A Bayesian Hierarchical method for fitting multiple health endpoints in a toxicity study

Taeryon Choi,[1] **Mark J. Schervish,**[1] **Ketra A. Schmitt**[2] **and Mitchell J. Small**[2]

[1] Department of Statistics, Carnegie Mellon University,
Pittsburgh, PA 15213

[2] Department of Engineering and Public Policy, Carnegie Mellon University,
Pittsburgh, PA 15213

September 22, 2004

Summary. Bayesian hierarchical models are built to fit multiple health endpoints from a dose-response study of a toxic chemical, perchlorate. Perchlorate exposure results in iodine uptake inhibition in the thyroid, with health effects manifested by changes in blood hormone concentrations and histopathological effects on the thyroid. We propose linked empirical models to fit blood hormone concentration and thyroid histopathology data for rats exposed to perchlorate in the 90-day study of Springborn Laboratories Inc. (1998), based upon the assumed toxicological relationships between dose and the various endpoints. All of the models are fit in a Bayesian framework, and predictions about each endpoint in response to dose are simulated based on the posterior predictive distribution. A hierarchical model tries to exploit possible similarities between different combinations of sex and exposure duration, and it allows us to produce more stable estimates of dose-response curves. We also illustrate how the hierarchical model allows us to address additional questions that arise after the analysis.

Key words: Dose-response study; Perchlorate; Hierarchical prior distribution; Logistic regression; Multivariate regression; MCMC; Optimal design point.

1. Introduction

In this paper, we develop a statistical approach to integrate and compare different toxicity studies in human and ecological risk assessment. Our proposed approach for this purpose is to apply Bayesian statistical techniques for harmonizing several different health endpoints studies or methods. A hierarchical Bayesian framework for fitting toxicological relationships in human and ecological receptors can be applied to evaluate relationships in dose-response among multiple health endpoints; different exposure conditions or durations; varying population characteristics such as age or gender within a given species; multiple target species. Hierarchical methods are applied using results for multiple health endpoints associated with exposure to a single chemical, perchlorate, for a single species, rats, allowing a hierarchy for gender and exposure period.

Perchlorate (ClO_4^-) is an oxidizing anion that originates as a contaminant in ground and surface water from the dissolution of ammonium, potassium, magnesium, or sodium salts. Perchlorate is exceedingly mobile in aqueous systems and can persist for many decades under typical ground or surface water conditions. Large-scale production of perchlorate-containing chemicals in the United States began in the mid-1940s and perchlorate began to be discovered at various manufacturing sites and in well water and drinking water supplies after 1997. There are 20 states with confirmed releases of perchlorate to ground or surface water, and there are 40 states that have confirmed the presence of perchlorate manufacturers or users. Increasing attention is being paid to perchlorate because of existing uncertainties in the toxicological database that make it difficult to adequately address the potential for perchlorate at low levels in drinking water to produce ecological and human health effects. The EPA recently listed perchlorate as a contaminant that required additional research and occurrence information before regulatory determinations could be made.

Exposure to perchlorate is known to inhibit the uptake of iodide in the thyroid of animals and humans, thereby causing a reduction in the hormones thyroxine (T3) and triiodothyronine (T4), and an increase in thyroid stimulating hormone (TSH) in blood. Circulating hormone levels (T3, T4 and TSH) and histopathology data on thyroid tissue can be monitored for a hazard identification. Other potential concerns involving perchlorate include carcinogenic, neurodevelopmental, reproductive, and immunotoxic effects that may result from changes in thyroid function.

Bayesian statistical techniques have proven useful in clinical and environmental epidemiological applications to evaluate and integrate available information (Berry and Stangl, 1996 ; Wilson, 2001). Because Bayesian techniques provide probabilistic estimates of effects, data expressed as incidence can be readily combined with continuous measures. Hierarchical Bayesian techniques can be used to elucidate common and divergent endpoints for use in risk assessment based on the consideration of taxonomy and the mode of action of perchlorate. We focus on one laboratory animal bioassay for ammonium perchlorate; the Springborn 90-day study (Springborn Laboratories Inc., 1998).

The Springborn 90-day study includes different experimental results at three time points during the study: day 14, day 90 and day 120 (after a 30-day recovery period); We analyzed the results from days 14 and 90.

In this study, assays for T3, T4, and TSH were performed and histopathology endpoints were measured. Crofton and Marcus (2001) reanalyzed the original data and report the checked data for T3, T4, and TSH, and we used this as our source of data for the blood hormone endpoints. The thyroid histopathology, as reviewed and reported by the pathology working group (PWG), can be found in Wolf (2001), and we used these results as our data source for these endpoints.

In Section 2, we introduce the data, the Springborn 90-day study, consisting of the 14-day sacrifices (i.e., rats within a 14-day exposure period) and the 90-day sacrifices. In Section 3, we describe two models that we consider in the analysis, including a mechanistic model based on expert knowledge and an empirical model constructed from various statistical models. In Section 4, we summarize the results using a hierarchical Bayesian approach with both models. In Section 5, we illustrate how to address a serious issue that arose from the analysis. In particular, there is a wide gap between the highest and next-to-highest dose in the study. Some of the dose-response curves that we fit increase significantly between those two doses. Hence, the data set gives us very little information about the shape of the dose-response curve in the region where it is changing the most. We illustrate how to use the posterior distribution of the model parameters to choose a new dose level that would most help to reduce the uncertainty about the shape of the dose-response curve. In Section 6, we conclude and discuss our results and future work.

2. Springborn 90-day study

The Springborn Laboratories Inc. (1998) study was a 90-day study of the effects of ammonium perchlorate (NH_4ClO_4) on the thyroid system of laboratory rats. The study was part of a bioassay testing strategy that consisted of oral administration of ammonium perchlorate via drinking water to male and female Sprague-Dawley rats. The rats were dosed for up to 90 days, with a 30 day recovery period for some of the rats. Several endpoints were measured at three different time points (14-day, 90-day and 120-day) and thyroid hormone analyses were performed at the 14, 90, and 120-day sacrifices.

In each laboratory experiment of the 14-day and 90-day sacrifices, 120 Sprague-Dawley rats were divided into 12 groups of 10 rats each. There were 6 groups of males and 6 groups of females. Each group received one of the following six doses of ammonium perchlorate in the drinking water: 0, 0.01, 0.05, 0.2, 1.0, and 10 mg/kg-day (Springborn Laboratories Inc., 1998). Measurements were taken of the concentration of several thyroid hormones in the blood and, among these thyroid hormones, our main interests are in the effects of perchlorate on T3, T4 and TSH. The other important endpoints were histopathology endpoints of the thyroid that included the incidence of colloid depletion, hypertrophy and hyperplasia. The histopathology measurements were reported as the numbers of incidences out of the (usually ten) rats in each dose group.

3. Models

3.1 *Mechanistic Model*

According to biological theory about the toxicological effects of perchlorate, there are cause-and-effect relationships among the toxicological endpoints that we consider, i.e. T3, T4, TSH, colloid depletion, hypertrophy and hyperplasia. We illustrate these effects in Figure 1. In structuring our model, we consulted toxicologists with expertise on perchlorate's mode-of-action in the thyroid, and the relationships in Figure 1 are compatible with the general mode of action for perchlorate as determined by the U.S.EPA (U.S.Environmental Protection Agency, 1998).

[Figure 1 about here.]

