9-2016

Regression or significance tests: What other choice is there?—An academic perspective

Michael C. Newman
Virginia Institute of Marine Science

Marcos Krull
Virginia Institute of Marine Science

In This Issue:

ET&C Perspectives

The Perspectives column is a regular series designed to discuss and evaluate potentially competing viewpoints and research findings on current environmental issues.

The Challenge: Statistical challenges in ecotoxicology

The statistical methodology required by newer test guidelines has been going through an evolution in recent years as newer methods become more accessible. This provides an opportunity for the statistical and scientific communities to reevaluate their approaches to the analysis of ecotoxicological data. It is important to replace methodology that is outdated while retaining what is valuable from existing approaches.

John W. Green
DuPont Applied Statistics Group
Newark, DE, USA

In Response: Challenges for statistical evaluation of ecotoxicological experiments—An industry perspective

There have been many calls within the ecotoxicological community to replace hypothesis testing methods to determine a no-observed-effect concentration (NOEC) with regression models to estimate an effects concentration (EC*x*) at which a specific percentage of effect, *x*, is expected to occur. Advances in statistical methodology and software have expanded the types of models that can be used. One such promising approach is generalized linear mixed models to capture the nested structure and overdispersion and to treat count data as such rather than through transforms to mimic normality-based models. Bayesian models have also been shown to provide good descriptions of the data from some types of responses.

It is critically important that the regression approach be evaluated carefully in each proposed application before abandoning the NOEC to avoid replacing what is widely perceived as a flawed approach by another approach that is also flawed and subject to abuse. Without intending to dismiss regression models for many common responses, Green et al. [1–3; J.W. Green et al., DuPont Applied Statistics Group, Newark, DE, USA, unpublished manuscript] provide examples of data that are problematic for the regression approach. Problems occur when the concentration–response shape is very shallow so that EC*x* estimates have great uncertainty, partly indicated by extremely wide confidence bounds. Point estimates of EC*x* in such cases are meaningless. Another problem exists when the control response is highly variable. If there is 20% standard error in the control mean, estimation of EC10 is absurd at face value. The fact that once a mathematical model has been fit, it is possible to estimate EC*x* for any positive value of *x* does not imply that all such estimates are meaningful or useful. Another problematic area is a response which has no pattern at all except at the highest 1 or 2 test concentrations. In such a situation, there might be no basis for proposing a model, yet the EC*x* estimate is highly model-dependent.

In evaluating fish early–life stage experiments for the revised Organisation for Economic Co-Operation and Development's test guideline 210, more than 100 studies were evaluated for size and mortality responses. For between 75% and 90% of the studies, good EC10 estimates for size and EC20 estimates for mortality could be obtained. "Good" in the Organisation for Economic Co-Operation and Development's evaluation of studies refers to estimates with tight confidence intervals and point estimates within a few percentages of the observed percentage effects in adjacent test concentrations based on regression models that agree well with the data over most of the concentration range, especially at the control, and do not exhibit significant lack of goodness-of-fit. For the remaining studies, either no model could be fit or the confidence interval for EC*x* spanned the entire range of tested concentrations including the control. In those cases, it was nonetheless possible to obtain a NOEC that appeared consistent with the data and corresponded to an observed effect of modest magnitude. A conclusion from the Organisation for Co-Operation and Development's investigation is that the NOEC approach must be retained as an alternative when regression fails to provide a useful result.

In proposing a modeling approach for a response from a specific type of study, it is important to do a computer simulation study based on a substantial, representative database of studies of the same type so as to capture the variability and concentration–response shapes likely to be encountered. Then it is possible to develop a distribution of EC*x* estimates that could be obtained so as to evaluate the viability of the proposed model. Such computer modeling is quite helpful in developing an understanding of what can be expected from experiments and

* Address correspondence to John.W.Green@dupont.com.

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/etc.3105

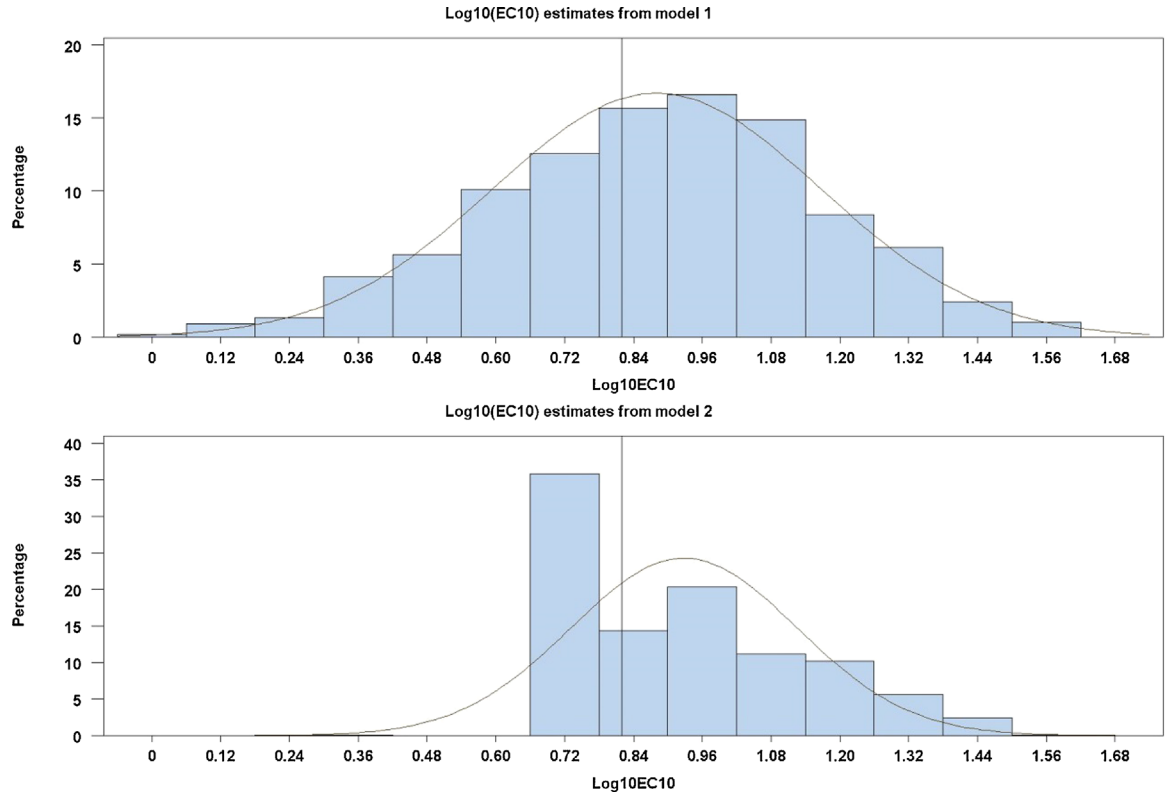

Figure 1. A comparison of 2 modeling approaches for snail reproduction.

for comparing alternative approaches. Figure 1 shows the distribution of log(EC10) estimates from 2 alternative models for snail reproduction. The true log(EC10) value is known in the simulation and shown as a vertical line. The distribution of model 1 estimates is symmetric about the true value and visually appealing; the distribution of model 2 estimates is skewed, but the model 2 estimates are much less variable. In particular, model 2 is much less likely to grossly underestimate EC10 and no more likely to overestimate. Such simulation studies must be set up with care to avoid unintentional bias, and all likely scenarios of shape and variability need to be modeled. It is this author's contention that many people have unrealistic confidence in the estimates from regression models that are not sufficiently grounded in real study experience.

One of the challenges in evaluating ecotoxicological studies is that at the present time there are responses for which no suitable regression models are available. One example is histopathological severity scores, which are required in 2 new test guidelines to be issued soon: the medaka multigeneration test, larval amphibian growth and development assay, as well as the fish short-term reproduction assay (Organisation of Economic Co-Operation and Development, test guideline 229)—and likely in the Japanese quail 2-generation toxicity test, which also should be issued in the near future. Severity scores are not numeric (other than as labels) and should not be treated as such. An analysis approach is given in Green et al. [3]. Currently, it is not clear whether an EC*x* can be defined for such scores that correspond in a meaningful way to EC*x* in other contexts.

