REP 1993-018

Animal Welfare and the Statistical Consultant

RICHARD M. ENGEMAN and STEPHEN A. SHUMAKE*

Animal welfare considerations and regulations serve to ensure that experimental animals are used wisely while minimizing stress. Substantial impacts on the use of animals in experiments have resulted, thereby increasing the challenges for designing studies and analyzing data to provide valid inferences. Statisticians should become more indispensable for involvement in study design (including Animal Care and Use Committees), application of appropriate analyses, prudent interpretation of results, and the development of new statistical techniques to meet these needs.

KEY WORDS: Animal care committees; Animal welfare regulations; Experimental design; Small sample size.

There has been an increasing focus worldwide on how animals are used in research experiments. The public and the scientific community generally concur that animals should be used efficiently and treated humanely when they are required in experimentation. Emphasis, in terms of attitudes and policies, has been placed on reducing the number of animals used in experiments, substituting nonanimal models (including mathematical models) for animal experiments where possible, and adjusting experimental methods to decrease animal suffering (for example, in vitro tissue irritation tests to replace eye irritant studies on rabbits). These original recommendations by Russell and Burch (1959) have been accepted as the standard by many oversight bodies. Their philosophy regarding animal research is summarized as the three R's: Refinement (of experimental technique), Reduction (of number of animals used), Replacement (of animals with substitutes). These general principles have been incorporated into animal welfare regulations and progress is being made in many areas of testing and experimentation towards the well-being of research animals (see, for example, Rao and Huff 1990; Rowan 1990).

Passage of the Animal Welfare Act and the amendments revising it (U.S. Congress 1989) have led to substantial impacts on the uses of animals in experiments in the United States to ensure both wise use and minimum stress to research animals. Similar regulations have been instituted in other countries. These laws have a direct effect on the role of consulting statisticians worldwide. The general areas in which the consulting statistician has been most affected are discussed in this article; both U.S. and Australian regulations are used as examples of the directions that the animal welfare laws have taken. Similar regulations have been instituted in many European countries in recent years (O'Donoghue 1992).

1. ANIMAL CARE COMMITTEES

The implementation of institutional animal care committees is the primary method for assuring that the well-being of research animals has been considered before any study is initiated. Both the Public Health Service (Office for Protection from Research Risks 1986) and the U.S. Department of Agriculture (1989) require Institutional Animal Care and Use Committees (IACUC) to be appointed by the directors of research institutions. Similarly, Animal Experimentation and Ethics Committees (AEEC) in Australia have been formed and they are likewise charged with ensuring the humane use and care for animals. All proposed experiments involving animals must be reviewed and approved by IACUC's or AEEC's according to these laws. Committees are usually comprised of members with expertise in veterinary science or laboratory animal medicine, practicing scientists from several disciplines, and persons representing community concerns (National Health and Medical Research Council, Commonwealth Scientific and Industrial Research Organisation, and Australian Agricultural Council 1990; Orlans, Simmonds, and Dodds 1987). They are responsible for ensuring the appropriate use of animals in experiments, which includes the numbers of animals used, how they are used, and whether they will be used to produce valid results that meet the experimental objectives. For this reason, a statistician would be useful as a permanent member of an institution's committee, or at least all research protocols could be separately reviewed by statisticians to provide input into the research design. Protocol review is already a primary responsibility of many staff and consulting statisticians, and this added responsibility would be strengthened by interaction with these committees.

The Australian Medical Research Committee of the National Health and Medical Research Council (1989) provides a guide to AEEC's for minimizing the number of animals used in each research project. One section in this guide, entitled "Ensuring the Best Statistical and Predictive Techniques Are Used," considers the degree of statistical knowledge and experience of the investigator and whether or not the investigator has sought appropriate statistical advice. However, there is no specific requirement that statisticians serve directly on animal care committees or serve as consultants in this

*Richard M. Engeman is Leader of the Quantitative Sciences Project and Stephen A. Shumake is Leader of the Sensory Characteristics Project and past Chairman of the Animal Care and Use Committee at the Denver Wildlife Research Center, USDA/APHIS/S&T, Building 16, Denver Federal Center, P.O. Box 25266, Denver, CO 80225-0266. Invaluable input for this article was provided by John Wilson of the Centre for Biological Population Management at the Queensland University of Technology, Lee Allen of the Rural Lands Protection Board of Queensland, and Allen McKinnon of the Queensland Department of Housing and Local Government.