Figure 2 shows the assumed causative relationships on which we base our overall model. Figure 2 is extracted from Figure 1, and we call this our *mechanistic model.* In spite of the relation between the mechanistic model and the assumed mode-of-action, the observed data are not fit well by the mechanistic model. This poor fit may result from several problems in the laboratory studies themselves. For example, all of the dependent measurements including hormone and histopathology data, were taken at the same time point and these measurements may not be suitable for supporting the underlying cause-and-effect relationships, which involve dynamic systems that may exhibit time delays. Instead of adhering to the mechanistic approach, we next consider *empirical models* based on the observed data.

[Figure 2 about here.]

3.2 *Empirical model*

A notable difference between the mechanistic and empirical models is that the latter proposes a direct relationship from dose to TSH. That is, the empirical models do not introduce intermediate steps in which T3 and T4 predict TSH as in the mechanistic model. This approach seems reasonable since it is known that T3 and T4 decrease with dose while TSH increases with dose, and we do not observe the proper sequencing from T3 and T4 to TSH in the observed data. In addition, there are some modifications at the final stage for predicting the histopathology measures (hypertrophy and hyperplasia) in response to dose, T4 or both. Although the assumed mechanistic relationship for the final stage is to predict histopathology data only from T4 hormone values, the observed data indicate that considering both the dose level and T4 hormone levels is useful. This is due in part to the fact that the histopathology data are reported only by dose group rather than for each individual rat.

The empirical models fall into three classes: the first model with every dependent measurement explained by dose alone; the second model, different from the first only in that hypertrophy and hyperplasia are predicted by T4 instead of dose; and the third model, a combination of first two models, allowing hypertrophy and hyperplasia to be predicted from both dose and T4. Within the three classes of models we have the issue of how to treat the explanatory variable dose. For this, we consider two sub-models: one with raw dose itself, and the other with a logarithmic transformation of the dose. In taking the logarithm of dose, we need a way to deal with the zero nominal dose group. We choose to add a small value, denoted δ, to all of the doses. This is a device that is necessary in order to take logarithms. We shall see later whether the extra parameter in the model provides a sufficiently better fit. Another alternative would be to add a small amount, such as δ, only to the zero nominal dose group. Our approach is to add δ to every dose level for the following reason. We interpret δ as if it were a background dose of contaminants that also affect the measured endpoints. The administered doses are added to this background to produce the total contaminant exposure. We treat the value δ as a parameter and estimate it in the model. Since δ is estimated, we can let δ have different values for different stages or different endpoints by regarding them as different parameters in the different model stages. For the remainder of this paper, we refer to these parameters as the "dose increments".

From this point of view, we have several empirical models proposed. To help choose between them, we use a model selection criterion approach combined with expert knowledge. Although cross-validation is an excellent method for assessing the predictive power of a model, the data set being modeled is not sufficiently large to make meaningful cross-validation comparisons. Among the other overall fit measures, we choose a quantity known as the Bayesian information criterion (BIC) (Schwarz, 1978). For parametric model k with maximum likelihood estimator (MLE) $\hat{\theta}_k$, number of parameters p, and sample size n, the BIC is defined by

$$\mathrm{BIC}(k) \equiv -2 \log L(\widehat{\theta_k}) + p \log n$$

Models with small values of the BIC are assumed to fit better than models with large BIC.

Finally, we choose two models to study more closely. The first model, which we call Model A, is the model that has the lowest BIC value while the second model, which we call Model B, has the second lowest BIC value and more closely reflects the expert opinion that T4 affects the incidence of HT and HP. For reference, Table 1 shows the BIC values for the two models that we use.

[Table 1 about here.]

As mentioned above, our data are incomplete in regard to the histopathology data, specifically the incidence of colloid depletion, hypertrophy and hyperplasia. These data were reported only by dose group while the thyroid hormones, T3,T4 and TSH were measured and reported for individual rats. We address this incomplete information by using summary statistics such as the average hormone levels for each dose group when we must model the histopathology data as a function of hormones.

We divide each of our final two models (A and B) into three steps: modeling colloid depletion, modeling the three hormone values, and modeling hypertrophy and hyperplasia. We assume independence between the parameters in the different steps.

According to our experts, the observed effects on the histopathology endpoints should not decrease with dose. Hence, we put constraints on the slopes in the logistic regression models for the histopathology data. For example, because the histopathology endpoints are known to increase with dose, we require the slopes of dose to be positive. Also, since $\log \mathrm{T4}$ is known to have negative effects on HT and HP, we require that the slope of $\log \mathrm{T4}$ be negative. In our descriptions of the models in the following subsections, the distribution denoted PN is a conditional normal distribution constrained to be nonnegative, and the distribution denoted NN is a conditional normal distribution constrained to be nonpositive.

Throughout the analysis, our approach is based upon Bayesian hierarchical modeling and we use WinBUGS software to fit our models via the Markov Chain Monte Carlo (MCMC) method. Since our approach is based upon Bayesian analysis, the prior distributions on parameters need to be specified. We consider two types of prior distributions. Noninformative priors are normal prior distributions with high variances. Hierarchical priors have hyperparameters for both sex and exposure time (14-day and 90-day). In the hierarchical structure (see Figure 3) we treat the parameters for female and male rats at the same time point as sampled from a distribution with common hyperparameters within each time point. Then, we introduce hyperparameters for the 14-day and 90-day time points as sampled from another distribution with common hyperparameters between exposure times.

[Figure 3 about here.]

3.2.1 Step 1 : Dose ⇒ Colloid Depletion The first step is to model the effect of dose on colloid depletion (CD). We use a logistic regression model for the probability of colloid depletion given dose. The explanatory and response variables, respectively dose and the number of rats with CD, were reported by group. That is, rather than a separate indicator of the presence of CD in each rat, we have only the 12 counts of rats that exhibited CD out of the ten rats in each of the 12 groups. These data are shown in Table 2 along with the hypertrophy (HT) and hyperplasia (HP) data that are relevant for Step 3.

[Table 2 about here.]

The following equations describe our models for Step 1.

- Logistic regression: response = CD, explanatory = dose. We have separate parameters for the four combinations of male/female and 14-day/90-day but otherwise the models are identical. In the superscripts, s stands for sex (male or female) and t stands for time (14-day or 90-day).

- Model structure:

  - For the ith dose group ($i = 1, \ldots, 6$), $r_i^{\mathrm{CD},s,t}$ is the number of incidences of CD, n_i is the number of rats observed, and $p_i^{\mathrm{CD,s,t}}$ is the probability of exhibiting CD.
  - $r_i^{\mathrm{CD},s,t} \sim \mathrm{Bin}(n_i, p_i^{\mathrm{CD},s,t})$
  - Model A : $\mathrm{logit}(p_i^{\mathrm{CD},s,t}) = \alpha_1^{\mathrm{CD},s,t} + \alpha_2^{\mathrm{CD},s,t}\mathrm{dose}_i$

    Model B : $\mathrm{logit}(p_i^{\mathrm{CD},s,t}) = \alpha_1^{\mathrm{CD},s,t} + \alpha_2^{\mathrm{CD},s,t}\log(\mathrm{dose}_i + \delta_1^{s,t})$,

- Prior distributions:

  - Noninformative prior distribution:

    $$\alpha_1^{\mathrm{CD},s,t} \sim N(0, 10^3), \quad \alpha_2^{\mathrm{CD},s,t} \sim PN(0, 10^3), \quad \delta_1^{s,t} \sim \mathrm{Exp}(5).$$

    All parameters are independent.