Another challenge is time-to-event data. Survival analysis methods have been well developed for decades [4]. Recent advances allow modeling that captures the replicate nature of most ecotoxicity studies [5–8]. The challenge is that the range of event times is often very limited in ecotoxicity studies. Instead of a 2-yr rodent study or long-duration human clinical trial, one has a span of 5 to 6 event times over a 14-d or 21-d *Daphnia* study or short reproduction time in a Japanese quail 2-generation toxicity test or a short time span to reach Nieuwkoop and Faber stage 62 in a larval amphibian growth and development assay study. With such limited time spans, models from survival analysis can be uninformative, and new approaches are needed if NOECs are to be replaced.

John W. Green
DuPont Applied Statistics Group
Newark, DE, USA

REFERENCES

1. Green JW, Springer TA, Staveley JP. 2012. The drive to ban the NOEC/LOEC in favor of EC*x* is misguided and misinformed. *Integr Environ Assess Manag* 9:12–16.
2. Green JW. 2014. Statistical design and analysis of studies. In Brock B, Mounho B, Fu L, eds, *The Role of the Study Director in Nonclinical Studies: Pharmaceuticals, Chemicals, Medical Devices, and Pesticides*. Wiley, Hoboken, NJ, USA, pp 191–224.
3. Green JW, Springer TA, Saulnier AL, Swintek J. 2014. Statistical analysis of histopathology endpoints. *Environ Toxicol Chem* 33: 1108–1116.
4. Lee ET, Wang JW. 2013. *Statistical Methods for Survival Data Analysis*, 4th ed. Wiley, Hoboken, NJ, USA.
5. Lee EW, Wei LJ, Amato DA. 1992. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In Klein JP, Goel PK, eds, *Survival Analysis: State of the Art*. Kluwer, Boston, MA, USA, pp 237–248.
6. Lin DY. 1994. Cox regression analysis of multivariate failure time data: The marginal approach. *Stat Med* 13:2233–2247.

7. Ripatti S, Palmgren J. 2000. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 56:1016–1022.
8. Sargent DJ. 1998. A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* 54: 1486–1497.

DOI: 10.1002/etc.3106

In Response: Regression or significance tests: What other choice is there?—An academic perspective

Both the no-observed-effect concentration and its null hypothesis significance testing foundation have drawn steady criticism since their inceptions [1–5]. Many in our field reasonably advocate regression to avoid conventional null hypothesis significance testing shortcomings; however, regression is compromised under commonly encountered conditions (Green, present *Perspective*'*s Challenge*). As the debate to favor null hypothesis significance testing or regression methods continues into the 21st century, a sensible strategy might be to take a moment to ask, Are there now other choices? Our goal is to sketch out 1 such choice.

So many misconceptions have amassed around null hypothesis significance testing–based methods that minor refinements to current practices seem unlikely to resolve serious errors in their application. Five of the worst issues with null hypothesis significance testing practices are detailed in Newman [3] and references therein. First, null hypothesis significance testing without a priori definition of type II error rate (β) and effect size cannot be used to infer that a significant difference exists. Only falsification of the null hypothesis (H_0) is possible because no "significant difference" alternative hypothesis exists without β. Second, although seldom done, α and β should be set based on the seriousness of making each type of error. Third, a meaningful effect size should be applied instead of the usual no effect. Given enough observations, there will always be a difference between treatments, so the no-effect size lacks meaning. Fourth, the tendency to publish significant outcomes more readily than nonsignificant ones creates a literature bias that befuddles meta-analyses. Fifth, crucially, a pervasive misconception exists that null hypothesis significance testing p values estimate the probability of the H_0 being true and that 1 minus the p value approximates the probability of the alternative hypothesis being true. Actually, the p value is the probability of getting the data, or more extreme data, if H_0 is true. A p value alone is a misleading measure of H_0 or alternative hypothesis plausibility [3]. More is needed to estimate $p(H_0|x)$ from $p(x|H_0)$ because $p(H_0|x) = [p(H_0)\ p(x|H_0)]/p(x)$. To document this confusion about p values, the first author surveyed environmental professionals and students during 6 presentations in 4 countries. The correct definition of the p value was chosen from among 7 options by only 10% of responders ($n =$ 374, 95% confidence interval 7–13%). Random picking of an answer would have produced 14% correct answers. The inescapable conclusion is that a deep-rooted misunderstanding exists among environmental professionals who likely use p values routinely.

Regression and conventional null hypothesis significance testing might have been the sole practical alternatives when environmental regulations were formulated, but several readily implemented methods are now available, such as hypothesis testing after a priori power analysis with meaningful error rates and effect size [1–3], inference with confidence or credible intervals [6], Bayes factors [7–9], and information theory methods [10]. The Bayes factor will be discussed in the present *Perspective* as 1 possible alternative to consider when null hypothesis significance testing application fails.

Bayes factors aid decisions to favor 1 hypothesis over another, such as H_A:$\theta_A = 0$ or H_B:$\theta_B = 0.35$, given a data set (x). The Bayes factor, or $p(H_A|x)/p(H_B|x)$, can be estimated from $p(x|H_A)/p(x|H_B)$ if no information is available prior to testing, that is, $p(H_A)/p(H_B) = 1$. It is the data-based probability of 1 hypothesis divided by that of the other. In the simplest case of 2 explicit probability density functions (Figure 2), the Bayes factor might be estimated as the likelihood ratio for the hypotheses given the data, $L(H_A)/L(H_B)$ [7].

Such simple situations are common only in textbooks. The Bayes factor approach was unfeasible when the null hypothesis significance testing versus regression debate began because its estimation often required difficult integrations:

$$BF = \frac{p(x|H_A)}{p(x|H_B)} = \frac{\int p(\theta_A|H_A)p(x|\theta_A, H_A)d\theta_A}{\int p(\theta_B|H_B)p(x|\theta_B, H_B)d\theta_B}$$

Now computer-intensive Markov chain Monte Carlo algorithms are widely available for this purpose, making the Bayes factor an attractive substitute for conventional null hypothesis significance testing methods.

If the estimate from the best-supported hypothesis is made the denominator, a minimum Bayes factor is produced that quantifies the degree of data-based support for that hypothesis relative to the alternative hypothesis [7,8]. Jeffreys [8] used minimum Bayes factors to categorize the amount of support for the best hypothesis: unhelpful (0.31 < minimum Bayes factor < 1), substantial (0.10 < minimum Bayes factor < 0.31), strong (0.031 < minimum Bayes factor < 0.10), very strong (0.01 < minimum Bayes factor < 0.031), or decisive (minimum Bayes factor ≤0.01) evidence. If the hypothesis in the numerator was H_0, a minimum Bayes factor ≤0.01 would result in its rejection. Such unencumbered interpretations of the minimum Bayes factor seem preferable to working around the pervasive misinterpretations of p values that consistently overestimate the evidence against the null hypothesis.

An unpublished study of activated carbon addition to sediments illustrates this point (M.C. Newman, unpublished data). Sediment had sorbent added or not added to create

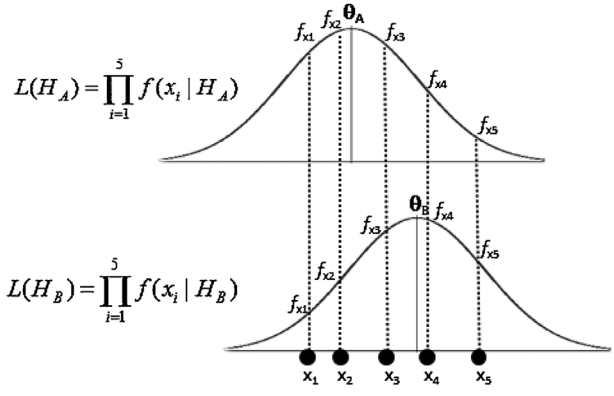

Figure 2. Bayes factor estimation with products of probability densities (f_{xi}) for 5 observations (•) using hypothetical distribution A or B: H_A and H_B have hypothesized means (θ_A and θ_B), with perhaps θ_A being a mean difference of 0.

