PENNY S. REYNOLDS

A Guide to Sample Size for Animal-based Studies

WILEY Blackwell

A Guide to Sample Size for Animal-based Studies

Penny S. Reynolds
Department of Anesthesiology, College of Medicine
Department of Small Animal Clinical Sciences
College of Veterinary Medicine
University of Florida, Gainesville
Florida, USA

WILEY Blackwell

This edition first published 2024

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data applied for

Paperback ISBN: 9781119799979

Cover Design: Wiley
Cover Images: © MOLEKUUL/SCIENCE PHOTO LIBRARY/Getty Images; MOLEKUUL/SCIENCE PHOTO LIBRARY/Getty Images; Verin/Shutterstock; n.tati.m/Shutterstock; Mariia Zotova/Getty Images; RF Pictures/Getty Images

Set in 11.5/13.5pt STIXTwoText by Straive, Pondicherry, India

To Nyx, Mel, Finnegan, and Fat Boy Higgins
Holly, Molly, and Abby
and all their nameless, uncounted kindred
who have done so much to advance science and medicine

Contents

Preface

'How large a sample size do I need for my study'? Although one of the most commonly asked questions in statistics, the importance of proper sample size estimation seems to be overlooked by many preclinical researchers. Over the past two decades, numerous reviews of the published literature indicate many studies are too small to answer the research question and results are too unreliable to be trusted. Few published studies present adequate justification of their chosen sample sizes or even report the total number of animals used. On the other hand, it is not unusual for protocols (usually those involving mouse models) to request preposterous numbers of animals, sometimes in the tens or even hundreds of thousands, 'because this is an exploratory study, so it is unknown how many animals we will require'.

This widespread phenomenon of sample sizes based on nothing more than guesswork or intuition illustrates the pervasiveness of what Amos Tversky and Daniel Kahneman identified in 1971 as the 'belief in the law of small numbers'. Researchers overwhelmingly rely on best judgement in planning experiments, but judgement is almost always misleading. Researchers choose sample sizes based on what 'worked' before or because a particular sample size is a favourite with the research community. Tversky and Kahneman showed that researchers who gamble their research results on small intuitively-based samples consistently have the odds stacked against their findings (even if results are true). They overestimate the stability and precision of their results, and fail to account for sampling variation as a possible reason for observed pattern. The result is research waste on a colossal scale, especially of animals, that is increasingly difficult to justify.

This book was written to assist non-statisticians who use animals in research to 'right-size' experiments, so they are statistically, operationally, and ethically justifiable. A 'right-sized' experiment has a clear plan for sample size justification and transparently reports the numbers of all animals used in the study. For basic and veterinary researchers, appropriate sample sizes are critical to the design and analysis of a study. The best sample sizes optimise study design to align with available resources and ensure the study is adequately powered to detect meaningful, reliable, and generalisable results. Other stakeholders not directly involved in animal experimentation can also benefit from understanding the basic principles involved. Oversight veterinarians and ethical oversight committees are responsible for appraising animal research protocols for compliance with best practice, ethical, and regulatory standards. An appreciation of sample size construction can help assess scientific and ethical justifications for animal use and whether the proposed sample size is fit for purpose. Funding agencies and policymakers use research results to inform decisions related to animal welfare, public health, and future scientific benefit. Understanding the logic behind sample size justification can assist in evaluation of study quality and reliability of research findings, and ultimately promote more informed evidence-based decision-making.

An extensive background in statistics is not required, but readers should have had some basic statistical training. The emphasis throughout is on the upstream components of the research process – statistical process, study planning, and sample size calculations rather than analysis. I have used real data in nearly all examples and provided formulae and code, so sample size approximations can be reproduced by hand or by computer. By training and inclination I prefer SAS, but whenever possible I have provided R code or links to R libraries.

Acknowledgements

Many thanks to Anton Bespalov (PAASP, Heidelberg, Germany); Cori Astrom, Christina Hendricks, and Bryan Penberthy (University of Florida); Cora Mezger, Maria Christodoulou, and Mariagrazia Zottoli (Department of Statistics, University of Oxford); and Megan Lafollette (North American 3Rs Collaborative), who kindly reviewed various chapters of this book whilst it was in preparation and provided much helpful feedback. Thanks to the University of Florida IACUC chairs Dan Brown and Rebecca Kimball, who encouraged researchers to consult the original 10-page handout I had devised for sample size estimation. And last, but certainly not least, special thanks to Tim Morey, Chair of the Department of Anesthesiology, University of Florida, who encouraged me to put that handout into book form.

Thanks are also due to the University of Florida Faculty Endowment Fund for providing me with a Faculty Enhancement Opportunities grant to allow me to devote some concentrated time to writing. A generous honorarium from Scientist Center for Animal Welfare (SCAW) and an award from the UK Animals in Science Education Trust enabled me to upgrade my home computer system, making working on this project immeasurably easier.

The book was nearing completion when I came across the Icelandic word *sprakkar* that means 'extraordinary women'. I have been fortunate to encounter many *sprakkar* whilst writing this book. In addition to the women (and men!) already mentioned, special thanks to researchers Amara Estrada, Francesca Griffin, Autumn Harris, Maggie Hull, Wendy Mandese, and Elizabeth Nunamaker, who generously allowed me to use some of their data as examples. And special thanks to Jane Buck and Julie Laskaris for their wonderful friendship and hospitality over the years. Jane Buck, Professor Emerita of Psychology, Delaware State University, and past president of the American Association of University Professors, continues to amaze and show what is possible for a statistician 'with attitude'. Julie advised me that the only approach to properly edit one's own work on a book-length project was to 'slit its throat', then told me to do as she said, not as she actually did. Cheers.

What is Sample Size?

1 The Sample Size Problem in Animal-Based Research

CHAPTER OUTLINE HEAD

Good Numbers Matter. This is especially true when animals are research subjects. Researchers are responsible for minimising both direct harms to research animals and the indirect harms that result from wasting animals in poor-quality studies (Reynolds 2021). The ethical use of animals in research is framed by the 'Three Rs' principles of Replacement, Reduction, and Refinement. Originating over 60 years ago (Russell and Burch 1959), the 3Rs strategy is framed by the premise that maximal information should be obtained for minimal harms. Harms are minimised by Replacement, methods or technologies that substitute for animals; Reduction, the methods using the fewest animals for the most robust and scientifically valid information; and Refinement, the methods that improve animal welfare through minimising pain, suffering, distress, and other harms (Graham and Prescott 2015).

