

Good Experimental Design and Statistics Can Save Animals, But How Can it Be Promoted?

Michael F.W. Festing

c/o MRC Toxicology Unit, University of Leicester, Leicester LE1 9HN, UK

E-mail: *mfwf1@le.ac.uk*

Summary — Surveys of published papers show that there are many errors both in the design of the experiments and in the statistical analysis of the resulting data. This must result in a waste of animals and scientific resources, and it is surely unethical. Scientific quality might be improved, to some extent, by journal editors, but they are constrained by lack of statistical referees and inadequate statistical training of those referees that they do use. Other parties, such as welfare regulators, ethical review committees and individual scientists also have an interest in scientific quality, but they do not seem to be well placed to make the required changes. However, those who fund research would have the power to do something if they could be convinced that it is in their best interests to do so. More examples of the way in which better experimental design has led to improved experiments would be helpful in persuading these funding organisations to take further action.

Key words: *animal experiments, experimental design, peer review, reduction, statistics.*

Introduction

Poor experimental design and the use of inappropriate statistical methods in biomedical research can result in misleading experiments which waste scientific resources and may cause unnecessary suffering of people and animals. For example, Altman (1) stated that, “In 1994 I observed that research papers in medical journals often exhibit the use of inappropriate design and analysis, incorrect use of appropriate techniques, incorrect interpretation of results, selective reporting of results, selective citation of the literature, and drawing unjustified conclusions. I commented that, “This is surely a scandal””. While Altman was mainly discussing clinical research, similar comments could apply equally well to animal experiments. For example, a recent meta-analysis of 44 papers on fluid resuscitation in animals found that only two said how animals had been allocated; none had sufficient power to detect reliably a halving in risk of death; there was substantial scope for bias; and there was substantial heterogeneity in the results, due to the method of inducing the bleeding, so that the odds ratios were impossible to interpret. The authors questioned whether these experiments had any relevance to human medicine (2).

Another Small Survey

An unpublished survey, conducted by the author, of a sample of 27 papers picked at random from a list of papers that used laboratory animals, published in 19 different journals in 1998, highlights some of the main problems. First, there were errors of omission.

Many papers failed to state whether the animals had been assigned to treatments at random. In cases where “blinding” would appear to have been appropriate (i.e. in cases where there was some subjective element in collecting the data), papers often failed to mention whether it had been used. None of the papers made any attempt to justify the sample sizes that they had used. It was not unusual for the papers to fail to record the age, sex and husbandry of the animals, and the exact way in which the experiment had been done. None of the papers mentioned the type of experimental design (e.g. completely randomised, randomised block, Latin square, etc.) and there was often no explanation for the choice of the animals (species and strain) that were used. In nearly a third of the papers, the exact number of animals used could not be determined.

About a third of the papers had statistical errors that were not serious enough to invalidate the conclusions, but a further third had errors that were more serious and might have invalidated the conclusions. For example, one paper used an unpaired t-test when the variances were clearly heterogeneous and the design was such that a paired t-test would have been appropriate. Thus, the p-values were unreliable. Another paper involved making several measurements on each experimental unit, and the results were mistakenly analysed as a factorial design, which is inappropriate. One paper scored the expression of a reporter transgene when injected into the eye on a subjective scale of 0, +, ++ and +++, but then incorrectly converted these to numerical values and estimated means and standard deviations. This is incorrect, because it assumes that the difference between 0 and + is the same as that between + and ++, etc., which is not

a valid assumption. Moreover, by expressing the data in this way, the number of animals in which the expression of the transgene was zero was obscured. The paper failed to do any statistical analyses using the numerical values and did not statistically compare a group of irradiated animals with un-irradiated ones.

Several papers apparently reported the results of more than one experiment, but often did not number the individual experiments, so it was difficult or impossible to discover which results came from which experiment. In one case, the authors did a formal statistical test to compare two means, but these were from different experiments, so any such comparison could have been seriously biased by environmental and technical factors acting differently between the two groups. The same paper used absence of a significant difference to claim that the treatment was having no effect. However, this could well have been because the experiment lacked sufficient power to detect an effect. Another paper made a somewhat similar error in that the aim was to compare the effects of two agonists, but for some reason, they had not been included in the same experiment, each had been compared with a control group.

Two papers using 40 rats and 28 primates, respectively, failed to generate any numerical data, so that the conclusions were based entirely on the author's subjective assessment of the histological findings. It seems unlikely that it was impossible to devise some method of scoring the results, even if subjectively. One paper concerned a $2 \times 2 \times 2 \times 2$ factorial design, but the authors had attempted to analyse it inappropriately, by using a one-way analysis of variance. They concluded that one of the factors was important in determining the outcome, when a re-analysis using more-appropriate techniques suggested that this was not the case.

Self-regulation, Power and Responsibility

Errors of these sorts must lead to wastage of animals, so they are ethically indefensible. How can they happen? Science is self-regulating in the sense that funding and the publication of research findings is subject to critical peer review by other scientists. However, it seems that this self-regulation can enter a downward spiral. Statistics is only a research tool and is not a fast moving, exciting, cutting edge science, at the level likely to impact on individual biomedical scientists. It has to compete for space in the scientific curriculum with subjects such as molecular biology and genetics, in which there has been an explosion of information recently. If it is not adequately taught, then the peer review system will gradually break down, because nobody will have the skill or training to question research proposals or papers submitted to journals.

