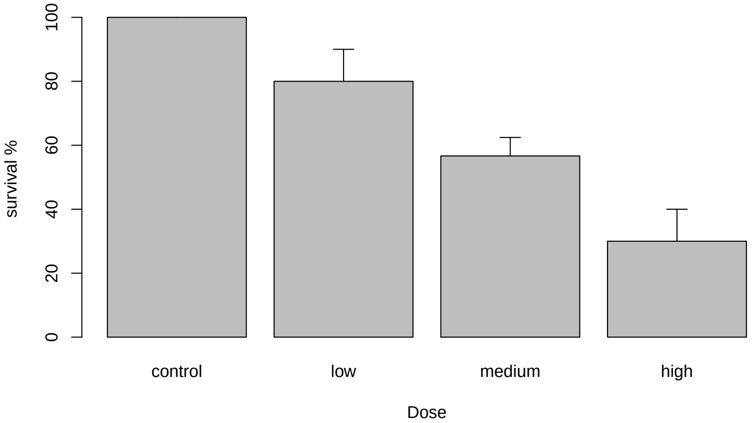




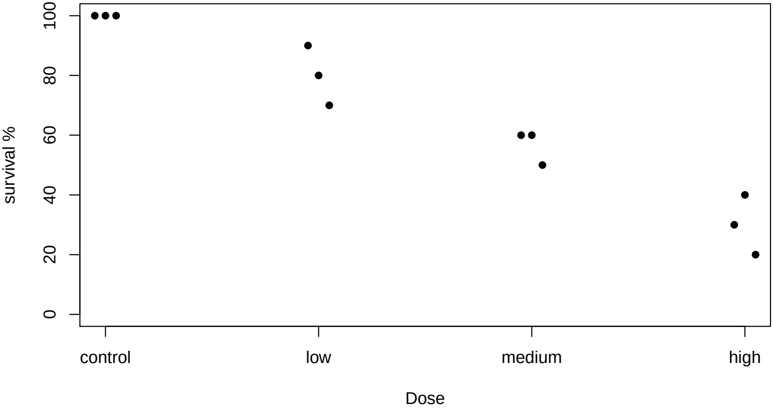
What improvement could we propose for the second figure? What do horizontal lines represent? Means or medians? What to prefer ?

Focus on a specific example

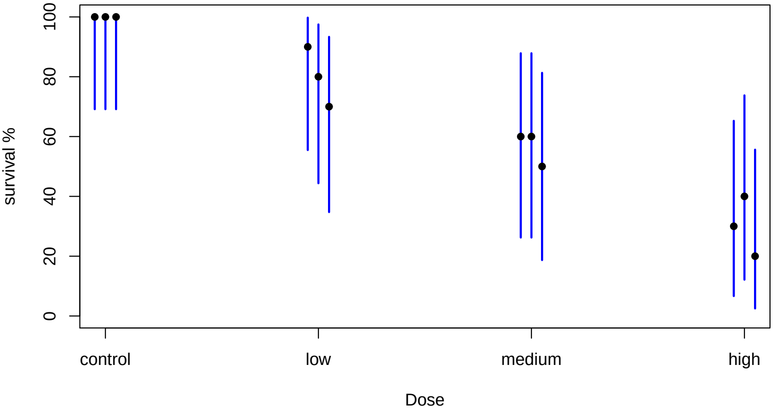
A proposed figure for a trial in ecotoxicology : survival percentages of organisms exposed to a pollutant at different concentrations. Three replicates per exposure condition = 3 beakers each containing 10 animals .



