THE HPV CHALLENGE PROGRAM AND PETROLEUM SUBSTANCES

VOLUME 2:

AN INVESTIGATION INTO THE RELATIONSHIP BETWEEN THE POLYCYCLIC AROMATIC COMPOUND CONTENT AND ACUTE, REPEAT-DOSE, DEVELOPMENTAL, AND REPRODUCTIVE TOXICITY OF 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

July 31, 2007

Executive Summary/Conclusions

In several Test Plans submitted by the Petroleum HPV Testing Group to U.S. EPA for the High Production Volume (HPV) Challenge Program it has been stated or implied that the repeat-dose, developmental, reproductive and genetic toxicities of some petroleum substances are related to their polycyclic aromatic compounds (PAC) content. Furthermore, the same Test Plans suggested that the PAC content of an untested petroleum substance could be used to predict its toxicity. U.S. EPA and others have asked for more support for this assertion. The Test Plans in question are for the categories: Aromatic Extracts, Crude Oil, Gas Oils, Heavy Fuel Oils, Lubricating Oil Base Stocks, and Waxes and Related Materials

A Task Group (TG) was commissioned by the Petroleum HPV Testing Group to study these claims further and was set these objectives:

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.**
2. **Identify and characterize relationships between PAC content and SIDS mammalian toxicity endpoints.**
3. **For any identified relationships, determine if they could be used to predict the toxicity of untested petroleum substances.**

The TG was provided with a sufficient number of studies on developmental toxicity and repeat-dose toxicity to address the project objectives for these two mammalian toxicity endpoints. A review of the relationships between PAC and genetic toxicity is reported separately in Volume 3 of this series "An evaluation of the relationship between the PAC content of selected categories of petroleum substances and their genetic toxicity."

The TG found that a previous published report of the correlations of total PAC content with repeat-dose and developmental toxicity was not robust enough to meet the objectives of this project. However, the TG found in their review of unpublished data that predictive models could be developed for effects on these endpoints by using the weight percent of each of the 1- through 7-ring compounds in the test substance (referred to as the PAC profile). As identified in this project the effects associated with the PAC profile are consistent with those reported for a number of individual PAHs, although the mechanism(s) of PAH-toxicity in this regard are themselves unclear.

In the repeat-dose studies, the TG found an association between the PAC profile and effects on thymus weight, liver weight, hemoglobin concentration and platelet count. In the developmental toxicity studies, associations were also found for effects on fetal weight, number of live fetuses/litter and percent resorptions in the prenatal studies (studies in which the pups were delivered by caesarean section) and pup weight, total litter size and number of live pups/litter in the postnatal studies (in which the 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 provided toxicity predictions on the basis of the PAC profile. The models were developed based on observed statistical relationships. No attempt was made to identify causal relationships. To do this would have required an understanding of the mechanisms of PAC-toxicity, an exercise beyond the scope of the TG's objectives.

For the assessment of the relationship between PAC content and fertility there were only two studies to review, neither of which was compliant with OECD-guidelines. There were no other data on petroleum substances that allowed an evaluation of the relationship between PAC and fertility. To address some endpoints related to fertility and other reproductive effects, the TG reviewed the repeat-dose toxicity studies for effects on reproductive organs and the developmental toxicity studies for effects on the fetus. The TG found that the developmental toxicity study endpoints, including both *in utero* and postnatal development, were affected even though no effects were observed on the reproductive organs in the repeat-dose studies and in this regard were considered more sensitive.

There were certain qualifications about the use of these models including the limitation to dermal studies in rats. Furthermore, the TG found that 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. interpolation) 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. extrapolation) were less certain. To evaluate this potential limitation, the models can be used to select new substances to study and help expand the models' coverage (if necessary). The models could then be adjusted and by an iterative process the utility of the models would be optimized. Similarly, by an iterative process the models can also be strengthened when new data become available on substances whose PAC profiles are within the bounds of the profiles of those substances used to develop the models.

TABLE OF CONTENTS

List of Appendices

List of Tables

List of Figures

1. Introduction

This is the second volume in a series of three volumes that reports the findings of a project to investigate the relationship between the polycyclic aromatic compound (PAC[1]) content and the potential mammalian toxicity (SIDS endpoints) of selected classes of petroleum substances. The first volume provided the background to the project and outlined the strategy adopted by the Petroleum HPV Testing Group to meet the mammalian toxicity requirements of the HPV challenge program. This, the second volume, describes the work undertaken to assess the relationship between the PAC content of selected petroleum substances and their repeat-dose, developmental and reproductive toxicities. The third volume describes the evaluation of the relationships between PAC content and the genetic toxicity of selected petroleum substances.

In some of the Petroleum HPV Testing Group's Test Plans, it was either stated or implied that the repeat-dose toxicity, genotoxicity, developmental toxicity and reproductive toxicity are associated with polycyclic aromatic compounds (PAC) content. It was also implied in the same Test Plans that the PAC content of selected petroleum substances could be used to predict the toxicity of similar, untested petroleum substances. The claims were made for Aromatic Extracts, Crude Oils, Gas Oils, Heavy Fuel Oils, Lubricating Oil Base Stocks, and Waxes and Related Materials. The basis for the claims was a publication by Feuston et al. (1994) that 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. The authors concluded:

> "*In general, toxicity was correlated with concentrations of polycyclic aromatic compounds (PAC) composed of 3, 4, 5, 6, and/or 7 rings (decreased thymus weight, increased liver weight, aberrant hematology and serum chemistry, increased incidence of resorptions, decreased fetal body weight), PAC containing nonbasic nitrogen heteroatoms (increased mortality, decreased body weight, decreased thymus weight, increased liver weight, decreased hemoglobin content, hematocrit level, decreased fetal body weight), and/or PAC containing sulphur heteroatoms (decreased red blood cell and platelet counts, increased sorbitol dehydrogenase). A relationship between 2- ring PAC and skin irritation was demonstrated. Severity of effect was ranked against concentration of component class and statistical significance determined by the rank order correlation of Spearman. For the 13 streams tested, the presence and severity of systemic and developmental toxicity were dependent upon the levels of PAC and nonbasic nitrogen PAC.*
>
> *It is reasonable to assume that refinery streams rich in 3- to 7-ring PAC, S-PAC, and nonbasic N-PAC (e.g., carbazole derivatives) would be toxic, not only to the adult animal, but to the fetus as well."*

The Petroleum HPV Testing Group recognized that the underlying data for the publication by Feuston et al (1994) were limited and a more sophisticated and robust analysis was needed to assess the possible relationship between the PAC content and SIDS mammalian toxicity endpoints of petroleum substances. Consequently, a Task Group (TG) comprised of experts in the fields of petroleum chemistry, toxicology, and biostatistics was commissioned. The objectives of the TG were:

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 mammalian toxicity endpoints required in the HPV Challenge.**

[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.

2. **Identify and characterize relationships between the PAC content and Screening Information Data Set (SIDS) mammalian toxicity endpoints of petroleum substances.**
3. **For any identified relationships, determine if they could be used to predict the toxicity of untested petroleum substances.**

To accomplish the three objectives with regard to repeat-dose, developmental and reproductive toxicity, the TG followed these steps:

1. Acquired company toxicology and compositional study reports that might contain data that would be useful in addressing the project's objectives.
2. Evaluated the reports for their data reliability and potential usefulness in addressing the project's objectives.
3. Identified all reports that were judged reliable and potentially useful in addressing the project objectives.
4. Captured, from the identified reports, any data judged potentially useful in addressing the project objectives.
5. Identified, for quantitative evaluation of possible dose-response relationship(s), those biological endpoints most often statistically significantly affected in the studies, and which would be considered biologically significant.
6. Developed, for the endpoints identified in Step 5, preliminary quantitative assessment(s) of the dose-response relationship(s) between PAC content and effects.
7. Selected the appropriate PAC assessment method (during Step 6).
8. Identified, from those biological endpoints assessed in Step 6, those which could be mathematically characterized with a high degree of accuracy (i.e. high r value) and were unique endpoints (e.g. among hematocrit, hemoglobin, and erythrocyte count only hemoglobin was identified for final modeling in Step 9).
9. Developed, for the endpoints selected in Step 8, final mathematical characterization(s) of the dose-response relationship(s) between PAC content and potentially adverse effects.
10. Developed, if possible, examples of percent changes from control values that could be considered potentially biologically meaningful for the endpoints for which final mathematical characterizations of the dose-response relationships had been developed in Step 9 and calculated the "Predicted Dose x (PDx)", where "x" indicates the degree of change from the control value.
11. Evaluated, for those endpoints for which examples of PDx values were developed, the utility of finalized dose-response models in predicting low effect and no-effect levels of untested petroleum substances.

Details of the process the TG followed are described in the remainder of this report, which is comprised of two parts. The essential information on the work of the TG, the approaches it has taken and its findings are presented in an abbreviated form in the body of the report. The appendices provide more detail on the key elements of the process.

2. Identify, Obtain, and Evaluate Available Information

The first objective for the TG was to 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 SIDS mammalian toxicity endpoints (OECD, 2006).

2.1 Identification of Useful Information

In the context of the HPV program, the SIDS mammalian toxicity endpoints are:

- acute toxicity
- repeat-dose toxicity
- genetic toxicity (gene mutations and chromosomal damage)
- developmental toxicity/teratogenicity
- reproductive toxicity.

The TG decided not to investigate an acute toxicity/PAC relationship since the reported oral LD_{50} values for aromatic extracts, crude oil, gas oils, heavy fuel oils, lubricating oil base stocks and waxes 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 the TG to conclude it was not worthwhile to investigate any possible relationship between acute toxicity and PAC content.

The evaluation of the relationship(s) between the PAC content of selected classes of petroleum substances and genetic toxicity is described in Volume 3 of this series of reports.

