Appendix 6: Statistical Evaluation of Data and Model Development

A data set for statistical analyses was formed by matching dose group data for the biological endpoints chosen for statistical analyses (**Section 3.1.1**, body of the report) with the corresponding compositional data. The criteria used to select the studies from which the dose group data was captured are discussed in **Section 3 2.4** (body of the report). Individual compositional and toxicity studies were matched using the sample identification number, ensuring that the toxicological and corresponding analytical information from the same samples had been used. See **Appendix 5** for a listing of the specific study matches and **Appendix 10**. for a tabulation of the toxicological data and corresponding analytical information that was used to develop the final models."

A6.1 Modeling Methods (**Section 3.4.1**, body of the report)

For each of the biological endpoints selected for modelling, a mathematical model of the dose response curve was developed. The model for each endpoint was developed independently, using an iterative process. Models were developed using general regression analysis methods with the biological endpoint (e.g. fetal body weight) as the dependent, or predicted variable, and relevant toxicological study design variables (e.g. control group response, litter size, sex) and the test material variables (e.g. PAC weight percentages) as the independent, or predicting, variables.

The development of each model went through several phases of exploratory data analysis. The initial phases were graphical, and sought to determine which transformations would be useful, how the control group would be utilized, which independent variables were available and useful.

The exploratory data analysis led to the selection of the forms of the final model.

Sections A61.1 through A61.3 describe the selection of the form of the dependent variables, the choice of the independent variables, and the final model forms. For clarity, these three subjects are described separately, in practice all three were done together since each influences the others. The fourth section explains why individual PAC terms are used in the model rather than the sum of the PAC weights. The fifth section explains the attempted use of a factor analysis to reduce the number of independent variables.

A6.1.1 Choice of Dependent Variables (**Section 3.4.1.1,** body of report)

The dependent variables were the responses of a dosed group (dose > 0) for each of the eleven endpoints selected for final modeling as described in **Section 3.3, body of report**. Final statistical models were developed for the biological endpoints shown in **Table A6-1**. The eleven endpoints selected for final modelling (see **Table A6-1**) were chosen based on biology and toxicology and not on statistical modelling. Control group responses were used as independent variables in the models (see **Section 3.4.1.2, body of report**).

For the repeat-dose studies, the dose-group response was the mean response of all the animals in the dose group in a specific study. For the developmental toxicity studies, the dose-group response was the mean of the means of all the litters in a dose-group in a specific study. Thus, if a study had 3 dosed groups, and data were available for each dose groups, there would be 3 data points for each modelled endpoint. The number of dependent variable data points used to develop the model for a specific endpoint is shown in **Table A6-2**.

The dependent variable was the observed response rather than either the ratio of the dose group response to the control group response, or the 'percent response relative to control'. The use of a covariate (the control group response as an independent variable) allowed more flexible modelling of the response and, in most cases, resulted in a more stable estimate. If the model was developed with percent response relative to control as the dependent variable, the response would be the ratio of two random variables. This ratio can vary widely, especially when the

control group value is likely to be small. For example, when measuring the number of resorptions a seemingly small change of the numerical value in the denominator can result in a large change in the ratio, i.e. if the number of control group resorptions decreases from 2 to 1 in a litter the percent of resorptions relative to control will double. All models were developed using both the covariate method and the percent response relative to control method. The covariate models were more stable and had regression fit diagnostics at least as good as the percent response relative to control models. The model-predicted responses from the covariate models can be converted to percent response relative to control predictions by dividing the predicted value by the control group response. This is demonstrated in **Appendix 8.**

Table A6-1. Biological Endpoints Selected for Final Mathematical Characterization

Study Type	Effect
Repeat-dose	Thymus weight (absolute)
	Platelet count
	Hemoglobin concentration
	Liver weight (relative)[a]
Developmental (Prenatal)	Maternal Thymus weight (absolute)[c]
	Fetal body weight
	Live fetuses/litter
	Percent Resorptions
Developmental (Postnatal)	Pup body weight (PND[b] 0)
	Total pups/litter (PND[b] 0)
	Live pups/litter (PND[b] 0)

[a] relative to terminal body weight
[b] PND = postnatal day
[c] Maternal thymus weight was selected as an endpoint for use in testing the statistical models using alternative data sources (**Section 3.4.3, body of report**). In order to do this it was necessary to develop final models for this endpoint, even though the TG had earlier decided not to characterize maternal endpoints in developmental toxicity studies.

A6.1.2 Choice of Independent Variables (**Section 3.4.1.2**, body of report)

Analytical Variables
The PAC content of the test samples in the various company toxicity reports had been determined using a variety of analytical techniques (see **Section 2.3.2,** body of the report and **Appendix 1)**. Preliminary models were built using four compositional data sets. Final models were developed using only Method 2-derived PAC data. The Method 2 data set was selected for use in the final models based on the model fit characteristics of the preliminary models. See **Section A.6.2** for details of the results of the modeling and the basis for the choice of Method 2 data set.

Toxicity Study Design Variables
A set of independent variables related to study design was included in each model. For the repeat-dose studies, the set included variables such as:

- dose level normalized to milligrams of applied compound per kilogram of animal body weight per day ($mg/kg_{bw}/day$),
- duration of dosing,
- control group response, based on the mean responses of the control groups in the TG's data set and,
- sex of the treated animals.

For the developmental toxicity studies, the independent variables included:

- dose level normalized to milligrams of applied compound per kilogram of animal body weight per day ($mg/kg_{bw}/day$),
- control group response, based on the mean responses of the control groups in the TG's data set,
- litter size,
- number of implantation sites,
- number of animals, or pregnant dams, or litters per dose group, and
- body weight.

Not all variables were eligible, available, or appropriate for all models; however, terms for dose level and control group response were always included in the model building process. All responses were means calculated in a similar manner to that described in **Section A61.1**.

A6.1.3 Model Forms (**Section 3.4.1.3**, body of report**)**

As noted in **Section A6.1**, the models for each endpoint were developed independently. The basic model form was a general linear regression model, with a possible transformation of the dependent variable, with the dose group response as the dependent variable, the control group response as an independent variable (covariate), and a selection of independent variables described in **Section A6.1.2**. In the model building process, for each endpoint, ordinary least squares and maximum likelihood methods were used. Several mathematical forms of each model were considered based on transformations of the dependent and independent variables. Based on the residuals pattern, several transformations were tested with the dependent variables including the natural logarithm, the exponentiation of the variable, several power transformations, and the probit transformation. Similar transformations were applied to the independent variables.

As previously noted (**Section A61.1**), the dependent variable was the observed response rather than the ratio of the dose group response to the control group response, or the percent response relative to control. The several forms of incorporating the control group were used in the model building process and the use of a covariate (the control group response as an independent variable) was shown to allow more flexible modelling of the response, and, in most cases, resulted in a more stable estimate.

During model development, models were developed based on both linear regression using ordinary least squares (OLS) methods (Draper and Smith, 1981) and mixed-effects models (Pinheiro and Bates, 2002) using maximum likelihood (ML) methods. These OLS methods assume all observations are independent. However, in our data the assumption of independence may not be achieved because there are usually from 2 to 10 dose group data points from a particular study (and the toxicological studies themselves may have had some commonality). The assumption of independence is important for assessing significance levels of terms in the model, but has little effect on the estimated coefficients. The mixed-effects models account for the relationships of dose groups within a study, and are theoretically preferable in the current situation.

The OLS method is widely known among researchers and software for expanding and applying the models is readily available. The ML methods are slightly more difficult to use and the consideration of accounting for within group variances in predictions may be difficult. We considered both models and found that, as expected, the models based on the two methods had similar forms, and coefficients, but the variance estimates for the mixed-models were smaller than for the OLS models. The difference in the overall variance estimates between the two will depend on the degree of difference between the petroleum substance study group means and the within petroleum substance study group variances.

For each model, both the fit of the data and a model, and the error term can be assessed by the correlation and the residual standard error, respectively. While it is known that the ML methods are not optimized for the correlation and standard error as are the OLS methods, the ML methods do provide a reasonable method of comparison. **Table A6-2** shows the correlation (r) and residual standard error (se) for the optimum models from the two estimation methods.

Table A6-2. Comparison of Model Fitting Characteristics for OLS and Mixed Model Analyses

Study Type	Dependent Variable	n	OLS		Mixed Model (ML)	
			r	se	r	se
Repeat –dose toxicity studies	Thymus Weight (absolute)	89	0.89	0.04	0.89	0.04
	Platelet Count	91	0.96	81.5[b]	0.96	69.0[b]
	Hemoglobin Concentration	104	0.95	0.55	0.96	0.42
	Liver Weight (relative[a])	103	0.94	0.20	0.95	0.16
Developmental Toxicity Studies (Prenatal)	Maternal Thymus Weight (absolute)[c]	34	0.91	0.04	0.97	0.02
	Fetal Body Weight	62	0.96	0.10	0.98	0.06
	Live Fetuses/Litter	62	0.99	0.84	0.98	0.71
	Percent Resorptions	62	0.97	0.25	0.99	0.04
Developmental Toxicity Studies (Postnatal)	Pup Body Weight (PND[d] 0)	62	0.93	0.16	0.93	0.13
	Total Pups/Litter (PND[d] 0)	62	0.96	1.09	0.96	0.85
	Live Pups/Litter (PND[d] 0)	62	0.96	1.17	0.96	0.91

a relative to terminal body weight
b The large se for platelets results from platelet counts being large absolute numbers, thus giving rise to a seemingly large standard error about the line of best fit for the data.
c Maternal thymus weight was selected as an endpoint for use in testing the statistical models using alternative data sources. (**Section 3.4.3**). In order to do this it was necessary to develop final models for this endpoint, even though the TG had earlier decided not to characterize maternal endpoints in developmental toxicity studies.
d PND = postnatal day

The equivalence of the fits of the models from the two methods can be seen in the similarity of the correlations, while the slightly smaller errors of predictions with the mixed models can be seen in the smaller se values. The differences between models from the two methods are small, and at the current stage of model development, the simplicity of the OLS methods is preferred over the mixed-effects models. At some later stage of development, it may be reasonable to estimate responses of the developed models with mixed-effects techniques, but currently there are no compelling reasons why the OLS methods should not be used.

