Regulatory Toxicology and Pharmacology 67 (2013) S10–S29

Contents lists available at SciVerse ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

The development of statistical models to determine the relationship between aromatic-ring class profile and repeat-dose and developmental toxicities of high-boiling petroleum substances

Mark J. Nicolich [a], Barry J. Simpson [b], F. Jay Murray [c], Randy N. Roth [d], Thomas M. Gray [e,*]

[a] COGIMET, 24 Lakeview Rd., Lambertville, NJ 08530, USA
[b] Simpson Toxicology Consulting, 4, Temple Farm Barns, Singledge Lane, Whitfield, Kent CT15 5AB, UK
[c] Murray & Associates, 5529 Perugia Circle, San Jose, CA 95138, USA
[d] Roth Toxicology Consulting, P.O. Box 6023, Thousand Oaks, CA 91359, USA
[e] American Petroleum Institute, 1220 L Street, N.W., Washington, DC 20005, USA

ARTICLE INFO

Article history:
Available online 13 December 2012

Keywords:
Biological models
Dermal
Developmental toxicity
High-boiling petroleum substances
HPV Chemical Challenge Program
Mixtures toxicity
Polycyclic aromatic compounds
QCAR modeling
Rat
Repeat-dose toxicity

ABSTRACT

The repeat-dose and developmental toxicities of certain petroleum refinery streams are related to their polycyclic aromatic compound (PAC) content (Feuston et al., 1994). Building on this foundation, and working within the context of the US EPA High Production Volume (HPV) Chemical Challenge Program, we:

(1) characterized relationships between PAC content and repeat-dose and developmental toxicities of high boiling petroleum substances (HBPS), and

(2) developed statistical models that can be used to predict critical effects of similar untested substances.

Data from 39 dermal toxicity studies of HBPS were used to develop statistical models to predict the dose–response relationships between the weight percent concentration of each of their 1–7 aromatic ring classes and 4 repeat-dose and 3 developmental endpoints (absolute thymus weight, hemoglobin count, platelet count, liver to body weight, live fetus count, fetal weight, and percent resorptions). The correlations between the observed and model-predicted values are >0.90. The predictive ability of the models was tested via a series of evaluation or corroboration methods.

As is shown in the paper, using only compositional data of untested HBPS, the models can be used to predict the effect at a given dose or the dose that causes an effect of a stipulated magnitude.

1. Introduction

A project was initiated to investigate the potential relationship between the polycyclic aromatic compound (PAC) content and the acute, repeat-dose, developmental, reproductive and genetic toxicities of HBPS. Two objectives of the project were to:

(1) identify and characterize relationships between PAC content and screening information data set (SIDS) mammalian toxicity endpoints, and
(2) determine if any identified relationships could be employed to predict the toxicity of untested high-boiling petroleum substances, i.e., substances whose final boiling point is ⩾ approximately 650 °F (343 °C).

This paper describes the data definition, selection, and collection procedures; the statistical model development, the final selected model forms, and the model testing and corroboration for the repeat-dose and developmental toxicity models.[1] Other papers in this series will describe the relationship between aromatic content and repeat-dose endpoints, the relationship between aromatic content and developmental toxicity results, the results of bacterial mutagenicity modeling and chromosomal aberration testing, the application of the repeat-dose and developmental toxicity model predictions to characterize samples for which toxicity data are not available, and how the model predicted values compare to traditional indices of toxicity. The available data indicated the sub-

* Corresponding author. Fax: +1 202 682 8270.
E-mail addresses: mark.nicolich@gmail.com (M.J. Nicolich), barryjsimpson@btinternet.com (B.J. Simpson), jmurray2@sbcglobal.net (F. Jay Murray), rroth@roth-tox.com (R.N. Roth), grayt@api.org (T.M. Gray).

[1] In this paper, and other papers in this Supplement, we have chosen to not use the term *validation* to refer to the process of demonstrating that the model predictions are similar to real-world observations. As pointed out by Oreskes and colleagues (1994), the intrinsic meaning of a *validated model* is that the model has been shown to be true or an accurate representation of reality when it is really meant to imply that there has been a demonstration of consistency between the model and reality. Based on the recommendation of Council for Regulatory Environmental Modeling (US EPA, 2009), we have chosen to use the word *corroborate* or *evaluate* rather than *validate*.

0273-2300/$ - see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.yrtph.2012.11.015

stances that were studied were associated with low acute toxicity; therefore we could not develop statistical acute toxicity models.

2. Background

The United States Environmental Protection Agency (US EPA), in partnership with industry and environmental groups, announced a voluntary chemical data collection effort called the High Production Volume (HPV) Challenge Program. In this program US manufacturers and importers of HPV chemicals were invited to voluntarily "sponsor" chemicals. Sponsorship involved the commitment to develop summaries of existing health and environmental effects data for HPV chemicals (those chemicals that are produced in or imported into United States in aggregate quantities of one million pounds or more per year) and to conduct new testing to fill any data gaps needed for the objectives of the program (US EPA, 2000).

Approximately 400 petroleum substances, defined by their Chemical Abstract Service Registry Number (CAS RN), were sponsored through the American Petroleum Institute (API) by companies belonging to the API managed Petroleum HPV Testing Group (PHPVTG), and, approximately 110 are potentially impacted by the analysis described in this and the subsequent reports.

The authors of the summaries of health effects and test plans for the HPV categories of aromatic extracts, crude oil, gas oils, heavy fuel oils, lubricating oil basestocks, waxes and related materials, and petroleum waste streams either stated or implied that the repeat-dose, genetic, developmental and reproductive toxicities of the category members were associated with PAC content and that the PAC content of these substances could be used to predict the toxicity of similar, untested petroleum substances.

The basis for these claims was a publication by Feuston et al. (1994) that examined the correlation between the weight percentage of several chemical classes of compounds in thirteen refinery streams and the biological endpoint effects in rats that were dermally exposed in repeat-dose and developmental toxicity studies. Feuston et al. (1994) showed that for these streams relationships existed between the rank of the lowest observed effect level (LOEL) in the toxicity studies and the rank concentration of various classes of refinery stream components measured using two different analytical methods. Significant rank correlations were found between the endpoints and the individual and combined PAC-ring classes containing three or more rings, but no significant rank correlations were found between the biological endpoints and the concentrations of non-aromatic, 1-ring class, 2-ring class, and 1- and 2-ring classes (with the exception of skin irritation). The relationships were lost when the correlations were based on the observed values rather than the ranks; an indication that the relationships are complex. Because the relationships were based on combined aromatic-ring classes (ARCs) they could not take advantage of the differential information from individual ARC concentration measures, which we later found to be important for predictions. Also, because the relationships were based on ranks, the responses for untested materials could not be predicted. Nonetheless, the Feuston et al. (1994) study was an important impetus that led to this expanded evaluation of laboratory studies that had both compositional data (aromatic content) and toxicity data on the same substance and to the development of the models presented in this paper.

The present evaluation of the relationship between PAC content and repeat-dose, developmental and reproductive toxicity has been completed and the resulting report has undergone a TERA peer consultation (Patterson et al., 2013; Simpson et al., 2007, 2008). This paper describes the data definition, selection, and collection, the thought processes applied in the model development, the final selected model forms, and the steps applied in model testing and corroboration for the repeat-dose and developmental toxicity models.

3. Materials and methods

3.1. Terminology

The following are the definitions of terms used in this publication:

Polycyclic aromatic hydrocarbons (*PAHs*): compounds of two or more fused aromatic rings consisting of only carbon and hydrogen atoms.

Polycyclic aromatic compound (*PAC*): a comprehensive term that includes PAHs and molecules in which one or more atoms of nitrogen, oxygen, or sulfur replace one of the carbon atoms in the ring system.

Aromatic-ring class (*ARC*) *profile*: the weight percent of each class of the DMSO-soluble 1–7 and larger aromatic-ring compounds present in a petroleum substance as determined by the Method II chemical characterization procedure (Blackburn et al., 1996; Gray et al., 2013; Roy et al., 1985, 1988, 1994), e.g. the ARC 3 value would be the weight percent of the DMSO-soluble 3-ring aromatic compounds within the petroleum substance.

3.2. Data sources

Four potential sources of information were identified for data that could be used in the analyses:

(1) a previously published report by Feuston et al. (1994),
(2) other published literature on the toxicity of individual PAHs and PAC containing materials,
(3) company laboratory reports of toxicity studies of petroleum substances that had accompanying PAC compositional data of the test sample, including those of the studies conducted on the thirteen samples in the Feuston et al. (1994) study, and
(4) laboratory reports of toxicity and analytical studies sponsored by API.

Of the four sources of information, only the company and API laboratory reports provided a sufficient number of studies and included sufficient detailed compositional data of the aromatic-ring content of the test samples to be of use in this evaluation. The materials that had been tested in the submitted studies covered a range of petroleum substances most of which were high-boiling. We limited the samples to HBPS with final boiling points ⩾ approximately 650 °F (343 °C). These substances contain fused aromatic-ring compounds with ⩾3 rings, which are the PAC compounds of interest for repeat-dose toxicity (Feuston et al., 1994). Substances with lower final boiling points are not expected to contain PAC compounds with ⩾3 aromatic rings.

The laboratory reports consisted of:

(1) 47 repeat-dose toxicity studies (nineteen 28-day and twenty-eight 90-day),
(2) 68 developmental toxicity studies,
(3) two reproductive toxicity studies, each with only a single sex dosed,
(4) one limited one-generation reproductive toxicity study,
(5) one exploratory dose range-finding study in non-pregnant female rats, and
(6) 157 analytical reports of compositional data on the tested substances.

All of the unpublished company laboratory reports (toxicity and analytical) were judged to be either "reliable without restrictions" or "reliable with restrictions," (reliability scores of 1 or 2 Klimisch et al., 1997).

Forty-six of the 47 repeat-dose toxicity studies had been carried out in Sprague Dawley rats exposed via the dermal route, the exception being a 10-week study in mice on sample 86001. This mouse study was not used in the current evaluation, but has been published in Feuston et al. (1997). We restricted the samples to those tested by the dermal route because, except for certain substances in the lubricating oil basestock and waxes categories, dermal contact is considered to be the most likely route of human exposure for HBPS. Note that one 13-week study (sample 86187) included two orally exposed groups (males) and four dermally exposed groups (males and females); only data from the dermally exposed animals were used in this evaluation. All of the studies used in our evaluation were conducted in Sprague Dawley rats, included one concurrent control group, and most included three dosed groups.

Of the 68 developmental toxicity studies, 64 involved dermal administration on Sprague Dawley rats. Among these studies, 29 were of traditional design in which the pregnant rats were exposed during gestation and the uterine contents were examined during a cesarean section just prior to birth. We termed these studies "Type I" developmental toxicity studies. The remaining 35 developmental toxicity studies were of a design in which pregnant rats were exposed during gestation, litters were allowed to be delivered naturally, and observations were made on the day of birth through postnatal day (PND) 4. These studies were termed "Type II" developmental toxicity studies.

Analytical data on the test substances from the 46 dermal repeat-dose and the 64 dermal developmental toxicity studies were captured from the corresponding analytical reports.

The toxicology data that were extracted were the standard measurements collected in guideline studies that were both biologically meaningful and could be used to define a point of departure, such as a Lowest Observed Effect Level (LOEL) in a dose response curve. The endpoints on which data were extracted are similar to those that are characterized during guideline studies for repeated dose dermal toxicity (OECD, 1981a), subchronic dermal toxicity (OECD, 1981b), prenatal developmental toxicity (OECD, 2001) or a combined screen for systemic and reproductive toxicity (OECD, 1996). For additional details on the selection of studies for use in the repeat-dose and developmental models building process, see Roth et al. (2013) and Murray et al. (2013), respectively.

There were 157 analytical reports that provided information on the PAC content of the test samples. A number of different analytical techniques were used in these reports, and as will be discussed in more detail below, the technique that was identified as being the most useful was "Method II." This method is described in Gray et al. (2013) as a method in which the extractable ring class aromatic components are extracted with DMSO. The weight percent of each of the seven ring classes of aromatic compounds present in the DMSO extract are separated by ring number with gas chromatography to give the aromatic-ring class profile (ARC profile).

3.3. Data selection

Several overlapping sets of criteria were applied to the toxicity studies used in this study to ensure the studies were of uniform quality, had similar study protocols, and were suitable for statistical modeling. Data were included in the final modeling effort if they met the following criteria[2]:

(1) The test material was a HBPS.
(2) Test substance compositional data was developed using Method II.
(3) Biological data on the test substance was from a repeat-dose or a developmental dermal toxicity study consistent with current international study guidelines.[3]
(4) Treatment groups in the repeat-dose studies had less than 50% mortality.[4]
(5) Daily dosing in developmental toxicity studies was for the total gestation period (i.e., gestation days [GD] 0-19).[5]
(6) Developmental toxicity data were from "remote" not "proximal" control groups.[6]
(7) Developmental toxicity data were from groups where there were four or more dams with viable fetuses or litters.[7]

Table 1 summarizes the availability of studies and the final number used in the modeling effort. For additional details on the selection of studies for use in the repeat-dose and developmental models building process, see Roth et al. (2013) and Murray et al. (2013), respectively. One sample from the Feuston et al. (1994) study, light catalytically cracked naphtha, was not included in this study because it is not a HBPS.