With regard to developmental, reproductive and repeat-dose toxicity data, earlier attempts to identify, during preparation of HPV Test Plans, published information other than Feuston et al. (1994) had not produced any additional information.

The TG concluded that if additional information were available that could be used to define PAC/ toxicity relationships (SIDS endpoints) it would most likely be in the form of unpublished studies sponsored by API member companies.

2.2 Acquisition of Information

Initially, two Petroleum HPV Group member companies provided the TG with copies of unpublished studies. This initial set of studies included:

- forty-six reports of repeat-dose toxicity studies,
- sixty-three reports, comprised of developmental toxicity (60 reports), reproductive toxicity (2 reports) and an exploratory dose range-finding study in non-pregnant female rats, and
- one hundred and fifty-three reports of accompanying compositional data.

To determine if additional studies were available, a letter was sent to all Petroleum HPV Group member companies. The letter asked for copies of reports of repeat-dose, developmental, and reproductive toxicity studies and corresponding compositional data. No request was made for acute toxicity studies, for reasons discussed previously. In the letter, the companies were asked to submit only studies that met the following minimum criteria:

- a complete copy of the laboratory report was available,
- the test sample was identified by CAS number,
- the PAC content of the test sample had been quantified; specifically the 3-7 ring PAC, S-PAC, and N-PAC content, and
- the test sample was a substance in one of the following API HPV categories:
  - Aromatic Extracts
  - Crude Oils

Gas Oils
Heavy Fuel Oils
Lubricating Oil Base Stocks
Waxes and Related Materials.

No additional reports of repeat-dose, developmental or reproductive studies were provided in response to the first solicitation letter.

Since only two reports of non-guideline reproductive studies were provided, the TG did not attempt to include fertility in this phase of the assessment due to lack of data. However the TG did have information on potential reproductive organ effects from repeated-dose studies as well as data on postnatal effects from some of the developmental toxicity tests. The TG noted that U.S. EPA guidelines indicate that an evaluation of the reproductive organs from a 90-day repeat-dose toxicity study together with an evaluation of the results from a guideline developmental toxicity study on the same substance can be sufficient to make an evaluation for reproductive toxicity. This topic is discussed in more detail in **Section 4.6** and **Appendix 8**.

A complete listing of the studies provided to the TG, together with details of any that were excluded from the evaluation can be found in **Appendix 2**.

2.3 Evaluation of Information

In general, the company reports submitted to the TG met the criteria set out in the solicitation letters. A number of the study reports did not identify the test sample by CAS number, but provided only a refinery substance name. Generally the refinery substance name was sufficient to allow the TG to assign an appropriate CAS number to the test material. Several studies were supplied that did not fall into the HPV categories specified in the solicitation letters. Nevertheless, the TG decided to use these studies if they could provide data for the assessment. These included studies that had been carried out on petroleum substances for which there was a description but for which a specific CAS number could not be assigned, and studies that had been carried out on substances other than those specified in the solicitation letters but which contained PAC; the TG categorized these as “Other”. The details of the compositional information on the test samples’ PAC content varied among studies depending on the analytical chemical methods that had been used. If a report (toxicity or compositional) appeared to be missing information, the company providing the report was asked to perform a further search in order to ensure the completeness of the data supplied to the TG.

Each of the submitted company reports was assessed for reliability of the data it contained according to Klimisch (Klimisch, et al. 1997). In this regard, all data (toxicity and analytical) were judged by the TG to be “reliable without restrictions”, i.e. a reliability score of 1.

2.3.1 Toxicological Studies

The materials that had been tested in the submitted toxicity studies covered a range of petroleum substances and HPV categories.

Repeat-dose toxicity studies
Of the forty-six reports provided to the TG, nineteen were of 28-day and twenty-seven were of 90-day repeat-dose studies. All but one of the repeat-dose toxicity studies were in rats exposed dermally, the exception being a 90-day mouse study involving both oral and dermal exposure routes. In all cases the test materials had been applied undiluted. In the 90-day studies the test materials had been applied to non-occluded skin, and the animals had been fitted with Elizabethan collars to prevent or minimize ingestion. In the 28-day studies the test material was occluded for 6 hours after application.

As noted above, the Task Group concluded that all the repeat-dose studies were highly reliable (Klimisch = 1).

Developmental toxicity studies
The sixty reports of developmental toxicity studies provided to the TG contained results of 67 developmental toxicity studies. Of the 67 studies, four were studies that had been carried out using the oral route. The remaining 63 studies were conducted in the rat using the dermal route of exposure.

The duration of dosing was not the same in all of the 63 dermal developmental toxicity studies. In many studies, dosing had been throughout gestational days 0-19, whereas in others dosing had been for a limited number of days during gestation, e.g. days 1-3, 4-6 etc. Group sizes varied from study to study. The number of mated females ranged from 8-25 per group. In all 63 studies, the test materials had been applied undiluted to non-occluded skin, and the animals had been fitted with Elizabethan collars to prevent or minimize ingestion

The 63 dermal studies were of two designs. Twenty-eight studies had been carried out in which the test material was applied to the skin of pregnant females during gestation after which the pregnant dams underwent a *caesarean* section on day 20 of gestation; such studies, designated “Prenatal studies”, assessed *in utero* development. In the other thirty-five studies (designated “Postnatal studies”) the dosing period was similar to the prenatal studies, but the dams were allowed to deliver and pups were monitored for 4 days of lactation. These studies provided information on postnatal effects including offspring viability and weight gain.

As noted above, the TG concluded all the developmental studies were highly reliable (Klimisch = 1).

2.3.2 Compositional Data

The one hundred fifty-three analytical reports available to the TG covered a range of analytical data and samples. The reported analytical data had been derived principally from five different methods; each identifying different chemicals or groups of chemicals (see **Table 1**). Information on individual PACs was available for only a limited number of samples. A further series of reports included information on class 1-7 ring PAH that had not been determined by analysis, but derived by calculation only. Results from the five analytical methods varied in their usefulness in developing predictive models see **Section 2.4.2**. The TG considered that all the analytical reports had a high degree of data reliability (Klimisch = 1)

A description of the various analytical techniques and the number of samples analyzed by each is given in **Appendix 1**.

Table 1. 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 Sulfur
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 3 Preparative Liquid chromatography and GC/MS to separate and identify individual compounds in six separated fractions	% Individual components within the following fractions: • Paraffins • Alkylnaphthalenes & Alkylbiphenyls • 3-Ring PACs • 3-Ring and 4-Ring PACs • 4-Ring PACs • Alkylcarbazoles and alkylbenzcarbazoles
Method 4 Preparative Liquid chromatography and GC/MS to separate and identify individual PAH	% 16 U.S. EPA PAHs % methylnaphthalenes
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.

2.4 Identification of Studies

The TG recognized that only comparable data could be used in meaningful statistical evaluation and modeling. Consequently, as discussed below, for each group of studies (repeat-dose, developmental, compositional) criteria were developed for the identification or exclusion of information. The TG identified studies only on the basis of availability of suitable biological and matching compositional information. It should be noted that identification was made before any statistical evaluation of data was undertaken. More details on the criteria that were used by the TG to identify or exclude studies from the evaluation are given in **Appendix 3**.

2.4.1 Toxicological Studies

Since the TG rated all the studies as highly reliable (Klimisch = 1), none were excluded for reasons of reliability or data quality. However, only toxicology studies accompanied by “sufficient” compositional data were selected for inclusion in the data set used to develop quantitative characterizations of relationships and predictive mathematical models. Concentrations of aromatic

compounds of ring classes 1 – 5 and 1 – 7, including S- and N-PAC, generated using either the Method 1 or Method 2 analytical methods (see **Section 3.1.2**), were the only empirically-derived data generated on a sufficiently large set of samples to provide a basis for comparison.

Repeat-dose toxicity studies

As noted above (**Section 2.3.1**), the repeat-dose studies were a mixture of 90- and 28-day studies. The TG agreed that data from both 28- and 90-day studies should be used in assessing the relationship between PAC content and toxicity. The difference in duration of dosing between 28- and 90-day results was considered in the statistical analysis (see **Section 3.4.1**).

In some repeat-dose studies, clinical chemistry and hematological data had been determined for animals that had not survived to study termination. The TG agreed that only data from animals surviving to study termination should be used.

The TG decided, because of possible species differences, not to use the previously mentioned 90-day mouse study in the evaluation.

Developmental toxicity studies

Animals had been exposed via the oral route in a limited number of studies. The TG did not think this limited number of studies comprised a robust data set on which to base statistical analysis. Consequently, the TG agreed that data from any orally administered doses should be excluded, and only results from studies using the dermal route of exposure should be used in developing predictive models.

The duration of dosing was not the same in all the developmental toxicity studies from which data were extracted. To ensure the modeling results were comparable, the TG decided to use only data from studies that included daily dosing on gestational days 0-19 as a minimum. Those studies that involved dosing on days 0-20 of gestation or from pre-mating day 7 through gestation day 20 were included. However, studies were excluded in which dosing was for only a portion of the gestation period (i.e. less than gestation days 0-19). Additional details of the identification of studies for use in the analysis can be found in **Appendix 2**.

Within individual studies, group sizes with three or fewer dams with viable fetuses (prenatal endpoints) or litters (postnatal endpoints) were excluded from the modeling and statistical analysis because the TG considered the group size inadequate. A small number of data points were excluded because the group size was three or less. Without the exclusion based on small group size, the slope of the modeled curve could be based heavily on a single data point representing only one or two litters. Thus, exclusion of data based on an inadequate group size provided a more scientifically defensible basis for modeling the data. This is discussed in greater detail in **Appendix 2**.

As a result of the study selection process, the number of studies that were used in the evaluation of the relationship between PAC content and toxicity are shown in **Table 2**.