The focus of this book is on Reduction and methods of 'right-sizing' experiments. A right-sized experiment is an optimal size for a study to achieve its objectives with the least amount of resources, including animals. However, simply minimising the total number of animals is not the same as right-sizing. A right-sized experiment has a sample size that is statistically, operationally, and ethically defensible (Box 1.1). This will mean compromising between the scientific objectives of the study, production of scientifically valid results, availability

BOX 1.1
Right-Sizing Checklist

Statistically defensible: Are numbers verifiable? (Calculations)

- Outcome variable identified
- Difference to be detected
- Expected variation in response
- Number of groups
- Anticipated statistical test (if hypothesis tests used)
- All calculations shown

Operationally defensible: Are numbers feasible? (Resources)

- Qualified technical staff
- Time
- Space
- Resources
- Equipment
- Funding

Ethically defensible: Are numbers fit for purpose? (3Rs)

- Appropriate for study objectives?
- Reasonable number of groups?
- Are collateral losses accounted for and minimized?
- Are loss mitigation plans described?
- Are 3Rs strategies described?

Source: Adapted from Reynolds (2021).

A Guide to Sample Size for Animal-based Studies, First Edition. Penny S. Reynolds.

of resources, and the ethical requirement to minimise waste and suffering of research animals. Thus, sample size calculations are not a single calculation but a set of calculations, involving iteration through formal estimates, followed by reality checks for feasibility and ethical constraints (Reynolds 2019).

Additional challenges to right-sizing experiments include those imposed by experimental design and biological variability (Box 1.2). In *The Principles of Humane Experimental Technique* (1959), Russell and Burch were very clear that Reduction is achieved by systematic strategies of experimentation rather than trial and error. In particular, they emphasised the role of the statistically based family of experimental designs and design principles proposed by Ronald Fisher, still relatively new at the time. Formal experimental designs customised to address the particular research question increase the experimental signal through the reduction of variation. Design principles that reduce bias, such as randomisation and allocation concealment (blinding) increase validity. These methods increase the amount of usable information that can be obtained from each animal (Parker and Browne 2014).

Although it has now been almost a century since Fisher-type designs were developed many researchers in biomedical sciences still seem unaware of their existence. Many preclinical studies reported in the literature consist of numerous two-group designs. However, this approach is both inefficient and inflexible and unsuited to exploratory studies with multiple explanatory variables (Reynolds 2022). Statistically based designs are rarely reported in the preclinical literature. In part, this is because the design of experiments is seldom taught in introductory statistics courses directed towards biomedical researchers.

Power calculations are the gold standard for sample size justification. However, they are commonly misapplied, with little or no consideration of study design, type of outcome variable, or the purpose of the study. The most common power calculation is for two-group comparisons of independent samples. However, this is inappropriate when the study is intended to examine multiple independent factors and interactions. Power calculations for continuous variables are not appropriate for correlated observations or count data with high prevalence of zeros. Power calculations cannot be used at all when statistical inference is not the purpose of the study, for example, assessment of operational and ethical feasibility, descriptive or natural history studies, and species inventories.

Evidence of right-sizing is provided by a clear plan for sample size justification and transparent reporting of the number of all animals used in the study. This is why these items are part of best-practice reporting standards for animal research publications (Kilkenny et al. 2010, Percie du Sert et al. 2020 and are essential for the assessment of research reproducibility (Vollert et al. 2020). Unfortunately, there is little evidence that either sample size justification or sample size reporting has improved over the past decade. Most published animal research studies are underpowered and biased (Button et al. 2013, Henderson et al. 2013, Macleod et al. 2015) with poor validity (Würbel 2017, Sena and Currie 2019), severely limiting reproducibility and translation potential (Sena et al. 2010, Silverman et al. 2017). A recent cross-sectional survey of mouse cancer model papers published in high-impact oncology journals found that fewer than 2% reported formal power calculations, and less than one-third reported sample size per group. It was impossible to determine attrition losses, or how many experiments (and therefore animals) were discarded due to failure to achieve statistical

BOX 1.2
Challenges for Right-Sizing Animal-Based Studies

Ethics and welfare considerations. The three Rs Replacement, Reduction, and Refinement should be the primary driver of animal numbers.

Experimental design. Animal-based research has no design culture. Clinical trial models are inappropriate for exploratory research. Multifactorial agriculture/industrial design may be more suitable in many cases, and they are unfamiliar to most researchers.

Biological variability. Animals can display significant differences in responses to interventions, making it challenging to estimate an appropriate sample size.

Cost and resource constraints. The financial cost of conducting animal-based research, including the cost of housing, caring for, and monitoring the animals, must be considered in estimates of sample size.

significance (Nunamaker and Reynolds 2022). The most common sample size mistake is not performing any calculations at all (Fosgate 2009). Instead, researchers make vague and unsubstantiated statements such as 'Sample size was chosen because it is what everyone else uses' or 'experience has shown this is the number needed for statistical significance'. Researchers often game, or otherwise adjust, calculations to obtain a preferred sample size (Schultz and Grimes 2005, Fitzpatrick et al. 2018). In effect, these studies were performed without justification of the number of animals used.

Statistical thinking is both a mindset and a set of skills for understanding and making decisions based on data (Tong 2019). Reproducible data can only be obtained by sustained application of statistical thinking to all experimental processes: good laboratory procedure, standardised and comprehensive operating protocols, appropriate design of experiments, and methods of collecting and analysing data. Appropriate strategies of sample size justification are an essential component.

1.1 Organisation of the Book

This book is a guide to methods of approximating sample sizes. There will never be one number or approach, and sample size will be determined for the most part by study objectives and choice of the most appropriate statistically based study design. Although advanced statistical or mathematical skills are not required, readers are expected to have at least a basic course on statistical analysis methods and some familiarity with the basics of power and hypothesis testing. SAS code is provided in appendices at the end of each chapter and references to specific R packages in the text. It is strongly recommended that everyone involved in devising animal-based experiments take at least one course in the design of experiments, a topic not often covered by statistical analysis courses.

This book is organised into four sections (Figure 1.1).

Part I *Sample size basics* discusses definitions of sample size, elements of sample size determination, and strategies for maximising information power without increasing sample size.

Part II *Feasibility*. This section presents strategies for establishing study feasibility with pilot studies. Justification of animal numbers must first address questions of operational feasibility ('*Can it work?*' Is the study possible? suitable? convenient? sustainable?). Once operational logistics are standardised, pilot studies can be performed to establish empirical feasibility ('*Does it work*?' is the output large enough to be measured? consistent enough to be reliable?)