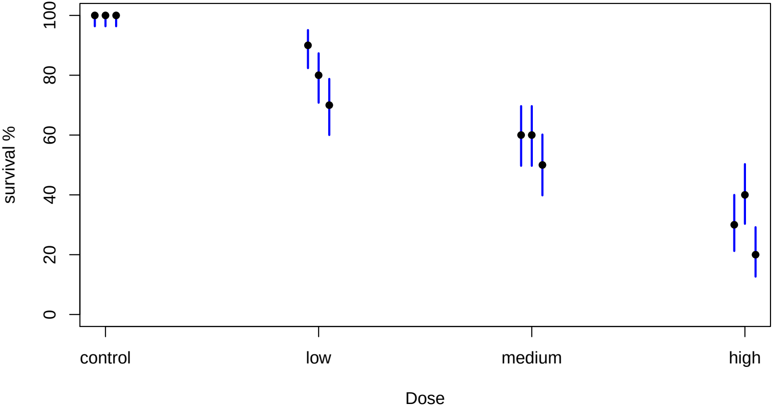
A simpler and more informative figure



An even more informative figure inluding 95% confidence intervals on each survival percentage

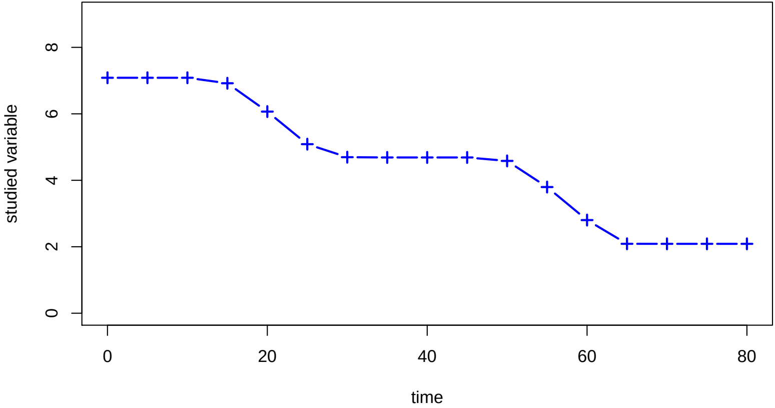


The same figure with the same survival percentages obtained with 100 organisms per beaker instead of 10



Another common but not always relevant figure

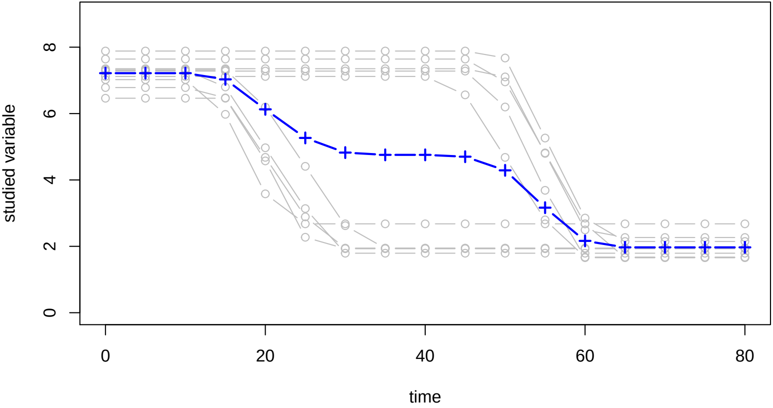
The mean curve = mean of animal responses at each time.



Another common but not always relevant figure

The mean curve = mean of animal responses at each time.

It is important to represent the curve of each animal to judge if the mean curve is a good summary.



To conclude about representation of data

- The mean and the standard deviation SD (or the standard error of the mean SEM) are not systematically good summaries of your data.
- A good figure should be simple but **informative**.
- What we need to see in a figure is a **good summary of the distribution of raw data**.

A quick reminder if necessary

For a **Gaussian distribution**,

$m \pm 2SD$ is a 95% **fluctuation interval** (expected to contain 95% of the observations)

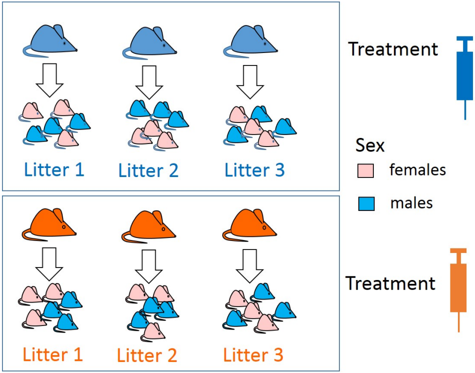
$m \pm SD$ is a 68% fluctuation interval (expected to contain 68% of the observations)

$m \pm 2SEM$ is a 95% **confidence interval** (with an expected coverage of 95%)

$m \pm SEM$ is a 68% confidence interval (with an expected coverage of 68%)

2.2 Statistical analysis of data

Which methods do we need to properly analyse data ?