In the Public Domain

capacity. This is probably another area where statisticians need to be assertive and actively involved so that as attempts are made to refine or reduce animal use, particularly in stressful experiments, the quality of proposed experimental designs can be maintained or enhanced accordingly. Just as it is essential for a consulting veterinarian on these committees to understand the nature of the problem being addressed in a study (Pakes 1990), it is essential that the consulting statistician gains this understanding also. By serving on these committees, statisticians have another opportunity to contribute directly to the quality of the research products.

2. DISALLOWED EXPERIMENTS

Certain experimental procedures may not be permitted legally or made almost impossible due to required bureaucratic approval processes. As an example, both LD_{50} evaluations and Draize tests are currently banned in Queensland, Australia, unless Ministerial approval is obtained (Blackshaw and Allan 1985; Government Gazette 1991). In such cases, the statistician must work with the investigator to determine what alternative experimental designs can best address the same or a similar question of interest. Frequently, another, more focused question may be of primary interest.

In the United States, approved registration of vertebrate pesticides frequently requires an estimate of the LD_{50}, even though confidence intervals or regression line slopes may not be required by the regulating agency (i.e., U.S. Environmental Protection Agency). But, instead of estimating the LD_{50}, the practical question of interest often posed is, "How effective is a particular dose?" The testing of only one or two doses obviously would require fewer animals than the total required to produce a valid estimate of the LD_{50}. Although these LD_{50} assessments are often required and permitted in the United States, the questions of interest could be refined in this manner, if and when restrictions similar to those in effect in Australia also are imposed in the United States.

3. REFINEMENTS REQUIRED FOR SMALL SAMPLE EXPERIMENTS

One of the effects that the animal welfare regulations could have on consulting statisticians would be the statistical constraints generated by reducing the number of animals used in experiments. In some cases, at marginally financed institutions, the number of animals used would be reduced even further for economic reasons related to routine animal care. Strict regulations concerning the care and use of animals, the quality of holding facilities, the number and training credentials of handlers, and record-keeping requirements, additionally could increase the direct costs of doing animal experiments.

Mann, Crouse, and Prentice (1991) provided a good discussion for researchers on sample size considerations in light of animal welfare considerations. These included reducing experimental variation, repeated sampling techniques, efficient use of control groups, refining baseline measurements, and in some instances, replacement of animals with in vitro tissue effects or computer simulations. To assist in these efforts, statisticians will be increasingly involved with designing, analyzing, and interpreting experiments with small sample sizes or will be asked to increase power by combining experimental groups to increase sample size, in association with an appropriately broadened hypothesis. They very well could be involved in the development of alternative animal tests where fewer animals are required. For example, such progress is one of the provisions of Directive 86/609/EEC adopted by the Council of European Communities, whereby member states are to encourage research into the production and validation of tests that yield the same level of information, but use fewer animals or cause less pain (O'Donoghue 1992).

Many data analysis procedures are based on asymptotic results. Their properties may not be well defined or they may not be reliable with small sample cases [Engeman, Otis, and Dusenberry (1986) have demonstrated this for bioassay experiments]. Even data structures as simple as 2×2 contingency tables continue to be a source of controversy in terms of which type of data analysis would be most appropriate in small sample situations (see, for example, D'Agostino, Chase, and Belanger 1988; Upton 1982). Here, and in other analytical areas, some tests are more conservative than others, and the statistician must decide which is most appropriate for the specific questions addressed in each experiment. The relative lack of theoretical guidance in the small sample size situation has added to the difficulty the statistician must face in recommending an analytical procedure.