  - Hierarchical prior distribution

    $$\alpha_1^{\mathrm{CD},s,t} \sim N(\mu_1^t, (\sigma_1^t)^2), \quad \alpha_2^{\mathrm{CD},s,t} \sim PN(\mu_2^t, (\sigma_2^t)^2), \quad \delta_1^{s,t} \sim \mathrm{Exp}(\gamma_t).$$

$$\text{For } i=1,2\text{: } \mu_i^t \sim N(\mu_i, \sigma_i^2), \quad (\sigma_i^t)^2 \sim \text{Gamma}(a_i, b_i), \ \gamma_t^{-1} \sim \text{Gamma}(a_\delta, b_\delta)$$

$$\mu_i \sim N(0, 10^3), \quad 1/\sigma_i^2 \sim \text{Gamma}(0.001, 0.001),$$

$$a_i \sim \text{Gamma}(2, 1.2), \quad b_i \sim \text{Gamma}(2, 1.2),$$

$$a_\delta \sim \text{Gamma}(2, 2), \quad b_\delta \sim \text{Gamma}(2, 0.2)$$

3.2.2 Step 2 : Dose ⇒ T3, T4 and TSH The second step in our model is to predict the effects on T3, T4 and TSH from dose. Expert knowledge suggests that T3 and T4 levels in the blood drop and the TSH level in the blood rises with increasing dose of perchlorate. Even though it is suggested by expert knowledge that the rates of change in the three hormone levels are quickened once CD has occurred, the data do not record which rats had CD, so we cannot compare hormone levels for rats with and without CD. Therefore, we predict the hormone variables by dose alone. Previous analyses of the hormone data, such as Crofton (1998) and Crofton and Marcus (2001), used an ANOVA approach. Our approach is to view the dose level as a continuous measurement and fit a regression model. In addition, the three hormones T3, T4 and TSH are related in the mechanistic model and exhibit noticeable correlation. Therefore, our proposed model for Step 2 is a multivariate linear regression model with the logarithmic transformation of dose level as the predictor. The dose measurements were 0, 0.01, 0.05, 0.2, 1.0 and 10 mg/kg-day with ten rats of each sex at each dose level. Taking the logarithm of the dose level adjusts the scale of dose level so as to improve the linear fitting of hormone variables. Once we take the logarithmic transformation, we must deal with the zero dose level. We address this issue by introducing additional parameters, namely the δ's that we described earlier. We consider two ways of introducing the δ's: three distinct δ's, one for each hormone, and one common δ for all three hormones. The choice between these two ways of introducing δ's are made as part of the overall model selection exercise described in Table 1.

We do not impose constraints on the slopes of the three hormones as we did in Step 1 because we did not know for sure at what point the sacrifices would occur in the feedback mechanism illustrated in Figure 1. When TSH increases, the thyroid tries to replenish the reduced levels of T3 and T4. Although we expected rats to have decreased levels of T3 and T4 and increased levels of TSH, we did not want to force the model to fit constrained slopes.

In the following description of the model, some notation appears repeatedly. A vector of all 0's is denoted $\underline{0}$, and the following two special matrices are used in several places:

$$\Sigma = \begin{pmatrix} 1.0 & -0.5 & 0.5 \\ -0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{pmatrix} \times 10^{-2}, \quad \Omega = \begin{pmatrix} 1.0 & 0.5 & -0.5 \\ 0.5 & 1.0 & -0.5 \\ -0.5 & -0.5 & 1.0 \end{pmatrix} \times 10^{-2}.$$

- Multivariate regression: response = (log(T3), log(T4), log(TSH)), explanatory = dose. We have separate parameters for the four combinations of male/female and 14-day/90-day but otherwise the models are identical. In the superscripts, s stands for sex (male or female) and t stands for time (14-day or 90-day).

- Model structure :

  - $\underline{Y}_j^{s,t}$: three-dimensional response vector for jth individual, $j = 1, \ldots, 60$ (with missing values).
  - $\underline{Y}_j^{s,t} \sim \text{MVN}_3\left(\underline{\mu}_j^{s,t}, \Sigma^{s,t}\right)$, trivariate normal distribution,

  - Model A : $\underline{\mu}_j^{s,t} = \underline{\nu}^{s,t} + \underline{\gamma}^{s,t} \log(\text{dose}_j^{s,t} + \underline{\delta}^{s,t})$,

    Model B : $\underline{\mu}_j^{s,t} = \underline{\nu}^{s,t} + \underline{\gamma}^{s,t} \log(\text{dose}_j^{s,t} + \delta_{\mathrm{c}}^{s,t})$.

    (Here, $\underline{\delta}_2^{s,t} = (\delta_{2,1}^{s,t}, \delta_{2,2}^{s,t}, \delta_{2,3}^{s,t})$ is a vector of three δ values for the three hormones, and $\delta_{\mathrm{c}}^{s,t}$ is a common δ value for all three hormones.)

- Prior distribution

  - Noninformative prior distribution

$$\underline{\nu}^{s,t} \sim \mathrm{MVN}_3\left(\underline{0}, \Sigma\right), \quad \underline{\gamma}^{s,t} \sim \mathrm{MVN}_3\left(\underline{0}, \Sigma\right), \quad (\Sigma^{s,t})^{-1} \sim \mathrm{Wishart}\left(\Omega, 3\right),$$
$$\delta_{2,i}^{s,t} \sim \mathrm{Exp}(5), \ i = 1, 2, 3, \quad \delta_c^{s,t} \sim \mathrm{Exp}(5).$$

  - Hierarchical prior distribution

$$\underline{\nu}^{s,t} \sim \mathrm{MVN}_3\left(\underline{\mu}_\nu^t, \Sigma_\nu^t\right), \quad \underline{\gamma}^{s,t} \sim \mathrm{MVN}_3\left(\underline{\mu}_\gamma^t, \Sigma_\gamma^t\right), \quad (\Sigma^{s,t})^{-1} \sim \mathrm{Wishart}\left(\Omega, 3\right),$$
$$\underline{\mu}_\nu^t \sim \mathrm{MVN}_3\left(\underline{\mu}_\nu, \Sigma_\nu\right), \quad (\Sigma_\nu^t)^{-1} \sim \mathrm{Wishart}\left(\Omega, 3\right), \quad \underline{\mu}_\gamma^t \sim \mathrm{MVN}_3\left(\underline{\mu}_\gamma, \Sigma_\gamma\right),$$
$$(\Sigma_\gamma)^{-1} \sim \mathrm{Wishart}\left(\Omega, 3\right), \quad \underline{\mu}_\nu \sim \mathrm{MVN}_3\left(\underline{\mu}, \Sigma\right), \quad \underline{\mu}_\gamma \sim \mathrm{MVN}_3\left(\underline{\mu}, \Sigma\right)$$
$$\delta_{2,k}^{s,t} \sim \mathrm{Exp}(\gamma_t), \ k = 1, 2, 3, \ \delta_c^{s,t} \sim \mathrm{Exp}(\gamma_t)$$
$$\gamma_t^{-1} \sim \mathrm{Gamma}(a_\delta, b_\delta), \ a_\delta \sim \mathrm{Gamma}(2, 2), \ b_\delta \sim \mathrm{Gamma}(2, 0.2)$$

3.2.3 Step 3 : (Dose , T4) ⇒ Hypertrophy and Hyperplasia The final step predicts the incidence of hypertrophy (HT) and hyperplasia (HP) from dose only, from T4 level only or from a combination of both dose and T4. In this step, we fit two logistic regression models as we did in the first step. The first logistic regression is for the probability of HT, and the second logistic regression is for the conditional probability of HP given the incidence of HT.