3.4. Data integration

The initial evaluation of which biological endpoints should be identified for statistical model building considered all the data captured from the repeat-dose and developmental toxicity studies. The three criteria used to select the endpoints were:

(1) most often reported as statistically significantly affected, and therefore most likely due to PAC exposure,
(2) affected most often at the study's LOEL (i.e., those effects that would be traditionally considered for choice as the chemical's critical effect) since they are most sensitive to PAC exposure, and
(3) consistent with reported effects of PACs or PAC-containing petroleum products.

[2] See Simpson et al. (2007, 2008) for descriptions of earlier versions of the models and the data used to build them.

[3] Repeat dose (OECD 410 and 411) and developmental (OECD 414) toxicity guidelines.

[4] The high mortality criterion was applied to remove groups where the maximum tolerated dose (MTD) had likely been surpassed and the surviving animals might not be representative of the general population or other animals in the study.

[5] The dosing period was not the same in all the developmental toxicity studies from which data were extracted, and in some cases not the same among dose groups within a study. To ensure the modeling results were comparable, only data from studies and dose groups that included daily dosing on GD 0-19, as a minimum, were used. This included studies with dosing on GD 0-19, or GD 0-20 or from pre-mating day 7 to GD 20. If a dose group was administered the test material every other day, the dose group was not included.

[6] Several of the developmental toxicity studies had two control groups: (1) a remote control group and (2) a proximal control group. In all studies with two control groups, only data from the remote control group was used for modeling. The remote control group was housed in a different animal room than the exposed animals in order to avoid inhalation exposure to the test material. The proximal control group, which was housed in the same animal room as the exposed animals, was excluded from modeling since this control group may have had some inadvertent or indirect inhalation exposure to the test material, even though the exposure was through the dermal route.

[7] Data from dose groups with three or fewer dams with viable fetuses in Type I developmental toxicity studies were excluded from the modeling and statistical analyses. Of the Type I developmental toxicity studies available in the modeling exercise, the number of mated females per group typically ranged from 10 to 25. Because the modeling weighted each data point (dose group) equally, it was important to exclude data that were based on a small group size. The variability associated with small group sizes is typically greater than that based on larger group sizes. Consequently, there is less confidence and greater uncertainty associated with these data points.

Table 1
Availability of repeat-dose and developmental toxicity studies.

	Repeat-dose			Developmental			Total
	28-day Studies	90-day Studies	Total	Type I Studies[a]	Type II Studies[b]	Total	
Studies reviewed by Petroleum HPV Testing Group	19	28	47	33	35	68	115
Studies from which data were extracted	19	27	46	29	35	64	110
Studies used for preliminary modeling	19	26	45	23	34	57	102
Studies used for final modeling	1	17	18	21	0	21	39

[a] Type I studies – pregnant females exposed during gestation, caesarean section on day 20 of gestation.
[b] Type II studies – pregnant females exposed during gestation, dams allowed to deliver and pups monitored through day 4 of lactation.

After completing the preliminary quantitative assessment, four repeat-dose and three developmental toxicity endpoints were selected (see Table 2) for final statistical characterization based on the following three criteria:

(1) whether an endpoint would be considered an adverse effect or predictive of an adverse effect,
(2) whether similar endpoints had also been characterized, thus making the analysis redundant (e.g. among hematocrit, hemoglobin concentration, and erythrocyte count, only hemoglobin concentration was identified for modeling), and
(3) the strength of the relationship of the preliminary statistical dose–response characterization.

The first two criteria were given more weight than the third.

The data sets for each of the seven endpoints averaged approximately 71 data points and 19 studies per endpoint. Sets of compositional (ARC profile) and effects data were developed for each of the seven endpoints. In addition to the seven ‘critical’ endpoints, a data set for maternal absolute thymus weight was developed and as a source for model testing (see Section 5).

Maternal toxicity endpoints were not selected for statistical analysis because the goal of the project was to determine whether developmental toxicity, not maternal toxicity, could be predicted based on PAC content. As a practical matter, among the available developmental toxicity studies maternal toxicity was not ideal for modeling. For example, it was not possible to get consistent data for maternal body weight, body weight gain, and food consumption because they were measured on different days of gestation in the different studies. Preliminary evaluations suggested that developmental toxicity was strongly associated with maternal toxicity (e.g. decreased maternal body weight, weight gain, and/or food consumption, skin irritation), and there was no strong evidence of developmental toxicity in the absence of maternal toxicity among the studies modeled. For purposes of this project, because the objective of the project was to predict developmental, not maternal, toxicity, it did not matter whether maternal toxicity played a role in producing developmental toxicity.

Table 2
Biological endpoints selected for final statistical characterization and number of dependent variable data points used in modeling.

Study type	Dependent variable	n[c]
Repeat–dose toxicity	Thymus weight (absolute)	84
	Platelet count	85
	Hemoglobin concentration	98
	Liver weight (relative[a])	90
Developmental Type 1 toxicity	Fetal body weight	59
	Live fetuses/litter	59
	Percent resorptions	59
Corroboration[b]	Maternal thymus weight (absolute)	29

[a] Relative to terminal body weight.
[b] Maternal thymus weights were utilized as an alternate data source when the models were tested (see Section 5).
[c] Number of dependent variable data points used in modeling.

4. Statistical model development

4.1. Modeling methods

A statistical model of the dose response curve was developed for each of the seven biological endpoints selected for final modeling and for the maternal thymus weight endpoint selected for model testing (Table 2). The models for each endpoint were developed independently, using an iterative process. Models were developed using general regression analysis methods with the biological endpoint (e.g. fetal body weight) as the dependent, or predicted variable, and relevant toxicological study design variables (e.g. control group response, litter size, sex) and the test material variables (e.g. the weight percent of each of the 1–7-ring ARC, the “ARC profile”) as the independent, or predicting, variables.

4.2. Choice of dependent variables

The dependent variable and number of dose groups used to develop the model for a specific endpoint are shown in Table 2. For each of the endpoints selected for final modeling, the dependent variables were the responses of a dosed group (dose > 0), while the control group response was used as an independent variable (covariate).

For the repeat-dose studies, the dose group response was the mean response of all the animals in the dose group in a specific study. For the developmental toxicity studies, the dose group response was the unweighted mean of the means of all the litters in a dose group in a specific study.

The modeled dependent variable was the observed response rather than either the ratio of the dose group response to the control group response, or the ‘percent response relative to control’. The use of a covariate (the control group response as an independent variable) allowed more flexible modeling of the response and, in most cases, resulted in a more stable estimate. If the models were developed with percent response relative to control as the dependent variable, the response would be the ratio of two random variables. The ratio of dose group to control responses can vary widely, especially when the control group value is likely to be small. For example, when measuring the number of resorptions a seemingly small change of the numerical value in the denominator can result in a large change in the ratio (i.e., if the number of control group resorptions decreases from 2 to 1 in a litter the percent of resorptions relative to control will double). All models were developed using both the covariate method and, as an alternative, the percent response relative to control. The covariate models were more stable and had regression fit diagnostics at least as good as the percent response relative to control models. If needed, the model-predicted responses from the covariate models can be converted to percent response relative to control predictions by divid-

Table 3
ARC profiles of 2 petroleum samples with similar total PAC content.

CAS RN	Sample No.	ARC-1	ARC-2	ARC-3	ARC-4	ARC-5	ARC-6	ARC-7[a]	Total
64741-59-9	8281	2.0	29.5	14.7	0.0	0.5	0.5	0.0	47.2
64741-62-4	86001	0.0	2.6	25.7	19.3	6.4	3.2	0.6	57.8

[a] The ARC 7 value is the weight percent of the 7 and >7 ring aromatic compounds within the petroleum substance as determined by the Method II chemical characterization procedure.

ing the predicted value by the control group response; we have not presented these converted models because we do not want to add more models to an already complex discussion. Note that the ratio discussed here is the ratio of a dose group response to a control group response, and is different from the ratio of liver weight to body weight (relative liver weight). The latter is the ratio of two different measures from the same animal while the former is not. The argument against the ratio of responses would apply to the ratio of dose group relative liver weights to control relative liver weights.

4.3. Choice of independent variables

4.3.1. Analytical variables (PAC measures)

The number of HBPS is large, with each containing at least thousands of structurally-related individual substances (Altgelt and Boduszynski, 1994; Gray et al., 2013; Potter and Simmons, 1998; Speight, 2007), including a wide variety of polycyclic aromatic compounds (PACs). The specific chemical composition of each sample of these HBPS is affected by both the source of the crude oil and the processing conditions used to create the substance (Speight, 2007). In addition, the composition of HBPS can vary substantially, even among substances with the same CAS RN (Gray et al., 2013). Fortunately, as will be shown, the models accurately predict the developmental and repeat-dose toxicity of these substances based on ARC profile and independent of the CAS RN.

The PAC content of the test samples in the individual studies had been determined using a variety of analytical techniques. As described in more detail elsewhere (Patterson et al., 2013; Simpson et al., 2007, 2008), there were four analytical methods with sufficiently large sets of samples to provide a basis for comparison. The four methods determined either the concentrations of aromatic compounds of ring classes 1–5 (noted as Method I), or extractable ring classes aromatic compounds of ring classes 1–7 ring and larger (Method II), or S-PACs from the Method I method, or N-PAC concentrations. The S-PACs are unalkylated and alkylated PACs in which the heteroatom is sulfur. These include the thiophenes and their benzologues (with additional aromatic rings fused to the thiophene structure). Dibenzothiophene is an example of an S-PAC. N-PACs are unalkylated and alkylated PACs in which the heteroatom is nitrogen. These are the pyrrolic (five-membered ring aromatic) and pyridinic (six-membered ring aromatic) structures and their benzologues. Benz[*a*]acridine is an example of a basic N-PAC.

Preliminary modeling with these four PAC measures clearly indicated that the models developed using the Method II data generally had better fit characteristics than those using the Method I measures and in many cases much better fit than the other two PAC measures. Furthermore, there were more toxicity studies available with corresponding Method II data on the test sample. Details can be found in the reports previously referred to (Patterson et al., 2013; Simpson et al., 2007, 2008). Because of the better model fit and the larger sets of available data the Method II data set was selected for use in the final model building.

The extractable concentrations of each of the 7 ring classes (the ARC profile) were used in the models. Using the Method II data we tested models using a reduced set of rings, or weighted averages of the 7 ring classes based on factor analyses, or other variable reduction techniques. None of these alternative models performed better than those based on the 7 individual ring-class concentrations, so the extractable concentrations of each of the 7 ring classes (the ARC profile) were used in the final models.

4.3.2. Individual ARC terms

It is not sufficient to consider the total percent weight of the 1–7 and larger aromatic-ring compounds because the total percent weight by itself does not adequately predict a description of the dose response of the petroleum substance, whereas consideration of the individual ring class concentrations (the ARC profile) predicts a sufficiently accurate dose response curve. For example, consider the ARC profile of two petroleum samples (Table 3) that have similar total PAC content (47.2 and 57.8) but different ARC profiles. The sample number refers to an internal sample number assigned to the data sets used in the development of the models.

The ratios of observed mean fetal body weight to the control mean fetal body weight for the two samples from Table 3 are plotted in Fig. 1. The fact that both materials have similar total PAC contents might lead to the expectation that they would have similar biological activity. However, there is a large difference in the observed biological responses of the samples. In contrast, the final statistical model predictions for these two samples closely agree with the observed data (Table 4) indicating the usefulness of the models when based on the ARC profiles. We used the ratio of the treated to control group body weights to simplify the interpretation of this point. The modeling efforts use only the treated group weights as the dependent variable.

4.3.3. Toxicity study design variables

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

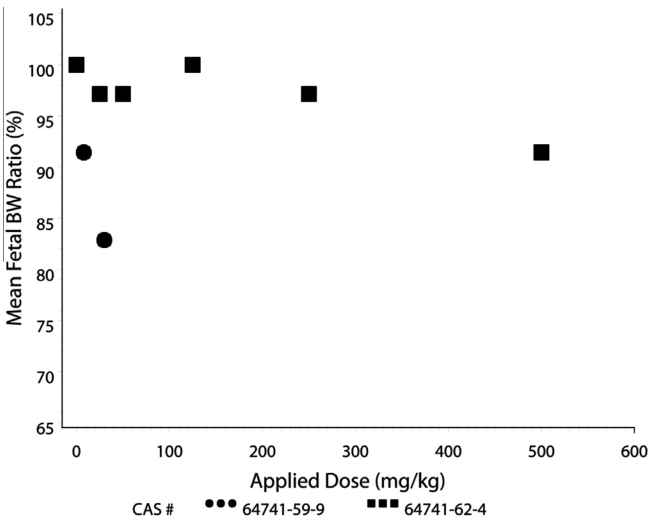

Fig. 1. Observed mean fetal body weight ratio vs. applied dose for two substances with total PAC extract weights of 47 and 58 percent.

