Ecotoxicology 5, 169–186 (1996)

Improving the quality of statistics in regulatory ecotoxicity tests

PETER F. CHAPMAN[1], MARK CRANE[2], JOHN WILES[3], FRANK NOPPERT[4] and EDDIE McINDOE[1]

[1]*Zeneca Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire, RG42 6EY, UK*
[2]*Division of Biology, School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, TW20 0EX, UK*
[3]*Department of Ecotoxicology, Huntingdon Life Sciences, PO Box 2, Huntingdon, Cambridgeshire, PE18 6ES, UK*
[4]*Ministry of Transport and Waterworks, PO Box 9070, 6800 ED Arnhem, The Netherlands*

Received: 30 October 1995; accepted 10 January 1996

The results of an international workshop on the use of statistics in regulatory ecotoxicology are presented. There are currently many errors of omission in the recommendations on statistical analysis given in test guidelines. These are identified and advice is given on how to incorporate best statistical practice. The use of the no observed effect concentration (NOEC) as a summary statistic is questioned, and an alternative is suggested. Several areas of research that would resolve uncertainty in the design and analysis of ecotoxicity tests are also identified.

Keywords: Ecotoxicity; test guidelines; statistical analysis; experimental design; NOEC.

Introduction

It is widely recognized by those working in the regulatory arena that there is considerable scope for improving the statistical content of regulatory guidelines for ecotoxicity testing. Most notably, the Organisation for Economic Cooperation and Development (OECD) have recently been actively taking steps to improve the statistical content of their guidelines and their current draft guidelines are a considerable improvement on earlier ones (N. Grandy, OECD, pers. comm.).

In order to gauge the severity of the problem the OECD initially commissioned a study of their own aquatic guidelines. The unpublished report of this study (Pack, 1993), together with another unpublished, but influential, report (Noppert *et al.*, 1994) was highly critical of the use of statistics by environmental toxicologists. The main concerns in both of these documents centre on the selection of summary statistics that adequately describe the results from toxicity tests, and on the optimal experimental design for achieving test objectives. The view of Pack (1993) and most of the contributors to Noppert *et al.* (1994) is that current test guidelines and procedures are sub-optimal in both respects.

As part of the continuing drive to improve the statistical content of guidelines, a two day workshop entitled *Asking the Right Questions: Ecotoxicology and Statistics* was held at Royal Holloway University of London from 26–27 April 1995 under the auspices of SETAC-Europe. Twenty four invited participants from the US, Canada, the UK, the Netherlands, Denmark, Germany and Italy were asked to consider key questions about the current description and use of statistics in toxicity test guidelines.

0963–9292 © 1996 Chapman & Hall

The workshop participants were divided into four groups and asked to review several guidance documents on toxicity tests with crustaceans, fish, terrestrial animals, and aquatic and terrestrial plants (see Table 1 for a list of the guidance documents). These guidelines were selected by the organisers as representative of those across the field of regulatory toxicity testing. Group members were asked to provide general comments on design and analysis, rather than a detailed critique of each guideline.

The purpose of the workshop was to review the statistical content of a selection of current guidelines and identify ways in which they can be improved. The intended outcomes of the workshop were: (i) to identify areas of agreement between biometricians working in environmental toxicology, and develop a plan for promoting wider acceptance of these ideas; and (ii) to identify areas of uncertainty, and develop a plan for research that resolves these uncertainties.

This paper summarizes the conclusions from the workshop. It provides recommendations for good statistical practice within current ecotoxicity testing guidelines and highlights areas where further actions and research are required.

Current status of statistics in ecotoxicity test guidelines

The workshop participants reviewed the statistical content of existing guidelines and observed that, whilst they varied greatly, in most cases there was scope for a great deal of improvement. The following points summarize the current situation:

Table 1. Guidelines reviewed in workshop

OECD Guideline for Testing of Chemicals, 201, 'Alga, Growth Inhibition Test', Adopted 7 June 1984.

OECD Guidelines for Testing of Chemicals, 202, 'Daphnia sp., Acute Immobilisation Test and Reproduction Test', Adopted 4 April 1984.

OECD Guideline for Testing of Chemicals, 203, 'Fish, Acute Toxicity Test', Adopted 17 July 1992.

OECD Guideline for Testing of Chemicals, 206, 'Avian Reproduction Test', Adopted 4 April 1984.

OECD Guideline for Testing of Chemicals, 207, 'Earthworm, Acute Toxicity Test', Adopted 4 April 1984.

OECD Guideline for Testing of Chemicals, 208, 'Terrestrial Plants, Growth Test', Adopted 4 April 1984.

OECD Guideline for Testing of Chemicals, 210, 'Fish, Early-life Stage Toxicity Test', Adopted 17 July 1992.

EPA Hazard Evaluation Division Standard Evaluation Procedure, 'Honey Bee – Toxicity of Residues on Foliage', EPA-504/9-85-003 June 1985.

EPA Hazard Evaluation Division Standard Evaluation Procedure, 'Acute Toxicity Test for Freshwater Invertebrates' EPA-540/9-85-005 June 1985.

EPA Hazard Evaluation Division Standard Evaluation Procedure, 'Fish Life-Cycle Toxicity Tests', EPA 540/9-86-137 July 1986.

EPA Hazard Evaluation Division Standard Evaluation Procedure, 'Daphnia Magna Life-Cycle (21-Day Renewal) Chronic Toxicity Test', EPA 540/9-86-141 July 1986.

European Commission Directive 67/548/EEC. C. 2. Acute Toxicity for Daphnia.

(1) The degree of attention given to different guidelines by the same organization varies greatly. This applies across the board, not only with respect to statistical content. Thus some guidelines have twenty or more pages and deal with topics thoroughly, whilst others have five pages or less and seem much more superficial.

(2) Objectives can be vague or non-existent. In some cases it is only possible to deduce the purpose of the test by studying the reporting requirements.

(3) Description of designs can be poor or non-existent. For example, in one plant test guideline (OECD Guideline 208) it is not clear whether there should be five plants per pot or five pots each with one plant.

(4) Recommendations on methods of analysis are in general poor and sometimes non-existent.

