THE RELATIONSHIP BETWEEN THE AROMATIC RING CLASS CONTENT AND SELECTED ENDPOINTS OF REPEAT-DOSE AND DEVELOPMENTAL TOXICITY OF HIGH-BOILING PETROLEUM SUBSTANCES

Report of the PAC Analysis Task Group

Sponsored by the Petroleum HPV Testing Group

W. Dalbey	ExxonMobil
J. Fetzer	Consultant
T. Gray	API
J. Murray	Consultant
M. Nicolich	ExxonMobil
R. Roth	Consultant
M. Saperstein	BP
B. Simpson*	Consultant
R. White	API

*Task Group Chair

American Petroleum Institute
1220 L. Street, N.W.
Washington, DC 20005

March 31, 2008

Executive Summary/Conclusions

This report describes the findings of a project undertaken by a multidisciplinary task force to investigate the potential relationship(s) between the polycyclic aromatic compound (PAC[1]) content and the acute, repeat-dose, developmental, reproductive and genetic toxicities of high-boiling petroleum substances, i.e. those with initial boiling points greater than approximately 300 °F. Specific objectives of the project were to:

1. Identify, obtain, and evaluate available information that could be used to assess the possible relationship between the PAC content and toxicity of petroleum substances for the Screening Information Data Set (SIDS) mammalian toxicity endpoints required in the HPV Challenge Program.
2. Identify and characterize relationships between PAC content and SIDS mammalian toxicity endpoints.
3. Determine if any identified relationships could be used to predict the toxicity of untested petroleum substances.

The stimulus for this project was a previously published report (Feuston et al, 1994) of the correlations for a number of petroleum substances of total PAC content with repeat-dose and developmental toxicity. This earlier paper was not robust enough because it was based on a limited data set, was qualitative in nature and did not allow the prediction of the toxicity of untested petroleum streams.

The current review and evaluation of previously-unpublished laboratory reports show that predictive models for effects on selected SIDS repeat-dose and developmental toxicity endpoints can be developed using the weight percent of each of the 1- through 7-ring compounds in the test substance (the "PAC profile"). The effects found to be associated with the PAC profile are consistent with those reported for a number of individual PAHs and PAC-containing materials, although the mechanism(s) of toxicity in this regard are unclear.

In the repeat-dose toxicity studies, associations were found and characterized between the PAC profile and effects on thymus weight, liver weight, hemoglobin concentration and platelet count. In the developmental toxicity studies, associations were found and characterized for effects on fetal weight, number of live fetuses/litter and percent resorptions in the prenatal studies (studies in which pups were delivered by caesarean section) and pup weight, total litter size and number of live pups/litter in the postnatal studies (studies in which pregnant females delivered their young).

For each of the endpoints of mammalian toxicity for which an association with PAC content was observed, mathematical models were developed that could be used to make toxicity predictions on the basis of the PAC profile. Predictions of the toxicity of substances whose PAC profiles and applied dose levels were within the bounds of the PAC profiles and dose levels of the substances that had been used to develop the models (i.e. interpolated predictions) worked very well. As in many modeling studies of this type, predictions of the toxicity of substances whose PAC profiles and applied dose levels were outside the bounds of the PAC profiles and applied dose levels of the substances that had been used to develop the models (i.e. extrapolated predictions) were less certain.

It should be noted, the models were developed based on observed statistical relationships. No attempt was made to identify causal relationships. To do this would have required a detailed understanding of the mechanisms of PAC toxicity, an exercise beyond the scope of the current evaluation.

[1] Polycyclic Aromatic Hydrocarbons (PAH) refers to compounds of two or more fused-aromatic rings consisting of carbon and hydrogen only. Polycyclic Aromatic Compounds (PAC) is a more inclusive term than PAH since in addition to the PAHs it also includes molecules in which one or more atoms of nitrogen, oxygen or sulfur (a heteroatom) replaces one of the carbon atoms in a ring system. See **Appendix 1** for additional comments on nomenclature.

TABLE OF CONTENTS

List of Appendices

List of Tables

List of Figures

1. Introduction

This report describes the findings of a project undertaken by a multidisciplinary task force to investigate the potential relationship(s) between the polycyclic aromatic compound (PAC[2]) content and the acute, repeat-dose, developmental, reproductive and genetic toxicities of high-boiling petroleum substances i.e. those with initial boiling points greater than approximately 300 °F.

The project was undertaken to evaluate further the observations made by Feuston et al. (1994) who examined the correlation between the weight percentage of various chemical classes of compounds in thirteen refinery streams and the magnitude of various effects produced in rats treated dermally with these substances in repeat-dose and developmental toxicity studies. In general, Feuston et al. found the toxicity of the thirteen refinery streams was correlated with concentrations of polycyclic aromatic hydrocarbons (PAC) composed of 3 to 7 rings

In the current project, four potential sources of information were reviewed: the publication by Feuston et al (1994), other published literature on the toxicity of individual PAH and PAC containing materials, studies sponsored by the American Petroleum Institute (API) and unpublished company laboratory reports. These unpublished laboratory reports consisted of:

- reports of twenty-six repeat-dose toxicity studies,
- reports of sixty-seven developmental toxicity studies, two reproductive toxicity studies and an exploratory dose range-finding study in non-pregnant female rats
- one hundred and fifty-three reports of accompanying compositional data.

Only the unpublished company laboratory reports had a sufficient number of studies and provided sufficient detailed compositional data of the PAC content of the test samples to be of use in this evaluation. Thus, The current report describes how the information in unpublished company reports was used for the evaluation of the relationship between PAC content and selected endpoints of repeat-dose and developmental toxicity. It also describes how the observed relationship might be used to predict some aspects of the toxicity of untested petroleum substances with initial boiling points above approximately 300 °F.

The relationship between acute toxicity and PAC was not investigated since the reported oral LD_{50} values for high-boiling petroleum substances are high, i.e., generally greater than the maximum doses tested, typically 5 g/kg and 2 g/kg for oral and dermal exposures, respectively (API 2001, 2002, 2003a, b, c & d, 2004). These high acute toxicity values lead to the conclusion that it was not worthwhile to investigate possible relationships between acute toxicity and PAC content.

Similarly, no attempt was made to include fertility in this phase of the assessment due to lack of data since only two reports of non-guideline reproductive studies were provided,.
The evaluation of the relationship between PAC and genotoxicity will be reported separately.

The essential information on the evaluation and associated findings are presented in an abbreviated form in the body of the report. The appendices provide more details on the key elements of the process.

[2] Polycyclic Aromatic Hydrocarbons (PAH) refers to compounds of two or more fused-aromatic rings consisting of carbon and hydrogen only. Polycyclic Aromatic Compounds (PAC) is a more inclusive term than PAH since in addition to the PAHs it also includes molecules in which one or more atoms of nitrogen, oxygen or sulfur (a heteroatom) replaces one of the carbon atoms in a ring system. See **Appendix 1** for additional comments on nomenclature.

2. Identification and Evaluation of Available Information (Creation of Modeling Data Set)

Information on the toxicity of PAC containing petroleum materials was available from four sources

- a previously published report by Feuston et al, (1994),
- other published literature on the toxicity of individual PAH and PAC containing petroleum materials,
- studies sponsored by the American Petroleum Institute (API), and
- unpublished laboratory reports from two companies.

A brief description and comments on the information from each source can be found in **Appendix 2.**

Of the four sources of information, only unpublished company toxicity reports had sufficiently detailed PAC compositional data and a sufficient number of studies to be of use in this evaluation. The materials that had been tested in the submitted company toxicity studies covered a range of PAC-containing petroleum substances that included gas oils, lubricating oil base stocks, aromatic extracts, heavy fuel oils and crude oil. Some reports on gasoline streams and kerosene were also submitted, but since these materials had initial boiling points below 300 °F, they were excluded from the evaluation (see below).

The unpublished company toxicity reports described:

- nineteen 28-day and twenty seven 90-day repeat-dose toxicity studies,
- sixty-seven developmental toxicity studies and two reproductive toxicity studies) and an exploratory dose range-finding study in non-pregnant female rats
- one hundred fifty-three analytical reports of compositional data on the test samples used in the toxicity studies .

A complete listing of the unpublished company laboratory reports can be found in **Appendix 3**

All unpublished company laboratory reports (toxicity and analytical) were judged to be either “reliable without restrictions” or “reliable with restrictions, i.e. reliability scores of 1 or 2 (Klimsch, et al. 1997). Since all the studies were judged to be reliable (Klimisch 1 or 2), none were excluded from use in this project for reasons of reliability or data quality. Data from both 90- and 28-day repeat-dose studies was used to assess the relationship between PAC content and toxicity with the difference in duration of dosing being considered in the statistical analysis.

All experimental observations/measurements and compositional data that were considered likely to be useful in subsequent evaluations were captured from the reports. All biological data captured from the unpublished company reports, were used in the subsequent statistical modeling subject with the following exceptions:

- Data from materials outside the “domain” of the materials of interest, i.e. initial boiling point below 300°F
- Data from toxicology studies that were not accompanied by “sufficient” compositional data (see **section 3** below),
- Data from studies conducted in species other than the rat,
- Data from studies conducted by routes other than the dermal route,
- Data from animals not surviving to study termination in repeat-dose studies,
- Data from groups in repeat-dose studies that had high mortality, >50%,
- Data from developmental studies in which daily dosing was for only a portion of the gestation period (i.e. less than gestation days 0-19), and
- Data from groups in development studies where there were three or fewer dams with viable fetuses (prenatal endpoints) or litters (postnatal endpoints), both of which were considered inadequate.

