The Logical Determination of "N" in Animal Experimentation

by David A. Crouse, Ph.D., Michael D. Mann, Ph.D. and Ernest D. Prentice, Ph.D.

The public sentiments to reduce the number of animals being used in research and the philosophies that underlie them are not modern phenomena, nor are the pleadings of respectable scientists and mathematicians who urge for adoption of a more responsible approach to experimental design. Indeed, the landmark work of Fischer (1935) soundly criticized the squandering of resources through poor design and inattention to the most fundamental principles of statistics. Fischer very bluntly stated that "*The waste of scientific resources in futile experimentation has, in the past, been immense in many fields.*" Over the years, this concern has been echoed by many others. Most scientists would agree that a major waste of animals in research occurs when results from experiments do not contribute to scientific knowledge because of poor design or lack of proper data evaluation. It may be equally wasteful to use animals or other precious resources in experiments that are never publicly reported or shared, even if they were well-designed and statistically valid.

In the late 50's, Russell & Burch (1959) presented their landmark paper which proposed some fundamental, responsible approaches to the proper utilization of animals in research protocols. Every investigator who uses animals in research should be familiar with the basic principles of **R**efinement, **R**eduction and **R**eplacement. As a matter of regulatory compliance (The Animal Welfare Act, 1985; PHS Policy on the Humane Use of Laboratory Animals, 1986; USDA APHIS Final Rules, 1989), IACUCs must consider these three R's in their reviews of animal research proposals.

Throughout Europe, the United Kingdom, Canada and the US, there has been a reduction in animal usage in research over the past several years (Festing, 1994; Orlans, 1994). On our own campus, there also has been a noticeable decline in the overall number of animal research protocols submitted for review. In the United States, the quality of the reporting data used to demonstrate the reduction in animal usage for research has been considered 'excellent' for less than the past 10 years, and the data do not extend to all species. Prior to that time, lack of reporting, poor follow-up, incomplete categorization and other variables led to 'poor' quality of the quantitative data on animal usage. Indeed, only since 1990 have reasonable numbers been obtained for some species now regulated by the USDA (mostly farm animals). Unfortunately, accurate and complete data on other unregulated species (rats, mice, birds) remains elusive, although recent efforts have had an impact on accounting. What is clear, however, is that the latter species probably make up over 80% of the total individual animals used (Orlans, 1994). New data collection procedures and requirements should give a much more complete and accurate representation of animal usage in the US. A survey of selected USDA data for animal usage since 1973 is presented in Figure 1 (see page 20).

In this brief paper, we review some basic elements in the research environment which can have a significant impact on animal usage, and, like that provided by other recent papers (Festing, 1992), we provide an overview of important statistical parameters which should be considered in the design of typical animal research protocols and their subsequent review

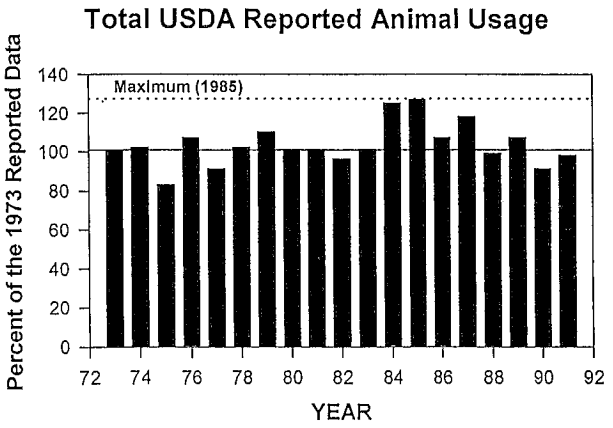

Figure 1. The use of animals in research protocols as reported to the USDA. The data are expressed as a percentage of the 1973 values. Data for recent years have had the number of reported fram animals used subtracted to make them comparable to earlier years. All data since 1985 have a greater level of accuracy (see text) (modified from Orlans, 1994).

by the IACUC. Significant contributions to the decline in use of animals in research protocols likely has come from many sources yet some change is a direct result of refinements in experimental protocols that lead to use of fewer animals.