The response variables are the incidences of HT and HP, and the explanatory variables are dose level and/or the amount of T4 (after a logarithmic transformation). The data on HT and HP are available only as counts within each group and dose level is measured by group, whereas T4 is measured on each individual rat. Because one explanatory variable (T4) is known separately for each rat, special attention is needed when T4 is used as a predictor. We choose to use the average of the ten T4 values for each group. In addition, we need to consider both HT and HP simultaneously because HT must occur in order for HP to occur. That is, the biological assumption from expert knowledge suggests that it is not plausible for a rat to have HP without having HT first.

The following equations show our models for the final step.

- Logistic regressions : responses = HT and HP, explanatory = dose and/or T4. We have separate parameters for the four combinations of male/female and 14-day/90-day but otherwise the models are identical. In the superscripts, e stands for endpoint (HT or HP), s stands for sex (male or female) and t stands for time (14-day or 90-day).

- Model structure:

  - For the ith group ($i = 1, \ldots, 6$), $r_i^{e,s,t}$ is the number of incidences of the endpoint, $n_i^{s,t}$ is the number of rats observed, $p_i^{e,s,t}$ is the probability of having the endpoint, (for e = HP it is the conditional probability of getting HP given that the rat has HT) and $\overline{\mathrm{T4}}_i^{s,t}$ is the average of 10 values of T4.

- $r_i^{\mathrm{HT},s,t} \sim \mathrm{Binomial}(n_i^{s,t}, p_i^{\mathrm{HT},s,t})$, $r_i^{\mathrm{HP},s,t} \sim \mathrm{Binomial}(r_i^{\mathrm{HT},s,t}, p_i^{\mathrm{HP},s,t})$ given $r_i^{\mathrm{HT},s,t}$.
- Model A : $\mathrm{logit}(p_i^{e,s,t}) = \alpha_1^{e,s,t} + \alpha_2^{e,s,t}\mathrm{dose}_i$.
- Model B : $\mathrm{logit}(p_i^{e,s,t}) = \alpha_1^{e,s,t} + \alpha_2^{e,s,t}\mathrm{dose}_i + \alpha_3^{e,s,t}\log(\overline{\mathrm{T4}}_i^{s,t})$.

- Prior distributions
  - Noninformative prior distribution

    $$\alpha_1^{e,s,t} \sim N\left(0, 10^3\right),\ \alpha_2^{e,s,t} \sim PN\left(0, 10^3\right),\ \alpha_3^{e,s,t} \sim NN\left(0, 10^3\right),\ \delta_5^{s,t} \sim \mathrm{Exp}(5).$$

  - Hierarchical prior distribution

    $$\alpha_1^{e,s,t} \sim N(\mu_1^{e,t}, \sigma^{2e,t}_{1}),\ \alpha_2^{e,s,t} \sim PN(\mu_2^{e,t}, \sigma^{2e,t}_{2}),\ \alpha_3^{e,s,t} \sim NN(\mu_3^{e,t}, \sigma^{2e,t}_{3}),\ \delta_5^{s,t} \sim \mathrm{Exp}(\gamma_t)$$

    $$\text{For } i = 1,2,3\text{: } \mu_i^{e,t} \sim N(\mu_i^e, \sigma^{2e}_{i}),\ \sigma^{2e,t}_{i} \sim \mathrm{Gamma}(a_i^e, b_i^e)$$

    $$\mu_i^e \sim N(0, 10^3),\ \tau_i^e = 1/\sigma^{2e}_{i} \sim \mathrm{Gamma}(0.001, 0.001),\ \gamma_t^{-1} \sim \mathrm{Gamma}(a_\delta, b_\delta)$$

    $$a_i^e \sim \mathrm{Gamma}(2, 1.2),\ b_i^e \sim \mathrm{Gamma}(2, 1.2),\ a_\delta \sim \mathrm{Gamma}(2, 2),\ b_\delta \sim \mathrm{Gamma}(2, 0.2)$$

3.3 *Implementation and prediction*

To make the units of predictors from different situations, i.e. dose and $\log \mathrm{T4}$, more similar in the hierarchical framework, we use a standardized version of each predictor. That is, we subtract the sample mean and divide by the sample standard deviation of each predictor before fitting the models. With multiple studies (e.g., 14-day and 90-day), the issue arises as to how to standardize the predictors. We prefer standardizing the predictors in each study separately rather than standardizing all of them together. That is, the sample means and standard deviations are calculated separately for the different data sets of predictors by sex (F/M) and exposure time (14/90). Another alternative is to standardize all of them together by pooling all data sets into one data set and calculating the sample mean and standard deviation for the combined set. This alternative has some drawbacks in the framework of multiple studies. For example, if a new study were to be considered, we would need to recalculate the overall sample mean and standard deviation. This would change the interpretations of the prior distributions used for the parameters of the earlier data sets.

After fitting the models, prediction of the future data can be made based on the posterior predictive distribution. We illustrate this by drawing plots based on such predictions.

4. Results

4.1 *Step 1*

We obtained 2000 posterior samples for all parameters for each combination of female or male rats in the 14-day or 90-day study under different prior distributions (noninformative and hierarchical prior distributions) for both Model A and Model B. For 101 dose levels between 0 and 10 mg/kg-day with a spacing of 0.1mg/kg-day and for each of the 2000 simulated parameter vectors from the MCMC output, we calculated the probability of getting CD at each dose for each sex in each study.

There were some differences apparent between not only males and females but also Models A and B and between the two different specifications of prior distributions. The most notable case that we could observe in the prediction plots where Models A and B and the priors matter was with female rats in the 14-day study. In this case there were no incidences of CD at any of the dose levels less than 10.0 mg/kg-day. (See Table 2.) In terms of model fitting, the six incidences of CD for female rats from the 14-day study at dose level 10.0 are not enough to estimate the intercept and slope of a logistic regression and the resulting parameters are unstable with high standard deviation. This instability was reduced by the hierarchical approach and the standard deviations for most parameters became smaller than in the case of a noninformative prior distribution. In particular, the less informative female rat data in the 14-day study is complemented by information from male rats and 90-day data in the hierarchical structure. Figure 4 illustrates predictions from the four combinations of model and prior for this case. In addition, the prediction plots from Model B show a greater effect of lower dosages on the prediction curve, due to the logarithmic transformation of dose.

[Figure 4 about here.]

4.2 *Step 2*

In Step 2, Model A uses the different dose increments (δ_i's) for each of the three hormones and Model B uses a single dose increment for all three hormones. If a small δ value is estimated, then this δ value makes the zero dose level very far from the other dose levels on the logarithmic scale. In contrast, when large δ values are estimated, the zero dose case is treated as closer to the nonzero doses on the logarithmic scale.

Figure 5 illustrates the difference between separate δ's and common δ. It is observed that prediction plots based on model A provide better fitting than those based on model B, which can be interpreted as the advantage of using separate δ's.

[Figure 5 about here.]