Additional details of the identification of studies/data for use in the analysis can be found in **Appendix 4**.

As a result of the study/data selection process, the number of studies that were used in the evaluation of the relationship between PAC content and toxicity are shown in **Table 1**.

Table 1. Number of Repeat-Dose and Developmental Toxicity Studies Used for Evaluation and Their HPV Categories

HPV Category	Repeat-dose toxicity studies				Developmental toxicity studies			
	28-day studies		90-day studies		Prenatal studies		Postnatal studies	
	Avail*.	Used	Avail.	Used	Avail.	Used	Avail.	Used
Crude Oil	0	0	2	2	4	2	4	2
Gas Oils	3	1	4	4	13	7	9	9
Heavy Fuel Oils	6	0	10	8	19	10	15	15
Lubricating Oils	0	0	2	1	1	0	1	1
Aromatic Extracts	0	0	5	1	2	1	2	0
Other	3	0	2	1	4	1	2	1
TOTAL	12	1	25	17	43	21	33	28
* Avail: Total number of studies made available for evaluation prior to selection process								

3. Preliminary Statistical Characterization(s) of Dose-Response Relationship(s)

From among those biological endpoints for which data had been captured, a number of endpoints were identified for a preliminary mathematical characterization of potential dose-response relationship(s) between PAC content (as measured by any of the 3 compositional methods described in **Table 3**) and endpoint-specific effects (listed in **Table 2**). The process used to identify the endpoints for this preliminary evaluation and statistical modeling is described in detail in **Appendix 4** and consisted of the following 3 steps:

1. identify those endpoints most often statistically significantly affected in the studies based on observed responses not statistical modeling,
2. identify those endpoints that were affected most often at the study's LOELs (i.e. those effects that would be predictive of a significant biological effect), and
3. Consistent with reported effects PACs or PAC containing petroleum products.

Dose group data for the biological endpoints chosen for characterization were matched to the appropriate compositional data to form a data set for analysis (see **Table 5A-1** in **Appendix 5**). Initial characterization efforts were made with linear models and a selection of dependent and independent variables. Modeling methods were similar to those described in detail in **Section 4.1.**

This preliminary assessment served two purposes:

1) to identify a smaller number of biological endpoints that would undergo final modeling, and
2) to evaluate the utility for final modeling of the three analytical data sets.

The results of this preliminary assessment can be seen in **Table 4**.

Table 2. Biological Endpoints Affected and Those Identified for Statistical Evaluation				
Endpoint	**Affected most often (statistically)**	**Sensitive Endpoint [d]**	**Good correlation in preliminary statistical evaluation**	**Used for final model development**
Repeat-dose toxicity studies				
Liver wt (abs)	√	√		
Liver wt (rel.) [a]	√	√	√	√
Thymus wt (abs)	√	√	√	√
Thymus wt (rel.) [a]	√			
RBC	√	√		
Hb conc.	√	√	√	√
Hematocrit	√	√		
Platelet count	√	√	√	√
Developmental toxicity studies (Maternal endpoints)				
Body wt	√	ND		
Body wt gain	√	ND		
Food consumption	√	ND		
Liver wt (rel.) [a]	√	ND		
Thymus wt (abs) [c]	√	ND		√
Thymus wt (rel.) [a]	√	ND		
Uterine wt (abs)	√	ND		
Developmental toxicity studies (Prenatal)				
Live fetuses/litter	√	√	√	√
Resorptions/litter	√	√	√	
% Resorptions	√	√	√	√
Fetal body wt	√	√	√	√
Delayed ossification	√	√	ND	
Developmental toxicity studies (Postnatal)				
Total pups/litter PND 0 [b]	√	√	√	√
Live pups/litter PND 0 [b]	√	√	√	√
Pup body wt PND 0 [b]	√	√	√	√
Pup body wt PND 4 [b]	√	√	ND	

[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 4.3.3**). 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] endpoint was among those most often statistically significant affected in the studies and affected most often at the study's LOELs (i.e. those effects that would be predictive of a significant biological effect).
ND= not determined
For cells that are blank an explanation of why this endpoint was not selected is given in Appendix 4

It should be noted that the endpoints selected for modeling are similar to/consistent with effects reported for both individual PACs and PAC containing materials (SCF, 2002, ATSDR, 1995; IPCS, 1998; IRIS 2007; RAIS, 2007). The endpoints selected are also supported from previously reviewed

information on PAC-containing petroleum substances (Robust summaries prepared by API to satisfy the requirements of the HPV challenge program). A more complete discussion of the toxicity of PAC containing petroleum materials can be found in **Appendix 2**.

Among the one hundred fifty-three analytical reports received, five different compositional analytical methods had been used. Only those reporting results of three of the five analytical methods were judged useful in developing predictive models. The three methods are briefly described in **Table 3**, and more fully in **Appendix 1**. Concentrations of aromatic compounds of ring classes 1 – 5 and 1 – 7, including S- and N-PAC, generated using one of these three methods were the only empirically-derived data generated on a sufficiently large set of samples to provide a basis for comparison. Therefore, only toxicity studies in which the test sample had been analyzed by at least one of these three methods were used in the statistical modeling.

Table 3. Methods of Chemical Analysis

PAC Analysis Method	Compositional Information Reported
Method 1 Separation of an aromatic fraction from the sample by silica gel followed by quantification of 1 to 5-ring aromatics content in the fraction	%Total aromatics % Mono-aromatics % Di-aromatics % 3-5 Ring PAC % S-PAC % Non-Basic N-PACs (calculated) % Basic N-PACS (calculated) Total and Basic Nitrogen Total Sulphur
Method 2 Extraction of sample by DMSO to produce an extract which is rich in PAH followed by quantification of 1-7-ring components in that extract	Total PAC content % 1-7 ring molecules in the DMSO extract [1] (often referred to as 1-7 ring PAC)
Method 5 Carbazoles	% Non-basic N-PAC

[1] By definition PAC are compounds with 2 or more rings. However, during the conduct of Method 2, the 1-7 ring structures in the PAH-rich extract are quantified. For simplicity throughout this report, results of this analysis are referred to as weight percent 1-7 ring PAC, even though it is understood that 1-ring compounds are not PAC.

As can be seen in Table 4, the preliminary evaluation found that models developed on measured S-PACs and carbazoles (Method 5) did not fit the data as well as the models that were developed using compositional data on 1-5 ring and 1-7-ring compounds (Methods 1 & 2 respectively). It was also found that ring-class compositional data derived from the Method 2 procedure (rings 1 – 7) generally produced models with a better fit than that derived using the Method 1 procedure (rings 1 – 5). See Appendix 6 for a more detailed discussion.

Table 4. Summary of Results for Preliminary Analysis Using 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			
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. PND 0	72	0.85	0.04	77	0.83	0.04	57	0.54	0.05	79	0.69	0.04

[a] relative to terminal 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

4. Final Statistical Characterization(s) of the Dose-Response Relationships

A detailed description of the development of the final mathematical characterizations of the dose-response relationships for the endpoints listed in **Table 5** can be found in **Appendix 6.** For a listing of the samples that were used for the final characterizations showing their PAC profiles and the report numbers for repeat-dose and developmental toxicity studies refer to **Appendix 7**.

Selection of Endpoints for Final Modeling

After completing the preliminary quantitative assessment of the dose-response relationship(s) (See **Preceding Section**), endpoints were selected for final mathematical characterization based on:

- whether the effect on an endpoint would be considered an adverse effect or predictive of an adverse effect,
- whether similar endpoints had also been characterized, thus making the analysis redundant, e.g. among hematocrit, hemoglobin, and erythrocyte count, only hemoglobin was identified for final modeling, and
- the degree of correlation of the preliminary mathematical dose-response characterization.

See **Table 5** for a listing and **Appendix 4** for a detailed discussion of how effects were identified for final mathematical characterization of the dose-response.

Endpoints of maternal toxicity observed in the developmental toxicity studies were not selected for final mathematical analysis with the exception of maternal absolute thymus weight. Maternal absolute thymus weights were selected in order to compare them to the results of the mathematical analysis of absolute thymus weights in the repeat-dose studies. Other maternal toxicity endpoints were not selected for mathematical analysis because the goal of the project was to determine whether endpoints of developmental toxicity could be predicted based on PAC profile. For purposes of this project, it does not matter whether maternal toxicity played a role in producing developmental toxicity. The model would have value if PAC profile accurately predicts developmental toxicity regardless of the mechanism of action (i.e., whether it is a direct effect or an indirect effect of maternal toxicity).

Developmental toxicity was strongly associated with maternal toxicity (e.g., decreased maternal body weight, weight gain and/or food consumption) and skin irritation in both the prenatal and postnatal studies. For example, among the 21 prenatal developmental toxicity studies, developmental toxicity was never observed in the absence of maternal toxicity. In addition, maternal skin irritation was observed in the vast majority of the developmental toxicity studies, although in 30% and 18% of the prenatal and postnatal studies, respectively, developmental toxicity was observed in the absence of maternal skin irritation. It is quite possible that maternal toxicity and skin irritation play a role in producing developmental toxicity. Skin irritation might cause developmental toxicity by causing pain and stress to the mother. Further, skin irritation could also alter the dermal absorption of the test materials, increasing or decreasing their potential to cause developmental toxicity (see **Appendix 4** for details on maternal toxicity and skin irritation).

Table 5. Endpoints Selected for Final Mathematical Characterization

Study Type	Endpoint
Repeat-dose toxicity studies	Thymus weight (absolute)
	Platelet count
	Hemoglobin concentration
	Liver weight (relative)[a]
Developmental toxicity studies (Prenatal)	Maternal Thymus weight (absolute)[c]
	Fetal body weight
	Live fetuses/litter
	Percent Resorptions
Developmental toxicity studies (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 weights were selected in order to compare the results of the mathematical analysis of thymus weights in the repeat-dose studies. (**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.

