Published in final edited form as:

Biometrics. 2011 September ; 67(3): 1153–1162. doi:10.1111/j.1541-0420.2011.01566.x.

Multilevel empirical Bayes modeling for improved estimation of toxicant formulations to suppress parasitic sea lamprey in the upper Great Lakes

Laura A. Hatfield¹, Steve Gutreuter², Michael A. Boogaard², and Bradley P. Carlin^{1,*}

¹Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

² U.S. Geological Survey, Upper Midwest Environmental Sciences Center, La Crosse, WI 54603

Summary

Estimation of extreme quantal-response statistics, such as the concentration required to kill 99.9% of test subjects (LC99.9), remains a challenge in the presence of multiple covariates and complex study designs. Accurate and precise estimates of the LC99.9 for mixtures of toxicants is critical to ongoing control of a parasitic invasive species, the sea lamprey, in the Laurentian Great Lakes of North America. The toxicity of those chemicals is affected by local and temporal variations in water chemistry, which must be incorporated into the modeling. We develop multilevel empirical Bayes models for data from multiple laboratory studies. Our approach yields more accurate and precise estimation of the LC99.9 compared to alternative models considered. This study demonstrates that properly incorporating hierarchical structure in laboratory data yields better estimates of LC99.9 stream treatment values that are critical to larvae control in the field. In addition, out-of-sample prediction of the results of *in situ* tests reveals the presence of a latent seasonal effect not manifest in the laboratory studies, suggesting avenues for future study and illustrating the importance of dual consideration of both experimental and observational data.

Keywords

Lethal concentration/dose; Markov chain Monte Carlo (MCMC); Non-linear model; Quantal-response bioassay

1. Introduction

Invasive alien species cause damages estimated at \$120 billion per year in the United States (Pimentel et al., 2005). Most are resistant to post-invasion control. A notable exception is the sea lamprey, a parasitic jawless fish that invaded the upper Great Lakes (Superior, Huron, Michigan and Erie) during the 1930's after shipping canals and locks bypassed Niagara Falls. In concert with commercial fishing, parasitism by adult sea lampreys eliminated or nearly eliminated many large-bodied native fishes from the upper Great Lakes during the middle of the 20th century (Smith, 1968). However, sea lampreys have a vulnerable non-parasitic larval stage, which has enabled ongoing population control using

* brad@biostat.umn.edu .

Supplementary Materials

The Web Appendix A referenced in Section 3.3 and the Web Appendix B and Web Figure 1 referenced in Section 6 are available under the Paper Information link at the Biometrics website, <http://www.biometrics.tibs.org>.

selective toxicants and successful reintroduction of large-bodied fishes, including recovery of native lake trout in Lake Superior (<http://www.glfc.org/lampcon.php>).

Sea lamprey control depends on periodic administration of formulations of the selective toxicants 3-trifluoromethyl-4-nitrophenol (TFM) and 2',5'-dichloro-4'-nitrosalicylanilide (niclosamide) to the tributary streams where larval sea lampreys reside before maturing and migrating into the Great Lakes (Smith and Tibbles, 1980). Those stream treatments are performed during May through October on a three- or four-year rotating schedule. The prescription of toxicant application rates requires care because over-application may kill non-target species, while under-application wastes labor and risks increased parasitism on native species. Treatment requirements vary among and within tributaries due to spatial and temporal variations in stream water chemistry. For example, the toxicity of TFM decreases as pH increases from 7 to 9 and also decreases with increasing alkalinity (Hunn and Allen, 1974; Bills et al., 2003). Additions of niclosamide, measured as the fraction of the mass concentration of TFM, are used to boost toxicity with increasing pH and alkalinity.

Prescriptions for stream treatment are based on the concentration (mg/l) of TFM required to kill 99.9% of the sea lampreys, denoted LC99.9. The estimates of this concentration are currently based on data from a set of toxicity tests that were conducted from 1989–90 (Bills et al., 2003). They derive from a two-stage approach prevalent in environmental toxicology (Gelderen et al., 2003) wherein LC99.9 estimates are obtained from unique combinations of pH and alkalinity. Those estimates are subsequently treated as observed data in regressions on pH and alkalinity. The effect of niclosamide is then included as a simple multiplier based on unpublished data. These LC99.9 estimates are arranged into streamside treatment tables that are used by treatment program staff in the field to try to achieve lamprey control targets.

This two-stage approach appears to lack both parsimony and coherence. Gutreuter and Boogaard (2007) (G&B) proposed a more parsimonious class of nonlinear models for the estimation of lethal concentration in the presence of covariates. Those authors showed that nonlinear, single-stage, mixed-effect models can yield substantially smaller root mean squared errors (RMSEs) and smaller bias than the conventional two-stage approach. Yet this method is still limited by the exclusion of relevant data.

In this article, we develop nonlinear hierarchical and empirical Bayes models that use historical data to inform a prior distribution and update it via contemporary data to produce a final posterior distribution. A primary motivation for this work is to create tables of LC99.9 values at various niclosamide augmentation rates that are tailored for pH and alkalinity. We also desire credible intervals for these LC99.9 estimates to inform *ad hoc* adjustments on the basis of expertise. To improve upon previous stream treatment tables, we incorporate all of the relevant data and use models that properly account for the clustered nature of the laboratory data. Including cluster-level parameters is a natural approach to modeling hierarchical data, though we ensure that global parameters completely determine quantities needed in the field, so they will be broadly applicable.

