

The ARRIVE 2.0 Essential 10: What they mean for NIH-sponsored research

Penny S Reynolds, PhD

Department of Anesthesiology, College of Medicine

Department of Small Animal Clinical Sciences, College of Veterinary Medicine

University of Florida

Statistics in Anesthesiology Research (STAR) Core

IACUC member, University of Florida

ARRIVE 2.0 International Working Group member

Outline: ARRIVE 2.0 Essential 10

- I. Why the Essential 10 (and why ARRIVE 2.0)?
- II. What are the Essential 10?
- III. Making the Essential 10 work for you



I. Why the Essential 10 (And why ARRIVE 2.0)?



NEW!! NOT-OD-23-057

10 Feb 2023

NIH Encourages the Use of the ARRIVE Essential 10 Checklist in all Publications Reporting on the Results of Vertebrate Animal and Cephalopod Research

Notice Number:

NOT-OD-23-057

Key Dates

Release Date:

February 10, 2023

<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-057.html>



What are the ARRIVE guidelines?



Animal **R**esearch: **R**eporting of **I**n **V**ivo **E**xperiments

International consensus best-practice reporting guidelines

Goal: Improve reporting of animal-based research

Shifts emphasis from *'sexy' results* to *rigorous methods*



ACD-NIH recommendations 11 June 2021

ACD WORKING GROUP ON
ENHANCING RIGOR,
TRANSPARENCY, AND
TRANSLATABILITY IN ANIMAL
RESEARCH

FINAL REPORT
June 11, 2021

When ARRIVE 2.0 should be followed:

Writing stage

AND

Entire research process

“Strengthening these elements **across the life of a study, from planning to execution and publication**, will result in a higher-quality knowledge base and will better inform future research.”

Main theme

“Good science must not only be done, it must be seen to be done”

High quality science is **valid** and **reliable**.

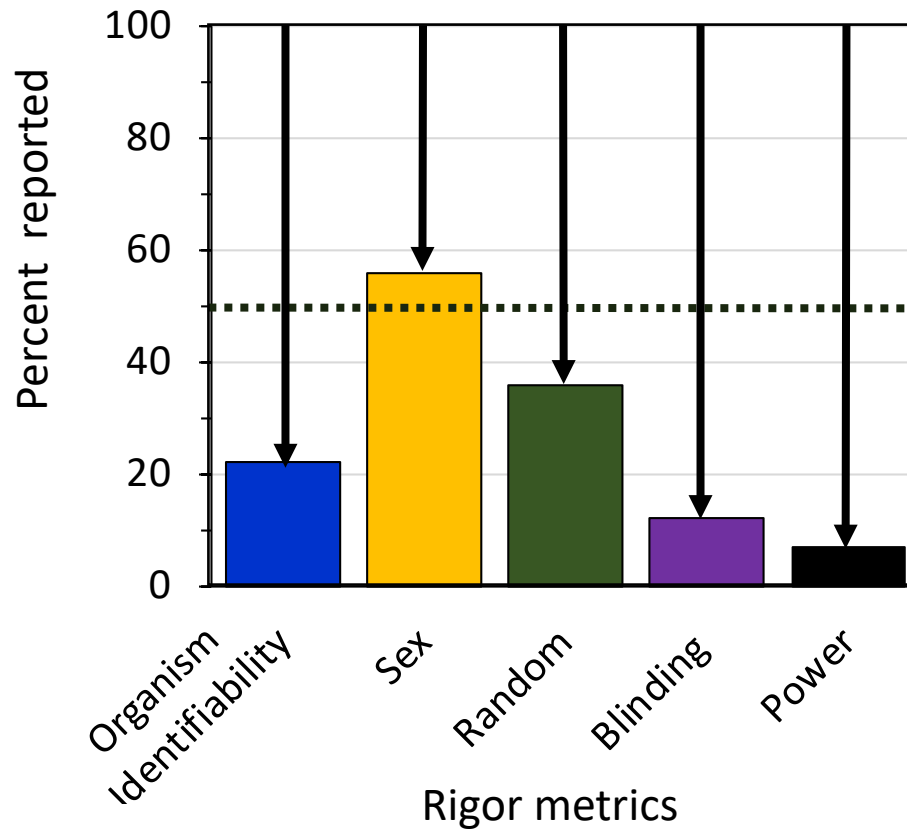
Validity and reliability are determined by **sound methodology**

- Good experimental design
- Bias minimization
- Appropriate statistical methods
- Transparent reporting

Why is it necessary?

The overwhelming majority of papers do not report basic metrics

Data from 51,312 animal-based studies, 2018



Reporting performance gaps →

Problems with reproducibility

Menke J *et al.* 2020 *iScience* 23(11): 101698

Reproducibility issues in preclinical research

Proxy:

Poor reporting indicates **poor experimental design and conduct**

Macleod *et al.* 2015 *PLoS Biol* 13(10): e1002273

Complacency:

Methodological illiteracy is the norm

van Calster *et al.* 2021. *J Clin Epidemiol* 138:219–226

Broken checks & balances:

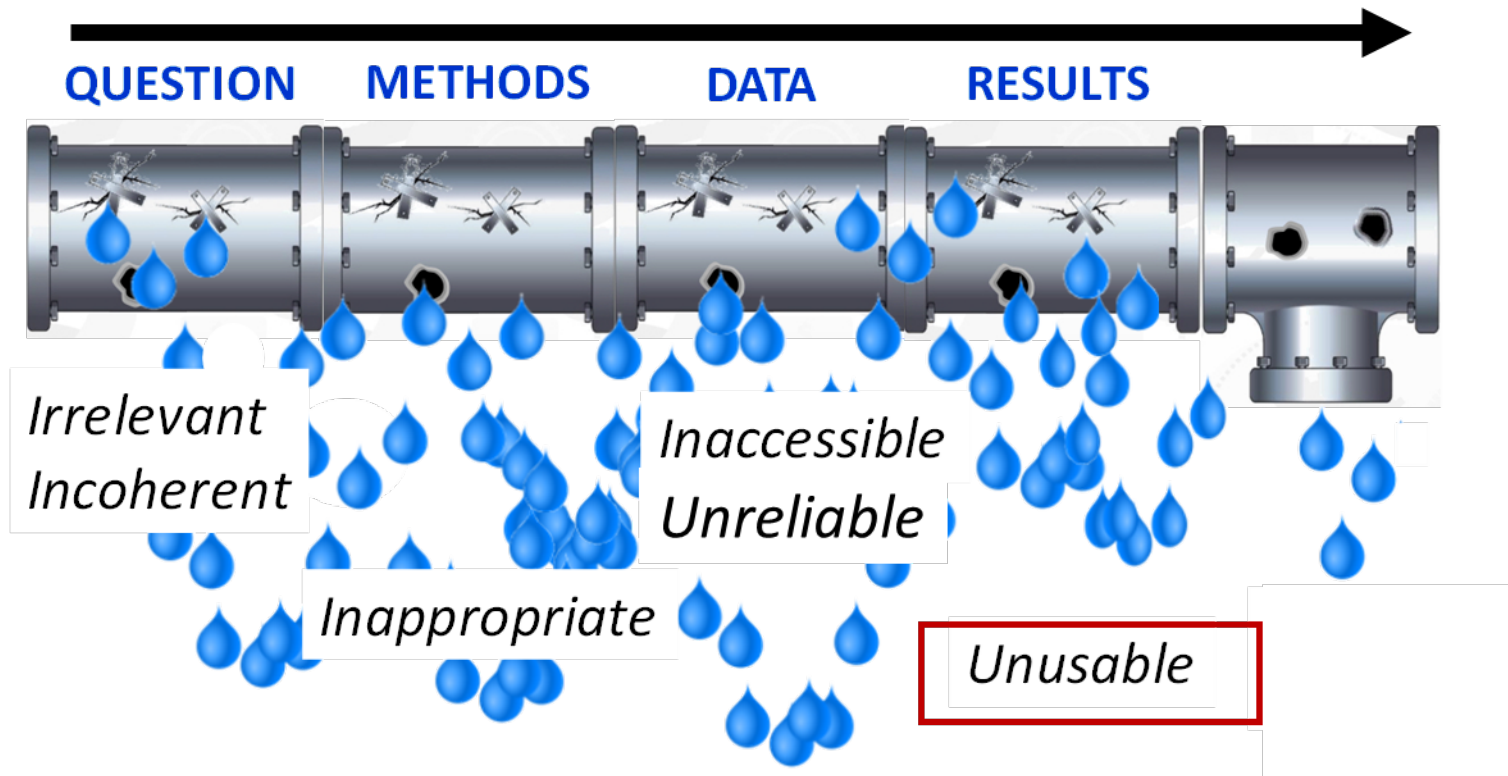
Editorial processes & peer review are not stringent, sufficient, or enforced

Moher *et al.* *BMC Medicine* (2015) 13:34

Hair *et al.* *Res Integr Peer Rev.* 2019; 4: 12. doi: 10.1186/s41073-019-0069-3

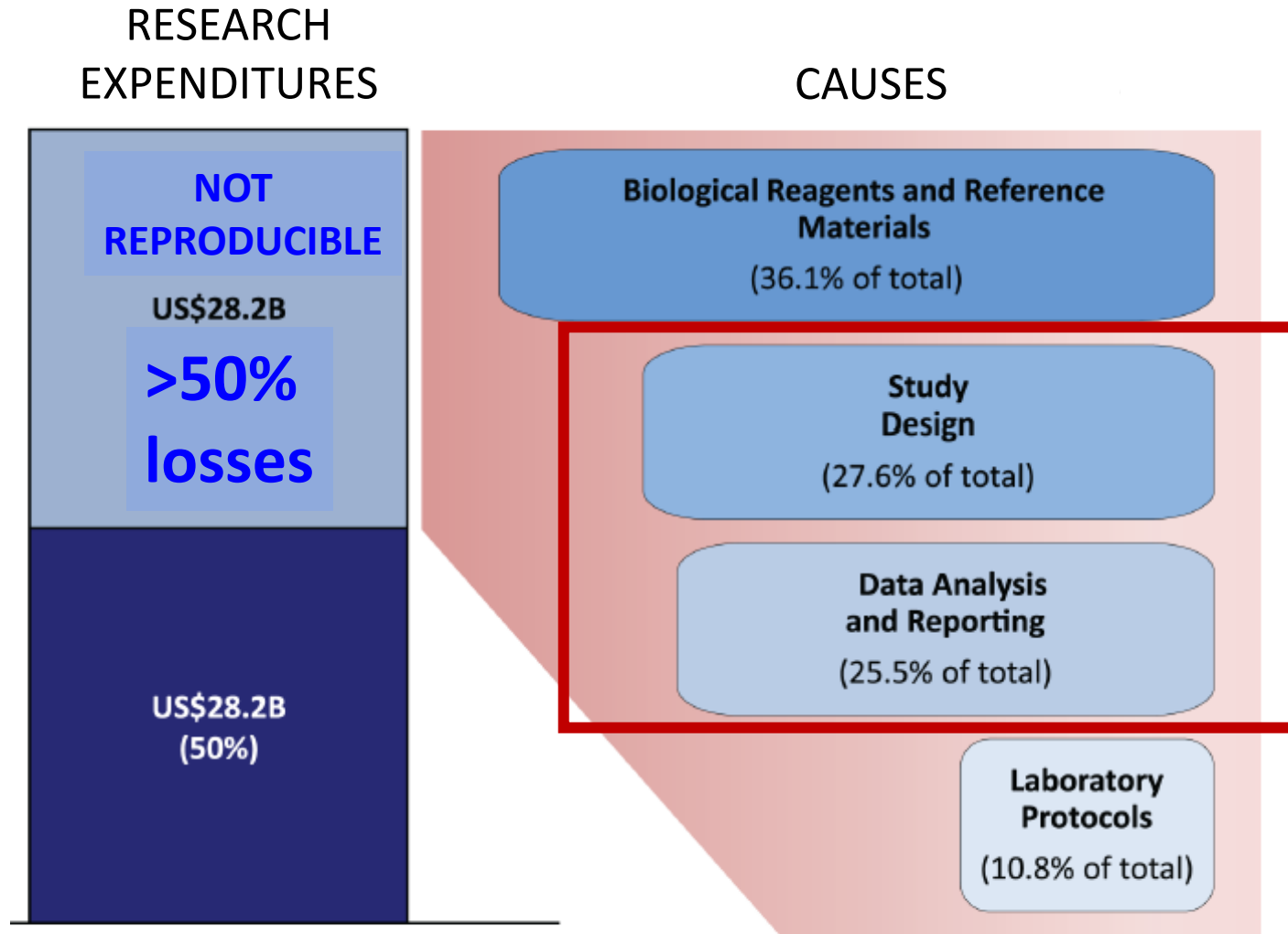
Why should we care?

Poor-quality research is wasted research



Chalmers et al *Lancet* January 8, 2014

1. Wasted research is a major loss of investment



At least 50% preclinical research **in the USA alone**

is not reproducible

Mostly due to

Poor

Study design,

Data analysis,

Reporting

2. Wasted research is an ethical issue

Collateral costs

- Hundreds of millions (billions) of animals wasted
- Thousands of humans injured and killed

Nearly all ALS drugs

Scott et al 2008. *Amyotroph Lat Scler.* 9: 4–15;
Nature Genetics 2012:611

Sildenafil for fetal growth restriction Symonds & Budge. 2018. *BMJ* 362:k4007

TB vaccine

Macleod. 2018. *BMJ* 360:k66

Alzheimer's & amyloid- β

Science 2022. 377(6604):358-363.

