Special Issue: Biostatistics

Laboratory Animals Limited

Laboratory Animals
2024, Vol. 58(5) 404–410

DOI: 10.1177/00236772241276806
journals.sagepub.com/home/lan

Sage

Study design: think 'scientific value' not '*p*-values'

Penny S Reynolds[1,2]

Abstract

Statistically based experimental designs have been available for over a century. However, many preclinical researchers are completely unaware of these methods, and the success of experiments is usually equated only with '$p < 0.05$'. By contrast, a well-thought-out experimental design strategy provides data with evidentiary and scientific value. A value-based strategy requires implementation of statistical design principles coupled with basic project management techniques. This article outlines the three phases of a value-based design strategy: proper framing of the research question, statistically based operationalisation through careful selection and structuring of appropriate inputs, and incorporation of methods that minimise bias and process variation. Appropriate study design increases study validity and the evidentiary strength of the results, reduces animal numbers, and reduces waste from noninformative experiments. Statistically based experimental design is thus a key component of the 'Reduction' pillar of the 3R (Replacement, Reduction, Refinement) principles for ethical animal research.

Keywords

Bias, experimental design, experimental unit, factor, pico, process control, randomisation, variance

Date received: 29 January 2024; accepted: 21 February 2024

> One general way in which great reduction [of animal numbers] may occur is by the right choice of strategies in the planning and performance of whole lines of research.
>
> —Russell and Burch, 1959.

A successful research study creates scientific value by maximising the amount of reliable and useful scientific information obtained for the minimum number of animals.[1] Unfortunately, many researchers equate the success of an experiment with '$p < 0.05$'. Chasing statistical significance as a research goal is often equated with an increased chance of the work being published and cited, but could foster questionable research practices, if not outright research malpractice.[2] Reliance on *p*-values alone as indicators of scientific merit will also result in the waste of all animals used if the experiment does not produce statistically significant results and hence is not published.[3]

Evidentiary strength is determined not by downstream *p*-values, but by upstream study validity. Because validity is methods-based, it must be built into the experiment before data are collected. Scientific value is created by a comprehensive and planned experimental strategy. Validity cannot be produced by large sample sizes or statistical analyses after the fact.[4,5]

There are three phases to the strategic planning of an experiment: quantitative framing of the research question (Will the data be relevant?), operationalising the research question (Will the data be interpretable?) and implementation of the design (Will the data be reliable?).

Framing the research question

A 'good' research question is simple, straightforward and, above all, answerable. It is not necessarily startlingly novel. The precise formulation of the research

[1]Department of Anesthesiology, University of Florida College of Medicine, Gainesville, USA
[2]Department of Small Animal Clinical Sciences, University of Florida College of Veterinary Medicine, Gainesville, USA

Corresponding author:
Penny S Reynolds, Department of Anesthesiology, University of Florida College of Medicine; Department of Small Animal Clinical Sciences, University of Florida College of Veterinary Medicine, 1600 SW Archer Rd, Gainesville FL 32610, USA.
Email: PReynolds@anest.ufl.edu

question renders the research hypothesis focused, specific and measurable, and ensures the data will be relevant to the hypothesis to be tested.

The format of a practical research question is described elsewhere in detail[6] and is therefore only briefly outlined here. The first step is to clearly define the input and output variables (I/O). Inputs are the explanatory or independent variables thought to affect the outputs (response variables, outcomes). The question is then refined by application of the PICOT acronym: platform (the animal model), intervention, comparators and controls, outputs, and time. Output variables are prioritised to discriminate those most critical to the central hypothesis and study objectives from those that may supply useful information but are otherwise lower priority and only add unnecessary complexity to the experiment. The study is both powered and interpreted off the output variable with the highest priority (primary outcome). The number of sampling times required (once, before–after, many) must be incorporated into the design. The time frame must also be clearly defined, especially if the outcome is survival to some predetermined time or humane endpoint.