Table 1. Comparison of null hypothesis significance testing and minimum Bayes factor assessments of activated carbon sorbents in sediment

Duration	L	*P value*	Minimum Bayes factor
5	–39	0.0061	0.1536
45	–53	<0.0001	0.0015
90	–47	0.0021	0.0367
180	–41	0.0003	0.0097

L = disk weight changes as a percentage.

2 treatments. A leaf disk and amphipod (*Hyalella azteca*) were added to 30 wells per sediment treatment, and the leaf disk weight loss was quantified after 10 d. Trials were run using sediment 5 d, 45 d, 90 d, and 180 d postamendment. The percentage of difference in detrital processing (disk wt change, L) in the amended sediments relative to those of the unamended sediment treatment were all negative and are shown in Table 1.

All H_0 of no difference in detrital processing were rejected based on Welch *t* test *p* values ($\alpha = \beta = 0.05$; effect size = 35%). Applying the minimum Bayes factor, the evidence against the H_0 relative to the H_A was judged to be substantial for the first duration and strong to decisive for later durations. Each *p* value was lower than its corresponding minimum Bayes factor and, if the conventional null hypothesis significance testing misinterpretation was applied, would have substantially overestimated support for H_A relative to that for H_0.

To answer the question posed in *The Challenge* for this *Perspectives* article, Bayes factors are now 1 alternative to conventional null hypothesis significance tests when regression fails. The common objections about Bayes subjectivity are irrelevant to Bayes factors because they are calculated only from the evidence [7]. Although the Bayes factor is not without its flaws, its wider use would foster movement away from the currently muddled application of null hypothesis significance testing.

Michael C. Newman
Marcos Krull
Virginia Institute of Marine Science
College of William & Mary
Gloucester Point, VA, USA

REFERENCES

1. Stephan CE, Rogers JW. 1985. Advantages of using regression analysis to calculate results of chronic toxicity tests. In Bahne RC, Hansen DJ, eds, *Aquatic Toxicology and Hazard Assessment: Eighth Symposium*. ASTM STP 891. ASTM International, Philadelphia, PA, USA, pp 328–338.
2. Newman MC. 2008. "What exactly are you inferring?" A closer look at hypothesis testing. *Environ Toxicol Chem* 27:1013–1019.
3. Newman MC. 2013. *Quantitative Ecotoxicology*, 2nd ed. Taylor & Francis/CRC, Boca Raton, FL, USA.
4. Gigerenzer G. 2004. Mindless statistics. *Journal of Socio-Economics* 33:587–606.
5. Nuzzo R. 2014. Statistical errors. *Nature* 506:150–152.
6. Cumming G. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, New York, NY, USA.
7. Goodman SM. 1999. Toward evidence-based medical statistics. 2. The Bayes factor. *Ann Intern Med* 130:1005–1013.
8. Jeffreys H. 1983. *Theory of Probability*, 3rd ed. Oxford University Press, Oxford, UK.
9. Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc* 90: 773–795.
10. Anderson DR. 2008. *Model Based Inference in the Life Sciences*. Springer, New York, NY, USA.

DOI: 10.1002/etc.3107

In Response: Biological arguments for selecting effect sizes in ecotoxicological testing—A governmental perspective

Criticisms of the uses of the no-observed-effect concentration (NOEC) and the lowest-observed-effect concentration (LOEC) and more generally the entire null hypothesis statistical testing scheme are hardly new or unique to the field of ecotoxicology [1–4]. Among the criticisms of NOECs and LOECs is that statistically similar LOECs (in terms of *p* value) can represent drastically different levels of effect. For instance, my colleagues and I found that a battery of chronic toxicity tests with different species and endpoints yielded LOECs with minimum detectable differences ranging from 3% to 48% reductions from controls [5].

For interpretations of field studies, recommendations for improved practices include using confidence intervals rather than hypothesis testing for group comparisons and evaluating whether apparent effects exceed predetermined "critical" effect sizes [2,6]. For interpretations of toxicity tests, recommendations for improved practices emphasize replacing NOEC and LOEC comparisons with either model-based true no-effect concentration estimates or curve fitting and from the fitted curve functions, reporting concentrations that produced *x*% of effects (EC*x*) [7–9]. These developments beg the question: What levels of toxic effect are of concern or can be considered negligible? Biologically based arguments for selecting *x* are scarce, and instead discussions for selecting *x* from curve-fitting approaches have emphasized statistical or test performance considerations for selecting *x* rather than biological implications. Confidence limits, variability of point estimates, model dependence, and comparisons of NOECs to EC*x* percentages have been evaluated [10,11]. These statistical considerations are of value but are not sufficient by themselves and may be circular. If a major shortcoming of NOECs is that they may actually correspond with fairly high adverse effects [9,12], why should it make sense to then turn about and ask what levels of effects are typically associated with NOECs to define the *x* in EC*x* values?

In contrast, the reasons for the existence of toxicity testing practices relate to making some estimate of safe or unsafe concentrations for aquatic populations or communities. Thus, judgments of what constitutes a negligible level of effect in toxicity tests should consider the consequences of similar effect levels in the wild. For example, the primary adverse effects studied in an early–life stage toxicity test with fish are reduced growth and survival. From a population biology perspective, survival and reproduction are the only endpoints that directly matter for viability. Reduced growth may indirectly influence reproduction by slowing the time to reproduction or because smaller females may produce fewer offspring. However, in the wild, subtle differences in size could have disproportionately large effects on survival. For fish, growth is closely linked to survival, in part because of the importance of size in competitive interactions and predator–prey relations. For

Figure 3. Subtle differences in size can lead to life or death outcomes. Predatory sculpin posed no risk to same-length juvenile salmon, but those with a 15% disadvantage in length were readily ambushed and eaten by the sculpin [13]. Photo by Jo Opdyke Wilhelm, King County Department of Natural Resources (Washington, USA). Used with permission.

example, adult sculpin (Cottidae) may prey on juvenile salmonids in streams and vice versa, depending on relative sizes. Torrent sculpin, *Cottus rhotheus* (Figure 3), that only had a 15% length advantage could readily ambush, subdue, and eat coho salmon, *Oncorhynchus kisutch*, yet when the sculpin and salmon were similar in length, the sculpin posed no threat [13]. With fish in temperate streams, contests for territory may determine profitable feeding locations, shelter from predation, and winter resting shelters. These in turn may determine whether fish have sufficient energy reserves for overwinter survival and eventual reproduction [14]. A size disparity of as little as 5% in body weight may tip the outcome of such contests [15]. Riverine survival of juvenile Chinook salmon (*Oncorhynchus tshawytscha*) in Idaho, USA, was disproportionately size-dependent, with a 10% difference in length associated with 33% to 70% reductions in migratory survival [16].

Similarly, with aquatic invertebrates, the same EC*x* values for different effect endpoints cannot be assumed to have the same level of effect. For instance, a 10% reduction in length of mussels would predict approximately a 19% to 44% reduction in fecundity, based on length–fecundity regressions from field studies with different freshwater mussel species [17]. In 28-d exposures with freshwater mussels and Cu, the maximum reductions in length in treatments in which at least some mussels survived to the end of the tests were only 13% to 29% [17]. Uncritical reliance on a single, fixed EC*x* value such as the EC20 for a growth endpoint for which the maximum range of response may not even reach 20% would lead to reporting test results as "greater than" values, which would incorrectly discount biologically important effects as being insensitive.

