


DISCUSSION

Open Access



Statistical analysis of avian reproduction studies

John W. Green^{1*} , Manousos Foudoulakis², Timothy Fredricks³, Thomas Bean⁴, Jonathan Maul⁵, Stephanie Plautz⁶, Pablo Valverde⁷, Adam Schapaugh³, Xiaoyi Sopko⁸ and Zhenglei Gao⁹

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 endpoint to be used in avian reproduction studies for 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 alternative 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.

Keywords: Avian reproduction, Hazard identification, Benchmark dose, Distribution, Model average, Diagnostics, NOAEL, Historical controls, Hormesis, Maximum safe dose

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 OCSP 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 [6–8] 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 can have as few as three test concentrations plus a negative control. Regression models that can be fit to such data are

*Correspondence: John@JohnWGreen-ecostats.com

¹ John W Green Ecostatistical Consulting LLC, Newark, DE, USA
Full list of author information is available at the end of the article

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 variances of these biological measures vary greatly. For measures with the lowest variances, very small difference between control and exposure groups can be statistically significant. For other biological measures, relatively large random differences between replicated groups are not unusual, and it may be difficult to distinguish real effects from background statistical “noise”. The statistical distributions characteristic of the different types of biological measures also vary greatly and require careful selection of distribution-appropriate models and statistical tests. Erratic concentration–response patterns, where there is little apparent relationship to concentration such as a saw-toothed appearance, add to the challenges to statistical interpretation. It should also be acknowledged that there is limited scientific basis to guide the risk assessor in choosing the size of effect for which benchmark doses should be estimated or which should be dependably detectable by studies designed to support hypothesis testing (e.g., determining NOECs). A rare exception to this is a conclusion that only an 18% decrease [7] or 22% [14] 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 concentration causing a 10% effect 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.

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 (BMDL10) and the size effect that can be estimated reliably are

discussed. However, as will be demonstrated in what follows, not all regulatory required responses from the current or any practical alternative experimental design are suitable for BMD methodology. Statistical methodology is recommended that is both more consistent with the nature of the data and is more consistent statistically to determine NOEC values as well. The intent is to make the best use of the data collected as well as to improve the experimental designs that generate the data. This can be done with little or no increase in the number of animals used in testing.

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 the highest test concentration approximately one-half of the LC10 determined by a prior dietary study (OECD TG 205), if such a study is available and delivers an LC10, but not to exceed 1000 ppm. This makes the test concentrations equally spaced on a logarithmic scale. There should be at least 12 replicates, each consisting of two or three birds with sex and number depending on the species tested. However, most studies are done with 16 to 18 replicates.

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):

$$MDD\% = CV * T \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}, \quad (1)$$

where $T = t_{(1-\alpha, df)} + t_{(1-\beta, df)}$, α and β are the false positive and false negative probabilities, CV is the coefficient of variation in the control expressed as a percent, n_0 and n_1 are the number of replicates in the control and each treatment group, respectively. Good discussion of MDD% is given in Duquesne [5]. For the table provided in this section, α and β are taken to be 0.05 and 0.2, respectively, corresponding to a power of 80% to detect an effect of the size MDD%. Staveley et al. [24] developed a method to estimate the minimum size effect (MSE%) that can be estimated reliably from typical regression models. That method was adapted here to show that MSE% is a multiple of MDD% as indicated by the following equation:

$$MSE\% = MDD\% \sqrt{\frac{h_z}{\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}} \frac{T^*}{T}, \quad (2)$$

where T^* is student's 2-sided t -statistic $t_{(1-\alpha/2, df)}$, h_z is the leverage associated with a given treatment group. The Supplement contains more details on Eq. (2), including an example calculation of the leverage h_z . (In the Supplement, the definitions of T and T^* are reversed.) 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 a different experimental design which includes 4 or 5 test concentrations and higher replication. Determining the multiplier of MDD% in equation (Eq. 2) that is applicable to avian studies will be further developed in a subsequent paper. For present purposes, a simulation study 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 historical control described below and experimental designs with 5 treatment groups plus control.

Using avian historical control data described in "Historical controls" section, it was shown that MDD% varied from 1 to 4% for percent eggs not cracked per eggs laid to 18–38% for eggs per hen, depending on the testing lab and species. A short summary table for selected responses is given in Table 1. It will be clear from Table 1 that for a few responses, it will be possible to detect effects or 5%, but often 10% or even 20% effects cannot be detected reliably. Similarly, Table 2 shows that reliable estimation of BMD10 is often unattainable. Table 2 only shows the point estimate. The properties of the lower confidence bound even more problematic. Based on these results, there is considerable challenge to the idea of replacing the NOEC by the BMD10.

As Eqs. 1 and 2 indicate, increased replication will reduce the size effect that can be detected or estimated, but practical designs will still not permit reliable BMD10 estimates for some responses. Even doubling the replication would reduce MDD% by only 30%. Whether a similar reduction in MSE% would also be obtained is an issue being addressed in a simulation study to be reported later. Analysis of the historical control databases suggests 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

Table 1 MDD% for mallard and quail

MDD%			Response	Abbrv			
Mallard					Quail		
p50	p75	p90			p50	p75	p90
18	23	35	Eggs laid per hen	EL	24	28	38
1	3	3	Eggs not cracked/number eggs laid	ENC_EL	2	3	4
8	16	24	Live embryos/number eggs set	LE_ES	8	16	25
2	4	9	Live embryos/number viable embryos	LE_VE	1	2	3
18	22	25	Number hatched/number eggs set	NH_ES	13	18	25
10	15	20	Number hatched/number live embryos	NH_LE	5	9	13
18	22	25	14-day survivors/number eggs set	HS_ES	14	19	26
2	2	3	14-day survivors/number hatched	HS_NH	4	7	10
26	29	35	Number of 14-day survivors per hen	HS	30	34	39
6	7	7	Hatchling body wt (g)	HATWT	5	6	7
6	7	8	14 Day survivor BW (g)	SURVWT	7	8	8
4	5	6	Eggshell thickness (mm)	THICK	5	6	6
8	9	12	Adult food consumption (g/bird/d)	FOOD	8	8	9
94	203	293	Adult male body weight gain (g)	WTGAINM	81	101	155
45	51	58	Adult female body weight gain (g)	WTGAINF	35	42	79
25	29	35	Number of hatchlings per hen (#/hen)	NH	29	34	38

pn = nth quantile of distribution of MDD% for indicated responses, $n = 50, 75, 90$