Operationalising the research question

The research question is operationalised by the experimental design. The correct design enables more reliable discrimination of the true effect (signal) from unwanted variation (noise). First, the design directs subsequent data collection; it cannot be imposed after data are collected. Second, both the design structure and the type of response variables determine the most appropriate statistical methods that can be used to analyse the resulting data, and ensure the correct error terms are used for testing statistical significance of differences.[7]

There are three steps to constructing an appropriate design: choosing appropriate inputs, identifying the unit of replication, and selecting the design structure.

Choice of input variable

In the design context, the independent variables are referred to as factors. Factors are the input variables chosen by the investigator specifically to study their effect on the response. Factor levels consist of a limited set of prespecified values for each factor. Qualitative or categorical factors may be nominal (e.g. sex: male, female) or ordinal (e.g. age class: neonate, juvenile, adult, aged). If the factor is continuous, a range of reasonable values should be chosen to bracket the biologically likely range of responses. For example, dose concentrations (levels) for a given drug (the factor) could be set at four levels: minimum (or zero) concentration (e.g. saline or vehicle control with no active drug), the maximum tolerable concentration, and two intermediate concentrations.[7]

Identify unit of replication

A treatment in the sense used by statisticians is one of each combination of input factors and factor levels. Treatments are randomly assigned by the researcher to the experimental units (EUs), the smallest independent entity to which a treatment is applied. The total number of EUs in an experiment is the sample size. A common problem when multiple measurements are made on a single EU is confusion of the sample size with the number of observational units (OUs), the smallest independent entity from which measurements are obtained. The conflation of sample size and OUs is pseudo-replication, resulting in falsely inflated sample sizes, spurious precision, and increased false positive rates.[8] Therefore, EUs and OUs must be clearly identified a priori because the number of EUs may not necessarily be the same as the total number of animals, or the number of OUs may not be the same as the number of EUs.

Select design structure

The design is the formal structuring of the factors and factor levels. Basic designs are briefly described below (Figure 1).

Completely randomised design. The completely randomised design (Figure 1(a)) typically consists of one input variable (factor) with k levels for k treatments. This design is not very practical for most animal studies when multiple factors are to be studied. It also lacks precision because any variation not accounted for by the model goes into the error term, inflating the error variance and reducing power. Because treatments are allocated to EUs by simple randomisation, when total sample sizes are small there is the potential for considerable imbalance of sample sizes across treatments, leading to 'undesirable' experimentation patterns, increased variance and reduced precision to detect true effects. Therefore, some form of restricted randomisation is usually recommended to obtain a better balance of sample sizes across treatment arms, or balance across potential confounders or nuisance variables. Nuisance variables may influence the response to some extent, but are not of direct interest for testing the central hypothesis.

Randomised complete block designs. Randomised complete block designs (RCBDs; Figure 1(b)) are a type of restricted randomised design. Randomisation is conducted independently for each block, and each