4.1 Modeling Methods

Models were developed using linear 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. dose, duration of dosing, and sex), biological variables (e.g. control group response, and litter size) and the test substance variables (e.g. PAC ring-class weight percentages) as the independent, or predicting, variables. The analyses were based on ordinary least squares (OLS) methods (Draper and Smith, 1998).

The predictive ability of the models was tested by three techniques that are discussed in detail in **Section 4.3**.

4.1.1 Choice of Dependent Variables

The dependent variables were the responses of a dosed group (dose > 0) for the eleven endpoints selected as described in **Section 3**. Control group responses were independent variables in the models (see **Section 4.1.2**).

A dose-group response was the mean of the responses of a dose group in a specific study. For the repeat-dose studies the dose-group response was the mean response of all the animals in a dose group. For the developmental toxicity studies, the dose-group response was the mean of the means of all the litters in a dose group. Thus, if a study had 3 dosed groups and data was collected from each dose group, there would be 3 data points for an endpoint. The number of dependent variable data points used to develop the model for a specific endpoint is shown in **Table 6**.

4.1.2 Choice of Independent Variables

Analytical variables
As noted in **Section 3** and **Appendix 1**, the PAC content of the test samples used in the various company toxicity studies had been determined using a variety of analytical techniques. Preliminary models were built using four compositional data sets (**Section 3**). 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 3** for details of the results of the modeling and the basis for the choice of the Method 2 data set.

Toxicity study design and biological 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 mg/kg/day), duration of dosing, control group response, and sex. The control group response values were based on the mean responses of the control groups in the Task Group's data set. For the developmental toxicity studies independent variables included control group response, dose level (normalized to mg/kg/day), litter size, number of implantation sites, number of animals or pregnant dams or litters per dose group. Not all variables were eligible or appropriate for all models. However, in the case of repeat-dose studies, terms for dose level, duration of dosing and sex were always included in the model building process. All responses were means calculated in a similar manner to that described in **Section 4.1.1**.

4.1.3 Model Forms

The basic model form was a linear regression model with a possible transformation of the dependent variable. The dose group response was the dependent variable, the control group response an independent variable (covariate), and a selection of additional independent variables as described in **Section 4.1.2**.

The models for the eleven endpoints were developed independently. In the model building process for each endpoint, several mathematical forms of the model were considered based on transformations of the dependent and independent variables. For each endpoint, the selection of the optimum model was based on a set of criteria and considerations. Among the direct statistical criteria were the overall model multiple correlation coefficient (r), the standard error (se, calculated as the square root of the error mean square, or EMS), the correlation among the independent variables, evaluation of the normality of the distribution of residuals by Wilk's test, and the set of standard statistics that indicate outliers and influence points. Other criteria used for model selection included visual inspection of the residuals against the independent variables, and plots of the observed vs. model predicted points for each endpoint. An overall goal in fitting the model to the data was to adhere to the principle of "parsimony", i.e. the simplest model that is adequate for the problem to be solved is used.

In developing the final models, 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. 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.

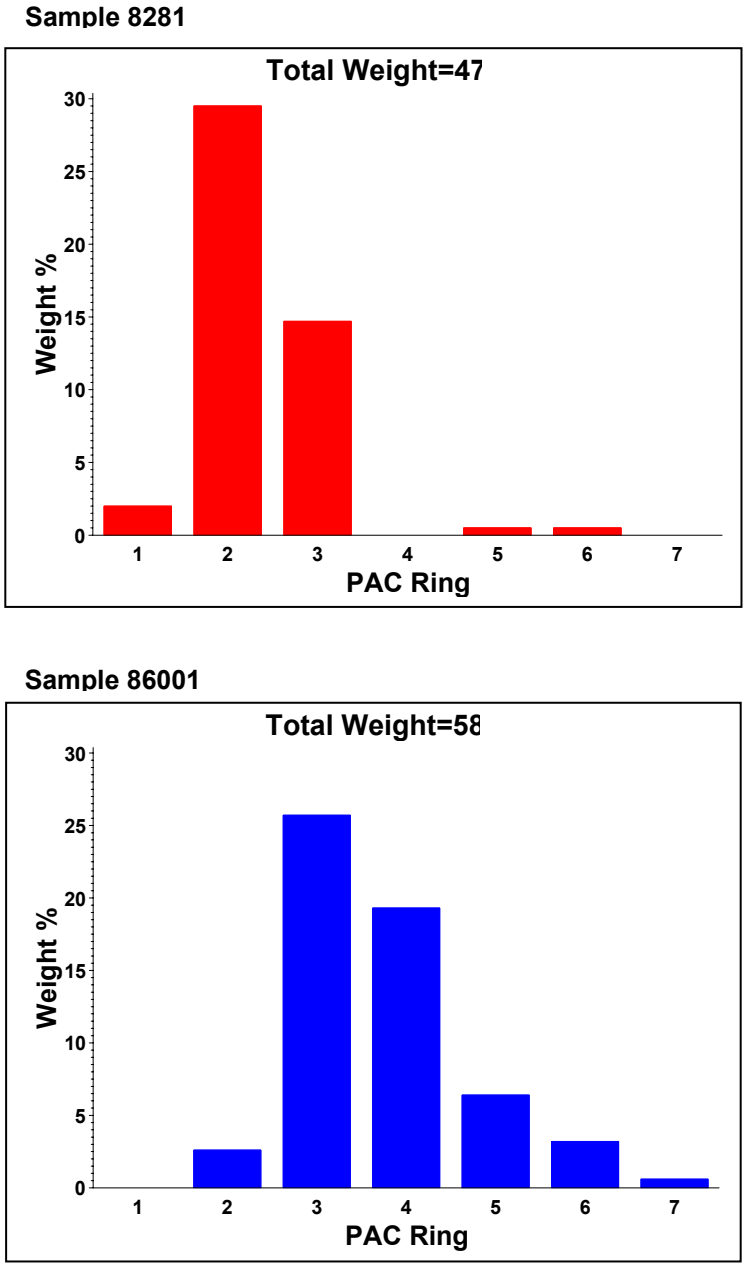
The preliminary models (results shown in **Table 4**) contained a term describing a compound's HPV group (Crude Oil, Aromatic Extracts, Gas Oils, Heavy Fuel Oils, Lubricating Base Oils or Waxes). It was recognized that the classification requirement for a specific compound would limit the overall applicability of the models, and indeed there might be cases where the specific category of a compound could be questionable. To ameliorate this potential problem, the final models (results shown in **Table 6**) did not include a term describing a substance's HPV group. Minor changes in model fit as a consequence of this change were considered acceptable consequences.

Table 6 shows the values of correlation (r) and residual standard error (se) for the final models, and provides a basis for the reader to compare model adequacy and fit. These two measures were selected from among the criteria used for model evaluation because, among their characteristics, the r value is a measure of the closeness of the observed and model predicted values and the se is related to the width of the confidence interval of the predicted value. By themselves the r and se values are not adequate to make final decisions. For example, if a few observations are far from the bulk of the data the correlation can be unrepresentatively large, or a few observations far from the prediction line can increase the se to make the model seem to be inadequate. Therefore, for the final models, the plots of the observed vs. the predicted values are presented and provide the most useful form for assessing model adequacy (**Figure 4**).

4.1.4 Individual PAC Terms

The final models were developed using the weight percent of each of the 1- through 7-ring compounds in the test substance (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 correlate with dose-response curve. For example, consider the weight percentage of the ring components in the two samples depicted in **Figure 1**. Both samples have similar total weight percent of 1-7-ring compounds but their PAC profiles differ.

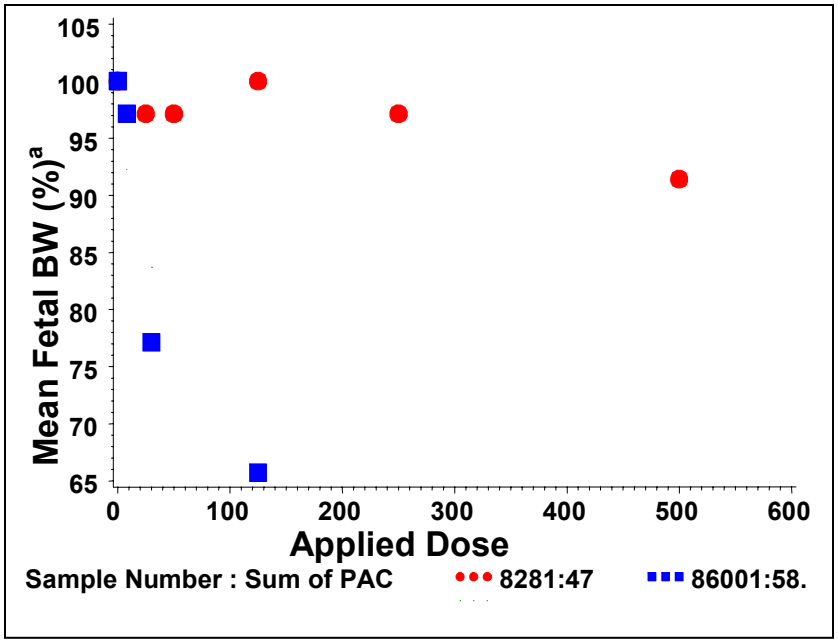
Figure 1. Weight Percent of 1- through 7-Ring Compounds of Two Petroleum Substances with Total PAC Extract Weights of 47 and 58 Percent



The biological responses to applied dose for substances with similar total weight percentage but with different PAC profiles can be very different, as shown in **Figure 2**. The observed mean fetal body weight ratio to the control group for each of the two substances from **Figure 1** are plotted in **Figure 2.** Results from samples 8281 (**Figure 1, top**) and 86001 (**Figure 1, bottom**), 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. The difference in the slope of the dose response curves indicates that total PAC weight alone is a poor predictor of response; rather it is apparent that biological activity of the sample with PAC constituents predominantly 3-6-membered rings is substantially greater than that of the sample in which the PAC constituents are predominantly 2-3-membered ring species. It should be noted that since the mechanism of action of PAC is not understood, the data should be viewed as indicating only that there is an observed relationship, and should not be used to assess whether any of the specific ring values are responsible for the response,.