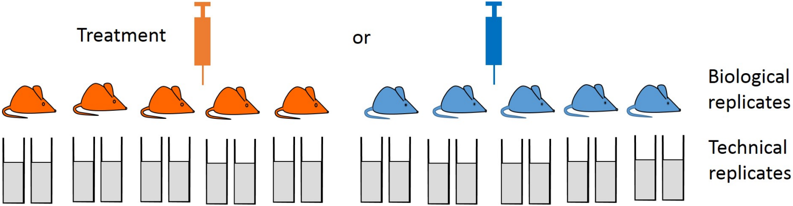
In the above example we need at least a **mixed linear model** taking into account two **fixed factors** (the treatment and the sex) and **their potential interaction** and a **random factor** (the litter).

Other examples from vaccine tests

- Respiratory challenge for comparing two or more vaccines with 15 mice per **vaccine group** (**fixed factor**), 5 mice analysed per **sampling time** (2 hours, 5 days, 8 days) (**fixed factor**), with measure of the **microbial response** in $log_{10}(UFC.lungs^{-1})$ (**continuous response**).
- Analysis of the data using at least a **linear model with two fixed factors** and their potential interaction.
- If the **mice are of both sexes** ⇒ **one more fixed factor**.
- To take into account a potential **cage effect** ⇒ **one more random factor**.
- For a **binary response** (e.g. proportion of dead mice) ⇒ **generalized linear model**.
- ...

Comparison of different vaccines on a quantitative response (e.g.)

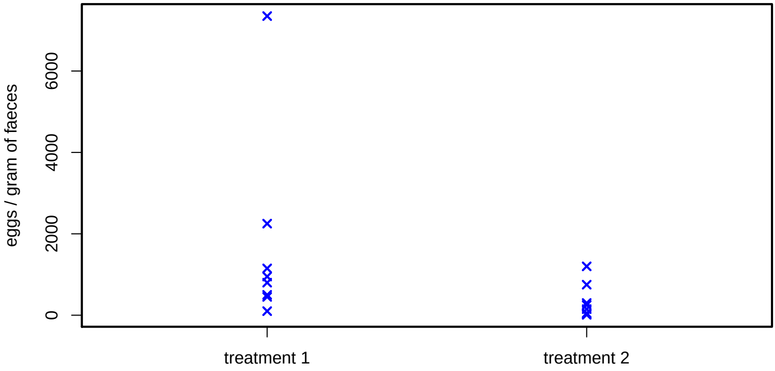
Is it reasonable to omit one level of variability (or equivalently to put at the same level both nested sources of variability) ?



Analysing such a design just by comparing two groups of size 10 assumes that biological variability could be neglected in front of technical variability.

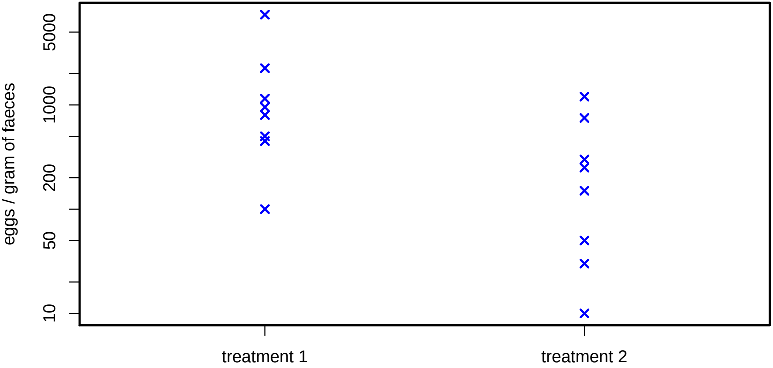
This is a **very strong and unlikely hypothesis** that should be checked using a mixed model to separately estimate inter-animal and intra-animal variability !

Can we remove outliers just because they are outliers ?



Of course not ! Outliers may be representative of a minority, or observations may appear as outliers just because of a bad chosen scale. . .