Using the OLS method, several preliminary models were developed for each endpoint. The final comparison among competing OLS models was based on the overall model multiple correlation coefficient (r) and the error mean square (EMS). As before, these measures were selected because, among their other characteristics, the r value is a measure of the closeness of the observed and model predicted values, while the EMS is related to the width of the confidence interval of the predicted value. During the model building process, we did not adhere strictly to the optimization of the correlation and standard error, but considered the overall reasonableness of the model, concentrating more on the fit of the model near the critical region (rather than near the no effect region), but not allowing a few outliers to drive the form of the model. In general, the goal was to develop a model that was both a good descriptor and a good predictor.

For each model, a series of technical tests were conducted to assure the model was optimum for the data set at hand. The building process was by definition, an iterative process where model forms were postulated and tested with various diagnostics. Based on the results of the diagnostics and an understanding of biology and toxicology, a model was then altered by adding or removing terms and/or transforming terms, or in some cases trying nonlinear model forms. The transformations included the standard set of logarithm, exponent, trigonometric, power, and probit transformations. The diagnostics included residual plots, and a statistical evaluation of the magnitude and effect of influence points. The influence points are data points that have a statistically large effect on the estimated coefficients and statistical significance of the coefficients. The residuals were tested for a normal distribution at the 0.01 significance level by the Shapiro-Wilk test (Shapiro and Wilk, 1965). All models met the Shapiro-Wilk criterion except the percent resorptions model and the total pup and live pup models. For the resorptions model there was one sample point (sample 86270, dose=30 mg/kg/day) that had a very low response; deleting the one value allowed the residuals to meet the criterion and decreased the se by about 30%. For the total pup and live pup models dropping same two sample points in each model (sample 89645, dose=1000 mg/kg/day and sample F-275, dose=250 mg/kg/day) allowed the residuals to almost meet the criterion. The unusual data points were not removed from the data set and all reporting includes these data points. Additionally, plots of the observed and predicted values from a model were developed to evaluate the adequacy of the model and to look for outliers and other possible anomalies, see **Figure A6-4**.

The initial model building included a categorical (nominal) term that described the HPV category of the test sample (aromatic extract, crude oil, etc). This term was an important term in almost all the models. Because this measure is not a physical property of the sample, and we wanted the terms to be measurable properties, we used logistic regression and discriminant function techniques to develop an alternative term that is based on the physical properties. The analyses indicated that a collection of terms involving the individual PAC concentrations and the interaction of PAC ring 4 with PAC ring 5 was a good predictor of the HPV category. Therefore, the models all contain this interaction term.

All data were tested for outliers from the model. A level I outlier was defined as an observation with a studentized residual greater than 2.58 in absolute value. The studentized residual considers the magnitude of the residual (difference between the observed data and model predicted value) and the standard deviation of the prediction (a smaller standard deviation yields a larger studentized residual). The choice of 2.58 results in flagging theoretically 1% of the

observations (because the residuals are normally distributed). Of the 793 residuals 24 were flagged. Many of these were flagged because of a very small standard error. A level II outlier was defined as a level I outlier where the prediction error was at least 50% of the observed value. **Table A6-3** lists the 24 Level I outliers and **Table A6-4** lists the 1 Level I outlier; only the Level II outliers will be discussed.

Table A6-3. Level I outliers

Obs	Endpoint	Sample No	Study No	Dose ($mg/kg_{bw}/day$)	Sex	Observed	Pred	se	Student residual	% Err
1	RD_Thymus	86181	64165	125	male	0.14	0.24	0.03	-2.90	68.96
2	RD_Thymus	86181	64165	125	female	0.14	0.23	0.03	-2.70	64.02
3	RD_Platelet	86187	61737	125	male	591.00	864.65	-77.98	-3.51	46.30
4	RD_Platelet	86187	61737	500	female	258.00	181.05	27.06	2.84	29.83
5	RD_LiverToBW	86181	64165	8	female	2.84	3.54	0.19	-3.65	24.48
6	RD_LiverToBW	86187	61737	500	female	5.60	5.83	0.07	-3.34	4.10
7	RD_LiverToBW	86187	61737	125	male	4.45	3.90	0.20	2.78	12.28
8	RD_LiverToBW	89645	63834	500	male	3.78	3.30	0.17	2.89	12.75
9	RD_Hemoglobin	86271	63456	500	male	11.50	12.92	0.40	-3.52	12.34
10	RD_Hemoglobin	86484	62710	30	male	13.00	14.19	0.40	-3.00	9.19
11	RD_Hemoglobin	89106	63266	1000	male	12.30	13.37	0.41	-2.58	8.70
12	RD_Hemoglobin	86271	63456	500	female	14.30	13.12	0.40	2.93	8.24
13	RD_Hemoglobin	86484	62710	30	female	15.20	14.05	0.40	2.90	7.58
14	RD_Hemoglobin	89106	63266	1000	female	14.80	13.57	0.41	2.96	8.30
15	Pre_Resorp	86270	62328	30	NR	0.00	0.08	0.24	-4.57	18520.05[a]
16	Pre_LiveFet	86193	64643	250	NR	12.60	14.48	0.62	-3.04	14.96
17	Pre_LiveFet	89106	63263	500	NR	2.80	4.06	0.47	-2.66	44.91
18	Pre_FetalWt	89646	63848	500	NR	3.40	3.50	0.04	-2.96	3.07
19	Post_TotalLit	89645	63837	2000	NR	8.60	10.29	0.53	-3.17	19.60
20	Post_TotalLit	F-275	66149	500	NR	4.40	5.87	0.53	-2.76	33.31
21	Post_TotalLit	F-275	66149	250	NR	13.30	10.77	0.98	2.59	19.06
22	Post_TotalLit	89645	63837	1000	NR	15.00	12.36	0.98	2.71	17.63
23	Post_LiveLit	89645	63836	2000	NR	7.80	9.63	0.57	-3.22	23.49
24	Post_LiveLit	89645	64282	1000	NR	14.80	11.95	1.04	2.72	19.23

[a] The resorptions percent error result is very large because the observed value is small (0.006).
NR = not relevant

Table A6-4. Level II outliers

Obs	Endpoint	Sample No	Study No	Dose ($mg/kg_{bw}/day$)	Sex	Observed	Pred	se	Student residual	% Err
1	RD_Thymus	86181	64165	125	male	0.14	0.24	0.03	-2.90	68.96
2	RD_Thymus	86181	64165	125	female	0.14	0.23	0.03	-2.69	64.02
3	Pre_Resorp	86270	62328	30	NR	0.01	0.08	0.24	-4.57	>1000[a]

[a] The resorptions percent error result is very large because the observed value is small (0.006).
NR = not relevant

For all 3 level II outliers, the model over predicted. For study 86181 the predicted thymus weight response was double the observed response at the 125 mg/kg/day dose for both sexes. The control group value for this study was low; the lowest applied dose response was 20 to 25% larger than the control response. The unusual control group value is the likely cause of the poor predictions for this study. The prediction for the percent resorptions for study 86270 appears to be too high, but it is similar to other materials; a possible explanation is an unusual response rather than an unusual prediction. The large percent error is based on the very small observed value.

Overall, these outlier data points appear to be reasonable data values (not recording or experimental errors), and are retained in the model building process, and we did not investigate the effect of these few data points on the models.

Using the criteria described above, the results of the various model forms indicated that linear models (models where the independent, or explanatory, variables are additive) provided a good description of the observed data and non-linear models would not improve the fit of the model to the data. The testing also indicated that the most stable models were based on predicting the dose group response directly (not as a ratio to the control group), with the control group response as an independent variable. The predicted ratio could be developed from the predicted direct dose group response by dividing by the control group response.

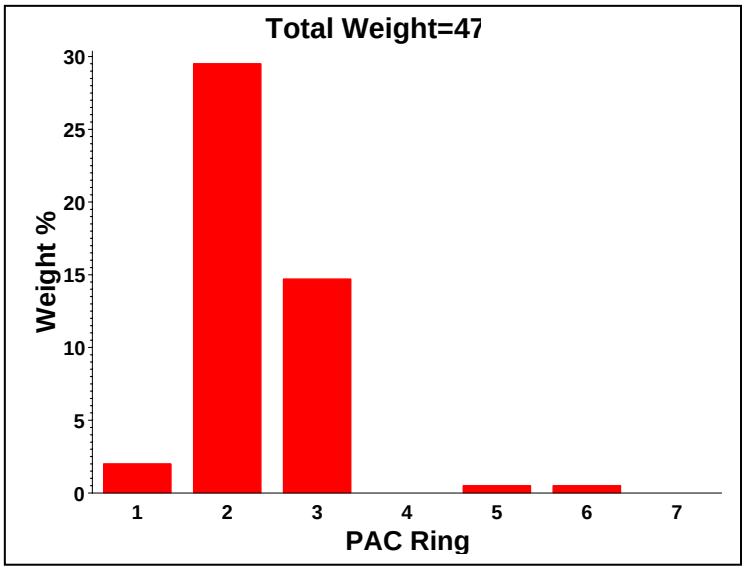
In summary, the primary goal of the statistical model building process was to identify and characterize the relationship between PAC content and SIDS endpoints (Objective 2 of the Task Group). The models were developed independently for each endpoint considering the biology, toxicology, and statistical aspects of the available data. The models were developed to be as simple as possible, but adequate. A model that fit the data well in the critical region was preferred to one that fit well at the extremes. The critical region is that region where the response changes from normal to adverse. After all the models were independently developed, some alteration was made to make them look similar while not sacrificing the integrity of the individual models. The amount of alteration was fairly small, which is an indication of the statistical consistency of the modelling process but is not meant to indicate anything about the underlying biological mechanism. The terms for all 7 of the individual PAC terms were kept for all models to avoid the problem of fitting the model to a specific data set and not have it generalizable to new data, and to minimize the tendency to inspect individual PAC terms for hints of the biological mechanism. The only exception to this is in the Pup Body Weight Model involving dose squared times PAC content that includes only 4 of the PAC terms; inclusion of the other 3 caused the model to be unstable.