(5) Reporting requirements are usually helpful in that some guidance is given, but often the terminology is confusing.

The purpose of statistics in ecotoxicity testing: accurate and precise estimates of effects

The purpose of environmental toxicity testing is to provide one half of the data package required for environmental risk assessment. Knowledge of the toxicity of a chemical obtained from toxicity tests is then combined with estimates of likely environmental exposure to the chemical in order to predict the likely effect of the chemical in the environment.

The purpose of statistical methods in support of individual toxicity tests is to enable the most precise estimate of toxicity to be obtained for that test, within the constraints of the resource allocated. There are at least three possible counter-arguments to this point of view, all of which have been expressed, and all of which are concerned with lack of precision elsewhere in the risk assessment procedure (de Bruijn, 1994).

One argument focuses on the fact that the results of a test, for example an LC50 (the 'lethal concentration' of the test chemical for 50% of the organisms tested) and associated confidence intervals, apply to that test alone and therefore fail to take into account the variation in environmental conditions likely to be encountered were several tests to be conducted. Nominally identical tests with the same strain of organism and identical environmental conditions, but carried out at different times, may give quite different results. A second argument focuses on the fact that estimates of environmental exposure used in risk assessment are subject to large errors. A third argument focuses on the fact that large safety factors are normally applied to toxicity data to take account of intra- and inter-specific variation in sensitivity. Improvements in test precision, according to this argument, may have a negligible impact on a final risk assessment when compared with the choice of safety factor.

Whilst the factual content of these arguments is recognized, they are not adequate justification for carrying out poor ecotoxicity tests. Errors are cumulative, so reducing an error in one part results in a lower aggregate error for the whole process. Thus, for a given level of resource, it makes sense to use optimal methods of experimental design and analysis to obtain the best possible estimate of toxicity from each individual test. Lack of precision elsewhere is a real issue that needs to be addressed, but is beyond the scope of this paper.

Typical ecotoxicity experiments from a statistical point of view

Ecotoxicity tests employ relatively straightforward experimental designs compared with experiments commonly encountered elsewhere. They can usually be classified under one of two headings:

(1) Control plus single concentration. This design comprises two 'treatments': a single concentration and an untreated control, each replicated a number of times. This design is used to study the effect of the chemical at the chosen concentration. The usual method of analysis is to estimate the difference in response between the two treatments together with confidence intervals. An alternative analysis is to carry out a test of statistical significance under the null hypothesis that the two treatments give identical responses. This design cannot usually be analysed by fitting a concentration-response curve.

(2) Multiple concentration design. This design comprises an untreated control and a number of increasing concentrations, each replicated a number of times. Data from this type of test can be analysed either by comparing means, for example via an analysis of variance (ANOVA), or by fitting a concentration-response curve from which quantities of interest, such as an LC50 can be estimated.

The types of data commonly encountered in ecotoxicological studies can be summarized as: (1) Counts or proportions with a known upper limit, such as mortality data in acute toxicity tests. Such data are likely to follow a binomial distribution, possibly with extra-binomial variation. This type of data is often called 'quantal' and will be referred to as such in the remainder of this paper. The other types of data listed below will be referred to as 'non-quantal'. (2) Counts with no known upper limit, such as number of offspring in sub-lethal studies. Such data can often be assumed to follow a poisson or negative binomial distribution, though normality can be assumed under some circumstances, or achieved via transformation. (3) Score data. This type of data results when a subjective judgement of some kind is made regarding the effect of a treatment on a particular test organism. It can vary from a very crude qualitative assessment to a reasonably fine quasi-continuous measurement, such as the percentage kill (or damage) inflicted by a chemical on a plant relative to the condition of an untreated control plant. The exact distributional nature of this type of data is unknown and, in any case, can vary depending upon the assessment being made. (4) Continuous measurements, such as height, weight and length of organisms under study. This type of data can often be assumed to follow a normal or log normal distribution.

Binomial data can be analysed by use of Fisher's exact test (Daniel, 1990) when there are two treatments, or by fitting a probit or logit concentration-response model when there are more than two treatments (Finney, 1971). Data from a normal distribution can clearly be analysed by the well known methods, e.g. ANOVA followed by a multiple comparison test to test the significance of the difference between treatment means, and linear or non-linear regression analysis to facilitate the estimation of EC (effective concentration) values. With data from other distributions, such as the poisson, some judgement is required. Often it will be possible to use methods of analysis based on the normal distribution by transforming the data or weighting the analysis (Armitage and Berry, 1987), and this is the most likely route for most analysts. Alternatively, a generalized linear model may be fitted (McCullagh and Nelder, 1989) which takes account of the true error distribution of the data.

Recommended topics for inclusion in a guideline – experimental design and study execution

Of the two statistical activities, experimental design and data analysis, design is by far the most important. This is because a good statistical design confers a high probability of success in achieving the objectives of a test, and ensures that estimated quantities of interest are free from bias and are sufficiently precise. Data analysis is merely a way of obtaining an estimate once an experiment has been completed. If insufficient attention is given to the requirements of good experimental design then the test may produce data that are incapable of providing the desired information. No amount of elegant analysis will then be able to salvage the situation. On the other hand, a poor initial analysis of data from a well-conducted experiment can easily be corrected at a later date.

This section of the paper briefly describes some of the issues that should be considered when designing an ecotoxicity test. For further general advice on experimental design, presented in a very practical way, see Cox (1958).

The need for clear objectives

Someone engaged in the statistical design of an experiment needs a clear view of how the data will be analysed and precisely what needs to be estimated. In designing an experiment therefore, it is important that the choice of treatments allows the estimation of summary statistics of interest; and the standard errors (or confidence intervals) for estimates fall within prescribed limits by considering the number of concentrations and their values, the number of replicates, whether or not the experiment should be blocked and so forth.

The first step in the design process is therefore to take the declared objective of the experiment and translate it into some quantitative measure that can be estimated, such as an EC50 (the 'effective concentration' or concentration of test chemical that affects 50% of the organisms tested). The experimental objective must be precise if it is to be of any value; it may not be possible to design an experiment capable of meeting a vague objective.