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

Table 2 Distribution of EC10 point estimates for shape ECPB = 1

EC10		Percentiles				YMAX	CV	ECPB	%Fit	P90/P10
Mean	Med	10	25	75	90					
26	26	22	23	28	30	60	5	1	100	1.4
26	25	18	21	30	34	60	10	1	100	1.8
25	24	13	17	32	38	60	20	1	87	2.9
24	23	10	14	32	41	60	30	1	50	4.1
63	21	7	12	33	46	60	40	1	26	6.6
37	37	31	33	41	43	70	5	1	100	1.4
36	35	26	30	42	47	70	10	1	100	1.8
34	33	17	23	44	55	70	20	1	73	3.2
36	30	11	18	45	60	70	30	1	40	5.4
68	27	7	14	45	61	70	40	1	18	9.0
55	55	46	50	60	64	80	5	1	100	1.4
53	53	38	44	62	70	80	10	1	99	1.8
68	48	22	33	64	77	80	20	1	57	3.5
70	39	12	23	60	81	80	30	1	24	6.7
56	35	5	16	58	89	80	40	1	8	17.6
100	96	83	89	106	122	90	5	1	97	1.5
150	92	63	77	118	207	90	10	1	77	3.3
246	77	30	50	131	341	90	20	1	48	11.3
162	60	8	27	94	219	90	30	1	35	28.4
139	48	0	15	84	251	90	40	1	24	5410.5

EC10: Mean=mean EC10 estimate from simulated data; EC10: Med=median EC10 estimate from simulated data; Percentiles: =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=Percent 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

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 modeling 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.

Simulation study to explore ECx/BMDx estimation for avian studies with modified design

The database of available avian studies is not large enough to develop a distribution of MSE% as was done in Staveley et al. [24]. Instead, concentration–response

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. [11], suggest generalized nonlinear mixed models (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 GNLMMs. Instead lower confidence bounds for the point estimates will tend to be extremely close to zero where a simulated continuous curve will give negative lower bounds.

The shapes of the concentration–response curves 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 concentration–response curve, and ECPB = 10 corresponds to a more delayed decrease or shallower

concentration–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 Eq. 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 details, including figures of the indicated concentration–response shapes and criteria for goodness of fit, are 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 are 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.

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.

Table 3 Distribution of EC10 lower bound (ECL10) estimates for shape $ECPB = 1$

ECL10		Percentiles				YMAX	CV	ECPB
Mean	Med	10	25	75	90			
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

ECL10: Mean = mean ECL10 estimate from simulated data; ECL10: Med = median ECL10 estimate from simulated data; Percentiles: = percentile of ECL10 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 ECL10 or another percentile indicate ECL10 estimate of little value for risk assessment. When $ECL10 < 0$ then the EC10 estimate is statistically indistinguishable from zero

- Better models, generally meaning the use of alternative distributions in familiar normality-based regression models (putting them formally in the category of generalized nonlinear mixed models or GNLMMs) can improve EC_x estimation when regression modeling is possible as well as making them more consistent with the nature of the data.
- MAXSD can provide a substitute for difficult to fit regression or improve rationale for NOEC approach. This is a technique for establishing an upper bound on the test concentration at which the percent effect is statistically significantly less than 10% (or another percentage effect chosen by the user).
- Model averaging takes model uncertainty into account and reduces effect of poorly fitting models.

Methodology

To achieve the objectives laid out in the introduction, the main tools were (1) an historical control data base (HCD) of such studies and a way of incorporating HCD in risk assessment; (2) an illustrative set of avian reproduction studies that were compiled by the Terrestrial Vertebrates ad hoc Team (TVaHT) of the European Crop Protection Association; (3) illustrations of some statistical methods for the various types of responses required in avian reproduction studies, including both older methods and newer approaches with a focus on exploring the benchmark dose (BMD) methodology. These tools serve as a partial motivation for the statistical protocols recommended for future avian reproduction studies that overcome some of the limitations of methods that have been used for many years. The resulting protocols are given in Sect. 3, Recommended Statistical Protocols, serve the function of a Results section.

Responses from the case studies were analyzed statistically using both standard and novel statistical methods and models. These analyses contributed to the development of detailed statistical protocols that cover the range of responses in avian studies. Detailed protocols are presented following summaries of the selected case studies that illustrate the concerns that commonly arise and that helped motivate the protocols. More detailed analyses of these and additional case studies are presented in Additional file 1.

Among the newer methods described are generalized linear mixed models (GLMMs) for NOEC determination and generalized nonlinear models (GNLMMs) for BMD_x estimation. On a conceptual level, the difference between GLMM and the classical ANOVA methodology for NOEC determination is simply the use of alternative distributions to describe biological responses that do not fit the usual normality paradigm. Similarly, the GNLMMs

used for BMD_x estimation are the standard nonlinear models that have been used for decades to model many ecotoxicology responses. These models are described in many publications, including OECD [16], Green et al. [11], Ritz et al. [25], Hothorn [18], and Shapiro [19]. Software to implement GLMMs and GNLMMs is readily available, for example, in Ritz and Streibig [26], Slob [21], Shao [20], BMDS [1], and [13].

Historical controls

The importance of historical control data (HCD) as an aid in distinguishing between real effects and statistical artifacts has been described above. Historical control data for avian reproduction studies done under OECD TG 206 or OCSPP 850.2300 has been made available from Eurofins (years 1976–2016 for quail and 1978–2016 for mallard) and by Smithers (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, a response 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 [9] recommended a 5-year period centered on the data of the current study with a minimum of 20 studies during that period. One challenge with this recommendation is that in the current market, it is estimated that globally at most six new studies will be done each year so that a single testing facility is unlikely to have 20 studies in a 5-year period. In displays of HCD, it is useful to put bounds indicating where most of the data reside. There is no hard rule about how to define these bounds. Throughout this manuscript, dashed lines are used to indicate the middle 95% of the HCD. That is, the upper bound is the 97.5% quantile of the HCD in whatever time interval is shown and the lower bound is the 2.5% quantile.

The main use of historical control data recommended for risk assessment is to demonstrate whether statistically significant trends or changes from the control mean response are the result of unusual concurrent controls or mild trends lying entirely within the HCD 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.

