



## The relationship between developmental toxicity and aromatic-ring class profile of high-boiling petroleum substances



F. Jay Murray<sup>a,\*</sup>, Randy N. Roth<sup>b</sup>, Mark J. Nicolich<sup>c</sup>, Thomas M. Gray<sup>d</sup>, Barry J. Simpson<sup>e</sup>

<sup>a</sup> Murray & Associates, 5529 Perugia Circle, San Jose, CA 95138, USA

<sup>b</sup> Roth Toxicology Consulting, P.O. Box 6023, Thousand Oaks, CA 91359, USA

<sup>c</sup> COGIMET, 24 Lakeview Rd., Lambertville, NJ 08530, USA

<sup>d</sup> American Petroleum Institute, 1220 L Street, N.W., Washington, DC 20005, USA

<sup>e</sup> Simpson Toxicology Consulting, 4 Temple Farm Barns, Singledge Lane, Whitfield CT15 5AB, UK

### ARTICLE INFO

#### Article history:

Available online 13 May 2013

#### Keywords:

Biological models  
Dermal  
Developmental toxicity  
High-boiling petroleum substances  
HPV Challenge Program  
Mixtures toxicity  
Polycyclic aromatic compounds  
QCAR modeling  
Rat  
UVCB

### ABSTRACT

In response to the US EPA HPV Challenge Program, this study was conducted to: (1) evaluate the relationship between PAC content and the developmental toxicity of high-boiling petroleum substances (HBPS) and (2) develop mathematical models to predict the developmental toxicity of similar untested substances based on their aromatic ring class (ARC) profiles. For this investigation, 68 developmental toxicity studies were reviewed. The ARC models relied on data from 21 rat dermal developmental toxicity studies conducted with similar experimental designs to ensure a consistent data set for comparison. The most sensitive general endpoints of developmental toxicity (i.e., decreased fetal survival and growth) were chosen for modeling. The ARC models demonstrated a strong correlation between the predicted vs. observed values for specific sensitive endpoints of these developmental toxicities (percent resorptions,  $r = 0.99$ ; live fetuses per litter,  $r = 0.98$ ; fetal body weight,  $r = 0.94$ ). Such associations provide a promising approach for predicting the developmental toxicity of untested HBPS. Efforts to corroborate the ARC models using test substances that were not used to build the ARC models produced mixed results, and further development and refinement of the ARC models is recommended before they can be reliably applied to all HBPS.

© 2013 Elsevier Inc. All rights reserved.

### 1. Introduction

High-boiling petroleum substances (HBPS), i.e., substances with final boiling points  $\geq$  approximately 650 °F (343 °C), include substances such as asphalt, aromatic extracts, crude oils, heavy fuel oils, gas oils, lubricating oil basestocks, waxes and related materials, and certain petroleum waste substances. Each HBPS contains at least thousands of structurally-related individual substances, including a wide variety of polycyclic aromatic hydrocarbons (PAHs) and polycyclic aromatic compounds (PACs) (Altgelt and Boduszynski, 1994; Potter and Simmons, 1998). 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 stream (Speight, 2007).

A limited number of developmental toxicity studies of HBPS have been published in the scientific literature (Feuston et al., 1989, 1997a, b; Feuston and Mackerer, 1996ab). Certain HBPS have been reported to cause evidence of developmental toxicity in animal studies. Among these substances, the observations of

developmental toxicity most often reported included an increased incidence of resorptions (and a corresponding decrease in the number of live fetuses per litter) and a decrease in fetal body weight.

A few individual PACs have been evaluated for their potential to cause developmental toxicity. For example, benzo(a)pyrene has been reported to cause an increase in the percentage of resorptions and a decrease in fetal body weight among the offspring of pregnant rats exposed by subcutaneous injection (Bui et al., 1986). In addition, decreased fetal survival was reported among the offspring of pregnant rats exposed by inhalation to benzo(a)pyrene (Archibong et al., 2002). In humans, maternal exposure to airborne PAHs has been associated with reduced birth weight among the offspring of women in Krakow and New York City (Choi et al., 2006, 2008). In another epidemiological study, exposure to PAHs from barbecued meat consumed during pregnancy was linked to a decrease in birth weight; no effect on the duration of pregnancy was observed (Jedrychowski et al., 2012).

Feuston et al. (1994) reported a correlation between the chemical composition of petroleum streams and certain endpoints of developmental toxicity. This work was based on the results of developmental toxicity studies of 11 different samples of petroleum streams applied dermally to pregnant rats; the basic experimental

\* Corresponding author. Fax: +1 408 239 0559.

E-mail address: [jmurray2@sbcglobal.net](mailto:jmurray2@sbcglobal.net) (F.J. Murray).

design of these studies has been described previously (Feuston et al., 1989). Feuston et al. (1994) performed a Spearman rank-order test to compare endpoints of developmental toxicity, as defined by the study Lowest Observed Effect Levels (LOELs) for increases in resorptions and decreases in fetal body weight, against the composition of refinery streams. These investigators found that developmental toxicity (i.e., increased resorptions and decreased fetal body weight) was correlated with 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 current study was conducted in order to: (1) evaluate the relationship between PAC content and the developmental toxicity of a range of HBPS, and (2) develop mathematically-based models to predict the developmental toxicity of similar untested substances. The current study extends the initial evaluation of Feuston et al. (1994) by assessing a greater number of developmental toxicity studies, and involving a larger number of HBPS. It uses sophisticated, mathematically-based computational models to evaluate the relationship between developmental toxicity and the aromatic-ring class (ARC) profile. The models developed for this project are not biologically-based; rather, they were developed without any preconceived notions about the mechanism of action of developmental toxicity associated with HBPS.