Three distinct phases of data analysis are presented. First, the *prior specification phase* uses studies conducted in 1989–90, previously reported by Bills et al. (2003), and two unpublished data sets from 1997 and 1998. Next, the *fitting phase* uses studies conducted in 2001–3 (Gutreuter and Boogaard, 2007) and 2009. Finally, in the *validation phase*, we consider streamside data collected during 1993–2008, in which we discover strong seasonal trends not previously documented. To perform out-of-sample validation of our methods, we seek a suitable subset comparable to the laboratory data.

The remainder of our paper proceeds as follows. In Section 2, we provide a careful description of the laboratory and streamside tests. Section 3 then presents the model we

advocate, as well as a few potentially attractive competitors. Section 4 gives the results of our analysis, including both our LC99.9 estimates from laboratory data and streamside differences. Section 5 offers a brief simulation study to further validate our choice of model, as well as sensitivity analyses to explore the effect of modeling assumptions. Finally, Section 6 discusses our findings and suggests avenues for further research.

2. Laboratory and Streamside Methods

All toxicity testing was conducted according to established protocols (American Society for Testing and Materials, 2000) using a modification of the continuous-flow serial-dilution apparatus described in detail by Bills et al. (2003). In those tests where their values were controlled, pH was regulated by addition of hydrochloric acid (HCl) or sodium hydroxide (NaOH) and alkalinity (mg/L bicarbonate equivalents) was regulated by addition of sodium bicarbonate. Toxicant concentrations were verified by *in situ* spectrophotometric (TFM) and high performance liquid chromatographic (niclosamide) measurements. Measurement error was relatively small; pH varied by 0.05 units, alkalinity by 5 mg/L, and TFM concentrations within 4% of the nominal values.

Laboratory Studies

Ten (or occasionally more) larval sea lamprey were placed in each tank and acclimated for 1–2 days prior to testing. During testing, toxicant was delivered continuously for a duration of 12 hours. The numbers of dead larvae were counted at the end of the testing.

The sample space covered by the laboratory studies is shown in Figure 1. The variables are shown on the transformed scales used in the model fitting, i.e., niclosamide (as % of TFM) and pH are centered around their means (0.53 and 8, respectively) and alkalinity is standardized (by its mean 147 and standard error 69), to put it on a scale similar to that of the other variables. This figure clearly shows the narrow ranges of covariates and collinearity in the historical data sets; see, for instance, the tight clustering of values for alkalinity in the data from 1997 and 1998.

The 1989–90 study used only TFM without niclosamide supplementation, at controlled values of both pH and alkalinity. The unpublished 1997 and 1998 studies and the 2001–03 study used both TFM and niclosamide at controlled pH values but did not control alkalinity, which varied within relatively narrow ranges among studies. The 2001–03 study was similar in design to the 1997 and 1998 studies except that the ranges of niclosamide augmentation rates, pH and alkalinity were broader. The 2009 study was designed to fill a gap in existing data: it is the only study that simultaneously tested varying TFM and niclosamide formulations at chosen combinations of both pH and alkalinity.

Hierarchical structure includes *clusters* of tanks in which testing was conducted at a particular place and time. All tanks in a cluster share values of pH, alkalinity, and niclosamide augmentation rate, but represent a gradient of TFM doses. The number of units at each level of the data hierarchy are summarized in the legend of Figure 1. Although more animals are used for prior specification than model fitting, only the contemporary 2001–03 and 2009 data sets are suitable for our purposes, given the design considerations above.

Streamside Data

We consider data collected during the course of stream treatments, which were conducted during the months from May through October during 1993–2008. Streamside toxicity tests were performed solely to verify requirements just prior to treatment. Those tests were conducted in mobile laboratories equipped with the two-sided serial diluters using methods identical to the laboratory tests, but used newly captured sea lampreys and the more

chemically complex and variable waters from adjacent streams. Observed pH values spanned 6.8–8.8 and alkalinity 6–268. Streamside data from 145 treatments (clusters) comprising 12,341 larvae were made available to us by the stream treatment crews.

Those streamside tests contain previously unexploited information about the applicability of the laboratory studies to stream-treatment performance. In Section 4.2, we document for the first time strong seasonal trends in streamside data not observed in the laboratory data. Thus we consider only the tests conducted in May and June suitable for model validation.

3. Statistical Models

The development of our models addresses three common challenges in hierarchical modeling. First, we wish to incorporate historical data arising from less-than-ideal designs; we do this by downweighting their contribution to the final analysis. Second, it is often difficult to estimate variance-covariance matrices for random effects, such as our cluster-level parameters (Stiratelli et al., 1984). Here we investigate the use of cluster-level variation in the historical data to empirically estimate the covariance matrix; we also fit a model in which the cluster-level effects are fixed. Finally, although we wish to incorporate the hierarchical structure of the data, we are ultimately interested not in cluster-level quantities, but in constructing dose requirements from global parameters. For this reason, we also consider an empirical Bayes approach wherein cluster-level parameters are treated as fixed once estimated. The three subsections below address these points in turn.