4.3 *Step 3*

The probabilities of getting HT and of getting HP conditional on having HT were computed as follows. Since Step 3 of Model B includes T4 as a predictor, we used the average of the log(T4) level of all rats in each dose group as a predictor for HT and HP. Model A includes only dose level as a predictor, so probabilities are calculated more simply. In drawing the prediction plots, we used the same strategy as for Step 1. For each of the 101 dose levels that we used in Step 1 and for each of the 2000 simulated parameter vectors from the MCMC output, we simulated the average of 10 log(T4) levels for rats at the given dose and used this average as the predictor for HT and HP in Model B. The probability of HP in the graphs refers to the marginal probability of HP, calculated as the product of the probability of HT and the conditional probability of HP given the occurrence of HT.

In Model A, some of the predicted probabilities for HT and HP at the lower levels of dose (0, 0.01, and 0.05) are indistinguishable, while the predicted probabilities for HT and HP under Model B reflect the use of log(T4) as a predictor. The effect arises because the decrease of log(T4) at the lower dose levels is so significant. This change in log(T4) helps to explain the increase in the probability of getting HT at the lower dose levels as in the case of male rats in the 90-day study. Although the differences between the two models are not very striking in the 14-day study, we see some slight improvement in the fit at low dose levels in the 90-study from the use of log(T4) in model B, as shown in Figure 6

[Figure 6 about here.]

In the 90-day study, male and female rats show similar patterns of HP occurrence, with only one non-zero observation at the highest dose level. As a result, the probability curves for HP for both sexes in the 90-day study get to be determined by the single dose at which HP occurred. Under the noninformative prior distribution, the estimates for the parameters of the HP model are rather unstable but, if we consider the hierarchical prior distribution, more stable results are obtained.

As we have noted several times, a serious problem with the histopathology data is the insensitivity to dose, i.e. the small numbers of HT and HP incidences at most of the dose levels in the 14-day study. Of course, if these dose levels and the 14-day time point are all that are of interest, then there is some evidence that perchlorate has little effect on the probability of HP for both male and female rats. Because there is little or no change in the incidence of HP in the 14-day study, it is virtually impossible to get stable estimates for the parameters of the models that relate dose to probability of HP in the 14-day study. In addition, it doesn't much matter which prior distribution or model we use, the data don't tell us much more than the fact that the probability of incidence is low and doesn't change much over the dose levels in the design. Figure 7 illustrates this result.

[Figure 7 about here.]

Once again, the observed data are not sufficiently informative to fit a logistic regression model, which is similar to the situation that arose in Section 4.1 when we tried to predict the probability of getting CD for female rats in the 14-day study. In Section 5, we discuss a way to use our model fit to suggest what further data could be collected to help alleviate this lack of information.

5. How to collect more informative data

Suppose that the same laboratory experiment could be performed with one additional dose level. We would like to determine the optimal additional dose level from the point of view of reducing some of the uncertainty in the model fit. We use a method based on a preposterior analysis of our fitted Bayesian models to evaluate the value (expected variance reduction) of additional information (Müller, 1999). For this purpose, we chose to look at the variation in model fit that one sees in Figure 4. When performing a Bayesian analysis, one hopes that different choices of prior distribution and similar models will produce similar predictions. If they produce vastly different predictions, that is evidence that the data are not sufficiently informative to provide a good model fit, as we have already noted with the CD data. As a quantitative measure of uncertainty, we chose the following measure of spread between the four plots in Figure 4. Label the four plots $i = 1, 2, 3, 4$. Label the mean curve in each plot $j = 1$, label the 5th percentile curve $j = 2$, and label the 95th percentile curve $j = 3$. Let $u_{i,j,k}$ be the height of curve j in plot i at dose equal to k for $k = 0, 1, \ldots, 10$. Let $\overline{u}_{.,j,k} = \frac{1}{4}\sum_{i=1}^{4} u_{i,j,k}$ be the average height for each curve/dose combination. Our measure of spread between the four plots is then

$$V = \sum_{j=1}^{3}\sum_{k=0}^{10}\frac{1}{3}\sum_{i=1}^{4}(u_{i,j,k} - \overline{u}_{.,j,k})^2, \qquad (1)$$

the total of the sample variances of the four heights for all of the curve/dose combinations. The smaller V is, the more similar the four plots look, and the less difference the different model/prior combinations make to the predictions. In our analysis, $V = 0.3587$ was observed.

Now, we ask how we could make the value of V smaller by collecting a little more data. Specifically, if we were to sample another set of ten rats with a new dose d, which value of d would reduce V the most? We addressed this question by simulating new data at different dose levels, fitting the model again including the new data, and computing a corresponding value of V from (1). To be precise, we chose doses from 1.5 to 9.5 in steps of 0.5 together with doses of 1.1, 5.3, 5.6, 6.7, and 9.9. For each dose, we simulated 200 new sets of rats at that dose as follows. First, we selected one of the four model/prior combinations at random, Second, we selected one of the 2000 vectors of parameters from the posterior sample that was obtained while fitting that model/prior combination. Third, we computed the probability p of CD given the selected parameter and the chosen dose, Fourth, we simulated a binomial random variable Y with parameters 10 and p, and treated Y as if it were part of the original data.

For each simulated data set, we fit all four model/prior combinations and computed a V from (1) as if Figure 4 had been plotted again. We then averaged the 200 values of V and plotted them against dose in Figure 8.

[Figure 8 about here.]

We see that a new dose around 5mg/kg-day would be expected to lower the value of V quite a bit, thereby making the four plots in the resulting version of Figure 4 more similar.

6. Conclusion and Discussion

We have developed several models to fit perchlorate dose-response functions using data from the Springborn 90-day study of Springborn Laboratories Inc.

(1998), based upon a mechanistic model derived from the assumed toxicological relationships between dose and the various endpoints. All the models were estimated and predictions were simulated in a fully Bayesian framework. Predictions about each endpoint in response to dose exposure are simulated based on the posterior predictive distribution. The hormone data, consisting of T3, T4 and TSH, were fit well by a multivariate regression model, using a logarithmic transformation of dose plus a dose increment δ to deal with zero doses. By introducing different δ values for each hormone, a more flexible model fitting is possible but it is not clear that different values provide much improvement over a common δ. Some of the histopathology end points were not very sensitive to the changes in dose that were represented in the Springborn study. For this reason, it was difficult to fit dose-response curves separately to each sex/sacrifice date combination. A hierarchical model that tries to exploit possible similarities between different combinations of sex and exposure duration was able to produce more stable estimates of dose-response curves.

There are some further issues yet to be resolved. First, we chose two empirical models to fit the data based upon a mechanistic model, but there is still the issue of model selection between two competitive models. Model A is simpler than Model B and superior to Model B based on the model selection criterion BIC; however, Model B may reflect expert knowledge better than Model A in terms of the use of the T4 level in Step 3. This difference resulted in different predictions of the probability of HT, in particular, for the predictions at lower dose levels. Second, we chose a particular hierarchical structure to combine different sexes and exposure durations. Other hierarchical structures are available, and some might be more reasonable based on expert opinion.

Incomplete or sparse information, particularly in the histopathology data, was a serious problem in the analysis. To improve the estimation, one might try to find the optimal dose level at which to conduct an additional experiment. We did this via a simulation version of *preposterior analysis.*

The model that we developed can be evaluated by experts and estimates can be used to identify the hazardous effects of perchlorate exposure. In Schmitt et al. (2004), we pursue this approach by computing quantities of interest to regulators together with associated uncertainties.

Acknowledgements

This work was supported by US EPA Offices of Research and Development National Center for Environmental Assessment. The project title was Development of Bayesian Updating Techniques to Incorporate Mechanistic Information Across Species. Grant Number : EPA award number R-82954701.