This study is part of a larger investigation by the authors to (1) evaluate potential relationships between PAC content and the toxicities of HBPS and (2) if any identified relationships are defined, use them to predict the toxicity of untested HBPS for the US EPA HPV Challenge Program (US EPA, 2000). The report of the investigation pertaining to acute, repeat-dose, developmental and reproductive toxicity of HBPS has undergone a Toxicology Excellence for Risk Assessment (TERA) peer consultation (Patterson et al., 2013). Additional aspects of the larger investigation, which also included genetic toxicity, are described in the accompanying articles (McKee et al., 2013; Murray et al., 2013; Nicolich et al., 2013).

## 2. Materials and methods

### 2.1. Terminology

The following is a list of terms and definitions used throughout this paper.

**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., 1988, 1994) e.g., the DMSO-soluble ARC 3 content would be the weight percent of the three ring aromatic compounds within the petroleum substance.

### 2.2. Developmental toxicity data

Participating companies of the Petroleum HPV Testing Group were requested to provide the original laboratory reports for any

developmental toxicity studies that had PAC compositional data available on the test sample. A boiling point criterion was added later in the analysis which excluded certain types of petroleum substances, such as gasoline refinery stream and kerosene that are not likely to contain aromatic compounds boiling above 650 °F (343 °C) and hence cannot be addressed by models based on PAC content. Initially, reports of 68 developmental toxicity studies were submitted to the authors (Table 1). Subsequent to the completion of the final models (i.e., the ARC models), the authors obtained dermal developmental toxicity studies of two HBPS, which were not used to build the models, that provided an opportunity to evaluate the predictive accuracy of the models, see Section 2.8. A few of the provided developmental toxicity studies had been published in the scientific literature, but for purposes of this project, the original laboratory reports (not the publications) were employed. The materials tested in the company studies covered a range of PAC-containing petroleum substances including asphalt, aromatic extracts, crude oils, gas oils, lubricating oil basestocks, heavy fuel oils, waxes and related materials, and certain petroleum waste substances.

All of the developmental toxicity studies submitted were conducted in Sprague–Dawley rats, and most used dermal application as the route of test material application. Most of the developmental toxicity studies are best described as screening studies because the group size was typically in the range of 10–20 mated females per group. However, some of the developmental toxicity studies were larger with group sizes of 25 mated females. All the studies had at least one concurrent control group, and most had at least three dose groups. All the reports were evaluated and given Klimisch reliability scores of either 1 (reliable without restrictions) or 2 (reliable with restrictions) (Klimisch et al., 1997). Ultimately, no studies were excluded for reasons of reliability or data quality.

Thirty-three of the developmental toxicity studies were of a traditional design in which the pregnant rats were exposed during gestation and the uterine contents were examined during a C-section just prior to birth. For this assessment, these studies were termed “Type I” developmental toxicity studies. In an additional 35 of the submitted developmental toxicity studies, pregnant rats were exposed during gestation (or in some cases from 7 days prior to mating through GD 20), 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.

The test material was administered dermally in all of the developmental toxicity studies with the exception of four Type I study reports, in which 13 HBPS were given by gavage on a single day of gestation. The gavage studies were designed to maximize the ability to detect potential teratogenicity by giving the test material

**Table 1**  
Availability of developmental toxicity studies.

Studies	No. of Type I studies <sup>a</sup>	No. of Type II studies <sup>b</sup>	Total studies
Studies received	33	35	68
Studies from which data were extracted	29	35	64
Studies used for preliminary modeling	23	34	57
Studies used for final modeling	21	0	21
Studies obtained after final models completed <sup>c</sup>	2	0	2

<sup>a</sup> Pregnant females exposed during gestation, Caesarean section on day 20 of gestation.

<sup>b</sup> Pregnant females exposed during gestation, dams allowed to deliver and pups monitored through day 4 of lactation.

<sup>c</sup> Studies used to evaluate the final models, see Section 3.7.

on a single day of gestation as a large bolus dose, usually 2000 mg/kg<sub>bw</sub>/day. Although most of the toxic effects were qualitatively similar in the dermal and oral studies for substances studied by both routes of administration, much higher doses were required to produce the effects in the oral studies since the test material was given on only a single day of gestation compared to the dermal studies with repeated, daily treatment during gestation. For most HBPS, dermal contact is considered the most likely route of human exposure.

### 2.3. Identification of toxicity endpoints and studies for modeling

The goal of the project was to model sensitive endpoints of developmental toxicity, not maternal toxicity; thus, endpoints of maternal toxicity were not generally considered. Further, the nature of the data on maternal toxicity among the developmental toxicity studies of HBPS was not ideal for modeling. For example, the data on maternal body weight, body weight gain, and food consumption was not easily comparable among these studies because the pregnant body weight and food consumption were measured on different days of gestation in different studies. Observations of skin irritation presented a similar challenge. Although the dams were monitored for skin irritation in most of the developmental toxicity studies, skin irritation was not measured or graded consistently across studies. Therefore, a subset of studies with comparable data on skin irritation was not readily available for purposes of quantitative modeling. Maternal thymus weights were modeled in order to corroborate the thymus weight model developed from the repeat-dose toxicity studies of HBPS, as described by Nicolich et al. (2013).