Table 2. 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
Crude Oil	0	2	2	4
Gasoline	2	0	1	2
Gas Oils	1	4	7	9
Heavy Fuel Oils	0	8	10	15
Lubricating Oils	0	1	0	1
Aromatic Extracts	0	1	1	0
Other	1	2	2	2
TOTAL	4	18	23	33

2.4.2 Compositional Data

The TG decided not to use data from analytical methods that had produced values for specific unalkylated PAHs (marker compounds) since the mechanism of PAH toxicity is unknown and in previous evaluations of the relationship between PAC content and dermal carcinogenicity the use of benzo(a)pyrene as a marker PAH was not found to be sufficiently useful (CONCAWE, 1994).. The TG also decided not to use the data from any method that had only been used on a limited number of samples or that that had been derived. The following data were therefore excluded:

- Data from individual analytical methods with insufficient number of samples examined, e.g. only one sample had been analyzed for benzo[a]pyrene,
- Data generated from Method 3 since it only identified the levels of 24 specific PAH/PACs,
- Data generated from Method 4 since it measured only 16 specific PAHs and methylnaphthalenes, and
- Data on class 1-7 ring PAH derived by calculation only (see **Appendix 1**)

2.5 Objective 1 Conclusions

The TG identified and obtained a large number of toxicity and compositional reports that it judged useful for meeting the project objectives. All the reports were judged by the TG to be “reliable without restrictions”, i.e. a reliability score of 1 (Klimisch, et al. 1997). The TG considered the size of the database used for the evaluation to be large and a particular strength of the investigation.

3. Characterize Relationships between PAC Content and Mammalian Toxicity (SIDS Endpoints)

The second objective was to identify and characterize relationships between PAC content and mammalian toxicity (SIDS endpoints).

The TG concentrated its efforts on two of the SIDS mammalian toxicity endpoints, repeat-dose and developmental toxicity. These were the two endpoints, for which the TG had a sufficient number of toxicity and analytical studies available. The identification of these studies was previously described in **Section 2**.

To accomplish objective 2, the TG completed the elements described below.

3.1 Capture of Potentially Useful Data

The TG captured, from all the identified repeat-dose and developmental studies (see **Section 2**) and the corresponding compositional reports, all those experimental observations/measurements that might be useful in subsequent evaluations. To this end, a series of spreadsheets was developed (**Table 3**) and populated with the appropriate information taken from each study report. Information from each study was tabulated by study, test material, sex (where appropriate), dose level and biological endpoint. A list of all the biological endpoints for which data was captured is given in **Tables A4-1** and **A4-2** in **Appendix 4**. For the repeat-dose studies, with the exception of clinical observations, necropsy findings and histopathological findings, the TG collected all quantitative/incidence data presented in the reports. With regard to clinical findings, the TG collected only data on the incidence of dermal irritation. Gross necropsy findings were not captured. Histopathological findings were captured in a qualitative manner, i.e. the TG simply noted the presence of any histopathology findings in liver, thymus or bone marrow. The developmental toxicity spreadsheets captured all of the endpoints of developmental toxicity evaluated in the studies.

Table 3. Spreadsheets Developed for Capture of Biological Data

Repeat-dose studies	Developmental toxicity studies
Hematology	Maternal endpoints
Clinical chemistry	Developmental endpoints – prenatal
Urinalysis	Developmental endpoints – postnatal
Other endpoints e.g. body and organ weights	

After initially being populated, each of the data extraction sheets (repeat-dose, reproductive/developmental, compositional) was checked for accuracy by a member of the TF that had not been involved in populating the sheet. The check involved comparing the values entered in the sheet with those found in the laboratory reports.

Subsequently, as the project progressed and models were being developed, the data sheets were "spot-checked". This was not an organized data check, but was usually done in response to questions/issues that arose as the models were being developed.

Finally, all values entered on the final data sheets used in the modeling effort were checked for accuracy by two members of the TF not involved in developing the models. This check involved comparing the entered data to the data on the previously checked data extraction sheets.

3.1.1 Identification of Biological Endpoints for Evaluation and Modeling

The TG recognized that it would be unnecessary to characterize the PAC content – toxicity relationship for all the biological endpoints on which it had collected data (see **Section 3.1**) since only those of biological relevance were of interest for the evaluation. Consequently, the TG identified a number of biological endpoints that would undergo preliminary quantitative assessment for possible dose-response relationship(s) between PAC content and endpoint-specific effects. During this preliminary quantitative evaluation the usefulness of the various compositional data sets was also assessed.

The process used by the TG to identify the relevant endpoints for final evaluation and statistical modelling is described in detail in **Appendix 3** and consisted of the following 3 steps:

1. identify those endpoints most often statistically significantly affected in the studies,

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. from the endpoints identified in steps 2 and 3, select those for which a relationship between PAC profile and effect had been confirmed statistically and develop the models further to improve their degree of correlation of the mathematical dose-response characterization.

The outcome of each step of the selection process is summarized in **Table 4**.

Table 4. Biological Endpoints Affected and Those Identified for Statistical Evaluation

Endpoint	Affected most often (statistically)	Sensitive Endpoint [1]	Good correlation in preliminary statistical evaluation [2]	Used for final model development
Repeat-dose toxicity studies				
Liver wt (abs)	√	√		
Liver wt (rel.)	√	√	√	√
Thymus wt (abs)	√	√	√	√
Thymus wt (rel.)	√			
RBC	√	√		
Hb conc.	√	√	√	√
Hematocrit	√	√		
Platelet count	√	√	√	√
Developmental toxicity studies (Maternal endpoints)				
Body wt	√			
Body wt gain	√			
Food consumption	√			
Liver wt (rel.)	√			
Thymus wt (abs)	√			√
Thymus wt (rel.)	√			
Uterine wt (abs)	√			
Developmental toxicity studies (Prenatal)				
Live fetuses/litter	√	√	√	√
Resorptions/litter	√	√	√	
% Resorptions	√	√	√	√
Fetal body wt	√	√	√	√
Delayed ossification	√	√		
Developmental toxicity studies (Postnatal)				
Total pups/litter PND 0	√	√	√	√
Live pups/litter PND 0	√	√	√	√
Pup body wt PND 0	√	√	√	√
Pup body wt PND 4	√	√		

[1] This endpoint was among those most often statistically significantly 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).
[2] $r > 0.75$
PND post natal day

3.1.2 Identification of Compositional Data for Use in the Quantitative Evaluation of Possible Dose-Response Relationship(s)

As noted in **Section 2.3.2,** the analytical reports that had been selected for use in assessing relationships between PAC content and mammalian toxicity contained compositional data that had been derived from several methods, each identifying different chemicals or groups of chemicals. In addition to identifying biological endpoints (see **Section 3.1.1.**), the TG undertook to assess whether different compositional data would have various degrees of usefulness in this assessment.

The evaluation of the utility of the various compositional data that were available to the TG was made with linear regression models and a range of dependent and independent variables. For a more detailed discussion of the development of the mathematical characterizations, see **Section 3.4.**

The results of the evaluation of the various compositional data sets are summarized in **Table 5**. The identification of those biological endpoints that were used in this evaluation is described in **Section 3.1.1**.

Based on these comparisons, it was found that models developed on measured S-PACs and carbazoles did not fit the data as well as the models that were developed using compositional data on 1-7-ring compounds. The TG also found that ring-class compositional data derived from the Method 2 procedure (rings 1 – 7) produced models with a better fit than that derived using the Method 1 procedure (rings 1 – 5). See **Section 3.4** for a more detailed discussion.

Table 5. Summary of 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			
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 post natal day

3.2 Preliminary Assessment(s) of Dose-Response Relationship(s)

For each of the endpoints identified as described in **Section 3.1.1**, a preliminary mathematical characterization(s) of the dose-response relationship(s) with PAC content was developed. Dose group data for the biological endpoints chosen for characterization were matched to the appropriate compositional data to form a data set for analysis. The analytical and toxicity studies were matched

using the sample identification number, thus ensuring that the same sample had been used in both analytical and toxicology studies (see **Table A6-1** in **Appendix 6**). Initial characterization efforts were made with linear models and a selection of dependent and independent variables. For a more detailed discussion of the development of the mathematical characterizations, see **Section 3.4**.

3.3 Identification of Biological Endpoints for Final Statistical Characterization

After completing the preliminary quantitative assessment of the dose-response relationship(s) for each of the endpoints affected most often at the studies' LOELs (**Table 4**), the TG considered the following:

- 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.

Based on these considerations, endpoints were selected for final mathematical characterization (see **Table 4**). These endpoints are also listed in **Table 6**.

Table 6. 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 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.

See **Appendix 3** for a detailed discussion of how effects were identified for final mathematical characterization of the dose-response.

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 the TG selected for modeling (see **Appendix 11** for further details). Further support that the selected endpoints are reasonable is found in API's robust summaries and test plans for petroleum streams where the spectrum of effects of PAC-containing streams is similar to the endpoints that the TG selected for modeling.

3.4 Development of 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 6** can be found in **Appendix 6**.

3.4.1 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 3.4.3**.

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 3.3**. Control group responses were independent variables in the models (see **Section 3.4.1.2**).

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 the 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, 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 5**.

3.4.1.2 Choice of Independent Variables

Analytical variables
As noted in **Section 2.3.2** 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 (see **Section 3.1.2**). 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 3.1.2** 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 TG'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 3.4.1.1**.

3.4.1.3 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 the independent variable (covariate), and a selection of independent variables as described in **Section 3.4.1.2**.

The models for each endpoint 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 the fitting was to adhere to the principle of "parsimony", whereby, the simplest model that is adequate for the problem to be solved is used.