A scale transformation may be necessary before use of statistical methods



Are missing data problematic ?

- As soon as the missing of an observation may be linked to the treatement, of course yes !
  - **Intention-to-treat** (ITT) analysis versus per protocol analysis
  - Proper **analysis of censored data** (e.g. methods for right censored survival data)
- Limitation of pain and distress (RRRs) procedures may induce missing data. A **difficult compromise** !

2.3 Interpretation of statistical results

Why some conclusions from the same results could differ ?

One paper, two discussions. Anaesthesia journal asks independent experts to draw their own conclusions from a same trial.

- Sieber et al. (2019). **Depth of sedation as an interventional target to reduce postoperative delirium: mortality and functional outcomes of the Strategy to Reduce the Incidence of Postoperative Delirium in Elderly Patients randomised clinical trial.** British journal of anaesthesia, 122(4), 480-489.
- Vlisides et al. (2019). **Hypnotic depth and postoperative death: a Bayesian perspective and an Independent Discussion of a clinical trial.** British journal of anaesthesia, 122(4), 421-427.

Extracts from Sieber et al. 2019

“Intention-to-treat analysis showed **no significant difference** between intervention groups in mortality up to 1 yr (Figure 2, log rank P = 0.96). Analysis using Cox proportional hazard model estimated the **hazard ratio of 1 yr mortality** for lighter vs heavier sedation being **0.85 (95% CI, 0.44-1.97)** after accounting for age and MMSE scores, the variables used for stratified randomisation.”

“In conclusion, the results from this analysis show that **there is no difference in mortality** or return to ambulation 1 yr after hip fracture surgery in elderly patients receiving heavier and lighter intraoperative sedation.”

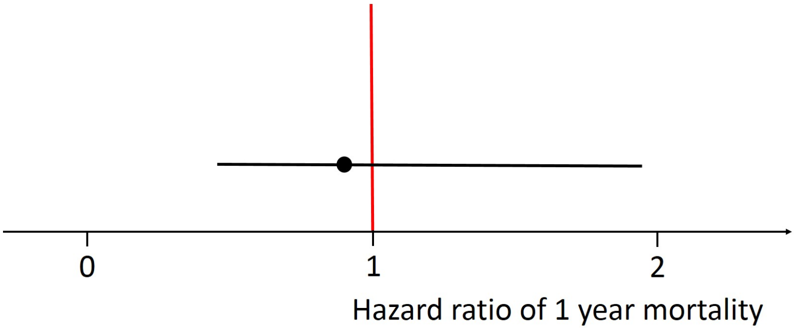
Extracts from Vlisides et al. 2019

"In designing a trial to **detect an absolute decrease in 1-yr mortality from 10% to 9% (10% relative reduction)** with ... a statistical significance level of <0.05, **an adequately powered (>80%) trial would require about 13 500 patients** per study group. **Yet a 1% absolute reduction in death should be considered clinically meaningful**, as this would mean that for every 100 patients treated with 'lighter' anaesthesia, one life would be saved. "

"We must also realise that **it is very difficult to detect and demonstrate a small effect of an intervention on a clinically important outcome such as death**, and conventionally designed clinical **trials of modest sample size might be inadequate for this purpose**. "

What can we conclude from the results of this trial ?

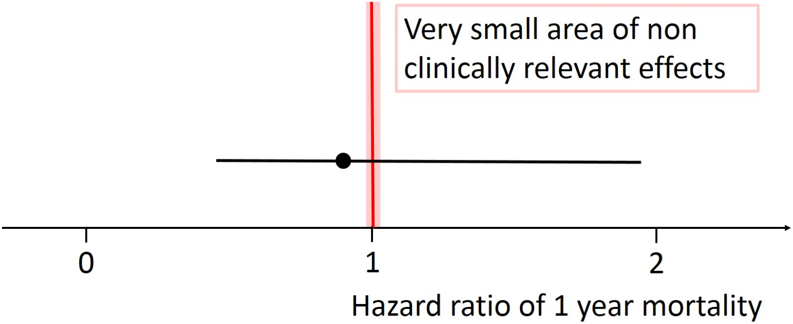
"...**hazard ratio of 1 yr mortality** for lighter vs heavier sedation being **0.85 (95% CI, 0.44-1.97)**..."