Although significant progress has been made on the fitting of Gaussian nonlinear mixed models (Davidian and Giltinan, 1995), extending the framework to generalized (non-Gaussian) nonlinear models remains challenging. We model our animal-level outcomes as Bernoulli trials, where we assume “successes” (deaths) in each tank are conditionally independent given each tank's characteristics; see e.g. equation (1) below. Our observed binomial counts are most frequently at the extremes (i.e., all successes or all failures). This sparse information challenges our ability to fit a nonlinear model, to say nothing of estimating random effects and the parameters controlling their distribution.

3.1 From Historical Data to Informative Priors

Let Y = number of dead animals, X = TFM, Z_1 = niclosamide, Z_2 = pH, and Z_3 = alkalinity. We then write a vague prior fixed effects model as

$$\begin{aligned} y_{ij} &\sim \text{Bin}(n_{ij}, p_{ij}) \\ \text{logit}(p_{ij}) &= \mu_1 + \frac{\mu_2 X_{ij}}{\exp(\mu_3 Z_{1ij} + \mu_4 Z_{2ij} + \mu_5 Z_{3ij} + \mu_6 Z_{1ij}^2 + \mu_7 Z_{2ij}^2 + \mu_8 Z_{3ij}^2)} \\ \mu_\ell &\stackrel{\text{ind}}{\sim} N(0, 100), \ell = 1, \dots, 8, \end{aligned} \quad (1)$$

where $i = 1, \dots, I_j$ indexes tank within cluster $j = 1, \dots, J$ and n_{ij} is the number of larvae in the $(i, j)^{\text{th}}$ tank (usually 10). The prior variance of the μ_ℓ (100) was chosen to be large enough not to influence the posterior distributions but still ensure good MCMC performance. The vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_8)^T$ contains the population-level mean structure parameters of interest. We suppose its joint posterior distribution has an estimable mean and covariance matrix that we denote by $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}_\mu$ respectively.

The functional form of our model was informed by investigations (Gutreuter and Boogaard, 2007) into the relationships among toxicant doses, covariate values, and outcomes. We assume that the effects of the covariates (pH and alkalinity) and toxicants (TFM and niclosamide) affect the probability of death via a logit link, though complementary log-log

and probit were also explored by these authors. Neither pH nor alkalinity are toxic within the observed ranges (pH: 6.5–9.5; alkalinity: 6–270), so to preclude influencing the probability of death when no TFM is administered, we use multiplicative terms for the effects of pH and alkalinity. These two covariates are included via an exponential function so that negative values (i.e., reversing the dose-response toxicity effect of TFM) are precluded. Niclosamide is added as a percentage of the mass concentration of TFM, so the effect on probability of death is multiplicative by design. Because of the non-linear form of the model and the scientific importance of the zero value for TFM dose, we do not center the TFM variable. One regrettable consequence of this is strong negative correlation between the parameters for intercept and TFM.

Gutreuter and Boogaard (2007) found that including quadratic niclosamide and pH effects improved model fit as measured by AIC (Akaike, 1973). Preliminary modeling (not shown) demonstrated that additive models containing these covariates provided substantially worse fit than the multiplicative exponential models. Given these considerations, we utilize the same functional form that provided superior fit to the 2001–2003 data among the alternatives evaluated by Gutreuter and Boogaard (2007).

3.2 Estimating Cluster-level Parameter Covariances

Our definition of cluster leads to constant values of pH, alkalinity, and niclosamide for all tanks in a cluster, so that only cluster-specific TFM and intercept terms may be estimated from these data. We would like to estimate the covariance matrix for the joint distribution of the cluster-level parameters. However, a mixed model that replaces μ_1 and μ_2 in model (1) by θ_{1j} and θ_{2j} (following a bivariate normal distribution) encounters numerical problems in WinBUGS and OpenBUGS (Lunn et al., 2009). Thus we modify (1) to obtain an overparameterized hierarchical model having cluster-level effects $\theta_j = (\theta_{1j}, \dots, \theta_{8j})^T$:

$$\begin{aligned} \text{logit}(p_{ij}) &= \theta_{1j} + \frac{\theta_{2j}X_{ij}}{\exp(\theta_{3j}Z_{1ij} + \theta_{4j}Z_{2ij} + \theta_{5j}Z_{3ij} + \theta_{6j}Z_{1ij}^2 + \theta_{7j}Z_{2ij}^2 + \theta_{8j}Z_{3ij}^2)} \\ \theta_j &\sim N(\mu, \Sigma), \Sigma^{-1} \sim \text{Wishart}((\rho\mathbf{R})^{-1}, \rho), \mu \sim N(\tilde{\mu}, c_\mu\tilde{\Sigma}_\mu), \end{aligned} \quad (2)$$

where R is an identity matrix of dimension 8 and ρ is chosen to make the Wishart as vague as possible, i.e., $\rho = \dim(R) = 8$. We call this model “fully Bayesian.”

Notice the multivariate normal prior distribution on μ having mean $\tilde{\mu}$ and covariance matrix that is a scaled version of $\tilde{\Sigma}_\mu$. The scaling factor allows appropriate down-weighting of the prior historical results since model (1) fails to account for expected clustering in the data. We expect the data to be positively correlated within cluster and study, but model (1) ignores this, producing posterior distributions with variances biased toward zero.