In order to identify a comparable dataset for modeling, only the studies conducted by the dermal route of exposure were considered for modeling. Additional criteria for including studies and dose groups in the modeling effort were established by the authors, as detailed by Nicolich et al. (2013). These criteria included factors such as duration of dosing, group size, and the method of compositional analysis. For example, the authors decided to use only data from studies that included daily dosing on gestation day (GD) 0–19, as a minimum. After application of the criteria for inclusion, a total of 23 and 34 Type I and II studies, respectively, were identified for the preliminary analysis (Table 1).

For the final modeling, only test substances with a final boiling point  $\geq$  approximately 650°F (343 °C) were considered and used. The purpose of the boiling point criterion is to exclude certain types of petroleum substances that are not likely to be addressed by models based on PAC content. Petroleum substances with high boiling points generally contain fused aromatic-ring compounds with  $\geq 3$  rings, which are the PAC compounds of interest for developmental toxicity (Feuston et al., 1994). Substances with lower boiling points are not expected to contain PAC compounds with  $\geq 3$  aromatic rings.

Only Type I developmental toxicity studies were used for the final modeling for several reasons. First, the correlations between developmental toxicity and ARC profiles in the preliminary modeling was slightly better with the Type I studies compared to the Type II studies. Second, the endpoints in the Type I and II studies are not directly comparable since there are no fetal data on GD 20 in the Type II studies; this precluded combining data from the Type I and II studies. Third, models based on the Type I studies were considered of greater utility than models based on Type II studies. More specifically, models based on Type I studies are designed to predict the effects observed in conventional developmental toxicity studies currently required by regulatory agencies around the world. In contrast, Type II studies, while scientifically valid, are less common because it is unusual to sacrifice the dams and offspring on PND 4 in developmental toxicity studies. Although

it is common to sacrifice offspring on PND 4 for reproductive toxicity screens (i.e., OECD 421/422), these tests do not typically limit exposure to the gestation period only, as occurred in many of the Type II studies. Therefore, the final models were developed based exclusively on the results of Type I developmental toxicity studies.

A total of 21 Type I developmental toxicity studies was identified for the final modeling. The majority of these studies involved test materials that fell into two broad categories: heavy fuel oils and gas oils. Of the 21 developmental toxicity studies used for final modeling, there were 10 and 7 studies of heavy fuel oils and gas oils, respectively. The remaining studies used for final modeling included two aromatic extracts, a lubricating oil basestock, and a petroleum waste. The 21 Type I developmental toxicity studies identified for final modeling included 10 of the 11 studies originally evaluated by Feuston et al. (1994).

### 2.4. Compositional analysis of petroleum substances evaluated for developmental toxicity

The HBPS considered in this paper are complex substances, containing at a minimum, thousands of individual hydrocarbons (Speight, 2007). Consequently, the precise composition of any given HBPS is not known. As a result, all of the materials considered in this report are defined as TSCA (Toxic Substances Control Act) Class II substances (Unknown or Variable Composition, Complex Reaction Products and Biological Materials, i.e., UVCBs).

Several analytical methods were used to broadly characterize the composition of the various test substances in the developmental toxicity studies. Because the results of the various analytical methods are not directly comparable, it was necessary to select studies that used the same analytical methodology to characterize the composition of the test material.

For most of the test substances used in the developmental toxicity studies, a common analytical method, known as the Method II chemical characterization procedure, was employed. The Method II chemical characterization procedure uses DMSO to extract the aromatic-ring compounds in HBPS and quantifies the weight percent of each class of 1–7 and larger aromatic-ring compounds present in the DMSO extract using gas chromatography with flame ionization detection or mass spectrometry to give the Aromatic-ring Class profile (i.e., the ARC profile) (see Gray et al., 2013 for more details on the extraction procedure). For example, using the Method II chemical characterization procedure, the ARC profile of one test substance (i.e., Sample F-179) was 0% of 1-ring, 0.7% of 2-ring, 10% of 3-ring, 30% of 4-ring, 20% of 5-ring, 6% of 6-ring, and 0% of seven and larger-ring substances. The Method II chemical characterization procedure does not differentiate between non-heterocyclic and heterocyclic PACs as it just defines the 1–7 and larger aromatic-ring distributions.

Final models were developed using only Method II-derived compositional data. It is important to recognize that the Method II chemical characterization procedure is not a model; the Method II chemical characterization procedure is the analytical method used to input the compositional data to develop the models. Preliminary modeling found the Method II chemical characterization procedure was more reliable and useful than any of the other analytical techniques (Nicolich et al., 2013). Because the final models were based on the Method II analytical results, it was necessary to exclude a small number of developmental studies from the final modeling because no compositional data using the Method II chemical characterization procedure was available on the test substance. For the models, seven separate values were input for the weight percentages of 1–7 and larger ring substances (i.e., the ARC profile). The models were not based on the total (combined) percent weight of the 1–7 and larger aromatic ring compounds because the total percent weight of all aromatic ring compounds did

not correlate with the dose–response curves. This finding suggests that some ring classes may be more important than other ring classes for purposes of predicting developmental toxicity.