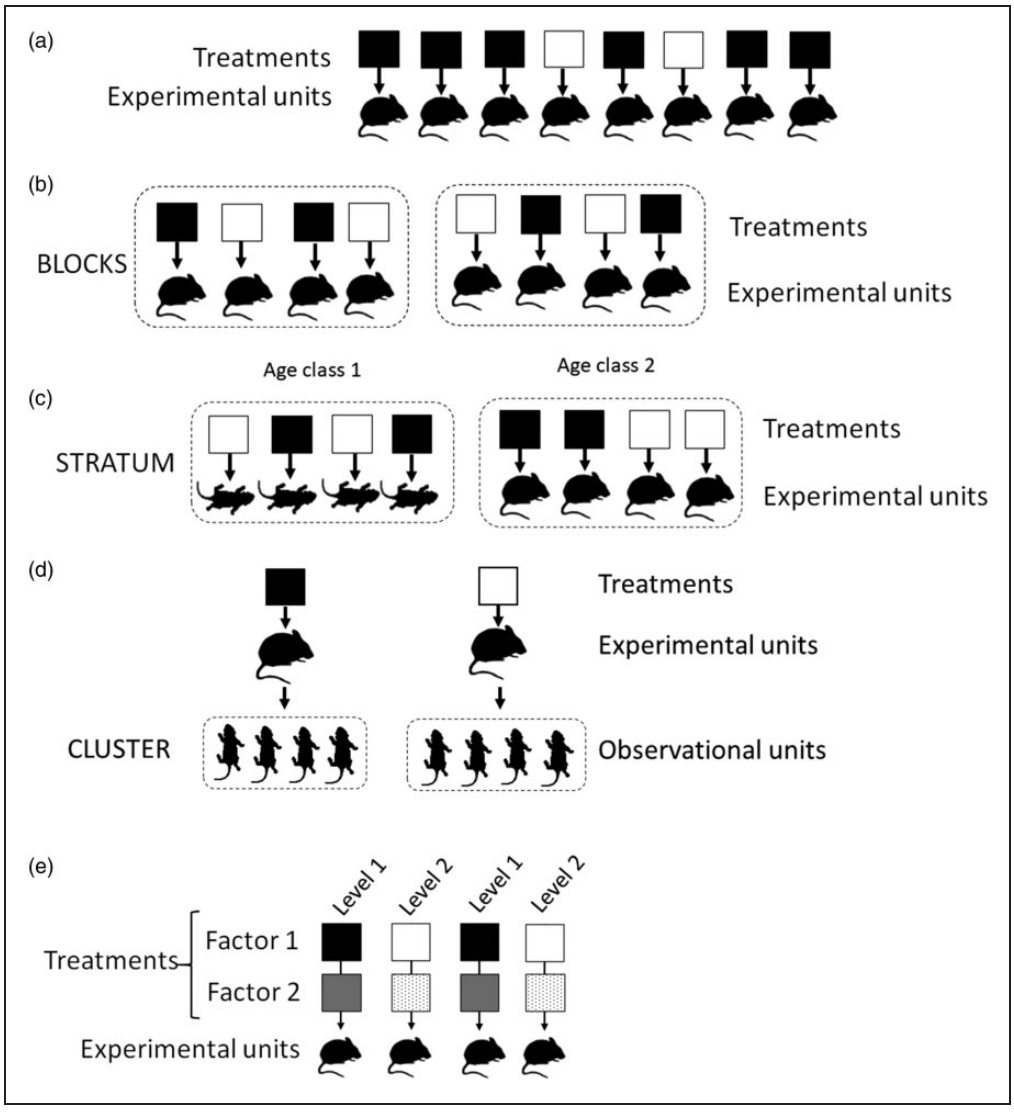

Figure 1. Schematics of basic experimental designs: (a) completely randomised design on one factor with two levels. When treatments are allocated to experimental units by simple randomisation, sample size imbalance may occur. (b) Randomised complete block design. Randomisation occurs within each block, ensuring equal sample sizes per treatment. (c) Stratified randomisation. The randomised design on one factor with two levels is stratified by age class (pup, adult) with treatments randomly allocated to experimental units within each stratum. (d) Nested (hierarchical) designs, with the experimental unit designated at the cluster level or within- the cluster and (e) Factorial design on two factors each at two factor levels for four treatments. Each treatment is randomly allocated to an experimental unit.

block contains all treatments, so sample sizes can be easily balanced. RCBDs improve power and precision by removing the contribution of variation between EUs from the experimental error. The blocks are subgroups of similar EUs classified by a categorical blocking variable. Blocking variables can be physical (e.g. donor, cage, litter, tank, laboratory, technician) or time-based (e.g. day, week). Stratified randomisation (Figure 1(c)) reduces imbalance across potential confounders, usually some attribute of the experimental subjects (such as age class, size category, tumour stage).