Of course not that there is no effect !

What could be considered as the area of non clinically relevant effect ?

“. . . Yet a **1% absolute reduction in death should be considered clinically meaningful**, as this would mean that for every 100 patients treated with ‘lighter’ anaesthesia, one life would be saved. . . .”



How to interpret a p-value ?

Wasserstein & Lazar (2016). **The ASA's statement on p-values: context, process, and purpose.** The American Statistician, 70(2), 129-133. (more than 2000 citations - Scholar Google sept. 2019)

- P-values can indicate **how compatible the data are with a specified statistical model**.
- P-values **do not measure the probability that the studied hypothesis is true**.
- Scientific **conclusions and decisions should not be based only** on whether a **p-value** passes a specific threshold.
- Proper inference requires **full reporting and transparency**.
- A **p-value does not measure the size of an effect** or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a hypothesis

We should be ready to question our habits

Conservatism is the enemy of progress !

Even in the use of statistics, nothing is set in stone !

If you want to go further, look at

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). **Moving to a world beyond “$p < 0.05$”**. The American Statistician, 73:sup1, 1-19.

a small extract of this paper:

“We conclude, based on our **review of the articles in this special issue and the broader literature**, that **it is time to stop using** the term **“statistically significant”** entirely. Nor should variants such as “significantly different,” “$p < 0.05$,” and “nonsignificant” survive, whether expressed in words, by asterisks in a table, or in some other way."

Stop focus on p-values and interpret model estimations !

- **Linear models** (continuous data, e.g. arterial blood pressure): model coefficients correspond to **additive effects** (differences) on the outcome.
- **Linear models on log scale** (continuous data in log scale): coefficients (after exponential transformation) correspond to **multiplicative effects** on the outcome.
- **Logistic models** (binary data, e.g. alive / dead): coefficients (after exponential transformation) correspond to **Odds Ratios (OR)** that often overestimate Risk Ratios (RR).
- **Cox models** (survival-time or time-to-event data): coefficients (after exponential transformation) correspond to **Hazard Ratios (HR)**, so ratios of hasard rates (similar to RR).
- **Accelerated Failure Time (AFT) survival models** (less common analysis for previous data): coefficients (after exponential transformation) correspond to **ratios of survival times**.

Some advices to design a good trial

1/ Define your objective(s)

One clear objective if possible !
The less, the better !

2/ Choose your outcome / endpoint / criteria (the studied variable)

A unique variable if possible !
The less, the better !

Is it continuous, dichotomous ?
This is important for the statistical analysis and thus the design.

3/ List the factors (qualitative) or covariates (quantitative) to take into account

- **Studied factors** (e.g. treatment, sex)
- **Nuisance factors** (e.g. litter, cage, lab, day of experiment) or **covariates** (e.g. age, initial measure of the outcome continuous variable)

4/ Choose a reasonable design while anticipating statistical analysis of data

- Choose an **relevant design taking into account all the listed factors** (both studied and nuisance factors),
- think about **randomization**,
- and determine reasonable **sample sizes** (power calculation that requires to know **which statistical method** will be used and **what minimum effect size** we want to be able to detect).

Do not wait the end of the trial to ask the help of a statistician if your case is not so simple.

To conclude in the context of RRRs (Reduce, Refine, replace)

Our global aim in designing preclinical trial should be to **improve animal welfare and scientific quality**.
So we have the obligation to build **good trials**,
that will drive science forward (so could be published),
whatever the results of the trial !

References that could help you to design good preclinical trials

- **ARRIVE**: Kilkenny et al. (2010). Animal research: reporting in vivo experiments: the ARRIVE guidelines. British journal of pharmacology, 160(7), 1577-1579.
- **PREPARE**: Smith et al. (2017). PREPARE: guidelines for planning animal research and testing. Laboratory Animals, 52(2), 135-141.
- Festing & Altman (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR journal, 43(4), 244-258.
- Landis et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. Nature, 490(7419), 187.