

Foundations for Evaluating Study Design and Statistical Approaches for the IACUC



National Institutes of Health
Office of Laboratory Animal Welfare

March 9, 2023

Penny S Reynolds, PhD
Department of Anesthesiology,
Department of Small Animal Clinical Medicine
Statistics in Anesthesiology Research (STAR) Core,
University of Florida



Outline

- Background and Rationale: Why design matters
- Design essentials
 - Bias minimization
 - The design skeleton
 - Screening for sample size





Learning objectives

At the conclusion of this activity, participants should be able to:

- Recognize the 5 common components of an experimental design
- Describe 3 essentials for right-sizing an experiment
- Identify red-flag items in protocol descriptions





Background: Why does it matter?



Enhancing Rigor, Transparency, and Translatability in Animal Research

NIH Advisory Committee to the Director (NIH-ACD)

Working Group Report, June 11, 2021



National Institutes of Health
Turning Discovery Into Health

<https://acd.od.nih.gov/working-groups/eprar.html>

Motivations

- High failure rates in therapeutics development
- ‘Documented problems’ in replication and translatability
- Need to adjust to evolving scientific best practice



Two themes:

1. Animal-based research quality is poor
2. Investigator understanding and training in statistics and statistical concepts are really poor

“Animal research needs good study design, statistical data analysis and results reporting – without them, even the best animal models are useless.”

- NIH (NIH-ACD) Report. June 11, 2021



Spoiler alert! None of this is new

2014

NIH Francis Collins and Lawrence A. Tabak outline restructuring initiatives to improve preclinical research. *Nature*, 505(7485):612-613

2009

NIH/OLAW (USA) & NC3Rs (UK)

Systematic survey and review of published, government-funded preclinical research → Major reporting deficiencies, omission of key information related to scientific value → ARRIVE (2010); ARRIVE 2.0 (2021)

1986

Bateson's cube: Study quality is the third dimension of harm:benefit. *New Scientist*, 109: 30–32, 1986

1959

Russell WMS, Burch RL. *The principles of humane experimental technique*. London: Methuen.

Foundation document for the 3Rs



The 3Rs and statistically-based study design



- 3Rs is the central ethical framework for animal-based research - **“Maximal information for minimal harms”**
- Reduction and refinement achieved by ‘good and appropriate’ statistically-based experimental design

Russell WMS, Burch RL. *The Principles of Humane Experimental Technique*. 1959



Barriers

NIH recommends IACUCs could be more pro-active as facilitators

BUT

1. Perception barrier: “Not our lane”
2. Perception barrier: “What, not MORE regulation”
3. Knowledge barriers: Attitudes & myths



Perception barrier I: Not our lane?

ANS. Yes it is: OLAW FAQ D.12

Scientific and technical merit is the purview of NIH Scientific Review Groups

BUT

The IACUC is expected to consider U.S. Government Principles

- Principle II: Evaluation of the relevance of a procedure to *human or animal health*, the *advancement of knowledge*, or the *good of society*.

Other PHS Policy review criteria:

- *Sound research design*
- *Rationale for involving animals*
- *Scientifically valuable research*

Presumably, a study that could not meet these basic criteria is inherently unnecessary and wasteful and, therefore, not justifiable.

<https://olaw.nih.gov/faqs#/guidance/faqs?anchor=50327>



Perception barrier II: More regulatory burden

Perceived as a:

- Set of externally imposed rules, a “regulatory burden”
- “Unnecessary busy work”

All best-practices (including design) are part of Responsible Conduct of Research (RCR)

Goals:

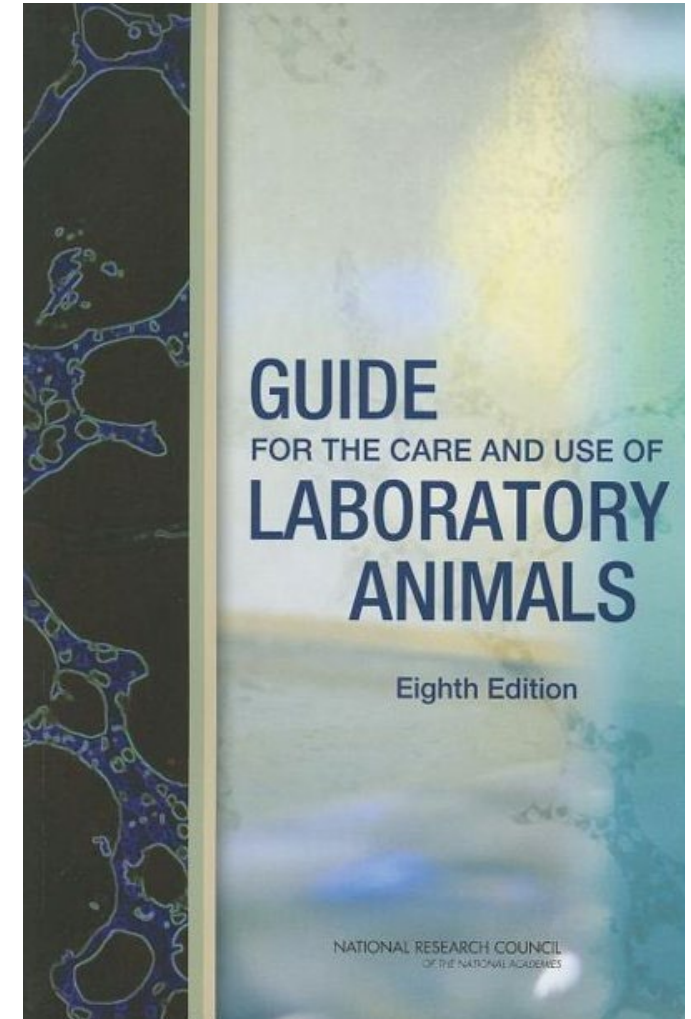
- Embed RCR in routine practice
- Develop skills & resources
- Alignment between proposal, the animal use protocol, and protocol implementation