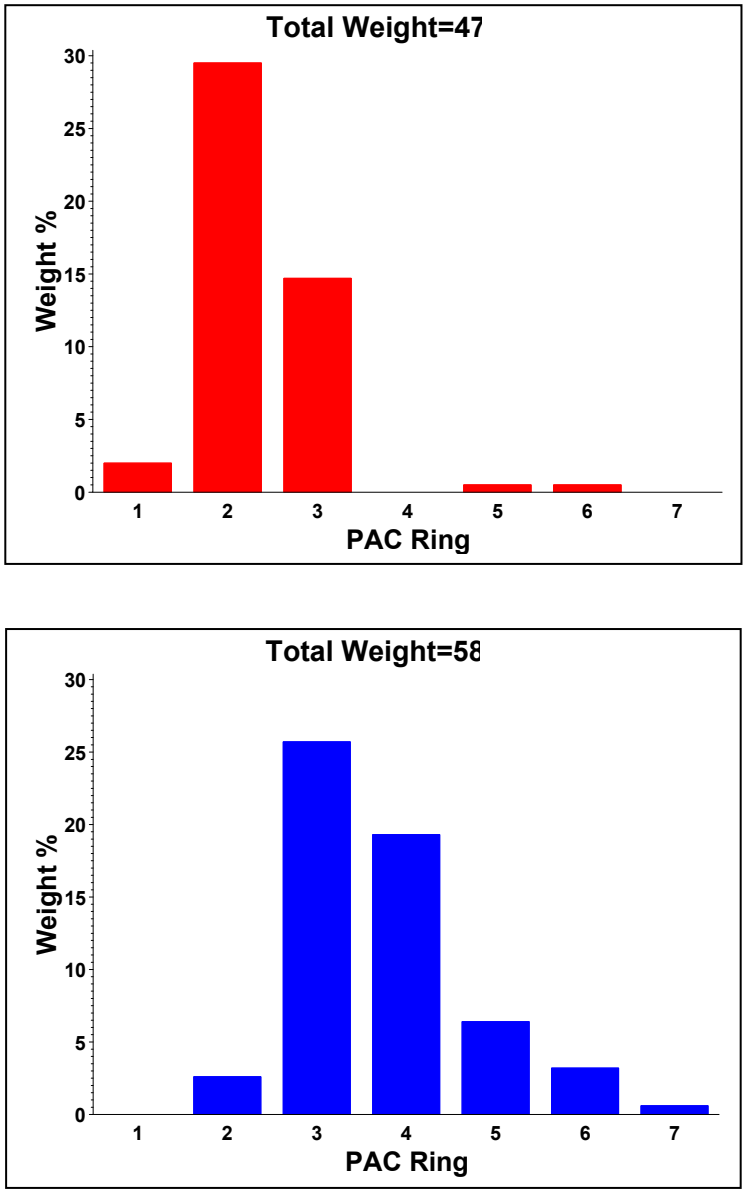
Table 7 shows only the values of r and se for the final models, and provides a basis for the reader to compare model adequacy and fit. These measures were selected from among the criteria used for evaluation because, among their other characteristics, the r value is a measure of the closeness of the observed and model predicted values and the se is related to the width of the confidence interval of the predicted value. Additionally, for the final models, the plots of the observed vs. the predicted values are presented and provide the most useful form for model adequacy. 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. To guard against these types of misinterpretations, plots of the observed and predicted values for each model are presented in Figure 4.

The basic model form was the linear model. 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.

3.4.1.4 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 (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 PAC profile of the petroleum substance. 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.

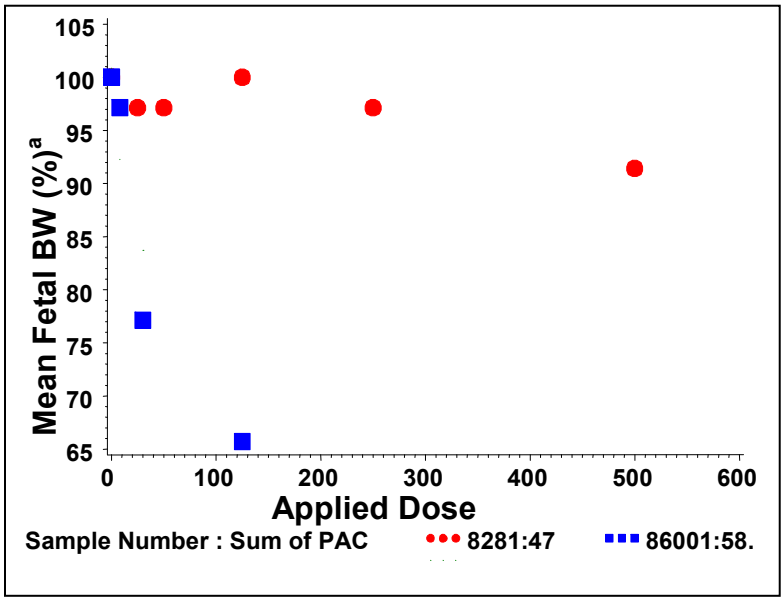
Figure 1. Weight Percent of 1- through 7-Ring Compounds of Two Petroleum Substances with Total PAC Extract Weights of 47 and 58 Percent



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 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, indicating 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. The mathematical model predictions for these two samples closely agree with the observed data indicating the usefulness of the models that were developed.

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

3.4.1.5 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). 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.

3.4.2 Final Model Results

The correlation and standard error (r and se) values in **Table 7** 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 from **Table 5**, comparisons cannot be made concerning the r and se from these two tables.

Table 7. Final Modeling Results with the Method 2 Results of PAC Weight %

Study Type	Dependent Variable	Transformation on Dependent Variable	n	r	se
Repeat –dose toxicity studies	Thymus Weight (absolute)	None	92	0.88	0.04
	Platelet Count	None	112	0.95	90.1[b]
	Hemoglobin Concentration	None	128	0.89	0.88
	Liver Weight (relative[a])	None	124	0.93	0.19
Developmental Toxicity Studies (Prenatal)	Maternal Thymus Weight (absolute)[c]	None	34	0.91	0.04
	Fetal Body Weight	None	67	0.96	0.10
	Live Fetuses/Litter	None	67	0.98	0.93
	Percent Resorptions	Probit	67	0.97	0.25
Developmental Toxicity Studies (Postnatal)	Pup Body Weight (PND[d] 0)	None	71	0.92	0.18
	Total Pups/Litter (PND[d] 0)	None	71	0.92	1.33
	Live Pups/Litter (PND[d] 0)	None	71	0.92	1.40

[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 initial models (results shown in **Table 5**) contain a term describing the compound's HPV group (Crude Oil, Aromatic Extracts, Gas Oils, Heavy Fuel Oils, Lubricating Base Oils or Waxes). The TG realized that the classification requirement for a specific compound would limit the overall applicability of the models and 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 7**) did not depend on a substance being assigned to an HPV group but relied only on the PAC profile. Minor changes in model fit as a consequence of this change were considered acceptable consequences. It should also be noted that the initial models were based on the ratio of the dose response to the control response whereas the final models were based only on the dose response; therefore the standard errors (se) are not comparable between **Tables 5** and **7**.

The magnitudes of the correlations in **Table 7** are large; the minimum correlation is 0.88 with the majority being above 0.90. The TG recognizes that these correlations are large for this type of data. Possible explanations for the large 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 correlations were not spurious, based on bias, confounding, or affected by model specifications, the final models were rigorously tested as described later in **Section 3.4.3**.

3.4.2.1 Model Equations

The final models for the eleven endpoints considered 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 8**. The table lists each dependent variable and its transformation (if any), the selection of biologically based independent variables and the analytic 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 8**, labeled "Interaction Term Included" indicates if the model included an interaction term, of the form:

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

The TG notes that the models are complex, with the number of coefficients ranging from 10 to 22. **Section A6.4.4** in **Appendix 6** provides the coefficients and complete forms for all the models listed.

Table 8. Forms of the Eleven Final Models

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

[a] CG = Control Group

[b] relative to terminal body weight

[c] Interaction term also included term of the form $\sum_{k=1,2,6,7} \upsilon_k \cdot dose \cdot PAC^2_k$

[d] 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 a full assessment of such endpoints and their relation to PAC content using the final model was outside the scope of this project.

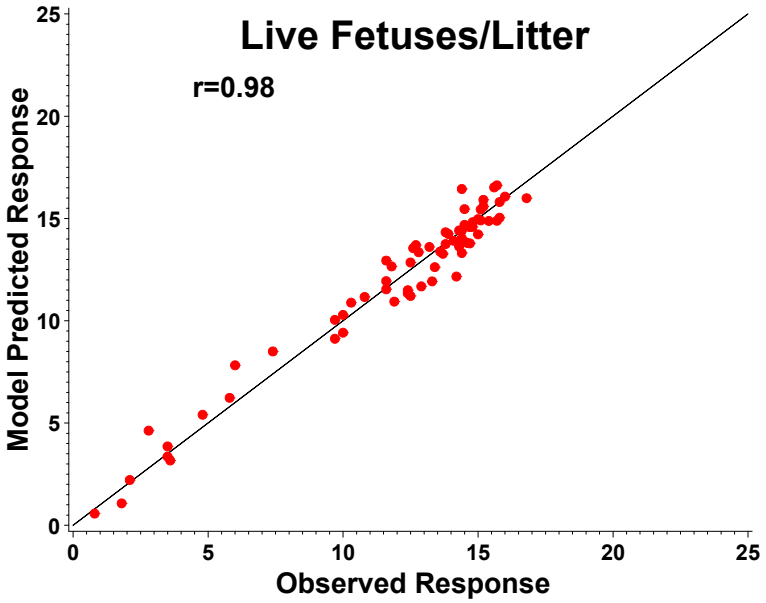
[e] PND = postnatal day

3.4.2.2 Model Fit

The accuracy of the fit of the models 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 live fetuses/litter model is shown in **Figure 3**. The correlation coefficient 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



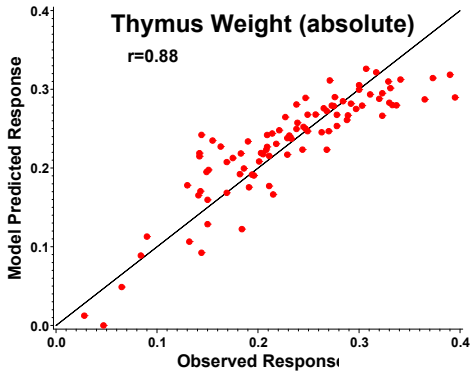
Similar plots for all eleven models are shown in **Figure 4**, with the live fetus/litter plot repeated for completeness.