As a general rule, experiments with small sample sizes have greater potential to lead to Type II inference errors. When the "magic" p value of .05 is not obtained, investigators frequently tend to conclude (outright or de facto in their discussions) that there was no effect and, therefore, they accept the null hypothesis. As fewer animals are used in experiments, Type II errors may become more common. The consulting statistician will need to place greater emphasis on explaining the power of tests with a given number of animals per treatment condition, and on helping the investigator describe "nonsignificant" results appropriately. An indication of the extent of this problem is given by Freiman, Chalmers, Smith, and Kuebler (1986), where they concluded that the results of 50 of the 71 human clinical trials they examined could not have detected a 50% improvement in treatment. Nevertheless, the results commonly were presented as though the trials statistically showed no meaningful improvement from the experimental treatments.

One could reasonably assume that this could become an increasingly widespread problem among experiments on animals as sample sizes are forced to decrease by animal welfare legislation. However, these considerations also indicate that there are statistical research

challenges for the development of new sensitive and reliable small sample analytical methods. In addition, the properties and limitations of exisitng analytical methods at small sample sizes need more refined clarification and definition.

4. SOURCES OF BIAS AND VARIABILITY

The health and welfare of the animals during experimentation has become the major concern in these regulations. The Animal Welfare Act of 1989 also had an indirect effect on improving the quality of data from experimental animals because indicators of health and stress had to be monitored at least daily. However, whenever distress- and pain-reducing methods are to be applied in a given study, the statistician should consider the effects of these methods on the outcome of experiments. For example, to what degree might analgesic medications or anesthetics confound or restrict the physiological or behavioral response of an animal, and when do these effects dissipate. In such cases, the statistician's role may be to assure that inferences from an experiment are restricted to the set of conditions under which the experiment was conducted. Thus the generality of the findings might be limited. These concerns make it even more essential that statisticians directly observe and understand the detailed procedures and data collection methods involved in the conduct of laboratory or field experiments.

Similarly, in cases where sample sizes are further restricted, the potential influence from an uncontrolled source of variability becomes greater. Methods such as reducing variability in sources of lab animals, controlling for sex, conducting the experiment to avoid the effects of breeding cycles, assigning animals to treatment groups randomized from weight classes, and controlling observer variability with the experimental design are all valuable for reducing variability and should not be overlooked by the investigator as a means of improving sensitivity of experimental procedures. Pilot studies can also be conducted to help identify some of the critical parameters that should be controlled or manipulated to optimize the effects before the formal experiment is initiated. Such use of pilot studies can yield a net savings of animals in the long run (Seidel 1990).

5. FOCUS OF EXPERIMENTS

As the concern for the well-being of experimental animals increases and, in some cases, their availability decreases, the experimenter needs to be even more focused towards the objectives of the work. Less complex statements of the research question(s) or hypotheses of interest often permit fewer animals to be used per experiment. In this regard, it has always been the consulting statistician's role to help the investigator define, simplify, and narrow his or her questions to improve experimental focus. In some cases, the statistician may help the investigator reduce the scope of an experiment in order to concentrate on narrower inferences that will be addressed with higher levels of confidence. For example, the number of treatments could be reduced and the experimental material distributed among fewer treatments, rather than having an insensitive comparison among a larger number of treatments. On the other hand, the statistician and/or the IACUC may recommend the use of more animals on the grounds that, in the long run, one major experiment resulting in a powerful test will ultimately lead to fewer total animals being used. Seidel (1990) argued that experiments where negative results are obtained because an insufficient number of animals were tested per experimental group can be a complete waste of animals. Such studies, upon obtaining "negative" treatment effects, also tend to discourage others from researching potentially useful areas of study.