Responsibility for quality in science depends on funding organisations, directors of institutes or companies, individual scientists, ethical review committees such as Institutional Animal Care and Use Committees (IACUCs), journal editors, regulatory bodies responsible for animal welfare and those, such as the Food and Drug Administration and the Environmental Protection Agency in the USA, and the European Medicines Commission, whose primary aim is to safeguard people. A distinction probably needs to be made between academic and commercial organisations, and the comments given below mainly refer to academic work.

Funding organisations, such as the National Institutes of Health in the USA and the Medical Research Council in the UK, usually submit research proposals for peer review, and this should assess the scientific validity and feasibility of the proposed project taking account of available resources and the track record of the applicant. They will usually also consider the ethical implications of the work, though this would rarely extend to detailed consideration of ways of reducing animal numbers, except where very large experiments are proposed. Once the work has been funded, there is usually no formal mechanism for assessing the scientific quality of the resulting work, apart from counting the number of papers published in prestigious journals. Thus, in a sense, quality control is delegated to the scientific journals. True, the work coming from some research institutes may be subject to periodic review, but apparently, the statistical quality of the work is rarely, if ever, assessed. However, the funding organisations probably have more power to change things than any of the other stakeholders. They could, for example, insist that, for some projects, the applicant should appoint a statistician. They could also audit the statistical quality of published papers produced by institutes or departments. More examples, both of bad design and of how better design has resulted in better science, are badly needed.

Directors of scientific institutes and professors and heads of departments in universities appoint staff, and should be able to ensure that the organisation has the right balance of bench scientists and support staff, such as statisticians. However, small departments may be unable to justify appointing a statistician, as there may be insufficient work for him or her. In some cases, the director may not even be aware of the need for statistical advice, and may be more interested in getting a job done than in getting it done efficiently. If twice as many animals are used than are necessary, but the paper gets published in *Nature*, most scientists will be entirely satisfied with the outcome.

Individual scientists clearly have an interest in doing their work efficiently and economically. If they can impress referees with the scientific quality of their work, there is more chance of getting it pub-

lished in a prestigious journal. However, they are often inadequately trained in experimental design and statistical methods, and in some cases, they have no access to statistical advice. As a result, they may be unaware that their work is not as good as it otherwise might be.

Ethical review committees, such as the IACUCs in the USA, do have a direct interest in ensuring that the minimum number of animals is used. However, the research proposals that they consider are often long-term projects involving many separate experiments, most of which cannot even be planned until the outcome of the first few experiments is known. Even if the IACUC has statistical advice, the statistician will only be able to help in general terms. What is really needed is one-to-one personal contact between the research scientist and a statistician who is familiar with the research subject matter. Ideally, the statistician should also have a biological training. This arrangement is rare in academia though it is more common in the larger commercial companies.

Editors will want to ensure that only high quality papers are published in their journals. If they have difficulty in getting statistical referees, then they should consider paying for it. Some large circulation journals may even employ a statistician.

How can Statistical Quality be Improved?

Training of scientists at undergraduate, graduate and mature research worker level needs to be improved. The main emphasis should be on research strategy and experimental design, rather than statistical methods, because if an experiment is well designed, the statistical analysis will normally present no problems (except in complex cases).

In view of the ethical implications of using too many animals, statistical advice should be available to all research workers who use animals. Ideally, this should be funded centrally, so that the service is free of charge to individual scientists. Scientists will not seek advice if it is going to cost them a lot of money. The possible place of the Web and electronic communication in offering advice remotely needs to be considered.

More statisticians need to be recruited into pre-clinical (i.e. animal) studies, and these statisticians will need to be offered a good career structure. They will also need a good grounding in biology and will need to be taught not to use mathematical jargon, as that is a sure way to alienate most biologists. Currently, only the funding organisations appear to have the power to begin to implement the necessary changes.

Conclusions

Better experimental design and statistical methods could lead to a reduction in animal use. The provision of teaching materials (3) and guidelines on best practice (4, 5) can help individual scientists. However, the peer review system, which is the main method of quality control in science, is currently unable to improve the situation much, due to a lack of statistically qualified referees. Organisations that fund scientific research have an interest in the quality of research and in ensuring that the minimum number of animals is used, and they also have the power to do something about it. There is an urgent need to make these organisations more aware of the likely benefits of improved experimental design and statistics.

References

1. Altman, D.G. (2002). Poor quality medical research: what can journals do? *Journal of the American Medical Association* **287**, 2765–2767.
2. 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. *British Medical Journal* **324**, 474–476.
3. Festing, M.F.W., Overend, P., Gaines Das, R., Cortina Borja, M. & Berdoy, M. (2002). *The Design of Animal Experiments*, 12pp. London, UK: Laboratory Animals Ltd.
4. Festing, M.F.W. & Altman, D.G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal* **43**, 233–243.
5. Festing, M.F.W. (2001). Guidelines for the design and statistical analysis of experiments in papers submitted to *ATLA*. *ATLA* **29**, 427–446.