Table 4
Observed and predicted mean fetal body weight (FBW) ratios based on the ARC profiles for two petroleum samples with similar total PAC content.

CAS RN	Sample No.	Dose or FBW ratio						
64741-59-9	8281	Dose (mg/kg$_{bw}$/day)	0	25	50	125	250	500
		Observed FBW Ratio	100	97	97	100	97	91
		Predicted FBW Ratio	100	100	99	98	96	92
64741-62-4	86001	Dose (mg/kg$_{bw}$/day)	0	8	30			
		Observed FBW Ratio	100	91	83			
		Predicted FBW Ratio	100	94	79			

(1) dose level normalized to milligrams of applied substance per kilogram of animal body weight per day (mg/kg$_{bw}$/day),
(2) duration of dosing, and,
(3) sex of the treated animals.

For the Type I developmental toxicity studies, the independent variables were selected from:

(1) dose level normalized to milligrams of applied substance per kilogram of animal body weight per day (mg/kg$_{bw}$/day),
(2) number of implantation sites,
(3) number of animals, or pregnant dams, or litters per dose group, and
(4) body weight.

Each model also included the control group response as an independent term, or covariate, in the model. Not all variables were eligible, available, or appropriate for all models; however, terms for dose level and control group response were always included in the model building process. All responses were means calculated in a similar manner to that described above.

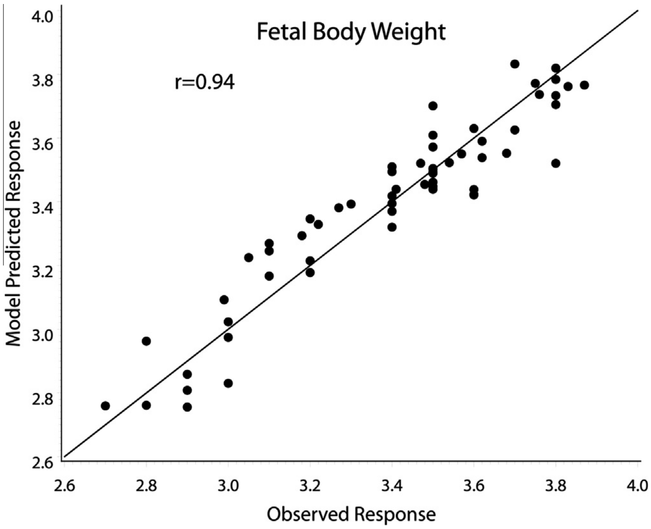

Fig. 2. Observed vs. model predicted data points for the fetal body weight model.

4.4. Models

As previously noted, the model for each of the endpoints was developed independently. The basic model form was a general linear regression model with the dose group response as the dependent variable, the control group response as an independent variable (covariate), and a selection of independent variables as described above.

The model building process was, by definition, an iterative process in which model forms were postulated and tested with various diagnostics. Based on the results of the diagnostics and an understanding of biology and toxicology, a model was then altered by adding or removing terms and/or transforming terms, or in some cases trying nonlinear model forms when these seemed justified.

The transformations included the standard set of logarithm, exponent, trigonometric, power, and probit transformations. The diagnostics included residual plots, and a statistical evaluation of the magnitude and effect of influence points (Belsely et al., 1980). The influence points are data points that have a statistically large effect on the estimated coefficients and statistical significance of the coefficients. The residuals were tested for a normal distribution at the 0.01 significance level by the Shapiro–Wilk test (Shapiro and Wilk, 1965). Plots of the observed and predicted values from a model were developed to evaluate the adequacy of the model and to look for outliers and other possible anomalies, see Fig. 2 for an example of such a plot.

The final comparison among competing models for an individual endpoint was based on the overall model multiple correlation coefficient (*r*) and the error mean square (EMS). These measures were selected because, among their other characteristics, the *r* value is a measure of the closeness of the observed and model predicted values, while the EMS is related to the width of the confidence interval of the predicted value. During the model building process, we did not adhere strictly to the optimization of the correlation and standard error, but considered the overall reasonableness of the model, concentrating more on the fit of the model near the critical region where an increase in dose was associated with a biologically important change in response (rather than near the no effect region or a region of extreme response), but not allowing a few outliers to drive the form of the model. In general, the goal was to develop a model that was both a good descriptor and one in which greater confidence could be placed in its predictions.

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 did not improve the fit of the model to the data. The testing also indicated that the most stable models were based on predicting the dose group response directly (not as a ratio to the control group), with the control group response as an independent variable. The predicted ratio could be developed from the predicted direct dose group response by dividing by the control group response.

During model evaluation, as described above, models were developed based on both linear regression using ordinary least squares (OLS) methods (Draper and Smith, 1998) and mixed-effects models (Pinheiro and Bates, 2002) using maximum likelihood (ML) methods. The OLS methods assume all observations are independent. However, in our data, the assumption of independence may not be achieved because there are usually from two to six dose group data points from a particular study (and the toxicological studies themselves may have had some commonality). The assumption of independence is important for assessing significance levels of terms in the model, but has little effect on the esti-

Table 5
Comparison of model fitting characteristics for OLS and mixed model analyses.

Study type	Dependent variable	*n* Studies	*n* Data points	OLS		Mixed effects model	
				r	se	*r*	se
Repeat–dose toxicity	Thymus weight (absolute)	16	84	0.91	0.03	0.94	0.03
	Platelet count	16	85	0.91	0.12	0.95	0.09
	Hemoglobin concentration	18	98	0.94	0.60	0.94	0.56
	Liver weight (relative[a])	17	90	0.94	0.19	0.97	0.15
Developmental Type 1 toxicity	Fetal body weight	21	61	0.94	0.11	0.98	0.07
	Live fetuses/litter	21	60	0.98	0.90	0.99	0.80
	Percent resorptions	21	60	0.99	0.24	0.99	0.23
Corroboration	Maternal thymus weight (absolute)[b]	10	29	0.92	0.03	0.98	0.02

[a] Relative to terminal body weight.
[b] Maternal thymus weights were utilized as an alternate data source when the models were tested (see Section 5).

mated coefficients. The mixed-effects models account for the relationships of dose groups within a study, and are theoretically preferable in the current situation.[8]

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

We assessed the efficacy of the OLS and mixed effect models by considering the model correlation (*r*) and residual standard error (se). While it is known that the mixed effect models are not optimized for the correlation and minimum standard error, as are the OLS methods, they do provide a reasonable method of comparison. Table 5 shows the correlation (*r*) and residual standard error (se) for the optimum models from the two estimation methods.

The equivalence of the fits of the models from the two methods can be seen in the similarity of the correlations (*r*), while the slightly smaller errors of predictions with the mixed models can be seen in the smaller "se" values. Given the small differences between models from the two methods, the simplicity of the OLS methods is preferred over the mixed-effects models.

The individual data points used in the models are means of individual dose groups from the repeat-dose studies and the means of mean litter responses of individual dose groups from developmental toxicity studies. Within a data set used for a specific model the number of studies averaged into a data point may vary. This variation in underlying sample size can engender a different variance for different data points. Unfortunately, this violates one of the assumptions of OLS and ML model building. However, because of the requirements specified in the initial data collection, the variation amongst studies is small and the resulting differences in variation will have little effect on the final variance of the estimator and the significance level of individual terms in the model. Since the goal is to develop a model form and not to test hypotheses or develop strict confidence limits on predictions, neither of these potential problems presents serious difficulties.

The initial model building included a categorical (nominal) term that described the Petroleum HPV Testing Group's (PHPVTG) category of the test sample (e.g. aromatic extracts, crude oil, etc.). This term was statistically significant in almost all the models. Because this measure is not a physical property of the sample but a descriptor and the goal was to develop terms that were measurable properties, logistic regression and discriminant function techniques were used to develop an alternative term that is similar to the category term but based on the chemical composition properties. The analyses indicated that a collection of terms involving the individual ARC concentrations and the interaction of ARC ring 4 with ARC ring 5 was a good predictor of the HPV category. Therefore, the interaction term was considered when building the models.

In summary, the models were developed independently for each endpoint considering the biology, toxicology, and statistical aspects of the available data. The models were developed to be as simple as possible, but still adequately describing the data. A model that fit the data well in the critical region, that is the region where the response changes from normal to abnormal, was preferred to one that fit well at the extremes. After all the models were independently developed, some alteration was made to have the models similar in their algebraic form while not sacrificing the integrity of the individual models. The amount of alteration was fairly small, which is an indication of the statistical consistency of the modeling process, but is not meant to indicate anything about the underlying biological mechanism. The terms for the individual ARC terms were kept for all models to avoid the problem of fitting each model to a specific data set and not have it generalizable to new data, and to minimize the tendency to inspect individual ARC terms for hints of the biological mechanism.

The models met the objective of characterizing the relationship between PAC content and SIDS endpoints as seen in the correlation between the observed and predicted data (a mean *r* of 0.94 and minimum *r* of 0.91).

The correlation and standard error (*r* and se) values in Table 5 are for the final OLS models that are based on the observed response, not the ratio of the response of the dosed group to control group.

The magnitudes of the correlations in Table 5 are large for this type of data; the minimum correlation is 0.91. Some plausible explanations for the large correlations are as follows:

(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, and increases 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.

4.5. Model equations

The models for the seven endpoints considered and the corroboration model are linear in the coefficients and of a similar form.

[8] The results for the mixed effects models are very similar to the results from the current OLS models. Readers interested in the specific results are referred to the API PAC Analysis Task Group Report (Simpson et al., 2008), or contact the corresponding author.

The forms of the models are described in Table 6; coefficients for the final fitted models are available in the on-line supplement.

4.6. Model fit

The accuracy of the fits of the selected models can best be seen in a plot of observed data points versus the predicted data points. In this type of plot, an individual data point would represent what is observed for a single dose group of an experiment and what is predicted from the statistical model. The optimum would have all points along the 45 degree line, meaning all predicted values equal the corresponding observed value. Fig. 2 is such a plot for fetal body weight, and it shows the observed and predicted points are in very good agreement across the range of data. Plots for the four repeat-dose models are found in Roth et al. (2013) and plots for the three the developmental toxicity models are found in Murray et al. (2013). All seven models show very good agreement between the observed and predicted data points, similar to the results seen in the fetal body weight model (Fig. 2).

5. Model testing and corroboration

The selected models were rigorously tested to ensure that the model results and corresponding correlations were not spurious because of over fitting or applicable only in a unique data region.

5.1. Interpolation and extrapolation

The concepts of interpolation and extrapolation are critical when using a statistical model to predict a new response data point from a new set of independent variables. The predicted data point is called an *interpolated* data point if the predicted data point is based on independent variables that are each within the range of the corresponding independent variables of the set of substances used to develop the model. Conversely, the new predicted data point is called an *extrapolated* data point if at least one of the independent variables is outside the range of the independent variables used to develop the model. We have more confidence in the

Table 6
Forms of the models.

Repeat-dose models:

$$\mathit{Absolute\ Thymus\ Weight} = \alpha + \beta_1 \cdot \mathit{Control\ Thymus\ Weight} + \beta_2 \cdot \mathit{sex} + \beta_3 \cdot \mathit{Body\ Weight} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{5} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

$$\mathit{Platelet\ Count} = \alpha + \beta_1 \cdot \mathit{Control\ Platelet\ Count} + \beta_2 \cdot \mathit{sex} + \beta_3 \cdot \mathit{Study\ Duration} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{3} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

$$\mathit{Hemoglobin\ Concentration} = \alpha + \beta_1 \cdot \mathit{Control\ Hemoglobin\ Concentration} + \beta_2 \cdot \mathit{sex} + \beta_3 \cdot \mathit{Study\ Duration} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{5} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

$$\mathit{Liver\ to\ Body\ Weight\ Ratio} = \alpha + \beta_1 \cdot \mathit{Control\ Liver\ to\ BW\ Ratio} + \beta_2 \cdot \mathit{sex} + \beta_3 \cdot \mathit{Body\ Weight} + \beta_4 \cdot \mathit{Study\ Duration} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{5} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

Developmental toxicity Type I models:

$$\mathit{Fetal\ Body\ Weight} = \alpha + \beta_1 \cdot \mathit{Control\ Fetal\ Body\ Weight} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{5} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

$$\mathit{Live\ Fetuses} / \mathit{litter} = \alpha + \beta_1 \cdot \mathit{Control\ Live\ Fetuses} / \mathit{Litter} + \beta_2 \cdot \mathit{Number\ of\ implants} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{5} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

$$\mathit{probit}(\mathit{percent\ resorptions}) = \alpha + \beta_1 \cdot \mathit{probit}(\mathit{Control\ percent\ resorptions}) + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{5} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

Corroboration model:

$$\mathit{Maternal\ Thymus\ Weight} = \alpha + \beta_1 \cdot \mathit{ControlMaternal\ Thymus\ Weight} + \eta \cdot ARC_4 \cdot ARC_5 + \sum_{i=1}^{7} \gamma_i \cdot \mathit{dose} \cdot ARC_i + \sum_{j=1}^{2} \xi_j \cdot \mathit{dose} \cdot ARC_4 \cdot ARC_5 \cdot ARC_j$$

expected accuracy and precision of interpolated predictions because they are in the range of experience and we have met the statistical assumptions for the models. We do not have the same confidence in the extrapolated predictions because they are based on data outside of our experience of the model and we do not know how the model will respond.