Better science ↔ Better animal welfare

It is **duty of care** to minimize harms caused by wasting animals in low-quality under-powered studies (*Guide* p.26)

- IACUC cannot evaluate specific experimental design
- IACUC can and should evaluate experimental design components





Knowledge & cognitive barriers

- Few IACUCs include a statistician
- Very few statisticians understand animal-based research (only interested in the maths)
- Very few IACUC members and researchers understand 'statistics'



Knowledge barriers: Attitudes and myths

Attitude: “We hate statistics”

Myth 1: ‘Statistics’ = Highly mathematical = ‘Magic’

Myth 2: ‘Statistics’ = ‘Analysis’;

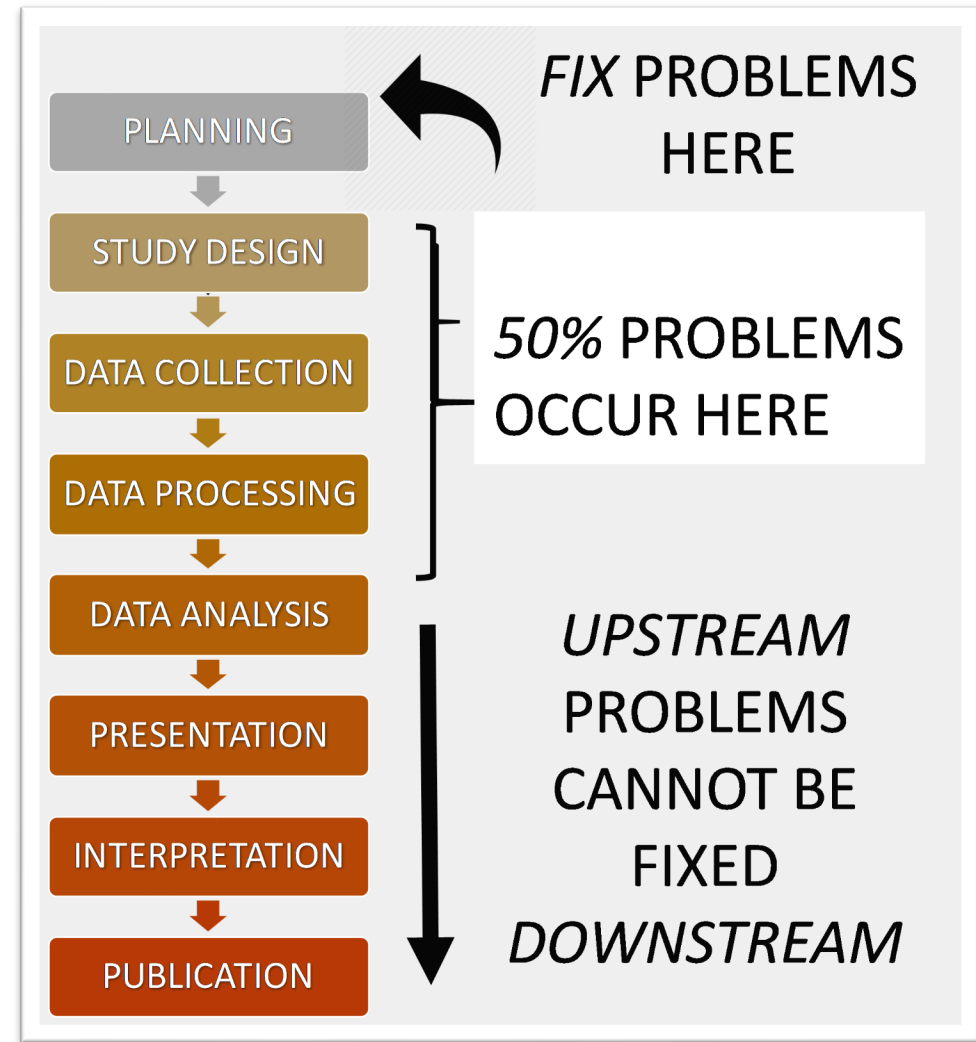
“We only need to consult a statistician after the data are collected”

Myth 3: ‘Statistics’ = “With the software, so easy anyone can do it”



Take home: “Statistics” is a Process not a Thing

- **Design** = planning, conduct, data collection, analysis, and interpretation
- **Must be built in EARLY BEFORE data are collected**
- Poor upstream basics → Unreliable downstream results
- Fancy analyses cannot rescue bad planning & design



Adapted from Sackett 1979,
Altman 1980 *BMJ* 281: 1182-4

Statisticians should be consulted at the beginning, not the end



“To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of”

- Sir Ronald A. Fisher



What is design?





What is (statistical) design of experiments (DOE)?