Figure 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

4.1.5 Factor Analysis

During model development, one of the goals was to minimize the number of independent variables and reduce the degree of correlation among the independent variables (the problem of multi-colinearity). Therefore, a factor analysis was done on the individual aromatic 1 to 7-ring weight percentage data. A three- factor solution was selected that accounted for 80% of the variance for the Method 2-derived aromatic 1 to 7-ring weight percentage data. Subsequent regression analysis models with the factor scores did not fit the data as well as the models using the individual ring weight percentages; this was seen in all models tested. Based on these results, the individual ring weight percentages and selected interactions among the weight percentages were used for model development.

4.2 Final Model Results

The correlation coefficient and residual standard error (r and se) values in **Table 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 different from the preliminary models used to generate the results shown in **Table 4**, comparisons cannot be made of the r and se values from these two tables.

Table 6. Final Modeling Results with the Method 2 PAC Weight % Results

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 4.3.3**). 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

r multiple correlation coefficient

se standard error (calculated as the square root of the error mean square, or EMS)

The magnitudes of the correlation coefficients presented in **Table 6** are large for this type of data, the minimum correlation being 0.89. Possible explanations for the large coefficient correlations are:

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.

To ensure that the model results and corresponding correlation coefficients were not spurious, based on bias, confounding, or affected by model specifications, the final models were rigorously tested as described later in **Section 4.3**.

4.2.1 Model Equations

The final models for the eleven endpoints are linear in the coefficients and of a similar form. An example of the algebraic form of a model based on the number of live fetus/litter 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 eleven final models are described in **Table 7**. The table lists each dependent variable and its transformation (if any), the selection of biologically based independent variables and the PAC terms. 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 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:

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

The models are complex, with the number of coefficients ranging from 10 to 25. **Section A6.5.4** in **Appendix 6** provides the coefficients and complete forms for the eleven final models.

Table 7. General Description 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	Log 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	CG[a] 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 4.3.3**). 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 $\sum_{j=1}^{7} \xi_j \cdot dose \cdot PAC_4 \cdot PAC_5 \cdot PAC_j$

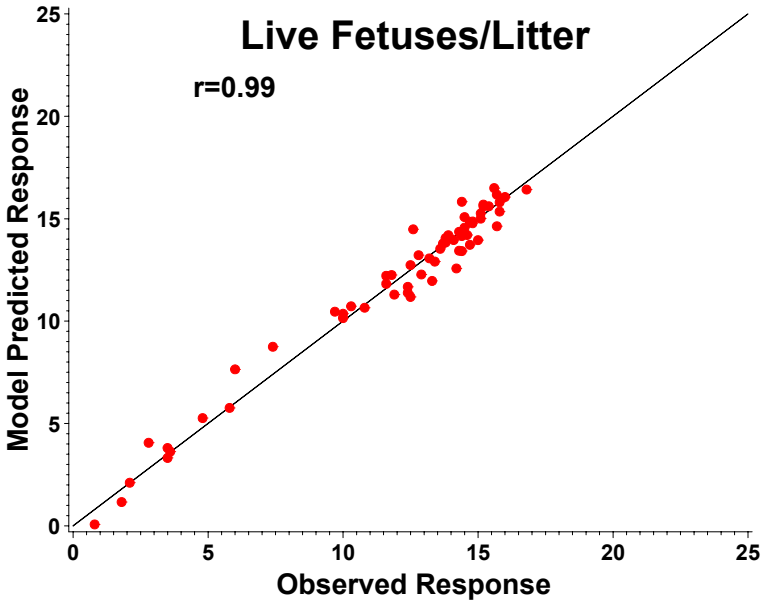
2 interaction term of the form $\sum_{k=1}^{7} \nu_k \cdot dose \cdot PAC_k^2$

4.2.2 Model Fit

The accuracy of the fit of a model can best be seen in a plot of the observed data points vs. the predicted data points. The optimum model would have all points along the straight line representing equal values of the observed and predicted data.

As an example, the plot for the model of live fetuses/litter is shown in **Figure 3**. The correlation coefficient (r) for this model is 0.98, which is an indication of a very good model fit.

Figure 3. Plot of Observed and Model Predicted Live Fetus/Litter Count



Similar plots for all eleven models are shown in **Figure 4**, with the live fetus/litter plot repeated for completeness.

Note that the r values (correlation coefficients) for all the models are equal to or greater than 0.89.

Note also that the r values in the figures are slightly different from the r values in **Table 4**. This difference is due to the fact that the r values in **Table 4** were derived from preliminary models, whereas those in **Table 6** and **Figure 4** were derived from final models.

Figure 4. Plots of Observed and Predicted values for Eleven Final Model Forms

Repeat-dose toxicity studies

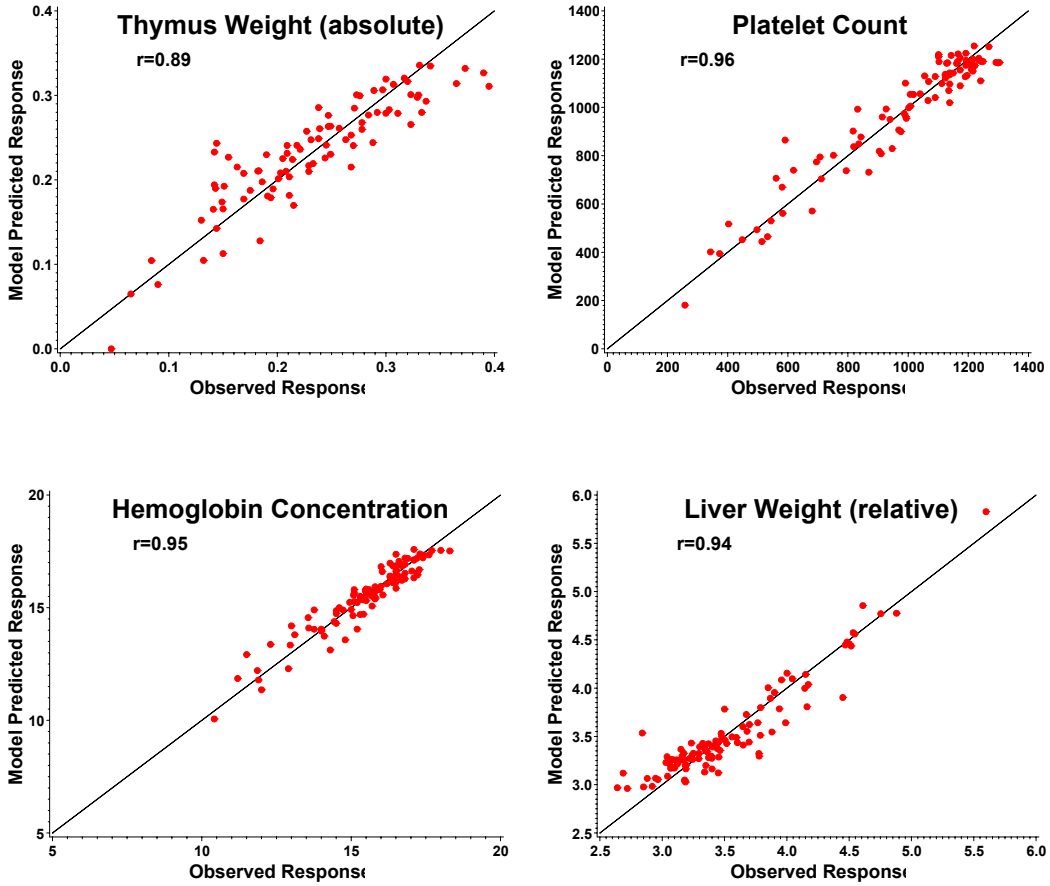
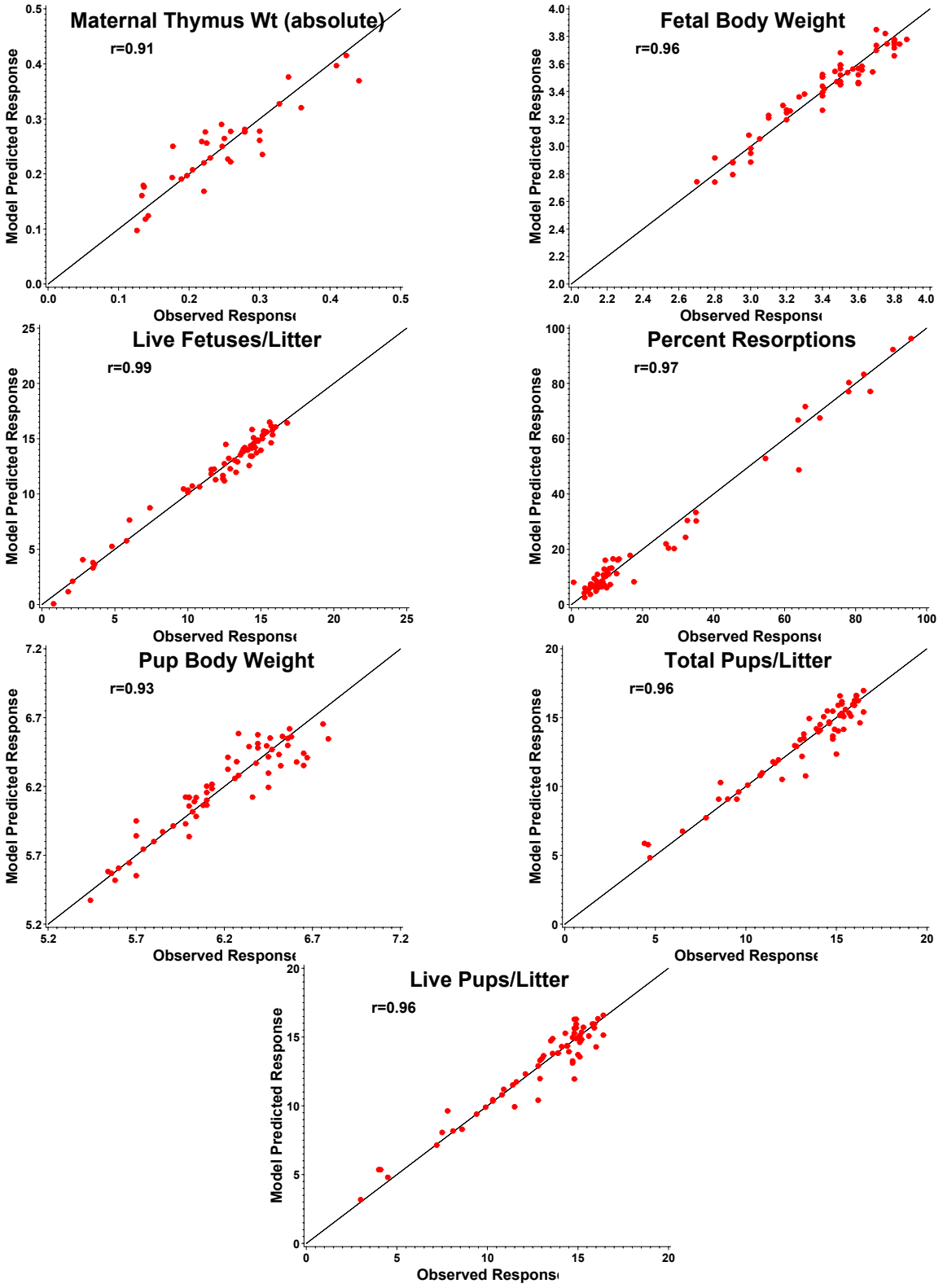