As a point of general background, one should understand that the number of animals used in an experiment or set of related experiments is not a direct indicator of the impact of the reported research on scientific progress. A couple of examples may help to illustrate this fact. In the late 50's a series of "megamouse" experiments by Russell and colleagues at the Oak Ridge National Laboratory were the first such studies that showed that the genetic effects of external radiation exposure in mammals (mice) were qualitatively comparable to those observed in Drosophila and single-celled organisms. In two key papers, which are now fundamental references describing genotoxic effects of radiation in mammals, 1,342,597 mice were used and results were detailed in only three brief tables (Russell *et al.*, 1958; Russell 1962). The large number of animals was a result of an experimental design capable of detecting relatively small increases in phenotypically expressed mutation frequencies which, even in controls, were rare events. These essential data have had an enormous impact and now form part of the basis for many of the radiation protection guidelines related to genetically significant doses of external radiation exposure (National Research Council, 1990).

In a more recent paper, detailing molecular studies conducted on individual experimental mice to characterize their genetic composition, results from only 21 mice (14 were controls) are specifically tabulated although it is clear that additional animals were also used (Palmiter, 1982). Only descriptive statistical analyses are presented, yet the impact of this project is clearly significant and wide-ranging. The data in this seminal paper were the first to demonstrate the successful introduction of foreign genes into mammalian embryos (e.g., mice transgenic for the human growth hormone gene).

It is unquestionable that there are many rational ways to reduce the number of animals used in research and many have nothing to do with the science of logic or with statistics. A thorough review of the literature is one of the easiest and most resource-conserving methods to assure that projects are not unnecessarily duplicative. It also facilitates the selection of "doses" for exogenous agents administered to research animals and very often, gives the best insights into experimental designs or approaches that are both efficient and effective. Such a review also appears to be mandated by the USDA (USDA APHIS Final Rules, 1989). Alternatively, in the review of previous literature and the data from his or her own laboratory, the researcher often finds that it is possible to refine the number of control animals for a given set of projects. In some instances, historical controls are not adequate, whereas in others, they can be quite suitable. When the endpoints of experiments are quantitatively reproducible, it may be appropriate to pool certain control values over time and thus achieve a significantly increased sample size in the control group. On occasion it may be necessary to statistically test controls against each other to validate such pooling of data. In any case, such pooling often allows significant reduction in animal usage and, as a practical point,

Table 1. Relationships between the null hypothesis and outcome.

	The Null Hypothesis is ACTUALLY TRUE	The Null Hypothesis is ACTUALLY FALSE
ACCEPT the Null Hypothesis	CORRECT Conclusion	TYPE II ERROR (β) "False Negative"
REJECT the Null Hypothesis	TYPE I ERROR (α) "False Positive"	INCORRECT Conclusion

makes more efficient use of technician time and other resources.

In a very similar fashion, one should consider the need for multiple levels of control groups. Some complex designs have controls for each of several different treatments or times of treatment. Careful analysis of previous data coupled with personal experience often allows the elimination or reduction of unnecessary or excessively large control groups (Mann *et al.*, 1991). Some experimental protocols may allow repeated sampling of animals without jeopardizing animal welfare. Such an approach should carefully consider all of the local and federal guidelines to assure that an appropriate balance between ethical cost and scientific benefit is present. Of course, protocols should not involve multiple major surgeries or other potentially distressing or painful procedures. However, with many other procedures (blood samples, minor biopsies, drug administration, etc.), this approach can greatly reduce the number of animals, allowing each experimental subject to serve as it's own control and the likely application of "paired" statistical approaches to increase the power of the test(s).