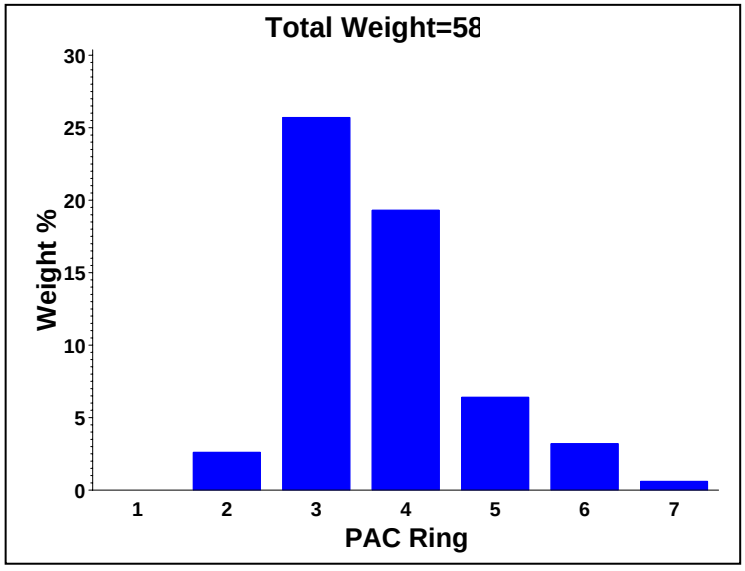
The models met the objective of characterizing the relationship between PAC content and SIDS endpoints as seen in the correlation between the observed and predicted data (a mean r of 0.94 and minimum r of 0.87). It is difficult to have models that are better descriptors, no matter what form they have.

A6.1.4 Individual PAC Terms (**Section 3.4.1.4**, body of report**)**

The final models were developed using the weight percent of each of the 1-ring through 7 ring compounds in the test material (referred to as the PAC profile). These values were obtained with analytical Method 2 (see **Appendix 1** for detailed description). It is not adequate to consider the total percent weight of the 1-7 ring compounds because the total percent weight does not describe the toxic response pattern when compared to the PAC profile of the petroleum substance. For example, consider the weight percentage of the ring components in the 2 samples from the dataset depicted in **Figure A6-1**. Both samples have similar total weight percent of 1-7 ring compounds but their PAC profiles differ.

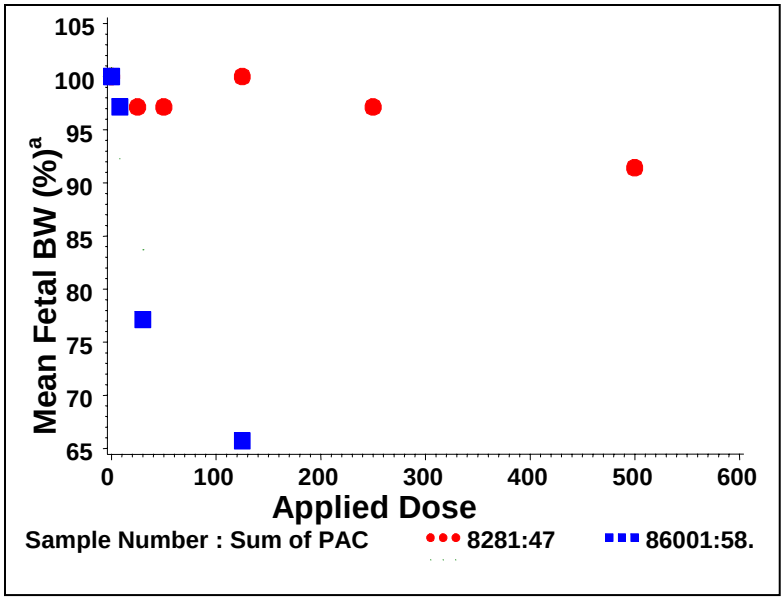
Figure A6-1. Weight Percent of 1- through 7-Ring Compounds of Two Substances with a total PAC extract Weights of 47 and 58 Percent





As shown in **Figure A6-2**, the biological responses relative to applied dose for materials with similar *total* weight percentage but with different PAC profiles are very different. The ratios of observed mean fetal body weight to the control mean fetal body weight for the two substances from **Figure A6-1** are plotted in **Figure A6-2**. Results from samples 8281 and 86001, which have similar total aromatic ring weight percentages, have different biological responses. Sample 8281 has a relatively shallow dose-response curve, whereas sample 86001 has a much steeper dose-response curve, indicating that total PAC weight alone is a poor predictor of response. The mathematical model predictions for these two samples closely agree with the observed data indicating the usefulness of the models, see the appropriate plots in Appendix 9.

Figure A6-2. Observed Mean Fetal Body Weight Ratio vs. Applied Dose for Two Substances with Total PAC Extract Weights of 47 and 58 Percent



[a] Mean fetal body weight is expressed as a percentage of the control values

A6.1.5 Factor Analysis

During model development one of the goals was to minimize the number of independent variables and reduce the degree of correlation among them (the problem of multicolinearity). A factor analysis was done on the PAC ring 1 through 7 weight percentage data. A three factor solution was selected that accounted for 80% of the variance for the Method 2 derived PAC rings 1 through 7 weight percentage data. Amongst all models tested, regression analyses models with the factor scores did not fit the data as well as the models using the individual ring weight percentages.. Based on these results, the individual ring weight percentages and two-term interactions among the weight percentages were used for model development.

A6.2 Preliminary Model Results (**Section 3.1.2**, body of report**)**

The analytical reports selected for use in assessing relationships between PAC content and mammalian toxicity contained compositional data derived from several methods, each identifying different chemicals or groups of chemicals. Consequently, the TG undertook to assess whether different compositional data would have varying degrees of usefulness in this assessment. As part of this effort, the TG evaluated whether data on S-PACs would prove more useful than data on 1-7 ring PACs.

To evaluate the utility of the various types of compositional data, preliminary mathematical characterizations were developed to provide an early indication of the degrees of accuracy that could be achieved with the different types of compositional data. These initial characterizations were made with linear regression models and a range of dependent and independent variables. **Table A6-5** provides summary results (n, r, and EMS) of these initial characterizations.

The data in **Table A6-5** indicate that for almost all endpoints, Method 2 model results have a larger r value and a smaller EMS value when there are a comparable number of data points on which to build the model. Consequently, the model results based on chemical analysis Method 2 were judged “better” than those produced by models using other compositional data sets.

Since the development of the models was an iterative process, the results from these preliminary model building efforts do not correspond exactly to the results of the final models seen in **Table A6-6**.

Table A6-5. Summary of Preliminary Results for Linear Regression Models with Four Compositional Data Sets

	Compositional Data Set											
	Method 1 (1- to 5-Ring Compounds)			**Method 2** (1- to 7-Ring Compounds)			**S-PAC** (From Method 1)			**Carbazoles** (From Method 5)		
Measure	n	r	se	N	r	se	n	r	se	n	r	se
Repeat-dose												
Liver wt. (relative) [a]	102	0.93	0.08	124	0.94	0.07	82	0.84	0.11	8	0.84	0.08
Thymus wt. (absolute)	70	0.85	0.13	92	0.90	0.11	68	0.75	0.15	8	0.89	0.09
RBC count	104	0.54	0.13	128	0.54	0.13	86	0.30	0.14	10	0.05	0.12
Platelet count	96	0.90	0.10	112	0.91	0.09	76	0.70	0.17	8	0.81	0.12
Hemoglobin concentration.	104	0.92	0.04	128	0.75	0.07	86	0.61	0.08	10	0.92	0.04
Hematocrit	104	0.54	0.17	128	0.60	0.17	86	0.30	0.20	10	0.06	0.12
Developmental (Prenatal)												
Percent resorptions	55	0.95	1.52	66	0.98	1.08	52	0.72	3.17	53	0.88	0.72
Resorptions/litter	55	0.96	1.48	66	0.98	1.07	52	0.75	3.01	53	0.89	0.76
Live fetuses/litter	55	0.92	0.12	66	0.98	0.07	52	0.68	0.20	53	0.90	0.05
Fetal body wt.	55	0.89	0.04	66	0.95	0.03	52	0.64	0.06	53	0.81	0.03
Maternal thymus wt (absolute).	28	0.94	0.10	35	0.95	0.09	28	0.74	0.17	0		
Developmental (Postnatal)												
Total pups/litter PND 0	72	0.87	0.11	77	0.93	0.09	57	0.50	0.20	79	0.84	0.13
Live pups/litter PND 0	72	0.89	0.11	77	0.92	0.10	57	0.50	0.21	79	0.83	0.14
Pup body wt. Day 0	72	0.85	0.04	77	0.83	0.04	57	0.54	0.05	79	0.69	0.04

[a] relative to body weight
wt weight
n number of dose groups
r multiple correlation coefficient
se standard error, calculated as the square root of the error mean square
PND postnatal day

The results of the various model forms indicated that linear models (models where the independent, or explanatory, variables are additive) provided a good description of the observed data and non-linear models did not improve the fit of the model to the data. The testing also indicated that the most stable models were based on predicting the dose group response directly (not as a ratio to the control group), with the control group response as an independent variable. The predicted ratio could be developed from the predicted direct dose group response by dividing by the control group response.

A6.3 Final Models

A6.3.1 Final Model Results (**Section 3.4.2**, body report**)**

The correlation and standard error (r and se) values in **Table A6-6** are for the final models that are based on the observed response, not the ratio of the response of the dosed group to control group. As these models are the next iteration of the models from **Table A6-5**, the r and se values from **Table A6-5 and Table A6-6** cannot be compared.

Table A6-6. Final Modeling Results Using the Method 2 Data for PAC Weight %

Study Type	Dependent Variable	Transformation on Dependent Variable	n	r	se
Repeat –dose toxicity studies	Thymus Weight (absolute)	None	89	0.89	0.04
	Platelet Count	None	91	0.96	81.5[b]
	Hemoglobin Concentration	None	104	0.95	0.55
	Liver Weight (relative)[a]	None	103	0.94	0.20
Developmental Toxicity Studies (Prenatal)	Maternal Thymus Weight (absolute)[c]	None	34	0.91	0.04
	Fetal Body Weight	None	62	0.96	0.10
	Live Fetuses/Litter	None	62	0.99	0.84
	Percent Resorptions	Probit	62	0.97	0.25
Developmental Toxicity Studies (Postnatal)	Pup Body Weight (PND[d] 0)	None	62	0.93	0.16
	Total Pups/Litter (PND[d] 0)	None	62	0.96	1.09
	Live Pups/Litter (PND[d] 0)	None	62	0.96	1.17

[a] relative to terminal body weight
[b] The large se for platelets results from platelet counts being large absolute numbers, thus giving rise to a seemingly large standard error about the line of best fit for the data.
[c] Maternal thymus weight was selected as an endpoint for use in testing the statistical models using alternative data sources (**Section A6.5.3**). In order to do this it was necessary to develop final models for this endpoint, even though the TG had earlier decided not to characterize maternal endpoints in developmental toxicity studies.
[d] PND = postnatal day

The magnitudes of the correlations in **Table A6-6** are large for this type of data; the minimum correlation is 0.89 with the remaining being above 0.90. Partial explanations for the large correlations are that:

1. Each data point is a group mean response often with at least 10 observations in the group. This reduces the variability of each point, hence amplifying the correlation.