6. MODELING AND SIMULATION

Computer models and simulations also often are suggested as alternatives to using animals in experiments. General modeling packages for populations or physiological systems are available. They frequently can serve as valuable tools for examining how biological systems function, and as an aid for suggesting the next connecting step in a series of experiments or hypotheses to be tested. The statistician may be asked to help build a mathematical model or to conduct a simulation of a biological system. Although computer models of incredible complexity can be developed, their utility is still based on how completely the biological systems under study currently are understood. The statistician must examine the derivation of the biological data used in the model; whether the model form is purely descriptive or implies a functional process; how sensitive the model is to changes in the parameter values; and under what conditions the data originally were gathered to generate the form and the parameterization of the model. The statistician should actively participate in model building and/or usage and also should insist on a validation process using actual biological data. Even with a proven operational model, continued cross-validation with animals is necessary, especially on those occasions when attempts are made to apply it to new situations or a closely related species or strain. The statistician needs to play an important role in suggesting and designing these validation tests, as well as in developing models and in designing modeling experiments.

The use of in vitro experiments to replace in vivo experiments also is emphasized as a means to reduce the number of animals used in experiments (see, for example, Weiss 1988). These studies can be considered as biological modeling or simulation experiments. Significant advancements have been made in developing in vitro tests to replace animals in experiments (Gad 1990); however, the statistician still is faced with similar challenges as for mathematical modeling and simulation. That is, how well do cells and tissues model what happens in a complete organism and how far can inferences from an in vitro experiment be applied towards a complete organism. The statistician, in designing in

vitro experiments, probably has the luxury of having greater resources in experimental material than if an animal experiment were to be conducted. However, presentation of the inferences when the underlying objectives may be towards the effects on an organism can prove challenging.

The appropriate application of inferences from in vitro studies delineates another area where the statistician may provide vital input. This could take the form of suggesting and designing experiments that relate results from in vitro studies to results from the in vivo studies they are attempting to replace, or comparing the results from two competing in vitro tests that, in reality, measure slightly different variables. For example, Wilsnack, Meyer, and Smith (1973) compared the results from using human WI-38 cells to the results from animal tests as indicators of toxicity, with outcomes indicating the potential for replacing some animal tests with in vitro tests. It is reasonable to presume that the degree of confidence that one could have for extrapolating from in vitro experiments to the in vivo situation would relate to how well a relationship had been defined in comparative studies.

7. MULTIDISCIPLINARY CHALLENGES

Contemporary scientific projects often require the collaboration of investigators from several disciplines working as a team. The physicists and chemists involved may use measures that essentially approach ratio scales with true absolute zero points (e.g., degrees Kelvin). Biologists may be working with interval scales in many of their measurements on animals (for example, centimeters of body length, grams of food consumed). Ethologists and histopathologists may be observing and taking frequency count data on qualitative variables (for example, incidence of animal aggression, abnormally shaped cell types). To conserve time and funding levels, as well as animals, it frequently becomes necessary to obtain many measures on the same groups of animals before, during, and after they have been exposed to a set of treatment conditions. In these cases, statisticians, in reviewing and providing input to project proposals, must exercise their own judgments as to the lowest common denominator for needed sample sizes, double-blind procedures that best remove or reduce bias, and other procedural or data analysis decisions that can strongly affect the quality and size of the research project.

In some studies that require very large sample sizes, attempts have been made to retroactively use data from many previously published and unpublished sources with meta-analysis techniques (Mann 1990). Although these techniques have the potential for reducing the number of actual studies performed, reducing research costs, and increasing sample sizes, their indiscriminant use could be fraught with many pitfalls (see, for example, Mosteller and Chalmers 1992). The use of published studies for meta-analysis leads to the potential for bias from the selective publication of only a subset of the experiments conducted in a particular area (Dear and Begg 1992; Hedges 1992). In addition, many assumptions must be made regarding some critical procedural details that can frequently affect the outcomes of individual independent studies. Those studies that involve the most control of independent variables and can achieve the most reliable measurement techniques on the dependent variables would, of course, be the best candidates for pooling for a meta-analysis of data from different sources. Although meta-analyses are being used with increasing frequency in clinical epidemiological studies on effects of different treatments for human diseases, caution is needed when it is applied to emerging animal research questions, and as with modeling questions, empirical studies are needed to validate this method of drawing inferences and to further advance meta-analytic statistical methods. However, one example of where the analyses of results from many studies seems to have provided information useful for reducing the number of animals used in testing is a study by Talsma et al. (1988) that indicated the ability of irritation scores from two-, three-, four-, and five-rabbit subsets to predict the outcome from a six-rabbit Draize test.