Enter ARRIVE



2009



National Centre
for the Replacement
Refinement & Reduction
of Animals in Research



National Institutes of Health
Office of Laboratory Animal Welfare

NC3Rs UK (funding from NIH/OLAW)

Systematic survey and review of published,
government-funded preclinical research

First ARRIVE guidelines 2010

Kilkenny et al 2009 *PloS ONE* 4(11): e7824. doi:10.1371/journal.pone.0007824



Enter ARRIVE

2017



National Centre
for the Replacement
Refinement & Reduction
of Animals in Research

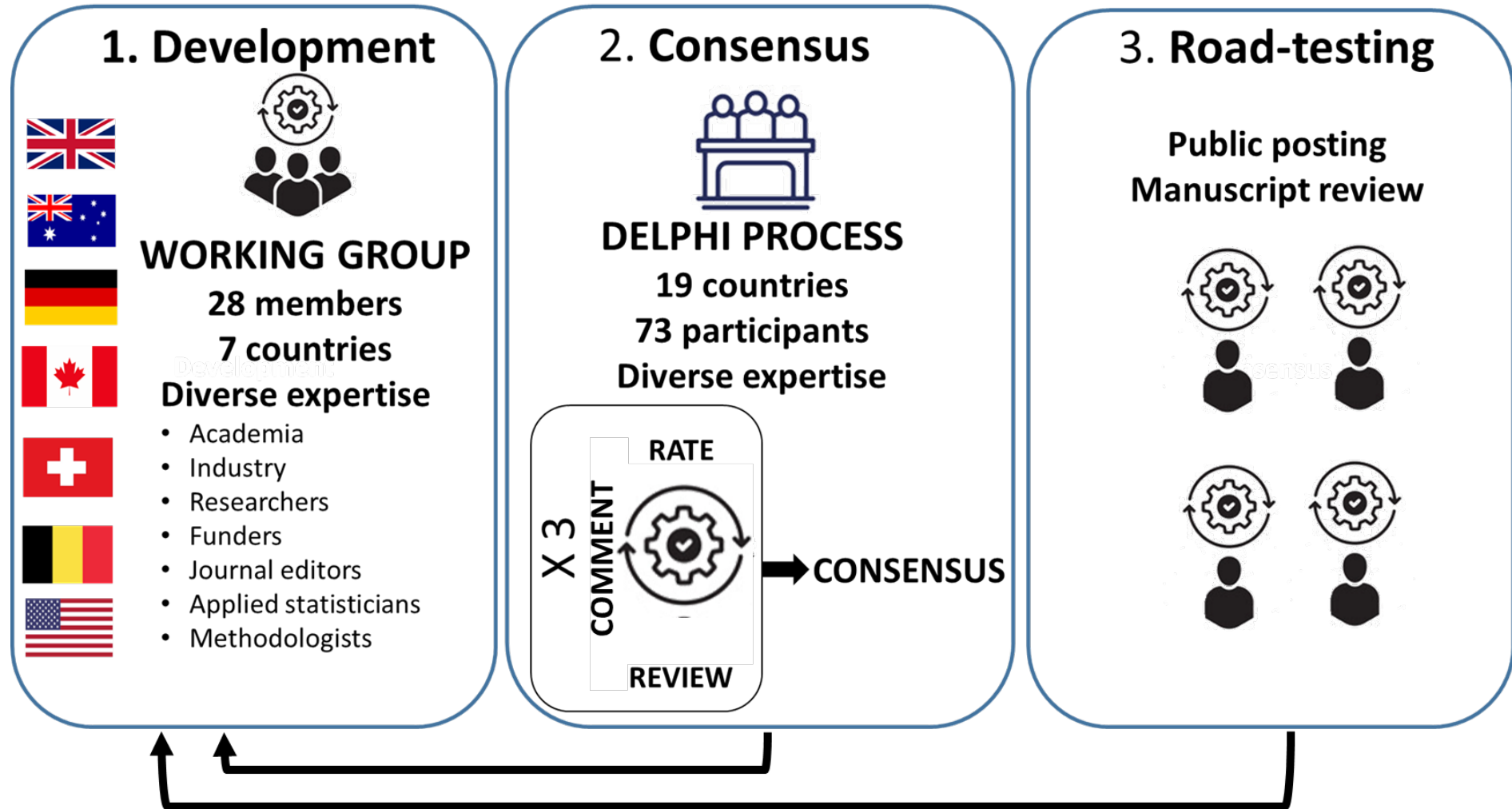


Second International Working Group ARRIVE 2.0

Goals:

- Revise, update, streamline
- Improve **implementation** by improving **utility**
- Multi-year international collaborative effort
- Rigorous methodology

I2C2: International, iterative, collaborative, consensus-driven



Product

<https://www.ARRIVEguidelines.org>

1. ARRIVE 2.0 2020 Checklist

Two-tiered

- ESSENTIAL 10:
 - Minimum information required for assessing rigour and reproducibility → INTERNAL VALIDITY
- RECOMMENDED 11:
 - Information required for assessing study-specific context → GENERALIZABILITY

2. ARRIVE 2.0 Explanation & Elaboration document

- User's manual

II What are the 'Essential 10' ?



'Essential Ten' reproducibility items

Expert consensus on priority information

Universal best-practice items for
Reliability,
Validity,
Reproducibility

Not *rank* order but *workflow* order

<https://arriveguidelines.org/arrive-guidelines>

ARRIVE The ARRIVE guidelines 2.0: author checklist

The ARRIVE Essential 10

These items are the basic minimum to include in a manuscript. Without this information, readers and reviewers cannot assess the reliability of the findings.