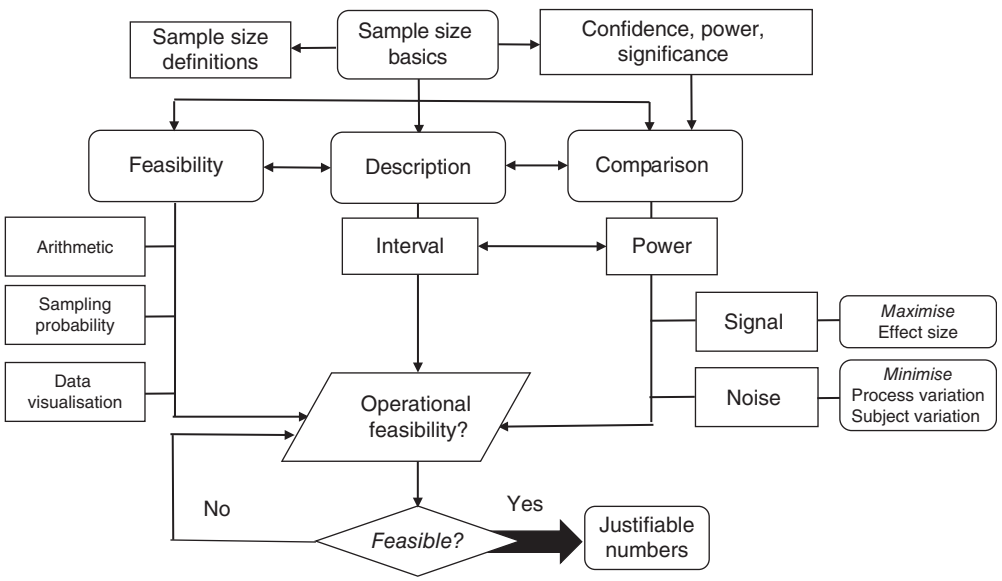

Figure 1.1: Overview of book organisation. For animal numbers to be *justifiable* (Are they feasible? appropriate? ethical? verifiable?), sample size should be determined by formal quantitative calculations (arithmetic, probability-based, precision-based, power-based) and consideration of operational constraints.

and translational feasibility ('*Will it work?*' proof of concept and proof of principle) before proceeding to the main experiments. Power calculations are not appropriate for most pilots. Instead, common-sense feasibility checks include *basic arithmetic* (with structured back-of-the-envelope calculations), simple probability-based calculations, and graphics.

Part III *Description*. This section presents methods for summarising the main features of the sample data and results. Basic descriptive statistics provide a simple and concise summary of the data in terms of central tendency and dispersion or spread. Graphical representations are used to identify patterns and outliers and explore relationships between variables. Intervals computed from the sample data are the range of values estimated to contain the true value of a population parameter with a certain degree of confidence. Four types of intervals are discussed: confidence intervals, prediction, intervals, tolerance intervals, and reference intervals. Intervals shift emphasis away from significance tests and *P*-values to more meaningful interpretation of results.

Part IV *Comparisons*. Power-based calculations for sample size are centred on understanding effect size in the context of specific experimental designs and the choice of outcome variables. Effect size provides information about the practical significance of the results beyond considerations of statistical significance. Specific designs considered are two-group comparisons, ANOVA-type designs, and hierarchical designs.

References

Button, K.S., Ioannidis, J.P.A., Mokrysz, C. et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 365–376.

Fitzpatrick, B.G., Koustova, E., and Wang, Y. (2018). Getting personal with the "reproducibility crisis": interviews in the animal research community. *Lab Animal (NY)* 47: 175–177.

Fosgate, G.T. (2009). Practical sample size calculations for surveillance and diagnostic investigations. *Journal of Veterinary Diagnostic Investigation* 21: 3–14. https://doi.org/10.1177/104063870902100102.

Graham, M.L. and Prescott, M.J. (2015). The multifactorial role of the 3Rs in shifting the harm-benefit analysis in animal models of disease. *European Journal of Pharmacology* 759: 19–29. https://doi.org/10.1016/j.ejphar.2015.03.040.

Henderson, V.C., Kimmelman, J., Fergusson, D. et al. (2013). Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Medicine* 10: e1001489.

Kilkenny, C., Browne, W.J., Cuthill, I.C. et al. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology* 8 (6): e1000412. https://doi.org/10.1371/journal.pbio.1000412.

Macleod, M.R., Lawson McLean, A., Kyriakopoulou, A. et al. (2015). Risk of bias in reports of *in vivo* research: a focus for improvement. *PLoS Biology* 13: e1002301. https://doi.org/10.1371/journal.pbio.1002273.

Nunamaker, E.A. and Reynolds, P.S. (2022). "Invisible actors"—how poor methodology reporting compromises mouse models of oncology: a cross-sectional survey. *PLoS ONE* 17 (10): e0274738. https://doi.org/10.1371/journal.pone.0274738.

Parker, R.M.A. and Browne, W.J. (2014). The place of experimental design and statistics in the 3Rs. *ILAR Journal* 55 (3): 477–485.

Percie du Sert, N., Hurst, V., Ahluwalia, A. et al. (2020). The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *PLoS Biology* 18 (7): e3000410. https://doi.org/10.1371/journal.pbio.3000410.

Reynolds, P.S. (2019). When power calculations won't do: fermi approximation of animal numbers. *Lab Animal (NY)* 48: 249–253.

Reynolds, P.S. (2021). Statistics, statistical thinking, and the IACUC. *Lab Animal (NY)* 50 (10): 266–268. https://doi.org/10.1038/s41684-021-00832-w.

Reynolds, P.S. (2022). Between two stools: preclinical research, reproducibility, and statistical design of experiments. *BMC Research Notes* 15: 73. https://doi.org/10.1186/s13104-022-05965-w.

Russell, W.M.S. and Burch, R.L. (1959). *The Principles of Humane Experimental Technique*. London: Methuen.

Schulz, K.F. and Grimes, D.A. (2005). Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 365 (9467): 1348–1353. https://doi.org/10.1016/S0140-6736(05)61034-3.

Sena, E.S. and Currie, G.L. (2019). How our approaches to assessing benefits and harms can be improved. *Animal Welfare* 28: 107–115.

Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. Publication bias in reports of animal

stroke studies leads to major overstatement of efficacy. *PLoS Biology*, 2010 8(3):e1000344. https://doi.org/10.1371/journal.pbio.1000344.

Silverman, J., Macy, J., and Preisig, P. (2017). The role of the IACUC in ensuring research reproducibility. *Lab Animal (NY)* 46: 129–135.

Tong, C. (2019). Statistical inference enables bad science; statistical thinking enables good science. *American Statistician* 73: 246–261.