Additional refinement of protocols through the use of sequential testing or appropriate replication can also lead to a reduction in the required number of experimental subjects. In sequential testing (Mann *et al.*, 1991; Roberts, 1991), the progress through blocks of the experimental design is dictated by the data from the most recently collected data set. For example, although five doses of an agent may be planned for a study, an initial challenge at a carefully selected mid-level dose is used to predict which set of subsequent doses will or will not be employed. In a very similar way, cumulative data can be collected using paired experimental subjects and an "untied pairs" statistical method. Both such designs can become quite complex and are described in much more detail in other more comprehensive literature (Mann *et al.*, 1991; Daly, *et al.*, 1991).

Once an overall experimental approach has been chosen, it is important to define the detailed design with some additional consideration of fundamental statistical principles. These include: an estimation of the kinds and magnitudes of error anticipated or set as thresholds (acceptable errors, criterion "p" values); the amount of difference between control and experimental groups (effect size) that will indicate a significant biological effect; and the determination of sample size. The appropriate sample size depends upon the former considerations. Each of these parameters will be discussed briefly.

Among the most fundamental parameters which influence sample size and thus the numbers of animals which enter research protocols are Type I versus Type II error. The basic interpretation of α- and β-errors is presented in Table 1. In the simplest sense, the Type I error or α-error can be considered the probability of having a "false positive" result, i.e., concluding that there is a treatment-related effect, when indeed there is none. One must always be aware that values for α seldom reach zero in biological experimentation; ranges of sample values from normally distributed control and experimental groups usually have some degree of overlap. Typically, researchers will accept a relatively low probability for Type I error, and it is called the "p value" or "level of significance." Most investigators use a p value of 0.05, which means that in approximately 1 of 20 comparative sets of similar data, they may actually accept a false hypothesis. On the other hand, 0.05 is an arbitrary value; there may be no real biological difference between results that have a p value of 0.06 and those with a p value of 0.05. For that reason, investigators should report the p value which results from a statistical evaluation of the data as a method for the

reader to formulate a more complete evaluation of the findings. Nearly all researchers realize that the smaller the value selected for α as an acceptance criteria for "significant difference," the harder it is to detect a real effect of the treatment; in other words, the "power" of the study is reduced. Clearly, if more subjects are entered into the control and treatment groups, it is easier to detect a significant effect for any value of α.

The second general category of errors is the Type II error or the β-error which represents the probability of having a "false negative," i.e., concluding that there is *not* a treatment-related effect when, indeed, there was an effect. The quantity $1-\beta$ is called the power of the test, which one normally desires to have as large as possible. The smaller the value of α is, the larger the value of $1-\beta$. A more stringent α criterion reduces the power of the study; a less stringent value increases it. The power of the test also is dependent upon sample size.

Because of these interrelationships, if the investigator can: (i) set the size of the difference between the control and the treatment group that is considered to be "biologically significant;" (ii) choose values for α (often 0.05) and β; and (iii) estimate the standard deviation of the test population (often based on experience), it is possible to accurately determine the minimum sample size that would allow the detection of a significant difference between the two populations. The actual computation of such estimates depends upon the statistical tests to be employed. There are a variety of tabular presentations or software packages which facilitate the exercise (Mann *et al.*, 1991; Daly, *et al.*, 1991). It is clear that a power analysis can result in a reduction in the number of animals used while optimizing the opportunity to statistically support or refute hypotheses. Many scientific journals now require evidence of such analysis to be included in manuscripts before they are accepted for publication.

Increased sensitivity of researchers to the need for power analysis, in part forced by IACUCs and scientific journals, has contributed to the decline in numbers of animal used for research but the decline also has been caused by other non-statistical elements. New regulatory guidelines have had an impact: careful considerations of proposed usage, explained and justified to a peer review group in detailed documentation and the concomitant paperwork needed to support the process, certainly have contributed to a reduction in animals used for research and in some cases to increases in projects not employing live animals. Having an equally significant impact has been the relative cost of animal-based research which has stayed well in front of inflation for most investigators. The increases in these basic costs (care and purchase) show few signs of abating, especially when one considers that cost per item (e.g., animal) and the total number produced by the supplier should be inversely related. The relatively fixed (or even declining) level of total research funding by federal agencies coupled with the increased cost of individual research projects has meant that conducting experiments with (or even without) animals approaches a no-win fiscal situation in too many cases. We and many of our colleagues in similar laboratories have gradually reduced our use of animals for primarily fiscal reasons. In some cases, investigators have chosen alternate experimental approaches, focusing on *in vitro* and molecular approaches which often utilize fewer animals or approaches that include a more direct move to the clinical setting with human research subjects/materials.