Item	Recommendation	Section/line number, or reason for not reporting
Study design	1 For each experiment, provide brief details of study design including: a. The groups being compared, including control groups. If no control group has been used, the rationale should be stated. b. The experimental unit (e.g. a single animal, litter, or cage of animals).	
Sample size	2 a. Specify the exact number of experimental units allocated to each group, and the total number in each experiment. Also indicate the total number of animals used. b. Explain how the sample size was decided. Provide details of any <i>a priori</i> sample size calculation, if done.	
Inclusion and exclusion criteria	3 a. Describe any criteria used for including and excluding animals (or experimental units) during the experiment, and data points during the analysis. Specify if these criteria were established <i>a priori</i> . If no criteria were set, state this explicitly. b. For each experimental group, report any animals, experimental units or data points not included in the analysis and explain why. If there were no exclusions, state so. c. For each analysis, report the exact value of <i>n</i> in each experimental group.	
Randomisation	4 a. State whether randomisation was used to allocate experimental units to control and treatment groups. If done, provide the method used to generate the randomisation sequence. b. Describe the strategy used to minimise potential confounders such as the order of treatments and measurements, or animal/cage location. If confounders were not controlled, state this explicitly.	
Blinding	5 Describe who was aware of the group allocation at the different stages of the experiment (during the allocation, the conduct of the experiment, the outcome assessment, and the data analysis).	
Outcome measures	6 a. Clearly define all outcome measures assessed (e.g. cell death, molecular markers, or behavioural changes). b. For hypothesis-testing studies, specify the primary outcome measure, i.e. the outcome measure that was used to determine the sample size.	
Statistical methods	7 a. Provide details of the statistical methods used for each analysis, including software used. b. Describe any methods used to assess whether the data met the assumptions of the statistical approach, and what was done if the assumptions were not met.	
Experimental animals	8 a. Provide species-appropriate details of the animals used, including species, strain and substrain, sex, age or developmental stage, and, if relevant, weight. b. Provide further relevant information on the provenance of animals, health/immune status, genetic modification status, genotype, and any previous procedures.	
Experimental procedures	9 For each experimental group, including controls, describe the procedures in enough detail to allow others to replicate them, including: a. What was done, how it was done and what was used. b. When and how often. c. Where (including detail of any acclimatisation periods). d. Why (provide rationale for procedures).	
Results	10 For each experiment conducted, including independent replications, report: a. Summary/descriptive statistics for each experimental group, with a measure of variability where applicable (e.g. mean and SD, or median and range). b. If applicable, the effect size with a confidence interval.	

1: Study design

Study design

1

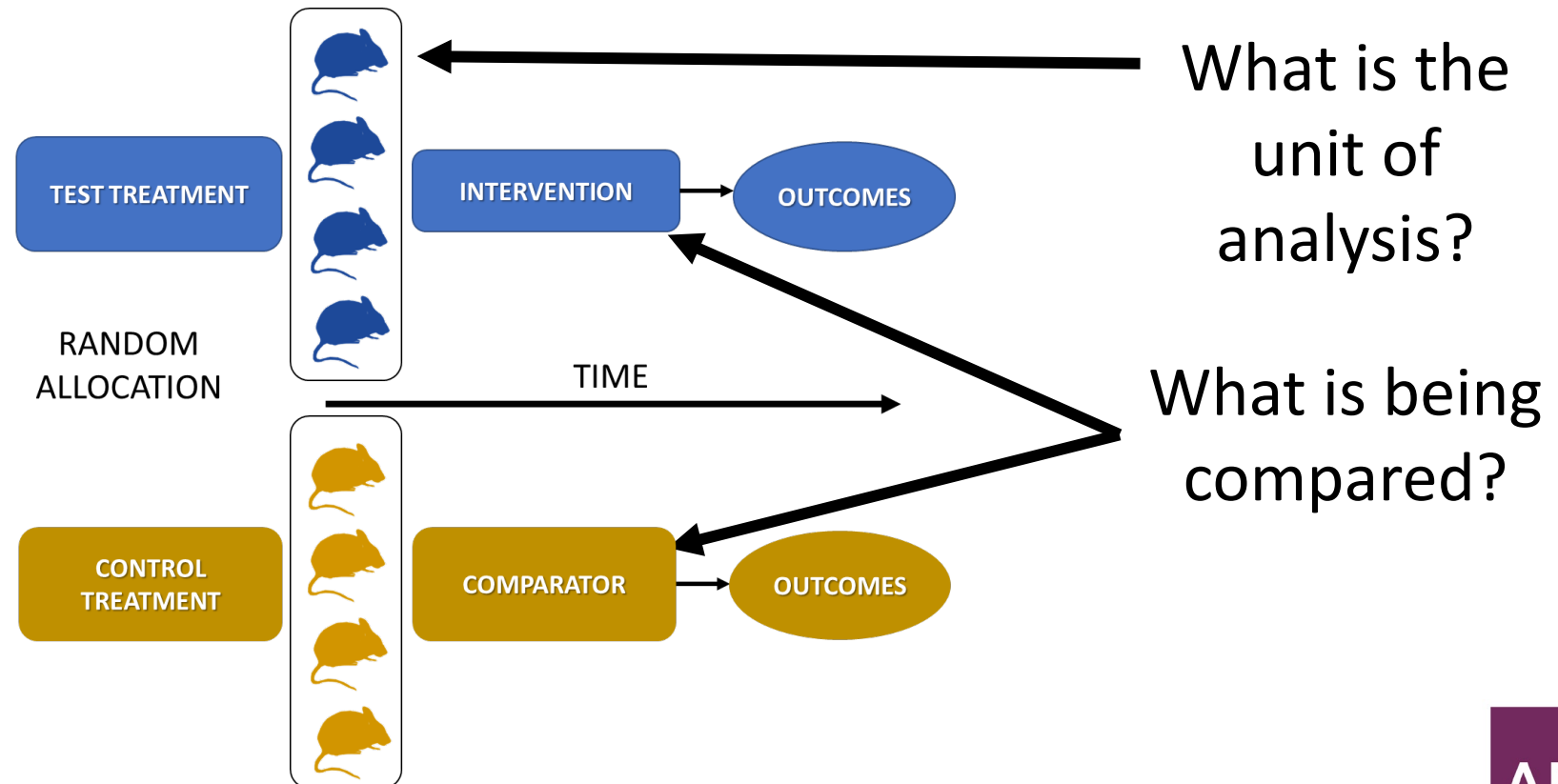
For each experiment, provide brief details of *study design* including:

- a. The **groups** being compared, including **control** groups. If no control group has been used, the **rationale** should be stated.
- b. The **experimental unit** (e.g. a single animal, litter, or cage of animals).