Vollert, J., Schenker, E., Macleod, M. et al. (2020). Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. *BMJ Open Science* 4 (1): e100046. https://doi.org/10.1136/bmjos-2019-100046.

Würbel, H. (2017). More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. *Lab Animal* 46: 164–166.

2 Sample Size Basics

CHAPTER OUTLINE HEAD

2.1 Introduction

Investigators frequently assume 'sample size' is the same as 'the number of animals'. This is not necessarily true. Reliable sample size estimates are determined by the correct identification of the experimental units, the true unit of replication (Box 2.1). Replication of experimental units increases both precision of estimates and statistical power for testing the central hypothesis. Replicates on the same subject over time provide an estimate of time dependencies in response. Technical replicates are used to obtain an estimate of measurement error and are essential for quality control of experimental procedures. Pseudo-replication is a serious statistical error that occurs when the number of data points (evaluation units) is confused with the number of independent samples, or experimental units (Hurlbert 2009; Lazic 2010). Incorrect specification of the true sample size results in erroneous estimates of the standard error, inflated type I error rates, and increased number of false positives (Cox and Donnelly 2011). Research results will therefore be biased and misleading.

Definitions of 'replicates' and 'replication' are frequently confused in the literature, and further conflated with study replication. Planning experiments using formal statistical designs can help differentiate between the different types of replicates and sampling units, and determine which is best suited for the intended study.

BOX 2.1
What Is Sample Size?

A *replicate* is one unit in one group.

Sample size is determined by the number of replicates of the *experimental unit*.

Experimental unit: Entire entity to which a treatment or control intervention can be independently and individually applied.

Biological replicate is a biologically distinct and independent experimental unit.

Technical replicate is one of multiple measurements on subsamples of the experimental unit, used to obtain an estimate of measurement error.

2.2 Experimental Unit

The *experimental unit* or *unit of analysis* is the smallest entire entity to which a treatment or control intervention can be independently and randomly

A Guide to Sample Size for Animal-based Studies, First Edition. Penny S. Reynolds.

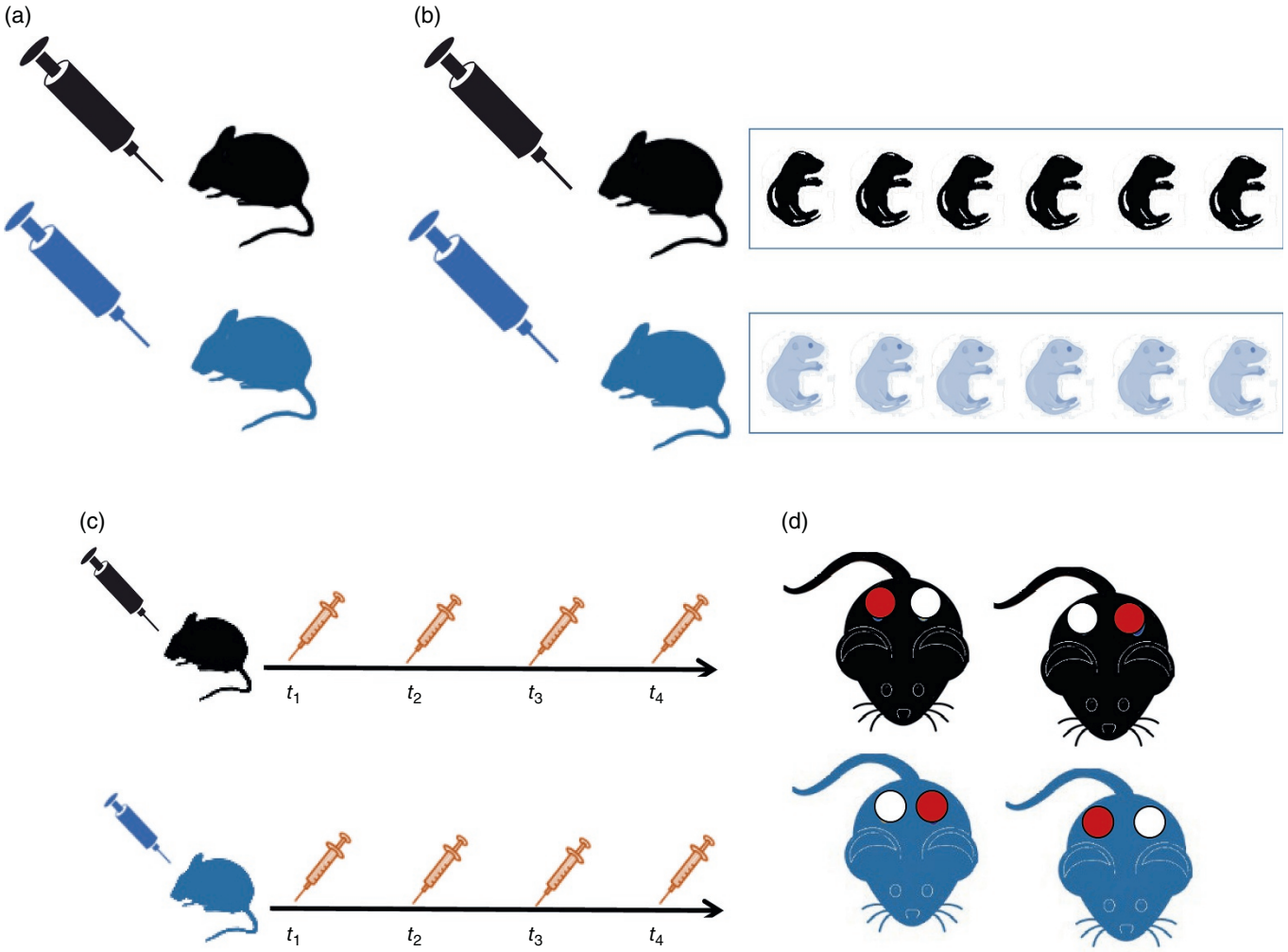

Figure 2.1: Units of replication. (a) Experimental unit = individual animal = biological unit. The entire entity to which an experimental or control intervention can be independently applied. There are two treatment interventions A or B. Here each mouse receives a separate intervention, and the individual mouse is the experimental unit (EU). The individual mouse is also the biological unit. (b). Experimental unit = groups of animals. There are two treatment interventions A or B. Each dam receives either A or B, but measurements are conducted on the pups in each litter. The experimental unit is the dam ($N = 2$), and biological unit is the pup ($n = 8$). For this design, the number of pups cannot contribute to the test of the central hypothesis. (c) Experimental unit with repeated observations. The experimental unit is the individual animal *(*= biological unit) with four sequential measurements made on each animal. The sample size $N = 2$. (d) Experimental unit = part of each animal. There are two treatment interventions A or B. Treatment A is randomised to either the right or left flank of each mouse, and B is injected into the opposite flank of that mouse. The experimental unit is flank ($N = 8$). The individual mouse is the biological unit. Each mouse can be considered statistically as a block with paired observations within each animal.

applied (Figure 2.1a). Cox and Donnelly (2011) define it as the 'smallest subdivision of the experimental material such that two distinct units might be randomized (randomly allocated) to different treatments.' Whatever happens to one experimental unit will have no bearing on what happens to the others (Hurlbert 2009). If the test intervention is applied to a 'grouping' other than the individual animal (e.g. a litter of mice, a cage or tank of animals, a body part; (Figure 2.1b–d)), then the sample size N will not be the same as the number of animals.