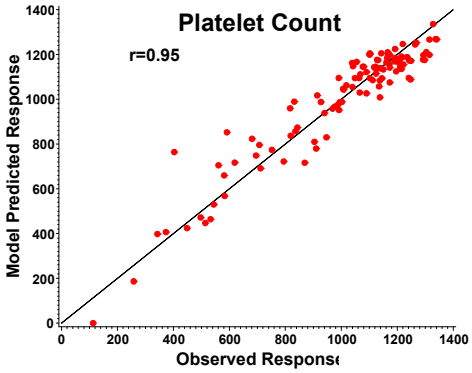
Note that the r values for all the models are equal to or greater than 0.88.

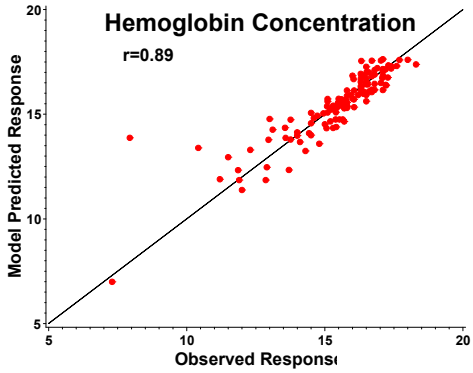
Note also that the r values in the figures are slightly different from the r values in **Table 5**. This difference is due to the fact that those in **Table 5** were derived from preliminary models whereas those in **Table 7** and **Figure 4** were derived from final models. Reasons for the difference in *r* values in **Tables 5** and **7** are given in **Section 3.4.2**.

Figure 4. Plots for Eleven Final Model Forms

Repeat-dose toxicity studies







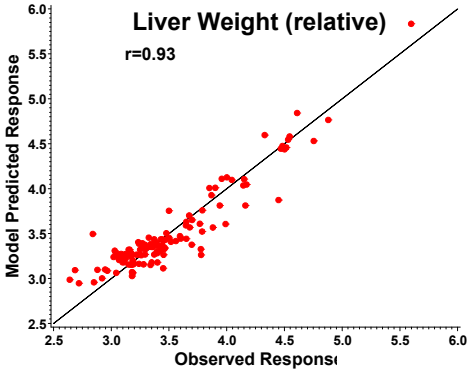
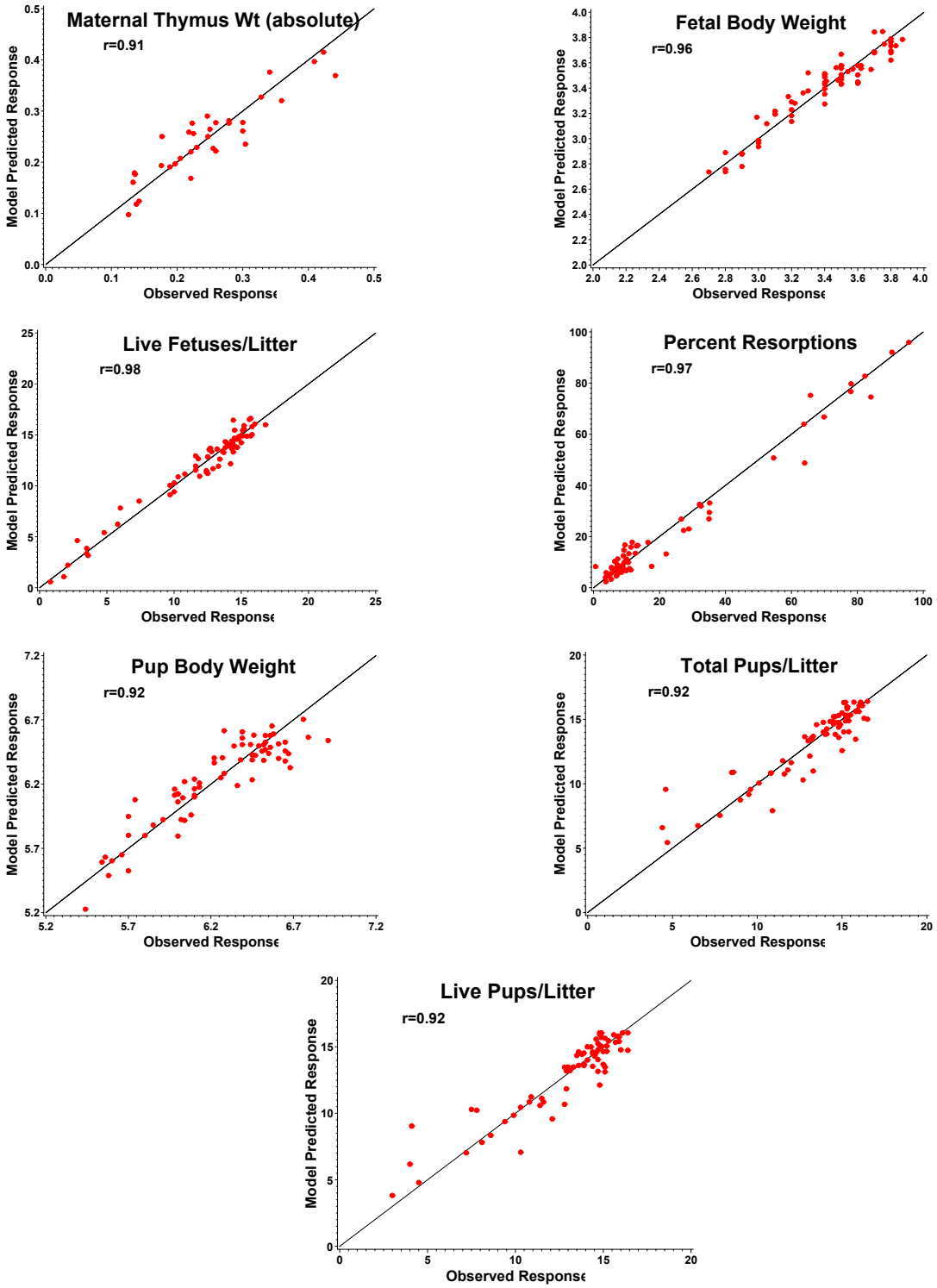


Figure 4 (cont.).

Developmental toxicity studies



3.4.3 Model Testing

An important component of model building is to test, or validate, the model's predictive ability. The models that were 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 models to be good predictors for data points that are interpolated from the existing data, but of questionable use for data points that are extrapolated. A summary of the results of the three model tests that were performed is given below.

3.4.3.1 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 hold-out 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 hold-out data. The data set that was used to develop the absolute thymus weight model was split where 30% of the data points were randomly selected and used as hold-out data and the model was developed from the remaining 70%. The model was then applied to the 30% hold out sample to see how well the model predicted the known results. This process was repeated 100 times.

Plots of the results in **Appendix 6** show the resulting models provide accurate predictions to the holdout samples when they are interpolated points, and are unreliable for points in the hold out sample that are extrapolated points.

The holdout samples indicated that 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 in a few instances 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 of interpolation and extrapolation appears in the "Limitations" section of this report (**Section 4.5**) and in **Appendix 6**.

3.4.3.2 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 outcome. The nonsense testing was applied to the hemoglobin concentration model. 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.
The results from the nonsense method of testing provided a clear indication that the model results are based on information in the data, and not from chance, or are related to the independent variables used in the model (see **Appendix 6**).

3.4.3.3 Model Testing Using Alternate Data Sources

The data available allows for using data from one source as a data set for predictions derived from models developed from a different source. 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. When the model is applied to these data the correlation between the observed and predicted data was 0.88 based on 92 observations. If the repeat-dose absolute thymus weight model (the same model form and the same coefficients) is used to predict the prenatal thymus weight data the correlation is 0.76 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 (prenatal 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 0.39. 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 0.36 (n=44). A fuller discussion with other examples can be found in **Appendix 6**.

3.5 Objective 2 Conclusions

A relationship between PAC content and certain toxicological endpoints has been identified and characterized. Preliminary statistical evaluations found compositional data generated using either Method 1 or 2 produced the most accurate models across a wide set of biological endpoints. The models that were developed fit the data used to develop them very well (r at least 0.88), and have been shown to be good predictors for data points that are interpolated relative to the existing data, but may not be useful for extrapolated data points. Preliminary statistical evaluations found compositional data generated using Method 2 generally produced the most accurate models.

4. Prediction of Toxicity of Untested Substances

The TG's third objective was to determine if any PAC-toxicity relationships could be used to predict the toxicity of untested petroleum substances.

This section describes how the models developed (**Section 3.4**) might be used for predictive purposes. Limitations on the utility of the predictive models are also discussed.

4.1 Prediction of Dose-Response Curves

As discussed in **Section 3.4**, eleven mathematical models have been 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 8**.

Predicted dose response curves may be generated with any of these models by following these steps:

1. Identify the specific endpoint model and data set used to develop that specific model.
2. Test the new sample against the first sample in the existing data set (i.e. the data set used to develop the specific model) by following these steps:
    a. Determine if the percent weight concentrations for the Aromatic Ring 1 concentration is *less than, equal to, or greater than* the corresponding concentration of the first sample in the existing data set.
    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. Determine if the applied dose of the new sample is less than or equal to the doses of the first sample in the existing data set (because there is always a control group it is not necessary to test for the applied dose being above some minimum value)
    d. If the answer is "less than" or "equal to" for all 7 PAC Ring concentrations and the applied dose, then the new sample is an interpolated point relative to the first sample in the existing data set.
    e. If any answer (dose or concentration) is "greater than", then the new sample is an extrapolated point relative to the first sample in the existing data set.
3. Test the new sample against the remaining samples in the existing data set as in step 2 above.
4. If the new sample is an interpolated point relative to ALL samples in the existing data set then the NEW SAMPLE IS AN INTERPOLATED POINT.
5. If the new sample is an extrapolated point relative to at least one sample in the existing data set, then the NEW SAMPLE IS AN EXTRAPOLATED POINT.

