Introduction

The Design and Statistical Analysis of Animal Experiments

Michael F. W. Festing

Key Words: animal experiments; experimental design; sample size; statistics; variation

Research scientists are strongly urged to read all of the papers in this issue if they want to improve the quality and quantity of their research and ensure that their papers are accepted in high-quality journals. There is always room for improvement; some experiments are performed quite badly (Festing 1994, 2000; Festing et al. 2002; Roberts et al. 2002). The main problem seems to be poor training in statistics and a general horror among some biologists of anything to do with mathematics. However, experience shows that most scientists have a keen interest in experimental design and are not nearly as frightened of statistical methods that are planned at the same time as their experiments. For this reason, I believe that courses for research workers should emphasize design aspects. After all, if an experiment is well designed, it is relatively easy to get help with the statistical analysis; but if it is incorrectly designed, it may be impossible to extract any useful information from it.

Animal experiments are usually done to discover something about the biology of, for example, the species, strain, or sex being studied and indirectly to infer something about humans or other target species. Thus, laboratory animals are usually used as "models" of some other species. The use of models involves the following three distinct steps: (1) choice of a suitable model, based on our current knowledge about disease processes in the target species and in potential model organisms; (2) one or more experiments to indicate how the model responds to any applied treatments; and (3) consideration of the relevance of the results for the target species. This issue of *ILAR Journal* is concerned with the second step, the design and analysis of experiments using animals that have already been chosen as likely to be informative indirectly about the target species.

Anyone planning an animal experiment should first consider the important ethical issues involved. A good framework for this consideration is to use Russell and Burch's "3Rs" (Russell and Burch 1959), namely, whether the animal model can be *replaced* by a less or nonsentient alternative such as an insect, nematode, cell culture, or computer simulation. If not, the possibility of *refining* the experiment in an effort to minimize pain and distress for each individual should be considered. Animals should be housed in good conditions, free of pathogens. Surgical techniques should use appropriate anesthesia and analgesia, and humane endpoints should be used. Finally, the number of animals should be *reduced* to the minimum required to achieve the scientific objectives of the experiment. Reduction can be achieved by choosing the most appropriate animal model (Herman 2002), allowing for the fact that it is not always necessary for the models to mimic the human condition exactly (Elsea and Lucas 2002), and by good experimental design and statistical methods, as outlined in this issue.

All animal models are subject to biological variation as a result of genetic and nongenetic variation and the interaction between them. Even genetically identical littermates will vary to some extent as a result of chance developmental effects, social hierarchy, and unequal exposure to environmental influences. Good experimental design aims to control this variation so that it does not obscure any treatment effect, with the statistical analysis being designed to extract all useful information and take into account any remaining variation.

A first step is to review likely sources of variation (Howard 2002). Species, strains, sexes, and individuals differ as a result of genetics; however, there are also many environmental influences that can either bias the results in some way or can contribute so much "noise" that it is impossible to identify the effects of any experimental treatments. Some of this variation is even introduced by the experimenter in the form of measurement error. One of the most serious causes of interindividual variation is clinical or subclinical disease, but fortunately (at least in the case of rats and mice), so-called "specific pathogen-free" animals have been available for many years. This availability does not mean that the problem has disappeared. Barriers break down due to new and existing pathogens somehow gaining entry, particularly when investigators exchange strains of animals; however, at least this source of variation is well

Michael F. W. Festing, M.Sc., Ph.D., D.Sc., CStat., BIBiol., is a Senior Research Scientist at the MRC Toxicology Unit, University of Leicester, UK.

[1]Abbreviation used in this Introduction: LD_{50}, median lethal dose.

known and is under control to the extent possible. Less well known are some of the effects of source of animals, husbandry, housing (including the effects of environmental enrichment), diet, and social interaction among grouped animals. Failure to account for some of these variables at the design stage will certainly reduce the efficiency of any animal experiment.

Practical considerations also must be taken into account. It is important to survey the literature, clarify research questions, select model systems and available skills, and review equipment and staffing. When these practical aspects fall short, it is advisable to identify collaborators and add a statistician to the research team if necessary. This is also the time to choose treatments and outcomes and to formulate a tentative design (Johnson and Besselsen 2002). Choosing the right number of animals (or other "experimental units") is particularly important. Choosing too few may result in missing biologically important effects, whereas choosing too many will waste animals. Statistical power and sample size calculations provide some guidance. Although the formulae are complex, particularly for more advanced designs, it is usually possible to simplify the questions so that a good approximation of the optimum sample size can be obtained using one of the available statistical packages or an interactive web site (Dell et al. 2002).

In any experiment, there are numerous variables, which include the following: (1) "ancillary" variables such as species, strain, sex, husbandry model preparation, and some aspects of the environment, which can be fixed by the experimenter; (2) variables such as diurnal variation, some of the measurement error, and observer differences, which can be taken into account if they are known; and (3) additional variables such as differences between animals, which are difficult or impossible to control. All of these variables form part of the record of the experiment, but all too often, such information is simply discarded. A range of methods are available for dealing with these variables, such as including them in the design by using randomized block or Latin square designs, using them in the statistical analysis in screening for outliers or using techniques such as covariance analysis to increase statistical power, or simply recording them so that other people know exactly what has been done (Gaines Das 2002).