Because covariance matrices in such models are notoriously difficult to estimate (Yang and Berger, 1994), we also consider using observed variation among cluster-level parameters in the historical data to estimate Σ . Taking the posterior means of the fixed effects, $\tilde{\mu}_3, \dots, \tilde{\mu}_8$, from fitting (1) to the 1989–1998 data, we treat these as known in a set of J ordinary generalized linear regression models. For the j^{th} cluster in the 1989–1998 data, we fit a binomial model using

$$\text{logit}(p_{ij}) = \beta_{1j} + \frac{\beta_{2j}X_{ij}}{\exp(\tilde{\mu}_3Z_{1ij} + \tilde{\mu}_4Z_{2ij} + \tilde{\mu}_5Z_{3ij} + \tilde{\mu}_6Z_{1ij}^2 + \tilde{\mu}_7Z_{2ij}^2 + \tilde{\mu}_8Z_{3ij}^2)}, \quad (3)$$

and save the resulting parameter estimates, $\widehat{\beta}_{1j}$ and $\widehat{\beta}_{2j}$. We then compute the empirical covariance matrix of zero-centered versions of these, $\widehat{\beta}_{1j} - \text{avg}(\widehat{\beta}_{1j})$ and $\widehat{\beta}_{2j} - \text{avg}(\widehat{\beta}_{2j})$, to obtain $\widehat{\Sigma}$, which we use below.

3.3 Empirical Bayes

Finally, we develop an empirical Bayes (EB) approach in which the cluster-specific parameters are estimated and then treated as known. Their estimation relies on a covariance matrix informed by the historical data, as described above. Specifically, we augment the 2×2 $\widehat{\Sigma}$ matrix to be 8×8 , $\Sigma^* = \text{BlockDiag}(\widehat{\Sigma}, I_6)$. Then replacing Σ in the “fully Bayesian” model by fixed Σ^* , we obtain a model we term “fixed sigma”:

$$\begin{aligned} \text{logit}(p_{ij}) &= \theta_{1j} + \frac{\theta_{2j}X_{ij}}{\exp(\theta_{3j}Z_{1ij} + \theta_{4j}Z_{2ij} + \theta_{5j}Z_{3ij} + \theta_{6j}Z_{1ij}^2 + \theta_{7j}Z_{2ij}^2 + \theta_{8j}Z_{3ij}^2)} \\ \theta_j &\sim N(\mu, \Sigma^*), \mu \sim N(\tilde{\mu}, c_\mu \tilde{\Sigma}_\mu). \end{aligned} \quad (4)$$

The cluster-specific parameters $\eta_{1j} = \theta_{1j} - \mu_1$ and $\eta_{2j} = \theta_{2j} - \mu_2$ are of primary interest. We take their posterior means $\widehat{\eta}_{1j}$ and $\widehat{\eta}_{2j}$ as fixed in an EB model fit to 2001–2009 data,

$$\text{logit}(p_{ij}) = (\widehat{\eta}_{1j} + \mu_1) + \frac{(\widehat{\eta}_{2j} + \mu_2)X_{ij}}{\exp(\mu_3Z_{1ij} + \mu_4Z_{2ij} + \mu_5Z_{3ij} + \mu_6Z_{1ij}^2 + \mu_7Z_{2ij}^2 + \mu_8Z_{3ij}^2)} \quad (5)$$

where $\mu \sim N(\tilde{\mu}, c_\mu \tilde{\Sigma}_\mu)$. From this model, we estimate means and credible intervals for LC99.9 at covariate combinations of interest. To compute LC99.9 values, fix $p = .999$ in (5) and solve for X :

$$\text{LC99.9} = \frac{\exp(\mu_3Z_1 + \mu_4Z_2 + \mu_5Z_3 + \mu_6Z_1^2 + \mu_7Z_2^2 + \mu_8Z_3^2)}{\mu_2} (\text{logit}(.999) - \mu_1). \quad (6)$$

LC99.9 depends only on μ so the resultant estimates are applicable in the field, where only global parameters are of interest. However, inclusion of the fixed cluster-specific $\widehat{\eta}_{1j}$ and $\widehat{\eta}_{2j}$ in model (5) ensure that expected correlation within the fitting data (2001–2009) is incorporated. Our Section 5 simulation studies show that there is an advantage to the EB model over a fixed effects version of (5) and the two-stage method.

The above models were fit using the R2WinBUGS (Sturtz et al., 2005) package for R v2.11.0, which calls WinBUGS v1.4 (Lunn et al., 2000), freely available at www.mrc-bsu.cam.ac.uk/bugs. The results below reflect 5000 iterations of burn-in followed by 5000 production samples from each of three parallel chains. Convergence was judged adequate via trace plots, auto- and cross-correlations, density plots, and Brooks-Gelman-Rubin statistics (Gelman and Rubin, 1992; Brooks and Gelman, 1997) of each chain. See Web Appendix A for the BUGS code to fit this model. The G&B and two-stage models were fit using PROC NLMIXED in SAS (Cary, NC).

4. Results

4.1 Historical and Contemporary Laboratory Results

The results from the historical data (not shown) indicate that concentration of TFM is positively related to mortality. The observed negative coefficient for niclosamide and the structure of the model indicate that an increase in niclosamide corresponds to an exponential multiplicative increase in the TFM effect. Conversely, pH and alkalinity have significantly positive coefficients, meaning exponential multiplicative decreases in the TFM effect. The quadratic terms indicate that the pH effect diminishes at high pH, and the alkalinity effect starts out moderate and drops off less dramatically. The empirical estimates of elements of Σ^* are 5.97 for the intercepts and 4.59 for the slopes, with extremely strong correlation of -0.97 . Using this in the fixed Sigma model (4), we obtain cluster-specific $\hat{\eta}_{1j}$ and $\hat{\eta}_{2j}$ estimates used below. A plot of these cluster-specific parameters over time (not shown) reveals no clear trends from the earlier to later studies.