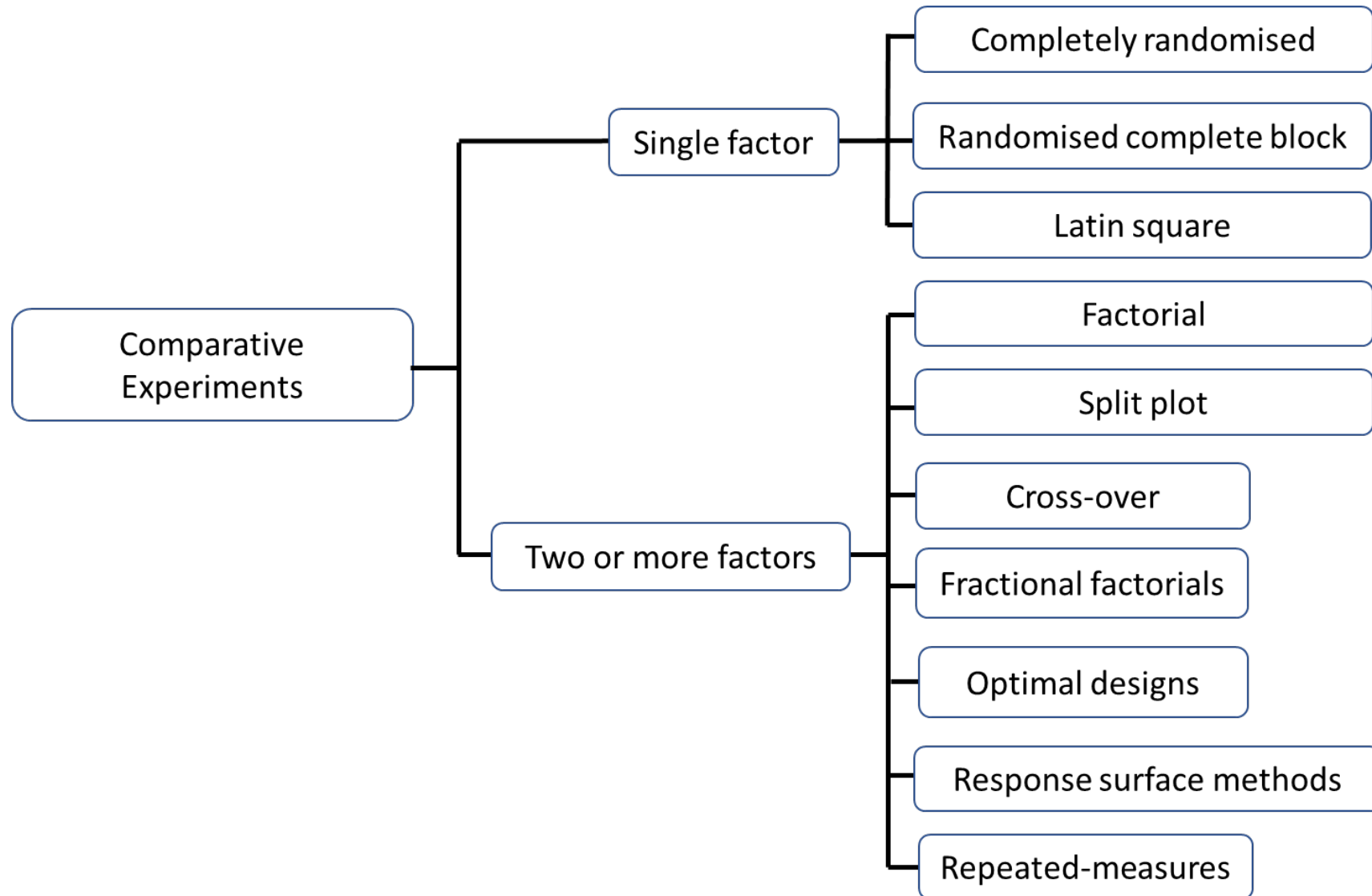
- Often confused with descriptions of technical methods and materials
- DOE = Formal logical & statistical structuring of the elements of the research question = independent variables/inputs

A “good” design:

- *Discriminates* ‘signal’ from ‘noise’ → power to detect real treatment differences
- *Detects* interactions → synergistic/inhibitory responses
- *Controls* ‘noise’ = unwanted variation (blocking, subsampling)
- Minimizes bias
- Is “right-sized”



DYK? Statistically-based designs have been available for over a century





A good design makes a valid study

Validity = “the degree to which a result from a study is likely to be true”
= strength of the cause-effect relationship

Internal validity

- Methods-based (“truth within the study”)
- *Purpose: Are results true? reliable?*

***Design* comes before inference**
***Data quality* comes before analysis**



Benefits of a statistically-based study design

1. **Require far fewer resources** for the amount of information obtained.

- More efficient
- More economical
- More informative
- Spares animal use

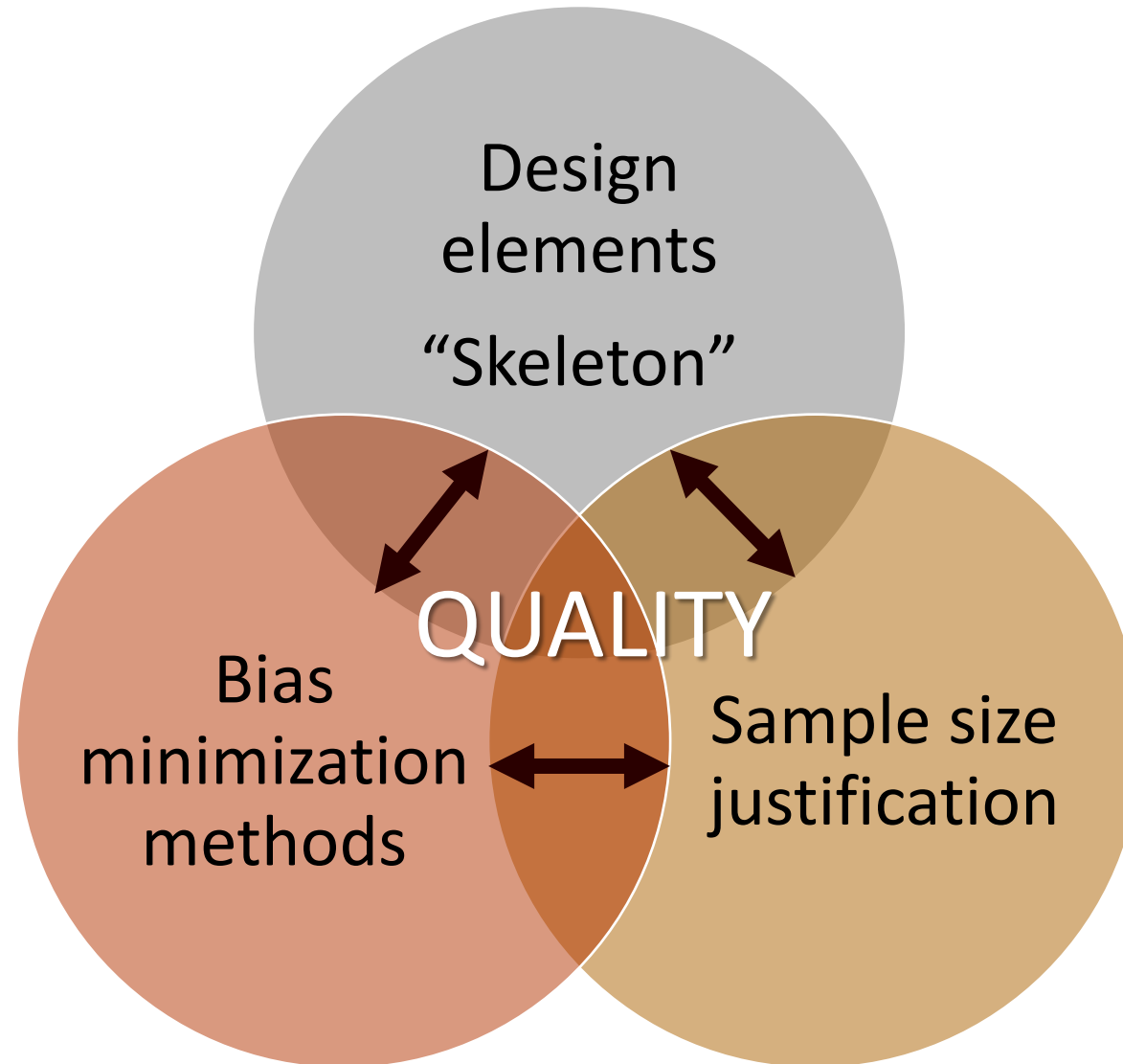
2. **Statistical tests rely on correct design**