What it means

What are you comparing? = Formal statistical structuring of predictor variables (test & control factors)

What is the unit of analysis? = Experimental unit



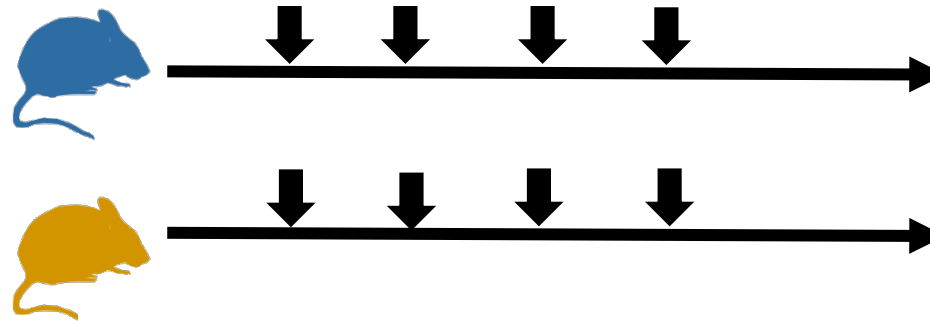
Key idea: Statistically-based designs control variation

Randomized complete block



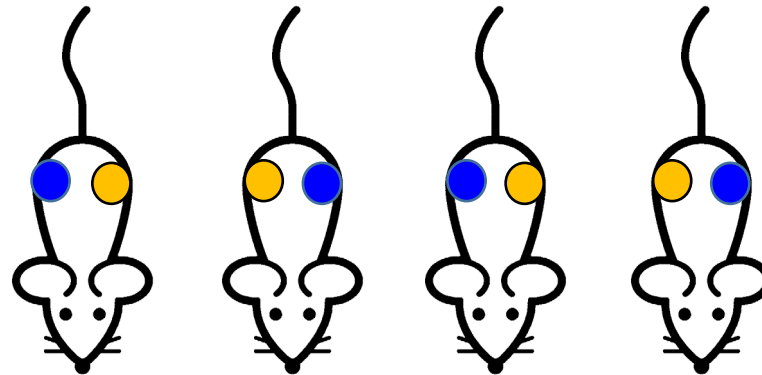
External source of variation

Repeated measures



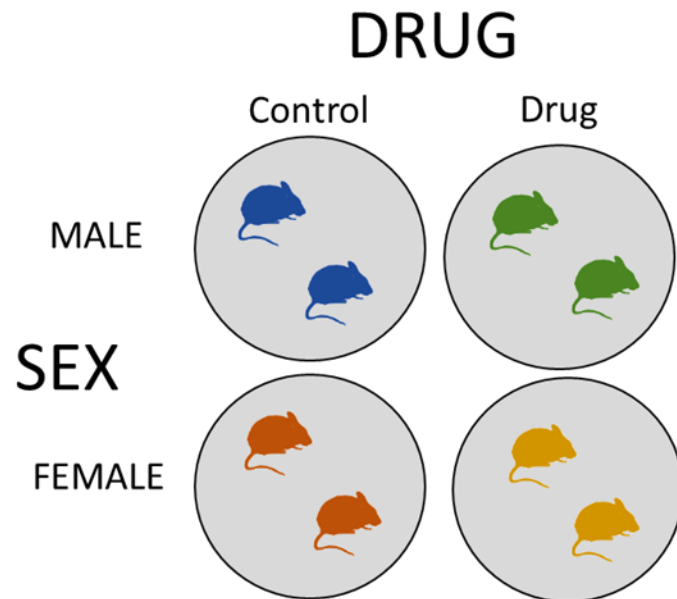
Source of variation between and within animals

Paired design



Factorial design are the best for multiple inputs

2 x 2 factorial with two replicates



DRUG	SEX	RUN
CONTROL	F	1
ACTIVE	F	2
CONTROL	M	3
ACTIVE	M	4



- Allows simultaneous evaluation of multiple input variables
- Identifies interactions
- Discriminates informative from non-informative inputs
- Shares N across multiple input variables

Results

- Saves time
- Saves \$\$\$
- Animal-sparing

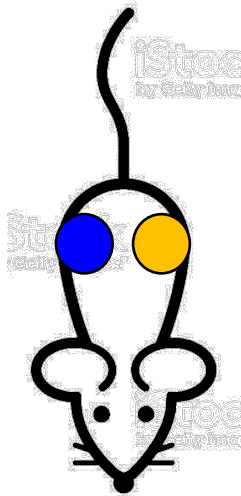
What is an experimental unit (EU)?

Experimental unit =

"smallest division of experimental material such that any two units receive different treatments in the actual experiment"

Paired design

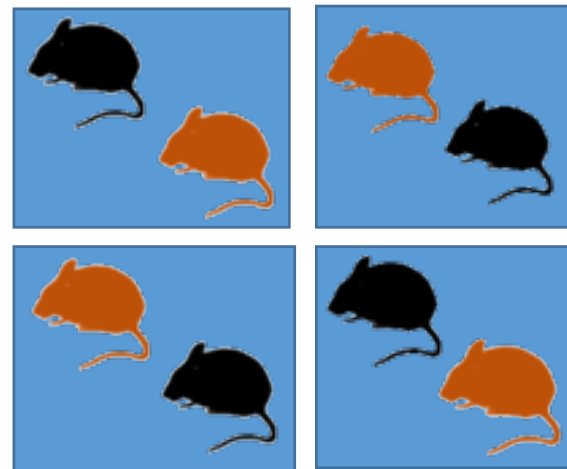
Blocking on mouse



EU = Flank

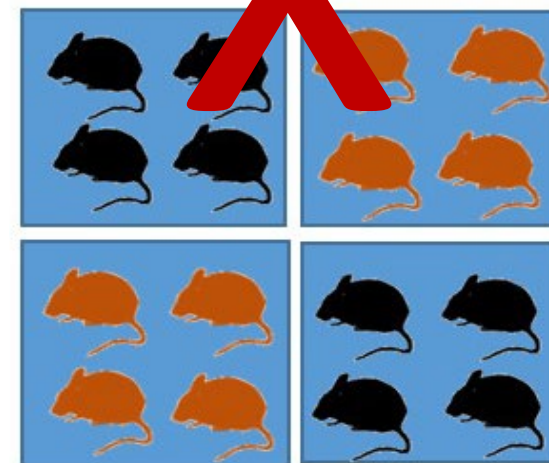
RCBD

Blocking on cage



EU = mouse
n = 4/group

~~Pseudo replication~~



EU = cage
n = 2/group (NOT
8/group)

Why it matters

1. “Backbone of good research”

Design cannot be imposed after data are collected!

The study design determines

- what and how data are collected,
- statistical analyses
- interpretation of the results.



Increases statistical power
Reduces noise
Increases information power



Reduces animal numbers

Why it matters

2. Poor understanding of design is the major limitation to research quality reform

Statistically-based designs have been available for
over a century

BUT

**MOST studies have NO formal design
Methods seldom taught**

Why it matters

3. Undesigned studies are grossly inefficient and wasteful

Vast majority of studies “organized” by “groups” or “cohorts”

- ***Statistically:*** Statistical methods misused

Miss true signals

- ***Logistically:*** Waste animals

Altman DG. Misuse of statistics is unethical. *BMJ* 281: 1182-1184, 1980

2. Sample size

Sample size

2

- a. Specify the **exact number of experimental units** allocated to each group, and the **total number in each experiment**. Also indicate the **total number of animals** used.
- b. Explain how the **sample size was decided**. Provide details of any *a priori* sample size calculation, if done.

What it means

Sample size = number of experimental units per group.