In attempts to standardize test methodology and to achieve high levels of control for quality assurance requirements, some major research prorams (for example, Buelke-Sam et al. 1985) have been conducted to compare and cross-validate experimental results obtained in different laboratories using essentially the same agreed-upon written standardized data collection and analysis procedures. The comparison standard methods (for example, startle response, body weight, animal activity, negative geotaxis) for evaluating the pharmacological effects of drugs in albino rats was an example of a large multilaboratory research endeavor that was conducted to assess sensitivity of the measures as well as intra- and interlaboratory reliability (Nelson, Felton, Kimmel, Buelke-Sam, and Adams (1985).

With this program of research, using computerized control of stimulus/response parameters, data handling, and data analysis routines, there was a fairly high degree of agreement on the effects of two drugs among six research laboratories. Although some laboratories consistently generated results that showed more (or less) of the drug effects than others, the effects for all laboratories were of the same form and in the same direction. In other words, no drug-by-laboratory interaction was detected.

Such large-scale research endeavors are expensive, and they require the use of many animals in complex replicated designs (Nelson et al. 1985). However, to achieve the necessary degree of confidence to develop advanced models or meta-analytical approaches, they may be required. Ultimately, these large-scale multidisciplinary, multilaboratory designs could reduce future usage of animals, as scientists and statisticians become more confident in their measurement and analytical techniques. When measurement methods need refinement, per se, to scientifically examine certain phenom-

ena in more detail, large-scale replicated experiments within and between laboratories are essential.

8. CONCLUSION

Animal welfare considerations and regulations can place more pressure on the investigator to get scientifically sound inferences when using fewer animals, especially in those studies that involve stressful or pain-producing procedures. This situation will increase the demand for creative input from the statistician. Consulting statisticians are practiced at working within given constraints to produce valid experimental designs and analyses that address the investigator's questions. Animal welfare considerations have had a constraining influence on animal research, and the statistician undoubtedly will become more indispensable for developing strong, valid protocol designs for applying appropriate statistical analyses, for prudent interpretation of the results, and for development of new techniques.

[*Received October 1991. Revised December 1992.*]

REFERENCES

Blackshaw, J. K., and Allan, D. J. (1985), *Principles of Laboratory Animal Management*, Queensland, Australia: Australasian Society for the Study of Animal Behaviour, University of Queensland.

Buelke-Sam, S. J., Kimmel, C. A., Adams, J., Nelson, C. J., Vorhees, C. V., Wright, D. C., St. Omer, V., Korol, B., Butcher, R. E., Geyer, M. A., Holson, J. F., Kutscher, C. L., and Wayer, M. J. (1985), "Collective Behavioral Teratology Study: Results," *Neurobehavioral Toxicology and Teratology*, 7, 591–624.

D'Agostino, R. B., Chase, E., and Belanger, A. (1988), "The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Proportions," *The American Statistician*, 42, 198–202.

Dear, K. B. G., and Begg, C. B. (1992), "An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis," *Statistical Science*, 7, 237–245.

Engeman, R. M., Otis, D. L., and Dusenberry, W. E. (1986), "Small Sample Comparison of Thompson's Estimator to Some Common Bioassay Estimators," *Journal of Statistical Computation and Simulation*, 25, 237–250.

Freiman, J. A., Chalmers, T. C., Smith, H., and Kuebler, R. R. (1986), "The Importance of Beta, the Type II Error, and Sample Size in the Design and Interpretation of the Randomized Controlled Trial. Survey of 71 "Negative" Trials. in *Medical Uses of Statistics*, eds. J. C. Bailar III and F. Mosteller, Waltham, MA: New England Journal of Medicine Books, pp. 289–304.

Gad, S. C. (1990), "Recent Developments in Replacing, Reducing, and Refining Animal Use in Toxicologic Research and Testing," *Fundamental and Applied Toxicology*, 15, 8–16.

Government Gazette, Queensland State Government (1991), "Animals Protection (Use of Animals for Scientific Experiments)" (Regulations 1991, No. 128), 2679–2680.