Fitting the EB model for values of $c_\mu \in \{1^2, 2^2, 4^2, 10^2, 20^2\}$ for 9 different pH-alkalinity pairs and a single niclosamide augmentation rate yields LC99.9 posterior medians and credible intervals shown in Figure 2. The values in this figure are those desired for the stream treatment tables, but at coarser resolution. Increasing the value of c_μ leads to slight widening of the credible intervals and changes in the point estimates, but these appear to taper off near $c_\mu = 10^2$, so we use this value in the final model fitting. We also fit the G&B likelihood-based model to either the full (1989–2009) or contemporary-only (2001–2009) laboratory data and display point estimates and asymptotic confidence intervals in Figure 2.

In most of the cases where the two G&B estimates differ, the complete data estimate (G&B '89-'09) is more similar to the EB estimate that also uses the historical data at full weight (EB $c_\mu = 1$), while the contemporary data-only estimate (G&B '01-'09) more closely resembles the EB estimates that use downweighted historical data (EB $c_\mu > 1$). The odd departure of the EB estimates from G&B when alkalinity = -1 (left column, Figure 2) may be due to the historical data's failure to include alkalinity values in this range (bottom row, Figure 1). The only historical clusters that include such values are a few from the 1989–90 study.

This ability to tailor the incorporation of historical data is one of the advantages of the EB approach that lead us to prefer it to G&B, despite similar frequentist operating characteristics in Section 5. Available software to fit G&B (i.e., PROC NLMIXED in SAS) was unable to fit more than random intercepts in these data, likely due to the computational burden of numerical integration. In contrast, the Bayesian hierarchical models including EB are limited to random intercepts and TFM slopes only by the experimental design. If more within-cluster combinations of factors were observed, we could easily extend the method to accommodate further cluster-specific effects, as evidenced by our ability to fit the overparameterized model (2). In addition, the Wald-type confidence intervals of G&B suffer from the usual problem of having a long-run frequency interpretation that may be misinterpreted by end users, and may be inaccurate when the corresponding likelihood departs markedly from normality.

The effect of alkalinity on the toxicity of TFM has been debated because it is often correlated with pH but, unlike pH, a mechanism for the effect remains unknown. Our hierarchical model results indicate that there is a DIC (Spiegelhalter et al., 2002) advantage of 199 for a model with alkalinity (DIC = 901, $p_D = 7.97$) versus one without (DIC = 1100, $p_D = 6.05$), strong evidence for ending the debate about its statistical importance.

4.2 Streamside Prediction Results

By fitting the EB model simultaneously with covariate values for the streamside data, we can obtain posterior predictive distributions for the proportion of animals killed. Most of the observed tanks had either zero or all animals killed, and many of the posterior predictive distributions were point masses at the extremes or bimodal with modes at the extremes.

Conditional on the parameters, we assume that the data come from a binomial distribution, which is unimodal. The posterior predictive distributions, however, are marginalized over the posteriors of the parameters and therefore need not resemble single binomials. The observed bimodality makes the mean a poor point predictor of the outcome; instead, we consider the posterior median (also not ideal) and the posterior probabilities that $Y = 0$ and $Y = n$.

We compute leave-one-out residuals for the 2001–2003 and 2009 data. To do this, we fit the hierarchical model 441 times, each time omitting one tank's outcome and producing a posterior predictive density for that excluded point. Again treating the posterior median as the point estimate, we can examine residuals for p_{ij} .

Preliminary modeling using all available streamside data (gathered in May through October) reveals strong seasonality in the residuals. The upper panel of Figure 3 displays the proportion of positive residuals for the streamside data by month. The EB model residuals use the posterior median as a point estimate of p_{ij} ; the G&B residuals use maximum likelihood point estimates. The seasonality is striking: in May and June, the model produces reasonably symmetric residuals, but as the summer progresses, the model exhibits increasing bias toward over-prediction of the proportion of animals killed. Late in the season, the residuals move back toward symmetry.

From the plot of the proportion of positive leave-one-out residuals for laboratory data, lower panel of Figure 3, we see that the monthly pattern in the streamside data is not evident in the laboratory. We note that the EB leave-one-out residuals were computed on the scale of the counts Y_{ij} and converted to proportions, so the apparent greater variability compared the G&B is in fact to do this discreteness. Still, in both EB and G&B fits, we see evidence that the field application of treatment prescriptions based on analysis of laboratory data alone may be problematic.

In order to perform out-of-sample validation of our models, we require data that arise from a similar process as the fitting data. Thus we subdivide the streamside data to obtain a subset free of seasonal effects (tests from May and June) and reserve the rest of the data (tests from July–Oct). Then we classify each tank according to which of the following posterior predictive probabilities is greatest: $Pr(Y = 0)$, $Pr(Y = n)$ or $Pr(Y \in (0, n))$. These predictive classifications versus the truth are shown in Table 1. For May and June, among tanks where $Y = 0$ or $Y = n$, the misclassification rate is 28% and almost always involves misclassifying into the middle category, $Y \in (0, n)$. In those tanks where $Y \in (0, n)$, the misclassification is more often in the direction of under-predicting than over-predicting. Compare this to the July–Oct tanks, where the overall misclassification rate is nearly the same (29%), but good classification of the $Y = n$ tanks is offset by only half of the remaining tanks being correctly classified. This reflects the same overprediction pattern also seen in Figure 3 for these months. An analogous classification based on the G&B results would require a new rule based on point predictions of the proportion killed rather than probability of binomial count outcomes, so we do not attempt that here. However, the residuals shown in Figure 3 indicate that G&B produces similar streamside predictions.