The total sample size N refers to the number of independent experimental units in the sample. The classic meaning of a 'replicate' refers to the number of experimental units within a treatment or intervention group. Therefore, replicating experimental units (and hence increasing N) contributes to statistical power for testing the central statistical hypothesis. Power calculations estimate the number

of experimental units required to test the hypothesis. The assignment of treatments and controls to experimental units should be randomised if the intention is to perform statistical hypothesis tests on the data (Cox and Donnelly 2011).

Independence of experimental units is essential for most null hypothesis statistical tests and methods of analysis and is the most important condition for ensuring the validity of statistical inferences (van Belle 2008). Non-independence of experimental units occurs with repeated measures and multi-level designs and must be handled by the appropriate statistically based designs and analyses for hypothesis tests to be valid.

2.3 Biological Unit

The biological unit is the entity about which inferences are to be made. Replicates of the biological unit are the number of unique biological samples or individuals used in an experiment. Replication of biological units captures biological variability between and within these units (Lazic et al. 2018). The biological unit is not necessarily the same as the experimental unit. Depending on how the treatment intervention is randomised, the experimental unit can be an individual biological unit, a group of biological units, a sequence of observations on a single biological unit or a part of a biological unit (Lazic and Essioux 2013; Lazic et al. 2018). The biological unit of replication may be the whole animal or a single biological sample, such as strains of mice, cell lines or tissue samples (Table 2.1).

2.4 Technical Replicates

Technical replicates or repeats are multiple measurements made on subsamples of an experimental unit (Figure 2.2). Technical replicates are used to obtain an estimate of measurement error, the difference between a measured quantity and its true value. Technical replicates are essential for assessing internal quality control of experimental procedures and processes, and ensuring that results are not an artefact of processing variation (Taylor and Posch 2014). Differences between operators and instruments, instrument drift, subjectivity in determination of measurement landmarks, or faulty calibration can result in measurement error. Cell cultures and protein-based experiments can also show considerable variation from run to run, so *in vitro* experiments are usually repeated several

Table 2.1: Units of Replication in a Hypothetical Single-Cell Gene Expression RNA Sequencing Experiment. Designating a given replicate unit as an experimental unit depends on the central hypothesis to be tested and the study design.

	Replicate 'unit'	Replicate type
Animals	Colonies	Biological
	Strains	Biological
	Cohoused animals in a cage	Biological
	Sex (male, female)	Biological
	Individuals	Biological
Sample preparation	Organs from animals killed for purpose	Biological
	Methods for dissociating cells from tissue	Technical
	Dissociation runs from given tissue sample	Technical
	Individual cells	Biological
	RNA-seq library construction	Technical
Sequencing	Runs from the library of a given cell	Technical
	Readouts from different transcript molecules	Biological or technical
	Readouts with unique molecular identifier (UMI) from a given transcript molecule	Technical

Source: Adapted from Blainey et al. (2014).

times. At least three technical replicates of Western blots, PCR measurements, or cell proliferation assays may be necessary to assess reliability of technique and confirm validity of observed changes in protein levels or gene expression (Taylor and Posch 2014).

The variance calculated from the multiple measurements is an estimate of the precision, and therefore the repeatability, of the measurement. Technical replicates measure the variability between measurements on the same experimental units. Repeating measurements increases the precision only for estimates of the measurement error; they do not measure variability either within or between treatment groups. Therefore, increasing the number of technical replicates does not improve power or contribute to the sample size for testing the central hypothesis. Analysing technical repeats as independent measurements is pseudo-replication.

High-dimensionality studies produce large amounts of output information per subject. Examples include multiple DNA/RNA microarrays; biochemistry assays; biomarker studies; proteomics; metabolomics; inflammasome profiles, etc. These studies may require a number of individual animals, either for operational purposes (for example, to obtain enough tissue for processing) or as part of the study design (for example, to estimate biological variation). Sample size will then be determined by the amount of tissue required for the assay technical replicates, or by design-specific requirements for power. Design features include anticipated response/expression rates, expected false positive rate, and number of sampling time points (Lee and Whitmore 2002; Lin et al. 2010; Jung and Young 2012).

Example: Experimental Units with Technical Replication

Two treatments A and B are randomly allocated to six individually housed mice, with three mice receiving A and three receiving B. Lysates are obtained from each mouse in three separate aliquots (Figure 2.2).

The individual mouse is the experimental unit because treatments can be independently and randomly allocated to each mouse. There are three subsamples or technical replicates per mouse. The total sample size is $N = 6$, with $k = 2$ treatments, $n = 3$ mice per treatment group, and $j = 3$ technical replicates per mouse. The total sample size N is 6, not 18.

2.5 Repeats, Replicates, and Pseudo-Replication

Confusion of repeats with replicates is a problem of study design, and pseudo-replication is a problem of analysis. Study validity is compromised by incorrect identification of the experimental unit. A replicate is

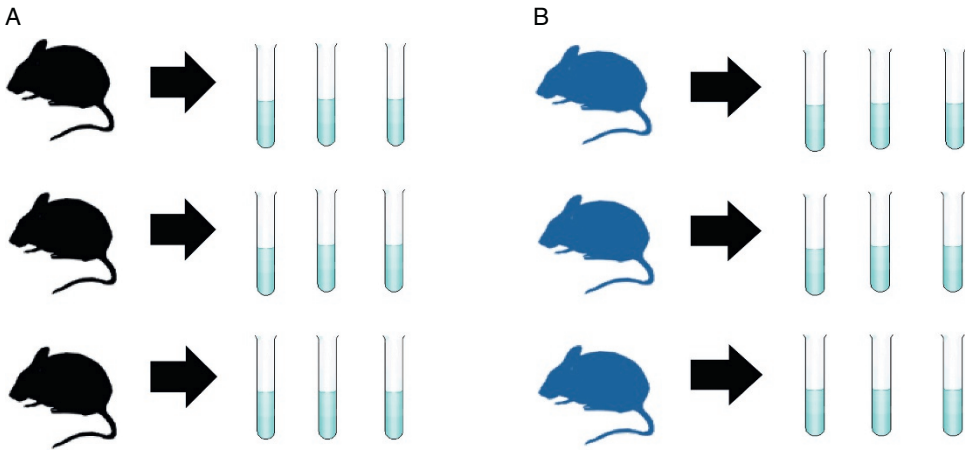

Figure 2.2: Experimental unit versus technical replicates. Two treatments A and B are randomly allocated to six mice. The individual mouse is the *experimental unit*. Three lysate aliquots are obtained from each mouse. These are *technical replicates*. The total sample size *N* is 6, not 18.