### 2.5. Identification of developmental toxicity endpoints for mathematical modeling

Developmental toxicity studies are designed to evaluate many measures of developmental toxicity. It was recognized that it would be difficult to characterize the relationship between ARC profile and effect for all of the developmental toxicity endpoints which are assessed in studies of standard design. Accordingly, it was decided to identify a smaller number of developmental toxicity endpoints on which to focus the modeling.

Endpoints of developmental toxicity were chosen in both Type I and II studies that would undergo preliminary quantitative assessment for dose–response relationship(s) between ARC profile and developmental effects. The preliminary assessment served two purposes: (1) to identify a smaller number of biological endpoints that would undergo final modeling, and (2) to evaluate the utility for final modeling of different compositional data sets derived by different analytical methods. The selected endpoints were: (1) those most often statistically significantly affected among the developmental toxicity studies from which data had been extracted and therefore, the effects most likely to be associated with exposure to PAC, and (2) the most sensitive endpoints (i.e., most often statistically significantly affected at the LOEL). Results of the preliminary evaluation indicated that models developed using compositional data from the Method II chemical characterization procedure produced the best fit of the data and the most promising approach for final analysis. Final models were developed for the following endpoints of developmental toxicity: percent resorptions per litter, live fetuses per litter, and fetal body weight; these endpoints were expressed as the mean values of the litter means.

### 2.6. Model development and evaluation

Mathematical models, termed the ARC models, were developed to calculate the predicted responses of HBPS on endpoints of developmental toxicity based on its ARC profile. Models were developed independently for each sensitive endpoint of developmental toxicity (i.e., fetal body weight, number of live fetuses per litter, percent resorptions). The numerical values for each endpoint for each dose group in each study were used to build the models. The dose group response was the un-weighted mean of the litter means. The compositional data used to build the models was the ARC profile data using the Method II chemical characterization procedure, as described earlier. Thus, for every test material used to develop the models, there were seven numbers entered into the models to describe the chemical composition of the test material.

Each model was based on least-squares linear regression 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) and the ARC profile of each test material as the independent, or predicting, variables. The models were developed to be as simple as possible, but still adequately describing the data. The final models were selected 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.

The ARC models were based on observed statistical relationships, not on biological knowledge or any presumed mechanism of action. Since the mechanism of action is not understood, the results only indicate that there is an association between the ARC profile and developmental toxicity. The methods used to develop

the mathematical models are described in greater detail by Nicolich et al. (2013).

The final ARC models were used to predict dose–response curves for the three sensitive endpoints of developmental toxicity. The ARC models were designed to predict the effect on a modeled endpoint at a given dose or the dose that causes a given effect. With respect to the latter, the ARC models were designed to predict the dose level that produces a predicted change from the controls, herein termed the Predicted Dose Response<sub>x</sub> (PDR<sub>x</sub>).

Throughout this paper, a 10% change from controls was selected for illustrative purposes. It should not be implied that a 10% response is necessarily the toxicologically appropriate degree of response for the selected endpoints. The ARC models were designed to predict any response level on a dose–response curve. In other words, based on the ARC profile, the models are designed to predict the dose of a specific HBPS that causes any designated magnitude of response for a given endpoint of developmental toxicity.

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 noted by Oreskes et al. (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 *evaluate* or *corroborate* rather than *validate*.

The ARC models were evaluated with four different methods: (1) using holdout sample data (developing the model on a subset of the data and evaluating the model with the remaining “holdout” data), (2) using nonsense data (demonstrating the poor performance of the model when using random or “nonsense” values that were *not* associated with the outcome or observed effect), (3) sensitivity analyses (determining which factors are important in the model and determining model behavior for changes in all inputs and parameters using a distribution of input values), and (4) using new data (described in Section 2.8). The methods and results of these evaluation techniques are described in detail by Nicolich et al. (2013).

### 2.7. Comparison of predicted values with estimates of observed toxicity

In order to demonstrate the predictive value of the PDR<sub>10</sub>, it was necessary to compare the predicted value from the ARC models against the responses observed in developmental toxicity studies. Using the observed data, there are several ways to develop an estimate of the dose associated with a 10% change in the response from the control group.

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<sub>10</sub> 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 four 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 empirical models. Because of the small number of dose groups in the studies that were used in this paper, the available empirical models chosen for current comparisons are the linear or quadratic regressions.

The method used to calculate the BMD<sub>10</sub> values in this paper has been described in detail by Nicolich et al. (2013). The EPA has provided framework of steps to be considered when calculating a BMD (Davis et al., 2011). The steps indicate that the BMD can be calculated if all the following conditions are met:

1. There are adequate data to assess the BMD (the individual “raw” data are available, or the dose group summary statistics needed for model estimation are available).
2. At least one dose group is statistically different from the control group (there is a LOEL) or there is a statistically significant dose trend.
3. At least one of the regression models adequately fits the data.
4. There are three or more dose groups, one of which is the control group.