Relating mortality rates of aquatic organisms in toxicity tests to corresponding effects in real populations from contaminant-induced mortality is difficult. Early–life stage mortality does not directly translate to proportional reductions in recruitment, largely because the survival of juvenile cohorts in nature is often density-dependent. That is, when densities are high, food and space become limiting, growth is stunted, and survival is low. When densities are low, food and space are abundant, growth is higher, and survival is higher [18]. In natural populations, different life stages have very different contributions to the population dynamics and, for example, the same percentage of mortality to young-of-year fish or sexually mature fish will have very different population implications. For instance, a bull trout, *Salvelinus confluentus*, population could withstand up to a 60% annual decrease in juvenile survival yet went into decline following a 5% increase in annual adult mortalities [19].

A closely related problem is what level *x* effect is most appropriate for use with the EC*x* when the policy goal is to allow no adverse effects. An oxymoronic response has often been to equate some low level of adverse effect to no effect. The EC10 has been used in this manner [7,9,10], although in a more extreme interpretation, a 25% toxic response from controls was considered to represent nontoxicity [20]. Although this may be logical if the policy goals are intended to allow some toxicity, it seems to me that if the goal is no toxicity, the only value for *x* in EC*x* that represents no toxicity is 0 (i.e., EC0).

Although calculating an EC0 estimate from distribution- or regression-based curve-fitting models is an impossibility when using a distribution with infinite tails such as the normal (gaussian) distribution, an EC0 can easily be estimated from curve-fitting routines that use finite distributions [21]. The triangular distribution can produce a threshold sigmoidal toxicity curve fit similar in shape to that from nonlinear logistic regression, and the rectangular distribution can produce a "broken stick" piecewise-linear fit. Whereas the logistic equation curve subtly angles downward from its start, illustrating the impossibility of an EC0, the threshold sigmoidal and piecewise-linear curves are flat until the no-effect thresholds (EC0) are reached (Figure 4). The piecewise-linear fit has the advantage of making visual explanation of the no-effect EC0 to laypeople or policy people because the break in the regression is the no-effect threshold, and its reasonableness can be interpreted relative to the underlying data points. These visual explanations may be easier than trying to explain that something (e.g., 10% effect) equates to nothing. If an objective of ecotoxicological testing and modeling is to estimate thresholds for the absence of effects, ecotoxicologists should not discount the EC0.

Growth data that have shallow slopes and limited ranges of response are less than ideal for nonlinear regression models, and I tend to place little confidence in confidence limits. For instance, in the piecewise-linear EC0 for mussel growth in Figure 4, the EC0 estimate appears eminently reasonable to me, breaking at a treatment with nearly identical responses as the controls, yet the calculated confidence limits encompass both the next lower (control) and higher treatments. Rather, my confidence in the EC*x* estimates is based on how well the models fit the underlying data. Model selection can matter, especially when interpolation is needed because effects occurred at the lowest concentration tested. In the mussel shell length example (Figure 4A), the EC10 estimate from logistic regression was lower than the EC0 estimate from piecewise-linear regression.

Finally, replicated exposure test designs such as those in Figure 4B are a legacy of null hypothesis significance testing and are inefficient for curve fitting and EC*x* point estimates. A gradient of 24 unreplicated exposures would be more capable of defining the thresholds and distributions of concentration responses than could 6 exposures replicated 4 times each. Acknowledging that proportional diluters or numbers of pump channels place practical limits on numbers of exposures, the point is that moving from null hypothesis significance testing to an EC*x* interpretation approach involves more of a mind-set change than simply also running the results of a null hypothesis significance testing–based test design through regression-fitting software.

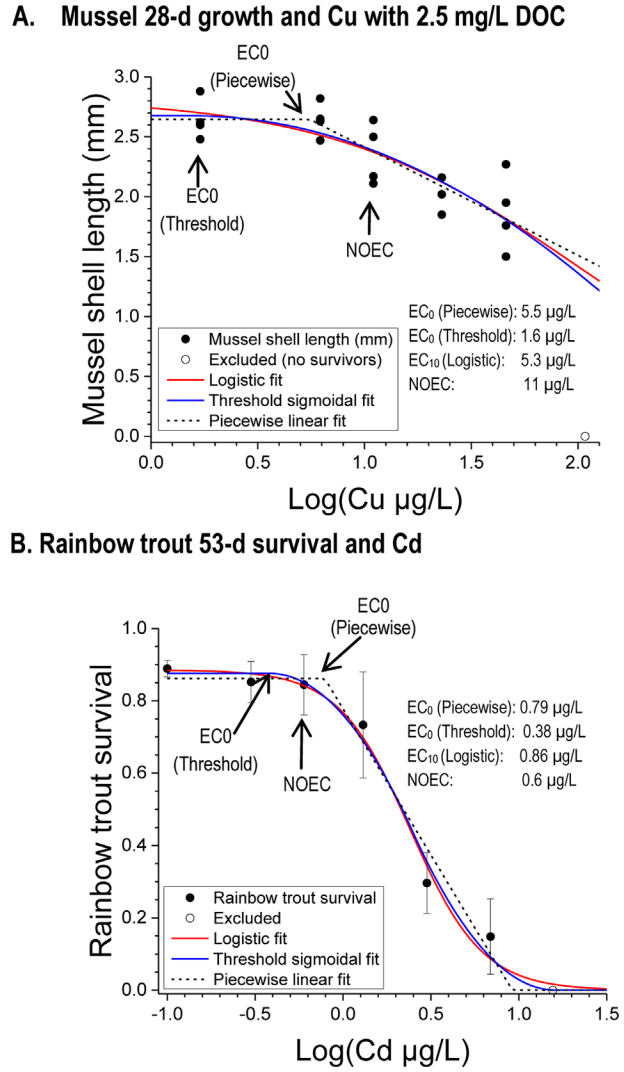

Figure 4. Two examples of estimating no-effect threshold values (EC0) from chronic tests with nonlinear regression using finite uniform and triangular distributions ("piecewise-linear" and "threshold sigmoidal" fits, respectively). Fits with the logistic equation using the normal distribution with its infinite tails are also shown. Although the EC0 estimates may appear reasonable relative to the underlying values, the examples illustrate the inefficiency of common replicated experimental designs for regression-based point estimates. Data from Mebane et al. [5] and Wang et al. [17], using models from Erickson [21]. NOEC = no-observed-effect concentration.

The main point of this *Response* is that a given EC*x* effect size percentage might have very different biological implications depending on the endpoint and ecological context. For instance, a 5% reduction in an endpoint with low inherent variability such as length-at-age could have comparable population-level implications to a 20% reduction in more variable and ecologically compensable endpoints such as early–life stage survival or fecundity. An underlying theme is that we should not just reduce species and biology to numbers and models and disconnect the "eco" from "toxicology." Considering the biological implications of toxicity testing in the context of species life histories and their ecological context is important, even though such considerations may be qualitative, uncertain, or even speculative.