For the models we are discussing there are three classes of independent variables: the seven ARC values, the applied dose, and the biological based variables such as the control group value or study duration. When using the models to predict responses for a new material the choice of ARC values and dose are the variables that could lead to extrapolated predictions. The biological variables are likely to be specified based on the same biological variables as were used to develop the models, and would not lead to extrapolated predictions in that region.

It is easy to understand how to choose a dose that will not lead to an extrapolated prediction: it only has to be within the range of doses used to develop the specific model being used. Similarly, for the single measure variables such as the control response, the body weight, or the study duration, it only has to be within the range of values used to develop the model being used. It is more difficult to understand interpolation and extrapolation for the ARC profile. Table 7 shows the ARC profiles for three hypothetical petroleum substances.

We can plot (Fig. 3) the ARC Profiles of the three hypothetical petroleum substances shown in Table 7 as a spider, or radar, plot. In the plot each substance is represented by a ring with the seven values plotted on the seven legs or rays.

Assume the ARC Profile of Substance A was used to develop a statistical model. The ARC Profile of Substance B is interpolated relative to that of Substance A because all the seven ring concentrations for Substance B are within the ARC values of Substance A. The Substance B biological value predicted from the model would be considered an *interpolated* predicted data point. In contrast, the ARC Profile of Substance C has a Ring 2 concentration that is greater than that of the original substance, Substance A. The Substance C biological value predicted from the model would be considered an *extrapolated* predicted data point. The concept of interpolation also requires that the new substance not only be within a substance used to develop the model but it must be outside a substance used to develop the model – it must be between two substances. Each of the models that were developed included a substance with the value zero for the seven ARC values, therefore when using these models only the 'outer' values need to be considered.

5.2. Model testing

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

(1) using holdout sample data,
(2) using 'nonsense' data,
(3) using new data, and
(4) sensitivity analyses.

Table 7
ARC Profiles of Three Hypothetical Petroleum Substances.

Substance	ARC–1	ARC–2	ARC–3	ARC–4	ARC–5	ARC–6	ARC–7[a]
A	7	5	15	8	8	6	9
B	5	5	12	3	5	2	7
C	5	13	12	3	5	2	7

[a] The ARC 7 value is the weight percent of the 7 and >7 ring aromatic compounds within the petroleum substance as determined by the Method II chemical characterization procedure.

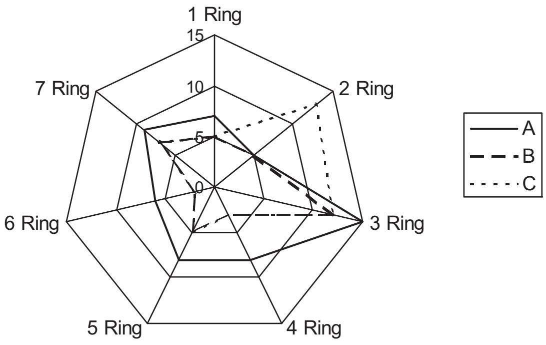

Fig. 3. Spider plot of the three ARC profiles.

5.2.1. Using holdout 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 the data not used to develop the model. This process is called holdout sample corroboration or data-splitting corroboration (Harrell, 2001). The data used to develop the model is called the training data, the remaining data is the test or holdout data.

To demonstrate the model accuracy, the data-splitting technique was expanded by having the method replicated 100 times; each replication used a different set of training and holdout data selected from the full data set. Consider absolute thymus weights from the repeat-dose studies. In the base data set used for the analysis there were 84 observations for the repeat-dose thymus weight. For each replication, approximately 70% of the data points were selected to build the model (training data), and the remaining, approximately 30%, were used as test data (holdout data). The percentages are approximate because the selection process chooses each point with probability 70% rather than choosing 70% of the sample. In each of the 100 replicates, the specific data points in the 70% and 30% groups were different.

The results from the 100 replications using the 84 data points (for a total of 8400 data points) are shown in the observed vs. predicted plots in Fig. 4a and b. Fig. 4a shows the model observed and predicted data for the training data ($n = 5857$), while Fig. 4b shows the plot of the model observed and predicted data for the holdout data ($n = 2543$).

As can be seen in Fig. 4b, some of the predicted data points in the holdout data set are "unreasonable" in that they are not close to the observed data point, as shown by their distance from the 45-degree line of equal values. Moreover, more data points exist outside the plotting boundaries, so the results are actually more extreme than shown {13 observations in the range (0.40, 0.61] and 100 observations in the range [−1.97, −0.10)}. However, some of the holdout data predicted values are interpolated points ($n = 2331$) and some are extrapolated data points ($n = 212$); the points are not identified as interpolated or extrapolated in these plots because of the large number of data points. If the interpolated and extrapolated holdout data points are plotted separately (Fig. 5a and b), the "unreasonable" data points can be seen to be the extrapolated data points, whereas the interpolated data points provide reasonable and accurate predictions.

These plots demonstrate that the predictions from the model for the original data set ('training data' in Fig. 4a) and for the *interpolated* holdout data (Fig. 5a) are good in that the predicted values are close to the observed values. However, model predictions for the *extrapolated* holdout data (Fig. 5b) are mixed, sometimes good and sometimes inaccurate. Note that in Fig. 5a all the interpolated data points are plotted, but in Fig. 5b there are still some extrapolated data points outside the plotting boundaries.

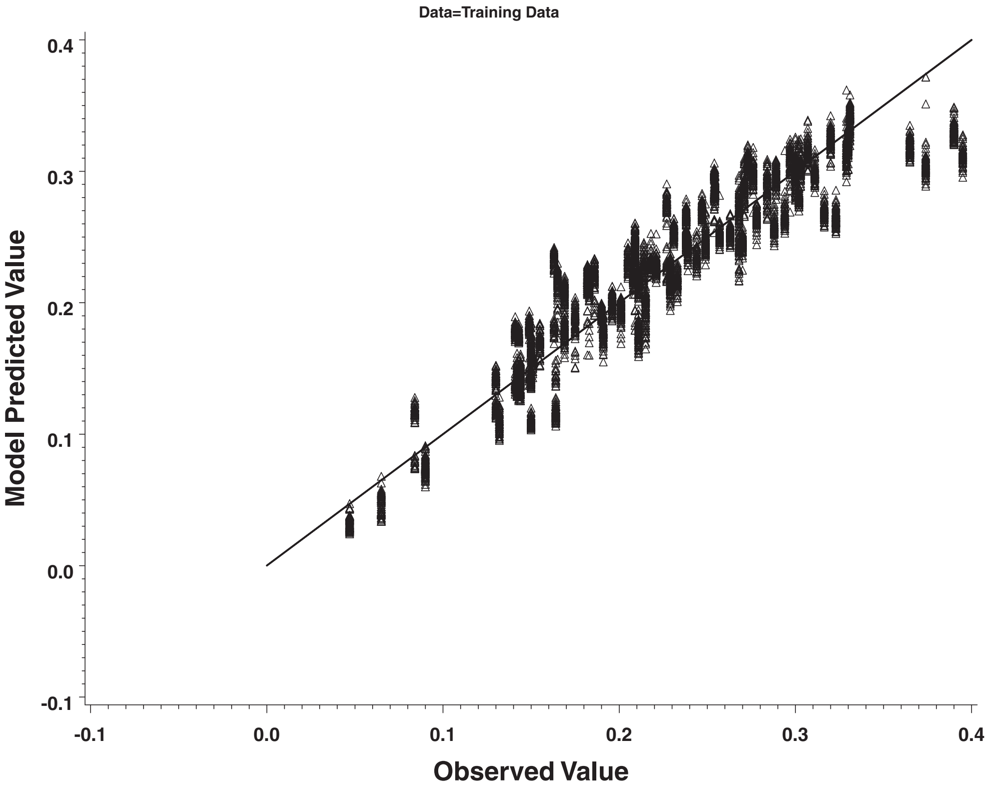

Fig. 4a. Training sample: observed and predicted points of all absolute thymus weight data from repeat-dose studies.

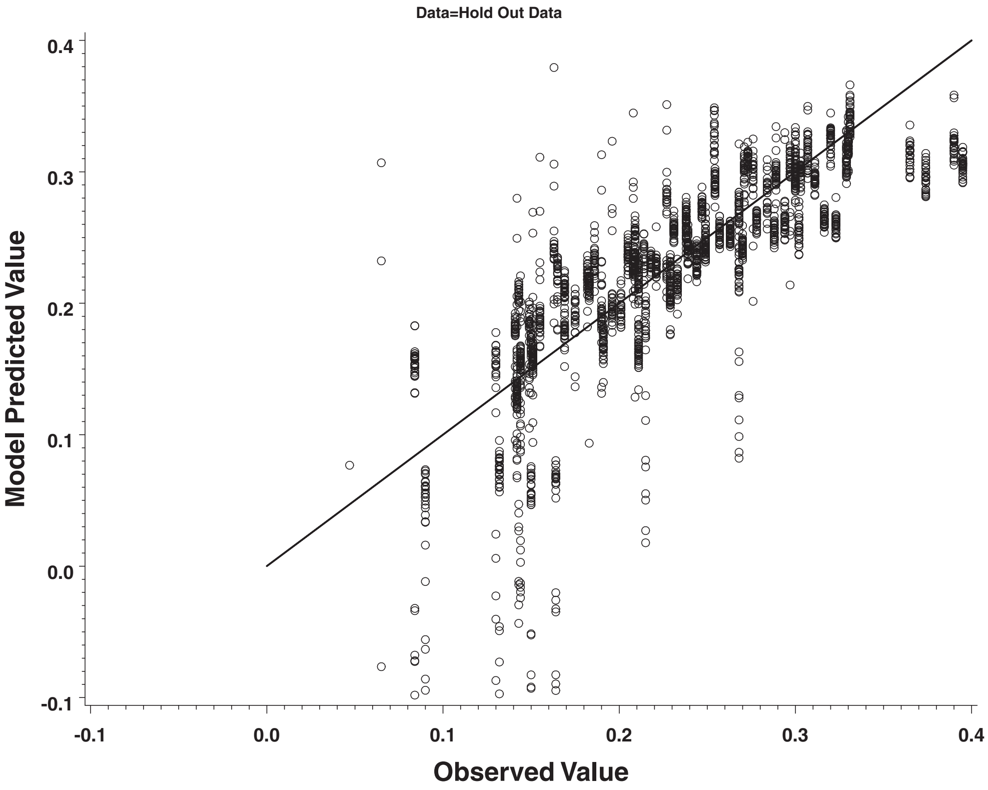

Fig. 4b. Holdout sample: observed and predicted points of all absolute thymus weight data from repeat-dose studies.

5.2.2. Using 'nonsense' data

A model's usefulness can be tested by determining model performance using values for the independent variables (PAC compositional data expressed as an ARC profile) that were *not* associated with the outcome or observed effect (Prajna, 2003). For example, if developing a model using rodent food consumption to predict body

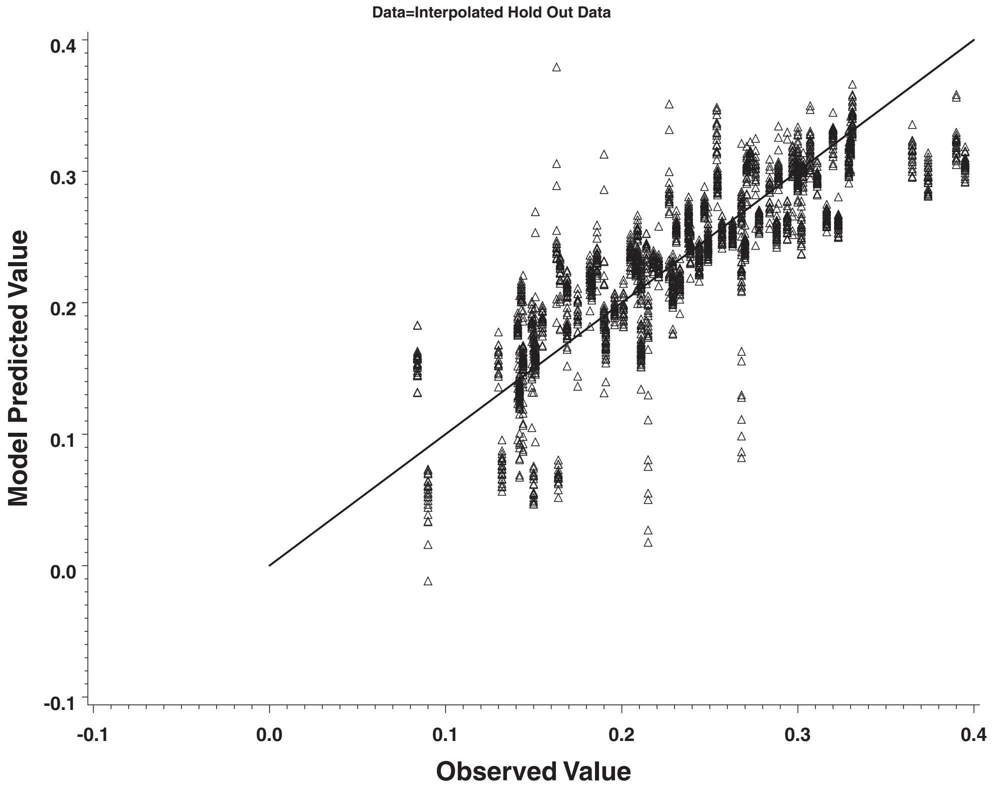

Fig. 5a. Holdout sample: observed and predicted points of the interpolated absolute thymus weight data from repeat-dose studies.