Models used for BMDx estimation

Since the case studies rely in part on fitting regression models, the list of models used for that purpose is provided here. Specific formulas for these models are given in the Supplementary material. In this manuscript, when one of the listed models is fit using one of the alternative distributions described, it is a GNLMM and whenever a reference is made to fitting a GNLMM, one of these models is being used.

For continuous responses, such as thickness and weight, the models recommended in OECD [16] and [17] are the Bruce–Versteeg, 3-parameter log-logistic, a set of four exponential models labeled OE2 (simple exponential), OE3 (simple exponential with a shape parameter), OE4 (simple exponential with a “floor” or lower bound on the response), and OE5 (simple exponential with a shape parameter and floor), and the Brain–Cousens hormetic model. All of these models can be fit using alternative distributions such as binomial, Poisson, or conditionally binomial, and can include adjustments for overdispersion or variance homogeneity or weights to accommodate increasing or decreasing patterns of concentration dependent variance.

For quantal data, the probit model can be fit using some normal approximations or using a binomial error distribution with the same modifications listed above.

Case studies

Each case study begins with a brief introduction of what it purports to show. Following that are analysis and discussion that describe what was done and justify the introduction. The analyses introduce in context most of the newer methods that are given in the protocols presented below. In these case studies, the term quail always means northern bobwhite quail.

Case study 1

Case study 1 illustrates how historical control data can help identify an extreme concurrent control result that can yield misleading significance in a statistical test. It also illustrates the importance of outlier identification. Table 4 shows the mean response (as a percent) in

Table 4 Case study 1

Group	Conc	Count	Mean	Median	Std
1	0	16	99.37	100.00	1.45
2	25	16	96.87	98.00	3.88
3	50	16	97.75	100.00	3.34
4	100	16	94.69	98.00	9.21

Quail percent eggs not cracked per eggs laid (PCL)

Count = number of replicates (breeding pairs), mean, median, and standard deviation are simple unweighted summary statistics; Conc = ppm

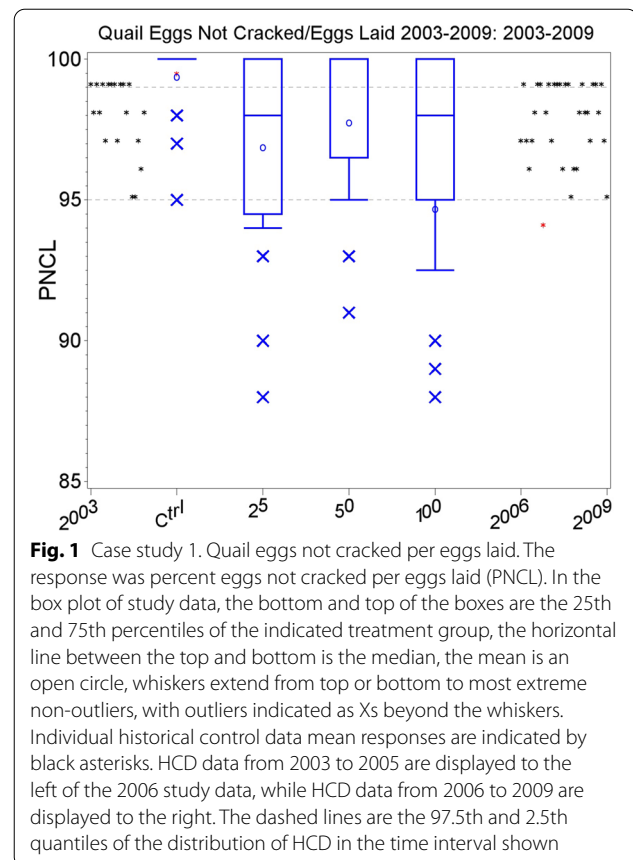


Fig. 1 Case study 1. Quail eggs not cracked per eggs laid. The response was percent eggs not cracked per eggs laid (PNCL). In the box plot of study data, the bottom and top of the boxes are the 25th and 75th percentiles of the indicated treatment group, the horizontal line between the top and bottom is the median, the mean is an open circle, whiskers extend from top or bottom to most extreme non-outliers, with outliers indicated as Xs beyond the whiskers. Individual historical control data mean responses are indicated by black asterisks. HCD data from 2003 to 2005 are displayed to the left of the 2006 study data, while HCD data from 2006 to 2009 are displayed to the right. The dashed lines are the 97.5th and 2.5th quantiles of the distribution of HCD in the time interval shown

all treatment groups with all observations included. The mean response, quail eggs not cracked per eggs laid, is notably lower in all treatment groups than in the control. Also relevant is that the concentration–response is not monotone.

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 $(LB, UB) = (95, 99)$, so the mean response at $\text{conc} = 100 \text{ ppm} > UB$. The Tukey outlier test [32] found 1 outlier in the 100 ppm treatment group. With that observation omitted, the mean response in that group was 96.73% which is well above LB. The control mean response, 99.38, is above the upper 95% CB of the HCD.

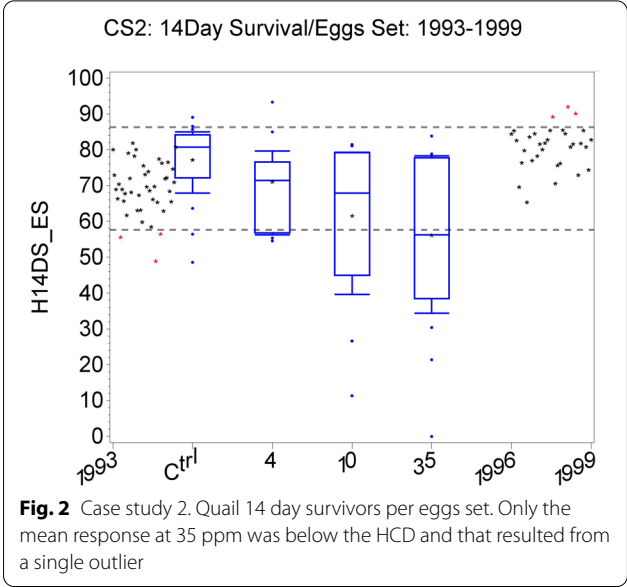
Statistical significance was assessed in several ways. A standard way to analyze such percentage data is to apply Dunnett’s test to arc-sine square-root transformed proportions [10]. A scientifically preferable analysis would be to analyze the number of not-cracked eggs as binomially distributed conditioned 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 accommodate it in the model failed. Overdispersion has a similar meaning for

quantal data as variance heterogeneity has for continuous data. Failure to take account of either can affect a NOEC determination or ECx estimation. 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 more a function of the low control than of real effects. A subsequent study done at the request of a regulatory reviewer appears to substantiate this. The mean responses in these two lower 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 entire 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.