Christopher A. Mebane
US Geological Survey,
Boise, ID

REFERENCES

1. Fidler F, Cumming G, Burgman M, Thomason N. 2004. Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics* 33:615–630.
2. Martínez-Abraín A. 2007. Are there any differences? A non-sensical question in ecology. *Acta Oecol* 32:203–206.
3. Berkson J. 1942. Tests of significance considered as evidence. *J Am Stat Assoc* 37:325–335.
4. Cohen J. 1994. The earth is round ($p < .05$). *Am Psychol* 49:997–1003.
5. Mebane CA, Hennessy DP, Dillon FS. 2008. Developing acute-to-chronic toxicity ratios for lead, cadmium, and zinc using rainbow trout, a mayfly, and a midge. *Water Air Soil Pollut* 188:41–66.
6. Munkittrick KR, Arens CJ, Lowell RB, Kaminski GP. 2009. A review of potential methods for determining critical effect size for designing environmental monitoring programs. *Environ Toxicol Chem* 28: 1361–1371.
7. van der Hoeven N, Noppert F, Leopold A. 1997. How to measure no effect. Part I: Towards a new measure of chronic toxicity in ecotoxicology. Introduction and workshop results. *Environmetrics* 8:241–248.
8. Green JW, Springer TA, Staveley JP. 2012. The drive to ban the NOEC/LOEC in favor of EC*x* is misguided and misinformed. *Integr Environ Assess Manag* 9:12–16.
9. Fox DR. 2008. NECs, NOECs and the EC*x*. *Australasian Journal of Ecotoxicology* 14:7–9.
10. Isnard P, Flammarion P, Roman G, Babut M, Bastien P, Bintein S, Essermeant L, Ferard J, Gallotti-Schmitt S, Saouter E, Saroli M, Thiebaud H, Tomassone R, Vindimian E. 2001. Statistical analysis of regulatory ecotoxicity tests. *Chemosphere* 45:659–669.
11. Moore DRJ, Caux P-Y. 1997. Estimating low toxic effects. *Environ Toxicol Chem* 16:794–801.
12. Crane M, Newman MC. 2000. What level of effect is a no observed effect? *Environ Toxicol Chem* 19:516–519.
13. Patten BG. 1977. Body size and learned avoidance as factors affecting predation on coho salmon, *Oncorhynchus kisutch*, fry by torrent sculpin, *Cottus rhotheus*. *Fish Bull* 75:457–459.
14. Cunjak RA, Power G. 1987. The feeding and energetics of stream-resident trout in winter. *J Fish Biol* 31:493–511.
15. Abbott JC, Dunbrack RL, Orr CD. 1985. The interaction of size and experience in dominance relationships of juvenile steelhead trout (*Salmo gairdneri*). *Behaviour* 92:241–253.
16. Mebane CA, Arthaud DL. 2010. Extrapolating growth reductions in fish to changes in population extinction risks: Copper and Chinook salmon. *Hum Ecol Risk Asses* 16:1026–1065.
17. Wang N, Mebane CA, Kunz JL, Ingersoll CG, Brumbaugh WG, Santore RC, Gorsuch JW, Arnold WR. 2011. Influence of DOC on toxicity of copper to a unionid mussel (*Villosa iris*) and a cladoceran (*Ceriodaphnia dubia*) in acute and chronic water exposures. *Environ Toxicol Chem* 30:2115–2125.
18. Scheffer M, Baveco JM, DeAngelis DL, Lammens EHRR, Shuter BJ. 1995. Stunted growth and stepwise die-off in animal cohorts. *Am Nat* 145:376–388.
19. Post JR, Mushens C, Paul AJ, Sullivan M. 2003. Assessment of alternative harvest regulations for sustaining recreational fisheries: Model development and application to bull trout. *N Am J Fish Manag* 23:22–34.
20. US Environmental Protection Agency. 2010. National pollutant discharge elimination system: Test of significant toxicity implementation document. EPA 833-R-10-003. Washington, DC.
21. Erickson RJ. 2013. *Toxicity Response Analysis Program*, Ver. 1.22. US Environmental Protection Agency, National Health and Environmental Research Laboratory, Mid-Continent Ecological Division, Duluth, MN.

DOI: 10.1002/etc.3108

In Response: Some species sensitivity distribution statistics revisited—A governmental perspective

Estimating toxicity thresholds for aquatic, sedimentary, and terrestrial environments present insurmountable challenges,

given the complexity of the systems involved. Since Kooijman [1] and Van Straalen and Denneman [2], the statistical species sensitivity distribution (SSD) method has provided a major shortcut in deriving hazardous concentrations (HC*x*, e.g., HC5) from a set of comparable laboratory toxicity endpoints. A 2002 volume on SSDs [3] documents overviews, applications, and extensions.

More recently, an European Centre for Ecotoxicology and Toxicology of Chemicals workshop on estimating toxicity thresholds for SSDs was held in Amsterdam [4], where current SSD–related issues were discussed and summarized. Identified research areas include ecological validation, species selection, statistical problems, and mode of action of the chemicals. Among these, statistical challenges are manifold as well.

In the present *Response*, it is shown that some statistical issues can be definitely addressed with present-day statistical methods. Three statistical topics are selected. The problem of data error is discussed firstly. Secondly, I will revisit predictive SSD fitting and extrapolation constants. Thirdly, the problem of censored data will be touched on.

Data error

My coworkers and I have been concerned with the problem of how to account for data points within a SSD, some of which may derive from quantitative structure–activity relationship (QSAR) estimates [5]. But as QSAR-based estimates bring in additional uncertainty, we investigated the effect of data error. But data error is not typical for model-injected data points: it is characteristic for experimental species data as well. The European Chemicals Agency's technical guidance says that SSD data points may be best conceived as species *means*, so it is advisable to average multiple sensitivity data points for the same species.

This carries over to QSAR-based points as expected model predictions. We briefly explored a Bayesian hierarchical model with 2 variance components: within-species error and between-species error [5]. Classical analysis of variance teaches us that high within-species error may mask the assessment of the SSD error of species means. If the SSD is mere species noise, the purported distribution of species means vanishes and becomes undetectable.

In a numerical experiment for $n = 5$, we found that the within-species noise—up to one-half of the SSD between-species mean variation—did not have much effect on the estimate of the SSD standard deviation (SD). Neglecting data error in fact leads to a conservative estimate because the SSD SD decreases slightly for increasing data noise. So, we tentatively concluded that neglecting moderate data noise may often be acceptable.

Predictive extrapolation constants

In the nonhierarchical Bayesian treatment [6], we have used the so-called noninformative prior, assuming the logarithm of the SSD sigma to be uniform. This is standard Bayesian practice. However, over the years, we have found this prior on sigma to be too optimistic (small). Why would the prior express that bigger SSD SDs are to be discounted?

There is another argument: for the hierarchical model, the noninformative prior for the SSD SD does not work, however small the data noise [5]. Hence, it seems unreasonable for an assumed value to be exactly 0.

Consequently, we decided to calculate revised extrapolation constants for a uniform prior. We call this "cautionary extrapolation," which we think amounts to more realistic extrapolation, especially for small sample sizes.

In the Bayesian hierarchical model, we also looked at the so-called predictive distribution of the SSD. This can be understood as the vertical average of SSD spaghetti plot threads. This holds for both the SSD density curves as well as the cumulative distribution plots.

The predictive distribution is single-threaded (1 curve) and, hence, a nice summary of the spaghetti plot and associated percentile curves. In our 2013 study [5], we compared lower quantiles (HC*x*) of the predictive distribution to former extrapolation values found in our 2000 study [6].

Statistical theory reveals that the predictive distribution of a normal SSD is a Student t distribution with $n - 1$ degree of freedom. Cautionary extrapolation makes this $n - 2$, so the more realistic prior on the SSD SD causes the loss of 1 more degree of freedom.

It turned out that predictive HC*x* values are more sensitive to low sample size than the former median estimates, doing more justice to small sample sizes.

In Table 2, the classical HC5 estimates [6] are compared with the predictive HC5 (all based on the noninformative prior for the SSD SD) and the cautionary predictive estimate for the uniform prior SSD SD [5].

Note that at the European Chemicals Agency–recommended sample sizes (Table 2, rows 10 and 15, printed in bold), the extrapolation constants for both priors converge to roughly –2. For sample size over 30, differences between median and both predictive estimates gradually disappear.

Censored SSD data

The statistics of censored data modeling are well developed, see Helsel [7]. We have applied this to SSD fitting of 14 acute antimony freshwater data [8]. Note that 4 species data "points" have a right-censored value (Figure 5). The same analysis applies to left-censored values (e.g., detection limits), mixtures of both, as well as interval data.