As an example consider the use of a model to generate dose-response curves for the live fetus/litter counts for two different samples. The model form is:

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

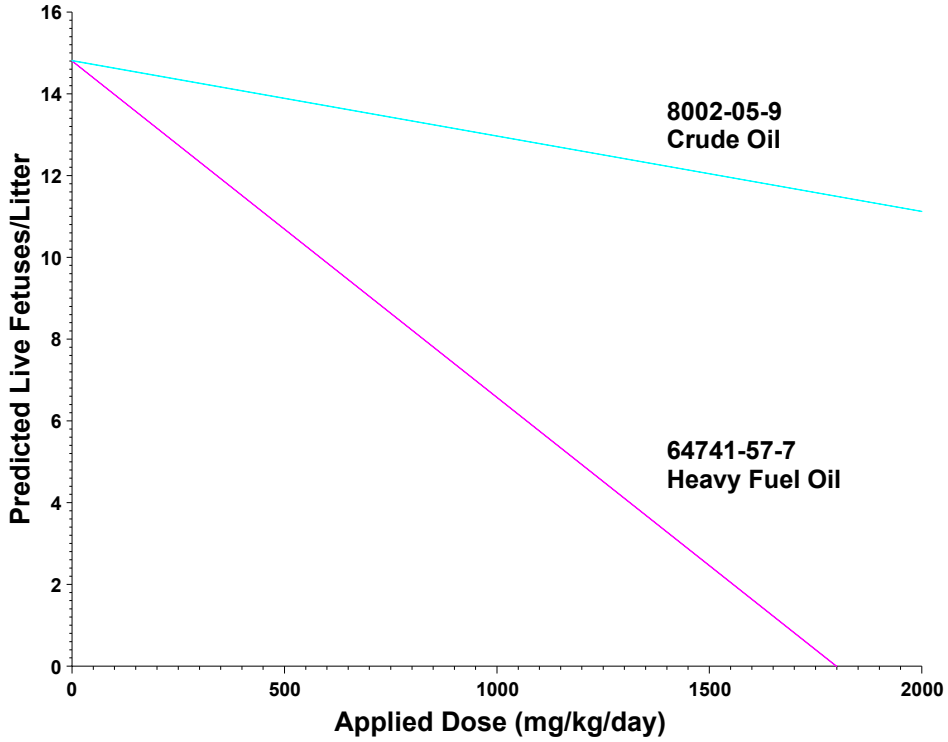
The PAC profiles for the two substances are:

	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. It should be noted that for any sample, the predictions for various endpoints may differ, with some being interpolations while others are extrapolations.

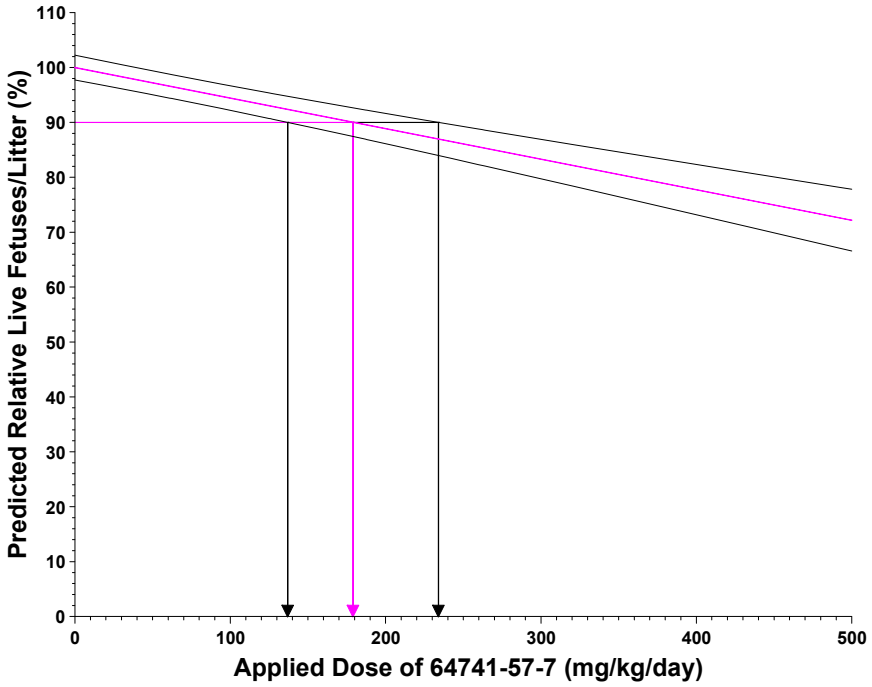
Based on control group data from the 23 prenatal studies used in model development, it can be assumed that the mean numbers of implantations and live fetuses in the control groups are 15.8 and 14.8, respectively. Assuming a dose of 500 mg/kg/day, and substituting the control values and the values of 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 10.7 for CAS 64741-57-7 and 13.9 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



4.2 Use of Models to Predict a Pre-Defined Change (PDx)

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.

With regard to the second use, the TG labelled the predicted dose that causes a defined effect as the "Predicted Dose x (PDx)", where "x" indicates the degree of change from the control value.

As an example, the dose associated with a 10% reduction, (PD_{10}) in the mean number of live fetuses per litter for CAS 64741-57-7 could be predicted. 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 (14.8 live fetuses), see **Figure 6**. The dose associated with a response that is 90% of control value (a 10% reduction) is estimated to be 179 mg/kg/day. Thus, 179 mg/kg/day would be identified as the PD_{10}.

In addition, confidence intervals (CI) can be developed and associated with the PD_x. Using the same example, the PD_{10} and associated 95% CI is 179 (137,234) mg/kg/day (see **Figure 6**). The confidence intervals are based on inverse regression methods also known as calibration methods (Draper and Smith, 1998).

4.3 Selection of the Degree of Change Needed to Demonstrate an Adverse Effect

A challenge arises in trying to identify the degree of change from the control value that would be considered indicative of an adverse effect for each of the endpoints for which models have been developed. In this regard, values must be selected that are both biologically and/or toxicologically meaningful. There are no definitive guidelines to assist in selecting the appropriate values, so the selection has to be based on "professional judgment" supported by any assistance that might be available in the scientific literature.

The TG has spent considerable effort in trying to identify relevant degrees of change that are toxicologically meaningful and could be used in a meaningful way. The values selected by the TG are shown in **Table 9**. A more comprehensive rationale for these selections is given in **Appendix 7**.

Table 9. Degrees of Change Selected by the TG as Toxicologically Meaningful

Study type	Dependent Variable	Degree of change
Repeat-dose toxicity studies	Thymus weight (absolute)	20%
	Platelet count	20%
	Hemoglobin concentration	10%
	Liver weight (relative[a])	ND[b]
Developmental toxicity studies (Prenatal)	Maternal Thymus weight (absolute)[c]	20%
	Fetal Body weight	10%
	Live Fetuses/Litter	15%
	% Resorptions	15%
Developmental Toxicity studies (Postnatal)	Pup Body Weight (PND[d] 0)	10%
	Total Pups/Litter (PND[d] 0)	15%
	Live Pups/Litter (PND[d] 0)	15%

[a] relative to terminal body weight

[b] Not Determined. An increase in liver/body weight ratio was a consistent effect apparently associated with PAC profile. However, as noted in **Section 3.1.1**, the TG decided that a change in absolute or relative liver weight alone i.e., in the absence of pathological changes in the liver is not an adverse effect. For this reason it was not possible to select a degree of change for relative liver weight that was toxicologically meaningful.

[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 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

4.4 Comparison of Predicted and Actual Effects

Two questions need to be addressed when comparing predicted with actual results of toxicity testing:

- how accurately do the predicted dose-response curves fit the observed data, and
- how do the predicted PDx effect levels compare with the endpoint LOAELs/LOELs observed in the actual studies?

4.4.1 Predicted Dose-Response Curves

The examples of the predicted dose-response curves (**Figures 5** & **6**) demonstrate how the models that have been developed may be used for predicting the toxicity of an untested petroleum

substance. However, it is important to understand how the predictions compared with actual results found in the company reports on which the models were developed.

The TG made this comparison 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 9**. 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 10**). The typical percent of observed data that was within the 95% confidence limits was, as expected, at least 95%. As illustrative examples, a plot for each dependent variable is shown in **Figure 7**.

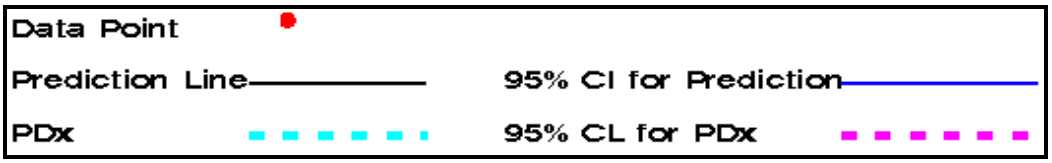
A low percentage of the curves were apparently poor or erroneous predictors of the expected dose response. Closer inspection of these revealed that in almost all cases an effect had not been demonstrated on the endpoint in the study in question. For example **Figure 7.9** (Postnatal pup weight, sample F-195) shows a slight increase in pup weight with dose, and is not consistent with the results of the experimental study in which no effect occurred. It is also inconsistent with the expectation that decreased pup weights would be associated with exposure to substances containing PAC.