2. *A priori* selection criteria for the data points resulted in a somewhat homogeneous data set that also reduced the variability.
3. Models were selected to maximize the correlation.

The final models were rigorously tested (**Section A6.4)** to ensure that the model results and corresponding correlations were not spurious, based on bias, confounding, or affected by model specifications.

A6.3.2 Final Model Equations

The final models for the 11 endpoints considered are linear in the coefficients and of a similar form. An example of the algebraic form of the model based on the live fetus/litter count is:

$$Live\,Fetus\,Count = \alpha + \beta_1 \cdot control\,live\,fetus\,count + \beta_2 \cdot number\,implants + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

where:

- α is the intercept,
- β_1 and β_2 are coefficients for the biologically based independent variables,
- PAC_i is the weight percent measure for i^{th} ring component of the PAC, and
- η, γ_i, and ξ_j are coefficients for the analytic based independent variables.

The forms of the eleven final models are described in **Table A6-7**. The table lists dependent variable and its transformation (if any), the selection of biologically-based independent variables and the selection of analytically-based independent variables. The models always include PAC concentration terms of the form:

$$\eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i$$

The last column in **Table A6-8**, labeled "Additional PAC Terms Included" uses an "**I"** to indicate if the model included an interaction term of the form:

$$\sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

and a "**2"** to indicate if the model included a PAC square term of the form:

$$\sum_{k=1}^{7} \nu_k \cdot dose \cdot PAC_k^2$$

Section A6.5.4 provides the coefficients and complete forms for all the models listed.

Table A6-7. Forms of the Eleven Final Models

Study Type	Dependent Variable	Transformation on Dependent Variable	Covariate (independent biological variable)	Other Independent Biological Variables	Additional PAC Terms Included
Repeat-dose toxicity studies	Thymus Weight (absolute)	None	CG[a] Thymus Weight	Body Weight, Sex	**No**
	Platelet Count	None	CG[a] Platelet Count	Sex, Duration	**I**
	Hemoglobin Concentration	None	CG[a] Hemoglobin Concentration	Sex, Duration	**I**
	Liver Weight (relative[b])	None	CG[a] Liver to BW Ratio	Body Weight, Sex, Duration	**I**
Developmental toxicity studies (Prenatal)	Maternal Thymus Weight (absolute)[c]	None	CG[a] Maternal Thymus Weight	None	**No**
	Fetal Body Weight	None	CG[a] Fetal Body Weight	None	**I**
	Live Fetuses/Litter	None	CG[a] Live Fetuses/Litter	N implants	**I**
	Percent Resorptions	Probit	Probit (CG[a] PctRes)	None	**I**
Developmental toxicity studies (Postnatal)	Pup Body Weight (PND[d] 0)	None	CG[a] Pup Body Weight	1/Total Litter Size	**I 2**
	Total Pups/Litter (PND[d] 0)	None	CG[a] Total Pups/Litter	N implants	**I 2**
	Live Pups/Litter (PND[d] 0)	None	Live Pups/Litter	N implants	**I 2**

[a] CG = Control Group

[b] relative to terminal body weight

[c] Maternal thymus weight was selected as an endpoint for use in testing the statistical models using alternative data sources (**Section A6.5.3**). In order to do this it was necessary to develop final models for this endpoint, even though it had been decided a full assessment of such endpoints and their relation to PAC content using the final model was outside the scope of this project.

[d] PND = postnatal day

I Interaction term of the form $\sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$

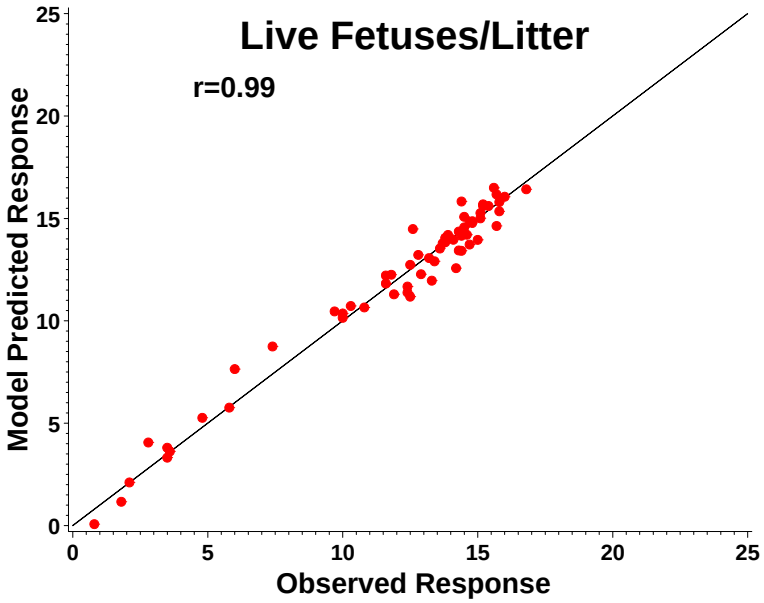
2 interaction term of the form $\sum_{k=1}^{7} \nu_k \cdot dose \cdot PAC_k^2$

A6.3.3 Final Model(s) Fit

The accuracy of the fits of the final models can best be seen in plots of observed data points versus the predicted data points. In these types of plots, an individual data point would represent what is observed for a single dose group of an experiment and what is predicted from the mathematical model. The optimum would have all points along the straight line, representing data points in which the observed value equals the predicted data.

As an example, the plot for the live fetus/litter model is shown in Figure **A6-3**. The correlation coefficient for this model is 0.99, which is an indication of a very good model fit.

Figure A6-3. Plot of Observed vs. Model Predicted Live Fetus/Litter Count



Similar plots for all 11 final models and their corresponding r values are shown in **Figure A6-4**, with the live fetuses/litter plot repeated for completeness. Note that for all of the models the r values are greater than 0.89.

Fig A6-4. Observed vs. Model Predicted Data Points for All Final Models

Repeat-dose toxicity studies

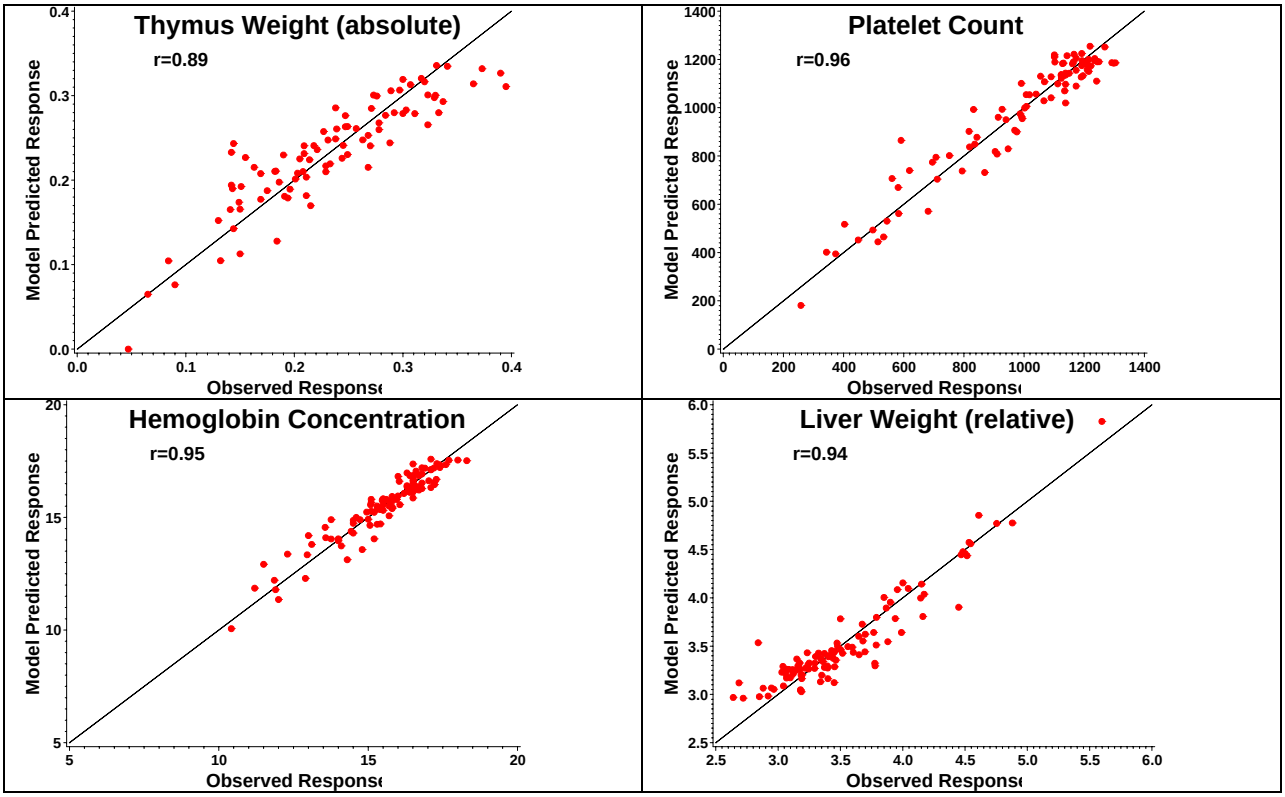
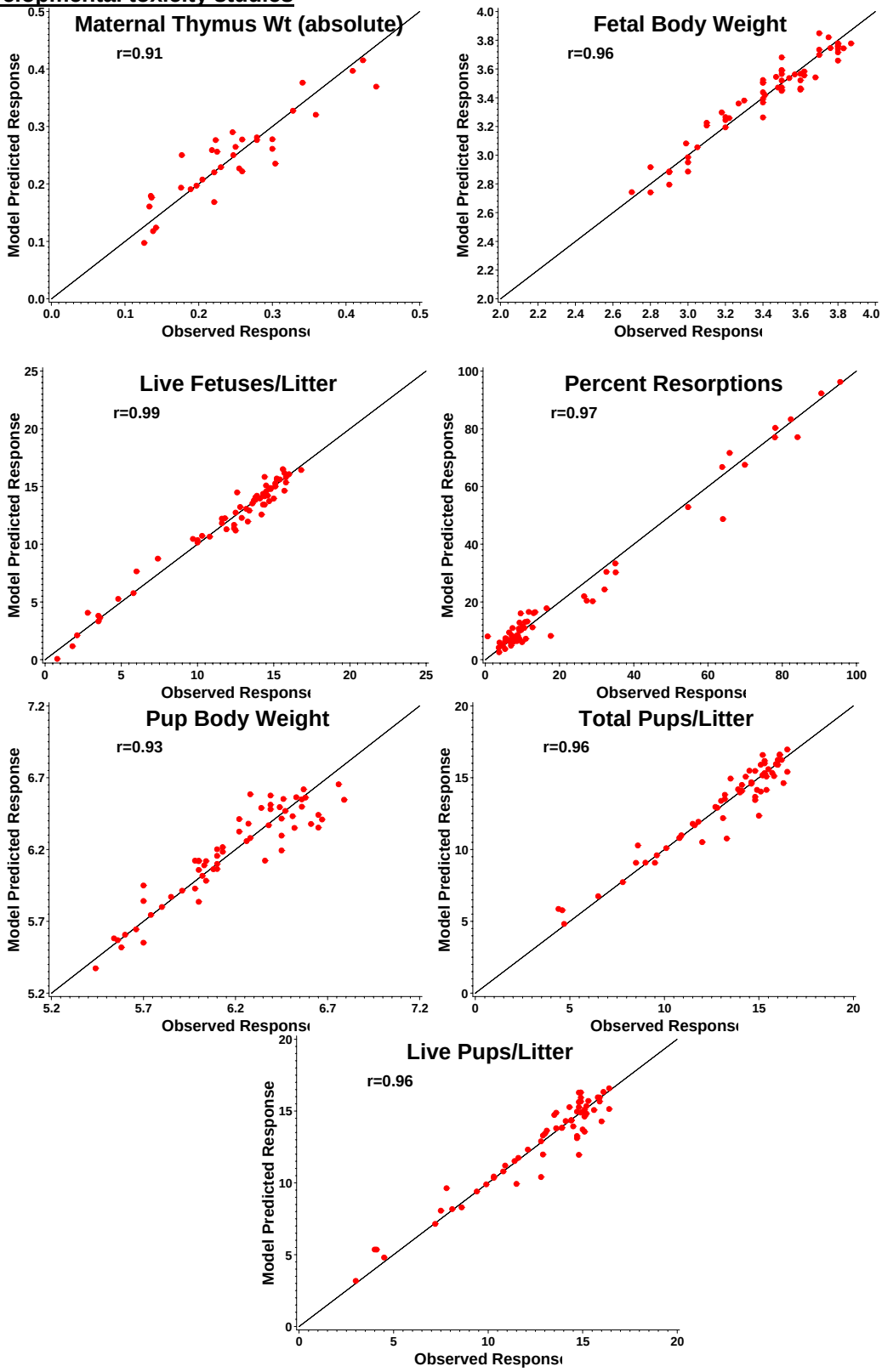