Nested (hierarchical) designs. Nested (hierarchical) designs (Figure 1(d)) are useful when OUs or EUs are contained within a larger naturally occurring organisational unit or cluster (e.g. pups clustered within litter, wells clustered within plate clustered within mouse). Randomisation can occur at either the cluster level (clusters are the EU) or the subject level (subjects within clusters are the EU), so the unit of randomisation and the level at which interventions are to be applied must be identified beforehand.[9]

Factorial designs. Factorial designs (Figure 1(e)) can consist of any combination of factors and number of levels, but the classic workhorse design is k factors at two levels each, resulting in 2^k treatments. In contrast to the traditional trial-and-error or one-factor-at-a-time (OFAT) approach, factorial designs allow simultaneous and direct examination of multiple factors and interactions, with fewer experimental runs. Factorial designs are not only more efficient and animal-sparing than conventional OFAT designs, but their potential for evaluating interactions makes this design more likely to flag up unexpected, potentially serendipitous, results, making them especially useful for exploratory studies.[10]

More complex experimental designs can involve a combination of two or more of the basic designs (e.g. factorial with blocking; split-plot RCBD). Modern computer-assisted optimal and definitive screening designs will be especially useful when there are large numbers of candidate factors to be studied and conventional designs are not applicable (for example, drug discovery).[11] Because of the increasing availability of easy-to-use commercial software, computer-assisted designs should find much wider application to exploratory animal-based research than is the case at present.

Implementing the experiment

Once an appropriate design for the study has been identified, the researcher must then consider several statistically based methods for its implementation: how treatments are to be allocated to the EUs, the order in which data are to be collected, how data are to be assessed, and how the results of the experiment can be used to reliably inform the next stages of the study. Bias minimisation methods ensure that the data are reliable by minimising systematic and cognitive biases in treatment allocation, data collection, analysis and interpretation. Variance minimisation methods ensure that the data are consistent by reducing as much variation as possible in the physical execution of the experiment. Sequential experimentation is a total strategic approach where the results of the previous experiment inform the design of the next set of experiments, increasing the certainty of success of the final definitive experiments.

Bias minimisation

Bias is the systematic deviation of estimates from the true value. Although bias can occur at all stages of the research cycle, most sources of bias can be minimised only during planning and design phases. Large sample size does not make an experiment any less biased, and bias cannot be removed statistically after data are collected. The two major methods of bias minimisation are *randomisation* and *allocation concealment*.[4]

Randomisation. Randomisation is the formal, technical, probability-based process of assigning a specific treatment to an EU with a given probability. 'Random' does not mean the allocation or sequence order is haphazard, ad hoc or unplanned. Treatments should be allocated to the EUs with the assistance of an appropriate computer-driven algorithm tailored to accommodate the specific experimental design for the study. Randomisation should also be deployed to the minimise biases associated with run order when EUs are processed sequentially (random sequence allocation).[4,12] Use of random number tables or coin tossing to determine either allocation or sequence order are vulnerable to human interference and manipulation and should be discouraged. Randomisation minimises systematic error and sampling bias in the conduct of the experiment. Critically, it is also the cornerstone of statistical hypothesis testing. The distribution of the test statistic is the appropriate reference for hypothesis testing only if the treatment allocation was randomised. No meaningful or valid hypothesis test can be based on non-random 'designs'. The unfortunately common practice of allocating 'groups' or 'cohorts' of animals to some intervention or control condition is usually not probability-based, and is also misleading and erroneous.[13] There is no scientifically valid justification for failing to randomise an animal-based experimental study.

Allocation concealment. Allocation concealment (blinding, masking) minimises the cognitive biases that can occur during any stage of the experiment: treatment allocation, experiment conduct, outcome assessment and data analysis. Allocation concealment involves disguising from some or all personnel which treatment was received by which EU. It is built into study procedures by the coded relabelling of EU and treatment identifiers. Blinding is especially critical for outcomes requiring subjective evaluation or judgement, such as histology, behaviour or clinical progress.[4]

Variance minimisation

Statistical methods for controlling variation include blocking, stratification and clustering. By grouping EUs into smaller homogenous subsets based on one or more classification – or nuisance – variables. These methods remove unwanted variation when incorporated into the experimental design.[10]