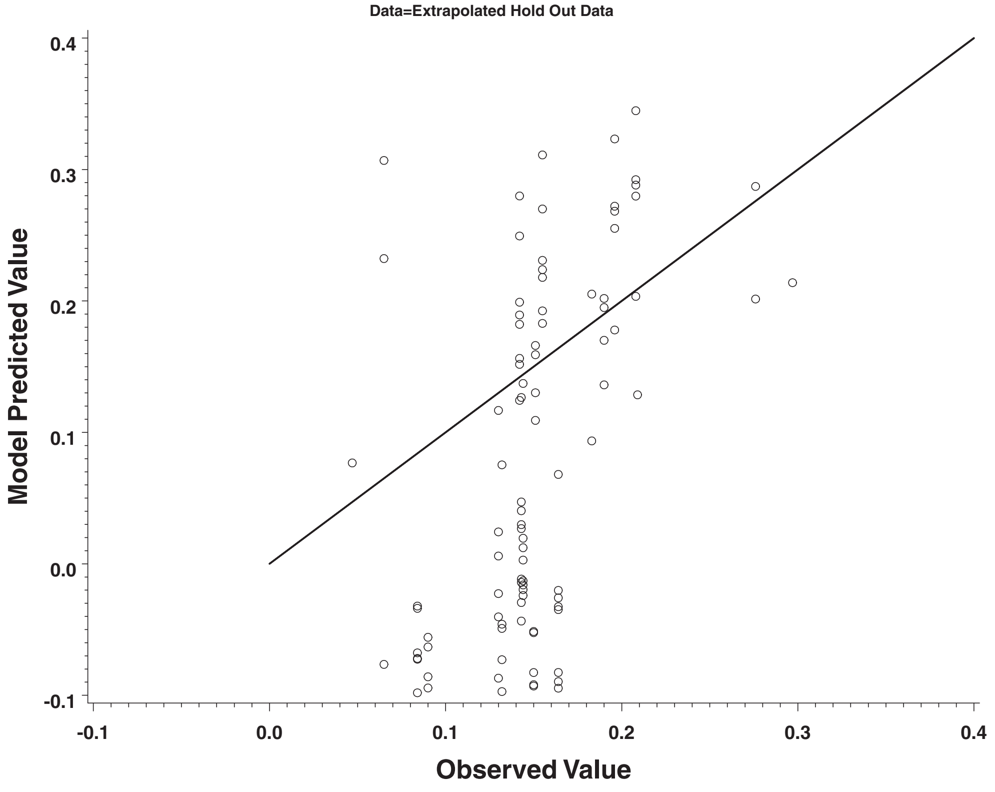

Fig. 5b. Holdout sample: observed and predicted points of the extrapolated absolute thymus weight data from repeat-dose studies.

weight, one could test the model by substituting the cage number in place of food consumption; or, associate animal A's food consumption with animal B's weight gain, which is analogous to what was done to test the current models.

If a model does not fit well using this ‘‘nonsense data’’ (i.e., produces relatively low r values), it is a clear indication that the model behavior is based on information in the data, and is not a result of chance. For example, in the original hemoglobin concentration model there were 98 data points with an r value of 0.94. The response data (hemoglobin concentration) and the corresponding values of the independent variables (ARC Profile) were randomly shuffled and a new model was fit. The process was repeated 100 times. The resulting models had a mean $r = 0.44$, with a minimum and maximum of 0.37 and 0.61, respectively. However, because the model incorporates the control group hemoglobin concentration value, part of the moderately large r (0.44) from the shuffled data is based on the relation between the control and dosed hemoglobin concentration in the ANCOVA model. Without the control group, the r value for the real data was 0.86 (lower than the 0.94 correlation developed with the ANOVA model) and for the 100 shuffled data runs the mean r value was 0.15 (minimum 0.06 and maximum 0.37). This is an indication that the model did not fit random data as well as it fit the real data. A similar series of shuffles was done, but the randomization was restricted to sets within the same petroleum category as the test substance and sex of the respondent. These shuffles selected from a smaller group of possible matches and resulted in some matches that were the same as the original ordering, so the resulting correlations should have been higher than the fully random shuffles, but less than the observed correction. For these restricted shuffles the mean and range of 100 replicates was 0.45 with a minimum and maximum of 0.39 and 0.57, respectively.

These results from the nonsense method of testing, while seemingly good, are still far from the observed r value of 0.94. These relatively low r values from the nonsense data are a clear indication that the model behavior is based on information in the data, and does not result from chance.

5.2.3. Using new data

The corroborative ability of a model can be determined by how well the model predicts results from a study that was not used in the development of the model. There are two types of data that are available for corroboration: direct corroboration and indirect corroboration. There are two studies that are available and can be used for a direct corroboration of the models. In addition, there are two data sets that can be used for indirect corroboration: (1) corroboration of the repeat-dose maternal thymus weight model using the developmental toxicity study maternal thymus weight data as a proxy and (2) corroboration of the developmental toxicity live fetuses/liter model using the number of pups delivered data from a Type II study.

5.2.3.1. Direct corroboration. Two samples that were not used to develop the ARC models were recently tested in standard rat developmental and repeat-dose toxicity studies that meet the ARC model requirements. The sample profiles are interpolations, and the samples final boiling point ⩾ approximately 650 °F (343 °C). Sample 20906 is a light paraffinic distillate aromatic extract and sample 120801 is an ultra-low sulfur diesel oil.

In these studies, the maximum dose was 150 mg/kg_{bw}/day for the repeat-dose study with 4 groups including the vehicle control, and the maximum dose was 450 mg/kg_{bw}/day for the developmental toxicity studies with 5 groups including the vehicle control (the sham control groups from these studies were not included). In both the repeat-dose and developmental toxicity studies, the light paraffinic distillate aromatic extract (sample 20906) produced statistically significant changes from control for at least 1 dose group for all endpoints except platelet count in males. In contrast, the ultra-low sulfur diesel fuel (sample 120801) did not cause a statistically significant effect on any of the parameters from the studies (maximum dose 600 mg/kg_{bw}/day) with 4 groups including vehicle control but not sham control for both studies.

To mimic making predictions for new samples, we used the ARC profile for each sample and the mean biological parameters from the studies used to develop the model (rather than the biological parameters from the new studies) and calculated the predicted endpoints for these 2 studies. Table 8 presents the correlations between the observed and model predicted values. The correlations are a reasonable description of the results, even though they are based on only 4 or 5 observations.

For the four repeat-dose endpoints for both sexes the model predictions for sample 20906 were very good, ranging from 0.76 to 1.00 with a median value of 0.96. There was not a dose response for sample 120801 for three of the four endpoints. The model predicted dose–response curves for sample 120801 were essentially flat (equal response at all doses) except for platelet count where, for each sex, the model predicted a modest decrease in count: a slope of approximately 2.5 units per mg/kg_{bw}/day, where 2.5 units is about 0.25% of the control value. As expected, the correlations between the randomly scattered observed values and the flat, or almost flat, predicted values varied from −0.90 to 0.82.

The fetal body weight predictions for sample 20906 yielded a high correlation, but the predicted dose response pattern was shallower than the observed data. For both samples the predicted model responses for live fetus count and percent resorptions were unusable because the predicted results were the reverse of what was expected based on the results of the other developmental toxicity studies of HBPS reviewed for this paper (increasing number of live fetuses per litter and decreasing percent resorptions with increasing dose). The fetal body weight predictions for sample 120801 were also reversed (increasing fetal body weight with increasing dose) but with a shallow slope.

These sample predictions for the developmental toxicity models for these two samples are not adequate. While a positive outcome is that they did not predict false negatives; they did not provide any estimates. The following is an explanation of why the developmental toxicity models at this point in their development were not adequate for these two samples but will be accurate for other samples, although not necessarily all other samples.

Both of the new test materials (samples 20906 and 120801) had ARC profiles that are interpolations. The assumption was that if a

Table 8
Correlations between observed and predicted data.

Study	Sex	Thymus weight (absolute)	Platelet count	Hemoglobin concentration	Liver weight (relative[a])	Fetal body weight	Live fetuses/ litter	Percent resorptions
20906	M	0.99	0.76	1.00	0.97	0.98	X	X
	F	0.96	0.88	0.96	0.98			
120801	M	0.45	0.82	0.72	−0.90	X	X	X
	F	−0.06	−0.31	0.60	−0.46			

X – model prediction unreliable.

[a] Relative to terminal body weight.

profile is an interpolation the model prediction would be accurate because the profile was surrounded by the profiles of samples used to develop the models. This assumption tacitly assumed the relations were linear, like the points on a line, and if a test number was greater than specific number and less than a third specific number the test number was "between" the other two values. However for the ARC models, the relations were not quite linear because there were 7 ARC values in the profile, and the space did not behave exactly as a straight line; they were more like a bent sheet of paper so that the concept of between was not easy to define.

The ARC 6 values for the two new materials were low (0.3 and 0.0 for sample 20906 and sample 12080, respectively). The samples used to develop the developmental toxicity models and that were the basis of the outer interpolation characteristic, have ARC 6 values that were larger than those of the new samples (0.5, 3.2, 4.9, and 6.0). Consequently, we hypothesized that the current developmental models did not predict the new data well when the ARC 6 value was low because the model had no experience in this region. As a demonstration of this idea, if the ARC value for either sample was numerically increased to 1.0 then the resulting live fetuses/litter and percent resorptions models would have had the predictions in the expected direction as the dose increases. This does not imply any mechanistic or biological importance to the ARC 6 concentrations.

The corroboration samples for the developmental toxicity models may not have provided useable results because the ARC profiles were in an area that was poorly represented by the samples used to develop the models. At a later time, when the models are updated with these, and other, samples we expect the results for these samples will improve. The empirical findings with these samples were predicted well by three of the four repeat-dose models. However, the current live fetuses/litter and percent resorptions models are not yet adequate to predict the developmental toxicity of all HBPS because of the limited number of ARC profile sample patterns of the materials used to develop these models. When additional data are available to fill out the domain of the ARC profiles of these models the problem of reverse predictions are expected to be ameliorated. These new models will likely have the same mathematical form as in Table 6, but with different coefficients.

5.2.3.2. Indirect corroboration – repeat-dose model. Consider the model for absolute thymus weight data from the repeat-dose studies. This model was applied to the developmental toxicity maternal absolute thymus weight corroboration data. If the repeat-dose absolute thymus weight model is adequate, the predictions of maternal thymus weight should be as accurate as predictions from the model developed for the developmental toxicity maternal thymus weight corroboration data. That is, the repeat-dose thymus weight model should work as well with the corroboration study as it did on the data for which it was developed.

Fig. 6 shows the plot of observed vs. predicted data points from the repeat-dose absolute thymus weight model applied to the repeat-dose absolute thymus weight data (the original data points used to develop the model) and the points from the developmental toxicity maternal absolute thymus weight study (the corroboration data) that were predicted by the repeat-dose absolute thymus weight model.

It can be seen from Fig. 6 that the repeat-dose model accurately predicted the corroboration data.

While not part of the model building for this project, we developed a model for the maternal absolute thymus weight. The results from applying the maternal absolute thymus weight model to the repeat-dose absolute thymus weight data were not as accurate as was seen from the reverse (the repeat-dose absolute thymus weight model applied to the maternal absolute thymus weight data). Fig. 7 shows the plot of observed vs. predicted data points from the corroboration maternal absolute thymus weight model applied to the corroboration data (the original data points used to develop the model) and the points from the repeat-dose absolute thymus weight data. This figure also indicates which of the predicted data points are interpolated and which are extrapolated. It can be seen that all of the poorly fitting data points are extrapolated, but that some of the extrapolated data points fit very well. This is a demonstration that, while interpolated data points can be reliable predictors, extrapolated data points may or may not be accurate.

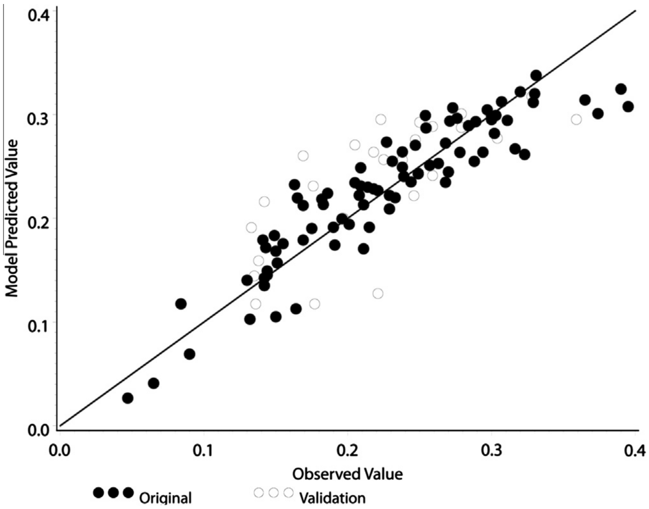

Fig. 6. Observed and predicted absolute thymus weight data based on the model developed from the repeat dose absolute thymus weight data applied to corroboration and repeat-dose data.