Case study 2

Case study 2 illustrates how a clear trend in the concentration–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.

An effort was made to apply standard statistical methods that require the data be normally distributed with homogeneous variances. The data collected did not meet that requirement and no transform of the response data were found that met those requirements. Consequently, a different analysis was done. The non-parametric JT test found all treatment mean responses significantly less than the control mean response. As Fig. 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 and 35 ppm). With those omitted, even the mean response at 35 ppm is within the HCD range. Setting the NOEC at 10 ppm, 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 ppm where there was only an 8% decrease is justified if one takes a 10% effect, the target of BMD10 estimation, as the cutoff for biological relevance. No acceptable regression model was found, as is common in studies with only 3 treatment groups in addition to the control.

Case study 3

Case study 3 illustrates that informal statistical reasoning can be misleading. As with other examples, the use of HCD 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 $\geq 10\%$ 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

Table 5 Case study 2

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

Quail proportion 14-day hatchling survivors/eggs set (H14DS_ES)

Count = number of replicates (breeding pairs), mean, median, and standard deviation are simple unweighted summary statistics; Conc = ppm

Table 6 Case study 3

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

Eggs hatched per eggs set (HATCH_ES)

Count = number of replicates (breeding pairs), mean, median, and standard deviation are simple unweighted summary statistics; Conc = ppm

a non-parametric analysis was indicated. There were only 6% and 1% decreases in the two treatment medians.

The response was Eggs Hatched per Eggs Set (HATCH_ES). The non-parametric Dunn and Jonckheere–Terpstra 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 shows 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.

Case study 4

Case study 4 illustrates a low-variability response (eggshell thickness) in which there is a sharp drop from relatively high control mean to a somewhat flat and non-monotone concentration–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.

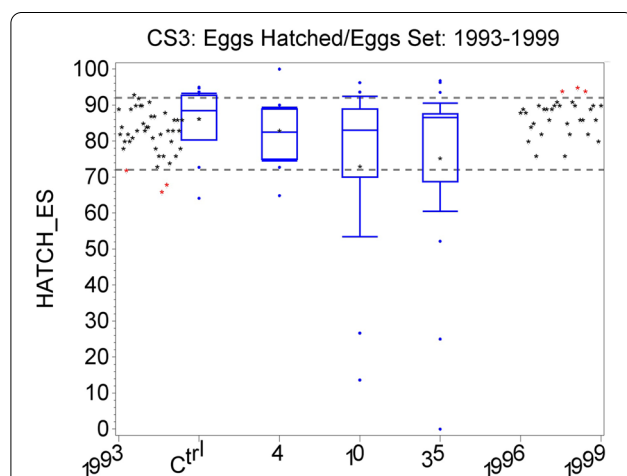


Fig. 3 Case study 3. Eggs hatched per eggs set (HATCH_ES), HCD data from 1993 to 1995 are displayed to the left of the concurrent study, while data from 1996 to 1999 are on the right. Two low outliers are evident at 10 and two more at 35 ppm. All treatment means and medians are within the HCD

Table 7 Case study 4

Group	Conc	Count	Mean	Median	Std
1	0	16	0.23	0.23	0.01
2	25	16	0.21	0.21	0.01
3	50	16	0.22	0.22	0.01
4	100	16	0.22	0.22	0.02

Eggshell thickness (ESThick)

Count = number of replicates (breeding pairs), mean, median, and standard deviation are simple unweighted summary statistics; Conc = ppm

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 Fig. 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

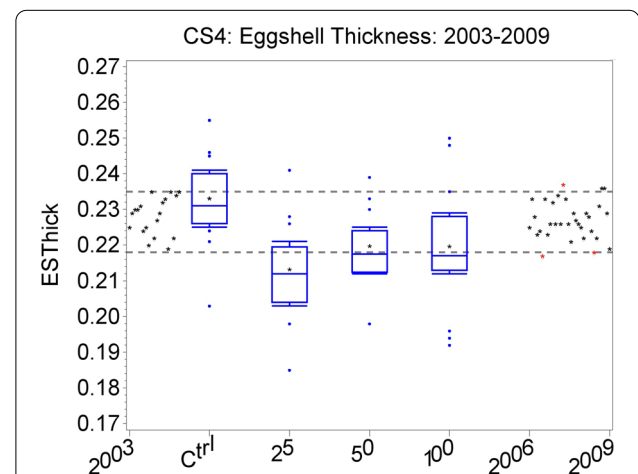


Fig. 4 Case study 4. Eggshell thickness (ESThick). HCD data from 1993 to 1995 are displayed to the left of the concurrent study, while data from 1996 to 1999 are on the right. One high outlier is evident in the control and two low outliers are evident at 100 ppm. The mean response at 25 ppm is below 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

to be biologically important. A decrease of less than 18% [7] or 22% [14] 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 responses. 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 concentration–response. However, MAXSD, the maximum safe dose analysis [4, 27–31] found significantly less than 10% effect in every treatment group, making MAXSD = 100. This means EC10LB > 100 ppm. Thus, if 10% is considered to be the minimum 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. A discussion of the MAXSD method is given in the Supplementary material as are more details for the application of this method to the current case study.

Case study 5

Case study 5 illustrates regression modeling 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 the proportion of 14-day survivors per eggs hatched 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.