Figure 4 (cont.).

Developmental toxicity studies



4.3 Model Testing

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

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

The full details of these tests are given in **Appendix 6**. In general, the tests showed the eleven models to be good predictors for data points that are interpolated from the existing data (were within the bounds of the PAC profiles and dose levels of the substances that had been used to develop the models). However, the tests also showed the models to be of questionable use for data points that are extrapolated (PAC profiles and applied dose levels were outside the bounds of the PAC profiles and applied dose levels of the substances that had been used to develop the models). A summary of the results of the three model tests that were performed is given below.

4.3.1 Model Testing with Hold-Out Sample Data

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 a separate set of the data that had not been used to develop the model. This process is called holdout sample validation or data-splitting validation. The data used to develop the model is called the training data; the remaining data are the test or holdout data.

In this project, the data set that was used to develop the repeat-dose absolute thymus weight model was split; 30% of the data points were randomly selected and used as holdout data and the model was developed from the remaining 70%. The model was then applied to the 30% holdout sample to see how well the model predicted the known results. This process was repeated 100 times, each time the data in the holdout data set varied.

Plots of the predicted vs. observed results in **Section A6.5.1**, **Appendix 6** show that the repeat-dose absolute thymus weight model provided accurate and robust predictions to the holdout samples when they are interpolated points, and in a few instances were not accurate for values were unreliable for points in the holdout sample that are extrapolated points. This problem is often found with these types of models and is called the problem of extrapolation; further discussion of interpolation and extrapolation appears in the "Limitations" section of this report (**Section 5.4**) and in **Appendix 6**, **Section A6.4**.

4.3.2 Model Testing Using Nonsense Data

A method for testing model usefulness is to determine model performance when the independent variables (PAC compositional data) were really *not* associated with the observed outcome. Conceptually, if a model fits well even though the independent data were not associated with the response, this is an indication that the model results were based on some structure not related to the postulated relationship.

This "nonsense testing" was applied to the hemoglobin concentration model, the original with an r value of 0.95. The response data (hemoglobin concentration) and the sets of independent variables were randomly shuffled and a new model was fit. The process was repeated 100 times. The resulting models had a mean r = 0.61, with a minimum and maximum of 0.44 and 0.88, respectively. The relatively low r values from the nonsense data are a clear indication that the model behavior is based on information in the data, and not from chance, or are related to the independent variables used in the model (see **Appendix 6, Section A6.5.2**).

4.3.3 Model testing Using Alternate Data Sources

The data available allowed for using data on one endpoint as a data set for predictions derived from models developed for a different endpoint. Examples include:

- repeat-dose absolute thymus weight and prenatal absolute thymus weight,
- prenatal fetal body weight and postnatal pup body weight, and
- prenatal live fetuses per litter and postnatal total pups per litter.

Consider a model developed from the repeat-dose absolute thymus weight data (the repeat-dose absolute thymus weights and the associated PAC 2 data). When the model is applied to these data the correlation coefficient (r value) between the observed and predicted data was 0.89 based on 89 observations. If the repeat-dose absolute thymus weight model (the same model form and the same coefficients) is used to predict the maternal thymus weight data using the PAC 2 compositional data from the developmental study samples the correlation is 0.77 based on 34 observations. This second step is a model validation with new data. It is a stronger test than just using new data because the new data are from a different type of study (developmental toxicity as opposed to repeat-dose).

The reverse fitting (prenatal model used to predict the repeat-dose data) was not as good: the correlation of the observed repeat-dose absolute thymus data with the predicted values using the prenatal model was 0.43. However, among the predictions that were based on interpolations the correlation was 0.77 (n=48); among the predictions that were based on extrapolations the correlation was 0.43 (n=41). A fuller discussion with other examples can be found in **Appendix 6**.

5. Prediction of Toxicity of Untested Substances

This section describes how the models developed might be used for predictive purposes, i.e. if any PAC-toxicity relationships could be used to predict the toxicity of untested petroleum substances. Limitations on the utility of the predictive models in this regard are also discussed. The different applications of the models included in this section are for demonstration only and are meant to show the wide applicability of the models. A more detailed description is given in **Appendix 8**.

5.1 Prediction of Dose-Response Curves

Eleven mathematical models were developed that describe the PAC-toxicity dose-response for a number of repeat-dose and developmental toxicity endpoints in the rat after dermal administration of certain classes of petroleum hydrocarbons. The models are summarized in **Table 7**.

Predicted dose response curves may be generated with any of the eleven models by:

1. selecting a sample and determine its PAC content data values (PAC profile)
2. selecting the endpoint of interest and its associated model
3. determining if the sample is interpolated or extrapolated relative to the data set used to develop the model that was selected (see **Section 5.4.1** for the required steps)
4. using the equation to estimate the predicted endpoint for a series of doses (the coefficients for the model are in **section A6.6** in **Appendix 6**, and the control group values and other needed values are provided in **Appendix 10**)

As an example consider the use of a model to generate dose-response curves for the live fetus/litter counts for two different samples. A complete example is provided in **section A8.2** in **Appendix 8.** The predictive model for live fetus/litter has the following form:

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

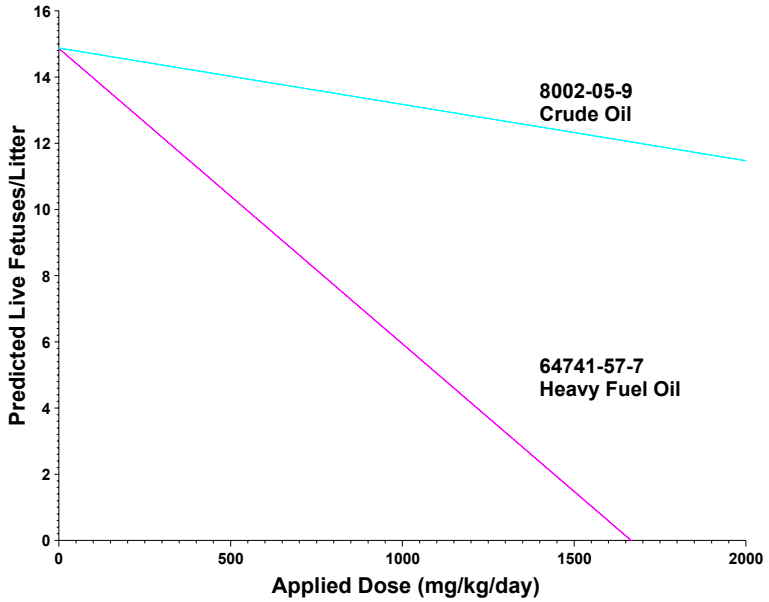
The PAC profiles for the two substances are:

	PAC rings (wt. %)						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Samples							
CAS 64741-57-7 (Heavy Fuel oil, sample 85244)	0.0	0.06	2.5	1.9	1.2	0.5	0.0
CAS 8002-05-9 (Crude oil, sample 89645)	0.0	6.4	1.6	0.4	0.0	0.0	0.0

For the purposes of this example, it will be assumed that the predictions of live fetuses/litter for both samples are interpolations. However, in the "real world", to determine if the predictions would be interpolations or extrapolations, the samples' PAC profile and dose would be compared to the PAC profiles and doses of the substances used to develop the live fetuses/litter model (see **Section 5.4.1**). It should be noted that for any sample, the predictions for various endpoints may differ, with some being interpolations while others are extrapolations.