Table 9 shows the prediction results for the repeat-dose and corroboration absolute thymus weight models used to predict alternate data with all alternate data points and with only the interpolated or extrapolated points. The column labelled "*r* for base model" is the correlation between the observed and predicted data points based on the original model and the original data used to develop it; the column labelled "*r* for *all* alternate data" is the cor-

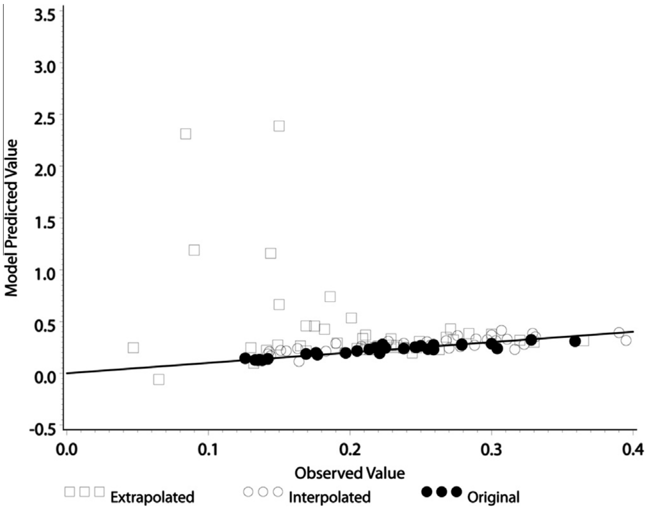

Fig. 7. Observed and predicted absolute thymus weight data based on the model developed from the corroboration absolute thymus weight data applied to corroboration and repeat-dose data with repeat dose data identified as interpolated or extrapolated.

Table 9
Model Correlations for Corroboration Analyses (Absolute Thymus Weight Models).

Model and alternate data set	*r* For base model (n^a)	*r* For *all* alternate data (n^a)	*r* For new data – interpolated predictions only (n^a)	*r* For new data – extrapolated predictions only (n^a)
Repeat-dose data model predicting corroboration study data	0.91 (84)	0.73 (28)	0.72 (25)	0.51 (3)
Corroboration maternal data model predicting repeat-dose study data	0.92 (29)	−0.22 (84)	0.79 (38)	−0.35 (46)

[a] Number of data points.

responding correlation for all the alternate test data using the original model. The last two columns subdivide the alternate data into the extrapolated and interpolated data.

5.2.3.3. Indirect corroboration – developmental toxicity model. For indirect corroboration of the Type I developmental toxicity studies (uterine contents examined during a cesarean section just prior to birth) we used data from the Type II developmental toxicity studies (litters allowed to be delivered naturally, and observations made on the PND 0–4) that have an ARC profile. We expected a moderate relationship between the number of pups delivered (Type II) and the number of fetuses (Type I), or between the day 0 weight of delivered pups (Type II) and the fetal weight (Type I).

The delivered pup count could not be greater than the fetal count, but the delivered count would be negatively influenced by any poor health of the dam and pups, or by any pup cannibalisation that might occur. The day 0 pup weight would naturally be greater than the fetal weight, but the pup weight would, like pup count, be negatively influenced to a greater extent by any poor health of the dam and pups. Because of recorded data limitations both the pup count and pup weight were restricted to the live pups, the relationship was further degraded.

While recognizing these limitations, we used Type II study data to examine their ability to corroborate the fetal count and fetal body weight Type I developmental toxicity models. There is no corresponding data set for the percent of resorptions.

There were 59 data points for the live pups per litter measure of which 41 were from materials not used in the development of the models. The predictions from the developmental toxicity model for fetal count are plotted in Fig. 8, with the original fetal count data used to build the model shown, and the live pup weight data points identified as interpolations or extrapolations.

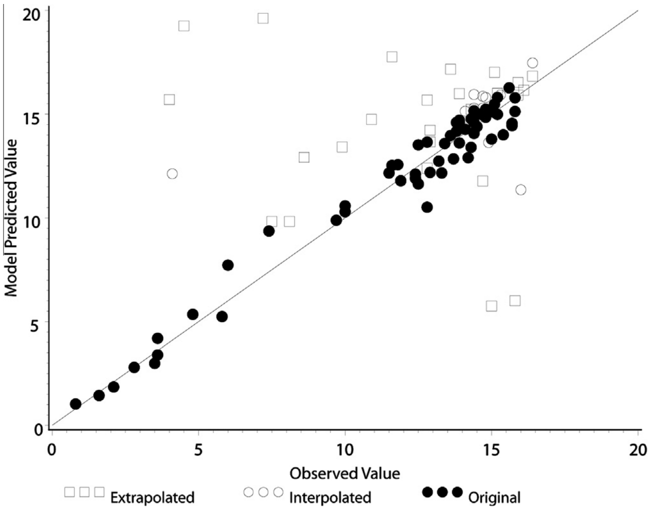

Fig. 8. Observed and predicted live pup or fetal count data identified as interpolated or extrapolated.

There were 53 data points for the live pups' weight per litter measure of which 38 were from materials not used in the development of the models. The predictions from the developmental toxicity model for fetus weight are plotted in Fig. 9, with the original fetal weight data used to build the model shown, and the live pup weight data points identified as interpolations or extrapolations.

The correlations between the observed and predicted data points are shown in Table 10 and the column headings are interpreted as in Table 9.

The correlation among the observed and predicted data for the interpolated fetal count model are reasonable ($r = 0.48$) considering the imposed limitations of using born live animals, the health of the animals, and cannibalisations. The plot indicates that there are more litters with a small fetal litter size compared to the sizes of the delivered pups; the narrower range for the live pups contributes to the low correlation. The correlation for the interpolated fetal weight data is low ($r = 0.37$), but the limitations from the count measure apply and in a more severe manner.

5.2.4. Sensitivity analyses

There are 3 major classes of sensitivity analyses (Saltelli et al., 2000):

(1) screening – determine which factors are important;
(2) local – determine model behavior for individual changes in input values or parameter estimates, usually one-at-a-time over a small range; and,
(3) global – determine model behavior for changes in all inputs and parameters using a distribution of input values.

The local methods were not used in our analysis because they examine variables one-at-a-time and are not adequate for complex

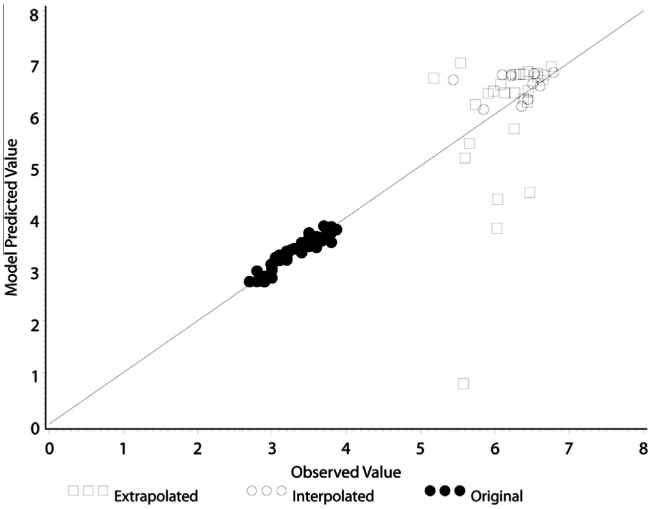

Fig. 9. Observed and predicted live pup or fetal weight data identified as interpolated or extrapolated.

Table 10
Model Correlations for Corroboration Analyses.

Model and alternate data set	*r* For base model (n^a)	*r* For *all* alternate data ([a])	*r* For new data – interpolated predictions only ([a])	*r* For new data – extrapolated predictions only ([a])
Fetal count model predicting live pup counts	0.98 (59)	0.11 (44)	0.48 (11)	0.08 (33)
Fetal weight model predicting live pup weight	0.94 (60)	0.33 (38)	0.37 (11)	0.35 (27)

[a] Number of data points.

models with many variables and with interactions among the variables.

5.2.4.1. Screening analyses. The screening analyses question involves the terms in the model and how they are related to the input data. For the screening analyses step, a set of data was simulated using a Monte-Carlo type simulation. A data set of several tens of thousands of observations was simulated with data for the ARC Profile values, the control values, and the biological parameters (e.g. body weight, sex, etc.).

The data were checked to allow only interpolated data. Each of the seven ARC ring values was generated independently, within its allowable range, but no checking was done to assure that there were not too many high values (except that the sum could not be greater than 100). There were between 39,000 and 40,000 observations that were eventually allowed in each simulation – each of the models had different data sets. The simulated data were classified into a low cumulative PAC set where each of the 7 individual PAC ring value was less than the median value for corresponding ring value in the data set used to develop the final model, or a high cumulative PAC set where it was greater than the median, or neither high or low.

A standard method of assessing the importance of a term in the model is to look at the correlation between the value of the term and the predicted value. If the correlation is high (either positive or negative) that means that term is important in the prediction, and conversely, if the correlation is low, the term is not important in the prediction.

For the repeat-dose platelet count model ($n = 39{,}213$), the correlations (non-parametric) are displayed in a 'tornado' diagram (Fig. 10a) – a bar graph that shows the size and direction of the correlation for all the independent variables. The shadings in the tor-

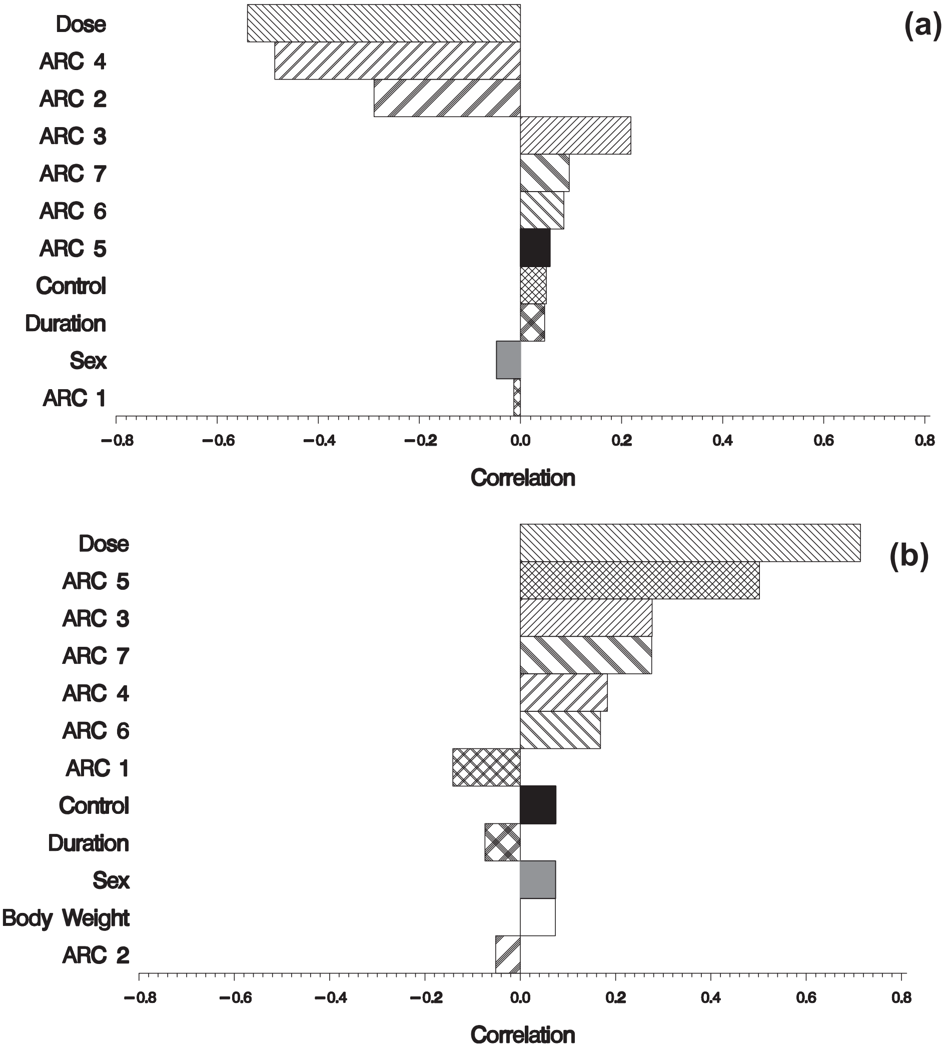

Fig. 10. Tornado diagram (a) repeat-dose platelet count and (b) repeat-dose liver-to-body weight ratio.