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

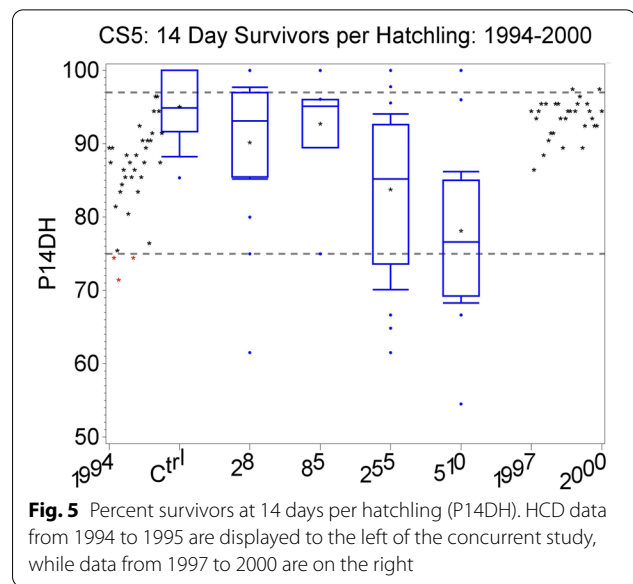


Fig. 5 Percent survivors at 14 days per hatchling (P14DH). HCD data from 1994 to 1995 are displayed to the left of the concurrent study, while data from 1997 to 2000 are on the right

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 for proportions or counts were fit to untransformed proportions or counts assuming normally distributed, homogeneous responses or 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 9 and 10 summarize the two modeling 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 10 summarizes approach 2 (GNLMM models for conditional counts). The BVP model parameters appear reasonable, but the estimated responses at positive test concentrations are in poor agreement with the observed data. 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}

Table 8 Case study 5

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

Percent 14-day survivors per hatchling (P14DH)

Count = number of replicates (breeding pairs), mean, median, and standard deviation are simple unweighted summary statistics; Conc = ppm

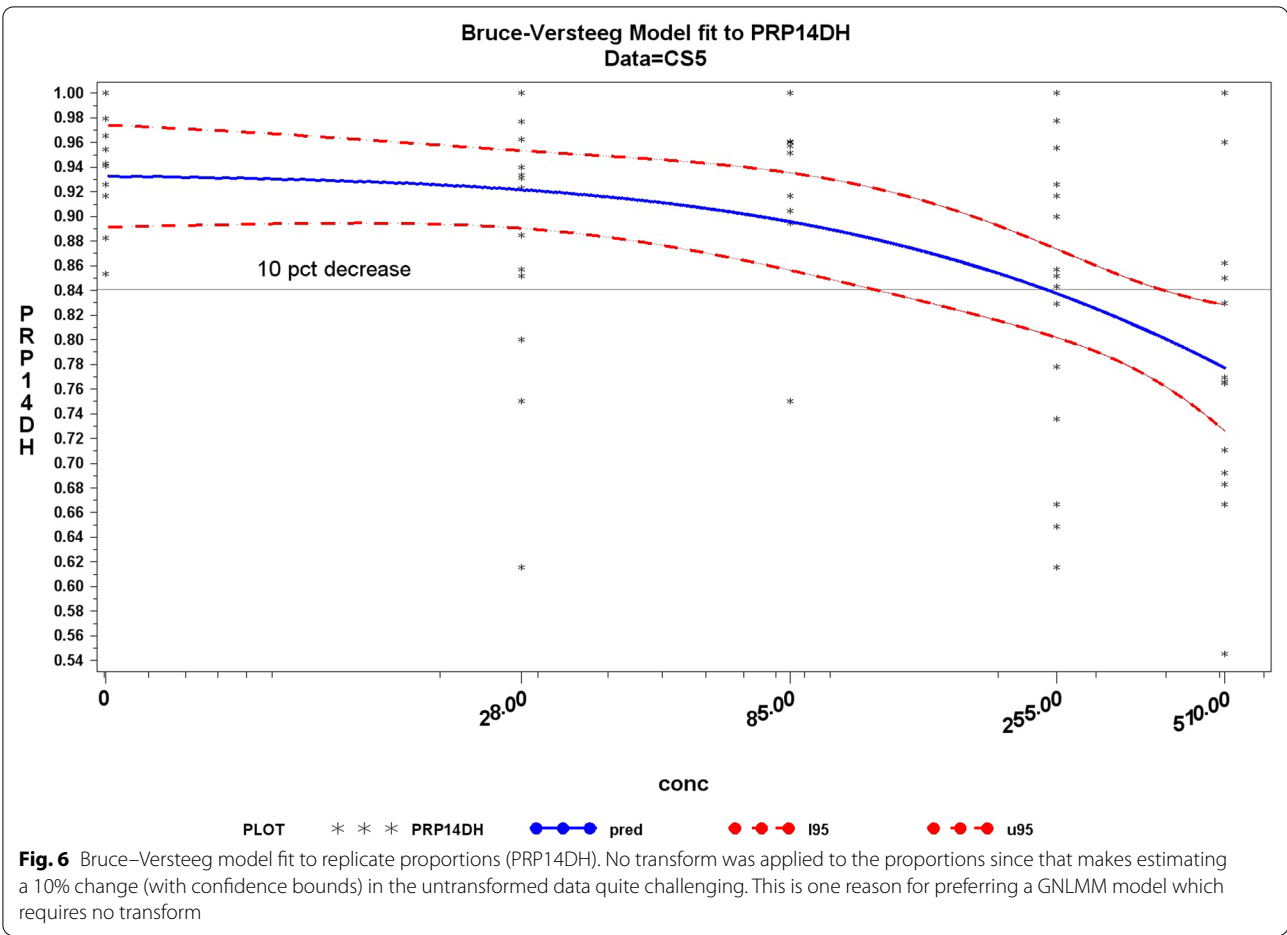


Table 9 Case study 5

Model	PARM	Estimate	LCB	UCB	AICc	Wgt
LL3	Y0	0.94	0.89	0.99	−313.49	0.12
	EC10	237.10	12.71	461.50		
	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		

Regression models for proportions

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 shape parameter, PARM = model parameter, AICc = Akiake information criterion with small sample correction, Wgt = Akaike weight, Estimate, LCB, UCB = point estimate, lower and upper 95% confidence bounds

Table 10 Case study 5

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

GNLMM regression models for counts conditioned on hatchlings

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 shape parameter 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

= 277, $EC10LB_{avg}$ = 183. A tighter lower bound reflects less uncertainty in this estimate compared to that of Approach 1.