Fig A6-4 (cont.). Observed vs. Model Predicted Data Points for All Final Models

Developmental toxicity studies



A6.4 Interpolation and Extrapolation

The concepts of interpolation and extrapolation need to be defined and understood for the discussion of model testing and prediction. The concept of interpolation and extrapolation applies when using a statistical model to predict a new response data point from a new set of independent variables. The predicted data point is called an *interpolated* data point if the predicted data point is developed from independent variables that are all within the range of the independent variables used to develop the model. Conversely, the new predicted data point is called an *extrapolated* data point if some, or all, of the independent variables are outside the range of the independent variables used to develop the model. For the models that have been developed the independent variables of concern are the 7 PAC concentrations and the applied dose. The biological variables (such as body weight, control group response, etc) are also of concern, but are largely at the discretion of the researcher when predicting new responses; that is, when the researcher has to assume a body weight value in the model, some historic value will be used.

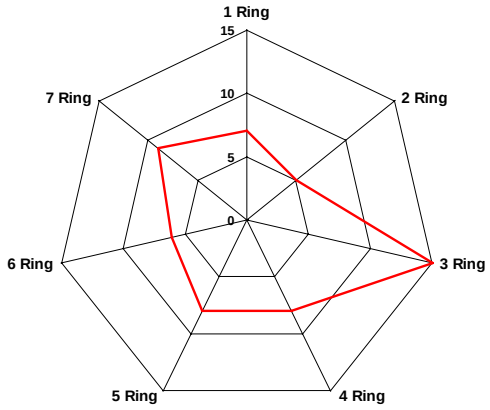
The concepts of interpolation and extrapolation can be illustrated by plotting the data on a spider or radar plot.

Consider a hypothetical petroleum substance with the following PAC ring weight percent concentrations:

PAC_1	PAC_2	PAC_3	PAC_4	PAC_5	PAC_6	PAC_7
7	5	15	8	8	6	9

This substance could be plotted as shown in **Figure A6-5.**

Figure A6-5. Spider Plot of the PAC Profile of a Hypothetical Petroleum Substance

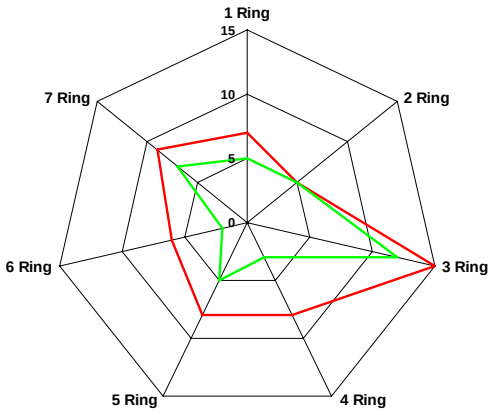


A different petroleum substance might have PAC ring weight percent concentrations of:

PAC_1	PAC_2	PAC_3	PAC_4	PAC_5	PAC_6	PAC_7
5	5	12	3	5	2	7

This substance is plotted in green as shown in **Figure A6-6.**

Figure A6-6. Spider Plot of the PAC Profile of a Hypothetical Petroleum Substance (green) that would Result in an *Interpolated* Predicted Data Value



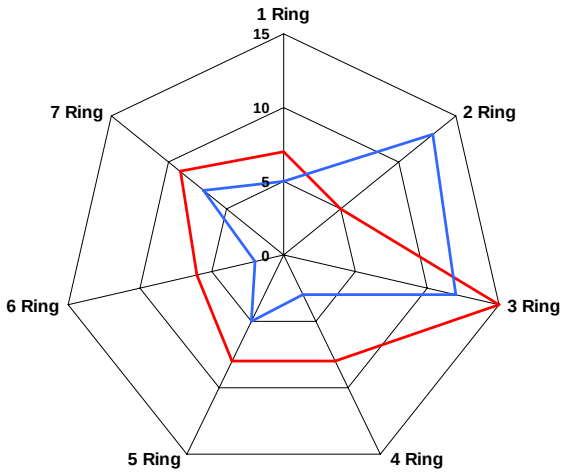
Because all the ring concentrations for the new substance (green plot) are within the range of the original substance (red plot) a biological value predicted from a model using this PAC profile would be considered an *interpolated* predicted data point Relative to the original (red) substance.

A third petroleum substance ring might have PAC weight percent concentrations of:

PAC_1	PAC_2	PAC_3	PAC_4	PAC_5	PAC_6	PAC_7
5	13	12	3	5	2	7

This third substance is plotted in blue as shown in **Figure A6-7.**

Figure A6-7. Spider Plot of the PAC Profile of a Hypothetical Petroleum Substance (blue) that would Result in an *Extrapolated* Predicted Data Value



Because the concentration for the PAC_2 ring in this third substance (blue plot) is greater than that of the original substance (red plot), a biological value predicted from a model using this PAC profile would be considered an *extrapolated* point relative to the original (red) substance.

The concepts of interpolation and extrapolation between two substances can be generalized to considering if the PAC profile of a new (untested) substance is interpolated or extrapolated relative to the set of substances that were used to build a model.
When classifying a new sample or substance as interpolated or extrapolated both the PAC profile and the applied dose of the new sample are assessed for interpolation or extrapolation.

For a new substance to be *interpolated* relative to an existing substance data set it must 'be between the largest and smallest existing substance data point'; so the new substance must

1. be interpolated in the 7-ring PAC sense to at least one substance in the data base (i.e. it must be smaller than the maximum)
2. be extrapolated in the 7-ring PAC sense to at least one substance in the data base (i.e. it must be larger than the minimum)
3. have applied dose values that are between the largest and smallest applied doses of all substances in the data base.

If the new substance violates any of these three criteria, any predictions made using the PAC profile of the substance will be an *extrapolated* predicted data point relative to the data base used to build the model.

Note that since each model developed in this study may have used a different base data set of substances when it was developed, a new substance may be extrapolated relative to one model's data set and interpolated relative to a different model's data set, or conversely.

A6.5 Model(s) Testing

An important component of model building is to test, or validate, the model's predictive ability. This testing is necessary to demonstrate the utility of the models. The models that were developed in this project were tested in three ways:

1. Using holdout sample data.
2. Using 'nonsense' data.
3. Using an alternate data set.

These tests are necessary to demonstrate the utility of the models. The holdout samples indicated the models were accurate and robust when predicting data not used in developing model coefficients when the predicted point was within the range of the observed data, and sometimes were not accurate for values very different from the base data set. This problem is often found with these types of models and is called the problem of extrapolation; further discussion appears in the "Limitations" section of the report (**Section 4.5** and **Section A6.4)**. The model results are firmly based on the input data as demonstrated by the poor results from nonsense data (a type of negative control). Finally, a model developed from data on one effect was able to accurately predict other outcomes subject to the limitations of the extrapolation problem as shown in the alternate data set section.