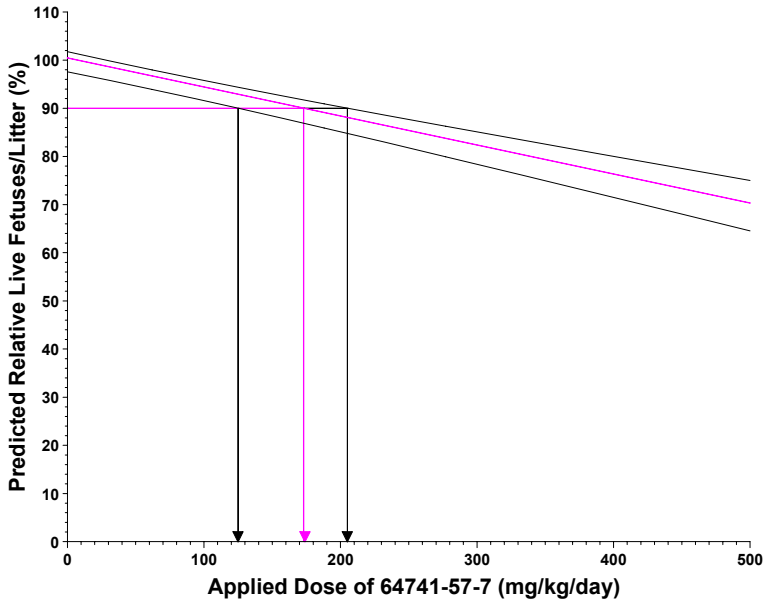
Based on control group data from the 21 prenatal studies used in final model development, it can be assumed that the mean numbers of implantations and live fetuses in the control groups are 15.8 and 14.9, respectively. Assuming a dose of 500 mg/kg/day, and substituting the control values and the values of the coefficients from the equation for the live fetuses/litter, the mean number of live fetuses/litter at 500 mg/kg/day is predicted to be 10.4 for CAS 64741-57-7 and 14.0 for CAS No. 8002-05-9. Repeating this calculation for different dose values would produce the two dose-response curves seen in **Figure 5**.

Figure 5. Predicted Dose-response Curves for Mean Number of Live Fetuses for Two Samples with Different PAC Profiles



To determine the live fetuses/litter relative to control, each of the values from **Figure 5** would be divided by the corresponding predicted control value, and then multiplied by 100. The result for CAS No. 64741-57-7 is shown in **Figure 6.**

Figure 6. Predicted Live Fetuses per Litter with 95% CI for CAS 64741-57-7



5.2 Use of Models to Predict a Pre-Defined Change (PACBMD)

The predicted dose-response curves that can be generated permit the prediction of either:
1) the effect at a given dose, or
2) the dose that causes a given effect.

Prediction of an effect at a given dose has been shown in the preceding section. The estimation of a dose associated with a specified effect is similar to a benchmark dose (BMD) (Crump, 1984). Using a specific PAC model and a defined change from the control value control, a dose can be calculated that would be associated with a defined change of that magnitude. To distinguish this value from the BMD, we will call the value determined from the PAC model the PAC_{BMD}, and let the dose associated with a 10% change from control be noted as PAC_{BMD10}.

The PAC_{BMD} is similar to the BMD, but the PAC_{BMD} relies on only one validated model, whereas the BMD can be developed from several competing models and the result is strongly dependent on the model selected (Gephart, et al, 2001). Because of the way the models have been developed for the two methods the PAC_{BMD} is usually based on an interpolated dose value from the models because of the large number of data points that have been used to develop the PAC model, while the BMD value is often an extrapolated dose value. An additional disadvantage is that the prediction error associated with the BMD is related to how near the observed data are to the critical response. The BMD cannot be used for untested materials, while the PAC_{BMD} can be. Another advantage of the PAC_{BMD} is that it is based on multiple studies while the BMD is based on a single study, usually with 3 to 5 data points.

As an example, the dose associated with a 10% reduction, (PAC_{BMD10}) in the mean number of live fetuses per litter could be predicted for the sample used as an example in the preceding section, CAS 64741-57-7,. To do this, the plot shown in **Figure 5** is simply replotted converting the absolute live fetus count values on the y axis to a percent relative to control. This is done by dividing the model predicted responses at each dose by the expected model predicted response at zero dose (14.9 live fetuses), see **Figure 6**. The dose associated with a response that is 90% of control value (a 10% reduction) is estimated to be 173 mg/kg/day. Thus, 173 mg/kg/day would be identified as the BMD_{PAC10}.

In addition, confidence intervals (CI) can be developed and associated with the PAC_{BMD10}. The confidence intervals are based on inverse regression methods also known as calibration methods (Draper and Smith, 1998). Using the same example (sample CAS 64741-57-7), the PAC_{BMD10} and associated 95% CI is 173 (125,206) mg/kg/day (see **Figure 6**).

5.3 Comparison of Predicted and Actual Effects

An indicator of a poorly fitting regression model is when the majority of the observed data points do not fit within the 95% confidence interval of the model-predicted points. Conversely, if the observed data points do fit within the 95% confidence interval of the model predicted points, although not a guarantee that the model is good, it is highly suggestive the model is a good fit.

This comparison (predicted vs observed) was made by generating a predicted dose-response curve for every endpoint modeled, for every sample that was used to develop the models. These curves are provided in **Appendix 8**. In each of the predicted curves, the 95% confidence limits are shown together with the actual values that had been determined in the studies that were used to develop the models.

When each predicted dose-response curve was compared with actual results of the study from which the information had been derived, it was found that the predictions were accurate in most, but not all cases (see **Table 8**). The typical percent of observed data that was within the 95% confidence limits was, as expected, at least 95%.

A low percentage of the curves were apparently poor or erroneous predictors of the expected dose response. Closer inspection of these revealed that in almost all cases an effect had not been demonstrated on the endpoint in the study in question. For example the total litter size response for sample F-215 (See **Fig 7**) shows a slight increase in total litter size with dose,. It is also inconsistent with the expectation that decreased litter size would be associated with exposure to substances containing PAC, or at least no change in response.

Figure 7 Unexpected Increase In Response Leads to Point Outside 95% CI

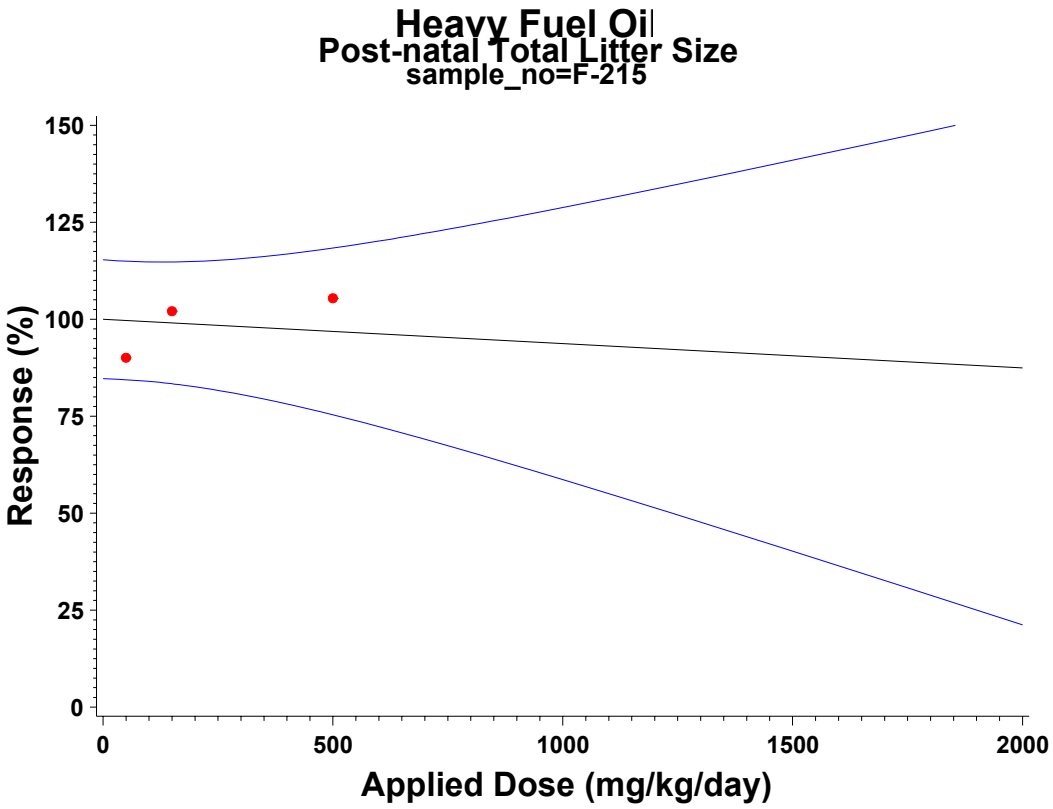


Table 8. Summary of the Proportion of Accurately Predicted Dose-Response Curves

Study type	Dependent Variable	% Correct predicted dose-response curves
Repeat-dose toxicity studies	Thymus weight (absolute)	97.8%
	Platelet count	97.8%
	Hemoglobin concentration	94.2%
	Liver to body weight ratio	96.1%
Developmental toxicity studies (Prenatal)	Maternal thymus weight (absolute)	100.0%
	Fetal body weight	98.4%
	Live fetuses/litter	96.8%
	Percentage resorptions	98.4%
Developmental Toxicity studies (Postnatal)	Pup body weight	100.0%
	Total pups/litter	93.6%
	Live pups/litter	96.8%

5.4. Potential Limitations/Restrictions on Model Use for Predictive Purposes

The models may be used immediately to assist in the selection of petroleum substances for further testing. As more Method 2 compositional information becomes available for the substances in PAC-containing petroleum categories, it will be possible to identify the compositional boundaries for each category. The models can then be used to identify those samples that would require a prediction by extrapolation, and these samples could be selected for testing. When additional compositional and toxicology data become available the strength of the models will be increased and the models can be used with increasing confidence. Limitations on model use are discussed below.