1. Numbers reporting:

Numbers through the study need to add up
Track attrition

2. Numbers justification

Are they adequate to answer the research question?

Are numbers

- ✓ Feasible?
- ✓ Verifiable?
- ✓ Ethical?

Reynolds *Nature-Lab Animal* 2021. 50: 263-271

Why it matters

NUMBER ONE REPRODUCIBILITY item

Vollert et al. *BMJ Open Science* 2020

NUMBER ONE ETHICAL principle

Three Rs = 'Minimal harm for maximum scientific value'

Too large a sample size wastes excess animals

Under-powered studies waste all animals

Majority of studies (>95%) do not either justify or report numbers in protocols or publications

These are not justifications

1. 'Magic' numbers that 'worked' in previous studies

".... based on our previous publications"

"In our experience this number is sufficient to obtain statistically significant results"

"What everyone else does"

2 . Making it up

"Unforeseen problems might happen"

"It is unknown how many animals we will require because this is an exploratory study"

[Personal favorite: 91,386,777 mice for 3-year project]

3. Passive-aggressive

"Power calculations are a necessary evil to satisfy the ethical oversight committee and reviewers" Fitzpatrick et al *Lab Animal* 2018. 47:175

3: Inclusion and exclusion criteria

Inclusion and exclusion criteria

3

- a. Describe any **criteria used for including and excluding animals** (or experimental units) during the experiment, and data points during the analysis. **Specify if these criteria were established a priori.** If no criteria were set, state this explicitly.
- b. For each experimental group, report any animals, experimental units, or data points **not included in the analysis** and **explain why**. If there were no exclusions, state so.
- c. For each analysis, **report the exact value** of n in each experimental group

What it means

Consistent, *a priori*, criteria for including or disqualifying animals and their data

Inclusion criteria = key features of the target population used to answer the research question

Exclusion criteria = features interfering with study goals (e.g. sick, failed instrumentation)

Why it matters

Effective criteria

Define the **subject pool** for obtaining the best data.

Clear & consistently **defined** study population

→ Representative

Minimizes bias resulting from arbitrary decisions

Reduces temptation to cherry-pick data & results (dishonest, unethical → research misconduct)

4. Randomisation

Randomisation 4

- a. State **whether randomisation** was used to allocate experimental units to control and treatment groups. If done, provide the **method** used to generate the randomisation sequence.
- b. Describe the strategy used to **minimise potential confounders** such as the order of treatments and measurements, or animal/cage location. If confounders were not controlled, **state this explicitly**

What it means

Formal, technical, probabilistic process of assigning interventions to experimental units & order of processing

NOT “haphazard”, “ad hoc”, “unplanned”

- Computer algorithms best practice method: unbiased, provides audit trail
- Specify method and algorithm used.

Why it matters

Randomisation is the NUMBER ONE VALIDITY item

- Minimises **systematic bias**
- AND**
- Ensures **validity of inferential tests**

If randomisation is NOT performed, your statistical hypothesis tests are **INVALID**

There are no good reasons not to randomize

Blinding

Blinding

5

Describe **who** was aware of the group allocation at the different stages of the experiment

during the **allocation**,
the **conduct** of the experiment,
the **outcome** assessment,
and the **data analysis**

What it means

Allocation concealment

= Hiding from some or all personnel which treatment was received by which subject

Logistic: must be built into study procedures

Can be at any or all stages

Assignment:	Personnel assigning treatments to EUs
Conduct:	Personnel performing the experiments
Assessment:	Personnel evaluating the outcomes
Analysis, interpretation:	Personnel analysing the data

Why it matters

Allocation concealment minimizes personnel cognitive biases

- Bias can be both conscious and unconscious
- Especially critical for outcomes requiring subjective evaluation
 - Histology
 - Behaviour
 - Clinical progress

6: Outcome measures

Outcome measures

6

- a. Clearly define *all outcome measures* assessed (e.g., cell death, molecular markers, or behavioural changes).
- b. For hypothesis-testing studies, specify the *primary outcome measure*, i.e., the outcome measure that was used to determine the sample size.

What it means

Specific, measurable variables that assess effects of intervention

→ Dependent, response variable

Primary outcome = most important relative to central hypothesis

- Must be clearly defined *a priori*, specific, measurable.

Other variables: “nice to know” = lower priority

Why it matters

Study is both *powered* and *interpreted* off the primary outcome

Investigators wish to measure many things to maximize information obtained from each animal.

Consequences of non-prioritized outcomes

1. Overly large and unfocused study → Impossible to interpret
2. Temptation to cherry-pick, 'chase significance'
→ False and non-informative positives

7: Statistical methods

Statistical methods

7

- a. Provide details of **the statistical methods** used for each analysis, including **software** used.
- b. Describe any methods used to assess whether the data met the **assumptions** of the statistical approach, and **what was done** if the assumptions were not met.

What it means

Statistical methods should be 'bespoke' not boilerplate

What did you do?

Was it appropriate?

There must be alignment between
Study-specific hypotheses, study design, variables
and
Best-practice statistical methods

Altman 1994 *BMJ* 1994;308:283

Diong et al. *PLoS ONE* 2018; 13(8): e0202121.

Analyses: Definitely not bespoke



Methods:	Mouse	Mostly t-tests (100%), rarely ANOVA
	Rat, swine	Mostly one-way ANOVA (85%)
Methods appropriate?		ALMOST NONE
Design specified?		ZERO
Sample size reported		<5%
Time dependencies apparent		MOST (75-90%)
	Accounted for	ALMOST NONE
Factor interactions apparent		MOST (>90%)
	Accounted for	ALMOST NONE (<2%)
Orphan inexact P-values		100%

Reynolds & Garvan. *Military Medicine* 185(S1): 88-95, 2020

Reynolds & Garvan. *Shock* 55: 573-580, 2021

Nunamaker & Reynolds *PLoS ONE* 17(10):e0274738, 2022

Why it matters

Valid statistical methods are essential for interpretation

Most errors are serious enough to **invalidate results**
Most errors are in **basic, not advanced**, statistical methods.

Altman 1994 *BMJ* 1994;308:283

Diong et al. *PLoS ONE* 2018; 13(8): e0202121.

8: Experimental animals

Experimental
animals

8

- a. Provide ***species-appropriate*** details of the animals used, including species, strain and sub-strain, sex, age or developmental stage, and, if relevant, weight.
- b. Provide further relevant information on the **provenance** of animals, health/immune status, genetic modification status, genotype, and any previous procedures.

What it means

“Who is in the study?”