References

Berry, D. A. and Stangl, D. K. (1996). *Bayesian Biostatistics.* Marcel Dekker, New York.

Crofton, K. M. (1998). Re-analysis of thyroid hormone data from subchronic perchlorate study submitted by Springborn Laboratories. (SLI study No. 3455.1)[memorandum with attachments to Annie Jarabek].

Crofton, K. M. and Marcus, A. (2001). Re-analyses of perchlorate hormone data from the 1998 ERD. USEPA, Office of Research and Development, Research Triangle Park, NC. [memorandum with attachments to Annie Jarabek], Oct. 15.

Müller, P. (1999). Simulation based optimal design. In Bernardo, J. M., Berger, J. O., Dawid, A. and Smith, A., editors, *Bayesian Statistics 6*. Oxford Univ. Press, Oxford, U.K.

Schmitt, K. A., Small, M. J., Choi, T., Schervish, M. J., Baird, S. J. S. and Jarabek, A. M. (2004). Estimating health effects of perchlorate using Bayesian hierarchical models for multiple endpoints. *In progress* .

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Springborn Laboratories Inc. (1998). A 90-day drinking water toxicity study in rats with ammonium perchlorate: amended final report. Springborn Laboratories, Inc., Spencerville, OH. study no. 3455.1.

U.S.Environmental Protection Agency (1998). Perchlorate Environmental Contamination : Toxicological Review and Risk Characterization based on emering information. Review draft. Washington, DC: Office of Research and Development; report no. NCEA-1-0503.

Wilson, J. D. (2001). Advanced methods for dose-response assessment: Bayesian approach-final report. Resources for the Future, Washington, DC. Discussion paper 01-15.

Wolf, D. (2001). Erratum to the report of the peer review of the thyroid histopathology from rodents and rabbits exposed to ammonium perchlorate in drinking water. USEPA, Office of Research and Development, Research Triangle Park, NC. [memorandum to Annie Jarabek], Oct. 26.

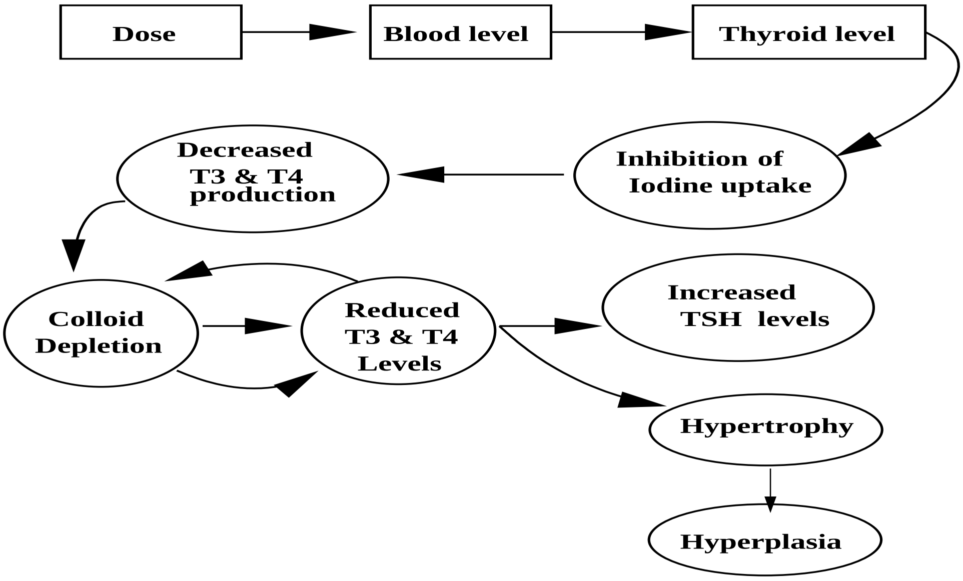

Figure 1. Perchlorate's effect

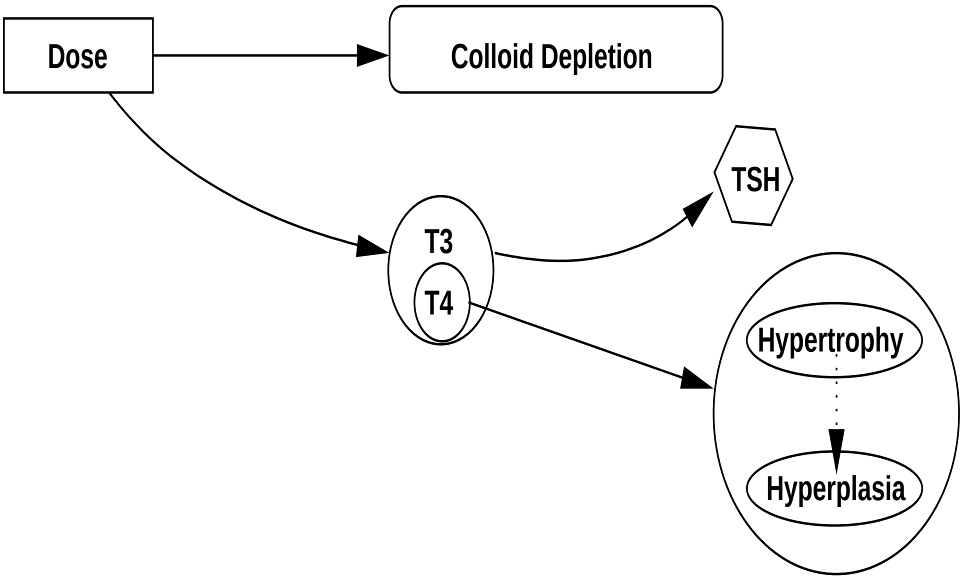

Figure 2. DAG (Directed Acyclic Graph) - mechanistic model

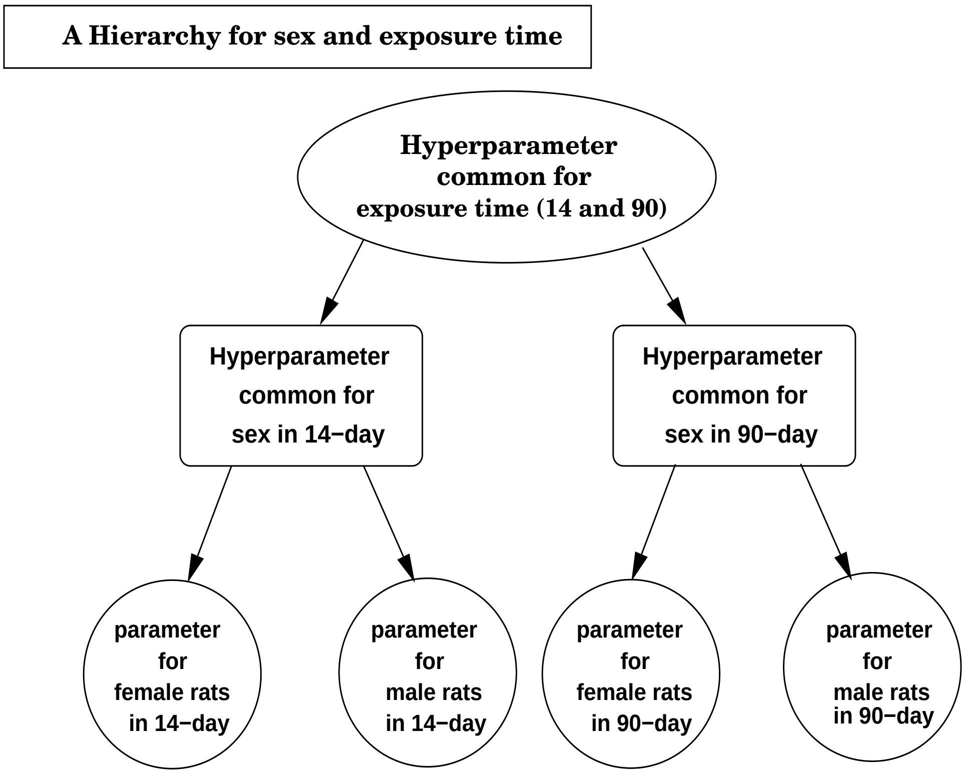
A Hierarchy for sex and exposure time
Hyperparameter common for exposure time (14 and 90)
Hyperparameter common for sex in 14−day
Hyperparameter common for sex in 90−day
parameter for female rats in 14−day
parameter for male rats in 14−day
parameter for female rats in 90−day
parameter for male rats in 90−day