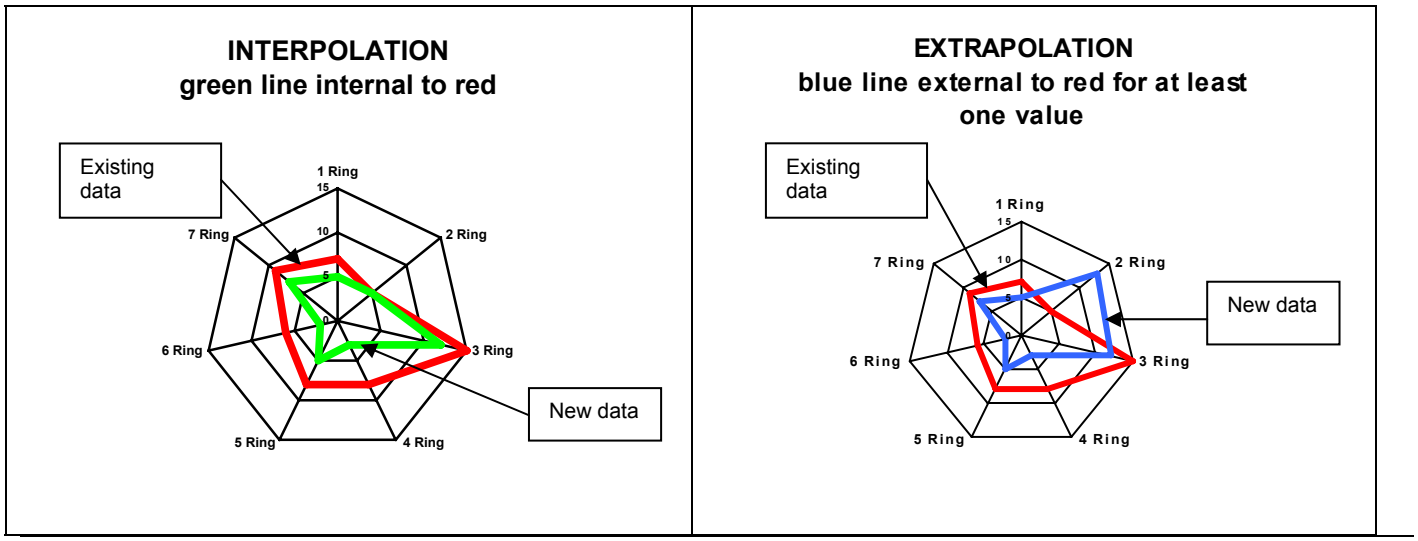
5.4.1. Interpolation and Extrapolation

As noted in **Section 4.3**, and demonstrated in more detail in **Appendix 6**, testing of the models shows they are all good predictors for samples whose PAC profile & applied doses would be interpolated but are not consistently accurate predictors among data points that would be extrapolated.

Specifically for this project, we need to determine if a new sample and its corresponding PAC profile (i.e. a sample that was not used to develop the model) is an extrapolated or interpolated sample relative to the samples that were used to develop the model. To make the determination one needs to know the applied dose and percent weight concentrations for the seven ring classes for the new sample along with the corresponding applied dose and percent weight concentrations for all the samples used to develop an endpoint-specific model (“the existing data set”).

As a simplification, consider the concepts of interpolation and extrapolation between two samples based on their 7 ring PAC weight percent concentrations (PAC profiles) as demonstrated pictorially in **Figure 8**.

Figure 8. Representation of the Difference Between Interpolated and Extrapolated Data



Note: Red line represents existing data points, the green and blue lines represent new data points

The process for determining the interpolation and extrapolation status of two samples consists of the following steps:

a. Determine if the percent weight concentrations for the Aromatic Ring 1 concentration of the new sample is *less than, equal to, or greater than* the corresponding concentration of the existing sample.
b. If the answer is “less than” or “equal to” then make a similar comparison for each of the PAC Ring concentrations 2 through 7.
c. If the answer is “less than” or “equal to” for all 7 PAC Ring concentrations then the new sample is an INTERPOLATED point relative to the existing sample.
d. If any answer for any Ring Concentration is “greater than”, then the new sample is an EXTRAPOLATED point relative to the existing sample.

For a new sample to be *interpolated* relative to an existing data set (the data base used to build the model) it must ‘be between the largest and smallest existing sample in the data base’ with regard to PAC content and dose; so it must

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

If the sample violates any of the three points above it is an *extrapolated* point relative to the data base used to build the model.

It should be noted that the assessment of whether a sample and its associated prediction is an extrapolation or interpolation are endpoint-specific for any given sample, since unique data bases were used to develop each endpoint model. Hence for a new compound the interpolation/extrapolation status is unique to the endpoint. Thus, for each new sample or substance, the extrapolation/ interpolation determination is made eleven times, once for each of the eleven biological endpoints. It is quite possible that the status of predictions (interpolation or extrapolation) for various endpoints for a single sample may differ, with some being interpolations while others are extrapolations.

An Excel® based spreadsheet has been developed that can be used to determine if a sample is an extrapolated or interpolated point, based on its percent weight concentrations for the seven ring classes (PAC profile) and its anticipated applied dose.

5.4.2. Compositional Data Set

The accuracy of the final models' predictions made using analytical data from methods other than Method 2 is not known because the current model(s) were developed using only Method 2 data. It may be possible to develop additional models, similar to the existing models, based on alternative analytical methods, but it would involve additional model development and testing.

5.4.3 Route of Exposure

The models in this project were developed using data from dermal toxicity studies. Without additional data, the predictive capacity and applicability of the models to other routes of exposure is unknown.

6.4.4 Species and Strain

The models in this project were developed using data from toxicity studies involving the Sprague-Dawley rat as the experimental animal. Without additional data, the predictive capacity and applicability of the models to other strains and species are unknown.

5.4.5 Coverage of Data Set

Although the various models were built using experimental data developed on samples from across a range of -petroleum categories, approximately 70% of the samples were from the gas oils and heavy fuel oils categories. Because the compositional component of the models is based only on PAC profile and not on specific category membership, the models are applicable to a wide range of petroleum-derived substances in which PAC may be the toxicity "driver". As further information becomes available from studies conducted on substances from HPV petroleum categories other than gas oils and heavy fuel oils, the models will be extended further, thus providing additional support for their use across all petroleum substances in which PAC may be the toxicity "driver".

5.4.6 Quantification of Degree of Change

The selection of a $BMD_{PAC\ 10}$ was solely for purposes of demonstrating how the models could be used to predict a dose that would be likely to be associated with a pre-defined adverse effect. Further consideration may need to be given to this issue to ensure that appropriate BMD_{PAC} values have been selected when attempting to predict the toxicity of an untested petroleum substance.

6. Discussion, Conclusions and Recommendations

The primary purpose of the present investigation was to determine whether there is a relationship between the PAC content of classes of petroleum substances boiling above approximately 300 °F and their mammalian toxicity. A secondary objective of the current investigation was to determine whether an association, if it existed, could be used to predict the toxicity of untested petroleum substances. The investigation was confined to repeat-dose and developmental toxicity endpoints.

It was found that there are indeed associations between some repeat-dose and developmental toxicity endpoints and the PAC content of selected petroleum substances. It has also been demonstrated that the toxicity of an untested substance can be predicted based on its PAC content.

6.1 Relationship between PAC and Effect

Repeat-dose toxicity

There was an association between a substance's PAC profile and effects on repeat-dose endpoints, including absolute thymus weight, hemoglobin concentration, erythrocyte count, hematocrit, platelet count and increased liver weight. Using linear regression techniques, predictive models were developed for absolute thymus weight, relative liver weight, hemoglobin concentration and platelet count. When the observed and predicted data were compared, the correlations were very good with correlation coefficients (r) of ≥0.87. For untested petroleum substances with PAC profiles similar to those of the samples used to develop a model, the dose that would be associated with a predefined quantitative change in one of the modeled endpoints could also be predicted.

Developmental toxicity

Associations were found between PAC profile and adverse effects on development, including reduced fetal bodyweight, reduced number of live fetuses/litter and increased resorptions/litter in the prenatal studies and reduced pup weight, total litter size and number of live pups/litter in the postnatal studies. Predictive models using linear regression techniques were developed for each of these biological endpoints. When the observed and predicted data were compared the correlations were very good with correlation coefficients (r) of ≥0.91. For untested petroleum substances with PAC profiles similar to those of the samples used to develop a model, the dose that would be associated with a predefined quantitative change in one of the modeled endpoints could be predicted.

