Statistical Analysis of Avian Reproduction Studies

John W Green (✉ John@JohnWGreen-Ecostats.com)  
John W Green Ecostatistical Consulting LLC  https://orcid.org/0000-0002-1524-6340

Manousos Foudoulakis  
Corteva Agriscience UK

Timothy Fredricks  
Bayer Crop Science United States

Tiffany Carro  
FMC Corporation

Jonathan Maul  
Syngenta Crop Protection AG

Stephanie Plautz  
BASF Corp

Pablo Valverde  
Corteva Inc Indianapolis

Adam Schapaugh  
Bayer CropScience AG

Xiaoyi Sopko  
Corteva Inc

Zhenglei Gao  
Bayer AG

Discussion

**Keywords:** Avian reproduction, hazard identification, benchmark dose, distribution, model average, diagnostics, NOAEL, historical controls, hormesis, maximum safe dose

**Posted Date:** November 30th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-1054353/v1

**License:** ☛ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

**Version of Record:** A version of this preprint was published at Environmental Sciences Europe on March 24th, 2022. See the published version at https://doi.org/10.1186/s12302-022-00603-5.
Abstract

Avian reproduction studies for regulatory risk assessment are undergoing review by regulatory authorities, often leading to requests for statistical re-analysis of older studies using newer methods, sometimes with older study data that do not support these newer methods. We propose detailed statistical protocols with updated statistical methodology for use with both new and older studies and recommend improvements in experimental study design to set-up future studies for robust statistical analyses. There is increased regulatory and industry attention to the potential use of benchmark dose (BMD) methodology to derive the point of departure in avian reproduction studies, to be used as the endpoint in regulatory risk assessment. We present benefits and limitations of this BMD approach for older studies being re-evaluated and for new studies designed for with BMD analysis anticipated. Model averaging is recommended as preferable to model selection for BMD analysis. Even for a new study following the modified experimental design analyses with BMD methodology will only be possible for a restricted set of response variables. The judicious use of historical control data, identification of outlier data points, increased use of distributions more consistent with the nature of the data collected as opposed to forcing normality-based methods, and trend-based hypothesis tests are shown to be effective for many studies, but limitations on their applicability are also recognized and explained. Updated statistical methodologies are illustrated with case studies conducted under existing regulatory guidelines that have been submitted for product registrations. Through the adoption of improved avian reproduction study design elements combined with the suggested revised statistical methodologies the conduct, analyses, and utility of avian reproduction studies for avian risk assessments can be improved.

1. Introduction

Avian reproduction studies for regulatory risk assessment are done under Organization for Economic Cooperation and Development (OECD) Test Guideline 206 or United States Environmental Protection Agency (USEPA) Guideline OCSPP 850.2300. Both guidelines were issued when risk assessment was based on hypothesis testing to derive a No Observed Effects Concentration (NOEC). Statistical guidance in these guidelines is minimally defined. Over time, both the USEPA and the European Food Safety Authority (EFSA) have issued guidance documents to supplement these original guidelines. Recent guidance (EFSA 2009a, EFSA 2009b, EFSA 2017) has promoted the use of regression or benchmark dose methods to derive estimates of effects concentrations, usually 10 and 20% effects concentration referred to as EC10 and EC20 or BMD10 and BMD20. When these statistical methods are followed, risk assessments can be based on the indicated estimate or on a lower confidence bound of that estimate.

Studies done under the indicated guidelines typically have only three test concentrations plus a negative control. Regression models that can be fit to such data are severely restricted by the small number of treatment groups (tested concentrations). Another complicating factor is that as many as 8 response variables are measured (and 15 are calculated) in such experiments. These responses include incidence data, such as survival, count data such as number of eggs laid, eggs hatched or eggs cracked, and continuous data such as hatchling body weight, eggshell thickness, and body weight gain (for adults and
hatchlings). The variance of these response variables varies widely, so that for some responses, very small changes might be statistically significant while for others, relatively large effects are not unusual and it may be difficult to distinguish real effects from statistical “noise”. The statistical distributions to describe these different types of responses vary and call for careful selection of statistical methods and models. Erratic concentration-response patterns add to the challenges to statistical interpretation. It should also be acknowledged that there is limited scientific basis to guide the risk assessor as to the size effect that should be estimated or for which to test. A rare exception to this is a conclusion that only an 18% decrease (EFSA 2009b) or 22% (Lincer 1975) in eggshell thickness is biologically important in terms of population level concerns. As a result, arbitrary decisions have been made, such as requiring an estimate of a 10% effects concentration or simply basing a risk assessment on whether the response in some treatment group is statistically significantly different from the control independent of whether the observed difference has biological relevance or population implications. In the absence of a scientific basis for the size effect of concern, historical control data can and should be used to help distinguish between real effects and mere statistical artifacts.

1.1. Objectives

The objective of this study is to indicate ways to improve the analysis and endpoint selection of avian reproduction studies. This is done partly through improved statistical analysis, the use of historical control data and the biological interpretation of the findings. Particular attention is given to regression or benchmark dose methodology where the experimental design should be modified and the relative merits of point estimates (BMD10) and lower confidence bounds (BMD10LB) and the size effect that can be estimated reliably are discussed.

1.2. Experimental Design

The current test guideline was designed for NOEC determination and requires at least three test concentrations plus a negative control. The spacing of test concentrations is geometric with a common ratio of 6, with the highest test concentration approximately one-half of the LC10 determined by a prior dietary study (OECD TG 205), in case this study is available and delivers an LC10, but not to exceed 1000ppm. This makes the test concentrations equally spaced on a logarithmic scale. There should be at least twelve replicates, each consisting of two or three birds with sex and number depending on the species tested.

For NOEC determination, the power of the statistical test for each response is a function of the variance of that response, the replication, and the specific test used. A crude but useful approximation to the power can be obtained using the minimum detectible difference, MDD, often expressed as the minimum detectible percent change from control, MDD%, defined in equation (eq. 1).

\[
MDD\% = CV \times T \sqrt{\frac{1}{n_0} + \frac{1}{n_1}},
\]  
(eq. 1)
where CV is the coefficient of variation in the control, T is the critical value of a T-distribution with appropriate degrees of freedom, \( n_0 \) and \( n_1 \) are the number of replicates in the control and each treatment group, respectively. For the sample sizes typical in an avian reproduction test, T is approximately 2. As demonstrated in Staveley et al (2018), the minimum size effect (MSE%) that can be estimated by one of the regression models recommended in this study is a multiple of MDD% as indicated by the following equation.

\[
MSE\% = MDD\% \sqrt{\frac{h_z}{\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}}
\]  

(eq. 2)

where \( h_z \) is the leverage associated with a given treatment group. The cited reference presented evidence for non-target terrestrial plant studies (NTTP) done under OECD test guidelines 208 and 227 based on many such studies using the same statistical models as recommended here. In the NTTP study, the MSE% varied between 0.46 and 1.99 times MDD%. For avian studies, only a very few studies have been done with 4 or 5 test concentrations and determining the multiplier of MDD% in equation (eq. 2) that is applicable to avian studies is more challenging. Instead, a simulation studies was done that indicates the properties of EC10 or BMD10 estimates and their lower confidence bounds. These simulation results are based on data simulated to have the observed levels of CV found in the avian database and 5 treatment groups.

For avian studies, MDD% varied from 1-3% for percent eggs not cracked per eggs laid to 7-23% for eggs per hen, depending on the testing lab and species. A short summary table for selected endpoints is given in Table 1. It will be clear from Table 1 that for a few responses, it will be possible to estimate BMD05, but for most it will not be possible even to estimate BMD10. The table only shows the point estimate. The properties of the lower confidence bound will be more problematic. Based on these results, there is considerable challenge to the idea of replacing the NOEC by the BMD10.
Calculations assume 18 cages of 2 birds each in every treatment group and within-study CVs from historical control data from two frequently used testing labs.