If conditions 1, 2, or 3 were violated, then the dose associated with a 10% change was estimated from a simple linear regression equation using ordinary least squares, and this value was reported for the “Estimate<sub>10</sub>” value, and the reason for not calculating a BMD was noted. If there were fewer than three dose groups, one of which was the control group, (condition 4), then a professional judgment was made as to whether a 10% change would have occurred below or above the response of the positive dose group used in the study and this estimated value was reported. It sometimes happens in a specific dataset that the response is in the unexpected direction (e.g., fetal pup weight decreases with increasing dose). In the EPA BMD program the user can restrict the response coefficient(s) to have the expected sign(s), but if the response data are in the unexpected direction then no estimate is provided. In this situation no “Estimate<sub>10</sub>” value was reported and the reason noted. In all cases if the “Estimate<sub>10</sub>” value was greater than 2000 mg/kg<sub>bw</sub>/day it was reported as “>2000 mg/kg<sub>bw</sub>/day” to avoid overly precise estimates on materials that are judged to be “non-toxic”.

We chose to define the two values as “consistent” if their relative percent difference was less than 100; this was equivalent to less than a 3-fold difference in the values (Felter and Dourson, 1998). The relative percent difference is defined as 100 times the absolute value of the difference in the two values divided by their average value (US EPA, 2012). For example if the two values were A and B, then the relative percent difference was

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

where the vertical lines represent the absolute value.

For example, if the PDR<sub>10</sub> and Estimate<sub>10</sub> values were 100 and 50 mg/kg/day, respectively, these values would be considered consistent since their relative percent difference of 67 is less than 100.

If one of the values being considered had a greater than or less than sign (> or <), then the value used in calculating the relative percent difference was the minimum value (the number without the ‘greater than’ sign) or was the maximum value (the number without the ‘less than’ sign). For example, if the number was >2000, “2000” was used to calculate the relative percent difference. If either a PDR<sub>10</sub> or Estimate<sub>10</sub> value for an endpoint was missing, no comparison was made.

### 2.8. Methods to corroborate the models with samples not used to build the final models

To further evaluate the model, a subset of the Type II developmental toxicity studies was used to evaluate the model for predicting litter size. There were six Type II developmental toxicity studies of HBPS that were (1) not tested in Type I developmental toxicity studies, (2) not used to build the models, (3) analyzed

for ARC profile using the Method II chemical characterization procedure, and (4) for which both PDR<sub>10</sub> and BMD<sub>10</sub> values were available. In the Type II developmental toxicity studies, the number of live pups per litter was recorded on the day of birth. In comparison, the model was designed to predict the PDR<sub>10</sub> for the number of live fetuses per litter at the time of C-section (gestation day 20). Although it was theoretically possible that additional offspring might die during the couple of days separating gestation day 20 and the day of birth, a gain in litter size is not possible between these events since fetal death is an irreversible effect. In other words, theoretically, an accurate PDR<sub>10</sub> from the data at the time of C-section should not be lower than the BMD<sub>10</sub> from the data at time of the live birth. Therefore, the model could be evaluated by comparing the PDR<sub>10</sub> for live fetuses per litter against the BMD<sub>10</sub> for live pups per litter actually observed in the study.

To further evaluate the models, two samples that were not used to develop the ARC models were recently tested in rat dermal developmental toxicity studies that meet the ARC model requirements. Sample 20906 is a light paraffinic distillate aromatic extract and Sample 120801 is an ultra-low sulfur diesel oil. The PDR<sub>10</sub> values predicted by the final models were compared to the Estimate<sub>10</sub> values observed in these two studies.

## 3. Results

### 3.1. Identification of sensitive endpoints of developmental toxicity

The most sensitive endpoints observed among the developmental toxicity studies of HBPS were endpoints of fetal/pup growth and survival (Table 2).

#### 3.1.1. Type I developmental toxicity studies

The endpoints of developmental toxicity most often affected among the Type I studies (i.e., those that terminated at C-section) included the number of resorptions per litter, the percentage of resorptions per litter, the percentage of dams with resorptions, the number of live fetuses per litter, fetal body weight, and skeletal variants.

**3.1.1.1. Fetal loss and resorptions.** Because the developmental toxicity studies started exposure on GD 0 or earlier (well before implantation occurs on approximately GD 6 in rats), it was possible to evaluate the potential for the test materials to produce pre-implantation loss. Among the submitted studies, a statistically significant increase in pre-implantation loss was observed at the LOEL in only one study. Therefore, the number of implantations per dam and the percent pre-implantation loss were not considered as sensitive endpoints and were not included in the preliminary modeling.

The incidence of resorptions reflects post-implantation loss (i.e., embryonal and fetal death following implantation). The percentage of resorptions (resorption sites/implantation sites) was statistically significantly increased in 66% (19/29) of the Type I studies from which data were extracted. Among the Type I studies used for the final modeling, the percentage of resorptions was statistically significantly increased in 62% (13/21) of studies. The mean number of resorption sites per litter was also reported. The mean number of resorption sites per litter was statistically significantly increased in 62% (18/29) of the Type I studies from which data were extracted and in 57% (12/21) of the Type I studies used for the final model development. The percentage of dams with one or more resorptions was also significantly increased in about half of the studies. All of these metrics appear to be measuring essentially the same effect (i.e., an increase in resorptions).