5. Simulation

To compare the frequentist performance of our EB model to that of the others described in Section 3, we simulate 500 data sets from the following model:

$$\begin{aligned} y_{ij} &\sim \text{Bin}(10, p_{ij}) \\ \text{logit}(p_{ij}) &= \mu_1 + \eta_{1j} + \frac{(\mu_2 + \eta_{2j})X_{ij}}{\exp(\mu_3 Z_{1ij} + \mu_4 Z_{2ij} + \mu_5 Z_{3ij} + \mu_6 Z_{1ij}^2 + \mu_7 Z_{2ij}^2 + \mu_8 Z_{3ij}^2)} \\ (\eta_{1j}, \eta_{2j}) &\sim N((-2, 2), \Sigma_{sim}) \quad \text{with } Pr=.75 \quad \text{or} \quad \sim N((2, -2), \Sigma_{sim}) \quad \text{with } Pr=.25 \end{aligned} \quad (7)$$

where Σ_{sim} is a 2×2 matrix with variances of 2 and correlation $-.9$. This “true model” is a version of the EB model (5) with the additional complication of a latent “study-level” effect such that 75% of the clusters have random slopes and intercepts centered at $(-2, 2)$ and 25% are centered at $(2, -2)$ so that the distribution is not symmetrical about zero. This violates the modeling assumptions of all the models and thus presents a challenging data analysis task.

In particular, the EB model assumes a single normal distribution for $(\eta_{1j}, \eta_{2j})^T$, and the G&B method assumes only random intercepts. We use model (7) to simulate both a set of historical data and a set of contemporary data, which are used as in the analysis above. The experimental design consists of 100 clusters of 10 tanks each in both data sets.

Covariate values comprise a balanced grid of all combinations of the following values: niclosamide (Z_1) $\in \{0, 0.5, 1.0, 1.5\}$, pH (Z_2) $\in \{7.0, 7.5, 8.0, 8.5, 9.0\}$, and alkalinity (Z_3) $\in \{50, 100, 150, 200, 250\}$. The covariates are centered (pH and niclosamide) or standardized (alkalinity) according to their means and standard deviations, as before. The true parameter values are set at $\mu = (-12, 10, -0.5, 1.3, 0.3, 0.15, 0.2, -0.1)^T$.

We follow the same multi-step procedure as above, first fitting model (1) to a set of *simulated historical data* generated from model (7) to obtain $\tilde{\Sigma}$, $\tilde{\mu}$, and $\tilde{\Sigma}$. Then we estimate cluster-specific parameters for *simulated fitting data* using model (4) and treat these as known in the EB model (5). For comparison, we also fit the fully Bayesian model (2), a fixed effects version of (5) where $\hat{\eta}_{j1} = \hat{\eta}_{j2} = 0$, the two-stage model (Bills et al., 2003), and the G&B likelihood-based model. Throughout, we use $c_\mu = 10^2$.

We compute $LC99.9_j^{\text{TRUE}}$ for each cluster $j = 1, \dots, 100$; recall that LC99.9 depends only on μ , not $(\eta_{1j}, \eta_{2j})^T$ and is thus the same across replicates. For each replicate $r = 1, \dots, 500$, we compute the root mean squared error (RMSE) and bias for LC99.9,

$$\begin{aligned} RMSF_r &= \sqrt{\frac{1}{J} \sum_{j=1}^J (LC99.9_j^{\text{TRUE}} - LC99.9_{jr}^{\text{model}})^2}, \\ \text{and bias}_r &= \frac{1}{J} \sum_{j=1}^J (LC99.9_j^{\text{TRUE}} - LC99.9_{jr}^{\text{model}}), \end{aligned}$$

where $LC99.9_{jr}^{\text{model}}$ is computed using the μ parameters. Based on encouragement from a referee, we also evaluate *relative* RMSE (RRMSE) and *absolute relative* bias.

Figure 4 displays the summary statistics for RRMSE, RMSE, bias, and absolute relative bias over the 500 replicates for the various models. The obviously worst model is the fixed effects model, banished from the figure due to the extremeness of its distributions of squared errors and biases. For that model, the RRMSE median was .36 with inter-quartile range (IQR) .023 – 8.56, and the median bias was .58 with IQR .03 – 21.0. Of the models shown

in the figure, the two-stage model (Bills et al., 2003) shows substantial bias compared to the other models and concomitantly large (R)RMSE. Among the remaining models, the EB model effectively ties with G&B and fully Bayesian models in terms of RRMSE and absolute bias, and enjoys a very slight RMSE advantage. However, the fully Bayesian and G&B models are slightly better for pure bias.