Figure 3. The hierarchical structure for the prior distribution

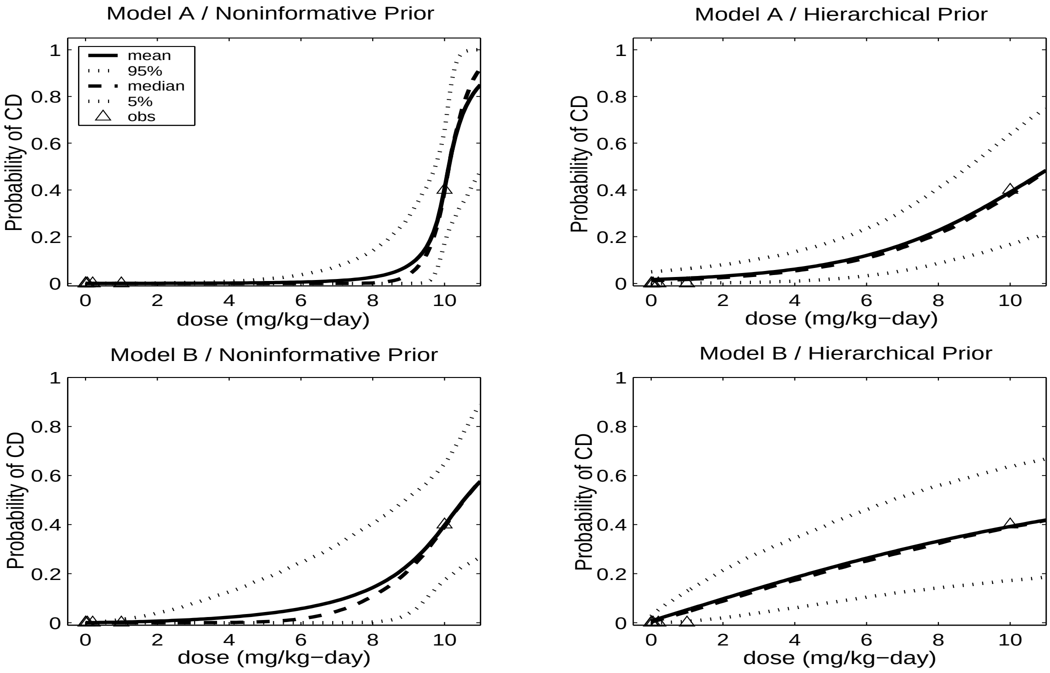

Figure 4. Probability of CD for female rats in 14 day study. First row is Model A, while second row is Model B. First column is noninformative prior, while second column is hierarchical prior. Solid lines are means, dashed lines are medians of the 2000 MCMC simulations. Dotted lines show pointwise 90% credible intervals for the probability of CD at each dose.

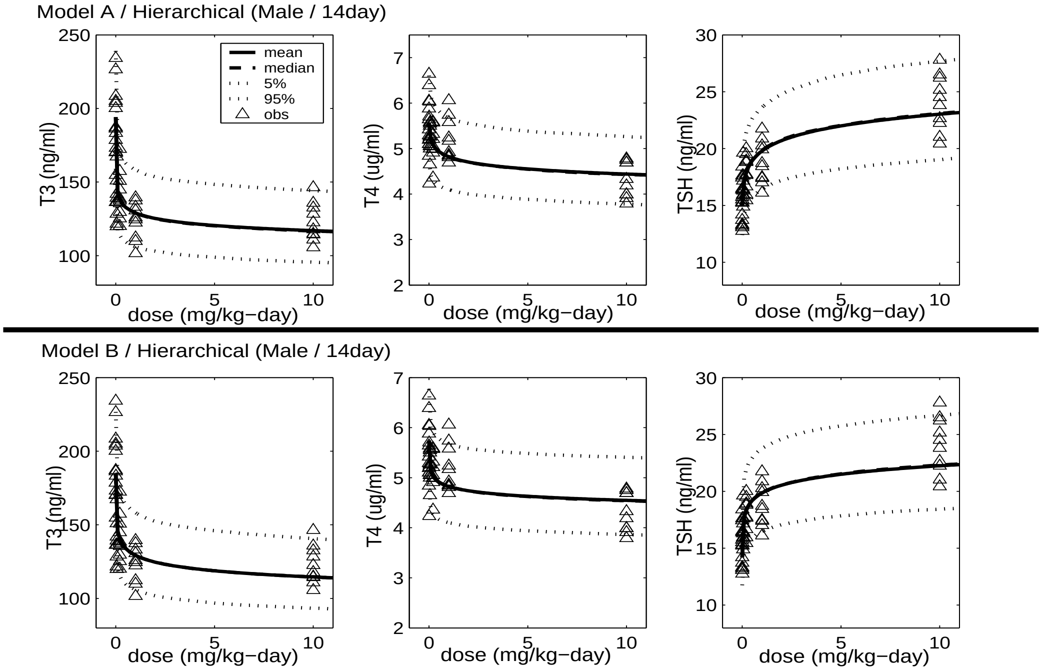

Figure 5. Effects of perchlorate dosage on T3, T4 and TSH for male rats in 14 day study. First row is Model A using separate δ's, while second row is Model B using common δ. Solid lines are means, dashed lines are medians of the 2000 MCMC simulations. Dotted lines show pointwise 90% credible intervals of MCMC simulations at each dose.