Most experiments are performed using relatively simple designs such as completely randomized, randomized block, and simple factorial designs. However, there are many specialized designs and techniques that can be used effectively in some circumstances. Some experiments are performed repeatedly with only minor changes in the applied treatments. Drug development often involves screening many related compounds using the same animal model and design, and safety and potency testing of vaccines involves testing different batches of product also using the same design. As noted above, large numbers of variables can influence the outcome. If such experiments are optimized by standardizing on the variables that result in the most clear-cut differences, sample size can be reduced or the experiments can be more powerful and able to detect more subtle differences. Advanced factorial design offers one method of assessing which, among a large number of controllable variables, are most important in determining the response and which can safely be ignored. Many combinations of factors such as strain, sex, age, weight, prior treatment, and measurement methods can be tried without using excessive numbers of animals (Shaw et al. 2002). When the optimum set has been identified, simpler designs using this combination of variables can be used.

The use of animals in acute toxicity tests to determine a median lethal dose (LD_{50}[1]) has always been a controversial procedure because it may involve excessive suffering. Several improved methods have now been developed that use fewer animals by relying on a sequential procedure. Such designs are used when the result of a treatment on an individual (or small group) of animals can be obtained quickly so that a decision on how to treat the next individual can be pursued. Typically, the first animal is given a dose thought to approximate the presumed LD_{50} level. If it dies, the next animal is given a lower dose; however, if it survives, the next animal is given a higher dose, and so on until the experiment has homed in on a sufficiently precise estimate of the LD_{50}, at which point the experiment is terminated. The so-called "up-and-down" method described herein uses fewer animals than the conventional test and has been optimized using computer simulations—a statistically and ethically better method than testing dose levels using real animals (Rispin et al. 2002). The "fixed dose procedure" (Whitehead and Curnow 1992) and the "acute toxic class" methods (Schlede et al. 1992) have somewhat similar objectives of classifying the toxicity of chemicals, although they do not result in estimates of the LD_{50}.

The articles in this issue cover so many aspects of what a scientist should and should not do to design a good animal experiment that some scientists may end up in total confusion. The final article titled *Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals* (Festing and Altman 2002) aims to clarify the issues by systematically going through the general principles involved in designing, analyzing, and presenting the results of an experiment. Starting with a clear specification of the objective of the experiment, the scientist should consider the ethical implication of the study, the validity of the model, sources of variation, and how some of these variables can be controlled using formal, appropriately randomized, experimental designs. It is important to choose carefully the independent variable(s) or treatments, the dependent variables (characters or outcomes), and the appropriate numbers of animals. The proposed method of statistical analysis should be determined at the design stage. However, the data should be studied for possible anomalies and may require transformation to a different scale before statistical analysis using parametric or nonparametric methods. Finally, the results should be presented to clarify exactly what was done and what was the outcome, "warts" and all.

The design, analysis, and interpretation of biomedical experiments are best performed with the aid of a good statistical textbook, dedicated statistical software, and advice from a statistician. Anyone reading and understanding the papers presented here cannot fail to do better, more humane animal experiments in the future.

References

Dell R, Holleran S, Ramakrishnan R. 2002. Sample size determination. ILAR J 43:207-213. ⟨http://www.national-academies.org/ilar⟩.

Elsea SH, Lucas RE. 2002. The mousetrap: What we can learn when the mouse model does not mimic the human disease. ILAR J 43:66-79. ⟨http://www.national-academies.org/ilar⟩.

Festing MFW. 1994. Reduction of animal use: Experimental design and quality of experiments. Lab Anim 28:212-221.

Festing MFW. 2000. Common errors in the statistical analysis of experimental data. In: Balls M, van Zeller A-M, Halder ME, eds. Progress in the Reduction, Refinement and Replacement of Animal Experimentation: Developments in Animal and Veterinary Science. Vol. 31a. Amsterdam: Elsevier. p 753-758.

Festing MFW, Altman DG. 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J 43:244-258. ⟨http://www.national-academies.org/ilar⟩.

Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. London: Royal Society of Medicine Press Limited.

Gaines Das R. 2002. The role of ancillary variables in the design, analysis, and interpretation of animal experiments. ILAR J 43:214-222. ⟨http://www.national-academies.org/ilar⟩.

Herman GE. 2002. Mouse models of human disease. ILAR J 43:55-56. ⟨http://www.national-academies.org/ilar⟩.

Howard BR. 2002. The control of variability. ILAR J 43:194-201. ⟨http://www.national-academies.org/ilar⟩.

Johnson PD, Besselsen DG. 2002. Practical aspects to experimental design in animal research. ILAR J 43:202-206. ⟨http://www.national-academies.org/ilar⟩.

Rispin A, Farrar D, Margosches E, Gupta K, Stizel K, Carr G, Greene M, Meyer W, McCall D. 2002. Alternative methods for the LD_{50} test: The up and down procedure for acute oral toxicity. ILAR J 43:233-243. ⟨http://www.national-academies.org/ilar⟩.

Roberts I, Kwan I, Evans P, Haig S. 2002. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. Br Med J 324: 474-476.

Russell WMS, Burch RL. 1959. The Principles of Humane Experimental Technique. London: Methuen & Co. Ltd. [Reissued: 1992, Universities Federation for Animal Welfare, Herts, England.] ⟨http://altweb.jhsph.edu/publications/humane_exp/het-het-toc.htm⟩.

Schlede E, Mischke U, Roll R, Kayser D. 1992. A national validation study of the acute-toxic-class method-an alternative to the LD_{50} test. Arch Toxicol 66:455-470.

Shaw R, Festing MFW, Peers I, Furlong L. 2002. The use of factorial designs to optimize animal experiments and reduce animal use. ILAR J 43:223-232. ⟨http://www.national-academies.org/ilar⟩.

Whitehead A, Curnow RN. 1992. Statistical evaluation of the fixed-dose procedure. Food Chem Toxicol 30:313-324.