Recommended statistical protocols

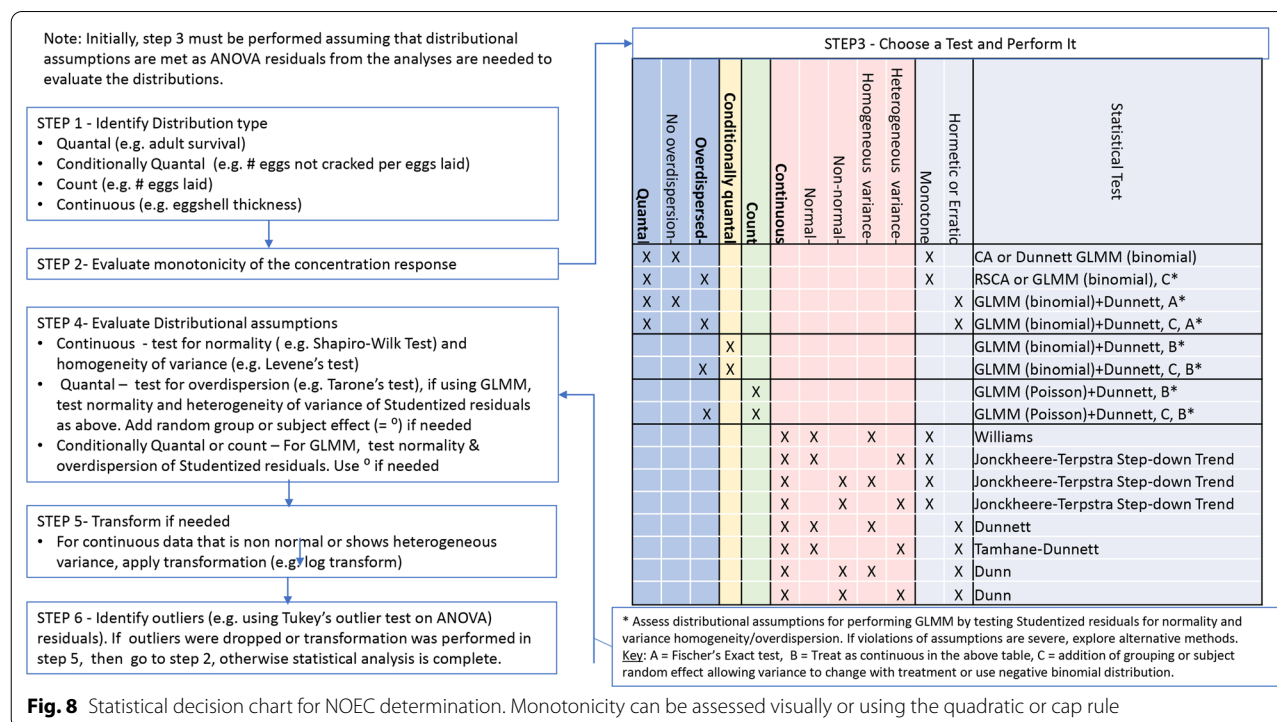
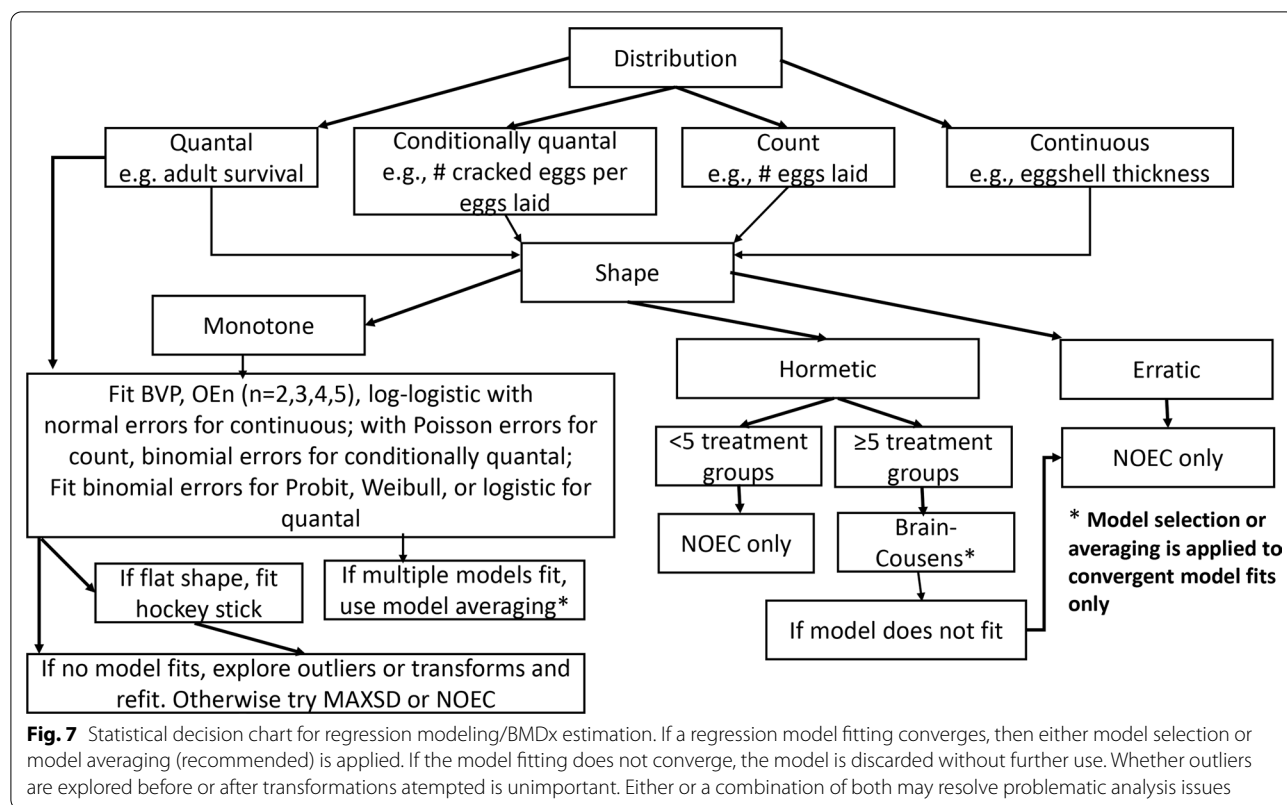
The case studies described above and in the supplementary material together with many combined years of experience with avian reproduction studies of the co-authors and the evolving regulatory requirements and statistical knowledge have led to a detailed statistical protocol which covers all types of responses currently required in regulatory test guidelines for avian reproduction studies. 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 eggs not cracked 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 not cracked eggs, or (d) a conditionally binomial response, such as number of not cracked eggs conditioned on the number eggs laid. A list of commonly reported responses with their distributions is given in Additional file 1.