As equations 1 and 2 indicate, increased replication will reduce the size effect that can be detected or estimated, but improvements from practical design changes will be modest. Even doubling the replication would reduce MDD% by only 30%. Analysis of the historical control databases suggest it might be possible that MDD% and MSE% could be reduced somewhat through refined laboratory techniques but not enough to detect 10% effects for all responses. For BMD estimation, increasing the number of test concentrations to 4, or preferably 5, would improve the ability to provide statistically sound BMD estimates for key responses through fitting models that better capture the shape of the concentration-response curve, but again practical designed will not permit BMD10 estimates for all responses. With 18 cages of 2 birds per cage and 4 treatment groups (control + 3 test concentrations), the current design requires 144 birds. An experimental design with 5 test concentrations plus a control, 12 cages per treatment, and 2 birds per cage would require the same number of birds, provide greater ability to calculate a BMD than possible under current designs and would reduce the power to determine a NOEC by only 12.5% when regression modelling fails for one or more responses. The proposed experimental design is estimated to increase the cost of a study by 11-12%. This experimental design was used in the simulation study described next.

### 1.3. Simulation study to explore ECx/BMDx estimation for avian studies

Table 1. MDD% for Mallard and Quail

|        | Mallard | Quail |
|--------|---------|-------|
|        | p50 | p75 | p90 | p50 | p75 | p90 |
| Response | Eggs laid per hen | Eggs not cracked / number eggs laid (%) | Live embryos / number eggs set | Live embryos / number viable embryos | Number hatched / number eggs set (%) | Number hatched / number live embryos |
| 14 | 19 | 28 | 19 | 22 | 30 | 1 | 2 | 3 |
| 6 | 13 | 19 | 6 | 12 | 20 | 1 | 3 | 7 |
| 14 | 17 | 20 | 10 | 15 | 20 | 8 | 12 | 16 |
| 8 | 17 | 20 | 11 | 15 | 21 | 14 | 17 | 20 |
| 1 | 1 | 2 | 3 | 5 | 8 | 21 | 23 | 28 |
| 5 | 5 | 5 | 4 | 5 | 5 | 24 | 27 | 30 |
| 3 | 4 | 5 | 4 | 5 | 5 | 75 | 161 | 233 |
| 7 | 7 | 9 | 6 | 7 | 7 | 65 | 81 | 123 |
| 7 | 7 | 9 | 28 | 33 | 94 | 36 | 41 | 46 |
| 20 | 23 | 28 | 23 | 27 | 31 | 20 | 23 | 28 |

p×=nth quantile of distribution of MDD% for indicated responses, n=50, 75, 90
The database of available avian studies is not large enough to follow the same procedure as used in Staveley et al (2018). Instead, data were simulated to follow one of three general shapes with a range of simulated CV (5 to 40) based on Table 1 for each shape. The simulations were set up for a continuous response. Previous simulation studies done by the lead author, some of which are given in Green et al (2018), suggest GNLMMs for conditionally binomial or Poisson responses will have comparable point estimates. Negative lower confidence bounds on such estimates, if calculated by exact methods rather than approximated using normality-based approximations, are not possible for GNLMM, but will tend to be extremely close to zero where a simulated continuous curve will give negative lower bounds. The shapes are characterized by the maximum effect simulated at the highest tested concentration and a shape parameter labeled ECPB, which varies from 1 to 10 in the simulations. ECPB=1 defines a concentration-response that decreases immediately from the control, while ECPB=5 corresponds to a moderately delayed decrease or shallower response curve, and ECPB=10 corresponds to a more delayed decrease or shallower response curve. All three shapes are observed in avian studies. Shallow concentration-response relationships are not uncommon in avian studies and that can make BMD estimates unattainable regardless of what the MDD% and MSE% figures indicate. As stated above, any percent change from control can be estimated once a regression model is fit. What equation 2 provides is the size effect for which a reliable estimate can be expected. A reliable estimate is one from a model that meets a set of criteria for model stability and for which the confidence interval is not overly wide. For the simulation study, the only criteria imposed were that a model result was used only if the model fitting algorithm converged and produced a lower 95% confidence bound greater than zero. All point estimates and confidence bounds reported are model average results, not from individual models. Further detail is provided in the supplementary material.

Tables 2 and 3 summarize EC10 point estimates and 95% lower confidence bounds. Table 2 indicates that if there is a 20% or greater observed effect at the highest tested concentration and the CV is 10 or less, then the EC10 point estimate is usually a reliable indicator of the size effect in the population being simulated. When there is only a 10% effect observed in the highest tested concentration and CV>5, the quality of the EC10 point estimate is seriously degraded. Table 3 shows that the lower confidence bounds is much more variable. For 8 of the 15 simulated conditions for shape ECPB=1, over 50% of the EC10 estimates have negative lower confidence bounds, making ECXLB of little or no value for risk assessment. The shape parameters ECPB=5 and 10 show worse results for risk assessment. Those results are given in the supplement. The simulation study shows that the BMD approach will be useful for avian studies only for limited responses unless an experimental design with greatly increased numbers of birds is used.
Table 2. Distribution of EC10 point estimates for shape ECPB=1

| TRUEC10 | Mean_ECX | Med_ECX | p10 | p25 | p75 | p90 | YMAX | CV  | %Fit | P90/P10 |
|---------|----------|---------|-----|-----|-----|-----|------|-----|------|---------|
| 27      | 27       | 27      | 27  | 27  | 27  | 27  | 27   | 27  | 1    | 1.4     |
| 27      | 26       | 25      | 25  | 24  | 23  | 23  | 22   | 21  | 1    | 1.8     |
| 27      | 25       | 24      | 24  | 23  | 22  | 22  | 21   | 20  | 1    | 2.9     |
| 27      | 24       | 24      | 23  | 23  | 23  | 22  | 22   | 21  | 1    | 4.1     |
| 27      | 25       | 21      | 21  | 21  | 17  | 17  | 12   | 13  | 1    | 6.6     |
| 27      | 25       | 21      | 21  | 21  | 17  | 17  | 12   | 13  | 1    | 8.7     |
| 27      | 34       | 34      | 34  | 34  | 34  | 34  | 34   | 34  | 1    | 100.1   |
| 38      | 37       | 37      | 37  | 37  | 37  | 37  | 37   | 37  | 1    | 1.4     |
| 38      | 36       | 35      | 35  | 35  | 35  | 35  | 35   | 35  | 1    | 1.8     |
| 38      | 34       | 33      | 33  | 33  | 33  | 33  | 33   | 33  | 1    | 2.2     |
| 38      | 36       | 30      | 30  | 30  | 25  | 25  | 25   | 25  | 1    | 4.0     |
| 38      | 38       | 30      | 30  | 30  | 25  | 25  | 25   | 25  | 1    | 9.0     |
| 57      | 55       | 55      | 55  | 55  | 55  | 55  | 55   | 55  | 1    | 1.4     |
| 57      | 55       | 55      | 55  | 55  | 55  | 55  | 55   | 55  | 1    | 1.4     |
| 57      | 53       | 53      | 53  | 53  | 53  | 53  | 53   | 53  | 1    | 1.8     |
| 57      | 53       | 53      | 53  | 53  | 53  | 53  | 53   | 53  | 1    | 1.8     |
| 100     | 100      | 100     | 100 | 100 | 100 | 100 | 100  | 100 | 1    | 1.5     |
| 100     | 150      | 150     | 150 | 150 | 150 | 150 | 150  | 150 | 1    | 3.3     |
| 100     | 246      | 246     | 246 | 246 | 246 | 246 | 246  | 246 | 1    | 11.3    |
| 100     | 162      | 162     | 162 | 162 | 162 | 162 | 162  | 162 | 1    | 28.4    |
| 100     | 139      | 139     | 139 | 139 | 139 | 139 | 139  | 139 | 1    | 540.5   |

TRUEC10 = value of EC10 for distribution simulated

Mean_ECX = mean EC10 estimate from simulated data

Med_EC10 = median EC10 estimate from simulated data

pN = Nth percentile of EC10 estimates from simulated data, N = 10, 25, 75, 90

YMAX = maximum % of control mean response at highest concentration in simulated data

CV = coefficient of variation simulated

%Fit = % of simulated datasets for which at least one model converged and had a positive lower confidence bound for EC10 estimate

P90/P10 = ratio of the indicated percentiles of the distribution of EC10 estimates. Larger values indicate more spread in the point estimates.