The next sections provide the details of these tests.

A6.5.1. Model(s) Testing – Hold Out Samples

A standard method of testing a statistical model is to develop the model on a subset of the available data, and then apply the model to the data not used to develop the model. This process is called hold-out sample validation or data-splitting validation (Harrell, 2001). The data used to develop the model is called the training data, the remaining data is the test or holdout data.

To demonstrate the model validity the data-splitting technique was expanded by having the method replicated 100 times; each replication used a different set of training and hold out data selected from the full data set.

The method is demonstrated with the absolute thymus weights from the repeat-dose studies. In the base data set used for the PAC analysis there were 92 observations for the repeat dose thymus weight. For each replication approximately 70% of the data points are selected to build the model (training data) and the remaining, approximately, 30% is used as test data (hold out data). The percentages are approximate because the selection process chooses each point with probability 70% rather than choosing 70% of the sample. In each of the 100 replicates, the specific data points in the 70% and 30% groups are different.

The results from the 100 replications are shown in the observed vs. predicted plots. **Figure A6-8** shows the model observed and predicted data for the training data (n=6,284). **Figure A6-9** provides a plot of the model observed and predicted data for the hold out data (n=2,616).

Figure A6-8. Observed and Predicted Points of the Training Sample of Absolute Thymus Weight Data from Repeat-dose Studies

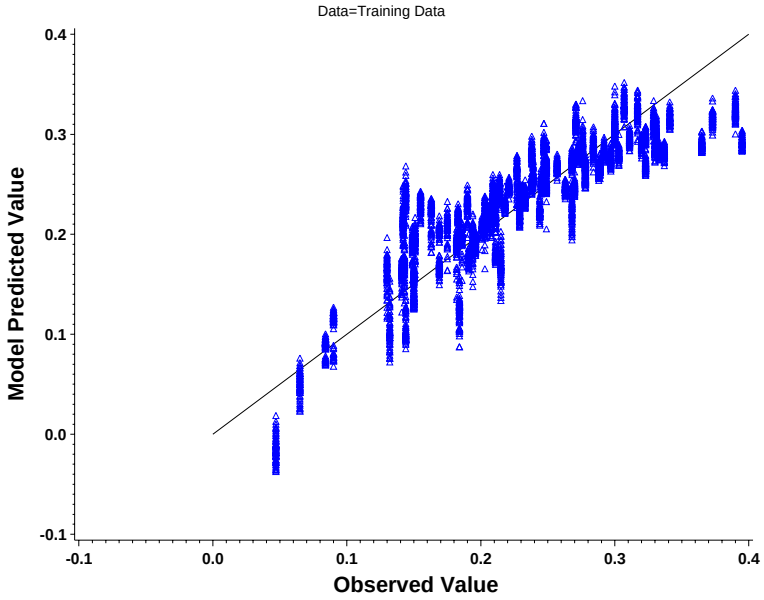
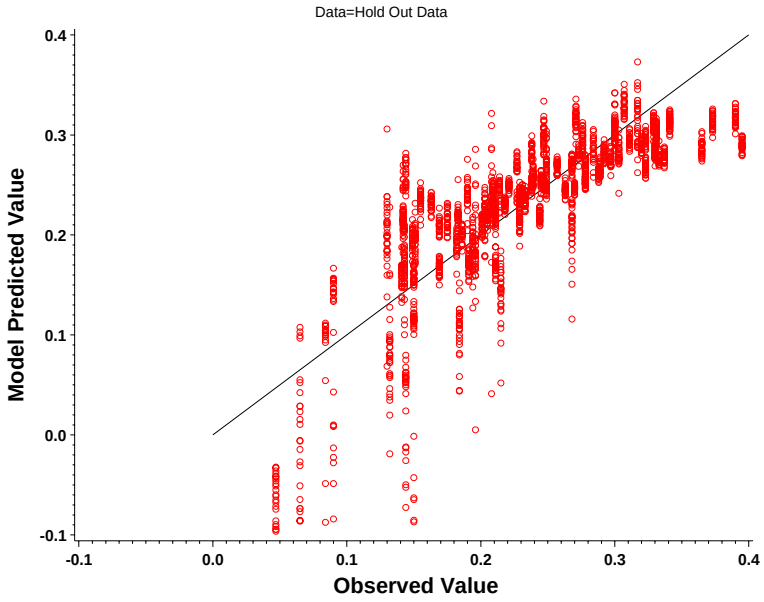
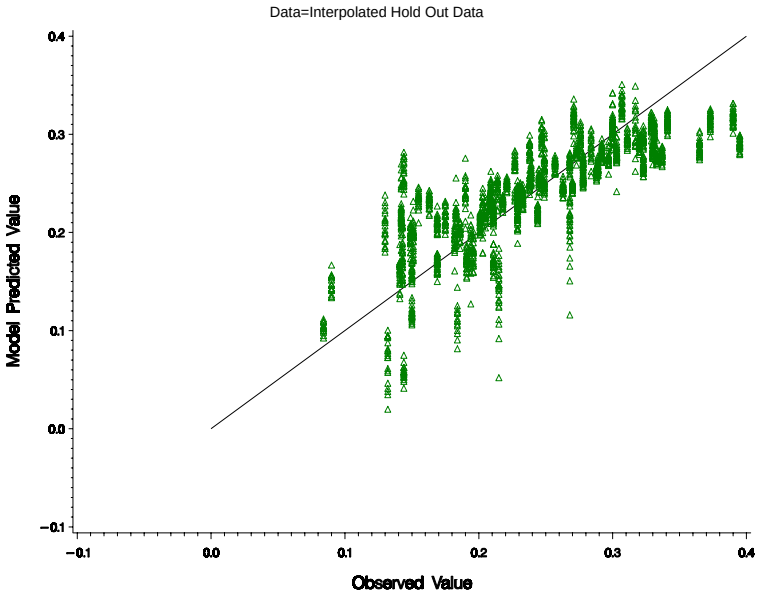


Figure A6-9. Observed and Predicted Points of the Hold-Out Sample of Absolute Thymus Weight Data from Repeat-dose Studies



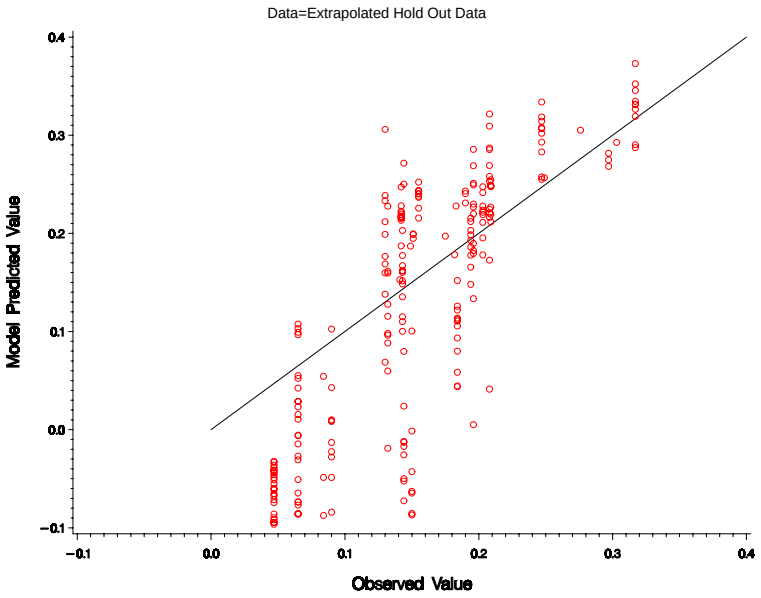
As can be seen in **Figure A6-9,** some of the predicted data points in the hold-out data set are "unreasonable" in that they are not close to the observed data point, as shown by their distance from the 45-degree line of equal values. However, because of the way the holdout data were sampled some of these holdout data points are interpolated points and some are extrapolated data points. **If** the interpolated and extrapolated holdout data points are plotted separately **(Figures A6-10** and **A6-11),** the "unreasonable" data points are the extrapolated data points, whereas the interpolated data points provide reasonable and accurate predictions.

Figure A6-10. Observed and Predicted Points of the Interpolated Holdout Sample of Absolute Thymus Weight Data from Repeat-dose Studies



.

Figure A6-11. Observed and Predicted Points of the Extrapolated Holdout Sample of Absolute Thymus Weight Data from Repeat-dose Studies



These series of plots (**Figures A6-8 to A6-11**) demonstrate that the predictions from the model for the original data set ('training data') and for the *interpolated* holdout data are good in that the predicted values are close to the observed values. However, model predictions for the *extrapolated* holdout data are mixed, sometimes good and sometimes inaccurate.

A6.5.2 Model(s) Testing – Nonsense Data

A model's usefulness can be tested by determining model performance using values for the independent variables (PAC compositional data) that were *not* associated with the outcome (observed effect).
If a model does not fit well using this "nonsense data" (i.e. produces relatively low r values), it is a clear indication that the model behavior is based on information in the data, and is not a result of chance.

The hemoglobin concentration model was tested using the Nonsense Method of model testing. In the original model there were 104 data points with an r value of 0.95.

1. The response data (hemoglobin concentration) and the corresponding values of the independent variables (PAC compositional data) were randomly shuffled and a new model was fit. The process was repeated 100 times. The resulting models had a mean r = 0.60, with a minimum and maximum of 0.35 and 0.81, respectively. However, because the model incorporates the control group hemoglobin concentration value, part of the seemingly large r (0.60) from the shuffled data is based of the relation between the control and dosed hemoglobin concentration in the ANCOVA model. Without the ANCOVA control group, the r value for the real data was 0.88 (lower than the 0.95 correlation developed with the ANOVA model) and for the 100 shuffled data runs was 0.36 (minimum 0.13 and maximum 0.66). This is an indication that the model will not fit random data well as it fit the real data.
2. A similar series of shuffles was done, but the randomization was restricted to sets within the same petroleum category (or class) and sex of the respondent. These shuffles selected from a smaller group of possible matches and resulted in some matches that were the same as the original ordering, so the resulting correlations should be higher than the fully random shuffles, but less than the observed correction. For these restricted shuffles the mean and range of 100 replicates was 0.55 with a minimum and maximum of 0.40 and 0.72, respectively.