With bad or no designs:

- Statistical tests are invalid
- Study quality cannot be rescued by analyses done after the data are collected



Design basics: Key components of internal validity





1. Bias minimization: Randomization

This is a statistical method

- “Random” IS NOT “haphazard,” “ad hoc,” “unplanned,” “alternating”
- Formal technical probabilistic process of:
 - (a) Assigning interventions to the experimental units
 - (b) Determining the order of processing/measurement

Purpose:

- Prevent systematic bias in treatment assignment,
- Prevent selection bias
- Ensure validity of inferential tests

SAS: *proc plan, proc surveyselect, proc factex*

R: *blockrand, randomizeR, pwr, experiment, clusterPower, CRTSize*



2. Bias minimization: Allocation & sequence concealment (sometimes known as “Blinding,” “sealed envelope”)

This is an operational (logistics) method:

- Methods that conceal which treatment was received by which subject
- Purpose: Minimizes conscious or unconscious bias in *allocation, assessment, interpretation*
- Not just for histology!
- Use the highest level possible
- Who? Operators, assessors, analysts



Design skeleton

Break the research question into basic components.

There should be sufficient information in the protocol to evaluate:

P – **P**articipants, Study **P**opulation, Test **P**latform

I – **I**ntervention

C – **C**omparator, control

O – **O**utcome

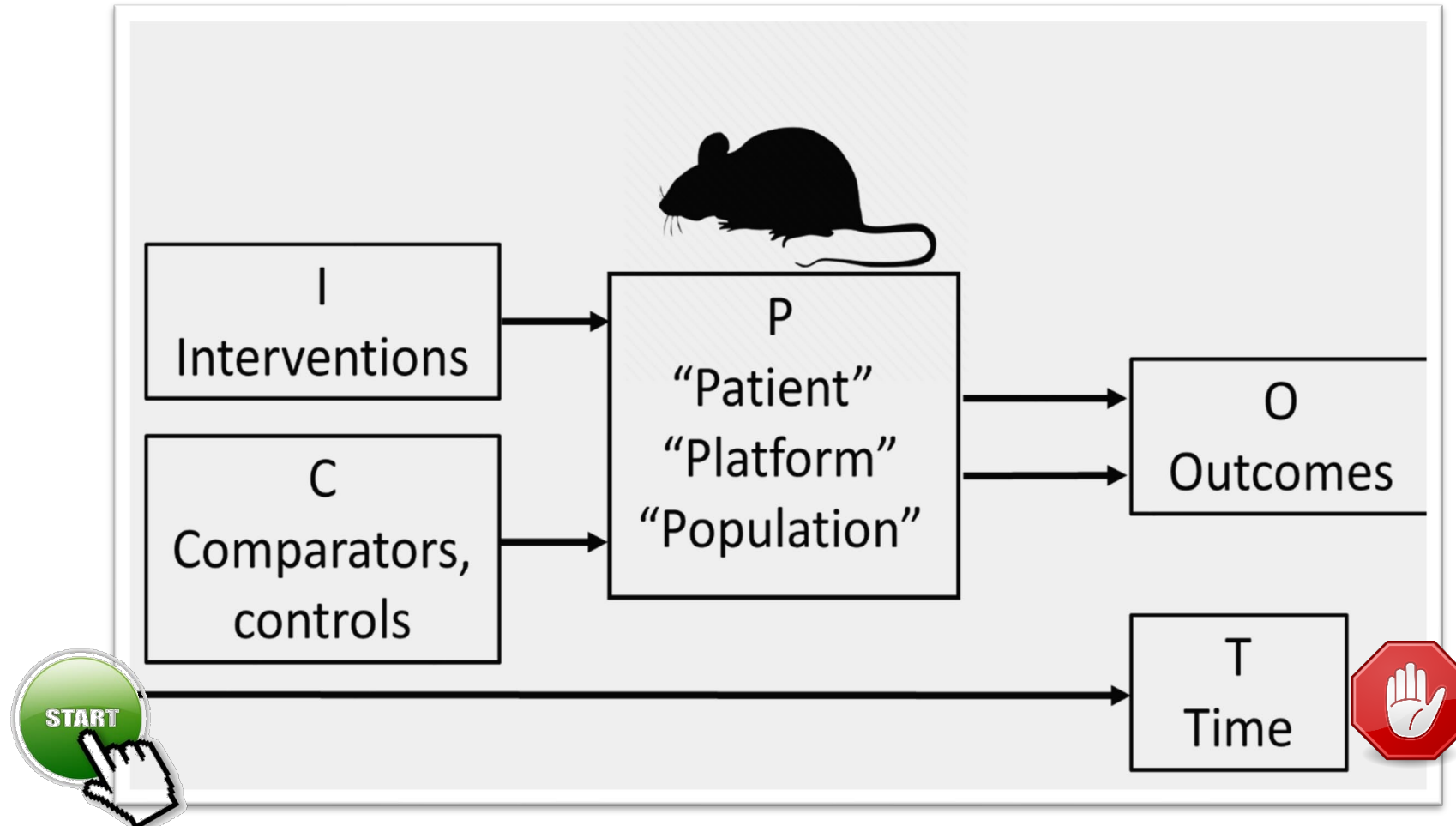