ECPB = shape parameter controlling the shape of the simulated curve
Table 3. Distribution of EC10 lower bound (ECXLB) estimates for shape ECPB=1

| mean_ECXL B | Med_EC10 | p10 | p25 | p75 | p90 | YMAX | CV | ECPB |
|-------------|----------|-----|-----|-----|-----|------|----|------|
| 20          | 20       | 17  | 18  | 22  | 24  | 60   | 5  | 1    |
| 15          | 15       | 9   | 12  | 18  | 21  | 60   | 10 | 1    |
| 7           | 7        | -1  | 2   | 11  | 16  | 60   | 20 | 1    |
| 0           | 0        | -9  | -4  | 5   | 10  | 60   | 30 | 1    |
| -7          | -4       | -21 | -9  | 1   | 5   | 60   | 40 | 1    |
| 29          | 29       | 23  | 26  | 32  | 35  | 70   | 5  | 1    |
| 21          | 21       | 11  | 17  | 26  | 31  | 70   | 10 | 1    |
| 7           | 9        | -6  | -1  | 14  | 22  | 70   | 20 | 1    |
| -33         | -3       | -22 | -11 | 5   | 10  | 70   | 30 | 1    |
| -29         | -9       | -47 | -18 | -2  | 3   | 70   | 40 | 1    |
| 42          | 42       | 34  | 38  | 47  | 50  | 80   | 5  | 1    |
| 29          | 30       | 11  | 22  | 37  | 45  | 80   | 10 | 1    |
| -29         | 6        | -29 | -11 | 15  | 23  | 80   | 20 | 1    |
| -73         | -10      | -94 | -31 | 0   | 8   | 80   | 30 | 1    |
| -115        | -16      | -132| -46 | -5  | 0   | 80   | 40 | 1    |
| 54          | 68       | 36  | 58  | 77  | 83  | 90   | 5  | 1    |
| -87         | 28       | -155| -10 | 45  | 58  | 90   | 10 | 1    |
| -285        | -29      | -574| -129| -3  | 14  | 90   | 20 | 1    |
| -269        | -41      | -664| -123| -11 | 0   | 90   | 30 | 1    |
| -418        | -40      | -786| -120| -10 | 0   | 90   | 40 | 1    |

Mean_ECXL B = mean EC10LB estimate from simulated data
Med_EC10 = median EC10LB estimate from simulated data
pN = Nth percentile of EC10LB estimates from simulated data, N=10, 25, 75, 90
YMAX = maximum % of control mean response at highest concentration in simulated data
CV = coefficient of variation simulated
ECPB = shape parameter controlling the shape of the simulated curve

A negative mean or median ECXLB or another percentile indicates EC10 estimates of little value for risk assessment. When ECXLB<0 then the EC10 estimate is statistically indistinguishable from zero.

1.4. Brief Summary of ways to improve avian reproduction hazard identification

- Careful test selection, diagnostics, attention to outliers, alternative distributions (GLMM) can provide improved NOEC determination.
- Historical control data can be very helpful in distinguishing between real effects and spurious statistical significance.
Better models, including alternative distributions (generalized nonlinear mixed models or GNLMM) can improve ECx estimation when regression modelling is possible.

MAXSD can provide a substitute for difficult to fit regression or improve rationale for NOEC approach.

Model averaging takes model uncertainty into account and reduces effect of poorly fitting models.

2. Methodology

A illustrative set of avian reproduction studies have been compiled by the Terrestrial Vertebrates ad hoc Team (TVahT) of the European Crop Protection Association. Endpoints from each study have been analysed statistically using both standard and novel statistical methods and models. This effort has contributed to the development of detailed statistical protocols that cover the range of endpoints in such studies. These protocols will be presented following summaries of selected case studies that illustrate the concerns that commonly arise and that helped motivate the protocols. Detailed analyses of these and other case studies are presented in the Supplementary material.

The importance of historical control data as an aid in distinguishing between real effects and statistical artifacts has been described above. Historical control data (HCD) for avian reproduction studies done under OECD TG 206 or OCSP 850.2300 has been made available from Eurofins (years 1976-2016 for quail and 1978-2016 for mallard) and by Smithers-Vincent (Years 2001-2020 for quail, 2004-2019 for mallard). The HCD consists of a single mean value for each recorded response and, in most cases, a within-study standard deviation for the response. In some instances, an endpoint of interest is a simple algebraic function of reported values. In those cases, no within-study standard deviation was available.

Case studies done by one of these labs were evaluated in part using the historical control data from the same lab. For use in evaluating an avian reproduction study done under the indicated guidelines, European Commission (2013) recommended a 5-year period centered on the data of the current study with a minimum of 20 studies during that period.

The main use of historical control data (HCD) recommended for risk assessment is to demonstrate whether statistically significant trends or changes from control are the result of unusual concomitant controls or mild trends lying entirely within the historical control data range or the result of true effects of the test substance that push treatments means beyond the range of historical control data.

Following the case study summary, statistical protocols are presented as charts with related discussion. Detailed descriptions of all tests and models are presented in the Supplementary material

2.1. Case Studies

2.1.1. Case Study 1.
Case study 1 illustrates how historical control data can help identify an extreme concurrent control result that can yield misleading statistical significance. It also illustrates the importance of outlier identification. Table 4 shows the mean response (as a percent) in all treatment groups with all observations included. The mean response, quail cracked eggs per eggs laid, is notably higher in all treatment groups than in the control. Also relevant is that the treatment-response is not monotone.

Table 4. Case Study 1. Quail Percent cracked eggs per eggs laid (PCL)

| Group | Conc | Count | Mean | Median | Std |
|-------|------|-------|------|--------|-----|
| 1     | 0    | 16    | 0.63 | 0.00   | 1.45|
| 2     | 25   | 16    | 3.13 | 2.00   | 3.88|
| 3     | 50   | 16    | 2.25 | 0.00   | 3.34|
| 4     | 100  | 16    | 5.31 | 2.00   | 9.21|

Count=number of replicates (breeding pairs)

Mean, median, and standard deviation are simple unweighted summary statistics.

Conc=ppm

Figure 1 shows the study data in relation to the historical control data from the same lab. The 95% confidence bounds on the HCD are given by $(LB, UB)=(1,5)$, so the mean response at Conc=100 ppm $>$ UB. The Tukey outlier test (Tukey 1977) found 1 outlier in the 100 ppm treatment group. With that observation omitted, the mean response in that group was 3.27% which is well below UB. The control mean, 0.62, is 38% below the lower 95% CB of the HCD.