a new experimental run on a new experimental unit. Randomisation of interventions to experimental units and randomising the order in which experimental units are measured (sequence allocation randomisation) minimises the effects of systematic error or bias. A repeat is a consecutive run of the same treatment or factor combination. It does not minimise bias and may actually increase bias if there are time-dependencies in the data. Repeats are not valid replicates.

Example: Replication Versus Repeats

In Figure 2.3, the experimental units are eight mice that receive one of two interventions. In the first scenario, both treatment allocated to mouse and the measurement sequence are randomised. Bias is minimised and treatment variance will be appropriately estimated. In the second scenario, treatment intervention may or may not have been randomly allocated to mice, but measurements were obtained for all mice in the first group followed by those in the second group. Bias results from confounding of outcome measurements with potential time-dependencies (for example, increasing skill levels or learning) and difference in assessment, especially if treatment allocation is not concealed (blinded).

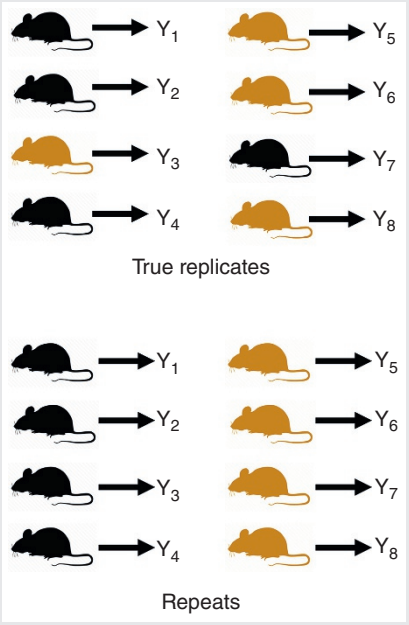

Figure 2.3: Replicates versus repeats. *True replicates* are separate runs of the same treatment on separate experimental units. Both treatment allocation to units and sequence allocation for the processing of individual experimental units are randomised. In this experiment, the eight measurements on eight mice are taken in random order. *Repeat* measurements are taken during the same experimental run or consecutive runs. Unless processing order is randomised, there will be confounding with systematic sources of variability caused by other variables that change over time. In this experiment, eight measurements on eight mice are obtained consecutively with units in the first treatment measured first.

2.5.1 Repeats of Entire Experiments

A common practice in mouse studies is to repeat an entire experiment two or three times. It has been argued that this practice provides evidence that results are robust. However, NIH directives are clear that replication is justifiable only for major or key results, and that replications be independent. Repeating an experiment in triplicate by a single laboratory is not independent replication. These repeats can provide only an estimate of the overall measurement error of that experiment for that lab. A major consideration is study quality. If the study is poorly designed and underpowered, replicating it only wastes animals. Unless the purpose of direct internal replications is scientifically justified, experiments are appropriately designed and conducted to maximise internal validity, and experimental, biological, and technical replicates are clearly distinguished, simple direct repeats of experiments on whole animals are rarely ethically justifiable. Chapter 6 provides practical guidelines for experiment replication.

2.5.2 Pseudo-Replication

In a classic paper, Hurlbert (1984) defines *pseudoreplication* as occurring when inferential statistics are used 'to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or experimental units are not statistically independent' (Hurlbert 1984, 2009). The extent of pseudoreplication in animal-based research is disturbingly prevalent. Lazic et al. (2018) reported that less than one-quarter of studies they surveyed identified the

correct repli-cation unit, and almost half showed pseudo-replication, suggesting that inferences based on hypothesis tests were likely invalid.

Three of the most common types of pseudo-replication are *simple, sacrificial*, and *temporal*. Others are described in Hurlbert and White (1993) and Hurlbert (2009).

Simple pseudo-replication occurs when there is only a single replicate per treatment. There may be multiple observations, but they are not obtained from independent experimental replicates. The artificial inflation of sample size results in estimates of the standard error that are too small, contributing to increased Type I error rate and increased number of false positives.

Example: Mouse Photoperiod Exposure

A study on circadian rhythms was conducted to assess the effect of two different photoperiods on mouse wheel-running. Mice in one environmental chamber were exposed to a long photoperiod with 14 hours of light, and mice in a second chamber to a short photoperiod with 6 hours of light. There were 15 cages in each chamber with four mice per cage. What is the effective sample size?

This is simple pseudo-replication. The experimental unit is the chamber, so the effective sample size is one per treatment. Analysing the data as if there is a sample size of $n = 60$ (or even $n = 15$) per treatment is incorrect. The number of mice and cages in each chamber is irrelevant. This design implicitly assumes that chamber conditions are uniform and chamber effects are zero. However, variation both between chambers and between repeats for the same chamber can be considerable (Potvin and Tardif 1988; Hammer and Hopper 1997). Increasing sample size of mice will not remedy this situation because chamber environment is confounded with photoperiod. It is, therefore, not possible to estimate experimental error, and inferential statistics cannot be applied. Analysis should be restricted to descriptive statistics only. The study should be re-designed either to allow replication across several chambers, or if chambers are limited, as a multi-batch design replicated at two or more time points.

Sacrificial pseudo-replication occurs when there are multiple replicates within each treatment arm, the data are structured as a feature of the design (such as pairing, clustering, or nesting), but design structure is ignored in the analyses. The units are treated as independent, so the degrees of freedom for testing treatment effects are too large. Sacrificial pseudo-replication is especially common in studies with categorical outcomes when the χ^2 test or Fisher's exact test is used for analysis (Hurlbert and White 1993; Hurlbert 2009).