Non-statistical methods should also be used to reduce variation in both experimental subjects and experimental processes. Housing animals in barren

so-called standardised housing without companions, enrichment or opportunities for normal activity and behaviours is inhumane, and the associated stress greatly increases uncontrolled variation.[14,15] In contrast, between- and within-animal variation can be minimised by routine implementation of refined husbandry, refined handling, and habituation to experimental procedures before experimentation begins.[16]

Methods for minimising process variation in the experiment itself are well documented in the quality improvement and project management literature. Experiments are usually complex with many moving parts, so it can be quite challenging to ensure consistent and reliable performance of all personnel. Performance variation can be greatly reduced by standardisation of protocols and operational procedures, training all personnel up to consistent best-practice standards, and performing regular quality assurance checks on operator performance, equipment calibration and drift, and data anomalies.[17,18] Simple non-technical management tools to check the stability of experimental processes and performance include process maps, written standard operating procedures, checklists and performance-tracking graphics.[17]

Sequential experimentation. Sequential experimentation is a high-information alternative to the usual practice of large, sprawling experiments consisting of numerous OFAT comparisons. This strategy involves running small, planned experiments in series, with each experiment providing feedback to inform the methodology of the next. Therefore, rather than the investigator resorting to trial-and-error or tweaking experiments, modifications to sample size and design can be intelligent and data-driven. Sequential feedback can also guide decisions to proceed with data-directed modifications or abort unpromising experiments. Sequential data-directed experimentation ultimately increases power and efficiency and reduces costs.[7,10]

A pilot phase is recommended to identify possible problems and standardise procedures before animals are used. Optimal screening and reduction designs are appropriate for intermediate studies with several input variables, with the objective of winnowing out the most promising factors from a larger number of potential candidates. Relative importance and effect sizes are easily assessed by graphing the results for main effects and interactions. Consequently, the final testing phase, involving appropriately powered and streamlined confirmatory trials,[19] will have a higher probability of success.

What comes next? Analysis

Both the experimental design structure and type of response variable will determine what analysis methods will be most appropriate for the resulting data ('design before inference'). Conventional analysis of variance (ANOVA) models may be appropriate for most analysis purposes if the design structure is preserved and the assumptions of normality, homogeneity of variance, and independence of observations are met. However, non-normally distributed response data and serially correlated observations are typical of much biological research. Examples include count data (Poisson or negative binomial distribution), time to event/survival/censored data (exponential, gamma, Weibull distribution), and severity rank or score category data (ordinal, multinomial distribution). Therefore, more sophisticated analysis methods will be necessary to accommodate the specific form of the response data to be analysed in combination with the design structure used for the experiment. Generalised linear mixed models are an extension of the linear additive models that underpin classic ANOVA models and can be adapted to accommodate most data types and study designs. Cox proportional hazards regression was specifically developed for time to event (survival time) responses where the fixed factors of the design and covariates such as age, weight or baseline measurements can be accommodated (unlike the commonly used log-rank test). Serially correlated observations require incorporation of a repeated measures component in one or more of the factors.[20]

Summary

Study design may be the least understood statistical concept in animal-based research. Unfortunately, poor or non-existent training of researchers in the principles of statistically based design is one of the major limitations to research quality reform. Many researchers still consider study design only in terms of the technical aspects of the experiment – the materials and methods – with statistical analyses applied indiscriminately and often inappropriately to the resulting data. Over-reliance on small *p*-values to interpret the results without considering the internal validity of the study further distorts the evidence base.

John Nelder once referred to significance tests and *p*-values as 'non-scientific statistics'. In contrast, 'scientific statistics' require a complete understanding of the statistical process, starting with an actionable question that directs construction of a statistically based experimental design, which is analysed by bespoke, rather than boilerplate, statistical methods.[21] Statistically based experimental designs and design principles have been available for over a century. For animal-based research to properly inform translational research, it is high time that there is a shift in focus from 'getting things done' to 'doing things right'. A good design

ensures that animal numbers are appropriate for meeting the scientific objectives of the study and that the study results will be valid and reliable. Therefore, statistically based experimental design is a key component of the Reduction pillar of the 3R (Replacement, Reduction, Refinement) principles for ethical animal research.