Although the relationship between PAC profile and endpoints of maternal toxicity was not analyzed mathematically, it is important to recognize that developmental toxicity was strongly associated with maternal toxicity in the developmental toxicity studies. This raises the possibility that the observed developmental toxicity may have been caused indirectly by maternal toxicity. For purposes of this project, however, it does not matter whether maternal toxicity and/or skin irritation causes the developmental toxicity of the test materials. The goal of the project is to determine whether developmental toxicity can be predicted on the basis of PAC profile. The model has value if PAC profile accurately predicts developmental toxicity regardless of the mechanism of action (i.e., whether it is a direct effect or an indirect effect of maternal toxicity). However, except for one questionable result in one study, none of the test materials are selective developmental toxicants (i.e., substances which cause developmental toxicity in the absence of maternal toxicity).

Biological plausibility/consistency
Identification of the repeat-dose and developmental toxicity endpoints that were modeled was carried out with considerable care. Confirmation that the endpoints identified for modeling were biologically plausible is provided in several reviews of the toxicity of PAH (SCF, 2002, ATSDR, 1995; IPCS, 1998; IRIS 2007; RAIS, 2007). In these reviews, the spectrum of effects attributable to PAH was similar to the endpoints that were selected for modeling.). Further support that the selected endpoints are reasonable is found in the robust summaries and test plans for petroleum streams prepared by API in their activities to fulfill the requirements of the HPV challenge program, where the spectrum of effects of PAC-containing streams is similar to the endpoints selected for modeling.

6.2 Model strengths

The statistical techniques used to develop the predictive models presented in this report are much more robust than the techniques used in the only previously published evaluation of the relationship between PAC content and toxicity of petroleum substances (Fueston et al, 1994). The current statistical techniques also made use of a larger data set, whereas the previous evaluation relied upon a more limited data set. The large number of data points used to develop the models is a particular strength of the current evaluation. The plots of the observed vs. predicted points shown in **Figure 4** demonstrate that the models are accurate descriptors of the data and are accurate predictors for interpolated substances (**Section 4.3**). The models are relatively simple linear models, all with a similar mathematical form across the endpoints, which provides a measure of the consistency of the models.

Analytical data required
To predict the toxicity of an untested substance using the models, the only compositional input that is required is the PAC profile of the substance as determined by a Method 2 compositional analysis. The essential features of the Method 2 analysis are extraction of the sample with DMSO to provide an unalkylated and low- to moderately-alkylated aromatics PAC-rich fraction, and the subsequent separation by gas chromatography and determination by flame ionization detection or mass spectrometry of the concentration of ring-classes 1 through 7. The models use the concentration of each ring-class rather than the total weight % of PAC or any subset of ring classes, e.g., 4-6 or 3-7-ring PACs. This approach was found to be essential as many substances with similar total weight % of PAC may be predicted to have significantly different toxicities.

Model limitations and data needs
A number of constraints were identified regarding the current versions of the predictive models.

Interpolation/extrapolation
As with most linear regression models of this form, the models were found to be good predictors if the PAC profile and dose of the untested petroleum substance fell within the PAC profiles and dose that had been used for model development (i.e. the prediction would be an interpolation). Not surprisingly, the models were sometimes less accurate predictors if the PAC profile and/or doses of the unknown petroleum substance fell outside the PAC profiles that had been used for model development (i.e. the prediction would be an extrapolation). To investigate and mitigate this recognized limitation requires more studies on substances whose Method 2 PAC profiles and doses are outside the profiles and doses that were used to develop these models, if any such substance can be found.

In the future, as new test data become available, they could be incorporated into the current models, further validating the models and strengthening their usefulness.

Domain of applicability
A spectrum of petroleum substances containing PAC was used in this investigation. Since the models were developed on the basis of information on the PAC profile of petroleum

substances, the models will apply to a wide range of petroleum substances that contain PAC. However, there may be other factors that should be considered before using the models, e.g. physical characteristics such as form or viscosity that could limit bioavailability. Therefore, the models should be used judiciously, to ensure that possible erroneous predictions are avoided.

Route of exposure
The largest toxicological data set available for evaluation was from studies conducted in rats using repeated dermal exposures. Those studies form the basis for all the predictive modeling work that was done. It follows, therefore. that application of the models to other routes of exposure and species is not justified at this time.

Mechanism of Action
The models were developed based on observed statistical relationships. No attempt was made to identify causal relationships. To do this would have required a detailed understanding of the mechanisms of PAC-toxicity, an exercise beyond the scope of the evaluation. No inferences should be made concerning which rings are responsible for adverse effects based on the form of the models or the magnitude of the coefficients in the models.

6.3 Use of Models

Selection of substances for testing
The models that have been developed can be used to intelligently select samples for biological testing. As more Method 2 compositional information becomes available for the substances in PAC-containing petroleum categories, it will be possible to identify the compositional boundaries for each category. The models can then be used to identify those samples that would require a prediction by extrapolation, and these samples could be selected for testing. The models could then be adjusted by an iterative process thereby improving the models' utility.

Prediction of toxicity for untested substances
The toxicity of an untested substance can be predicted with confidence provided that the prediction is an interpolation and the physical parameters of the untested material are similar to those materials used to build the models. Untested materials, whose PAC values are within the range of tested substances, might have characteristics, such as viscosity, that might influence the bioavailability of the substance, therefore altering the biological response. Initially these should be considered on case by case basis using best professional judgment as to the effect such physical differences could have on the model results and application. As new toxicological and compositional data become available the confidence in the models will increase.

7. References

American Petroleum Institute (API) Petroleum HPV Testing Group; Heavy Fuel Oils Category HPV Category Test Plan, June 17, 2004; Posted to U.S. EPA website July 2, 2004; http://www.epa.gov/oppt/chemrtk/heavyfos/c15368tc.htm

American Petroleum Institute (API) Petroleum HPV Testing Group; Kerosene/Jet Fuel Category HPV Test Plan, December 31, 2003; posted to U.S. EPA website March 3, 2004; http://www.epa.gov/oppt/chemrtk/kerjetfc/c15020tc.htm

American Petroleum Institute (API) Petroleum HPV Testing Group; Crude Oil Category HPV Test Plan, November 21, 2003; posted to U.S. EPA website December 19, 2003; http://www.epa.gov/oppt/chemrtk/crdoilct/c14858tc.htm

American Petroleum Institute (API) Petroleum HPV Testing Group; Gas Oils Category HPV Test Plan, November 3, 2003; posted to U.S. EPA website December 16, 2003; http://www.epa.gov/oppt/chemrtk/gasoilct/c14835tc.htm

American Petroleum Institute (API) Petroleum HPV Testing Group; Lubricating Oil Basestocks Category HPV Test Plan, March 24, 2003; posted to U.S. EPA website April 4, 2003; http://www.epa.gov/oppt/chemrtk/lubolbse/c14364tc.htm

American Petroleum Institute (API) Petroleum HPV Testing Group; Waxes and Related Materials Category HPV Test Plan, August 6, 2002; posted to U.S. EPA website August 22, 2002; http://www.epa.gov/oppt/chemrtk/wxrelmat/c13902tc.htm

American Petroleum Institute (API) Petroleum HPV Testing Group; Gasoline Blending Streams Category HPV Test Plan, December 20, 2001; posted to U.S. EPA website January 25, 2002; http://www.epa.gov/oppt/chemrtk/gasnecat/c13409tc.htm

ATSDR, (1995)
Toxicological profile for polycyclic aromatic hydrocarbons (PAH)
Agency for Toxic Substances and Disease Registry (ATSDR), Atlanta, GA., US Department of health and human services, Public health services

CONCAWE (1994)
The use of the dimethylsulphoxide (DMSO) extract by the IP 346 method as an indicator of the carcinogenicity of lubricant base oils and distillate aromatic extracts
CONCAWE report 94/51
CONCAWE, Brussels, February 1994

Draper, NR, and Smith, H, (1998)
Applied Regression Analysis, 3rd ed, Wiley and Sons, NY,

Feuston, M. H., Low, L. K., Hamilton, C. E. and Mackerer, C. R. (1994)
Correlation of systemic and developmental toxicities with chemical component classes of refinery streams.
Fundamental and Applied Toxicology 22, 622-630

Gephart, L. A., Salminen, W. F., Nicolich, M. J. and Pelekis, M. (2001)
Evaluation of subchronic toxicity data using the benchmark dose approach
Regulatory Toxicology and Pharmacology, 33, 37-59

IPCS (International Programme on Chemical Safety) (1998)
Environmental Health Criteria: 202: Selected non-heterocyclic polyaromatic hydrocarbons.

WHO, Geneva, www.inchem.org

IRIS (Integrated Risk Index System) (2007)
U.S. EPA Office of Research and Development, National Center for Environmental Assessment,
http://www.epa.gov/iris/index.html

Klimisch, H. J., Andreae, M, and Tillman, U. (1997)
A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data
Regulatory Toxicology and Pharmacology, 25, 1-5

OECD (2004)
Manual for Investigation of HPV Chemicals
OECD Secretariat, September 2004

OECD (2006)
OECD SIDS Manual Sections 3.4 and 3.5
OECD Secretariat, October 19, 2006

RAIS (The Risk Assessment Information System) (2007)
U.S. Dept. of Energy. http://rais.ornl.gov

SCF (Scientific Committee on Foods) (2002)
Opinion of the Scientific Committee on Food on the risks to human health of polycyclic aromatic hydrocarbons in food
Scientific Committee on Food, Brussels, Belgium
http://europa.eu.int/comm/food/fs/sc/scf/index_en.html

U.S. EPA (2002)
A review of the reference dose and reference concentration process, EPA/630/P-02/002F
Risk Assessment Forum, December, 2002, Pg 4-11