Hedges, L. V. (1992), "Modeling Publication Selection Effects in Meta-Analysis," *Statistical Science*, 7, 246–255.

Mann, C. (1990), "Meta-Analysis in the Breech," *Science*, 249, 476–480.

Mann, M. D., Crouse, D. A., and Prentice, E. D. (1991), "Appropriate Animal Numbers in Biomedical Research in Light of Animal Welfare Considerations," *Laboratory Animal Science*, 41, 6–14.

Mosteller, F., and Chalmers, T. C. (1992), "Some Progress and Problems in Meta-Analysis," *Statistical Science*, 7, 227–236.

National Health and Medical Research Council, Medical Research Committee (1989), *Animal Experimentation Ethics Committee Newsletter*, No. 3, October, 1989.

National Health and Medical Research Council, Commonwealth Scientific and Industrial Research Organisation, and Australian Agricultural Council (1990), *Australian Code of Practice for the Care and Use of Animals for Scientific Purposes*, Canberra: Australian Government Publishing Service.

Nelson, C. J., Felton, R. P., Kimmel, C. A., Buelke-Sam, J., and Adams, J. (1985), "Collaborative Behavioral Teratology Study: Statistical Approach," *Neurobehavioral Toxicology and Teratology*, 7, 587–590.

O'Donoghue, P. N. (1992), "European Regulation of Animal Experiments," *Lab Animal*, 21, 20–27.

Office for Protection from Research Risks (1986), *Public Health Service Policy on Human Use of Laboratory Animals*, Bethesda, MD: National Institutes of Health.

Orlans, F. B., Simmonds, R. C., and Dodds, W. J. (eds.) (1987), "Effective Use of Animal Care and Use Committees," *Laboratory Animal Science*, Special Issue, 11–13.

Pakes, S. P. (1990), "Contributions of the Laboratory Animal Veterinarian to Refining Animal Experiments in Toxicology," *Fundamental and Applied Toxicology*, 15, 17–24.

Rao, G. N., and Huff, J. (1990), "Refinement of Long-Term Carcinogenesis Studies," *Fundamental and Applied Toxicology*, 15, 33–43.

Rowan, A. N. (1990), "Refinement of Animal Research Technique and Validity of Research Data," *Fundamental and Applied Toxicology*, 15, 25–32.

Russell, W. M. S., and Burch, R. L. (1959), *The Principles of Humane Experimental Technique*, Springfield, IL: Charles C Thomas.

Seidel Jr., G. E. (1990), "Basic Principles of Experimental Design," in *The Experimental Animal in Biomedical Research, Vol. I. A Survey of Scientific and Ethical Issues for Investigators*, eds. B. E. Rollin and M. Lynne, Boca Raton, FL: CRC Press.

Talsma, D. M., Leach, C. L., Hatoum, N. S., Gibbons, R. D., Roger, J.-C., and Garvin, P. J. (1988), "Reducing the Number of Rabbits in the Draize Eye Irritancy Test: A Statistical Analysis of 155 Studies Conducted Over 6 Years," *Fundamental and Applied Toxicology*, 10, 146–153.

Upton, G. J. G. (1982), "A Comparison of Alternative Tests for the 2 × 2 Comparative Trial," *Journal of the Royal Statistical Society*, Ser. A, 145, 86–105.

U.S. Congress (1989), "The Animal Welfare Act. 1966 PL 89–544," U.S. Code 2131-2156, Vol. 7.

U. S. Department of Agriculture, Animal and Plant Health Inspection Service (1989), "9 CFR Parts 1, 2, and 3 Animal Welfare Final Rules," *Federal Register, Part IV*, 54, 36151–36153.

Weiss, R. (1988), "Test Tube Toxicology," *Science News*, 133, 42–45.

Wilsnack, R. E., Meyer, F. J., and Smith, J. G. (1973), "Human Cell Culture Toxicity Testing of Medical Devices and Correlation to Animal Tests," *Biomaterials, Medical Devices, Artificial Organs*, 1, 543–562.