This is only a brief overview of the principles and procedures involved in the design of animal-based experiments. The researcher is encouraged to consult the resource list in Supplementary File 1 for more information.

Acknowledgements

The author thanks Scott Hunter, University of Florida College of Medicine, Department of Anesthesiology Communications & Publishing Office for editorial assistance, and the two anonymous reviewers for their thoughtful and constructive comments that greatly improved the article.

Conflict of interests

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical statement

The subject matter did not involve human or animal subjects, therefore, no ethical board approval was necessary.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Penny S Reynolds https://orcid.org/0000-0001-7480-6275

References

1. Russell WMS and Burch RL. *The principles of humane experimental technique*. London, United Kingdom: Methuen, 1959.
2. Ware JJ and Munafò MR. Significance chasing in research practice: causes, consequences and possible solutions. *Addiction* 2015; 110: 4–8.
3. ter Riet G, Korevaar DA, Leenaars M, et al. Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. *PLoS One* 2012; 7: e43404.
4. Bespalov A, Wick K and Castagné V. Blinding and randomization. In: Bespalov A, Michel M and Steckler T (eds) *Good research practice in non-clinical pharmacology and biomedicine*. USA: Springer Cham, 2019.
5. Sargeant JM, Brennan ML and O'Connor AM. Levels of evidence, quality assessment, and risk of bias: evaluating the internal validity of primary research. *Front Vet Sci* 2022; 9: 960957.
6. Reynolds PS. The well-built research question. *Lab Anim (NY)* 2023; 52: 221–223.
7. Czitrom V. Guidelines for selecting factors and factor levels for an industrial designed experiment. In: Khattree R and Rao CR (eds) *Handbook of statistics*. Amsterdam, the Netherlands: Elsevier Science, 2003, 3–32.
8. Lazic SE, Clarke-Williams CJ and Munafò MR. What exactly is 'N' in cell culture and animal experiments? *PLoS Biol* 2018; 16: e2005282.
9. Aarts E, Dolan CV, Verhage M, et al. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neurosci* 2015; 16: 94.
10. Box GEP, Hunter JS and Hunter WG. *Statistics for experimenters: design, innovation, and discovery*. 2nd ed. Hoboken NJ: Wiley-Interscience, John Wiley & Sons, 2005.
11. Trutna L, Spagon P, del Castillo E, et al. Process improvement In: Croarkin C and Tobias P (eds) *NIST/SEMATECH e-handbook of statistical methods*. Gaithersburg, MD: National Institute of Standards and Technology, 2013.
12. Altman DG. Randomisation. *BMJ* 1991; 302: 1481–1482.
13. Festing MFW. The 'completely randomised' and the 'randomised block' are the only experimental designs suitable for widespread use in pre–clinical research. *Sci Rep* 2020; 10: 17577.
14. Martin B, Ji S, Maudsley S, et al. 'Control' laboratory rodents are metabolically morbid: why it matters. *Proc Natl Acad Sci USA* 2010; 107: 6127–6133.
15. Richter SH, Garner JP and Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat Methods* 2009; 6: 257–261.
16. Gouveia K and Hurst JL. Improving the practicality of using non-aversive handling methods to reduce background stress and anxiety in laboratory mice. *Sci Rep* 2019; 9: 20305.
17. Reynolds PS, Fisher BJ, McCarter J, et al. Informing efficient pilot development of animal trauma models through quality improvement strategies. *Lab Anim* 2019; 53: 394–404.
18. Preece DA. The design and analysis of experiments: what has gone wrong? *Util Math* 1982; 21: 201–244.
19. Mogil JS and Macleod MR. No publication without confirmation. *Nature* 2017; 542: 409–411.
20. Stroup WW. Rethinking the analysis of non-normal data in plant and soil science. *Agronomy J* 2015; 107: 811–827.
21. Nelder J. Statistics for the millennium: from statistics to statistical science. *Statistician* 1999; 48: 257–269.