Statistically significance was assessed in several ways. A standard way to analyze such percentage data is to apply Dunnett’s test to arc-sine square-root transformed proportions (Fleiss et al 2003). A scientifically preferable analysis would be to analyze the number of cracked eggs as binomially distributed conditioned on the number of eggs laid, followed by Dunnett’s test in a generalized linear mixed model (GLMM). There was evidence of overdispersion in the study data, but all attempts to capture it failed. The Jonckheere-Terpstra (JT) trend-based test would also be appropriate except for the non-monotonicity observed in both the complete and outlier omitted data. The increases found by either Dunnett test at 25 ppm and 100 ppm were a function of the low control, not real effects. The mean responses in the low and intermediate treatment groups were within the historical control range and the only reason the high treatment mean response was above the historical control range was due to a single hen where 36% (8 of 22) of eggs laid were cracked. Only three hens in the study laid fewer eggs than were laid by this hen. The biological importance of this single observation or the resulting high proportion of cracked eggs is not clear. It should be noted that a regulatory reviewer requested another avian reproduction study on the test substance based on the significant increase found in the low treatment group. The historical control database suggests that this was not warranted and a new study supported this conclusion.

2.1.2. Case Study 2.
Case study 2 illustrates how a clear trend in the treatment response requires care to distinguish statistical from biological significance. The test and model selection present challenges as well. The response analyzed was 14-day hatchling survivors/eggs set (H14DS_ES). Summary data are given in Table 5, where clear downward trends in the mean and in the median are evident.

| Group | Conc | Count | Mean  | Median | Std  |
|-------|------|-------|-------|--------|------|
| 1     | 0    | 15    | 76.35 | 80.77  | 11.63|
| 2     | 4    | 14    | 70.18 | 71.40  | 11.66|
| 3     | 10   | 13    | 60.69 | 67.92  | 22.92|
| 4     | 35   | 15    | 55.23 | 56.25  | 24.50|

Count=number of replicates (breeding pairs)

Mean, median, and standard deviation are simple unweighted summary statistics.

Conc=ppm

No normalizing, variance stabilizing transform was found. The non-parametric JT test found all treatments significantly less than control. As Figure 2 shows, only the mean response at 35 ppm is outside the HCD range and that at 4 ppm is in the middle of that range. The Tukey outlier test identified 2 low outliers (1 each at 10 & 35 ppm). With those omitted, even the mean response at 35 ppm is within the HCD range. Setting the NOEC at 10, where there was a 20% decrease in the mean (16% decrease in the median) is justifiable in terms of the HCD. In terms of biological significance, setting the NOEC at 4 where there was only an 8% decrease is justified. No good regression model was found, as is common in studies with only 3 treatment groups in addition to the control.

2.1.3. Case Study 3

Case study 3 illustrates that an incorrect statistical test can lead to a wrong conclusion. As with other examples, the use of historical control data and outlier detection help to clarify the analysis. Summary data are given in Table 6. In regulatory review the NOEC was set at 10 ppm on the grounds that a ≥10% decrease was observed at the two higher treatment groups. By comparing the treatment means and medians, a skewness was deduced in the two highest treatment groups. Moreover, the standard deviations in the two highest treatment groups were much higher than in the control and low treatment. The data were found inconsistent with normality and variance homogeneity so a non-parametric analysis was indicated. There were only 6% and 1% decreases in the two treatment medians.
Count=number of replicates (breeding pairs)

Mean, median, and standard deviation are simple unweighted summary statistics.

Conc=ppm

The response was Eggs Hatched per Eggs Set (HATCH_ES). The non-parametric Dunn and Jonckheere-Terpstra (JT) tests found no treatment group significantly different from the control. Even though the full data did not justify a parametric analysis, the Dunnett and Williams’ tests were done anyway and reached the same conclusion, namely that no treatment differed significantly from the control. Figure 3 showed four low outliers, two in each of the two highest treatment groups. That figure also shows that all treatment means were within the HCD.

With the 4 outliers omitted, the means and medians were consistent and the data were found to be normally distributed with homogeneous variances. The Dunnett, Williams, and JT tests still found no significant effect at any dose. A NOEC =35 ppm is fully justified.

### 2.1.4. Case Study 4

Case study 4 illustrates a low-variability endpoint (eggshell thickness) in which there is a sharp drop from relatively high control mean to a somewhat flat and non-monotone treatment-response where all treatment means are significantly lower than the control but the actual percent change from the control is small and biologically unimportant. This example also illustrates an alternative to regression for estimating a 10% effects level when no acceptable regression model can be found. The data are summarized in Table 7.

| Group | Conc | Count | Mean  | Median | Std  |
|-------|------|-------|-------|--------|------|
| 1     | 0    | 15    | 85.29 | 88.46  | 8.70 |
| 2     | 4    | 14    | 82.04 | 82.43  | 9.01 |
| 3     | 10   | 13    | 72.02 | 83.02  | 25.86|
| 4     | 35   | 15    | 74.30 | 86.54  | 27.87|

Count=number of replicates (breeding pairs)

Mean, median, and standard deviation are simple unweighted summary statistics.
The data were found inconsistent with normality so non-parametric tests were used. All treatment group means were found significantly lower than the control mean by the JT and Dunn tests. The Tukey outlier test identified 3 outliers, one in the control and two in the highest treatment group. These can be observed in Figure 4. When those were omitted, the data were found normally distributed and homogeneous and Williams and Dunnett tests reached the same conclusion.

Despite the statistical significance of the decreases in all treatment groups, it should be noted that the maximum observed decrease was only 6%, which is unlikely to be biologically important. A decrease of less than 18% (EFSA 2009b) or 22% (Lincer 1975) in eggshell thickness, is not biologically important in terms of population effect. This is a rare instance when a specific size effect of biological importance is documented in the scientific literature for avian studies. It would be beneficial to hazard identification and risk assessment to have such information on more key endpoints. Note, however, the mean response at 25 mg/L is below HCD lower bound and the other treatment means are close to the HCD lower bound and the control mean is near the HCD upper bound.

No acceptable regression model was found for this non-monotone response. However, MAXSD, the maximum safe dose analysis (Tamhane et al 2001, 2002, 2004) found significantly less than 10% effect in every treatment group, making MAXSD=100. This means EC10LB>100 ppm. Thus, by setting 10% as a more biologically meaningful effect, the MAXSD is a more relevant measure of hazard than a simple NOEC and is a substitute for EC10 when no suitable regression model can be fit.

2.1.5. Case Study 5.

Case study 5 illustrates regression modelling that can be done when the data justify it. The emphasis is on model averaging. Two regression approaches were followed and a recommendation is made. The two approaches were to model proportions treated as continuous and to model 14-day survivors as binomially distributed conditioned on number hatched. The second approach is scientifically sounder since it treats the data as it was collected and this approach has better statistical properties. The data are summarized in Table 8.

| Group | Conc | Count | Mean | Median | Std |
|-------|------|-------|------|--------|-----|
| 1     | 0    | 14    | 94.60| 94.87  | 4.90|
| 2     | 28   | 16    | 89.73| 93.10  | 10.42|
| 3     | 85   | 11    | 92.28| 95.12  | 6.63|
| 4     | 255  | 15    | 83.34| 85.19  | 12.15|
| 5     | 510  | 13    | 77.69| 76.60  | 12.42|

Count=number of replicates (breeding pairs)

Mean, median, and standard deviation are simple unweighted summary statistics.
Figure 5 indicates a control mean response near the HCD upper confidence bound and the two lowest treatment means and medians not much different. The two highest treatment means are clearly lower, but still near or above the HCD lower confidence bound.