The objective of most ecotoxicity tests in which mortality is the endpoint is to fit concentration-response curves to the mortality data, and to summarize this as an LC50. It is not clear that the LC50 on its own is a sensible summary of the results of a test. Since the whole of the concentration-response curve can be characterized by two parameters, namely the LC50 and the slope of the curve, it makes sense to estimate both. Then LC points other than the LC50 can be estimated and used in risk assessment, if required.

For sub-lethal studies the emphasis is usually on calculating the no observed effect concentration (NOEC: the highest concentration of chemical evaluates in a test that does not cause statistically significant differences from the controls), a controversial summary statistic discussed later at greater length.

Randomization

For all experiments in which the experimental material is highly variable, and in which the variation tends to be large relative to the signal being measured – a condition affecting all ecotoxicological experiments – the treatments should be allocated to the experimental units in a random fashion. Furthermore, all handling of experimental units after the allocation of treatments, such as taking measurements, should also be done in a random way. This requirement follows from the fact that it is randomization that ensures that estimates are unbiased, so failure to apply this fundamental rule may lead to a false conclusion. Without randomization one might falsely conclude that a chemical has a toxic effect when there is none, or fail to detect a toxic effect when one is present. Alternatively, the estimate of the magnitude of a toxic effect may be incorrect. If, for example, experimental material were to be dosed or handled in concentration order, then increasing operator tiredness or shifts in machine performance could lead to an effect at higher concentrations that is greater or lesser than that caused by the chemical alone.

It is recognized that for practical reasons it is not always possible to follow this advice. For example, a test compound at a specific concentration is often made up once and applied to all replicates at the same time. Where possible, however, the principle of randomization should be adopted and any major deviation that could affect test results should be reported.

Replication

Replication allows the power or precision of a test to be controlled: larger numbers of replications lead to more precise estimates or tests with greater power. Precision and power are related statistical concepts: precision refers to the width of a confidence interval around an estimate, with narrower intervals giving more precise estimates; power is the probability of detecting a difference between two treatments via a formal significance test, with a higher probability corresponding to greater power. The number of replications is not the only factor affecting power and precision: both are also affected by the natural variation in the experimental material, and the number of degrees of freedom for the estimate of residual error. In addition, power is affected by the choice of multiple comparison method, the choice of type I error in the test (also known as the alpha or significance level) and the size of the treatment difference it is desirable to detect.

When ANOVA is the likely form of analysis, all guidelines should specify the precision or power required in order that an adequate number of replications can be chosen. Alternatively, the number of replications should be specified and in this case it should be justified by stating the desired power or precision. Statements on power tend to be of the form ‘the test should have an 80% chance of detecting a difference of 20% from the control using test X with a 5% significance level’. For analysis of variance a useful discussion on power, including instructions on how to do the required calculations, is given in Cohen (1988) and Sokal and Rohlf (1995). These references also give required calculations for measures of precision, statements of which tend to take the form ‘confidence intervals should be no more than plus or minus X% of the estimated value’.

If a dose response model is likely to be fitted then the situation is more complicated because the precision of an estimate of LCx of ECx is also affected by the number of concentrations and their values.

Blocking

The purpose of blocking is to increase the power or precision of a test without the need to allocate extra resource. Blocking may not always be necessary but in some situations it will confer greater precision or power on the test. This is likely to be the case if the test results are affected by some known or measurable factor of the test other than the treatment itself. In this case the replicates should be arranged in blocks so that the factor in question exhibits large variation between blocks and small variation within blocks. If such blocking is successful then some of the natural variation in the experimental material can be allocated to a block effect in the analysis of results, resulting in a smaller estimate of residual error variance and hence increased power and precision.

There are many possible factors that could be used to assign experimental units to blocks. Some examples are environmental variation in the laboratory, initial weight of the experimental organism, machine used in making measurements, and variation between operators. Often it may be wise or necessary to block with respect to more than one factor.

Though blocking can greatly increase the power of an experiment, it can also reduce the power if done badly. It therefore calls for a great deal of judgement and, preferably, advice from a statistician.

Number of organisms per experimental unit

Having more than one organism per experimental unit will usually improve power and precision, although the extent of the improvement will depend on the size of the within-unit variation and how large it is relative to the between-unit variation. If the within-unit variation is likely to be large compared to the between-unit variation then worthwhile improvements in power are possible by housing more than one organism per unit.

Therefore, in addition to specifying the number of replicates per treatment, guidelines should specify the number of organisms per unit, taking account of the requirements for power and precision discussed above. A description of the required calculations can be found in Cohen (1988) and Sokal and Rohlf (1995).

Individual organisms housed together in a single experimental unit are strictly not replications, although it is fairly common to see tests analysed as if they were. This is referred to as pseudoreplication (Hurlbert, 1984) and is highly undesirable. At worst it can give the impression that a test is much more precise or powerful than it really is.

It is recognized that pseudoreplication does significantly reduce the cost of toxicity testing, for example by reducing laboratory space per test, and in some circumstances a test would not be possible without it. In these circumstances it is absolutely essential that preliminary work to validate the methodology be carried out prior to the recommendation of pseudoreplication in a guideline.

Number and spacing of concentrations

The choice of concentrations, both the number of them and their value, affects the precision of LC and EC estimates. Guidelines currently often require four or five concentrations that are geometrically spaced, in addition to an untreated control. In acute studies, they also often require a minimum response of 0% and a maximum response of 100% in quantal studies.

It has been pointed out that five concentrations are not strictly essential in order to be able to fit a concentration-response model. Three concentrations may be adequate for a reasonably precise estimate of an LCx or ECx (Pack, 1993). Thus estimates of ECx or LCx should not be disregarded simply because they are based on few concentrations. The size of the confidence intervals should indicate how acceptable the estimates are. Furthermore, the requirement for 0 and 100% response in quantal studies is not necessary, and perfectly acceptable estimates of LC50 can be obtained without this restriction. However, it would be unwise to design a test with only three concentrations, since the risk of failing to obtain a precise estimate would be very high. As a general guide, therefore, one should aim for at least five concentrations that will not give 0 or 100% mortality.