These results from the nonsense method of testing, while seemingly good, are still far from the observed r value of 0.95. These relatively low r values from the nonsense data are a clear indication that the model behavior is based on information in the data, and do not result from chance.

A6.5.3 Model Testing – Alternate Data Sources

A model's usefulness can be tested by determining the model performance with data from another data source similar to the source used to develop the model. That is, use test data from a different but related effect. . Examples include:

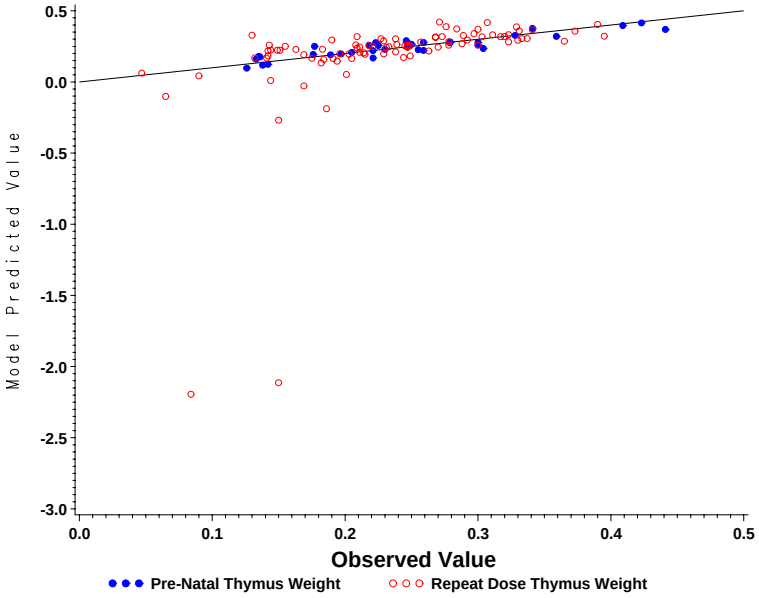
- using the model for repeat-dose absolute thymus weight to predict prenatal maternal absolute thymus weight,
- using the model for prenatal fetal body weight to predict postnatal pup body weight, and
- using the model for prenatal live fetuses per litter to predict postnatal total litter size

and the reverse order of each of these 3 examples.

Consider a model for maternal absolute thymus weights that was developed using the data from the *prenatal* studies. This model can be applied to the thymus weight data from the *repeat-dose* studies. If the prenatal thymus model is adequate, the repeat-dose data predictions should be as accurate in predicting the repeat-dose data as the original predictions were in predicting the prenatal data. That is, the prenatal model should work as well with repeat-dose study data as it did on the data it was developed for.

Figure A6-12 shows the plot of observed vs. predicted data points (blue dots) for the prenatal maternal absolute thymus weight data (the points used to develop the model) and the points from the repeat-dose studies (red circles) that were predicted by the prenatal model.

Figure A6-12. Observed and Predicted Prenatal Maternal Thymus Weight Data Points Based on the Model Developed from the Prenatal Data Applied to Prenatal and Repeat-dose Data



Several of the points in **Figure A6-12** have very poor predicted values, some are even negative (a biological impossibility). The poorly predicted points are (new) repeat-dose data points that were predicted by the pre-natal model. The reason some of the (new) repeat-dose data points are poorly predicted is that they are extrapolated points relative to the pre-natal model. Recall, a data point is identified as extrapolated if the PAC weight percent for any ring is greater than the corresponding ring for all substances in the base data set used to develop the original model (in this case, the data used to develop the pre-natal model), or lower than the corresponding ring for all substances, or if the applied dose is greater than the largest applied dose for all substances of the group (see **Section A6.4** for a fuller explanation).

Figure A6-13. Observed and Predicted Prenatal Maternal Thymus Weight Data Points Based on the Model Developed from the Prenatal Data with Repeat-dose Data Identified as Interpolated or Extrapolated

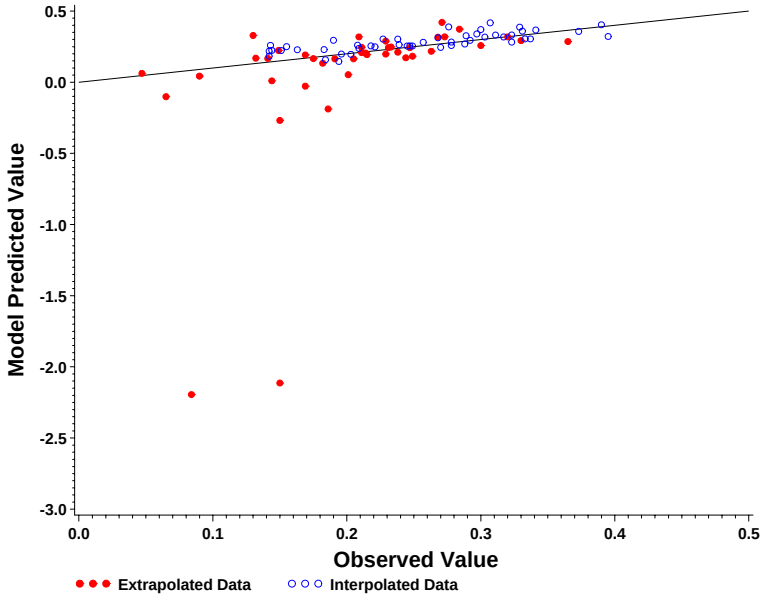


Figure A6-13 shows which of the predicted data points from the repeat-dose data are interpolated and which are extrapolated. It can be seen that all of the poorly fitting data points are extrapolated, and that some of the extrapolated data points fit very well. This is a demonstration that, while interpolated data points can be reliable predictors, extrapolated data points may or may not be accurate.

The pattern relating to interpolation and extrapolation works well for predicting thymus weights using the original model and an alternate data set. **Table A6-8** shows the prediction results for models used to predict alternate data with all alternate (new) data points and with only the interpolated or extrapolated points. The column labelled "r for base model" is the correlation between the observed and predicted data points based on the original model and the original data set used to develop it; the column labelled "r for *all* new data" is the corresponding correlation for all the alternate (new) test data using the original model. The last two columns subdivide the alternate (new) data into the extrapolated and interpolated data.

For example, the second row of **Table A6-8** shows the correlation of the observed and predicted data for the prenatal maternal absolute thymus model applied to the actual prenatal maternal thymus data (i.e. the model fitted to its own data) has an r of 0.91. When the prenatal maternal absolute thymus model is applied to the repeat-dose thymus data the r is 0.43. When the prenatal maternal absolute thymus model is applied to only the interpolated data the r increases to 0.77. When the prenatal maternal absolute thymus model is applied to only the repeat-dose data judged to be "extrapolated", the r is 0.43. With the exception of the pup weight models (rows 4 and 5), **Table A6-8** shows that the models predict data from alternate interpolated data very well (r>0.62) but can be very poor for extrapolated data (often with r<0).

The forms of the models involved in the testing with alternate data sources can explain some of the differences/similarities in the correlations seen in **Table A6-8**. The repeat-dose and prenatal

thymus models are similar (the basic model with no interaction term), so it would be expected that the predictions using data

The forms of the models involved in the testing with alternate data sources can explain some of the differences/similarities in the correlations seen in **Table A6-8**. The repeat-dose and prenatal thymus models are similar (the basic model with no interaction term), so it would be expected that the predictions using data from alternate sources should work well (the correlations for the alternate source interpolated data are 0.84 and 0.77). The same holds true for the prenatal fetus count and postnatal litter size: both utilize the basic model with an interaction term, so again it would be expected that the predictions using data from alternate sources should work well (correlations for the alternate source interpolated data are 0.62 and 0.80). However, the models for the prenatal fetal weight and postnatal pup weight are different. The prenatal fetal weight model is the basic model with an interaction term, but the postnatal pup weight is the most complex model of the set (it has the model form of the prenatal fetal weight, but includes additional terms of the reciprocal of litter size and the PAC concentration squared). The more complicated postnatal pup weight model reflects the complex data associated with the model, so it is not unreasonable to have the simpler prenatal fetal weight model not do well with the complex postnatal pup weight data, and conversely. Even so, the correlations of 0.50 and 0.24 for the cross model interpolated data are acceptable, they may be considered moderately poor only in relation to the very good correlations seen in the other four situations.

Two conclusions can be drawn from these analyses:

1. the models fit the data used to develop the models very well (r at least 0.89), and
2. the models can predict new interpolated data well (r usually greater than 0.50), and can be very poor for extrapolated data.

Table A6-8. Prediction Results for Models Used to Predict Alternate Data

Endpoint	Alternate Data Set	r for base model (n[a])	r for *all* new data (n[a])	r for new data – interpolated predictions only (n[a])	r for new data – extrapolated predictions only (n[a])
Thymus	Repeat-Dose Model Predicting Prenatal Data	0.89 (89)	0.81 (34)	0.84 (30)	0.99 (4)
	Prenatal Model Predicting Repeat-Dose Data	0.91 (34)	0.43 (89)	0.77 (48)	0.43 (41)
Weight	Prenatal Fetal Weight Model Predicting Postnatal Pup Weight	0.96 (62)	-0.14 (62)	0.50 (36)	-0.21 (26)
	Postnatal Pup Weight Model Predicting Prenatal Fetal Weight	0.93 (62)	-0.35 (62)	0.24 (34)	-0.34 (28)
Count	Prenatal Fetus Count Model Predicting Postnatal Litter Size Data	0.99 (62)	-0.20 (62)	0.62 (36)	-0.25 (26)
	Postnatal Litter Size Model Predicting Prenatal Fetus Count	0.96 (62)	0.26 (62)	0.80 (34)	0.23 (28)

[a] number of data points used

A.6.5.4 Model Coefficients

The following section presents the algebraic model forms and coefficients for the models described in **Table A6-7**.

The models were developed for describing specific endpoints. The independent (predicting) variables were selected from a set of variables that were eligible for inclusion in the model. Because the data do not come from a complete statistically planned experiment, the estimated of the coefficients are correlated amongst themselves. Because these coefficient estimates are correlated, comparisons of coefficients within a model or between models cannot be made. That is:

1. While an individual coefficient in a model may indicate the relative effect of an independent variable on the response, it is not correct to assume that changing the value of the independent variable will change the response in the direction of magnitude associated with the coefficient. This point indicates that the models can not be used to 'engineer' a petroleum product with required characteristics. For example, if a product has a high concentration of 3-Ring PAC, and a large positive coefficient for Ring 3 in the model, it is not necessarily true that if the concentration of the 3-Ring PAC is reduced the response will in fact be reduced. This is a consequence of the model being a descriptive model rather than a predictive one.
2. Comparisons of the sign and magnitude of coefficients of an independent variable across model endpoints are not meaningful because (a) the predicted endpoints are not always on the same scale or may be transformed, and (b) for reasons similar to those described in the preceding point.