Table 10. 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)	96.7%
	Platelet count	98.2%
	Hemoglobin concentration	98.4%
	Liver to body weight ratio	96.0
Developmental toxicity studies (Prenatal)	Maternal thymus weight (absolute)	100.0%
	Fetal body weight	98.5%
	Live fetuses/litter	98.5%
	Percentage resorptions	98.5%
Developmental Toxicity studies (Postnatal)	Pup body weight	98.6%
	Total pups/litter	98.6%
	Live pups/litter	97.2%

Figure 7. Plots for Eleven Final Model Forms Showing Predicted and Actual Responses

The key for the following curves is:



Repeat-dose toxicity studies

Figure 7.1 Absolute Thymus Weight

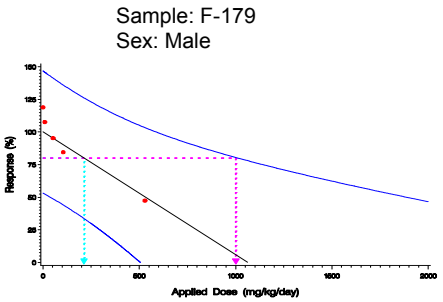


Figure 7.2 Platelet Count

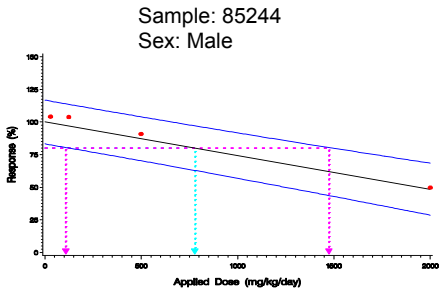


Figure 7.3 Liver/Body Weight Ratio

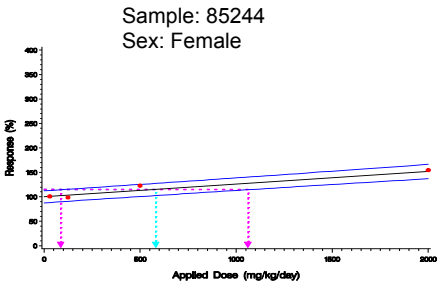
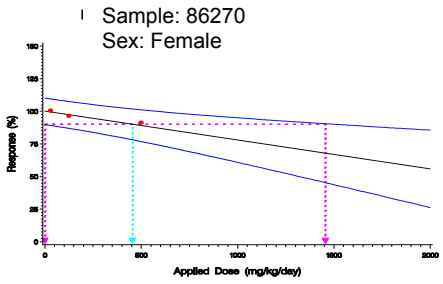


Figure 7.4 Hemoglobin Concentration



Prenatal developmental toxicity studies

Figure 7.5 Maternal Thymus Weight (absolute)

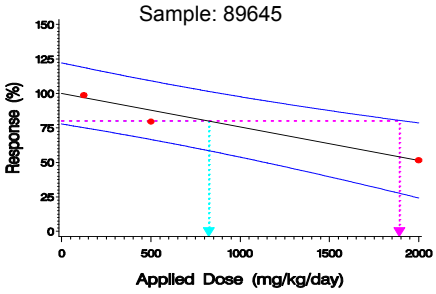


Figure 7.6 Fetal Weight

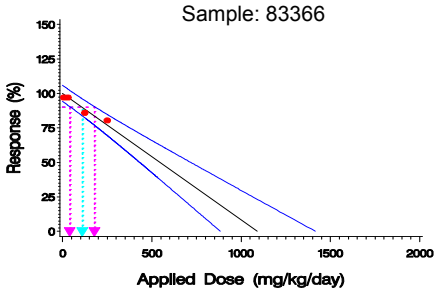


Figure 7.7 Live Fetuses/Litter

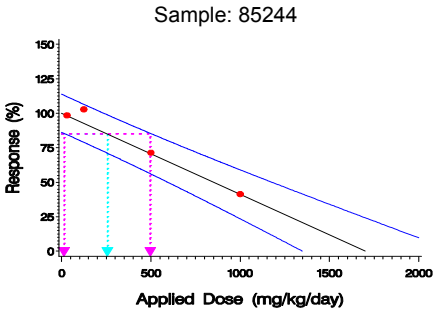
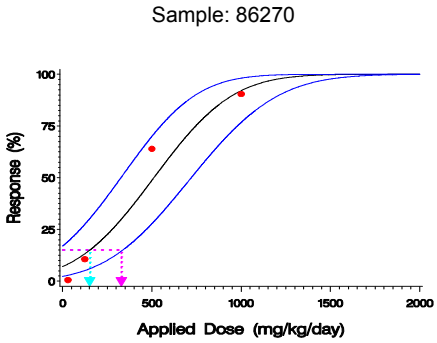


Figure 7.8 Percent Resorptions



Postnatal developmental toxicity studies

Figure 7.9 Pup Weight Day 0

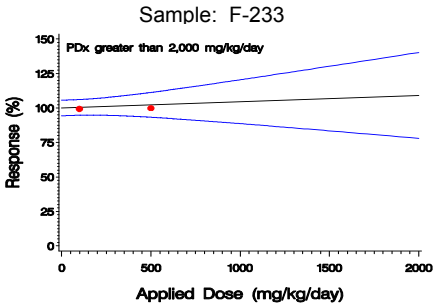


Figure 7.10 Total Litter Size

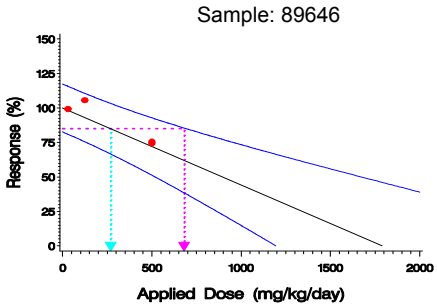
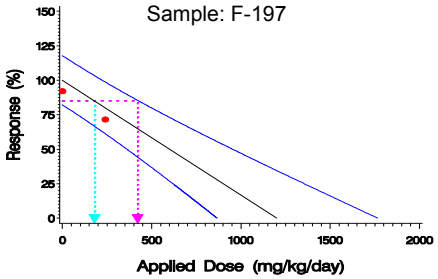


Figure 7.11 No. Live Pups/Litter



4.4.2 PDx Levels

The TG compared the predicted PDx effect levels for each endpoint with the endpoint LOELs observed in the actual studies. It should be noted however, that the determined LOELs are functions of the doses that were selected for a particular study and are not dependent on the degree of change at any particular dose level. Furthermore the degrees of change for each endpoint were selected by the TG only to demonstrate a possible use of the models that were developed. PDx values could be considered an "adverse" effect dose. Nevertheless, bearing these reservations in mind it is interesting to note that the predicted PDx values and observed compared reasonable favourably (**Tables A7-7 – A7-9** in **Appendix 7**).

4.4.3 PDx Levels and the Bench Mark Dose

As another way to check whether the selected PDx values were realistic, the TG considered comparing the PDxs against Bench Mark Doses (BMD) calculated from the actual studies. The purpose of developing a comparison for the PDx was to determine whether use of the suggested PDx values would miss identifying a potential LOAEL. However, the TG concluded that the use of a BMD had several disadvantages. The primary disadvantage of the BMD is that the BMD is based on a selected model and the result is strongly dependent on the model selected (Gephart, et al, 2001). 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. It is likely that differences between the PDx and BMD would be due to inaccuracies or model misspecification in the BMD calculations. For these reasons, it would be impossible to establish whether any differences between the PDx and BMD were artefacts or real. The LOEL is preferred because it is based on the data and is independent of model specification and is therefore the better basis for comparison.

4.5 Potential Limitations/Restrictions on Model Use

The TG believes that the models may be used immediately for petroleum substances falling within the appropriate model domains. 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.

4.5.1 Interpolation and Extrapolation

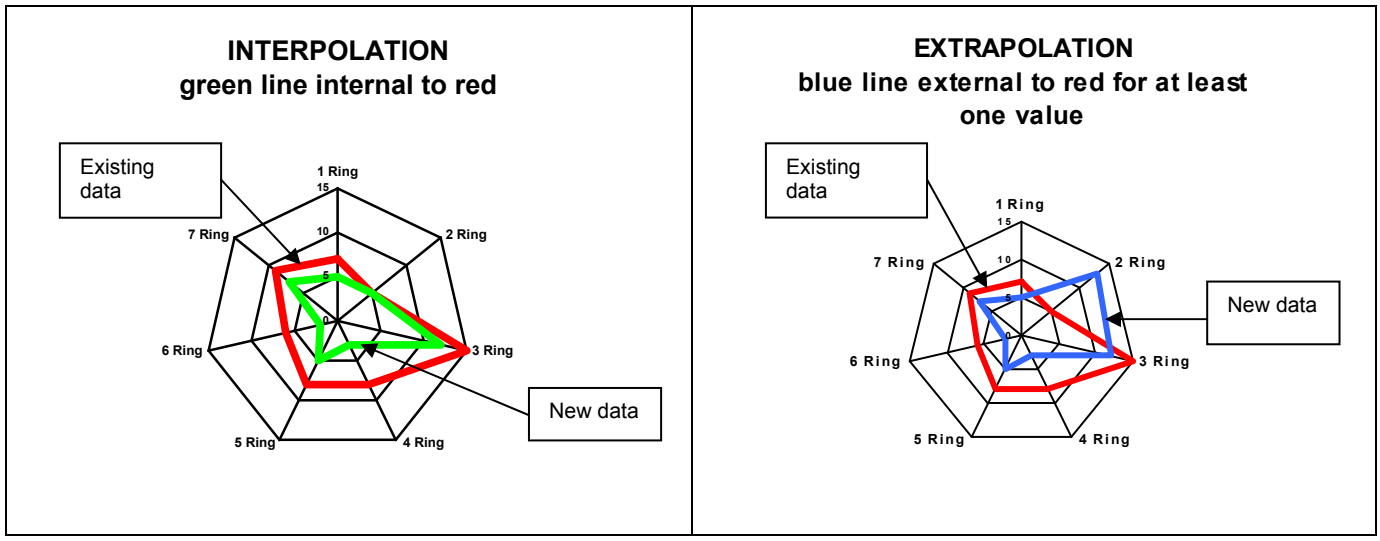
As noted in **Section 3.4.3**, and demonstrated in more detail in **Appendix 6**, testing of the models shows they are all good predictors for data points that are interpolated from the existing data, but are not consistently accurate predictors among data points that are extrapolated. Consequently, the concepts of interpolation and extrapolation need to be defined and understood for the discussion of model testing and prediction, and modelling limitations. In statistical modelling, when predictions or forecasts of a new data value are based on independent variables that are all within the range of data used to develop the model it is called an *interpolated* point. If some, or all, of the independent variables are outside the range of data used to develop the model the new data value is called an *extrapolated* point (see **Appendix 6**).