To determine the NOEC, the Dunnett, Williams, and Jonckheere-Terpstra tests were applied to the proportions. All tests found the NOEC=85 ppm, where a 2% decrease was observed. Only the Dunnett test could be applied in a GLMM model for the count of survivors conditioned on the number of hatchlings and the same NOEC was found. Figure 6 shows the Bruce-Versteeg (BVP) model fit to proportions, as this provides the simplest graphical representation. All regression models were fit to untransformed proportions assuming normally distributed, homogeneous responses and also to counts conditioned on the number of hatchlings using generalized nonlinear mixed models (GNLMM). This allowed direct comparisons of the two approaches. Mathematical descriptions of these models and model weighting schemes are given in the Supplementary material.

Tables 8 and 9 summarize the two modelling approaches. Table 9 summarizes approach 1 (models for proportions), where EC10 estimates are found reasonably tight, varying from 235 to 287. However, lower confidence bound (LCB) estimates vary widely from 9 to 166. Model averaging estimates: EC10_{avg}=258 and EC10LB_{avg}=76.7.

Table 9. Case study 5. Regression models for proportions

| MODEL | PARM | ESTIMATE | LCB | UCB | AICc   | Wgt |
|-------|------|----------|-----|-----|--------|-----|
| LL3   | Y0   | 0.94     | 0.89| 0.99| -313.49| 0.12|
|       | b    | 0.81     | 0.02| 1.60|        |     |
| OE4   | Y0   | 0.93     | 0.89| 0.97| -315.69| 0.35|
|       | C    | 0.69     | -0.12| 1.51|        |     |
|       | EC10 | 250.50   | 35.32| 465.60|        |     |
| OE2   | Y0   | 0.93     | 0.90| 0.96| -315.52| 0.32|
|       | EC10 | 286.60   | 166.50| 406.60|        |     |
| BVP   | EC10 | 244.10   | 99.10| 601.20| -313.31| 0.11|
|       | gamma| 2.31     | 0.12| 4.49|        |     |
|       | y0   | 0.93     | 0.89| 0.98|        |     |
| OE3   | Y0   | 0.94     | 0.89| 0.99| -313.51| 0.12|
|       | EC10 | 235.30   | 9.06| 461.60|        |     |
|       | D    | 0.76     | 0.02| 1.50|        |     |

LL3=3-parameter log-logistic
OE4=exponential model with a floor

OE2=simple exponential model

BVP=Bruce-Versteeg probit-type model

OE3=exponential model with curvature

PARM=model parameter

AICc=Akiaki information criterion with small sample correction

Wgt=Akaiki weight

Estimate, LCB, UCB=point estimate, lower and upper 95% confidence bounds

Table 10 summarizes approach 2 (GNLMM models for conditional counts). The BVP model parameters appear reasonable, but the estimated responses at positive test concentrations are very poor. This is evidently what caused the large value of AICc. The model average gives 0 weight to that model. Model averaging estimates were EC10_{avg}=277, EC10LB_{avg}=183. A tighter lower bound reflects less uncertainty in this estimate compared to that of Approach 1.

Table 10. Case study 5. GNLMM regression models for counts conditioned on hatchlings

| MODEL | PARM | ESTIMATE | LCB | UCB | AICc | wgt |
|-------|------|----------|-----|-----|------|-----|
| LL3B  | P0   | 0.94     | 0.92| 0.96| -34.7| 0.161|
|       | EC10 | 281      | 181.43| 380.57|      | 0.161|
|       | b    | 1.2      | 0.51| 1.89|      |     |
| OE4B  | P0   | 0.94     | 0.92| 0.96| -34.9| 0.178|
|       | C    | 0.25     | -6.20| 6.70|      |     |
|       | EC10 | 272      | 132.69| 411.31|      | 0.178|
| OE2B  | P0   | 0.94     | 0.92| 0.96| -37.1| 0.535|
|       | EC10 | 274      | 199.43| 348.57|      | 0.535|
| OE3B  | P0   | 0.94     | 0.92| 0.96| -34.2| 0.126|
|       | EC10 | 289      | 186.09| 391.91|      | 0.126|
|       | D    | 1.2      | 0.48| 1.92|      |     |
| BVPB  | P0   | 0.94     | 0.93| 0.947| 5412 | 0.000|
|       | B0   | -4.28    | -4.93| -3.63|      |     |
|       | B1   | 0.54     | 0.43| 0.65|      |     |
|       | EC10 | 261.86   | 228.16| 295.56|      | 0.000|

LL3B=3-parameter log-logistic GNLMM
OE4B=exponential model with a floor GNLMM
OE2B=simple exponential model GNLMM
BVPB=Bruce-Versteeg probit-type model GNLMM
OE3B=exponential model with curvature GNLMM
PARM=model parameter
AICc=Akaike information criterion with small sample correction
Wgt=Akaike weight
Estimate, LCB, UCB=point estimate, lower and upper 95% confidence bounds

3. Statistical Protocols

Statistical analysis should begin with careful consideration of the data for each response to be analyzed. A basic step is determining the appropriate distribution for a response. This begins by assessing whether the data are best realized as (a) a continuous response, such as egg shell thickness, that might come from a normal distribution, or (b) a proportion such as number of cracked eggs per eggs laid, which could be treated as normally distributed after a normalizing, variance stabilizing transform (usually an arc-sine square-root transform), or (c) a count response, such as total number of cracked eggs, or (d) a conditionally binomial response, such as number of cracked eggs conditioned on the number eggs laid. A list of commonly reported responses with their distributions is given in the supplementary material.

The most appropriate statistical methodology should be determined in order best to distinguish between real effects and mere artifacts of statistical probability by more properly reflecting the nature of the data and experimental design. To the extent possible, statistical analysis should be consistent with visual assessment of data. Only in limited situations, such as assessment of normality and variance homogeneity, and only then with expert judgment, a visual assessment may be sufficient without formal testing. Where visual assessment and formal tests are in conflict, the cause should be explored.

The ideal statistical methodology is a regression approach to estimate an appropriate percent effect of biological importance and its associated measure of uncertainty. This ideal is hampered by the small number of treatment groups in typical avian reproduction guideline studies.

The statistical tests listed in the case studies and in the decision flow diagrams are intended to be implemented as described in the cited references, especially Green et al (2018). Not all software packages that offer these tests implement them in equivalent fashion. For example, the R package mcp has a procedure that may appear to be Williams’ test. In fact, Williams’ test as described in Williams (1971, 1972) and Green et al (2018) and recommended here and in some OECD guidelines is quite different from the test in the mcp package. The StatCHARRMS R package provides a good, but not perfect,
approximation to Williams’ test as developed by Williams. A similar precaution is needed for regression models. For example, the software ToxRat does a preliminary transformation of the data prior to fitting regression models that, if not disabled by the user, distorts the concentration-response relationship and can result in seriously misleading BMD/ECx estimates. It is not a purpose of this manuscript to critique software packages that might be used to carry out the recommended protocols but some recommendations are provided below. In addition, the website associated with Green et al (2018) does offer programming code in SAS and R to carry out the tests and regression models discussed.

Figure 7 gives a decision chart that captures the highlights of the regression modeling steps. Figure 8 provides the same for a NOEC determination. Detailed NOEC decision charts are given in the Supplement for each type of response (e.g., quantal, continuous, conditionally binomial, count).

In Figure 8 the Conover test can be configured as a non-parametric alternative to the Dunn test, but the Dunn test is recommended in numerous OECD test guidelines and guidance documents (e.g. OECD 2006, 2014) and its power properties are documented more completely (e.g., Green et al 2018 and in documents supporting OECD test guidelines).

If there are outliers, at most 6 in total should be removed and preferably 4 or fewer. Otherwise, the outlier-omitted data may no longer truly represent the data collected. All data should be re-analyzed after outliers are omitted. If the NOEC or BMDx changes, then care should be taken interpreting results.

If a transform removed non-normality or variance heterogeneity/overdispersion, then the results of the transformed data are generally preferred. A check of distribution fit for GLMM models is assessed through studentized residuals and a non-significant normality test for these residuals means the data fit the modeled distribution.