T – **T**imeline







Think of it as a system diagram

Shows how the investigator plans to investigate *cause and effect*
'What does what to what and what happens'?






P = Animal model

-  1. Are animals needed at all? (e.g., training)
-  2. Platform: Why that animal?

DON'T say “lowest phylogenetic representative” “least sentient”

DO scientifically justify model choice

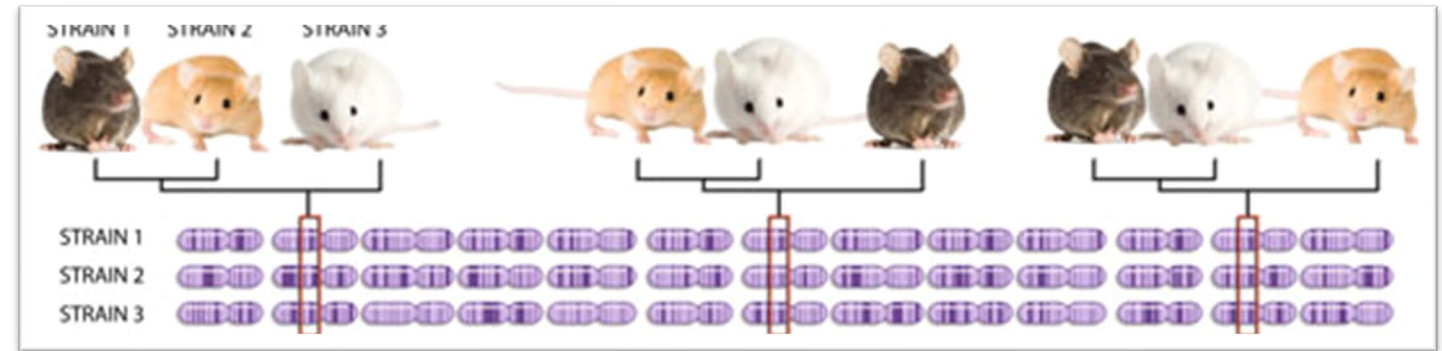
- most appropriate for testing the scientific hypotheses → anatomy, physiology, behavior, genome
- postulated targeted effects or mechanism of action
- assessment of competing models of the disease condition, advantages and disadvantages
-  • Test condition :Spontaneous or induced?



Model choice affects results and interpretation

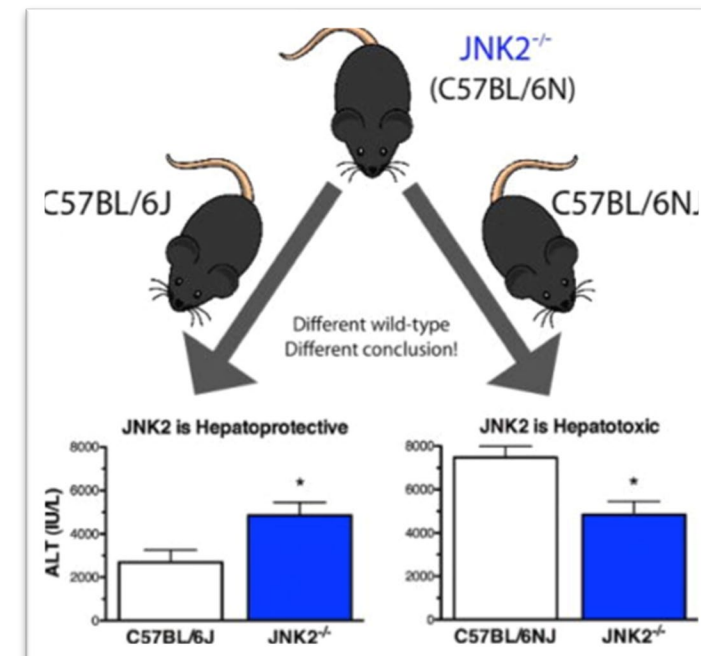
Poor choice of model results in

- Inappropriate techniques
- Misleading conclusions
- Translation issues
- Welfare issues



No such thing as a C57BL/6 mouse!

Conclusions differ depending on control strain selection.





I = Interventions

- “Treatment” administered to study subjects
- What does the researcher plan to do to/for the subject?
- Specific test, procedure, therapy, drug, etc.

Identify:

- What is it? Agent? Drug? Procedures?
- How much?
- How often?



C = Comparators and controls

- Used to establish cause-and-effect between interventions and responses
- Controls are the method to distinguish experimental 'signal' from 'noise.'