Table 2. Classical lower, median, and upper 5% hazardous concentration extrapolation constants (after Aldenberg and Jaworska [6]) compared with predictive and cautionary predictive 5% hazardous concentration extrapolation constants (from Aldenberg and Rorije [5])[a]

n	Lower	Median	Upper	Noninformative predictive	Cautionary predictive
6	−3.71	−1.75	−0.87	−2.18	−2.57
8	−3.19	−1.72	−0.96	−2.01	−2.23
10	**−2.91**	**−1.70**	**−1.02**	**−1.92**	**−2.07**
15	**−2.57**	**−1.68**	**−1.11**	**−1.82**	**−1.90**
30	−2.22	−1.66	−1.25	−1.73	−1.76
100	−1.93	−1.65	−1.41	−1.67	−1.68

[a]At the European Chemicals Agency–recommended sample sizes (bold), the extrapolation constants for both priors converge to roughly –2.

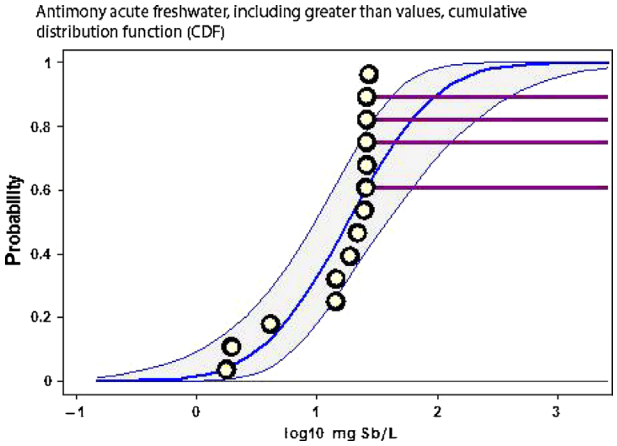

Figure 5. Species sensitivity distribution fit for antimony (acute, freshwater), including 4 right-censored values.

When the 4 "greater thans" are removed, the median HC5 equals 1.7 mg Sb/L. When included, the median HC5 is equal to 2.0 mg Sb/L. Note the asymmetry of the normal SSD percentile curves.

A similar approach to censored values using Bayesian concepts is followed by Kon Kam King et al. [9] with the R-based web-tool MOSAIC_SSD. Figure 5 shows the lower bound of the censored values. It does not convey how the method of maximum likelihood takes censoring into account. For that, Kon Kam King et al. [9] and Helsel [7] supply the details.

We hope to have shown that the SSD method, although incapable of modeling the full complexity of environmental systems, is by no means "done" yet and is amenable to extensions addressing a range of statistical challenges.

Tom Aldenberg
National Institute for Public Health and the Environment
Bilthoven, The Netherlands

REFERENCES

1. Kooijman SALM. 1987. A safety factor for LC_{50} values allowing for differences in sensitivity among species. *Water Res* 21:269–276.
2. VanStraalen NM, Denneman CAJ. 1989. Ecotoxicological evaluation of soil quality criteria. *Ecotoxicol Environ Saf* 18:241–251.
3. Posthuma L, Suter GW II, Traas TP, eds. 2002. *Species Sensitivity Distributions in Ecotoxicology*. Lewis, Boca Raton, FL, USA.
4. European Centre for Ecotoxicology and Toxicology of Chemicals. 2014. *Estimating Toxicity Thresholds for Aquatic Ecological Communities from Sensitivity Distributions, 11–13 February 2014, Amsterdam.* Brussels, Belgium.
5. Aldenberg T, Rorije E. 2013. Species sensitivity distribution estimation from uncertain (QSAR-based) effects data. *Altern Lab Anim* 41:19–31.
6. Aldenberg T, Jaworska JS. 2000. Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicol Environ Saf* 46:1–18.
7. Helsel DR. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. Wiley-Interscience, Hoboken, NJ, USA.
8. Van Leeuwen LC, Aldenberg T. 2012. Environmental risk limits for antimony. RIVM Letter Report 601357001/2012. Bilthoven, The Netherlands.
9. Kon Kam King G, Veber P, Charles S, Delignette-Muller ML. 2014. MOSAIC_SSD: A new web tool for species sensitivity distribution to include censored data by maximum likelihood. *Environ Toxicol Chem* 33:2133–2139.

DOI: 10.1002/etc.3109

In Response: Challenges when weighing evidence about environmental risks—An industry perspective

Asking the right scientific questions, designing appropriate tests of hypotheses based on these questions, and applying the correct quantitative techniques to the resulting data are, of course, important goals in environmental toxicology and chemistry [1]. However, 1 overarching challenge that has been largely neglected to date is how to use all available information both efficiently and fairly to come to decisions about the environmental risks and benefits of chemical substances.

Reduction of uncertainty through the use of multiple lines of evidence within an overall weight-of-evidence approach is encouraged by regulatory organizations such as the European Chemicals Agency [2]. Unfortunately, practical guidance on how to perform such an analysis is currently rather thin, probably because some definitions of "weight of evidence" are vague and ambiguous [3]. Recently, Suter and Cormier [4] proposed a useful distinction between "genuine weighing of commensurable pieces of evidence" and "interrelating heterogeneous evidence," which they call "building a case." Distinguishing between these 2 activities may help to remove ambiguity and vagueness from the process.

Environmental risk assessment of a chemical usually involves genuine weighing of commensurable pieces of evidence. This evidence may comprise information on the concentration of the substance in different environmental compartments or its toxicity to different biological taxa. Evidence is presented as statistical summaries such as time-weighted average concentrations, 90th percentile environmental concentrations, no-observed-effect concentrations, 5% hazardous concentrations, or 50% lethal concentrations. However, it remains a challenge to use these summaries in a way that does not exclude large amounts of useful data by only comparing (reasonable) worst-case exposure concentrations with a single "most sensitive" or "representative" toxicity value. It is also rare to find full use of information on both the disadvantages and the benefits of a substance in an assessment of risk in a way that really does weigh up all of the available evidence, both "for" and "against" the substance.

Some may argue that "expert judgment" is sufficient and that all substances should be addressed on a case-by-case basis. Unfortunately, case-by-case assessment is a recipe for inconsistency and inefficiency. Inconsistency arises because different individuals bring their own views on how to weight the value of different pieces of data. Thus, they are likely to come to different conclusions on the basis of the same data. Inefficiency arises because substance registrants and evaluators need to develop or understand a new weighting system for each substance, even when many similar substances have been assessed previously. The reality is that most so-called weighting systems are likely to be implicit and value-laden if developed case by case. In other words, the way in which data have been weighted is often informal and not described clearly, and the faith that is placed in each data value is associated with one's role in the process (i.e., industry, governmental, academic, nongovernmental organizations, and consultancy representatives will bring different value judgments and personal biases to bear on their interpretation of the same data). This then becomes a foundation for lengthy and intractable arguments between vested interests.

It would be helpful to develop a more explicit and formal approach to weight-of-evidence assessment, which goes beyond

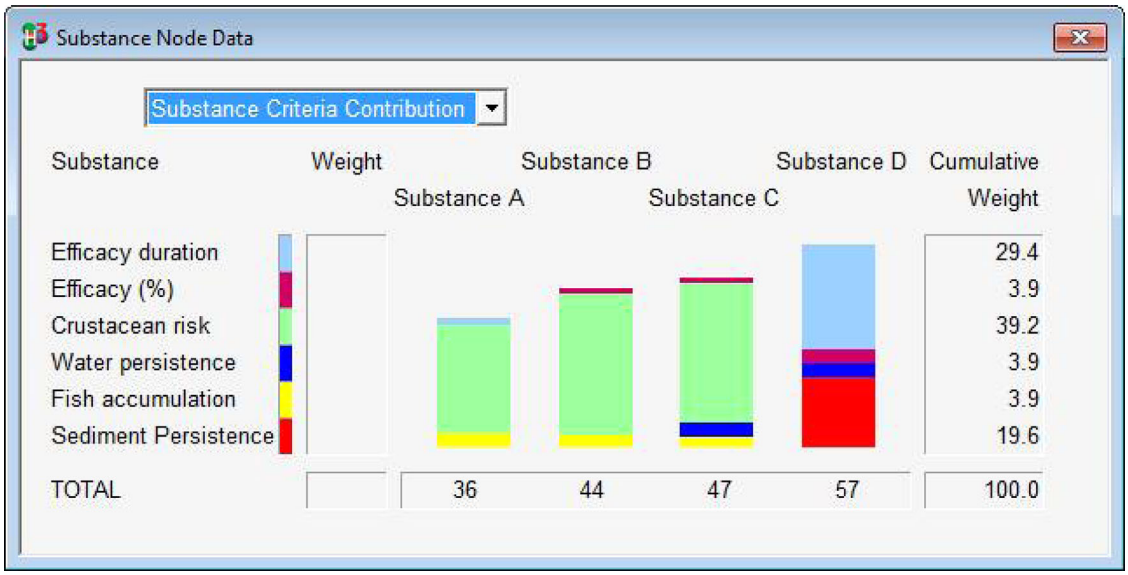

Figure 6. Multicriteria decision analysis output from HiView3 software.