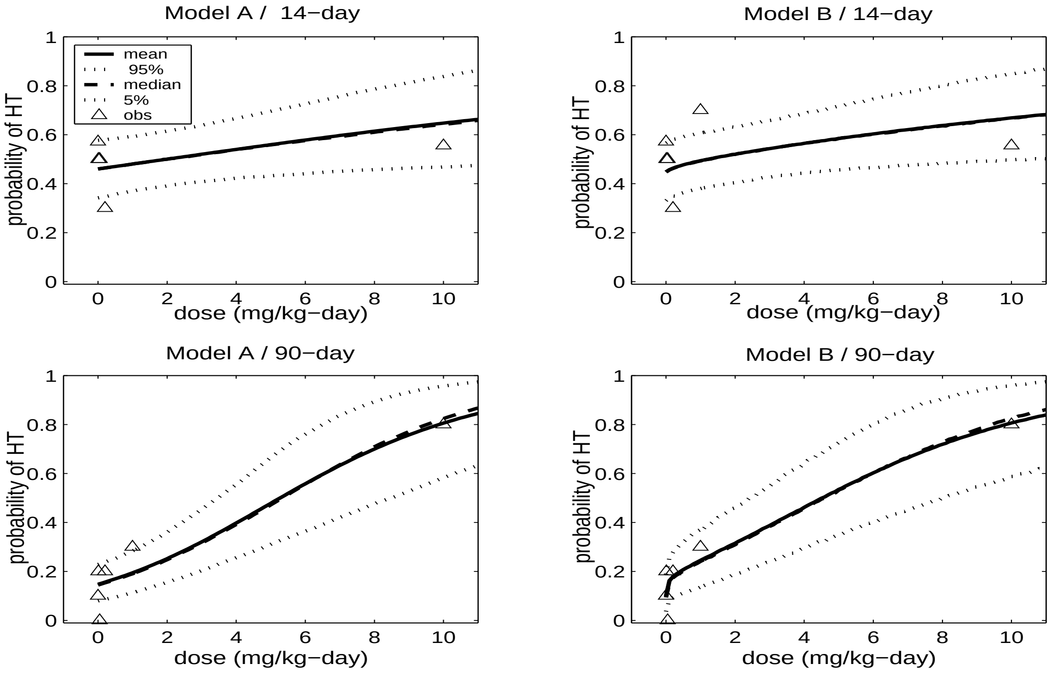

Figure 6. Probability of HT for male rats in 90-day study for Models A and B in both 14 and 90-day studies

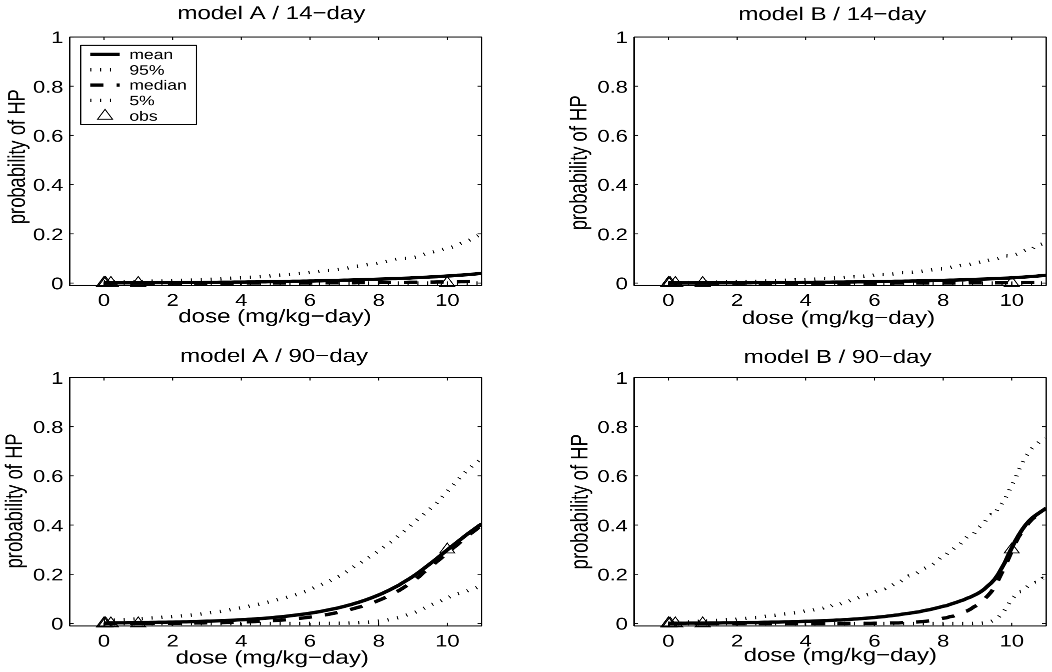

Figure 7. Probability of HP for female rats in the 14-day study and the 90-day study crossed with different models

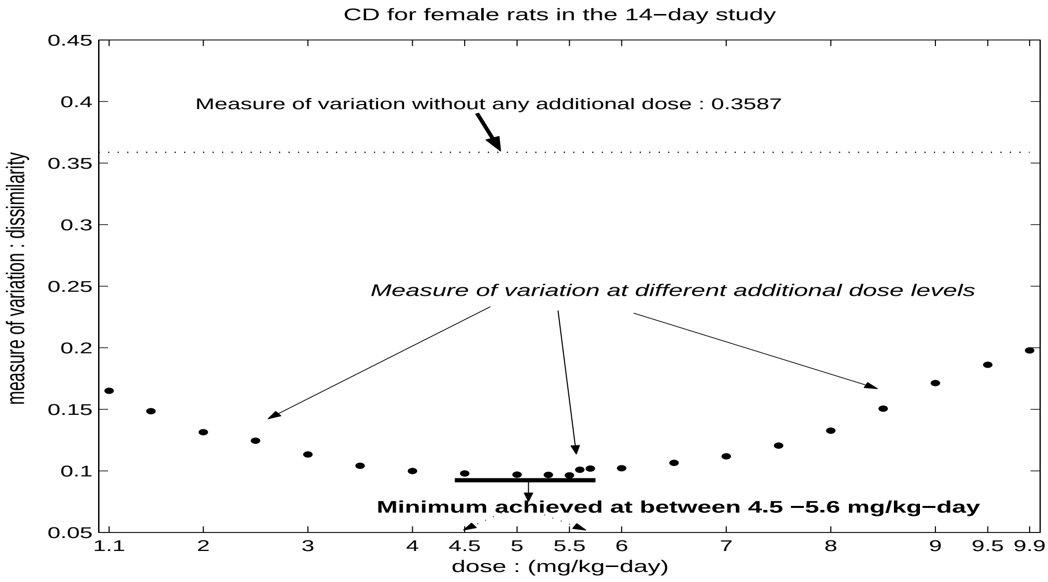

Figure 8. Finding the optimal design point : e.g. CD for female rats in the 14-day study

Table 1

BIC values for two best models at each step

Step	Model	BIC (14 day)	BIC (90 day)	BIC (Total)
1	dose	42.76032	40.01666	82.77698
	$\log(\text{dose} + \delta)$	55.63924	46.00789	101.6471
2	$\log(\text{dose} + \delta)$	-502.4542	-528.8523	-1031.307
	$\log(\text{dose} + \delta_c)$	-496.8823	-530.1983	-1027.081
3	dose	71.83565	65.3227	137.1583
	$\text{dose} + \log(\text{T4})$	82.73969	76.22593	158.9656

Note : In step 2, δ denotes a separate dose increment for each blood hormone endpoint, whereas δ_c denotes a common dose increment for all three.

Table 2

Histopathology data from Wolf (2001). Here, n stands for the number of rats with observed histopathology

	14-day								90-day							
	Female				Male				Female				Male			
dose	n	CD	HT	HP	n	CD	HT	HP	n	CD	HT	HP	n	CD	HT	HP
0	10	0	1	0	7	3	4	3	10	0	0	0	10	0	1	0
0.01	10	0	1	0	10	0	5	1	10	0	0	0	10	1	2	0
0.05	10	0	0	0	10	5	5	2	10	1	3	0	10	1	0	0
0.2	10	0	0	0	10	5	3	0	10	1	2	0	10	1	2	0
1.0	10	0	1	0	10	4	7	4	10	1	1	0	10	2	3	0
10.0	10	4	8	0	9	7	5	2	10	4	5	3	10	9	8	4