**Table 2**  
Sensitive endpoints in developmental toxicity studies.

Endpoint <sup>a</sup>	Sensitive Endpoint <sup>b,c</sup>	Chosen for Preliminary Model Development <sup>c</sup>	Good Correlation in Preliminary Modeling <sup>c</sup>	Chosen for Final Model Development <sup>c</sup>
<i>Developmental toxicity studies (Type I)</i>				
Implantation sites/ dam				
% Preimplantation loss				
Resorptions/litter	✓	✓	✓	
% Resorptions/litter	✓	✓	✓	✓
% Dams with resorptions	✓			
Live fetuses/litter	✓	✓	✓	✓
Fetal body wt.	✓	✓	✓	✓
External anomalies				
Visceral anomalies				
Skeletal variants	✓			
<i>Developmental toxicity studies (Type II)</i>				
% Dams delivering litters				
Implantation sites/ dam				
Total pups/litter PND <sup>d</sup> 0	✓	✓	✓	
Live pups/litter PND <sup>d</sup> 0	✓	✓	✓	
Total pups/litter PND <sup>d</sup> 4	✓			
Live pups/litter PND <sup>d</sup> 4	✓			
Pup body wt. PND <sup>d</sup> 0	✓	✓	✓	
Pup body wt. PND <sup>d</sup> 4	✓			

<sup>a</sup> Key endpoints evaluated in the developmental toxicity studies.

<sup>b</sup> In the reviewed studies, endpoint was among those most often affected (statistically significant) and affected (statistically significant) most often at the studies' LOELs (i.e., those effects that would be predictive of a significant biological effect).

<sup>c</sup> Blank cells represent endpoints that were judged "not sensitive" or not chosen for evaluation.

<sup>d</sup> PND, postnatal day.

Both the number of resorptions per litter and the percentage of resorptions per litter were chosen for preliminary modeling. The percentage of dams with resorptions was not chosen because it was statistically significantly affected less frequently and is not as responsive an endpoint (i.e., dams with one resorption or many resorptions are counted the same). For the final modeling, the percentage of resorptions was selected over the number of resorptions per litter because this parameter has the advantage of "normalizing" the results based on the number of implantation sites, which can vary among dams.

**3.1.1.2. Live fetuses per litter.** The number of live fetuses per litter was statistically significantly decreased in 66% (19/29) of the Type I studies from which data were extracted. Among the Type I studies used for the final modeling, the mean number of live fetuses per litter was statistically significantly decreased in 62% (13/21) of the studies.

**3.1.1.3. Fetal body weight.** Fetal body weight was statistically significantly decreased in 66% (19/29) of the Type I developmental toxicity studies from which data were extracted. Among the Type I studies used for final model development, fetal body weight was statistically significantly decreased in 67% (14/21) of the studies. Fetal body weight was chosen for both the preliminary and final modeling.

**3.1.1.4. Fetal malformations.** Malformations were not selected as an endpoint for preliminary or final modeling because, among the Type I studies, increases in malformations were infrequent, seldom statistically significant, and not clearly attributable to any test material. As such, no malformation, individually or collectively,

was identified as an endpoint most often statistically significantly affected in the studies, and no increase in any malformation was an endpoint affected at the study's LOAEL.

**3.1.1.5. Skeletal variants and delayed ossification.** The incidence of fetal skeletal variants (usually delayed ossification) observed at the time of caesarean section was statistically significantly increased in 63% (17/27) of the Type 1 studies from which data were extracted (2/29 studies did not evaluate the fetal skeletons). Among the Type I studies used for the final modeling, the incidence of skeletal variants was statistically significantly increased in 71% (15/21) of the studies. However, this endpoint was not included in the modeling for several reasons:

1. The increased incidences of delayed ossification were highly correlated with decreased fetal body weights in the Type I developmental toxicity studies. An increase in skeletal variations was not observed at lower doses than those that caused decreased fetal body weight. Both fetal body weight and delayed ossification are probably indicators of a similar effect, i.e., an effect on growth. It is unlikely that significant additional information would be gained by adding skeletal variants or delayed ossification to the list of endpoints evaluated in the modeling.
2. The skeletal examination procedures and reporting varied among the studies and among the laboratories. Therefore, it is difficult to compare the incidences of skeletal variants and delayed ossification across different studies and laboratories. In contrast, fetal body weight is easily compared across studies since the method for determining fetal body weight is standardized.

**Table 3**  
preliminary analyses: developmental toxicity endpoints using linear regression models and compositional data determined by method II.<sup>a,b</sup>

Endpoint	Number of dose groups	Multiple correlation coefficient, <i>r</i>	SE <sup>d</sup>
<i>Developmental toxicity studies (Type I)</i>			
Percent resorptions	66	0.98	1.08
Resorptions/litter	66	0.98	1.07
Live fetuses/litter	66	0.98	0.07
Fetal body wt	66	0.95	0.03
<i>Developmental toxicity studies (Type II)</i>			
Total pups/litter	77	0.93	0.09
PND <sup>e</sup> 0			
Live pups/litter	77	0.92	0.10
PND <sup>e</sup> 0			
Pup body wt PND <sup>e</sup> 0	77	0.83	0.04

<sup>a</sup> Analyses based on 57 studies, see Table 1.

<sup>b</sup> Method II chemical characterization procedure, see Section 2.4.