most currently available regulatory guidance. This can draw on useful work by several authors [4–8]. Multicriteria decision analysis modeling seems currently to be the most promising approach when performing the type of weight-of-evidence assessment of all available data required in environmental risk assessment. Multicriteria decision analysis has the following features: 1) it allows comparisons of dissimilar favorable and unfavorable effects to be compared, with each measured quantity converted to a preference value on a 0 to 100 scale; 2) conversion for each effect can be accomplished by a linear or nonlinear translation, called a "value function," which is an assessment of the relevance of various levels of the measured quantity; 3) the units for the preference value scales are equated through a process known as "swing weighting," which requires judgments of the relevance of scale differences, enabling weighted effects to be summed to give an overall risk–benefit balance.

Figure 6 shows a simple multicriteria decision analysis output from HiView 3 software [9] in which 4 hypothetical chemical substances are compared across both "favorable" and "unfavorable" criteria to determine which of them scores highest. This type of multicriteria decision analysis can incorporate both quantitative statistical data and semiquantitative or qualitative information, and it makes explicit the value judgments (i.e., the swing weights) that are used to weight the different types of information. The graphical outputs from an analysis like this are easy to understand, and sensitivity analyses are included in the assessment software so that the effect of subjective differences between data evaluators can be fully explored.

As stated by the European Medicines Agency, the European regulatory agency responsible for the authorization of human and veterinary medicines [10],

> A key feature of decision modelling is its ability to distinguish facts from value judgements, and to combine both features into an overall assessment. Because value judgements are necessarily subjective, it is important to make them explicit and subject to discussion, debate and peer review. Even if agreement cannot be reached, a quantitative model can be used to explore whether or not disagreements matter to the final result. If they do, then further information or data might be required, or [there may need to be] further exploration of the reasons for the disagreement [that has] surfaced.

The challenge for chemical risk assessors is to develop decision-modeling approaches such as multicriteria decision analysis, which are simple enough for wide use yet sufficiently sophisticated to capture the most important criteria when assessing environmental risks. Rising to this challenge would lead to more holistic and consistent decision making and substantial savings in time, money, and other scarce resources.

Mark Crane
AG-HERA
Faringdon, Oxfordshire, United Kingdom

REFERENCES

1. Crane M, Chapman PF. 1996. Asking the right questions: Ecotoxicology and statistics. *Ecotoxicology* 5:137–138.
2. European Chemicals Agency. 2011. Evaluation of available information. In *Guidance on Information Requirements and Chemical Safety Assessment*. Helsinki, Finland.
3. European Chemicals Agency. 2010. *How to Report Weight of Evidence*. Practical guide 2. Helsinki, Finland.
4. Suter GW II, Cormier SM. 2011. Why and how to combine evidence in environmental assessments: Weighing evidence and building cases. *Sci Total Environ* 409:1406–1417.
5. Jaworska J, Gabbert S, Aldenberg T. 2010. Towards optimisation of chemical testing under REACH: A Bayesian network approach to integrated testing strategies. *Regul Toxicol Pharmacol* 57:157–167.
6. Linkov I, Loney D, Cormier S, Satterstrom FK, Bridges T. 2009. Weight-of-evidence evaluation in environmental assessment: Review of qualitative and quantitative approaches. *Sci Total Environ* 407: 5199–5205.
7. Linkov I, Welle P, Loney D, Tkachuk A, Canis L, Kim JB, Bridges T. 2011. Use of multicriteria decision analysis to support weight of evidence evaluation. *Risk Anal* 31:1211–1225.
8. Smith EP, Lipkonch I, Ye K. 2002. Weight-of-evidence (WOE): Quantitative estimation of probability of impairment for individual and multiple lines of evidence. *Hum Ecol Risk Assess* 8:1585–1596.
9. Catalyze. 2015. Hiview 3 Software. Hursley, Winchester, UK. [cited 2014 November 1]. Available from: http://www.catalyze.co.uk/index.php/software/hiview3/
10. European Medicines Agency. 2011. Benefit–risk methodology project. Work package 3 report: Field tests. EMA/718294/2011. London, UK.

DOI: 10.1002/etc.3110

In Response: Benefits of the ARRIVE guidelines for improving scientific reporting in ecotoxicology—An academic perspective

The challenge of adequate reporting in journal articles

As the number of life science journals grows year by year, there are concerns that in some areas of biomedicine and toxicology the reporting of research is often inadequate and is therefore of limited value for reliably informing scientific practice, regulatory decision making, or policy needs. As observed by Kilkenny and colleagues [1], this has important implications for animal research, one of the most controversial aspects of the life sciences. For example, a review conducted by the United Kingdom's National Centre for the Replacement, Refinement and Reduction of Animals in Research (an independent scientific organization) concluded that only 59% of the 271 randomly chosen published articles assessed stated the study objectives and the number and characteristics of the animals used (i.e., species, strain, sex, and age or weight). Most of the papers surveyed also did not report using randomization (87%) or blinding (86%) to reduce bias in animal selection and outcome assessment [2]. Kilkenny et al. [2] also reported that only 70% of the publications that used statistical methods fully described them and presented the results with a measure of precision or variability. As noted by Kilkenny et al. [1], these findings

	ITEM	RECOMMENDATION
Title	1	Provide as accurate and concise a description of the content of the article as possible.
Abstract	2	Provide an accurate summary of the background, research objectives, including details of the species or strain of animal used, key methods, principal findings and conclusions of the study.
INTRODUCTION		
Background	3	a. Include sufficient scientific background (including relevant references to previous work) to understand the motivation and context for the study, and explain the experimental approach and rationale. b. Explain how and why the animal species and model being used can address the scientific objectives and, where appropriate, the study's relevance to human biology.
Objectives	4	Clearly describe the primary and any secondary objectives of the study, or specific hypotheses being tested.
METHODS		
Ethical statement	5	Indicate the nature of the ethical review permissions, relevant licences (e.g. Animal [Scientific Procedures] Act 1986), and national or institutional guidelines for the care and use of animals, that cover the research.
Study design	6	For each experiment, give brief details of the study design including: a. The number of experimental and control groups. b. Any steps taken to minimise the effects of subjective bias when allocating animals to treatment (e.g. randomisation procedure) and when assessing results (e.g. if done, describe who was blinded and when). c. The experimental unit (e.g. a single animal, group or cage of animals). A time-line diagram or flow chart can be useful to illustrate how complex study designs were carried out.
Experimental procedures	7	For each experiment and each experimental group, including controls, provide precise details of all procedures carried out. For example: a. How (e.g. drug formulation and dose, site and route of administration, anaesthesia and analgesia used [including monitoring], surgical procedure, method of euthanasia). Provide details of any specialist equipment used, including supplier(s). b. When (e.g. time of day). c. Where (e.g. home cage, laboratory, water maze). d. Why (e.g. rationale for choice of specific anaesthetic, route of administration, drug dose used).
Experimental animals	8	a. Provide details of the animals used, including species, strain, sex, developmental stage (e.g. mean or median age plus age range) and weight (e.g. mean or median weight plus weight range). b. Provide further relevant information such as the source of animals, international strain nomenclature, genetic modification status (e.g. knock-out or transgenic), genotype, health/immune status, drug or test naïve, previous procedures, etc.