7 types of controls:

1. Positive
2. Negative
3. Standard of care
4. Sham
5. Vehicle
6. Matched
7. Strain




C = Comparators and controls

Look for unknown, inappropriate, and/or redundant control groups

- 🚩 **Historical:** Not valid for hypothesis tests against new test data
 - Cannot be randomized
 - Valid only if they meet stringent 'Pocock criteria': all experimental conditions must be identical
- 🚩 **Shams:** Animals left untreated after acute, well-established procedures with highly predictable experimental endpoints, spontaneous recovery is not possible
- 🚩 **Redundant:** Many one-factor-at-a-time (OFAT) two-group trials, each with its own 'control' → inefficient, low precision, cannot detect interactions



O = Outcomes

- The outcomes are the answers to research questions = measured responses
 - Drives sample size calculations
 - Determine the methods!
-  • Must be relevant, clearly-defined, specific, measurable

3 criteria:

1. *What* is being measured?
2. *How* will it be measured?
3. *How often* will it be measured?



Outcomes: What, how, when, how often?

What to watch out for:

1. Vague, unmeasurable, undefined metrics
 - “Survival” “Function” “Protection” → No meaning
2. Vague, unmeasurable, undefined success criteria: How do will you know if the experiment “worked”
 - “Better” “improved” → No standard of comparison
3. Technical methods: Invasive/noninvasive? Excessive instrumentation? Multiple surgeries? Excessive sampling?
4. 3Rs considerations: Pain, distress? Alleviation/welfare plans in place?

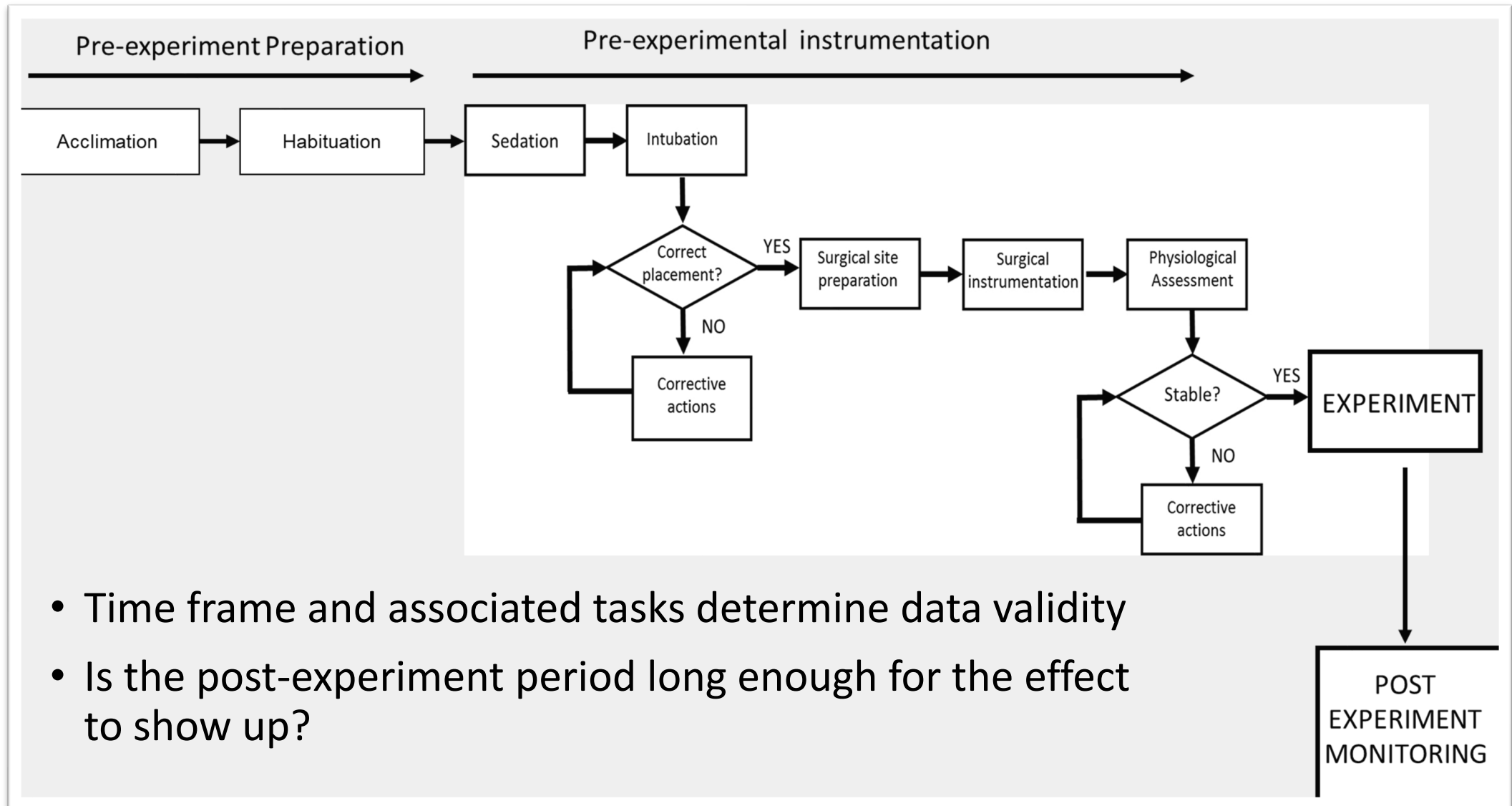
Humane endpoints



- Choice and measurement of outcome variables AND time frame depend on humane endpoints
- *Welfare indicators* = one or more behavioural and physical criteria that determine the point at which the animal will be removed from the study
- Ethical mandate to minimise pain, suffering, and distress of animals used in experiments.
 - Predetermined
 - Objective
 - Easily-recognised
 - Adhered to
- Consult **published guidelines** for specific research models:
 - Oncology: Workman *et al* 2010
 - Ischaemia: Percie du Sert *et al* 2017
 - Sepsis: Zingarelli *et al* 2019



T = Experiment time frame



- Time frame and associated tasks determine data validity
- Is the post-experiment period long enough for the effect to show up?