3.1. Steps in the recommended statistical protocol. Additional detail is given in the Supplement.

1) Assess the distribution

Once the conceptually appropriate distribution is determined, it is important to assess the fit of that distribution (e.g., normality), variance homogeneity or overdispersion. Dunnett and Williams tests and various regression models assume normally distributed data with homogeneous variances. Both are assessed through residuals from an ANOVA model. Normality of the residuals can be assessed using the Shapiro-Wilk or Anderson-Darling test. In the case of generalized linear mixed models (GLMM), studentized residuals are used to assess agreement of the data to the modelled distribution. Variance homogeneity for a normally distributed response can be assessed using Levene’s test. For incidence and count data, overdispersion (also called extra-binomial variance) can be assessed using Tarone’s C(α) test or a method based on GLMMs.

2) Determine the presence, meaning, and impact of outlier
Careful consideration of outliers is advised since outliers can sometimes show that a statistically significant effect is the result of a small number of observations or the lack of statistical significance may be the result of high variability caused by one or more outliers. It should be emphasized that outliers are statistically detected unusual observations, not “bad” observations to be discarded. The primary purpose of outlier detection is to determine to what extent a small number of unusual observations influences the statistical tests and models. These observations may also be important indicators that merit further investigation. The Tukey outlier rule is recommended for continuous responses. But formal outlier rules need to be supplemented by consideration of other data anomalies. For example, 0 fertile eggs out of 1 egg laid is very different from 0 fertile eggs out of 36 eggs laid. A weighted analysis or treatment of fertile eggs as binomially distributed conditioned on the number of eggs laid is a potential way of dealing with some outlier issues. Decision trees for NOEC determination given in the Supplementary material indicate when consideration of outliers is applied. It should be noted that if the NOEC changes after outliers are removed or a normalizing, variance stabilizing transform is found, then scientific judgment is needed to resolve the difference.

3) Assess concentration-response monotonicity

Monotonicity in the dose-response should be assessed to determine whether a trend test (e.g., Williams, Jonckheere-Terpstra, Cochran-Armitage) should be used. Use of a trend test where it is not justified can obscure a real effect or indicate an effect that is not justified. Failure to use a trend test where it is justified ignores relevant biology and can miss an important effect or lead to confusion when a low dose is found statistically significant but higher doses are not.

In general, if a chemical affects a biological response, the effect increases with increasing concentrations of the chemical. That is, one expects a monotonic concentration-response. This is not a strict requirement, but serious deviations from monotonicity rule out the use of trend tests and should prompt careful exploration of the data. Much additional discussion of trend tests and ways to assess monotonicity are given in Green et al (2018) and Springer and du Hoffmann (2018). For normally distributed data with homogeneous variances, Williams’ test is recommended, but with cautions. This test uses a pool-the-adjacent-violators (PAVA) algorithm to smooth the data by forcing monotonicity. If the data deviates greatly from monotonicity, there can be too much smoothing which distorts the interpretation of the data. Green et al (2018) contains further discussion of this, as does OECD TG 248 (OECD 2019). As a rough guide, if three or more mean responses from positive test concentrations are merged by the PAVA algorithm, then the data may not be suitable for Williams’ test. A test for monotonicity is given in the Supplementary material.

For continuous response data that do not meet the requirements of normality and variance homogeneity, the Jonckheere-Terpstra test is a non-parametric trend test that has similar power as Williams’ test to detect effects. Like Williams’ test this is a step-down trend test but unlike Williams, it does not use a smoothing algorithm and so does not have the same tendency as Williams’ test to mask departures from monotonicity. For incidence data, the Cochran-Armitage test is very useful step-down trend test. Where
overdispersion is found, a robust version of that test using the Rao-Scott adjustments can be used. All these tests are discussed in detail in Green et al (2018), where additional references are also given. Of those several deserve additional mention, including OECD (2006, 2014).

4) Use historical control data if available

Valverde et al (2018) investigated the utility of historical control data for interpreting avian reproduction studies, including power analyses to document the size effect that could be expected to be found statistically significant. The work reported here continues and, to some degree, extends that work. If historical control data are available, such data could provide information on which observations indicate real effects, which observations are well within the historical control range, and can alert the investigator to the presence of an unusual control that may skew statistical analysis. By examining the study data in the context of historical control data, some responses may be found not to require further statistical analysis. Once statistical analysis is done to determine a NOEC or estimate an ECx value, the study data again can be compared to relevant historical control data to help interpretation for hazard identification and risk assessment.

The most appropriate historical control data is from the same laboratory that does the concurrent study and uses data within a time interval centered on the date of the concurrent study. Historical control data from other laboratories can be used if appropriate inter-laboratory comparisons have been done. A span of 2 - 5 years on each side of the date of the concurrent study is recommended. However, European Commission (2013) recommended a 5-year span centered on the starting date of the study. The span will depend in part on the number of studies in the database. It would be best to have 20 or more studies from the HCD, approximately equally split on both sides of the concurrent study date where possible. Once the span of time to include in the HCD is determined, extreme observation should be discarded to avoid skewing the interpretation. It is suggested that a concurrent treatment mean response between the 5th and 95th percentiles of the HCD is not indicative of a real effect. These percentiles are dependent on the number of studies in the HCD and a reality check would include assessing the data using several time spans, such as ± 2, ± 3, and ± 5 years in the HCD to make sure these percentiles are not overly influenced by the size or the time span of the HCD. Note also that 5% of 20 is 1, so the 5% and 95% bounds on a smaller HCD are of questionable relevance.

5) Transform responses to meet test requirements or use generalized (non-)linear mixed models

Transformation of responses must be order-preserving. For example, the Freeman-Tukey transform of proportion data need not be order preserving and its use can distort or even reverse some concentration-response relationships and produce misleading results. If regression models are used to estimate ECx, the meaning of an x% change in the transformed response is unlikely to be equivalent to an x% change in the original, untransformed response.

For proportion responses such as viable eggs per eggs set, the traditional way to analyze is to treat these responses as continuous responses, often with a normalizing, variance stabilizing transformation such
as the arc-sine square-root transform. That remains a viable method, but another method can be more informative and is more consistent with the nature of the data. This is the use of a generalized linear mixed model that treats the numerator, viable eggs in the illustration, as binomially distributed conditioned on the denominator, eggs set in the illustration. Count data, such as eggs laid, can likewise be analyzed by treating the data as continuous, usually following a square-root transform, or using a GLMM with a Poisson distribution. Where overdispersion is found, an adjustment is recommended, such as using a negative binomial distribution or allowing variance to vary by treatment group. See Green et al (2018) for additional details and references on all the statistical recommendations.

6) Use regression or BMD methodology where supported by data

Where more treatment groups are available in a study and regression modeling is feasible, model selection criteria are important. Criteria are described in the Supplement. Simulation studies reported by Burnham and Anderson (1998) among others, demonstrate that if the same model selection procedure is followed in repeat studies using the same test concentrations and study design, then different models from the set of models used will be selected in different studies. To compensate for this model uncertainty, a model averaging technique can be implemented. There are two main ways to approach model averaging. Benchmark dose (BMD) methodology outlined in (EFSA 2009a) indicated that the lowest point and interval estimates be used from all models in a set of standard models. This recommendation was updated in EFSA (2017) to use a combination of bootstrap sampling and weighted averages. The most common weighting scheme is based on a single information criterion, such as AIC or BIC. Details are given in the Supplement. With either approach, care must be taken to identify the set of models to use, as clearly both model average and model selection are highly dependent on the models utilized. In addition, one should not rely solely on an automated procedure, such as Akaike weight, that down weights contributions from poorly fitting models or focuses on only one selection criterion. It is also important to understand the limitations of regression modelling. Once a model is fit to a dataset, it is mathematically possible to estimate ECx for any positive value of x up to 100 for a decreasing model. Not all such estimates are statistically reliable. The dangers of extrapolation much beyond the range of tested positive concentrations are well understood. Also, an estimate of ECx for x < 10 is often beyond the capability of the data. For example, a 1% or 5% change in adult body weight or proportion of eggs laid that hatch or survive 14 days is rarely possible. The impracticality of such estimates is often indicated by a wide confidence interval or a confidence interval extending below 0. More detail on this in given in the Supplement under the heading of model fitting criteria.