Example: Sunfish Foraging Preferences

Dugatkin and Wilson (1992) studied feeding success and tankmate preferences in 12 individually marked sunfish housed in two tanks. Preference was evaluated for each fish for all possible pairwise combinations of two other tankmates. There were 2 groups $\times$ 60 trials, per group $\times$ 2 replicate sets of trials, for a total of 240 observations. They concluded that feeding success was weakly but statistically significantly correlated with aggression ($P < 0.001$) based on 209 degrees of freedom, and that fish in each group strongly preferred ($P < 0.001$) the same individual in each of the two replicate preference experiments, based on 60 observations.

The actual number of experimental units is 12, with 6 fish per tank. The correct degrees of freedom for the regression analysis is 4, not 209. Suggested analyses for preference data included one-sample *t*-tests with 5 degrees of freedom or one-tailed Wilcoxon matched-pairs test with $N = 12$. Correct analyses would produce much larger *P*-values, suggesting that interpretation of these data requires substantial revision (Lombardi and Hurlbert 1996).

Temporal (or spatial) pseudo-replication occurs when multiple measurements are obtained sequentially on the same experimental units, but analysed as if they represent an individual experimental unit. Sequential observations (or repeated measures) are correlated within each individual. Repeated measures increase the precision of within-unit estimates, but the number of repeated measures do not increase the power for estimating treatment effects.

Example: Tumour Proliferation in Mouse Models of Cancer

Sequential measurements of solid tumour volume in mice are commonly reported as a measure of disease progression or response to an intervention. Mull et al. (2020) tested the effects of low-dose UCN-01 to promote survival of tumour-bearing mice with lower tumour burden. Mice in four treatment groups were weighed daily for 30 days, then twice weekly to day 75. Differences in tumour volume between groups were assessed by t-tests and one-way ANOVA at five time points.

This is temporal pseudo-replication because the same groups of mice are repeatedly sampled over time, but separate hypothesis tests were performed at different time points. However, successive observations on the same mice are correlated, and sample size is expected to decline as mice die or are humanely euthanised at different times during the study. Traditional ANOVA or repeated-measures ANOVA methods cannot handle missing data or imbalance in the number of repeated responses and do not incorporate the actual correlation structure of the data. Mixed models are much more appropriate, because the true variation in the repeated measurements can be modelled directly by incorporating time dependencies and allowing customisation of the correlation structure; they can also accommodate missing data due to subject loss.

References

Blainey, P., Krzywinski, M., and Altman, N. (2014). Replication. *Nature Methods* 11: 879–880.

Cox, D.R. and Donnelly, C.A. (2011). *Principles of Applied Statistics*. Cambridge: Cambridge University Press.

Dugatkin, L.A. and Wilson, D.S. (1992). The prerequisites for strategic behaviour in bluegill sunfish, *Lepomis macrochirus*. *Animal Behaviour* 44: 223–230.

Hammer, P.A. and Hopper, D.A. (1997). Experimental design. In: *Plant Growth Chamber Handbook* (ed. R. W. Langhans and T.W. Tibbitts), 177–188. Iowa State University NCR-101 Publication No. 340. https://www.controlledenvironments.org/wp-content/uploads/sites/6/2017/06/Ch13.pdf.

Hurlbert, S.H. and White, M.D. (1993). Experiments with freshwater invertebrate zooplanktivores: quality of statistical analysis. *Bulletin of Marine Science* 53: 128–153.

Hurlbert, S. (2009). The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology* 123 (4): 434–443.

Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187–211.

Jung, S.-H. and Young, S.S. (2012). Power and sample size calculation for microarray studies. *Journal of Biopharmaceutical Statistics* 22: 30–42.

Lazic, S.E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience* 11: 5. https://doi.org/10.1186/1471–2202–11–5.

Lazic, S.E. and Essioux, L. (2013). Improving basic and translational science by accounting for litter-to-litter variation in animal models. *BMC Neuroscience* 14: 37.

Lazic, S.E., Clarke-Williams, C.J., and Munafò, M.R. (2018). What exactly is 'N' in cell culture and animal experiments? *PLoS Biology* 16: e2005282.

Lee, M.-L.T. and Whitmore, G.A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine* 21: 3543–3570.

Lin, W.-J., Hsueh, H.-M., and Chen, J.J. (2010). Power and sample size estimation in microarray studies. *BMC Bioinformatics* 11: 48. https://doi.org/10.1186/1471–2105–11–48.

Lombardi, C.M. and Hurlbert, S.H. (1996). Sunfish cognition and pseudoreplication. *Animal Behaviour* 52: 419–422.

Millar, R.B. and Anderson, M.J. (2004). Remedies for pseudoreplication. *Fisheries Research* 70: 397–407. https://doi.org/10.1016/j.fishres.2004.08.016.

Mull, B.B., Livingston, J.A., Patel, N. et al. (2020). Specific, reversible G1 arrest by UCN-01 *in vivo* provides cytostatic protection of normal cells against cytotoxic chemotherapy in breast cancer. *British Journal of Cancer* 122 (6): 812–822. https://doi.org/10.1038/s41416–019–0707–z.

Potvin, C. and Tardif, S. (1988). Sources of variability and experimental designs in growth chambers. *Functional Ecology* 2: 123–130.

Taylor, S.C. and Posch, A. (2014). The design of a quantitative Western Blot. *BioMed Research International* https://doi.org/10.1155/2014/361590.

van Belle, G. (2008). *Statistical Rules of Thumb*, 2nd edition. New York: Wiley.

Vaux, D., Fidler, F., and Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant? *EMBO Reports* 13: 291–296.

3 Ten Strategies to Increase Information (and Reduce Sample Size)

CHAPTER OUTLINE HEAD

3.1 Introduction

Reduction of animal numbers is a key tenet of the 3Rs strategy, but at times may seem to conflict with the goal of maximising statistical power. Large power results in part from increasing sample size. However, a large sample size does not guarantee adequate power, and high power alone does not ensure that results are informative. This section outlines ten complementary strategies for maximising experimental signal and reducing noise, and therefore increasing the information content of study data. Highlighted are strategies for reducing experimental variation before, rather than after, the experiment is conducted. Incorporating all ten strategies will also increase experimental efficiency – the ability of an experiment to achieve study objectives with minimal expenditure of time, money, and animals.

The ten strategies are as follows:

1. 'Well-built' research questions
2. Structured inputs (statistical study designs)
3. Reduce variation I: Process control
4. Reduce variation II: Research animals
5. Reduce variation III: Statistical control
6. Appropriate comparators and controls
7. Informative outcomes
8. Minimise bias
9. Think sequentially
10. Think 'right-sizing', not 'significance'

3.2 The 'Well-Built' Research Question

Once the investigator has identified an interesting clinical or biological research problem, the challenge is to turn it into an actionable, focused, and

A Guide to Sample Size for Animal-based Studies, First Edition. Penny S. Reynolds.

testable question. A well-constructed research question consists of four concept areas: the study population or problem of interest, the test intervention, the comparators or controls, and the outcome. Format is modified according to study type (Box 3.1 and Figure 3.1).