Research to identify the optimal number and location of concentrations for LC and EC estimation has been carried out by a number of authors but, because of a lack of consistency in the results, further work is needed. Muller and Schmitt (1990) carried out simulations for a number of different designs, each with 48 experimental units in total. They concluded that more concentrations and fewer replications per concentration led to more precise estimates of LC50, but that there was relatively little gain when the number of concentrations exceeded twelve. In contrast, Robertson *et al.* (1984) carried out simulations in the context of insect toxicology, concluding that for lethal tests the number of concentrations is unimportant, and that the minimum number of organisms is 120 for reliable concentration-response experiments. However, since they only compared designs with eight and five concentrations their results may lack generality. They also show that precise LC50 tests require concentrations to be equally spaced on a log scale between concentrations giving 25 and 75% effects. Furthermore, precise LC10 estimates were shown to require one or two responses above 90% and the majority between 5 and 25%. Again these results are conditional upon certain aspects of the simulations, such as extreme response values, and may not be generally true. Finney (1971), Adbelbasit and Plackett (1983), Kooijman (1983), Chaloner and Larntz (1989), Kalish (1990), and Sitter (1992) have also made contributions to this debate.

For sub-lethal studies producing non-quantal data, the authors are unaware of any research that recommends an optimal choice of concentrations for estimating EC points.

Optimum times for taking measurements

Guidelines often require measurements to be taken at more than one time although there is currently no statistical basis for this. With time-to-response models or with methods of longitudinal data analysis (Diggle *et al.*, 1994; Fahrmeir and Tutz, 1994) it should be possible to recommend both optimum number and location of measurement times in order to maximize precision for a given input of resource.

Blind assessing

It is often advisable that the operator taking measurements on the experimental units should be unaware of which treatment was applied to each unit. In some situations this can help to eliminate bias because the operator cannot then contaminate the test results with his or her own expectations of the outcome. This is related to the issue of randomization since, in a non-randomized experimental layout, it may be obvious which treatments were applied to which units.

Whether or not this is worth doing depends very much on the test endpoint. If the endpoint is mortality and there is no chance of confusing living and dead organisms then there will be no need for blind assessment. However, if the endpoint is more subjective, then blind assessing may be advisable.

Recommended topics for inclusion in a guideline – methods of statistical analysis

The following discusses the details and methods of statistical analysis that might usefully be included in ecotoxicity test guidelines.

A description of the analysis to be carried out. The recommended analysis will depend upon the design and the type of data being collected. Examples of appropriate analyses are as follows:

(1) Control/single concentration design with quantal data. Data from this test can be analysed as a contingency table using Fisher's exact test (Daniel, 1990) to test for an effect of treatment. The estimated mean percentage mortalities for the control and treated group should be calculated together with their confidence limits.

(2) Multiple concentration design with quantal data. A concentration-response model that takes account of the binomial distribution of the data at a given concentration should be fitted, and both the LC50 and the slope of the response should be estimated together with confidence limits. The tolerance distribution of the organism could be selected from the logit, probit, Weibull, or Gompertz (Newman, 1995). If extra-binomial variation is present, then estimates of variance should be adjusted accordingly. For a full description of this method of analysis see Finney (1971), Collett (1991), or Morgan (1992). If there are only one or two responses that are not 0 or 100% kill, then the Spearman-Karber non-parametric method (Finney, 1971; Hamilton *et al.*, 1977) or the moving average method (Stephan, 1977) can be used. If there are no responses other than 0 and 100% then the geometric mean of the highest concentration giving 0% kill and the lowest concentration giving 100% kill should be used as an estimate of the LC50. In this last case, confidence intervals can be estimated as described in Williams (1986) or van der Hoeven (1991).

(3) Control/single concentration design with non-quantal data. The treatment means and a pooled estimate of residual error should be calculated. The standard errors and confidence intervals of the treatment means should then be calculated. A significance test, such as a *t*-test, can be carried out to compare the treatment mean with that of the untreated control. Alternatively, the confidence interval for the difference between the control and treatment means can be calculated. If the confidence interval does not include zero this implies a statistically significant result, i.e. the single concentration has a clear effect on the test organism.

(4) Multiple concentration design with non-quantal data. An analysis of variance could be carried out, and a multiple comparison method then used to determine which concentrations have mean responses that are significantly different from the control. This is the approach that would be used if a NOEC was to be determined or if an estimate of the difference between pairs of treatments was required. A preferable approach would be to fit an appropriate concentration-response model and estimate an ECx value together with confidence limits. If necessary, the data should be transformed prior to analysis or a weighted analysis should be carried out.

The methods of analysis for non-quantal data discussed above are based on the normal distribution. Such methods are familiar to ecotoxicologists and there is a wide choice of software available that enables them to be used. Methods based on the normal distribution are therefore likely to be very popular. However, equivalent methods based on different distributions, such as the poisson or negative binomial could be used. For a comprehensive description of these methods see McCullagh and Nelder (1989). For an example of poisson regression in ecotoxicology see Bailer and Oris (1993).

Model checking – testing the assumptions of the analysis

There are a number of ways in which a fitted model or analysis may be inadequate and it is therefore essential that certain checks be made. If a concentration-response model is fitted then, for a variety of reasons, the functional form of the model may be incorrect. The data may contain outliers (Barnett and Lewis, 1979) or observations called 'influential values' (Atkinson, 1985) that have an undue impact on the conclusions drawn. The assumption that the observations come from a particular probability distribution may also be incorrect. The following describes some of the tests of assumptions that can be made for the types of data collected in typical ecotoxicological tests.

When ANOVA is the preferred form of analysis, three assumptions need to be considered: independence of errors, normality of errors, and homogeneity of variance between treatments. Of these, two are less important: independence of errors is usually guaranteed by randomization; and ANOVA is reasonably robust to non-normal errors (Scheffé, 1959). This leaves homogeneity of variance as the main assumption which must be satistified. Formal tests of homogeneity of variance such as Bartlett's test (Winer, 1971) are not recommended. This is because normality of data is a rather strict condition of such tests but is not required for ANOVA. A less formal graphical method is perfectly acceptable. This involves calculating the residuals (i.e. the individual observations minus the treatment means) and plotting them on a graph with concentration on the x-axis. Inspection of the graph easily reveals whether or not the variance is non-homogeneous. Non-homogeneity of variance can be corrected by using transformations or by weighting the analysis. Commonly used transformations are $\log_{10}$ or square root for counts, and arcsine for proportions and percentages. If a weighted analysis is carried out, the weight should be proportional to the reciprocal of the variance. For example, if the distribution of observations at a given concentration is assumed to be poisson, then the weight would be the mean response or the fitted value at that concentration. For more detail on transformations and weighting refer to Armitage and Berry (1987).