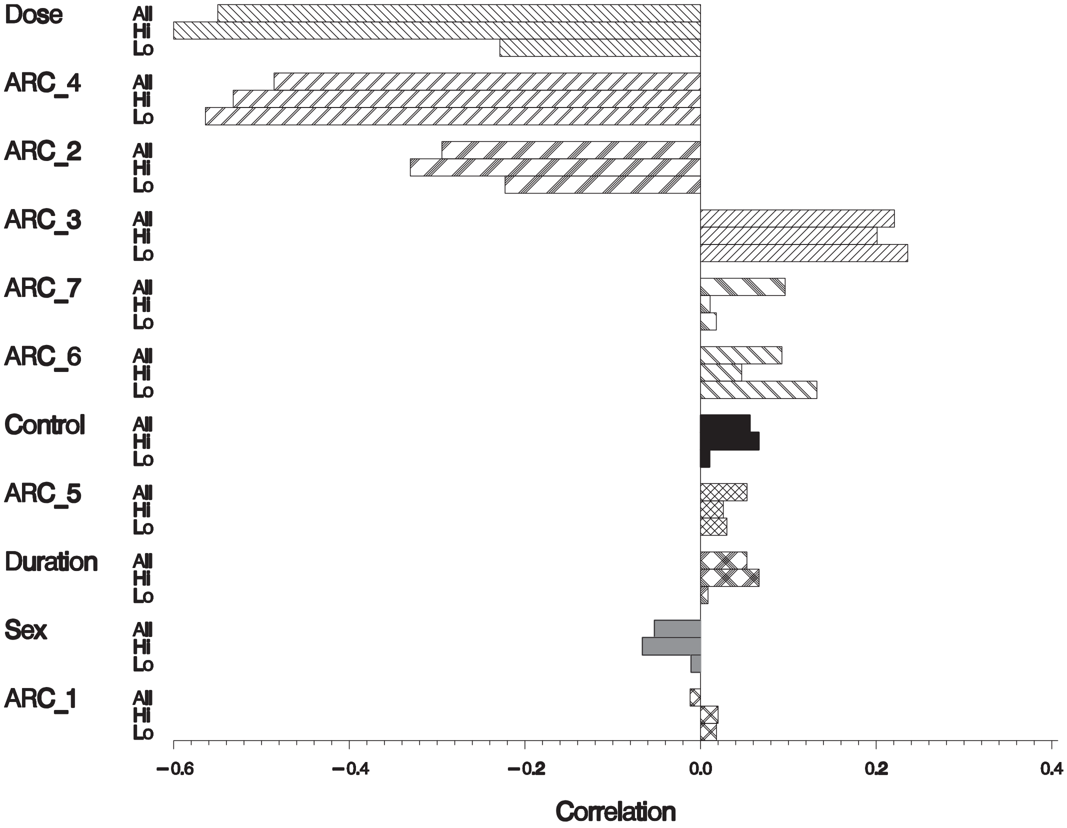

Fig. 11. Repeat-dose platelet count model tornado diagram, total, high and low cumulative ARC values.

nado diagram are used to distinguish the response among the variables. The pattern for the repeat-dose liver-to-body weight ratio model ($n = 36,896$) is shown in Fig. 10b.

For these two models the strongest influence is dose, and the least important is the ARC 1 for the platelet count model and the ARC 2 concentration for the repeat-dose liver to body weight model. If the models are effective then it is expected that dose would have the strongest influence on the response, and that is the case. As expected as dose increases the platelet count decreases (correlation is negative) and as dose increases the liver to body weight ratio increases (correlation is positive). For platelet count the concentrations of ARC 4 and 2 have the largest influence among the concentrations, for the liver to body weight model it is the concentrations of ARC 5 and 7 that have the largest influence among the concentrations. The covariates such as the control value, study durations, and sex have less influence.

However, these tornado diagrams show the influence of each independent variable on the predicted response, but it is averaged over the whole range of simulated data. Recall, the data were simulated in two parts, with high cumulative PAC values (sum of the concentrations of all 7 rings) and low cumulative PAC values.

As seen in Fig. 11 for repeat-dose platelet count the important variables (large correlations) change if low cumulative ARC or high cumulative ARC values are considered. For example, with low cumulative ARC values the influence of dose is very small, likely because the slope of the dose response curve is relatively flat. There is a similar pattern for the ARC 2 concentrations, but not for the other variables. The "sensitivity" of the variable depends on the range of the data used in the simulation. Tornado diagrams for other models, not presented here, show various patterns indicating the different responses for different endpoints.

5.2.4.2. Global analyses. The screening analyses examined the sensitivity of the model to individual data points used in model building and determined which individual points were considered to be "influential" in determining model coefficients. The global analyses explored the sensitivity of the model response to the individual terms in the model when varying simulated input data were used.

The method used for the global analyses was to determine which data points had a strong influence on the estimated model coefficients. The 'strong influence' was based on three well-known summary statistics of individual data points used in building the model (Belsely et al., 1980):

(1) h*i*, or the leverage statistic, is a measure of how unusual the values of the particular independent variable are (is the point an 'outlier' with respect to the independent variables?),
(2) DFFITS*i* measures how much the predicted value of the *i*th observation would change if the *i*th observation was not in the model (it measures influence of the *i*th observation), and
(3) Studentized residual*i* is a measure of how far the model predicted value is from the observed value (is the prediction poor for this data point?).

We first determined which observations from the four repeat-dose models had at least one 'strong influence' statistic in at least three models: this process determined 18 data points. A similar process with the three developmental toxicity models determined 12 data points that met the criteria.

Using these "influential" data points, we determined if the model results were substantially altered when these observations were removed. If the model results were not changed when the points were removed, then the model was insensitive to the data points

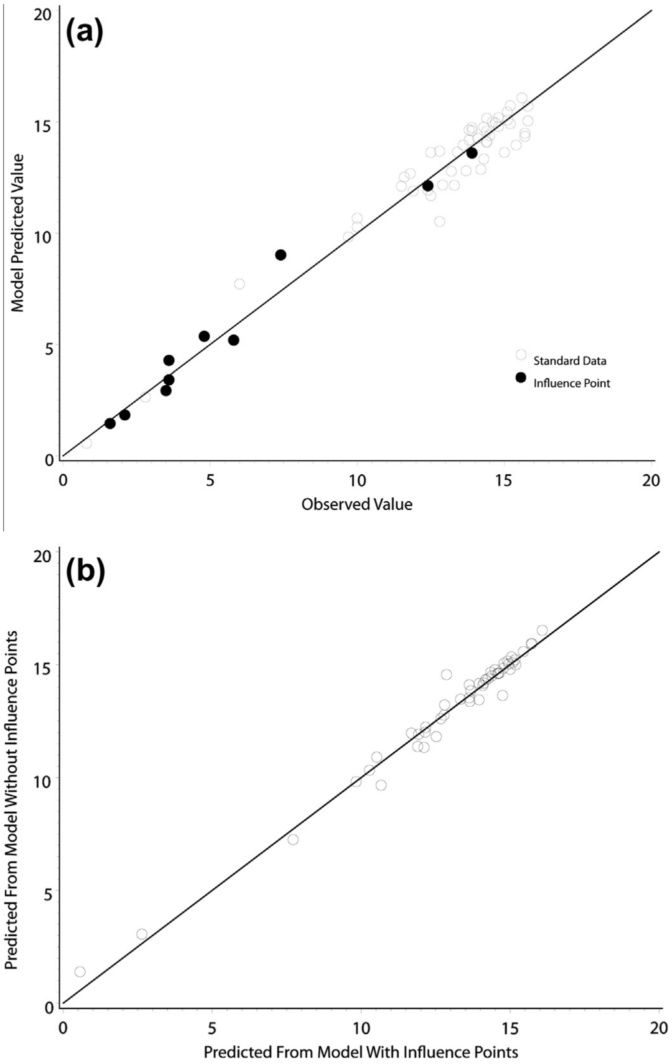

Fig. 12. Developmental toxicity live fetus count model (a) indication of influence points (b) comparison of predicted points from model with and without influence points.

that met the above criteria – if the model was insensitive then we can conclude the model was stable.

For example, Fig. 12a is a plot of the observed and model predicted values for the developmental toxicity live fetus count model that fit with all values, highlighting the 10 influence points to see where they are (note there are only 10 of the 12 influence points included in the developmental toxicity live fetus count data). Fig. 12b shows how the model fits using data without influence points by plotting the predicted value using the model *with* the influence points against the predicted value using the model *without* the influence points. From the two plots, it can be seen that the influence points apparently do not drastically alter the model or model fit and the model is stable. The plots for the other six models not presented here, show similar results.

6. Model comparisons

This section will show that the seven mathematically-based computational models are consistent with standard measures that would be used to describe observed data from a series of repeat-dose or developmental toxicity studies.

A common measure of relative toxicity from a standard toxicity study is the BMD (Crump, 1984). BMDs can only be calculated for samples that have existing toxicity data and therefore cannot be used to characterize the dose–response of untested materials. The BMD is defined as the dose that causes a defined change from control value, e.g. the BMD_{10} is the estimated dose that would cause a 10% change from control value. Essentially, the method uses a set of data from a single, standard toxicology experiment (usually 4 dose groups), fits a maximum likelihood estimation regression model to the data to predict response from dose, then uses an inverse regression estimate of the dose associated with a fixed change to calculate the BMD. The regression model used is usually the best fitting from among a standard set of available models.

The models developed in this paper can be used to calculate the predicted dose response (PDR). The PDR is the model-predicted dose associated with a specified change from control group mean for a particular multi-dose experiment; for example, the PDR_{10} represents the dose associated with a 10% change from control. The PDR_{10} is similar in concept to the BMD_{10} (Crump, 1984), but the PDR is predicted and is calculated by fitting a model that is derived from a series of toxicity studies covering a range of materials, in contrast to the BMD that is generally limited to only one study.

The goal was to assess the degree of consistency between the statistical model based estimate of relative toxicity (PDR_{10}) and the standard measure of relative toxicity based on the observed data (BMD_{10}). The EPA has detailed a set of conditions for calculating a BMD_{10} and when a BMD_{10} can be calculated (Davis et al., 2011), and there are specific conditions for the calculation and interpretation of the PDR_{10}. When the BMD_{10} could not be calculated, we developed an alternate estimate based on either a simple linear regression, or failing that, we developed an estimate based on professional judgment; all three derived values were called the $Estimate_{10}$ and the preferred choice was in the order of the BMD_{10}, linear regression, or judgment. Based on these sets of conditions we developed an algorithm to determine if the estimated PDR_{10} and $Estimate_{10}$ were judged to be consistent for a particular substance and the associated toxicity data. The conditions for developing an estimate and the algorithm were designed to be conservative, that is to minimize any bias in favor of an assessment of consistency. A full explanation of the conditions and the algorithm can be found in Roth et al. (2013), or Murray et al. (2013); the former also provides detailed results for the repeat-dose models and the latter the detailed results for the developmental models. Basically, the measure of consistency is the relative percent difference, defined as 100 times the absolute value of the difference in the two estimates divided by their average value. If the 2 values are A and B, then the relative percent difference is

$$100\left|\frac{A-B}{(A+B)/2}\right|$$

where the vertical lines represent the absolute value. If one value is 3 times as large as the other, the relative percent difference is 100%. We defined the PDR_{10} and the $Estimate_{10}$ to be consistent if the relative percent difference is less than 100%.

Of the 173 possible comparisons between PDR_{10} and $Estimate_{10}$ values derived for the samples that were used in building the repeat-dose and developmental toxicity models, 148 (86%) were assessed to be consistent by the algorithm; (82% for the repeat-dose measures and 96% for the developmental models Roth et al., 2013; Murray et al., 2013).

A major application of these models is for screening new materials. As detailed in Simpson et al. (this issue), in the screening of a sample material each endpoint will likely not be evaluated separately, but the lowest PDR_{10} from among the endpoints estimated for a sample material would be used. The lowest PDR_{10} would indicate the lowest dose among all estimated endpoints associated with a 10% change. Among the 24 sample materials for which we could develop sets of PDR_{10}s and $Estiamtes_{10}$s, we found only 1 sample where the lowest PDR_{10} value and the lowest $Estimate_{10}$ value was not consistent. The sample in question was 87213, with a minimum PDR_{10} of 167 mg/kg_{bw}/day and a minimum $Estimate_{10}$ of 40 mg/kg_{bw}/day, for a relative percent difference of 123%. This is a reasonable comparison when considering this is the most extreme difference when comparing the results of model predictions to observed data for 24 HBPS.

Overall, when compared to an observed toxicity study data the models had an 86% agreement or consistency rate for individual predictions, and over a 95% agreement rate when considering a toxicity characterization of a HBPS. The models did very well when compared to standard measures from observed study data, especially when considering the known variability of rodent toxicity studies (Haseman et al., 1989).

7. Discussion

The primary purpose of the present investigation was to determine whether there is a relationship between the PAC content of HBPS and their repeat-dose and developmental toxicity endpoints. 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 with similar physical and chemical properties.