Structuring the research question enables clear identification and discrimination of *causes* (factors that are manipulated or serve as comparators), *effects* (the outcomes that are measured to assess causality), and the *test platform* (the animals used to assess cause and effect). Breaking the research question into components allows the identification and correction of metrics that are otherwise poorly defined or unmeasurable.

BOX 3.1
The 'Well-Built' Research Question

***Experimental/intervention studies*: PICOT**

Population/**P**roblem
Intervention
Comparators/**C**ontrols
Outcome
Time frame, follow up

***Observational studies*: PECOT**

Population/**P**roblem
Exposure
Comparators
Outcome
Time frame, follow up

***Diagnostic studies*: PIRT**

Population/**P**roblem
Index test
Reference/gold standard
Target condition.

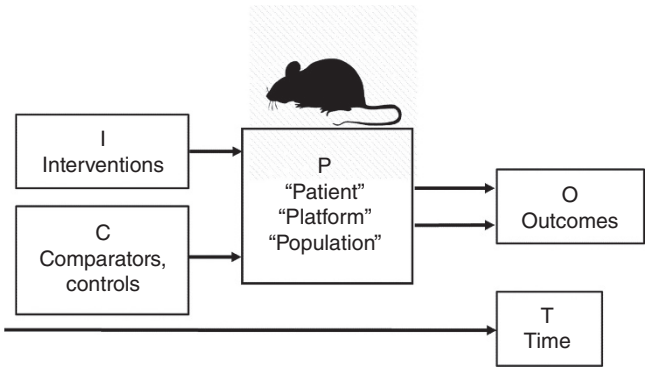

Figure 3.1: System diagram for the 'well-built' research question.

A well-constructed research question is essential for effective literature searches. Comprehensive literature reviews provide current evidence-based assessments of the scientific context, the research gaps to be addressed, suitability of the proposed animal and disease model, and more realistic assessments of potential harms and benefits of the proposed research (Ritskes-Hoitinga and Wever 2018; Ormandy et al. 2019). Collaborative research groups such as CAMARADES (`https://www.ed.ac.uk/clinical-brain-sciences/research/camarades/about-camarades`) and SYRCLE (`https://www.syrcle.network/`) are excellent resources for certain specialities such as stroke, neuropathic pain, and toxicology, and provide a number of e-training resources and tools for assessing research quality. Construction of the research question in the PICOT framework was originally developed for evidence-based medicine. Information on constructing research questions and designing literature searches can be obtained from university library resources sections and the Oxford Centre for Evidence-based Medicine website.

The research question dictates formation of both the *research hypothesis* and related *statistical hypotheses*. These are often confused or conflated. The *research hypothesis* is a testable and quantifiable proposed explanation for an observed or predicted relationship between variables or patterns of events. It should be rooted in a plausible mechanism as to why the observation occurred. One or more testable predictions should follow logically from the central hypothesis ('If A happens, then B should occur, otherwise C'). A description of the scientific hypothesis provides justification for the experiments to be performed, why animals are needed, and rationale for the species, type or strain of animals, and justification of animal numbers.

The *statistical hypothesis* is a mathematically-based statement about a specific statistical population

parameter. Hypothesis tests are the formal testing procedures based on the underlying probability distribution of the relevant sample test statistic. Choice of statistical test will depend upon the statistical hypothesis, the study design, the types of variables (continuous, categorical, ordinal time to event), designated as inputs or outcomes. Statistical hypotheses should be a logical extension of the research hypothesis (Bolles 1962). However, the research hypothesis may not immediately conform to any one statistical hypothesis, and multiple statistical hypotheses may be required to adequately test all predictions generated from the research hypothesis.

Example: Research Versus Statistical Hypotheses

Based on a comprehensive literature review, an investigator determined that ventricular dysrhythmia after myocardial infarction is associated with high risk of subsequent sudden cardiac arrest in humans (*clinical observation, clinical pattern*). The investigator wished to design a series of experiments using a mouse model of myocardial infarction to test the effects of several candidate drugs with the goal of reducing sudden cardiac death.

Scientific hypothesis. Pharmacological suppression of ventricular dysrhythmia should result in clinically important reductions in the incidence of sudden cardiac death.

Research question. In a mouse model of dysrhythmia following myocardial infarction (P), will drug X (I) when compared to a saline vehicle solution (C) result in fewer deaths (O) at four weeks post-administration (T)?

Quantified outcomes. Number of deaths (n) in each group and proportion of deaths (p) in each group.

Statistical hypothesis. The null hypothesis is that of no difference in the proportion of deaths for mice treated with drug X (p_X) versus the proportion of deaths for mice treated with control C (p_C) is H_0 $p_X = p_C$ or H_0: $p_X - p_C = 0$.

3.3 Structured Inputs (Experimental Design)

The design of an animal-based study will affect estimates of sample size. Good study designs are an essential part of the 3Rs (Russell and Burch 1959; Kilkenny et al. 2009; Karp and Fry 2021; Gaskill and Garner 2020; Eggel and Würbel 2021). Rigorous, statistically-based experimental designs consist of the formal arrangement and structuring of independent (or explanatory) variables hypothesised to affect the outcome. (Box 3.2). The optimum design will depend on the specific research problem addressed. However, to be fit for purpose, all designs must facilitate discrimination of signal from noise, by identifying and separating out contributions of explanatory variables from different sources of variation (Reynolds 2022). By increasing the power to detect real treatment differences, a properly designed experiment requires the use of far fewer time, money, and resources (including animals) for the amount of information obtained.

Well-designed studies start with a well-constructed research question and well-defined input and output variables. A good design also incorporates specific design features that ensure results are reliable and valid. These include correct specification of the unit of analysis (or experimental unit), relevant inclusion and exclusion criteria, bias minimisation methods (such as allocation concealment and randomisation; Section 3.10), and minimisation of variation (Addelman 1970). Useful designs for animal-based research include completely randomised design, randomised complete block designs, factorial designs, and split-plot designs (Festing and Altman 2002; Montgomery 2017; Festing 2020; Karp and Fry 2021).

BOX 3.2
Statistical Design of Experiments: Components

Study design. Formal structuring of input or explanatory variables according to statistically-based design principles

Study design features. Unit of analysis (experimental unit), inclusion/exclusion criteria, bias minimisation methods, sources of variation.