When a concentration-response model based on the normal distribution is fitted to non-quantal data, a check for homogeneity of variance is needed (as above). In addition, a check must be made of the validity of the fitted model. This can be done again by plotting a graph of the residuals against the concentration, although in this case the residuals are the observations minus the prediction from the model (as opposed to the treatment mean). A more formal check of the validity of the model can be made by dividing the error mean square into two components, one called 'lack of fit' and the other called 'pure error' (Draper and Smith, 1981). A significance test for lack of fit can then be carried out.

When a concentration-response model based on a distribution other than the normal is fitted, then both over-dispersion and model adequacy need to be checked for. The procedure is similar to that described for quantal data. For a full description of the methods see McCullagh and Nelder (1989).

When Fisher's exact test is the form of analysis there is little checking to be done.

The distribution of observations in each of the two treatment groups is assumed to be binomial, in which case the variance of the estimated percent mortality for each treatment is a function of the mortalities themselves. In some situations, such as when organisms are held in groups and there are several groups per treatment, the observed variance may be greater than we would expect from the observed mortalities, a condition known as over-dispersion or extra-binomial variation. Over-dispersion should be tested for by a simple chi-square test (Collett, 1991) and, if found to be present, standard errors, confidence limits and significance tests should be adjusted.

When a concentration-response model is fitted to quantal data a check for over dispersion, as described above, should be carried out in addition to a check of the validity of the functional form of the chosen model. The difficulty in doing this is that the chi-square test described above could indicate either lack of fit or over-dispersion or a mixture of both. The recommended procedure is thus as follows. First fit the model and carry out the chi-square test for lack of fit and over-dispersion, and if the result of the test is non-significant assume the model fits well. If the test gives a significant result a check of the adequacy of the model should be made by plotting a graph of standardized residuals against concentration. There are a number of different types of standardized residuals but the 'likelihood' or 'deviance' residuals are preferred (Collett, 1991). If it seems as though the model does not fit well then repeat the process trying a new model. Finally, if the model seems to fit well and the chi-square test still gives a significant result it may be assumed that over-dispersion is present and standard errors and confidence intervals should be adjusted accordingly.

The methods for model checking described so far are appropriate if a model is being fitted to some data for the very first time. However, some would argue that this approach is not correct in the context of ecotoxicology, in which a fairly routine process of experimentation, data generation and analysis takes place. According to this line of thought, for example, if over-dispersion occurs in 90% of tests then it should be corrected for in 100% of analyses, even when it is not detected in specific tests. Whilst this argument deserves serious consideration, it is a subject for debate and it is doubtful that all statisticians would agree with it. And since all statistical analysis calls for judgement, it should be left to the analyst to decide. Therefore, in order to give as much help as possible to the analyst when taking this decision, guidelines should give as much information as possible about experiences in analysing data likely to be generated. Such experience is likely to be gained, for example from ring tests.

Outliers and influential observations

Outliers are observations that seem to be extreme when compared with other observations. Either they come from the tail of the distribution exhibited by the other observations or they are from a different distribution. Either way they are likely to violate the assumptions of the analysis that is carried out and may have a large and undesirable effect on the analysis.

Checking for outliers should be a standard requirement in all guidelines, but it is not easy and calls for both skill and judgement. In doubtful situations the analysis should be performed and presented both with and without suspected outliers. Many formal tests for outliers are available (Barnett and Lewis, 1979) but they should not be relied on alone. Simple graphical methods, such as plotting replicate data against concentration are probably just as useful.

Influential observations are values that affect the result of an analysis more than is usual. Examples are observations at the extremes of the concentration range, which have a much larger effect on the slope of a fitted line than more central observations. For a full discussion of influential observations in statistical models see Atkinson (1985). If an observation is suspected of having an undesirable effect on an analysis the analysis should be carried out both with and without the observation and, if the results are different, both analyses should be reported.

Non-parametric methods

In situations in which parametric methods would otherwise be used, some guidelines advocate the use of non-parametric methods after a test for normality has shown that data are non-normal. However, parametric methods based on the normal distribution that are likely to be used in analysing toxicity data are fairly robust in the sense that lack of normality has little effect on their performance. Thus, non-parametric methods should not be used simply because data is non-normal.

Of course, non-parametric methods are extremely useful for analysing ecotoxicity data in specific situations. Examples are Fisher's exact test and the Spearman-Karber method.

Multiple comparison methods

Multiple comparison methods are used for comparing treatment means in the analysis of variance. There are many different tests to chose from, but those most frequently encountered in ecotoxicology are the *t*-test (or LSD method), Williams' test and Dunnett's test. An unfortunate feature of multiple comparison methods is that each method gives a different result. For a comprehensive discussion of these techniques see Day and Quinn (1989).

The use of multiple comparison methods is highly unsatisfactory in practical situations and should be avoided, where possible, in ecotoxicology. They are only used in regulatory studies for the purposes of calculating NOECs and are one of the many reasons that NOECs themselves are unsatisfactory.

Measured versus nominal concentrations

It is not clear whether nominal or measured concentrations should be used when fitting concentration-response curves. Nominal concentrations should be used if errors in measuring concentrations are larger than errors in applying test chemicals to experimental units. This is a potential area of research and so at present it is not possible to make firm recommendations.

Confidence intervals

All guidelines should insist on confidence intervals being estimated and reported.