‘Table 1 information’ = describes characteristics of animals in the sample

Hayes-Larson et al. 2019 *J Clin Epidem* 114: 125e132

- 1. Animal signalment:** Details about the animals used (species, strain/breed, age, sex, weight, reproductive status, health)
- 2. Source:** *Verifiable* identification (strain/stock numbers, source)

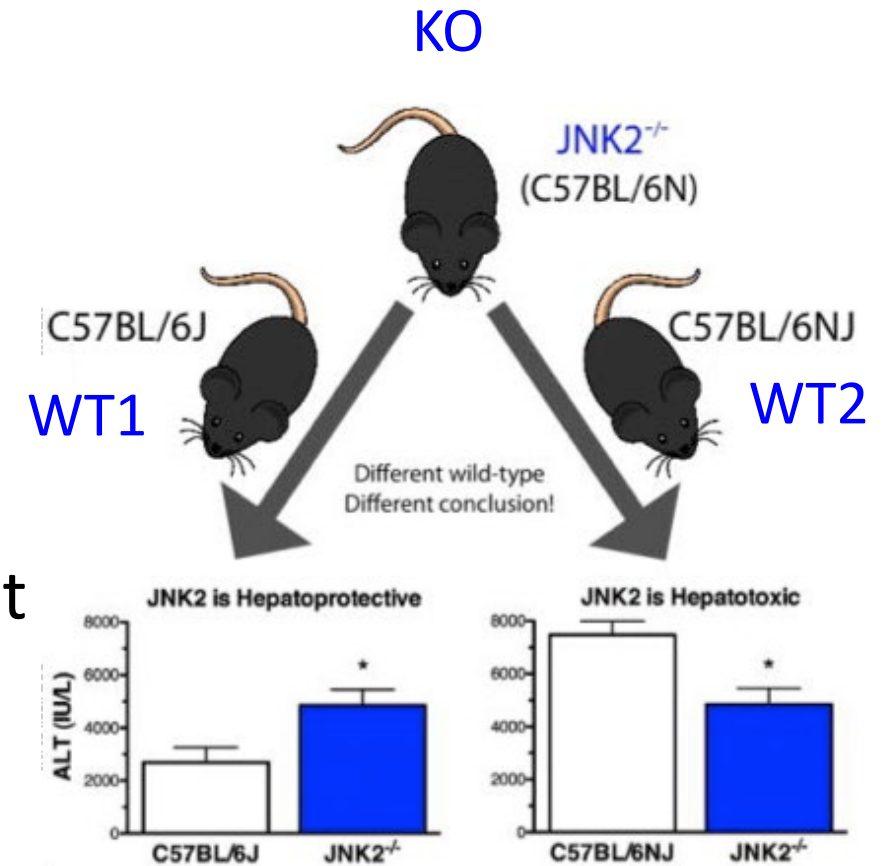
Why it matters

Necessary to assess study validity

- Is the sample appropriate?
- Is the sample representative?
- Can the results be extended?

Signalment = analogous to human patient demographic data

Source: Mice from different vendors or different sublines can show very different responses!



Rasmussen et al *Viruses*. 2019 11(5): 435.

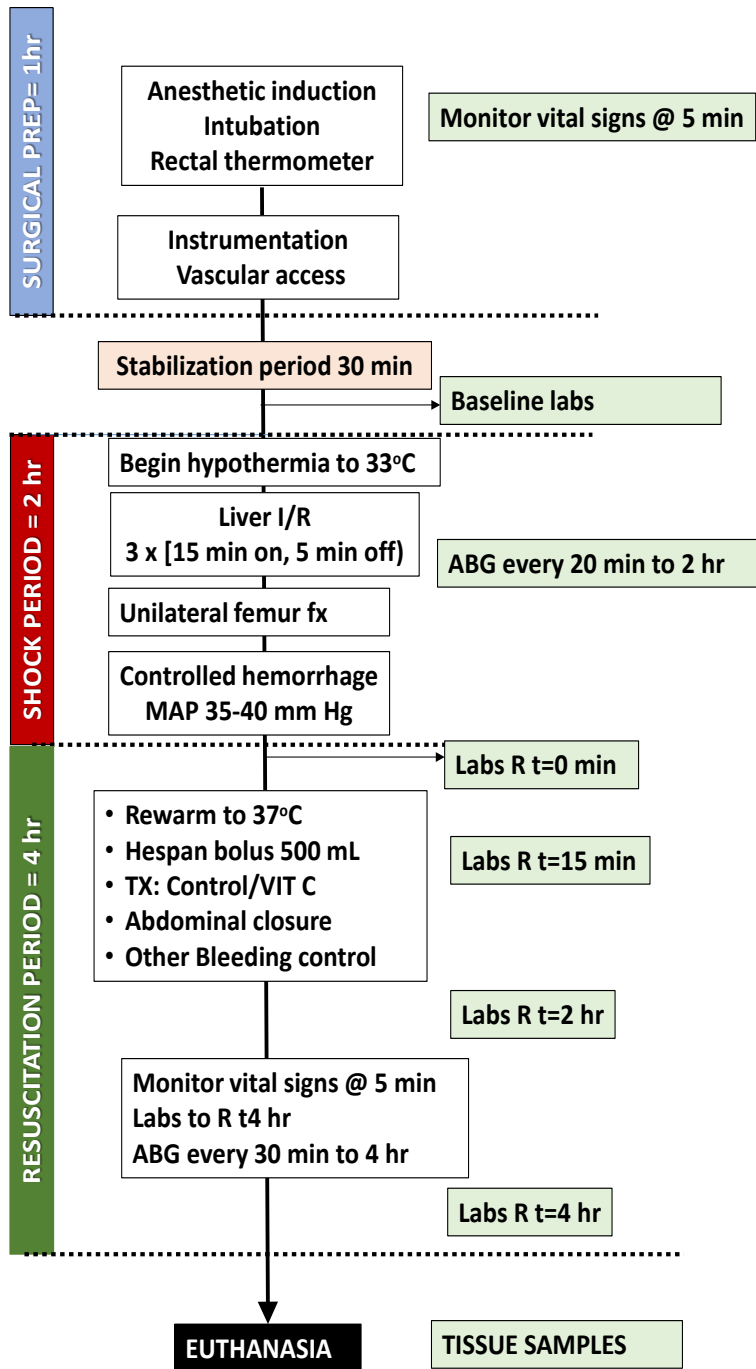
9: Experimental procedures

Experimental
procedures

9

For each experimental group, including controls, **describe procedures in enough detail to allow others to replicate them**, including:

- a. What** was done, **how** it was done, and **what** was used.
- b. When** and **how often**.
- c. Where** (including detail of any acclimatisation periods).
- d. Why** (provide rationale for procedures)



What it means

Describe ALL procedures used to develop the model

Not just the experiment itself

- Pre-experimental
- Preparation/induction of the pathology,
- Experiment proper
- Post-experimental
- Termination: Euthanasia

Why it matters

ALL manipulations can affect the experimental outcome.