Specifically for this project, to determine if a new sample (i.e. a sample that was not used to develop the model) is an extrapolated or interpolated point 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 concentrations for all the samples used to develop an endpoint-specific model ("the existing data set"). To determine whether the modeling of the new sample will be an extrapolation or interpolation, a test is done to determine if the weight percent concentrations of each of groups of 1-7 rings are within the range of the concentrations of the base data set and then test that the dose level of the new material is within the dose levels of the materials of the base data materials that bracket the new material:

1. Identify the specific endpoint model and data set used to develop that specific model.
2. Test the new sample against the first sample in the existing data set (i.e. the data set used to develop the specific model) by following these steps:
    a. Determine if the percent weight concentrations for the Aromatic Ring 1 concentration is *less than, equal to, or greater than* the corresponding concentration of the first sample in the existing data set.
    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. Determine if the applied dose of the new sample is less than or equal to the doses of the first sample in the existing data set (because there is always a control group it is not necessary to test for the applied dose being above some minimum value)
    d. If the answer is "less than" or "equal to" for all 7 PAC Ring concentrations and the applied dose, then the new sample is an interpolated point relative to the first sample in the existing data set.
    e. If any answer (dose or concentration) is "greater than", then the new sample is an extrapolated point relative to the first sample in the existing data set.
3. Test the new sample against the remaining samples in the existing data set as in step 2 above.
4. If the new sample is an interpolated point relative to ALL samples in the existing data set then the NEW SAMPLE IS AN INTERPOLATED POINT.
5. If the new sample is an extrapolated point relative to at least one sample in the existing data set, then the NEW SAMPLE IS AN EXTRAPOLATED POINT.

The concepts of interpolation and extrapolation are described in detail in **Appendix 6.** The concept of interpolation and extrapolation for the PAC ring weight percent concentrations is 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

It should be noted that the assessment of whether a prediction is an extrapolation or interpolation are endpoint-specific for any given sample. Therefore, the predictions for various endpoints for a single sample may differ, with some being interpolations while others are extrapolations.

For each test sample used in this project, the TG evaluated whether predictions for the sample for each of the eleven biological endpoints would be an interpolation or extrapolation. The evaluation consisted of the following steps:

1. A PAC profile was developed for each sample that was used in the development of an endpoint specific model. The model for each endpoint was developed using data from a unique set of samples. To verify that a prediction of an effect of an untested substance is an interpolation it is necessary to ensure that the PAC profile of the untested substance falls within the PAC profile of at least one of the samples that was used in the development of the endpoint specific model.
2. The TG developed an Excel® based computer program that can be used to determine if a substance is an extrapolated or interpolated point, based on its percent weight concentrations for the seven ring classes. Because the predictive model for each endpoint is based on a unique data set, hence for a new compound the interpolation/extrapolation status is unique to the endpoint. Thus, for each new substance, the extrapolation/ interpolation determination is made eleven times, once for each of the 11 biological endpoints.

4.5.2 Compositional Data Set

If the PAC concentrations derived from analytical methods other than Method 2 are entered into the existing final models, the accuracy of the resulting predictions is not known because the current model(s) were developed using only Method 2 data. The results using other analytical methods would be different in defining the PAC content results and therefore may alter the final models' predictive abilities. 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.

4.5.3 Route of Exposure

All 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.

4.5.4 Species and Strain

Since all the models in this project were developed using data from toxicity studies involving the Sprague-Dawley rat as the experimental animal, the predictive capacity and applicability of the models to other strains and species are unknown.

4.5.5 Coverage of Data Set

Although the various models were built using experimental data developed on samples 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 TG believes that 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 validated further and provide additional support for their use across all PAC-containing petroleum substances.

4.5.6 Quantification of Degree of Change

The selection of an appropriate PDx for each of the endpoints that were modeled was based on the TG's best professional judgment. The values that were ultimately selected by the TG were 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 PDxs have been selected.

4.6 Implications for Reproductive Toxicity

As noted earlier, only two non-guideline (OECD) reproductive toxicity studies were provided to the TG for review. These two studies were not considered sufficient to assess effects on fertility. However, other endpoints within the developmental toxicity studies (e.g., decreased survival and growth of fetuses/pups) and of repeat-dose studies (e.g., sex organ weights and histology) provide an indication of potential reproductive toxicity. Indeed, under the HPV Challenge Program, U.S. EPA has provided guidance on the requirements for evaluating reproductive toxicity. According to the U.S. EPA guidance, the combination of (1) a 90-day repeat-dose study and (2) a screening or full developmental toxicity study adequately address the requirements for the reproductive toxicity endpoint. Since no significant effects on reproductive organs were observed in the repeat-dose studies evaluated in this project, the effects on postnatal survival and development are likely to be the most sensitive endpoints of reproductive toxicity. Furthermore, limited data suggest that fertility is not a sensitive effect of exposure to benzo(a)pyrene. Therefore, the TG thinks the PDx for developmental toxicity will be a reasonably good predictor of the PDx for reproductive toxicity. A more detailed explanation of the basis for the TG's conclusion can be found in **Appendix 8**.

-4.7 Objective 3 Conclusions

The TG's third objective was to determine if the characterization(s) of any PAC-toxicity relationships could be used to predict the toxicity of similar/related untested petroleum substances. The TG considers that it is reasonable to apply the models developed in this project to untested substances falling within the appropriate model domains. However, as with all models, it is important to understand their limitations and these are outlined in **Section 4.5**.

5. Discussion, Conclusions and Recommendations

The primary purpose of the present investigation was to determine whether there is a relationship between the PAC content of selected classes of petroleum substances and their mammalian toxicity. A relationship between toxicity and PAC content had been asserted or implied in six of the thirteen Petroleum HPV Test Plans originally submitted to U.S. EPA. Therefore, 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 non-acute mammalian endpoints within the HPV Challenge program and included repeat-dose toxicity, reproductive toxicity and developmental toxicity. Genetic toxicity is the subject of Volume 3 in this series of reports.

The TG found that there are indeed associations between some repeat-dose and developmental toxicity endpoints and the PAC content of selected petroleum substances. The TG has also been demonstrated that the toxicity of an untested substance can be predicted based on its PAC content. The TG also demonstrated the utility of these associations within the context of the HPV program.

5.1 Relationship between PAC and Effect

Repeat-dose toxicity

The TG found 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 coefficients of $r \geq 0.88$. 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.

Developmental toxicity

The TG found associations 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 coefficients of $r \geq 0.91$. 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.

Reproductive toxicity

The TG had only two non-guideline (OECD) reproductive toxicity studies available for its review. Although samples in both these studies had high PAC content, no reproductive effects were observed. Given the limited data set, the TG has not attempted to model the relationship between PAC content and fertility. However, the TG reviewed all 45 repeat-dose toxicity studies for effects on reproductive organs as well as the developmental studies for signs of reproductive toxicity. Since no significant effects on reproductive organs were observed in the repeat-dose studies, the TG considered that the effects on postnatal survival and development are likely to be the most sensitive endpoints of reproductive toxicity.

Biological plausibility/consistency
Identification of the specific repeat-dose and developmental toxicity endpoints that were modeled was carried out with considerable care. Identified endpoints were those that were more often affected than any others in the study reports that were provided to the TG. 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 the TG selected for modeling.). Further support that the selected endpoints are reasonable is found in API's robust summaries and test plans for petroleum streams where the spectrum of effects of PAC-containing streams is similar to the endpoints that the TG selected for modeling.

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. The current statistical techniques also make use of all the available data, whereas the previous evaluation relied upon a more limited data set. The TG considered the large number of data points a particular strength of the current evaluation.

5.2 Model Validation

Model limitations and data needs
The TG has identified a number of constraints regarding the extent to which the predictive models can be reliably used.

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.

The TG notes that 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 by the TG in its investigation. Since the models were developed on the basis of information on the PAC profile of petroleum substances, the TG believes that the models will apply to a wide range of petroleum substance that contains 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 TG thinks the models should be used judiciously, to ensure that possible erroneous predictions are avoided.

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

5.3 Use of Models to Satisfy HPV Requirements

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.

Reproductive toxicity
The TG believes that the general lack of toxicologically significant effects on reproductive organs seen in the repeat-dose studies, in combination with the results of the developmental toxicity studies reviewed by the TG, should satisfy the requirements for the reproductive toxicity endpoint under U.S. EPA's High Production Volume (HPV) Challenge Program (see **Appendix 8** for additional details**)**. Furthermore, the TG also believes that the PDx values for the developmental toxicity endpoints are likely to be reasonably good predictors of the PDx values for these indicators of reproductive toxicity.

Selection of substances for testing
The TG considers that within the context of the HPV program, 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 petroleum HPV 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 TG considers that 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.

6. 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) (1993)
Environmental Health Criteria 150: Benzene.
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