Example: Design skeleton

An investigator wanted to study effects of a new drug thought to improve motor function in rats with osteoarthritis. Measurements were to be made once a week for 3 weeks.

P	Study population	Rats with osteoarthritis	[Animal justification, model creation]
I	Intervention	New drug	[Justification, formulation, administration]
C	Comparator	?	[Details]
O	Outcome	Motor function	[?, details]
T	Time frame	3 weeks	[Justification, number of samples, endpoints]

If any elements or details missing, **STOP!**



Example: What is the outcome?

Motor “function”: how is it defined?

Several metrics:

- Step length (cm) → Non-invasive, animal could be measured several times
- Peak tetanic force of isolated soleus muscle (N) → invasive/terminal

Include all materials and methods in the description.



Screening animal numbers: Why does it matter?

Animal numbers



What are “justifiable” numbers?

Power calculations are the gold standard, but often misapplied

- Often incorrect or unverifiable
- Often gamed
 - To obtain a favorite number
 - Tick a box: “Power calculations are a necessary evil to satisfy the ethical oversight committee and reviewers”

- Fitzpatrick *et al Nature-Lab Animal* 2018. 47:175

- Ignores logistic constraints






Instead consider 'Right-sizing' the experiment

Sample size justification means that numbers are *statistically, operationally, ethically* justifiable.

The numbers requested must:

- Be sufficient to address the research question
- Ensure animals are not wasted

3 criteria:

-  • Feasible?
-  • Verifiable?
-  • Ethical?

Feasible?



Do the numbers match lab capability?

Personnel, resources, space, time,
budget...

OR ARE THEY

 **Absurd, Preposterous**

- Usual requests: thousands, tens of thousands, hundreds of thousands
- My favourites
 - “one million” mice
 - 91,386,777 mice

What this communicates: NO planning, mice as disposable “furry test-tubes”



Verifiable

1. Do they show their work?
2. Do the (simple) maths:
 - How many experiments?
 - How many groups?
 - How many animals per group?
 - Attrition? Expected losses? (researchers must include loss mitigation plans?)

e.g., 2 drugs x 3 dose levels = 6



These statements cannot be verified...



- “Based on previous publications”
- “In our experience....”
- “This number is sufficient to obtain statistically-significant results”
- “Unforeseen problems might happen”
- “It is unknown how many animals we will require because this is an exploratory study”
- “What everyone else does”; “it is the industry standard”

Take home 2: Sample size justification is bespoke, not boiler-plated!
Numbers must be aligned & customized to the actual study

Ethical?



1. 3Rs plans? Can numbers be reduced? Better designs, methods?
Welfare checks, mitigation plans?
2. Collateral losses? Check end-use/disposition
Example: In-house breeding supply:
 - PI requested 400 mice of a desired genotype
 - But ~1200 “unwanted” genotypes to be euthanised without use



Example: Conventional vs. statistical study design





Trial of vaccine efficacy in mice

Investigators proposed a series of experiments designed as a series of two-group comparisons or “one-way ANOVAs” on 6 strains, 3 doses, 3 dosing intervals, 3 age classes, both sexes.

- 5 mice per group “because that was necessary for statistical significance,” “what everyone else does”
- They requested 1,620 mice for 324 experimental ‘groups,’ 5 animals/group
- Strain x dose x interval x age x sex = $6 \times (3 \times 3 \times 3 \times 2) = 324 \times 5 = 1,620$



Why is this a bad “design”?

-  1. Unmanageable → large undetectable sources of variation swamp true experimental signals
-  2. Likely to miss important results
 - Most important drivers
 - Synergisms or interactions → where most discoveries occur
 - Optimum ‘best’ response
-  3. Probably unfeasible
 - Can the numbers requested be processed given lab capabilities?
 - How long would it take to process 1,600+ animals?
-  4. Almost certainly wasteful
 - Animals age out of the study before they can be used
 - Discarding of non-statistically significant results



Compare with appropriately-designed screening experiments

Run	Block	Dose	Interval	Age	Sex
1	1	-1	-1	1	female
2	1	-1	-1	1	male
3	1	0	1	1	female
4	1	0	-1	-1	male
5	1	-1	1	0	male
6	1	1	-1	0	female
7	1	0	0	0	male
8	1	1	1	-1	male
9	1	1	1	1	female
10	2	1	1	-1	female
11	2	-1	0	-1	male
12	2	0	0	0	female
13	2	1	-1	1	male
14	2	-1	-1	-1	male
15	2	-1	1	1	male
16	2	-1	1	-1	female
17	2	1	0	1	female
18	2	1	-1	-1	female