A6.6 Repeat-dose Final Models:

$$Thymus\,Weight = \alpha + \beta_1 \cdot Control\,Thymus\,Weight + \beta_2 \cdot Body\,Weight + \beta_3 \cdot sex + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i$$

$$Platelet\,Count = \alpha + \beta_1 \cdot Control\,Platelet\,Count + \beta_2 \cdot sex + \beta_3 \cdot Study\,Duration + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

$$Hemoglobin\,Concentration = \alpha + \beta_1 \cdot Control\,Hemoglobin\,Concentration + \beta_2 \cdot sex + \beta_3 \cdot Study\,Duration + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

$$Liver\,to\,Body\,Weight\,Ratio = \alpha + \beta_1 \cdot Control\,Liver\,to\,BW\,Ratio + \beta_2 \cdot Body\,Weight + \beta_3 \cdot sex + \beta_4 \cdot Study\,Duration + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

Table A6-9. Repeat-dose Final Model Coefficients

	Model Coefficients			
	Thymus Weight (absolute)	Platelet Count	Hemoglobin Concentration	Relative Liver weight[a]
Intercept	-0.1427299571	285.1076574	-2.707705095	0.8835433748
Control Value[b]	0.3398713256	0.7135032	1.161820602	0.7439996114
Sex	0.1067550657	9.4178397	0.318079711	-0.0338966562
Study Duration		0.7800292	-0.004567399	0.0017545319
Body weight	0.0007258891			-0.0001755696
ARC_4*ARC_5	-0.0000278109	-0.3569536	-0.000624229	0.0001350048
dose*ARC_1	0.0001672821	0.0532450	-0.000230556	-0.0000368764
dose*ARC_2	0.0000108139	0.0719436	0.000284364	-0.0000520017
dose*ARC_3	-0.0000449663	-0.4126377	-0.000471310	0.0001509607
dose*ARC_4	0.0000077882	0.6278025	-0.000563086	0.0003866033
dose*ARC_5	0.0001125392	0.9929705	0.002160929	-0.0007689318
dose*ARC_6	-0.0003628083	-3.7285191	-0.007680771	0.0023529070
dose*ARC_7	-0.0004779950	-1.5916091	-0.008674676	0.0037556119
ARC_4*ARC_5*dose*ARC_1		2.8106290	0.003580681	-0.0012804307
ARC_4*ARC_5*dose*ARC_2		-0.1955823	-0.000812552	0.0003485840
ARC_4*ARC_5*dose*ARC_3		0.0052847	0.000270571	-0.0000792582
ARC_4*ARC_5*dose*ARC_4		-0.1119368	-0.000546831	0.0001324717
ARC_4*ARC_5*dose*ARC_5		-0.0480042	-0.000064921	0.0000592812
ARC_4*ARC_5*dose*ARC_6		0.7301261	0.002599385	-0.0007713856
ARC_4*ARC_5*dose*ARC_7		-0.8686524	-0.003599790	0.0010881481

[a] relative to terminal body weight
[b] control group response for the model under consideration
Note: “ARC x” terms refer to the percent weight concentrations of the Ring *x* material (x = 1 through 7)
Note: “dose” is the applied daily dose in mg/kg/day.

Developmental Toxicity (prenatal) Models:

$$Maternal\,Thymus\,Weight = \alpha + \beta_1 \cdot Control Maternal\ Thymus\,Weight + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i$$

$$Fetal\ Body\ Weight = \alpha + \beta_1 \cdot Control\,Fetal\ Body\,Weight + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

$$Live\,Fetuses/litter = \alpha + \beta_1 \cdot Control\,Live\,Fetuses/Litter + \beta_2 \cdot Number\,of\ implants + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

$$probit(percent\,resorptions) = \alpha + \beta_1 \cdot probit(Control\ percent\,resorptions) + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$$

Table A6-10. Developmental (prenatal) Final Model Coefficients

	Model Coefficients			
	Maternal Thymus Weight (absolute)	Fetal Body Weight	Live fetuses/litter	probit (resorptions/ implants)
Intercept	-0.049351014	0.5733426390	1.721611494	-1.074170983
Control Value[a]	1.137022882	0.8477840952	0.122260599	0.267806492
Number Implants			0.717062149	
ARC_4*ARC_5	0.000073246	-0.0007145352	0.000110287	-0.000011360
dose*ARC_1	-0.000313981	0.0003499551	0.004644289	-0.000203040
dose*ARC_2	-0.000018800	-0.0000435989	-0.000196448	0.000081688
dose*ARC_3	0.000049689	0.0000687564	0.000711925	-0.000286902
dose*ARC_4	-0.000158042	-0.0001873954	-0.003954569	0.001092787
dose*ARC_5	-0.000125555	-0.0014968290	0.018878324	-0.003085551
dose*ARC_6	0.000371121	-0.0007386610	-0.054180024	0.009867294
dose*ARC_7	0.000550250	0.0043864601	-0.052092080	0.006890499
ARC_4*ARC_5*dose*ARC_1		-0.0041012221	-0.009190232	0.002761363
ARC_4*ARC_5*dose*ARC_2		0.0000550464	-0.001682758	0.000333902
ARC_4*ARC_5*dose*ARC_3		-0.0001301287	-0.000133492	0.000066827
ARC_4*ARC_5*dose*ARC_4		0.0002725663	0.000700435	-0.000234635
ARC_4*ARC_5*dose*ARC_5		0.0000183848	-0.000710801	0.000131140
ARC_4*ARC_5*dose*ARC_6		0.0011112570	0.001022511	-0.000406666
ARC_4*ARC_5*dose*ARC_7		-0.0097219896	-0.009734692	0.004480131

[a] control group response for the model under consideration
Note: "ARC x" terms refer to the percent weight concentrations of the Ring *x* material (x = 1 through 7)
Note: "dose" is the applied daily dose in mg/kg/day.

Developmental Toxicity (postnatal) Models:

$$Pup\ Body\ Weight = \alpha + \beta_1 \cdot Control\ Pup\ Body Weight + \beta_2 \cdot Total\ Litter\ Size^{-1} + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j + \sum_{k=1}^{7} \nu_k \cdot dose \cdot PAC^2{}_k$$

$$Total\ Pups / Litter = \alpha + \beta_1 \cdot Control\ Total\ Pups / Litter + \beta_2 \cdot Number\ of\ implants + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j + \sum_{k=1}^{7} \nu_k \cdot dose \cdot PAC^2{}_k$$

$$Live\ Pups / Litter = \alpha + \beta_1 \cdot Control\ Live\ Pups / Litter + \beta_2 \cdot Number\ of\ implants + \eta \cdot PAC_4 \cdot PAC_5 + \sum_{i=1}^{7} \gamma_i \cdot dose \cdot PAC_i + \sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j + \sum_{k=1}^{7} \nu_k \cdot dose \cdot PAC^2{}_k$$

Table A6-11. Developmental (postnatal) Final Model Coefficients

	Model Coefficients		
	Pup body weight (PND[a] 0)	Total pups/litter (PND[a] 0)	Live pups/litter (PND[a] 0)
Intercept	0.946714149	2.239670171	2.944585024
Control Value[b]	0.881434343	-0.050280888	-0.054234799
Number of Implants		0.868019895	0.813933476
1/Total Litter Size*	-3.723731802		
ARC_4*ARC_5	-0.000192728	-0.000967677	-0.000925710
dose*ARC_1	0.000194923	0.000168421	0.000123784
dose*ARC_2	-0.000203264	-0.001062111	-0.000981506
dose*ARC_3	-0.000351392	0.000819004	0.000597959
dose*ARC_4	0.000114875	-0.006449753	-0.005834799
dose*ARC_5	-0.003222217	-0.052122141	-0.055359841
dose*ARC_6	0.008932519	-0.006977416	0.008599625
dose*ARC_7	-0.041387314	-0.084871247	-0.017979122
ARC_4*ARC_5*dose*ARC_1	0.001160907	-0.017287359	-0.015502838
ARC_4*ARC_5*dose*ARC_2	-0.000384566	0.005115255	0.005289680
ARC_4*ARC_5*dose*ARC_3	0.000528932	0.000242632	0.000000847
ARC_4*ARC_5*dose*ARC_4	-0.001175097	0.005033753	0.004034059
ARC_4*ARC_5*dose*ARC_5	-0.001966090	0.021396336	0.017079188
ARC_4*ARC_5*dose*ARC_6	0.011726490	-0.101375752	-0.081365511
ARC_4*ARC_5*dose*ARC_7	0.006604587	0.017648662	0.005400796
Dose*Dose*ARC_1	-0.000062133	-0.000068470	-0.000065306
Dose*Dose*ARC_2	0.000039520	0.000131793	0.000116127
Dose*Dose*ARC_3	-0.000017572	-0.000169348	-0.000150152
Dose*Dose*ARC_4	0.000187121	0.001637961	0.001522181
Dose*Dose*ARC_5	0.003954292	0.011587510	0.016611812
Dose*Dose*ARC_6	-0.062097211	0.256415324	0.177961635
Dose*Dose*ARC_7	0.137879002	0.307715208	0.134096377

[a] PND = postnatal day
[b] control group response for the model under consideration
Note: "ARC x" terms refer to the percent weight concentrations of the Ring *x* material (x = 1 through 7)
Note: "dose" is the applied daily dose in mg/kg/day.

A6.7 Conclusions

1. Preliminary statistical evaluations found compositional data generated using either Method 1 or 2 produced the most accurate models.
2. The eleven final models developed in this project fit the data used to develop them (observed data) very well (r values at least 0.89).
3. The eleven final models were shown to be good predictors of biological effects of untested materials when using input data that are interpolated relative to the existing data.
4. However, as discussed in **Sections A6.4** and **A6.5,** the eleven final models may not be useful or accurate when generating predictions using extrapolated input data points.