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. [11]. 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 [35] and Green et al. [11] 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. [11] 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).

Note that when the response is transformed, the meaning of x in BMDx is changed and care is needed to specify the x after transform to correspond to x on original scale. With regard to the effect of outliers, both regression and pure error outliers can contribute to model instability or lack of fit. Plotting of data with model overlay is highly recommended. Heterogeneity / overdispersion can be handled by weighting or allowing different variances in different treatment groups.

Transform here includes the addition of model terms or weighting to adjust for overdispersion/heterogeneity.

Transforms, as used in the chart, include for all distributions indicated the possibility of adding a grouping or subject random variable that allows the within-treatment variance to differ across treatment groups. An alternative for non-continuous responses is to specify a negative binomial distribution.

For count data to be treated as continuous as an alternative in the chart, a square-root transform is usually needed to approximate a normal distribution. For quantal or conditionally quantal data to be treated as continuous, this usually requires analysis of replicate proportions and requires an arc-sine square-root transform to normalize the response and stabilize the variance.

For a continuous response following a monotone concentration–response, the step-down Jonckheere–Terpstra test can be used regardless of whether the variances are homogeneous. The power properties of that test are generally similar to those of Williams' test.

In Fig. 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., [16, 17] and its power properties are documented more completely (e.g., [11] and in documents supporting OECD test guidelines).

If there are outliers, only a small number should be omitted for re-analysis for outlier effect. 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.

Steps in the recommended statistical protocol. Additional details are 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 GLMMs, studentized residuals are used to assess agreement of the data to the modeled 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 $C(\alpha)$ 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 and for studentized residuals from GLMMs and GNLMs. 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 a

response such as 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 Figure 8 and in Additional file 1 indicate when consideration of outliers is applied. It should be noted that if the NOEC or BMDx 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 concentration–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 response is found statistically significant but higher dose responses 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. [11] and Springer and du Hoffmann [23]. 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. [11] contains further discussion of this, as does OECD TG 248 [15]. 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 quantal 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. [11], where additional references are also given. References deserving additional mention include OECD [16, 17]. The focus of the above discussion is on trend tests when the concentration–response is monotone. However, as indicated in the decision chart, the power properties of GLMMs with Dunnett's test are competitive with, and in some ways, superior to, these trend tests and should be considered.

4. Use historical control data if available

Valverde et al. [35] 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 HCD 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 [10] 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 where possible, approximately equally split on both sides of the concurrent study date. 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 EC_x, 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 GLMM 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. [11] for additional details and references on all the statistical recommendations.

If a transform removed non-normality or variance heterogeneity/overdispersion, then the results of the transformed data are generally preferred for NOEC determination. 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.

6. Use regression or BMD methodology where supported by data

Where sufficient 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 [2] 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 described in item 7) can be implemented. It is also important to understand the limitations of regression modeling. Once a model is fit to a dataset, it is mathematically possible to estimate EC_x 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.

Extrapolation beyond the observed range of test concentrations cannot be assessed merely in terms of the width of the confidence interval because confidence interval calculations assume the model is correct. Outside the observed range there is no basis for assuming the model fit to the data range describes the unobserved range. There is nothing novel in that view. Problems with low and high dose extrapolation have been well reported in the scientific literature.

Also, a reliable estimate of EC_x for $x < 10$ is often beyond the capability of the data. For example, obtaining a meaningful estimate of 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 details on this are given in the Supplement under the heading of model fitting criteria.

7. Use Model Averaging where possible for BMD_x calculations

When a study is repeated under the same conditions, the data are different, results from statistical tests are often different, and different models are often fit to the same response. For NOECs, it is not uncommon for them to differ by an order of magnitude between such studies. This has been observed many times in developing test guidelines where inter-laboratory studies are conducted. Differences in EC_x or BMD_x of such magnitude are also found. The real relationship between the concentrations of a test substance and measured responses to it is unknown. Models are our attempts to determine that relationship but under the best of circumstances, they can fall short. Model averaging is an attempt to take such model uncertainty into account. Confidence intervals for model parameters or predictions do not capture model uncertainty because they assume that the model is correct.

A consistent set of models appropriate for the type of data are defined. Such a set of models is described in "Models used for BMD_x estimation" section. That set is intended to cover all the usual general shapes likely

to be encountered for several categories of responses. This set of models is fit to the data. Some model convergence or goodness of fit criteria may rule out one or more. The other models then are used in an objective weighting scheme to arrive at what are called model average estimates of BMDx and BMDLx.

There are two main ways to approach model averaging. Benchmark dose (BMD) methodology outlined in [6, 7] 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 [8] to use a combination of bootstrap sampling and weighted averages. The discussions in [3, 33, 34] also contain valuable insight into model averaging. 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.

8. 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.

9. Consider an alternative to NOEC and BMD.

For BMD estimation, a 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.

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 [25, 26]; Proast [21, 22] which was developed specifically for regulatory risk assessment under the auspices of RIVM, BBMD [19, 20] which provides a Bayesian implementation. Also notable is the BMD software developed by the United States EPA (<https://www.epa.gov/bmds>) 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 Akaike 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 BMDLx 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 [13], which was developed for the United States EPA, implements all the standard statistical tests recommended, but not the GLMM tests. The R package PMCMRplus [18] 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 [12].

Biological relevance

Real improvement in hazard identification and risk assessment requires scientifically based criteria for what constitutes a hazard. According to [7], in determining a NOAEL there may not be a consideration of the effect or its biological relevance. Therefore, it is proposed to use responses 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 response 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. Despite the statistical significance of the decreases in all treatment groups, it should be noted that the maximum observed decrease was only 6%. This response is a rare instance

when scientific evidence is available for this purpose. According to [7] and Lincer [14] population effects in the wild tend to come about after thinning of 18% or more. 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, Fig. 4).

The regulatory process would be much enhanced by developing data-based estimates of the sizes of effects on key biological measures that result in biologically meaningful consequences such as population decline. As it is, largely arbitrary effect sizes, such as 10% change, or a statistically significant change, are assumed to represent the demarcation between acceptable and unacceptable effects, regardless of biological importance.