Direct = Technical (molecular, laboratory etc) → Mostly reported
AND

Indirect = What was done to the animals → Poorly reported

- Husbandry, handling
- Habituation
- Disease/injury model;
- Surgery;
- Monitoring, sampling
- Drugs, analgesia, anesthesia, palliative/welfare care
- Euthanasia

10: Results

Results

10

For **each experiment** conducted, including independent replications, report:

- a. **Summary/descriptive statistics** for each experimental group, with a **measure of variability** where applicable (e.g., mean and SD, or median and range).
- b. If applicable, the **effect size** with a **confidence interval**

Explicitly statistical

What it means

1. Describing “Who was studied”

What:

- Summary data for the sample
- Signalment, baseline/pre-intervention/clinical/laboratory characteristics

What to report: Sample statistics

- *Sample size* per group n ,
- *Point estimates*: Mean, median, counts (percent)
- *Measure of variation*: SD, IQR (NOT SEM)

What it means

2. Describing “What was found”

What:

- Summary of major results for each study group
- Effect size applicable to results of **hypothesis tests**
- Population-based **measures of precision**

What to report:

Sample size per group n ,

Point estimate: means, mean differences

Measure of variation: confidence intervals

Why it matters

Results of hypothesis tests are used to

- **interpret data**
- **make inferences about the larger population.**

Descriptive statistics summarise sample properties

Confidence intervals describe size, direction, uncertainty of the observed **effect**

- provide **useful, actionable, and interpretable information about the population**

NB: P-values do not! P-values have NO clinical or biological meaning

Gardner MJ, Altman DG (1986) *British Medical Journal*, 292(6522), 746–750.



III. Making the Essential 10 work for you

When should ARRIVE be used?

1.
Study
planning

To *design* experiments

2.
Study
conduct

To *identify and record* critical information

3.
Manuscript
Writing

To *report* all critical information
(memory aid)

4.
Manuscript
review

To *check* that all relevant information included

1. During planning and protocol development

Build quality & reproducibility into the study
during planning

Takes the guesswork out of determining what practices need to be included for a high-quality study

- You cannot report what wasn't done
- Ignorance as a justification of omission is not a justification

2. During manuscript writing

**Identifies reliability, validity, reproducibility items
to be *reported***

Takes the guesswork out of **prioritizing** and **organizing** massive amounts of complex information

Papers & grants are

- Easier to **write**
- Easier to **review**

A high-quality study is more likely to be funded and published

3. After publication

Reliable, valid, reproducible data have a longer shelf life

High-quality, well reported data

- Contribute to databases
- Contribute to systematic reviews

Reliably inform further research

FAQ and common misunderstandings

ARRIVE guidelines simply tell you
to **report** what you did do,
and **justify** what you didn't do.

Most misunderstandings occur because researchers do not
understand the difference between
conducting research
and
reporting research

FAQ: Won't these guidelines stifle creativity?

NO

ARRIVE guidelines do not prescribe research topics.

- ARRIVE helps you report your methods and results
- If important information is missing, the article is useless.

FAQ: “What if I don’t do those items?”

Reporting of each item should still be COMPLETE

If not performed, say so

Some may not be possible e.g. Allocation concealment

If key reproducibility items are NOT performed

- Report omission honestly
- Justify omission (if scientifically warranted)
- List as a study limitation
- And don’t lie!

FAQ: “What if I just say I did all that?”

Research misconduct is a continuum

Bad practices lead to misconduct

*Questionable
research
practices*



*Scientific
fraud*

Laziness = Too much bother to find out about and incorporate best practices

→ Irresponsible, Negligent

Liar = Deliberately misrepresent & distort research =

→ Scientific fraud

- You are a very bad person.

Ignorance, incompetence, and lies are not good looks

Examples of box-ticking

Red flag claim 1

- “Experiments were performed according to the National Institutes of Health guidelines and ARRIVE guidelines on the use of laboratory animals
- “The animal experimental protocol was in accord with the ARRIVE guidelines”.
- The experiments were conducted in compliance with (ARRIVE) guidelines for animal models and National Institutes of Health guidelines on the use of laboratory animals.

ARRIVE is for **disclosure**, not study-specific conduct

ARRIVE does not dictate or mandate experimental protocols

[And they haven't read *The Guide* either]

Red flag claim 2

- “We followed the ARRIVE guidelines for the **care of animals** in biomedical research in performing these experiments”
- “All efforts were made to **minimize the number of animals used and the suffering of animals** in accordance with the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines.
- “This protocol was performed in accordance with the **Animal Welfare Act** and other Federal statutes and regulations relating to animal experiments, as well as the **Guide for the Care and Use of Laboratory Animals** of the National Academy of Sciences and the **ARRIVE guidelines**”.

ARRIVE is NOT a statement of investigator compliance with ethical care and use standards!

An informal test

- One journal Nov 2022- Jun 2023;
- 9 papers EXPLICITLY claimed ARRIVE compliance
 - “Procedures” 6/9 ; “Ethical oversight” 2/9; Reporting 1; Checklists 2
- Design: “Groups” 2; Design 0
- Sample size: Total 3; Per group 3; Justification 0
- “Randomisation” 5; Method 0; checklists “N/A”
- Outcomes: 0, Primary 0; checklists “N/A”
- Statistical methods: 9; appropriate 0
- Results: Orphan inexact P-values 9; other 0
- Study positive? 9

Concluding thoughts

Incorporating the Essential 10
will not be 'business as usual' for
researchers or reviewers

Implementation will be disruptive

1. Researchers require new skills

Experimental design

Updated, more relevant statistical analysis methods

Better more relevant instruction in basic statistical methods

2. Grant and journal reviewers must do better due diligence

Sound methodology over small P-values

(Checklist standards actually expedite reviews)

Summary

1. Good science relies on reliable, valid, and transparently reported information
2. Research quality depends on experimental validity
3. Reporting guidelines help us get there
4. Understanding the essentials enables you to build in quality from the beginning
 - Fewer animals used, less research waste

Where to find ARRIVE 2.0 guidelines

ARRIVE 2.0 website <https://www.ARRIVEguidelines.org>

1. Checklist, overview *PLoS Biology* 18(7): e3000410

Simultaneous release in multiple journals

BMJ Open Science, Br J Pharmacol., BMC Vet Res., J Physiol., J Exp Physiol., J Cerebr Blood & Met., Vet Clin Pathol, BMJ Open Science

2. Explanation & Elaboration document

PLoS Biology 18(7): e3000411

Acknowledgements

Dr Nathalie Percie du Sert, NC3Rs

NC3Rs staff

ARRIVE 2.0 International Working Group colleagues



National Centre
for the Replacement
Refinement & Reduction
of Animals in Research





Questions?
Thank you for
your attention