Threshold, hormesis and time-to-response models

A number of models have been proposed that allow direct estimation of no-effect concentrations (NECs). A No Effect Concentration is the highest concentration of test chemical at which there is no zero effect on the organisms tested. This is in contrast with the NOEC, which might allow a large, non-zero effect on test organisms so long as it does not differ significantly from the controls. So far these models seem to have had little impact in regulatory work, which is unfortunate because they seem to offer real benefits. Before they can be recommended in guidelines more research needs to be carried out. This work should involve collating the results of existing research, identifying gaps in knowledge and then doing the work necessary to fill the gaps.

Cox (1987) reviews a number of different threshold models for analysing quantal data; in these models the NEC is included as an explicit parameter and it is assumed that at concentrations less than the NEC there is no response to the toxicant. Brain and Cousens (1989) propose a hormesis model for non-quantal data; hormesis occurs when very low concentrations of test chemical lead to a stimulatory effect and so the NEC can be estimated as the concentration at which the response is equal to the response of the untreated control.

A number of authors have proposed time-to-response models for both quantal and non-quantal data (Kooijman, 1993; Pack, 1993; Newman, 1995; Newman and McCloskey, 1996). These models, although different in detail, have two common features in that they permit all of the data from all times of assessment to be included in a single analysis and include the NEC as an explicit parameter. They make use of techniques such as survival analysis, failure time analysis, and life data analysis that are widely used in medical and engineering research but not in ecotoxicology (Collett, 1994). Such models offer real prospects for enhancing the power of tests and research is needed to confirm their potential.

Statistical reporting of results

The following information should always be reported:

(1) The raw data

(2) a graph of the replicate data should be plotted with concentration on the x-axis. Ideally the treatment means and any fitted curve should be plotted, either on the same graph as the raw data, or on a separate graph

(3) a full description of the design and methods of statistical analysis employed

(4) parameter estimates of interest, such as an EC50, together with confidence limits

(5) when a curve is fitted, the slope of the concentration-response curve plus confidence interval

(6) if a NOEC is required, the treatment means, their standard errors, the error degrees of freedom and the least significant difference, or standard error of the difference.

In short, the report should provide enough information to enable the analysis to be repeated in an identical fashion, should provide evidence that statistical models or analyses are adequate for the data, and should give useful measures of toxicity.

The no observed effect concentration (NOEC)

Test guidelines, particularly for chronic tests, often ask for both an EC estimate and for a NOEC to be determined for sub-lethal endpoints. This can lead to difficulties in designing ecotoxicity experiments because of the different design requirements of dose-response modelling and hypothesis-testing, a fact recognized by ecotoxicologists for some time (Stephan and Rogers, 1985). We have already touched upon the issue of the no observed effect concentration in ecotoxicology, and would like to continue by reviewing the reasons why the NOEC is a poor summary statistic.

The NOEC has been severely criticized, on both theoretical and practical grounds, in a number of other publications (Kooijman, 1981; Skalski, 1981; Stephan and Rogers, 1985; Hoekstra and van Ewijk, 1993; Pack, 1993; Noppert *et al.*, 1994; Laskowski, 1995; Kooijman, 1996). Most biometricians favour EC estimation over calculation of NOECs for the following reasons.

(1) The NOEC must be one of the concentrations used in an experiment since hypothesis testing does not allow interpolation between test concentrations. Thus, an important determinant of the NOEC is the choice of test concentrations;

(2) the NOEC tends to increase as the precision of an experiment decreases. Since a larger NOEC implies a safer chemical, the approach rewards those who perform poor experiments;

(3) confidence intervals cannot be calculated for the NOEC. It is therefore not possible to compare the accuracy of NOEC values from different experiments;

(4) a NOEC is not always obtainable. In particular, it cannot be determined when the lowest test concentration produces a statistically significant effect when compared with the control. If the calculation of a NOEC is a regulatory requirement, the experiment may have to be repeated, probably unnecessarily;

(5) NOECs may occur at concentrations which actually cause large effects because high experimental variability reduces statistical sensitivity, thus preventing these effects being detected as statistically significant (Barnthouse *et al.*, 1987; Suter *et al.*, 1987; Masters *et al.*, 1991; Leisenring and Ryan, 1992). The NOEC cannot therefore be considered an estimate of a safe dose;

(6) the NOEC contravenes one of the basic rules of modern scientific and statistical method, by attempting to 'prove' the null hypothesis of 'no effect', instead of disproving the presence of effects (Skalski, 1981; Hoekstra and van Ewijk, 1993);

(7) calculation of a NOEC does not provide information on the range of sensitivity of an organism to the test compound (Bruce and Versteeg, 1992), information which can easily be obtained by other standard methods of analysis. The NOEC is therefore very wasteful of data;

(8) the NOEC depends upon the choice of α (type I error rate) used in a significance test and also on the choice of test. Thus a 5% t-test may produce a different NOEC to a 1% t-test. Similarly, the t-test, Dunnett's test and Williams' test may produce different NOECs; and

(9) it may be difficult to determine a NOEC if the observed treatment means do not follow a monotonic trend. Thus, for example, if there are five concentrations in a test and the third and fifth are significantly different from the control but the fourth is not, then we could choose the second or the fourth concentration as the NOEC.

EC estimates have the advantage that they generally overcome all of the above criticisms. The particular strengths of this approach are given.

(1) The ECx is not restricted to be one of the test concentrations since regression analysis permits the estimation of effects at untested concentrations;

(2) the value of an ECx does not depend upon the precision of the experiment;

(3) the precision of the ECx can be estimated and can be reported as a confidence

interval, thus allowing ECx estimates and their associated confidence intervals to be meaningfully compared between tests;

(4) an ECx should always be obtainable provided sufficient care has been taken in designing the experiment, particularly in the choice of concentrations;

(5) the interpretation of the ECx is straightforward in that it is the concentration expected to result in an x% effect;

(6) the ECx and its confidence intervals provide a range of plausible values for a safe dose, allowing investigators to judge whether or not these values are of concern;

(7) the regression model used to estimate an ECx allows the entire toxic response of an organism to be characterized;

(8) the choice of α (type I error rate) affects only the confidence limits on the ECx, not the ECx itself;

(9) regression modelling is sufficiently flexible to be able to model a wide range of concentration-response situations including non-monotonic relationships such as hormesis;

(10) regression modelling allows the analysis of both lethal and sub-lethal data to be handled using the same basic approach;

(11) regression modelling utilizes data from all concentrations, whilst only one is used for determination of the NOEC; and

(12) both measured and nominal concentrations can be used to estimate ECx values.