We found that there are indeed associations between sensitive repeat-dose and developmental toxicity endpoints and the PAC content (expressed as the ARC profile) of selected petroleum substances. We have also demonstrated that numerical estimates of these sensitive repeat-dose and developmental toxicity endpoints can be predicted for an untested substance based on its ARC profile.

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 (Feuston et al., 1994). In comparison to the previous evaluation, the current statistical technique makes use of a larger data set and is based on observed numerical values as opposed to ranks. The large number of data points used to develop the models is a particular strength of the current evaluation. The plots of the observed vs. predicted points for samples used to build and to corroborate the models demonstrate that the models are both accurate descriptors of the observed data and accurate predictors for interpolated substances. The models are relatively simple linear models, all with a similar mathematical form across the endpoints, which provides a measure of the concordance of the models.

To predict the toxicity of an untested substance using the models, the only compositional input that is required is the ARC profile of the substance as determined by the Method II chemical characterization procedure (see Gray et al., 2013 for details). Based on the sensitivity analyses we can see that the model is not based on the influence of only a few points; rather, it is stable, and adding more data will likely improve the fit and the range of applicability. We also can see that all the terms in the model are likely to be important over some domain of the input data. The simulations show how the degree of importance of an independent variable changes with the domain of the input data.

The predictive models described here have a number of constraints. As with most linear regression models of this form, the models were found to be good predictors of treatment effects in the majority of comparisons we made if the ARC profile and dose of the petroleum substance fell within the ARC profiles and doses that had been used for model development (i.e., the prediction was an interpolation). Not surprisingly, the models were sometimes less accurate predictors if the ARC profile and/or doses of the unknown petroleum substance fell outside the ARC profiles that had been used for model development (i.e., the prediction was an extrapolation). To investigate and mitigate the extrapolated data limitation requires that more biological studies be conducted on substances with ARC profiles (derived using Method II) and doses that are outside the profiles and doses that were used to develop these models.

There are two other circumstances where the models may give seemingly inaccurate results. In one situation the untested material is inherently relatively non-toxic, that is, it has a flat or relatively flat dose response curve. In this situation the model may either predict a flat, slightly increasing, or slightly decreasing dose response because of random variation around the flat slope. If the model selects the dose response that is "contrary" to the expected effect (slightly in the wrong direction, say a slope of 1.01 where a slope of 1.0 or less is expected) then the model may appear to be in error even though this is just a slight variation. The other situation is when the ARC model predictions are in fact in error and result in an unreasonable dose response model. For example, if for an untested material the ARC model predicts a 500% *increase* in fetal body weight for every 100 mg/kg_{bw}/day increase in dose, in this case the prediction is contrary to what is expected and the predicted effect is large. As previously noted, the ARC models are complex and have been built with a relatively small number of materials (individual ARC profiles), there may be areas within the ARC profile region where there is little or no biological information, causing the model to falter. The second situation will be ameliorated when additional biological studies and associated PAC determinations are conducted in the data poor regions. In the future, as new test data become available, the results can be incorporated into the current models, further corroborating them and expanding the domain of applicability.

The domain of the data used to develop the models described in this paper included dermal studies in rats, and the models cannot be applied to other routes of exposure or other species. If data become available for other routes or species then a similar model development exercise may develop useful models for the additional route and/or species.

Although the various models were built using experimental data developed on samples from across a range of petroleum categories, approximately 70% of the samples were from the gas oils and heavy fuel oils categories. Because the compositional component of the models is based only on ARC profile and not on specific category membership, the models are applicable to a wide range of petroleum substances in which PAC may be in some way related to the effects of interest, with the proviso that the ARC profile of the new substance is interpolated relative to the profiles of the substances used to develop the models.

It should be stressed that the models developed herein are based strictly on observed statistical relationships, not on biological knowledge or any presumed mechanism of action. No attempt was made to identify causal relationships. Since the mechanism of action is not understood, the data should be viewed only as indicating that there is an association between the ARC profile and repeat-dose or developmental toxicity. The data should not be used to draw conclusions about whether any of the specific PAC ring structures, individually or collectively, is the cause of the observed repeat-dose or developmental toxicity, only that the pattern of ARCs (ARC profile) in a petroleum substance may allow an estimate of toxicity through modeling.

8. Conclusions

The current review and evaluation of the unpublished company laboratory toxicology and analytical reports show that predictive models for effects on the selected most sensitive SIDS repeat-dose and developmental toxicity endpoints can be developed using the weight percent of the array of the 1–7 and larger aromatic-ring compounds in the test substance (the "ARC profile"). The effects found to be associated with the ARC profile are consistent with those reported for a number of individual PAHs and PAC-containing materials, although the mechanism(s) of toxicity in this regard are not known and have not been investigated in this study.

In the repeat-dose toxicity studies, associations were found and characterized between the ARC profile and effects on absolute thymus weight, relative liver weight, hemoglobin concentration and platelet count. In the developmental toxicity studies, associations were found and characterized for effects on fetal weight, number of live fetuses/litter and percent resorptions. These were the biological endpoints most commonly affected in the dataset and often affected the determination of the LOEL. As part of a corroboration exercise, and to show the wider applicability of the modeling process, a model for the absolute maternal thymus weight from developmental toxicity studies was also successfully developed. The overall four-part model testing and corroboration effort demonstrated the ability of the models to predict results for HBPS with a usable degree of accuracy.

It should be noted, the models were developed based on observed statistical relationships. No attempt was made to identify causal relationships. To do this would have required a more detailed understanding of the mechanisms of PAC toxicity, or at least a general understanding of the underlying mode of toxic action that was beyond the scope of the current evaluation.

Conflict of interest

Four of the coauthors (RNR, BJS, MJN, FJM) are paid consultants to the Petroleum High Production Volume Testing Group Three (RNR, BJS, MJN) are former employees of companies that manufacture petroleum products. One co-author (TMG) is employed by the American Petroleum Institute.

Role of the funding source

The authors received financial support for the research, authoring and publication of this article from the Petroleum High Production Volume Testing Group.

Acknowledgments

This project was sponsored and funded by the Petroleum HPV Testing Group (PHPVTG), an unincorporated group of manufacturers and importers affiliated by contractual obligation to fund a voluntary data disclosure and toxicity testing program on certain petroleum-related chemical substances in response to the US EPA HPV Challenge Program. The American Petroleum Institute (API) manages the PHPVTG's activities.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.yrtph.2012.11.015.

References

Altgelt, K.H., Boduszynski, M.M., 1994. Composition and Analysis of Heavy Petroleum Fractions. Marcel Dekker, New York.

Belsely, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics. Wiley and Sons, New York.

Blackburn, G.R., Roy, T.A., Bleicher Jr., W.T., Reddy, M.V., Mackerer, C.R., 1996. Comparison of biological and chemical predictors of dermal carcinogenicity of petroleum oils. Polycyclic Aromatic Compounds 11, 201–210.

Crump, K., 1984. A new method for determining allowable daily intakes. Fund. Appl. Toxicol. 4, 854–871.

Davis, J.A., Gift, J.S., Zhao, J., 2011. Introduction to benchmark dose methods and US EPA's benchmark dose software (BMDS) version 2.1.1. Toxicol. Appl. Pharmacol. 254, 181–191.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. Wiley and Sons, New York.

Feuston, M.H., Low, L.K., Hamilton, C.E., Mackerer, C.R., 1994. Correlation of systemic and developmental toxicities with chemical component classes of refinery streams. Fund. Appl. Toxicol. 22, 622–630.

Feuston, M.H., Hamilton, C.E., Mackerer, C.R., 1997. Oral and dermal administration of clarified slurry oil to male C3H mice. Int. J. Toxicol. 16, 561–570.

Gray, T.M., Simpson, B.J., Nicolich, M.J., Murray, F.J., Verstuyft, A.W., Roth, R.N., McKee, R.H., 2013. Assessing the mammalian toxicity of high-boiling petroleum substances under the rubric of the HPV program. Regul. Toxicol. Pharmacol. 67 (2S), S1–S3.

Harrell Jr., F.E., 2001. Regression Modeling Strategies. Springer-Verlag, New York.

Haseman, J.K., Huff, J.E., Rao, G.N., Eustis, S.L., 1989. Sources of variability in rodent carcinogenicity studies. Fund. Appl. Toxicol. 12, 793–804.

Klimisch, H.J., Andreae, M., Tillman, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25, 1–5.

Murray, F.J., Roth, R., Nicolich, M., Gray, T., Simpson, B., 2013. The relationship between developmental toxicity and aromatic-ring class profile of high-boiling petroleum substances. Regul. Toxicol. Pharmacol. 67 (2S), S46–S59.

OECD (Organisation for Economic Co-Operation and Development), 1981a. OECD Guidelines for the Testing of Chemicals/Section 4: Health Effects – No. 410: Repeated Dose Dermal Toxicity: 21/28-day Study. Paris, France.

OECD (Organisation for Economic Co-Operation and Development), 1981b. OECD Guidelines for the Testing of Chemicals/Section 4: Health Effects – No. 411: Subchronic Dermal Toxicity: 90-day Study. Paris, France.

OECD (Organisation for Economic Co-Operation and Development), 1996. OECD Guidelines for the Testing of Chemicals – No. 422: Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test. Paris, France.

OECD (Organisation for Economic Co-Operation and Development), 2001. OECD Guidelines for the Testing of Chemicals – No. 414: Combined Prenatal Developmental Toxicity Study. Paris, France.

Oreskes, N.M., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation and confirmation of numerical models in the earth sciences. Science 263, 641–646.

Patterson, J., Maier, A., Kohrman-Vincent, M., Dourson, M.L., 2013. Peer consultation on relationship between PAC profile and toxicity of petroleum substances. Regul. Toxicol. Pharmacol. 67 (2S), S86–S93.

Pinheiro, J.C., Bates, D.M., 2002. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York.

Potter, T.L., Simmons, K.E., 1998. Composition of Petroleum Mixtures: Total Petroleum Hydrocarbon Criteria Working Group Series Volume 2. Amherst Scientific Publishers, Amherst, MA.

Prajna, S., 2003. Barrier Certificates for Nonlinear Model Validation, Control and Dynamical Systems. California Institute of Technology, Technical Report 04-001. <http://www.cds.caltech.edu/~prajna/papers_pdf/modinv.pdf> (accessed 18 January 2010).

Roth, R.N., Simpson, B.J., Nicolich, M.J., Murray, F.J., Gray, T.M., 2013. The relationship between repeat-dose toxicity and aromatic-ring class profile of high-boiling petroleum substances. Regul. Toxicol. Pharmacol. 67 (2S), S30–S45.

Roy, T.A., Johnson, S.W., Blackburn, G.R., Deitch, R.A., Schreiner, C.A., Mackerer, C.M., 1985. Estimation of mutagenic and dermal carcinogenic activities of petroleum fractions based on polynuclear aromatic hydrocarbon content. In: Cooke, M., Dennis, A.J. (Eds.), Polynuclear Aromatic Hydrocarbons: A Decade of Progress. Batelle Press.

Roy, T.A., Johnson, S.W., Blackburn, G.R., Mackerer, C.M., 1988. Correlation of mutagenic and dermal carcinogenic activities of mineral oils with polycyclic aromatic compound content. Fund. App. Toxicol. 10, 466–476.

Roy, T.A., Blackburn, G.R., Mackerer, C.R., 1994. Evaluation of analytical endpoint to predict carcinogenic potency of mineral oils. Poly. Arom. Comput. 5 (1), 279–287.

Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity Analysis. Wiley & Sons, New York.

Simpson, B., Dalbey, W., Fetzer, J., Gray, T., Murray, J., Nicolich, M., Roth, R., Saperstein, M., White, R., 2007. 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 (report for peer consultation), Sponsored by the Petroleum HPV Testing Group, July 31, 2007. <http://www.tera.org/peer/API/APIWelcome.htm> (accessed 23 July 2012).

Simpson, B., Dalbey, W., Fetzer, J., Gray, T., Murray, J., Nicolich, M., Roth, R., Saperstein, M., White, R., 2008. The relationship between the aromatic ring class content and selected endpoints of repeat-dose and developmental toxicity of high-boiling petroleum substances. Report of the PAC Analysis Task Group, Sponsored by the Petroleum HPV Testing Group, March 31, 2008. <http://www.petroleumhpv.org/pages/publications.html> (accessed 16 January 2012.

Simpson, B.J., Murray, F.J., Roth, R.N., Nicolich, M.J., Gray, T.M., this issue. Application of statistical models to characterize the repeat-dose and developmental toxicity of high-boiling petroleum substances. Regul. Toxicol. Pharmacol.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611.

Speight, J., 2007. The Chemistry and Technology of Petroleum, fourth ed. CRC Press, Taylor & Francis Group, Boca Raton.

US EPA (US Environmental Protection Agency), 2000. Data collection and development on high production volume (HPV) chemicals. Federal Register 65(248), 81686–81698.

US EPA (US Environmental Protection Agency), 2009. Guidance on the Development, Evaluation, and Application of Environmental Models, EPA/100/K-09/003. <http://www.epa.gov/crem/library/cred_guidance_0309.pdf> (accessed 21 April 2012).