Conception de l'étude: penser « valeur scientifique » et non « valeurs P »

Résumé

Les conceptions expérimentales reposant sur des statistiques existent depuis plus d'un siècle. De nombreux chercheurs précliniques ignorent cependant complètement ces méthodes, et le succès des expériences n'est généralement assimilé qu'à « $P < 0,05$ ». Une stratégie de conception expérimentale bien pensée fournit pourtant des données ayant une valeur probante et scientifique. Une stratégie fondée sur la valeur exige la mise en œuvre de principes de conception statistique associés à des techniques de base de gestion de projet. Cet article décrit les trois phases d'une stratégie de conception basée sur la valeur: le bon cadrage de la question de recherche, l'opérationnalisation basée sur les statistiques grâce à une sélection et une structuration minutieuses des intrants appropriés, et l'incorporation de méthodes qui minimisent les biais et les variations de processus. Une étude ayant une conception appropriée aura davantage de validité et ses résultats auront une plus grande force probante, ce qui permettra de minimiser le nombre d'animaux utilisés et de réduire les déchets provenant d'expériences non informatives. La conception expérimentale basée sur des statistiques est donc un élément clé du pilier réduction des principes des 3R (remplacement, réduction, raffinement) pour la recherche animale éthique.

Studiendesign: Priorisierung von ,,wissenschaftlichem Wert ' statt ,,P-Werten'

Abstract

Statistisch fundierte Versuchspläne gibt es schon seit über einem Jahrhundert. Vielen in der präklinischen Forschung Tätigen sind diese Methoden jedoch gänzlich unbekannt, und der Erfolg von Experimenten wird meist nur mit ,,$P < 0,05$ ' gleichgesetzt. Hingegen liefert eine gut durchdachte Versuchsplanungsstrategie Daten von beweiskräftigem und wissenschaftlichem Wert. Eine wertorientierte Strategie erfordert die Umsetzung statistischer Planungsprinzipien in Verbindung mit grundlegenden Projektmanagementtechniken. In diesem Artikel werden die drei Phasen einer wertbasierten Planungsstrategie umrissen – die richtige Formulierung der Forschungsfrage, die statistisch fundierte Operationalisierung durch sorgfältige Auswahl und Strukturierung geeigneter Inputs und die Einbeziehung von Methoden zur Minimierung von Verzerrungen und Prozessvariationen. Ein angemessenes Studiendesign erhöht die Validität der Studie und die Beweiskraft der Ergebnisse, reduziert die Zahl der Versuchstiere und verringert die Verschwendung durch nichtinformative Experimente. Eine statistisch fundierte Versuchsplanung ist daher eine Schlüsselkomponente der Komponente ,,Reduktion ' des 3R-Prinzips (Replacement, Reduction, Refinement) für ethische Tierforschung.

Diseño del ensayo: Piense en 'valor científico' y no en 'valores P'

Resumen

Los diseños experimentales basados en la estadística llevan estando disponibles desde hace más de un siglo. No obstante, muchos investigadores preclínicos desconocen por completo estos métodos y el éxito de los experimentos suele equipararse únicamente con '$P < 0,05$'. Por el contrario, una estrategia de diseño experimental bien pensada proporciona datos con valor probatorio y científico. Una estrategia basada en el valor requiere la aplicación de principios de diseño estadístico junto con técnicas básicas sobre gestión de proyectos. Este artículo describe las tres fases de una estrategia de diseño basada en el valor: el encuadre adecuado de la pregunta de investigación, la operacionalización basada en estadísticas mediante la selección y estructuración cuidadosa de entradas apropiadas, y la incorporación de métodos que minimicen el sesgo y la variación del proceso. Un diseño adecuado de los estudios incrementa la validez de los mismos y la solidez probatoria de los resultados, reduce el número de animales y el despilfarro de experimentos no informativos. El diseño experimental basado en la estadística es, por tanto, un componente clave del pilar de Reducción de los principios de las 3R (Reemplazo, Reducción, Refinamiento) para la investigación ética con animales.