6. Discussion

This study demonstrates that by incorporating the hierarchical structure of laboratory data, we can produce accurate and precise estimates of extreme quantal response values such as LC99.9. Our treatment tables offer an improvement over the existing stream treatment charts produced using the two-stage method, as demonstrated by the simulation study results in Figure 4, where our EB approach yields smaller absolute bias and mean squared error. Our method also proved adept at predicting historical streamside outcomes (see e.g. Table 1), illustrating the applicability of our results to real-world treatment scenarios. A further advantage to this method is that we can produce uncertainty bounds for any quantities of interest, including LC99.9 values and outcome predictions. By contrast, the G&B method produces only asymptotic, Wald-type bounds, while the two-stage method cannot produce confidence bounds at all.

A substantial new insight revealed by our work is the apparent seasonal change in the toxicity of TFM and niclosamide mixtures in stream treatments (Figure 3). A seasonal effect was previously suspected: Scholefield et al. (2008) found that some larval sea lamprey captured from the wild during April and May were more sensitive to TFM than those captured during July through September. The concave seasonal pattern of the streamside residuals has clear implications for ongoing stream treatments. While laboratory studies yield accurate prescriptions of treatment requirements during May and June, they underestimate the actual stream treatment requirements outside of these months. The mechanism that underlies this seasonal effect in the streamside tests remains unknown. Any direct effect of seasonality can be eliminated because the effect was observed from the streamside tests, but not the laboratory tests. The toxicity of TFM is likely due to profoundly reduced serum glucose levels (hypoglycemia) that lead to terminal neurological failure (Wilkie et al., 2007). Therefore, one hypothesis is that replenishment of glycogen (a glucose storage molecule) reserves during peak summer feeding helps delay the onset of profound hypoglycemia.

Of course, like any method, our EB approach has its limitations. Some of our model diagnostics in Table 1 reveal inadequacies we could not fully address. However, these inadequacies are present in all the competitors as well, including the G&B method which produces very similar (and similarly accurate) LC99.9 predictions. Our assumption of conditional independence for the deaths within each tank may be a bit strong; modest positive dependence would likely increase the RMSE and RRMSE of our predictions. Also, our approach of downweighting the contribution of historical data (via c_μ) until the posterior quantities of interest stop changing might or might not generalize well in other nonlinear hierarchical modeling scenarios, but at least follows standard applied Bayesian practice of using as vague a hyperprior as the MCMC algorithm will tolerate for the given model and dataset. Our Section 5 simulation used a slightly more complicated “truth” than assumed by any of our preferred models, and one might wonder about the implications of this for the generalizability of our results. However, the true model was chosen as a “level playing field” on which to compare models that range from overly simple (two-stage) to overly complex (fully Bayesian); we would expect our preferred EB model to perform no worse using a less challenging “truth.”

In view of very similar frequentist operating characteristics, sea lamprey control scientists may prefer the EB approach since it delivers credible sets that are almost impossible for practitioners to misinterpret. By contrast, the G&B method excludes relevant data, and produces only asymptotic, Wald-type confidence intervals that are more difficult for non-statisticians to interpret and may be inaccurate when the corresponding likelihood departs markedly from normality.

One obvious trade-off for treating parameters as known is the potential sensitivity to poor estimation. In our approach, we have treated Σ^* or $\hat{\eta}_{1j}$ and $\hat{\eta}_{2j}$ as known in the fixed Sigma and EB models, respectively. We conducted a small additional simulation study, described in Web Appendix B, which indicated that poor estimation of Σ^* in the fixed Sigma model leads to the poorer operating characteristics of that model compared to EB and G&B. Using the true η_{1j} and η_{2j} to estimate Σ^* improves performance (see Web Figure 1), but obviously, such an approach is impossible in a real data analysis scenario.

From a theoretical standpoint, one might prefer the fully Bayesian model. Hyperpriors closely resembling the truth for Σ lead to improved performance for that model in our simulation study (see again Web Appendix B), but not enough to provide a significant advantage over EB (Web Figure 1). Further, this advantage is artificial in the sense that such prior information is unlikely to be available in a real data analysis setting. Indeed, one of the chief challenges of our analysis is to incorporate historical data when previous studies' designs preclude reliable estimation of cluster-specific parameters (or higher-level parameters controlling these) in the model applied to the contemporary data.

Finally, it is true that we initially adopted our EB approach on pragmatic grounds: hierarchical logistic response models are an area where WinBUGS is known to struggle, and the high dimension and nonlinearity of our mean structures further exacerbate this. Ideally, we would fit a properly parameterized (relative to the study design) fully Bayesian model. But technical problems in WinBUGS prevented us from comparing the performance of such a model to the overparameterized fully Bayesian, fixed Sigma, EB, and G&B models. However, we do find the EB approach attractive in its own right, largely since it helps improve overall model identifiability and thus produces possibly safer and more sensible estimates.

Overall, our analyses also provide a useful example for laboratory science more broadly. Laboratory studies are valued because experimental conditions can be tightly controlled. Although laboratory conditions enable isolation of hypothetical mechanisms from background variation, models based solely on data from tightly controlled experimental studies can fail to predict responses in nature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge financial support from the Great Lakes Fishery Commission Sea Lamprey Research Program under contract 08E32282000002 with the U.S. Geological Survey (USGS). Collaboration between the USGS and the University of Minnesota was enabled under USGS contract 08ERSA0351/0001. The contributions of M.A.B. were funded by the Great Lakes Fishery Commission Technical Assistance Program. We thank the Great Lakes Sea Lamprey Control Program for providing data from historic streamside tests. Use of trade, product, or firm names does not imply endorsement by the U.S. Government.