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 [8] 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 responses as does careful attention to identifying outliers. BMDx estimation can be improved using revised modeling 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 responses that indicate population level effects and distinguish such effects from mere statistical significance or a percentage change from control without associated biological significance.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12302-022-00603-5>.

Additional file 1. Statistical Analysis of Avian Reproduction Studies - Supplementary Material.

Acknowledgements

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, MF 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. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹John W Green Ecostatistical Consulting LLC, Newark, DE, USA. ²Corteva™ Agriscience, 152 32 Halandri, Greece. ³Bayer U.S. Crop Science, Chesterfield, MO, USA. ⁴FMC Corporation, 2929 Walnut Street, Philadelphia, PA, USA. ⁵Syngenta Crop Protection LLC, Greensboro, NC, USA. ⁶BASF Corporation, 26 Davis Dr, Research Triangle Park, NC, USA. ⁷Corteva™ Agriscience, Indianapolis, IN, USA. ⁸Corteva™ Agriscience, Wilmington, DE, USA. ⁹Bayer AG, Monheim, Germany.

Received: 17 November 2021 Accepted: 19 February 2022

Published online: 24 March 2022

References

1. BMDs 2020. BMDs 3.2 User Guide (epa.gov) User Guide (epa.gov). https://www.epa.gov/sites/default/files/2020-09/documents/bmds_3.2_user_guide.pdf
2. Burnham KP, Anderson DR (1998) Model Selection and Multimodel Inference A Practical Information-Theoretic Approach. Second. edition. Springer-Verlag New York, inc. ISBN 0-387-95364-7
3. Cade BS (2015) Model averaging and muddled multimodel inferences. *Ecology* 96:2370–2382. <https://doi.org/10.1890/14-1639.1>
4. Dunnett CW (1989) Algorithm AS 251: Multivariate Normal Probability Integrals with Product Correlation Structure. *J R Stat Soc Series C (Applied Statistics)* 38:564–579
5. Duquesne S, Alalouni U, Gräff T, Frische T, Pieper S, Egerer S, Gergs R, Wogram J (2020) Better define beta—optimizing MDD (minimum detectable difference) when interpreting treatment-related effects of pesticides in semi-field and field studies. *Environ Sci Pollut Res* 2020(27):8814–8821. <https://doi.org/10.1007/s11356-020-07761-0>

6. EFSA (2009) 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
7. EFSA (2009) Risk Assessment for Birds and Mammals. *EFSA J* 7(12):1438
8. EFSA (2017) Update: use of the benchmark dose approach in risk assessment. *EFSA J* 15(1):4658
9. 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
10. Fleiss JL, Levin B, Paik M (2003) *Statistical Methods for Rates and Proportions*, third edition. Wiley. ISBN 0–471–52629–0. Hoboken, NJ
11. Green JW, Springer TA, Holbech H (2018) *Statistical Analysis of Ecotoxicity Studies*. Wiley. ISBN: 978–1–119–48881–1
12. Hothorn, L.A. 2018. *Statistics in Toxicology Using R*. CRC Press. ISBN-13 number 978–1–4987–0127–3.
13. Ives MA (2021) Comprehensive Environmental Toxicity Information System (CETIS), Version 2.0. Tidepool Scientific LLC, McKinleyville, CA. <http://www.tidepool-scientific.com/Cetis/CetisStats.html>
14. Lincer JL (1975) DDE-induced eggshell-thinning in the American kestrel: a comparison of the field situation and laboratory results. *J Appl Ecol*. 12:781
15. 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>.
16. 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
17. OECD (2014) Fish Toxicity Testing Framework. OECD Publishing, Paris
18. .PMCMRPlus 2021.PMCMRplus.pdf (r-project.org). PMCMRPlus 2021. PMCMRplus.pdf (r-project.org)., 2021 PMCMRPlus 2021.PMCMRplus.pdf (r-project.org)
19. Shao K, Shapiro AJ (2018) A Web-Based System for Bayesian Benchmark Dose Estimation. *Environ. Health Perspect.* 126. CID: 017002 <https://doi.org/10.1289/EHP1289>
20. Shao K (2021) Bayesian BMD (benchmarkdose.org)
21. Slob W (2019) PROAST. A general software tool for dose-response modelling. RIVM, Bilthoven PROAST MANUAL GUI version.pdf (rivm.nl)
22. Slob W (2018) Joint project on Benchmark Dose modelling with RIVM. EFSA supporting publication 2018:EN-1497. 14 pp. doi:<https://doi.org/10.2903/sp.efsa.2018.EN-1497> ISBN: 2397–8325
23. Springer TA, du Hoffmann G (2018) Evaluation of Monotonicity of Concentration Response in Avian Reproduction Studies – Summary of Findings. Project 857B-101. Eurofins Laboratory
24. Staveley JP, Green JW, Nusz J, Edwards D, Henry K, Kern M, Deines AM, 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
25. Ritz C, Jensen SG, Gerhard D, Streibig JC (2020) *Dose-response analysis using R*. CRC Press, Boca Raton, FL
26. Ritz C, Streibig JC (2016) [drc.pdf \(r-project.org\)](https://r-project.org/doc/drc.pdf)
27. Schervish MJ (1984) Algorithm AS 195: Multivariate Normal Probabilities with Error Bound. *J R Stat Soc Series C (Applied Statistics)* 33:81–94
28. Tamhane AC, Logan BR (2004) Finding the maximum safe dose level for heteroscedastic data. *J Biopharm Stat* 14:843–856
29. Tamhane AC, Logan BR (2002) Multiple Test Procedures for Identifying the Minimum Effective and Maximum Safe Doses of a Drug. *JASA* 97:293–301
30. Tamhane AC, Dunnett CW, Green JW, Weatherington JD (2001) Multiple Test Procedures for Identifying the Maximum Safe Dose. *JASA* 96:835–843
31. Tarone RE, Gart JJ (1980) On the robustness of combined tests for trends in proportions. *JASA* 75:110–116
32. Tukey JW (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA
33. USEPA 2015. https://www.epa.gov/sites/production/files/2015-11/documents/support_material_for_model_averaging_workshop-11_06_2015-508.pdf
34. USEPA 2021. <https://www.epa.gov/bmds>.
35. Valverde-Garcia P, Springer T, Kramer V, Foudoulakis M, Wheeler JR (2018) An avian reproduction study historical control database: A tool for data interpretation. *Regul Toxicol Pharmacol* 92:295–302

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)