<sup>d</sup> SE, standard error calculated as the square root of the error mean square.

<sup>e</sup> PND, postnatal day.

3. Delayed ossification of skeletal bones is not usually considered a malformation but rather a minor skeletal variation since it is usually reversible and does not affect the quality of life.

### 3.1.2. Type II developmental toxicity studies

The endpoints of developmental toxicity most often statistically significantly affected among the Type II developmental toxicity studies (i.e., those that terminated on PND 4) included decreased litter size and decreased pup body weight. These affected endpoints are consistent with the sensitive endpoints identified in the Type I studies that were chosen for the final modeling. A detailed description of the sensitive endpoints in the Type II developmental toxicity studies is available in the online [Supplemental Information](#).

### 3.1.3. Maternal toxicity

Endpoints of maternal toxicity were not the subject of modeling in this study. However, it is noteworthy that developmental toxicity was strongly associated with maternal toxicity (e.g., decreased maternal body weight, weight gain, and/or food consumption) and skin irritation in both the Type I and II developmental toxicity studies of HBPS, suggesting that the developmental effects may have been secondary to maternal toxicity. However, determining the mechanism of developmental toxicity was beyond the scope of this study. More importantly, the models would be considered to have value regardless of whether the developmental effect was a direct action of the HBPS or an indirect effect of maternal toxicity.

Of note, among the Type I studies, developmental toxicity was never observed in the absence of maternal effects. Further, mater-

nal thymus weights were measured at necropsy in roughly two-thirds of the Type I developmental toxicity studies of HBPS, and decreased maternal thymus weight was observed in about 75% of those studies. Decreased thymus weights typically reflect elevated endogenous corticosterone levels, which is an indicator that the dams were stressed. Every dose level of a HBPS that produced a high ( $\geq 40\%$ ) incidence of resorptions in a Type I developmental toxicity study also produced a significant decrease in maternal thymus weight. Additional details of the extent of maternal toxicity in both the Types I and II studies are provided in the online [Supplemental Information](#).

### 3.2. Preliminary modeling of developmental toxicity

The results of the preliminary modeling using Method II derived compositional data are presented in Table 3. Among the Type I studies, the magnitude of the correlations (*r*) between the values predicted by the preliminary models and the values observed in the studies are large, ranging from 0.95 to 0.98 for the four developmental toxicity endpoints evaluated. The correlation coefficients for the developmental toxicity endpoints evaluated among the Type II studies were high (range: 0.83–0.93), but not as high as those observed in the Type I studies. The results of the preliminary analysis strongly suggested a relationship between ARC profile, as determined by the Method II chemical characterization procedure and the most sensitive endpoints of developmental toxicity identified for HBPS.

### 3.3. Final modeling of developmental toxicity

After completing the preliminary modeling of the dose–response relationships, sensitive endpoints of developmental toxicity were selected for final mathematical characterization. Only Type I developmental toxicity studies were considered for the reasons detailed in Section 2.3.

The final modeling evaluated three sensitive endpoints of developmental toxicity: percent resorptions, number of live fetuses per litter, and fetal body weight. While the number of resorptions per litter also correlated with ARC profile in the preliminary modeling, this endpoint was excluded from the final modeling because fetal survival was evaluated by two other endpoints being modeled (i.e., the percent resorptions and the number of live fetuses per litter).

For the ARC models, the independent (i.e., predicting) variables consisted of relevant study design features, biological parameters (e.g., control group response), and test substance variables (e.g., chemical classes based on the ARC profile), as shown in Table 4. The analyses were based on ordinary least squares (OLS) methods (Draper and Smith, 1998) and the mixed effects model. The forms of the ARC models and the coefficients are described in greater detail by Nicolich et al. (2013).

**Table 4**  
Variables for final models of developmental toxicity.

Dependent variable	Transformation on dependent variable	Covariate (independent biological variable)	Other independent biological variables	Additional Method II terms included <sup>c</sup>
Fetal body weight	None	CG <sup>a</sup> fetal body weight	None	1 <sup>b</sup>
Live fetuses/litter	None	CG <sup>a</sup> live fetuses/litter	N implants	1 <sup>b</sup>
Percent resorptions	None	CG <sup>a</sup> percent resorptions	None	1 <sup>b</sup>

<sup>a</sup> CG, control group.

<sup>b</sup> Interaction term of the form  $\sum_{j=1}^5 \xi_j \cdot \text{dose} \cdot \text{ARC}_4 \cdot \text{ARC}_5 \cdot \text{ARC}_j$ .

<sup>c</sup> Method II chemical characterization procedure, see Section 2.4.

**Table 5**  
Final models: correlation between observed and predicted values for developmental toxicity endpoints.<sup>a</sup>

Dependent variable	Studies (n)	Data points (n)	OLS <sup>b</sup>		Mixed effects model	
			Correlation coefficient (r)	Standard error (SE)	Correlation coefficient (r)	Standard error (SE)
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

<sup>a</sup> Based on the 21 samples used to build the final models, see Table 1.

<sup>b</sup> OLS, ordinary least squares.