References

- Akaike, H. Information theory and an extension of the maximum likelihood principle.. In: Petrov, B.; Csaki, F., editors. Second International Symposium on Information Theory. Akademiai Kiado; Budapest: 1973. p. 267-281.
- American Society for Testing and Materials. *Annual Book of ASTM Standards*, volume 11.05, chapter Standard guide for conducting acute toxicity tests with fishes, macroinvertebrates and amphibians, E729-96. American Society for Testing and Materials; Philadelphia, PA: 2000. p. 213-233.
- Bills T, Boogaard M, Johnson D, Brege D, Scholefield R, Westman R, Stephens B. Development of a pH/alkalinity treatment model for applications of the lampricide TFM to streams tributary to the Great Lakes. *Journal of Great Lakes Research*. 2003; 29(Suppl. 1):510-520.
- Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*. 1997; 7:434-455.
- Davidian, M.; Giltinan, D. *Nonlinear Models of Repeated Measurement Data*. Chapman & Hall/CRC; Boca Raton: 1995.
- Gelderen EV, Ryan A, Tomasso J, Klaine S. Influence of dissolved organic matter source on silver toxicity to *Pimephales pomelas*. *Environmental Toxicology and Chemistry*. 2003; 22:2746-2751. [PubMed: 14587917]
- Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457-472.
- Gutreuter S, Boogaard M. Prediction of lethal/effective concentration/dose in the presence of multiple auxiliary covariates and components of variance. *Environmental Toxicology and Chemistry*. 2007; 26:1978-1986. [PubMed: 17705653]
- Hunn J, Allen J. Movement of drugs across the gills of fishes. *Annual Review of Pharmacology*. 1974; 14:45-57.
- Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*. 2009; 28:3049-3067. [PubMed: 19630097]
- Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*. 2000; 10:325-337.
- Pimentel D, Zuniga R, Morrison D. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*. 2005; 52:273-288.
- Scholefield R, Slaght K, Stephens B. Seasonal variation in sensitivity of larval sea lampreys to the lampricide 3-Trifluoromethyl-4-Nitrophenol. *North American Journal of Fisheries Management*. 2008; 28:1609-1617.
- Smith B, Tibbles J. Sea lamprey (*Petromyzon marinus* lakes Huron, Michigan and Superior: History of invasion and control, 1936-78. *Canadian Journal of Fisheries and Aquatic Sciences*. 1980; 37:1780-1801.
- Smith S. Species succession and fishery exploitation in the great lakes. *Journal of the Fisheries Research Board of Canada*. 1968; 25:667-693.
- Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*. 2002; 64:583-639.
- Stiratelli R, Laird N, Ware J. Random-effects models for serial observations with binary response. *Biometrics*. 1984; 40:961-971. [PubMed: 6534418]
- Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*. 2005; 12:1-16.
- Wilkie M, Holmes J, Youson J. The lampricide 3-trifluoromethyl-f-nitrophenol (TFM) interferes with intermediary metabolism and glucose homeostasis, but not with ion balance, in larval sea lamprey (*Petromyzon marinus*). *Canadian Journal of Fisheries and Aquatic Sciences*. 2007; 64:1174-1182.
- Yang R, Berger J. Estimation of a covariance matrix using the reference prior. *Annals of Statistics*. 1994; 22:1195-1211.

Figure 1. Covariate values represented in the five studies. Each point is a tank, with plotting symbol indicating study. Niclosamide and pH are centered; alkalinity is standardized. The legend presents the number of units at each level of the model hierarchy.

Figure 2. Posterior means and 95% credible intervals for LC99.9 from the EB model fit to 2001–2009 data at various c_μ values (left cluster), G&B model fit to only the contemporary data (2001–2009), and the same fit to all the laboratory data (1989–2009). Each subplot represents a different combination of pH and alkalinity (alk) values, but all are at mean niclosamide augmentation (.52).

Figure 3. Proportion of positive residuals (truth – posterior median of proportion killed) by month for streamside validation data (upper panel) and leave-one-out fitting data (lower panel) plotted against study month. EB model is in black;G&B is in grey.

Figure 4. RRMSE, RMSE, bias, and absolute for LC99.9 in 500 simulated data sets. Estimates are based on posterior medians of the μ parameters in the Bayesian models. “EB” is model (5), “Fixed Sigma” is model (4), “Fully Bayesian” is model (2), “Two-stage” is the Bills et al. (2003) two-stage model, and “G&B” is the Gutreuter and Boogaard (2007) likelihood-based model. The “Fixed” model (5) with $\hat{\eta}_{1j}=\hat{\eta}_{2j}=0$ is excluded for the sake of the scale of the figures.

Table 1

Classification of streamside outcomes according to the largest posterior predictive probability among $Y = n$, $Y = 0$, and $Y \in (0, n)$.

Truth	May & June Predictions			July-Oct Predictions		
	Y = n	Y ∈ (0, n)	Y = 0	Y = n	Y ∈ (0, n)	Y = 0
Y = n	164 (75%)	52 (24%)	3 (1%)	383 (97%)	12 (3%)	0 (0%)
Y ∈ (0, n)	7 (13%)	31 (56%)	17 (31%)	65 (49%)	69 (51%)	0 (0%)
Y = 0	1 (1%)	30 (24%)	92 (75%)	18 (5%)	152 (46%)	163 (49%)