Figure 7. The Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines reporting checklist for in vivo studies in ecotoxicology [1].

Housing and husbandry	9	Provide details of: a. Housing (type of facility e.g. specific pathogen free [SPF]; type of cage or housing; bedding material; number of cage companions; tank shape and material etc. for fish). b. Husbandry conditions (e.g. breeding programme, light/dark cycle, temperature, quality of water etc for fish, type of food, access to food and water, environmental enrichment). c. Welfare-related assessments and interventions that were carried out prior to, during, or after the experiment.
Sample size	10	a. Specify the total number of animals used in each experiment, and the number of animals in each experimental group. b. Explain how the number of animals was arrived at. Provide details of any sample size calculation used. c. Indicate the number of independent replications of each experiment, if relevant.
Allocating animals to experimental groups	11	a. Give full details of how animals were allocated to experimental groups, including randomisation or matching if done. b. Describe the order in which the animals in the different experimental groups were treated and assessed.
Experimental outcomes	12	Clearly define the primary and secondary experimental outcomes assessed (e.g. cell death, molecular markers, behavioural changes).
Statistical methods	13	a. Provide details of the statistical methods used for each analysis. b. Specify the unit of analysis for each dataset (e.g. single animal, group of animals, single neuron). c. Describe any methods used to assess whether the data met the assumptions of the statistical approach.
RESULTS		
Baseline data	14	For each experimental group, report relevant characteristics and health status of animals (e.g. weight, microbiological status, and drug or test naïve) prior to treatment or testing (this information can often be tabulated).
Numbers analysed	15	a. Report the number of animals in each group included in each analysis. Report absolute numbers (e.g. 10/20, not 50%[2]). b. If any animals or data were not included in the analysis, explain why.
Outcomes and estimation	16	Report the results for each analysis carried out, with a measure of precision (e.g. standard error or confidence interval).
Adverse events	17	a. Give details of all important adverse events in each experimental group. b. Describe any modifications to the experimental protocols made to reduce adverse events.
DISCUSSION		
Interpretation/ scientific implications	18	a. Interpret the results, taking into account the study objectives and hypotheses, current theory and other relevant studies in the literature. b. Comment on the study limitations including any potential sources of bias, any limitations of the animal model, and the imprecision associated with the results[2]. c. Describe any implications of your experimental methods or findings for the replacement, refinement or reduction (the 3Rs) of the use of animals in research.
Generalisability/ translation	19	Comment on whether, and how, the findings of this study are likely to translate to other species or systems, including any relevance to human biology.
Funding	20	List all funding sources (including grant number) and the role of the funder(s) in the study.

Figure 7. Continued.

are a cause for scientific concern and are consistent with observations for several research areas published in recent years. The same challenge faces ecotoxicology, especially with respect to experimental studies using a range of aquatic and terrestrial species. For example, it is known that different strains (clones) of the model test organism *Daphnia magna* may show variation in response to toxicants [3]. Other areas of relevance to ecotoxicology also relate to the use of fish and other nonmammalian species in comparative pharmacology and physiology. For instance, the zebrafish *Danio rerio* shows evidence of interstrain differences in responses to ethanol (a still widely used solvent in Organisation for Economic Co-Operation and Development test guidelines for ecotoxicology) [4,5]. These examples of interspecies variability emphasize the importance of reporting such information, considering it can have a significant bearing on experimental outcomes. It is for this reason that in Europe it is now a legal requirement that laboratories seeking to work with zebrafish must use defined, purpose-bred strains from established cultures [6,7].

Improving the reporting of animal experiments—The ARRIVE guidelines

No ecotoxicology journals currently provide specific guidance on the reporting of research involving the use of animals. Kilkenny et al. [2] reported back in 2009 that 4% of the 271

articles evaluated did not report the numbers of animals used anywhere in the methods or results sections. Clearly, such a serious omission has negative implications for other researchers and potentially for decision makers where the conclusions may have had toxicological or biomedical implications. The importance of adequate reporting in toxicology and ecotoxicology has also been emphasized by Klimisch et al. [8] and Küster et al. [9]. Against this background, Kilkenny et al. [1] have proposed a set of highly valuable guidelines termed Animals in Research: Reporting In Vivo Experiments (ARRIVE). The ARRIVE guidelines focus on a helpful checklist of 20 items that describe the minimum information for all scientific publications involving the use of animals (Figure 7). These guidelines have been endorsed by 449 journals across various scientific disciplines. Although the ARRIVE guidelines have been described with applications to mammalian species closely in mind, in our view the same scientific principles apply equally to laboratory studies using birds, fish, and other nonmammalian species.

To enable uptake across the ecotoxicology community, the nuances particular to this area of research and testing will need to be encompassed within any guidelines that are ultimately developed and applied to such studies. For example, relevant to items 3 and 19, researchers would report on the study's relevance to wildlife populations, rather than to humans; relevant to item 7, the reporting of where studies are carried out will include detail on whether they are conducted within a laboratory or outdoor mesocosm; and regarding baseline data, there would be benefits of reporting characteristics such as fecundity and hatching success, as well as historical control and solvent data where relevant (see Hutchinson et al. [10], relevant to item 14).

The adoption of guidelines such as these could have far-reaching implications; for example, there are increasing requirements for open literature ecotoxicological data to be included in submission dossiers for regulatory chemical safety assessment purposes—in this context, better-quality published studies could contribute to a waiving of additional in vivo studies. In conclusion, it is hoped that the 20 items recommended in the ARRIVE guidelines will serve to improve the reporting of all types of ecotoxicological study and that they will be considered by the community when establishing the most appropriate standards applicable to this type of research.

Thomas H. Hutchinson
School of Biological Sciences, University of Plymouth
Plymouth, United Kingdom

Natalie Burden
National Centre for the Replacement, Refinement and Reduction of Animals in Research
London, United Kingdom

REFERENCES

1. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol* 8(6):e1000412.
2. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 4:e7824.
3. Baird DJ, Barber I, Calow P. 1990. Clonal variation in general responses of Daphnia magna Straus to toxic stress. I. Chronic life-history effects. *Funct Ecol* 4:399–407.
4. Loucks E, Carvan MJ. 2004. Strain-dependent effects of developmental ethanol exposure in zebrafish. *Neurotoxicol Teratol* 26:745–755.
5. de Esch C, van der Linde H, Slieker R, Willemsen R, Wolterbeek A, Woutersen R, De Groot D. 2012. Locomotor activity assay in zebrafish larvae: Influence of age, strain and ethanol. *Neurotoxicol Teratol* 34:425–433.
6. European Communities. 2010. Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. *Official Journal of the European Union* L 276/33. [cited 2015 December 15]. Available from: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:276:0033:0079:EN:PDF
7. UK Home Office 2014. *Code of Practice for the Housing and Care of Animals Bred, Supplied or Used for Scientific Purposes*. Presented to Parliament pursuant to Section 21 (5) of the Animals (Scientific Procedures) Act 1986. [cited 2015 December 15]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/388535/CoPanimalsWeb.pdf
8. Klimisch H-J, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol* 25:1–5.
9. Küster A, Bachmann J, Brandt U, Ebert I, Hickmann S, Klein-Goedicke J, Maack G, Schmitz S, Thumm E, Rechenberg B. 2009. Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. *Regul Toxicol Pharmacol* 55:276–280.
10. Hutchinson TH, Shillabeer N, Winter MJ, Pickford DB. 2006. Acute and chronic effects of carrier solvents in aquatic organisms: A critical review. *Aquat Toxicol* 76:69–92.

DOI: 10.1002/etc.3111