However, whilst NOECs can be severely criticized on statistical grounds and regression modelling seems to offer distinct advantages, the use of ECx estimates is not a panacea for all of the statistical problems associated with ecotoxicity testing. There are a number of difficulties associated with regression modelling which need to be resolved.

The difficulty in choosing an appropriate model is often put forward as a disadvantage of regression modelling. However, most statisticians would see it as part of their daily routine to fit a variety of models to a set of data and choose the one with best fit. To help achieve this there is a wealth of literature on regression diagnostics such as Draper and Smith (1981), Cook and Weisberg (1982) and Atkinson (1985). Some research is necessary in order to identify suitable classes of model that are appropriate to sub-lethal studies. In particular, the usefulness of threshold and hormesis models needs to be investigated since these models offer the potential benefit of being able to estimate a NEC. Also requiring some further examination are models proposed by Newman (1995), Kooijman (1993), and Newman and McCloskey (1996) that allow all of the data from a test to be utilized in one analysis. Furthermore, choosing an appropriate model is no more difficult than selecting a test for performing multiple comparisons.

Another difficulty in the use of ECx estimates is that it requires an x to be specified. This is both a biological and a statistical issue. On the one hand, it requires the biologist to think in terms of effects (rather than of no-effects). There is currently no general agreement on the appropriate choice of x and a large number of possible values have been proposed in the literature. It is highly likely that different organisms or endpoints will require different values of x. From the statistical viewpoint, it may be impractical to try to estimate very low ECx values, such as EC5, as the confidence intervals around the estimates may be very wide. In addition, it is in this region of the concentration-response curve that different models can yield very different ECx

estimates. This is therefore another topic requiring both some biological and statistical research. ECx estimates may also be difficult to obtain in some cases, particularly if there are few responses between 0 and 100%. In these situations, it is possible to obtain a crude estimate of an EC50 and Crump (1984), Williams (1986) and van der Hoevan (1991) have all shown how to derive confidence intervals in these situations.

Finally, finding suitable user-friendly software for fitting the necessary range of regression models may be a problem for a great many non-statisticians. Routines are widely available in professional statistical packages but their successful use can often require some detailed knowledge of statistics. Providing suitable user-friendly software to non-statisticians is therefore seen as an important stage in the replacement of NOECs with ECs or parametric NECs.

Conclusions

Statistical advice in current ecotoxicity test guidelines is in need of improvement. More advice should be given on experimental design, statistical analysis and reporting of results. This might most usefully be in the form of a manual that specifically addresses the needs of the biologists responsible for ecotoxicity testing. This paper has identified the items that should be included in guidelines or an accompanying manual.

The use of the NOEC as a summary statistic has been criticized since the early 1980s. We summarize these criticisms and show why most biometricians favour a move away from NOECs and towards EC estimation.

Several areas for further statistical research can be identified. There should be further validation of threshold, hormesis and time-to-response models to determine whether their use in regulatory toxicity testing would be beneficial and cost-effective. The statistical implications of selecting particular values of x in ECx and LCx should also be determined. For example, is it practical to attempt to estimate EC points for low values of x, such as EC5, even if there is biological justification for doing so? If not, how large does x need to be before it does become practical? The answer to these questions will probably differ for different types of models. Research is required into the effect on statistical accuracy and precision of the number and spacing of concentrations. This needs to be done for both quantal and non-quantal data. The feasibility of setting statistical quality control criteria on allowable test variability should also be examined. For example, should test results be declared invalid if control mortality exceeds a certain value and, if so, what should that value be? This requires a study to be made of existing regulatory data. Finally, the effect on statistical accuracy and precision of using nominal versus measured chemical concentrations should be investigated.

None of these research needs require further experimentation. Existing data sets or simulations can be used to answer all of these questions.

Acknowledgements

The authors thank the workshop participants for contributing their time and expertise: Glenn Atkinson, Ian Barber, Jacques J.M. Bedaux, Francesco Bertani, Gregory J. Carr, G. Peter Dohmen, John S. Fenlon, Michael Harrass, Helle Holst, Simon Hoy,

S.A.L.M. Kooijman, Robert M. Mulliss, Mike Newman, Niels Nyholm, Andrew Riddle, Tim Sparks, Bruce Stanley, Nelly van der Hoeven and Barry Zajdlik.

References

Adbelbasit, K.M. and Plackett, R.L. (1983) Experimental design for binary data. *J. Amer. Stat. Assoc.* **78**, 90–8.

Atkinson, A.C. (1985) *Plots, Transformations and Regression*. Oxford: Oxford University Press.

Armitage, P. and Berry, G. (1987) *Statistical Methods in Medical Research*. Oxford: Blackwell.

Bailer, A.J. and Oris, J.T. (1993) Assessing toxicity of pollutants in aquatic systems. In *Case Studies in Biometry* N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest and J. Greenhouse (eds), pp. 25–40. New York: John Wiley.

Barnett, V. and Lewis, T. (1979) *Outliers in Statistical Data*. New York: Wiley.

Barnthouse, L.W., Suter II, G.W., Rosen, A.E. and Beauchamp, J.J. (1987) Estimating responses of fish populations to toxic contaminants. *Environ. Toxicol. Chem.* **6**, 811–24.

Brain, P. and Cousens, R. (1989) An equation to describe dose responses where there is stimulation of growth at low doses. *Weed Res.* **29**, 93–6.

Bruce, R.D. and Versteeg, D.J. (1992) A statistical procedure for modeling continuous toxicity data. *Environ. Toxicol. Chem.* **11**, 1485–94.

Chaloner, K. and Larntz, K. (1989) Optimal Bayesian design applied to logistic regression experiments. *J. Stat. Plan. Inf.* **21**, 191–208.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences*. New Jersey: Lawrence Erlbaum Associates, Hillsdale.

Collett, D. (1991) *Modelling Binary Data*. London: Chapman & Hall.

Idem. (1994) *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. London: Chapman & Hall.