7) Assess the need for special regression models

When there is a flat response in the treatment groups but all such groups differ significantly from the control, a “hockey-stick” model may be helpful in describing the data and providing ECx estimates where more standard decreasing models fail. If hormesis is evident a hormetic model, such as Brain-Cousens, should be considered. Such models usually require more test concentrations than commonly found in avian reproduction studies.
8) Consider an alternative to NOEC and BMD

The small number of treatment groups can sometimes be overcome by a statistical methodology designed to test for a specified level of effect of biological or regulatory importance. For example, a maximum “safe dose” or MAXSD can be identified at and below which the effect of the test substance is significantly less than 10%. This method can also be applied when there are more treatment groups but no acceptable regression model can be found.

3.2. Software

While it is not the intent to give a survey of software available to carry out the recommended statistical tests to determine a NOEC or models to estimate ECx or BMDx, it still seems appropriate to provide brief descriptions of some software packages useful for the two approaches. For regression model fitting, including model averaging, there are at least three good choices. These are the R package drc (Ritz et al 2005, 2006); Proast (Slob 2018, 2019) which was developed specifically for regulatory risk assessment under the auspices of RIVM; BBMD (Shao and Shapiro 2018; Shao 2021) which provides a Bayesian implementation. Also notable is the BMD software developed by the United States EPA (BMDS Benchmark Dose Tools | US EPA) which is an Excel-based application. The current version (3.2) provides model averaging only for dichotomous responses, which limits its utility for avian reproduction studies. The first two cited packages use the Akaaike information criteria to obtain weights for model averaging. The third and fourth cited packages use weights based on prior distributions but otherwise follow the same idea of estimating both BMDx and BMDxLB on these weights. One should be aware that Bayesian model averaging can produce notably different results from the information criteria approach and the list of models used in averaging can also have a strong impact on results. The criteria (e.g., all convergent models from a fixed list or only those meeting some additional criteria) used to decide which model fits to include can also impact results.

For NOEC determination, Cetis (Ives 2021), which was developed for the United States EPA, implements all the standard statistical tests recommended, but not the GLMM tests. The R package PMCMRPlus (2021) provides all tests described for continuous responses, including non-parametric rank-based tests, but it does not include GLMM tests or tests for quantal data. SAS software has very useful procedures for GLMM models but these require programming. There are numerous R packages for GLMM but results from different packages compared to each other or to SAS will often not agree. A good resource for relevant GLMM models in R is Hothorn (2016).

3.3. Biological relevance

Real improvement in hazard identification and risk assessment requires scientifically based criteria for what constitutes a hazard. According to EFSA 2009b, in determining a NOAEL there may not be a consideration of the effect or its biological relevance. Therefore, it is proposed to use endpoints that are based on a consideration of the biological and/or ecological relevance. Consequently, the biological
relevance should be always considered for the final toxicological endpoint selection as a higher tier refinement option.

For example, Case study 4 illustrated the importance of having an agreement on the size effect on eggshell thickness in evaluating a statistical finding as despite the statistical significance of the decreases in all treatment groups, it should be noted that the maximum observed decrease was only 6%. This endpoint is a rare instance when scientific evidence is available for this purpose. According to EFSA 2009b population effects in the wild tend to come about after thinning of 18% or more. Eggshell breakage increases when eggshells become more than 22% thinner than unaffected eggshells (Lincer 1975). Overall, the maximum observed decrease of 6% should not be considered biologically important and thus the final NOEL should be set to the maximum concentration tested (Table 7, Figure 4).

The regulatory process would be much enhanced by the establishment of such information on other key responses. As it is, an arbitrary rule is adopted, such as a 10% change, or a statistically significant change is used regardless of biological importance.

4. Conclusions

Current test guidelines and guidance emphasize purely statistical methodology for hazard identification. The focus of Test Guideline 206 is on whether a statistically significant change is observed in one or more test concentrations compared to the concurrent control. More recent EFSA guidance (EFSA 2017) emphasizes BMD10 or its lower 95% confidence bound. Other relevant information is often either ignored, such as historical control data, or not available, such as no biological basis for the size of effect important to be able detect or estimate (biologically significant effect level).

Evidence has been presented on ways to improve NOEC determination through improved statistical test selection, diagnostics, and the use of historical control data. In particular, generalized linear mixed models take the natural distribution of the response variables and the sources of random variability into account, leading to more appropriate corresponding statistical tests for several endpoints as does careful attention to identifying outliers. BMDx estimation can be improved through the use of revised modelling techniques including generalized nonlinear mixed models, implementation of model selection criteria and model averaging in addition to adopting improvements to the experimental design. Historical control databases from years of avian reproduction studies demonstrate that BMD estimates, especially BMD10, will not be possible for some study response variables, so that NOECs will continue to be required for use in risk assessments. Explicit proposals for statistical tests, models and experimental designs are provided that require no more birds per study than currently required in TG 206 studies but nonetheless are more statistically sound and robust, for deriving the endpoints used for risk assessment.

A clear correlation has been found from the laboratory studies showing a decreased hatching and population decline was associated with 18 to 22% reduction in eggshell thickness. This illustrates the need for additional information to quantify the level of effect for key endpoints that indicate population
level effects and distinguish such effects from mere statistical significance or a percentage change from control without associated biological significance.

**Declarations**

- **Ethics approval and consent to participate:**

  No human tissue was involved in this work and no animals or plants were used expressly for this study. All data analyzed in this work were extracted from studies previously submitted to regulatory authorities and followed the ethical approvals indicated in OECD Test Guideline 206 or OCSPP 850.2300: Avian Reproduction Test. The historical control datasets came from two testing laboratories, Eurofins and Smithers Vincent and consisted of data obtained from studies performed at those facilities done under the same two regulatory guidelines.

- **Consent for publication:**

  Not applicable

- **Availability of data and materials:**

  The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request, except the historical control data which belong to the testing laboratories.

- **Competing interests:**

  The authors declare that they have no competing interests.

- **Funding:**

  Funding for this research was provided by European Crop Protection Association (Crop Life Europe)

- **Authors’ contributions:**

  The lead author did all statistical analyses, developed the statistical protocol, and wrote the manuscript. Each other author contributed datasets, identified responses of special interest, helped shape the research and manuscript, made valuable comments and edits throughout the process and especially for the final manuscript. In addition to the above, Manousos Foudoulakis was the leader of the Terrestrial Vertebrates ad hoc Team (TVahT) of the European Crop Protection Association and provided overall guidance for the work reported here.

- **Acknowledgements:**

  Not applicable
References

1. Burnham, K.P. and Anderson, D.R. 1998. Springer-Verlag New York, Inc. ISBN 0-387-95364-7.
2. Cade, B.S. 2015. Model averaging and muddled multimodel inferences. Ecology 96: 2370-2382. https://doi.org/10.1890/14-1639.1

3. Dunnett, C.W. 1989. Algorithm AS 251: Multivariate Normal Probability Integrals with Product Correlation Structure. J R Stat Soc Series C (Applied Statistics) 38: 564-579.