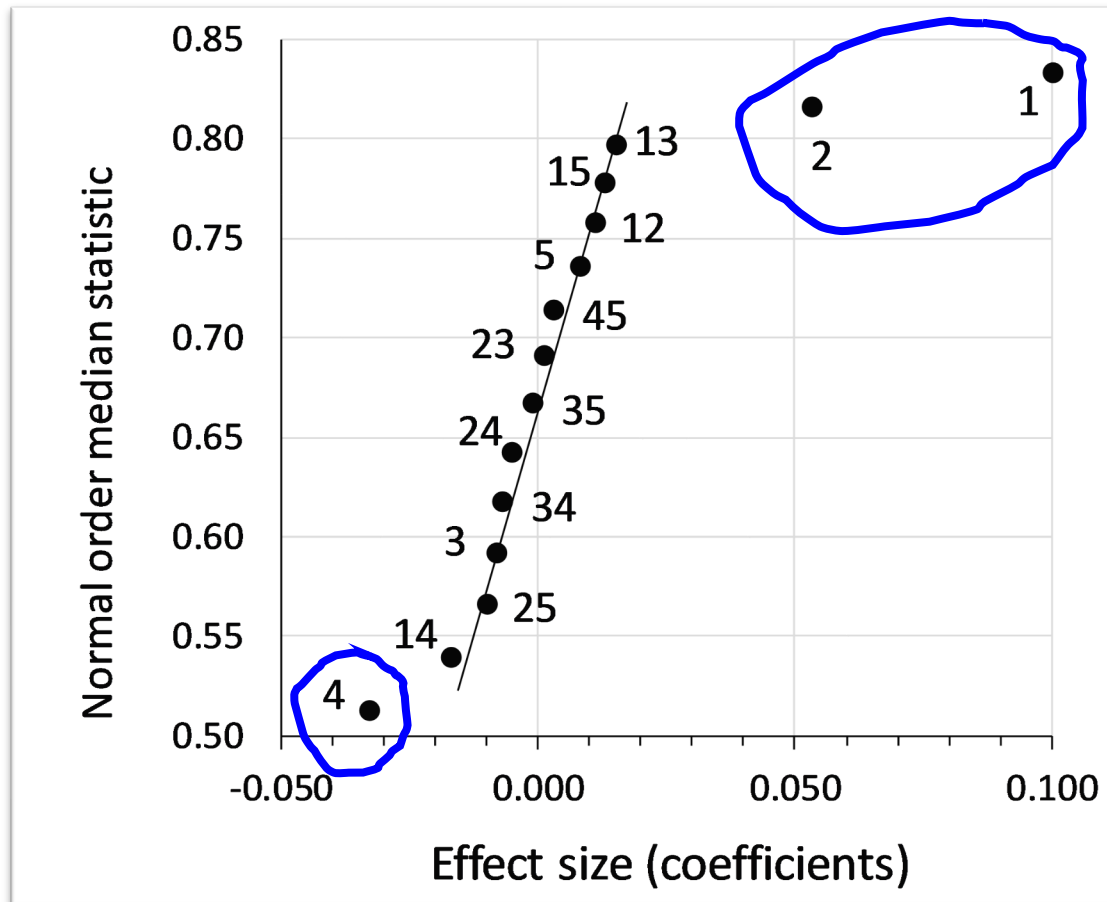
Design generated in SAS JMP Pro 16.

Each experiment is run for each strain separately

- Two blocks (controls day-to-day variation)
- Factors dose, interval, age, sex
- 18 runs performed in random order = 9 males + 9 females = 18 mice/strain
- Centre points (0,0,0) provide variance estimates for assessing significance of main effects
- 6 strains x 18 runs = 108 mice



Data visualization identifies most important effects



- Plot regression coefficients on a half-normal plot
- Identify the most important factors for each strain
- Design a new experiment with only the most important factors



Summary



Take homes

1. Appropriate statistically-based study designs are fundamental tools for:
 - 3Rs
 - Translation
2. Statistics is a process, not a “thing”
 - *Design* before inference
 - *Data quality* before analysis
3. Three design basics:
 - Bias minimization
 - Design skeleton: 5 elements
 - Numbers: 3 screens

For both reviewers & researchers



Advocates for better science = better animal welfare

Promote CULTURE CHANGE

1. Become familiar with best-practice standards (e.g., ARRIVE 2.0)
2. Facilitate & promote best research practices (statistical DOE)
3. Evaluate protocols and publications in line with best-practice standards
4. Continuing education in *evolving* best scientific and welfare practices



Parting thoughts

- Poorly-designed experiments and misuse of statistics are irresponsible, negligent, and unethical
 - Animals suffer & are killed
 - Animals wasted in non-informative studies
 - Humans are harmed and killed

It is **duty of care** to minimize harms & promote best-scientific practice.

Altman D. *BMJ* 281:1182-4, 1980

MacCallam, C.J. *PLoS Biol* 8, e1000413, 2010

Brønstad, A., et al *Laboratory Animals* 50, 1-20, 2016



Questions?
Thank you for
your attention

Question 1:

What practical suggestions do you have for ways that IACUCs can positively impact the design of experiments in the initial stages (before the grant proposal) to include more robust statistical approaches without overstepping or micromanaging?

What institutional processes or outreach have you seen that have been effective?

Question 2:

- Can you provide some examples of when your IACUC has felt the need to reach out to the PI regarding study design and animal numbers when reviewing a protocol?
- How have you broached these conversations in a productive way?
- What was the outcome of these conversations?

Question 3

For OLAW: There is a lot of overlap between the IACUC's role in evaluation of the science and the SRG. What balance should IACUCs strike when evaluating the statistics behind animal numbers requested in protocols, especially if the protocol is submitted to the IACUC at the just-in-time (JIT) stage?

Question 3

Answer: See OLAW FAQ D12:

<https://olaw.nih.gov/faqs#/guidance/faqs?anchor=50327>

- IACUCs not expected to conduct peer review of research proposals, **BUT** are expected to consider U.S. Government Principles.
- Principle II: evaluation of the relevance of a procedure to human or animal health, the advancement of knowledge, or the good of society.
- Other PHS Policy review criteria refer to sound research design, rationale for involving animals, and scientifically valuable research.

Question 3

- Presumably, a study that could not meet these basic criteria is inherently unnecessary and wasteful and, therefore, not justifiable.
- The primary focus of the SRG is scientific merit and the primary focus of the IACUC is animal welfare, but they overlap!
- SRGs may raise concerns about animal welfare and IACUCs may question the scientific rationale or necessity for a procedure.

Next Webinar: Summer 2023

Topic TBD



National Institutes of Health
Office of Laboratory Animal Welfare