Cox, C. (1987) Threshold dose-response models in toxicology. *Biometrics* **43**, 525–35.

Cox, D.R. (1958) *Planning of Experiments*. New York: Wiley.

Crump, K.S. (1984) A new method for determining allowable daily intakes. *Fundam. Appl. Toxicol.* **4**, 854–71.

de Bruijn, J. (1994) The political acceptability of the ECx. In How to Measure No Effect: Towards a Measure of Chronic Toxicity. F. Noppert, A. Leopold and N. van der Hoeven (eds). Unpublished workshop report, BKH Consulting Engineers, Delft, The Netherlands.

Daniel, W.W. (1990) *Applied Nonparametric Statistics, 2nd edition*. Boston: PWS-KENT Publishing Company.

Day, R.W. and Quinn, G.P. (1989) Comparisons of treatments after an analysis of variance in ecology. *Ecol. Monogr.* **59**, 433–63.

Diggle, P.J., Liang, K. and Zeger, S.L. (1994) *Analysis of Longitudinal Data*. New York: Oxford University Press.

Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis, 2nd edn*. New York: Wiley.

Fahrmeir, L. and Tutz, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer Verlag.

Finney, D.J. (1971) *Probit Analysis, 3rd edn*. Cambridge: Cambridge University Press.

Grundy, N. OECD, Personal communication.

Hamilton, M.A., Russo, R.C. and Thurston, R.V. (1977) Trimmed Spearman-Karber method for estimating median lethal concentrations in toxicity bioassays. *Environ. Sci. Technol.* **11**, 714–9.

Hoekstra, J.A. and van Ewijk, P.H. (1993) Alternatives for the no-observed effect level. *Environ. Toxicol. Chem.* **12**, 187–94.

Hurlbert, S. (1984) Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**, 187–211.

Kalish, L.A. (1990) Efficient design for estimation of median lethal dose and quantal dose-response curves. *Biometrics* **46**, 737–48.

Kooijman, S.A.L.M. (1981) Parametric analyses of mortality rates in bioassays. *Water Res.* **17**, 749–59.

idem. (1983) Statistical aspects of the determination of mortality rates in bioassays. *ibid.* **19**, 107–19.

idem. (1993) *Dynamic Energy Budgets in Biological Systems: Theory and Applications in Ecotoxicology.* Cambridge: Cambridge University Press.

idem. (1996) An alternative for NOEC exists but the standard model has to be abandoned first. *Oikos*, in press.

Laskowski, R. (1995) Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *Oikos* **73**, 140–4.

Leisenring, W. and Ryan, L. (1992) Statistical properties of the NOAEL. *Reg. Tox. and Pharm.* **15**, 161–71.

McCullagh, P. and Nelder, J.A. (1989) *Generalised Linear Model, 2nd edition.* London: Chapman & Hall.

Masters, J.A., Lewis, M.A., Davidson, D.H. and Bruce, R.D. (1991) Validation of a four-day *Ceriodaphnia* toxicity test and statistical considerations in data analysis. *Environ. Toxicol. Chem.* **10**, 47-55.

Morgan, B.J.T. (1992) *Analysis of quantal response data.* London: Chapman & Hall.

Muller, H-G. and Schmitt, T. (1990) Choice of number of doses for maximum likelihood estimation of the EC50 for quantal dose-response data. *Biometrics* **46**, 117–29.

Newman, M.C. (1995) *Quantitative Methods in Aquatic Ecotoxicology.* Boca Raton: Lewis Publishers.

Newman, M.C. and McCloskey, J.T. (1996) Time-to-event analyses of ecotoxicity data. *Ecotoxicology* **5**, 187–196.

Noppert, F., Leopold, A. and van der Hoeven, N. (1994) How to Measure No Effect: Towards a Measure of Chronic Toxicity. Unpublished workshop report, BKH Consulting Engineers, Delft, The Netherlands.

Pack, S. (1993) A Review of Statistical Data Analysis and Experimental Design in OECD Aquatic Toxicology Test Guidelines. Report to the Organisation for Economic Cooperation and Development, Paris.

Robertson, J.L., Smith, K.C., Savin, N.E. and Lavigne, J.L. (1984) Effects of dose selection and sample size of the precision of lethal dose estimates in dose-mortality regression. *J. Econ. Entomol.* **77**, 833–7.

Scheffé, H. (1959) *The Analysis of Variance.* New York: Wiley.

Skalski, J.R. (1981) Statistical inconsistencies in the use of no-observed-effect-levels in toxicity testing. In *Aquatic Toxicology and Hazard Assessment: Fourth Conference, ASTM STP 737* D.R. Branson and K.L. Dickson (eds), pp. 377–87. Philadelphia: American Society for Testing and Materials.

Sitter, R.R. (1992) Robust designs for binary data. *Biometrics* **48**, 1145–55.

Sokal, R.R. and Rohlf, F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research, 3rd edition.* New York: W.H. Freeman and Company.

Stephan, C.E. (1977) Methods of calculating LC50. In *Aquatic Toxicology and Hazard Evalution, ASTM STP 634* F.L. Mayer and J.L. Hamelink (eds), pp. 65–84. Philadelphia: American Society for Testing of Materials.

Stephan, C.E. and Rogers, J.W. (1985) Advantages of using regression to calculate results of chronic toxicity tests. In *Aquatic Toxicology and Hazard Assessment: Eighth Symposium, ASTM STP 891* R.C. Bahner and D.J. Hansen (eds), pp. 328–38. Philadelphia: American Society for Testing and Materials.

Suter II, G.W., Rosen, A.E., Linder, E. and Parkhurst, D.F. (1987) Endpoints for responses of fish to chronic toxic exposures. *Environ. Toxicol. Chem.* **6**, 793–809.

van der Hoeven, N. (1991) LC50 estimates and their confidence intervals derived for tests with only one concentration with partial effect. *Water Res.* **25**, 401–8.

Winer, B.J. (1971) *Statistical Principles in Experimental Design, 2nd edition.* New York: McGraw Hill.

Williams, D.A. (1986) Interval estimation of the median lethal dose. *Biometrics* **42**, 641–6.