As a closing comment, in spite of the credibility that statistics can bring to efficient use of animals in research protocols, there is also no question that inappropriate application of these tools can lead to faulty conclusions or can mask important information. In this regard, researchers must be aware of the problem of using too few experimental subjects. Some studies, for example, have suggested that more than 50% of clinical trials concluding "no effect of the treatment," could not have detected an effect if it existed because they used too few subjects. Few estimates of this nature exist for animal research (Festing, 1992), but clearly, similar experimental designs and reporting would be wasteful of animals. From a lighter perspective, it also has been proposed that there are considerable similarities between the stereotypical views of statistics and politicians.

Some individuals see both as: admittedly and often grudgingly essential for what we do; frequently very slick, colorful and complicated; usually set-up to tell us what we already know and believe; rarely willing to tell us things that we don't want to hear; and always capable of hiding or obfuscating the most interesting and essential information.

With these caveats in mind it is possible, even commendable, for investigators to accept a joint responsibility with their IACUC to arrive at the most appropriate numbers of animals entering into research protocols. Experience in our own IACUC has shown that this cooperative process has led researchers to realize that it may be possible to use fewer animals in some protocols and still obtain equal or better data. The emphasis must always be on using the appropriate number of animals for each experiment.

References

Daly, L.E., G.J. Bourke and J McGilvray. 1991. *Interpretation and Uses of Medical Statistics, 4th Ed.* London: Blackwell Scientific.

Festing, M.F.W. 1992. The scope for improving the design of laboratory animal experiments. *Laboratory Animals* 26:256-267.

Festing, M.F.W. 1994. Reduction of animal use: Experimental design and quality of experiments. *Laboratory Animals* 28:212-221.

Fischer, R.A.F. 1935. *The Design of Experiments*. New York: Hafner Press.

Mann, M.D., D.A. Crouse and E.D. Prentice. 1991. Appropriate animal numbers in biomedical research in light of animal welfare considerations. *Laboratory Animal Science* 41:6-14.

National Research Council. 1990. *Health Effects of Exposure to Low Levels of Ionizing Radiation, The BIER V Report.* Washington, DC: National Academy Press.

OPRR (Office for the Protection from Research Risks). 1986. *Public Health Service Policy on Humane Use of Laboratory Animals*. Bethesda, MD: NIH.

Orlans, B.F. 1994. Data on animal experimentation in the United States: What they do and do not show. *Perspectives in Biology and Medicine* 37:217-231.

Palmiter, R.D., R.L. Brinster, R.E. Hammer, *et al.* 1982. Dramatic growth of mice that develop from eggs microinjected with metallothionein-growth hormone fusion genes. *Nature* 300:611-615.

Prentice, E.D., D.A. Crouse and M.D. Mann. 1992. Scientific merit review: The role of the IACUC. *ILAR News* 34:15-19.

Roberts, C.D. 1991. Statistical model in tests for eye irritants. *Food and Chemical Toxicity* 29:463-468.

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

Russell, W.L., L. Brauch and E.M. Kelly. 1958. Radiation dose rate and mutation frequency. *Science* 128:1546-1550.

Russell, W.L. 1962. An augmenting effect of dose fractionation on radiation-induced mutation rate in mice. *Proceedings of the National Academy of Science (USA)* 48:1724-1727.

U.S. Congress. *The Animal Welfare Act.* 1966. PL 89-544, Vol. 7, U.S. Code 2131-2156. Amended, December 17, 1985.

USDA, Animal and Plant Health Inspection Service (APHIS). 1989. Part IV, 9 CFR Parts 1, 2 and 3 Animal Welfare; Final Rules. *Federal Register* 54(168):36153.