4. EFSA 2009a. Guidance of the Scientific Committee on a request from EFSA on the use of the benchmark dose approach in risk assessment. EFSA J. 1150: 1-72.

5. EFSA 2009b. Risk Assessment for Birds and Mammals. EFSA J 2009; 7(12):1438.

6. EFSA 2017. Update: use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1):4658.

7. European Commission 2013. Commission regulation (EU) No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market.

8. Fleiss, J.L., Levin, B., Paik, M.C. 2003. Statistical Methods for Rates and Proportions, third edition. Wiley. ISBN 0-471-52629-0. Hoboken, NJ.

9. Green, J.W., Springer, T.A., Holbech, H. 2018. Statistical Analysis of Ecotoxicity Studies. Wiley. ISBN: 978-1-119-48881-1.

10. Hothorn, L.A. 2018. Statistics in Toxicology Using R. CRC Press. ISBN-13 number 978-1-4987-0127-3.

11. Ives, M.A. 2021. Comprehensive Environmental Toxicity Information System (CETIS), Version 2.0. Tidepool Scientific LLC, McKinleyville, CA. http://www.tidepool-scientific.com/Cetis/CetisStats.html

12. Lincer, J.L. (1975) DDE-Induced Eggshell-Thinning in the American Kestrel: A Comparison of the Field Situation and Laboratory Results. J Appl Ecol. 12:781

13. OECD 2019. Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, https://doi.org/10.1787/a13f80ee-en.

14. OECD 2006. Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application, OECD Series on Testing and Assessment, Number 54, ENV/JM/MONO(2006)18, Environment Directorate, Organisation for Economic Co-Operation and Development, Paris.

15. OECD 2014. Fish Toxicity Testing Framework, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264221437-en.

16. PMCMRPlus 2021.PMCMRplus.pdf (r-project.org)

17. Shao, K. and Shapiro, A.J. 2018. A Web-Based System for Bayesian Benchmark Dose Estimation. Environ. Health Perspect. 126. CID: 017002 https://doi.org/10.1289/EHP1289

18. Shao, K. 2021. Bayesian BMD (benchmarkdose.org)
19. Slob, W. 2019. PROAST. A general software tool for dose-response modelling. RIVM, Bilthoven. PROAST MANUAL GUI version.pdf (rivm.nl)

20. Slob, W. 2018. Joint project on Benchmark Dose modelling with RIVM. EFSA supporting publication 2018:EN-1497. 14 pp. doi:10.2903/sp.efsa.2018.EN-1497 ISSN: 2397-8325.

21. Springer, T.A., du Hoffmann, G. 2018. Evaluation of Monotonicity of Concentration Response in Avian Reproduction Studies – Summary of Findings. Project 857B-101. Eurofins Laboratory.

22. Staveley, J.P., Green, J.W., Nusz, J., Edwards, D., Henry, K., Kern, M., Deines, A.M., Brain, R., Glenn, B., Ehresman, N., Kung, T., Ralston-Hooper, K., Kee, F., McMaster, S. 2018. Variability in Non-Target Terrestrial Plant Studies Should Inform Endpoint Selection. Integr. Environ. Assess. Manag. 14: 639-648.

23. Ritz, C., Jensen, S.G., Gerhard, D., Streibig, J.C. 2020. Dose-response analysis using R. CRC Press. Boca Raton, Fl.

24. Ritz, C. and Strebig, J.C. 2016. drc.pdf (r-project.org)

25. Schervish, M.J. 1984. Algorithm AS 195: Multivariate Normal Probabilities with Error Bound. J R Stat Soc Series C (Applied Statistics) 33: 81-94.

26. Tamhane, A.C. and Logan, B.R. 2004. Finding the maximum safe dose level for heteroscedastic data. J Biopharm Stat. 14: 843-56.

27. Tamhane, A.C. and Logan, B.R. 2002. Multiple Test Procedures for Identifying the Minimum Effective and Maximum Safe Doses of a Drug. JASA 97: 293-301.

28. Tamhane, A.C., Dunnett, C.W., Green, J.W., Weatherington, J.D. 2001. Multiple Test Procedures for Identifying the Maximum Safe Dose. JASA 96: 835-843.

29. Tarone, R.E. and Gart, J.J. 1980. On the robustness of combined tests for trends in proportions. JASA 75: 110-116.

30. Tukey, J.W. 1977. Exploratory Data Analysis. Reading, MA: Addison-Wesley.

31. USEPA 2015. https://www.epa.gov/sites/production/files/2015-11/documents/support_material_for_model_averaging_workshop-11_06_2015-508.pdf

32. USEPA 2021. https://www.epa.gov/bmds.

33. Valverde-Garcia, P., Springer, T., Kramer, V., Foudoulakis, M., Wheeler, J.R. 2018. An avian reproduction study historical control database: A tool for data interpretation. Regul. Toxicol. Pharmacol. 92: 295-302.

34. Williams, D.A. 1971. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 27: 103-117.

35. Williams, D.A. 1972. The comparison of several dose levels with a zero dose control. Biometrics 28. 519-531.

Figures
Case Study 1. The response was percent eggs cracked per eggs laid (PCL). In the box plot of study data, the bottom and top are the 25th & 75th percentiles of the indicated treatment group, the horizontal line between the top & bottom is the median, the mean is the blue asterisk, whiskers extend from top or bottom to most extreme non-outliers, with outliers indicated as blue dots beyond the whiskers. Individual historical control data mean responses are indicated by black asterisks. HCD data from 2003-2005 are displayed to the left of the 2006 study data, while HCD data from 2006-2009 are displayed to the right.
Figure 2

Case Study 2. Quail 14 Day survivors per eggs set Only the mean response at 35 ppm was below the HCD and that resulted from a single outlier.
Case Study 3. Eggs Hatched per Eggs Set (HATCH_ES) HCD data from 1993-1995 are displayed to the left of the concurrent study, while data from 1996-1999 are on the right. Two low outliers are evident at 10 and two more at 35 ppm. All treatment means and medians are within the HCD.
Figure 4

Case Study 4. Eggshell thickness (ESThick) HCD data from 1993-1995 are displayed to the left of the concurrent study, while data from 1996-1999 are on the right. One high outlier is evident in the control and two low outliers are evident at 100 ppm. The mean response at 25 ppm is below the HCD lower confidence bound. The control mean response is near the HCD upper bound and mean responses in the two highest treatments are near the HCD lower bound. The median responses in all treatment groups were below the HCD lower bound. There is no question about the statistical significance of the decreased thickness in the three treatment groups. Nor is there any question about whether the mean or median responses are below the HCD lower bound. The question is whether such small differences are biologically important. EFSA guidance indicates not.
Figure 5

Case Study 5. Percent survivors at 14 days per hatchling (P14DH) HCD data from 1994-1995 are displayed to the left of the concurrent study, while data from 1997-2000 are on the right.
Figure 6

Bruce-Versteeg model fit to replicate proportions (PRP14DH) No transform was applied to the proportions since that makes estimating a 10% change (with confidence bounds) in the untransformed data quite challenging. This is one reason for preferring a GNLMM model which requires no transform.
Figure 7

Statistical decision chart for regression modeling/ECx estimation. If a regression model fitting converges, then either model selection or model averaging (recommended) is applied. If the model fitting does not converge, the model is discarded without further use.
Figure 8
Statistical decision chart for NOEC determination NHomV=normally distributed with homogeneous variances NHetV=normally distributed with heterogeneous variances BNOV=Binomially distributed with no overdispersion BOV=Binomially distributed with overdispersion Red wavy curve on chart indicates that for the indicated tests, the tests are repeated with outliers omitted to determine whether outliers affected NOEC.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialrev2.docx