The results of the final modeling are presented in Table 5. The magnitude of the correlations (*r*) between the values predicted by the ARC models and the values observed in the studies are large, ranging from 0.94 to 0.99. These results indicate that for those samples used to build the models for developmental toxicity endpoints, the predicted and actual values were highly correlated. In contrast to the preliminary models, the ARC models were based on actual observed responses, not on the relative response (i.e., ratio) vs. the control group. Therefore, the *r* and SE values from the ARC models (Table 5) cannot be directly compared with those generated under the preliminary models (Table 3).

The accuracy of the fit of the models can best be seen in plots of the observed data points vs. the predicted data points (Fig. 1). The optimum model would have all the points along the 45 degree line representing equal values of the observed and predicted data. The ARC models show an excellent fit for all three endpoints of developmental toxicity.

#### 3.4. Use of the models to predict developmental toxicity

The ARC models can be used to generate dose–response predictions. Fig. 2 shows the results of using the model to generate dose–response curves for percent resorptions for two different HBPS samples with different ARC profiles (i.e., Sample 86001 [heavy fuel oil] and Sample 86187 [distillate aromatic extract]). The ARC profiles of these two substances are shown in Table 6. Based on the control group data from the 21 Type I developmental toxicity studies used in the ARC model, the background incidence of resorptions is assumed to be 6.85% on average. The curves are generated by using the equation for percent resorptions (Fig. 1), along with the ARC profile (derived using the Method II chemical characterization procedure), the coefficients for percent resorptions, and the average percent resorptions among the control groups. At a dose of 50 mg/kg<sub>bw</sub>/day, the estimates of percent of resorptions are predicted to be about 100% and 28% for Samples 86001 and 86187, respectively. Repeating this calculation for different dose levels produces the two dose–response curves in Fig. 2.

Fig. 3 presents an example of how the model may be used to predict the dose level that produces a 10% increase in the percent resorptions relative to controls (PDR<sub>10</sub>) using the same sample of distillate aromatic extract that appears in Fig. 2. To determine the predicted percent resorptions relative to controls, each of the values in Fig. 2 is divided by the corresponding predicted control value, and then multiplied by 100. In this example, the critical value is the control incidence (4.86%) plus 10% of the difference between the control value and 100%. The critical value for a 10% increase in the percent resorptions is then 14.37%. Thus, the response (percent resorptions) relative to the control value would be 296% (i.e., 14.37/4.86). For this sample of distillate aromatic extract, a 10% increase in resorptions (which is 296% of the control value) is predicted to occur at a dose of approximately 27 mg/kg<sub>bw</sub>/day, with 95% confidence intervals of approximately 17 and 39 mg/kg<sub>bw</sub>/day (Fig. 3).

#### 3.5. Comparison with existing predictive methods for samples used to build the ARC models

For the 21 studies used to build the developmental toxicity ARC models, Table 7 provides comparisons of the dose associated with a 10% change in response from the control value derived using either the ARC model (PDR<sub>10</sub>) or the observed data (Estimate<sub>10</sub>) for studies that have the appropriate observed data. Table 7 indicates that the ARC models generate values that are consistent with other standard measures.

To avoid overly precise estimates on materials that are judged to be relatively inactive, PDR<sub>10</sub> values that are greater than 2000 mg/kg<sub>bw</sub>/day are shown in the Table 7 as “>2000 mg/kg<sub>bw</sub>/day.” Values that are extrapolations of the doses from the studies used to build the models are noted. Also noted are values based on model predictions for which the dose–response slopes (1) are not in the appropriate direction, i.e., a direction inconsistent with the expected biological effect on the specific endpoint based on results of reviewed studies, or (2) are nearly flat, i.e., the magnitude of the slope is small (less than or equal to the absolute value of control value divided by 10,000). The choice of the control values divided by 10,000 is somewhat arbitrary and corresponds to an approximate 20% change from control at a dose of 2000 mg/kg<sub>bw</sub>/day. For the three developmental toxicity endpoints, the appropriate directions for an adverse effect on the dose–response slopes are: an increase for the percent resorptions and decreases for fetal body weight and live fetuses per litter. PDR<sub>10</sub> values were considered unreliable and are not reported when the slope of the model prediction is large (greater than the absolute value of the control value divided by 10,000).

As shown in Table 7, there was consistency between the PDR<sub>10</sub> and the Estimate<sub>10</sub> for fetal body weight and live fetuses per litter in 100% (17/17 and 16/16, respectively) of the studies for which a comparison could be made. For the percent resorptions, there was consistency between the PDR<sub>10</sub> and the Estimate<sub>10</sub> in 87% (13/15) of the studies for which a comparison could be made. The two exceptions were Samples 86193 and F-199. In both cases, the Estimate<sub>10</sub> for the percent resorptions was well above the highest dose level tested, and the Estimate<sub>10</sub> was based on a projection from a slight (not statistically significant) increase in resorptions at the highest dose tested. Thus, the apparent inconsistency in the values for percent resorptions observed with two studies may simply reflect a less reliable Estimate<sub>10</sub> value for this endpoint due to the choice of dose levels in these two studies. It is important to recognize that in both cases, the model erred on the side of safety by over-predicting the effect of the test substance on the percent resorptions.

#### 3.6. Comparison with existing predictive methods for samples not used in building the ARC models

There were six Type II developmental toxicity studies of HBPS that were (1) not tested in Type I developmental toxicity studies, (2) not used to build the ARC models, (3) analyzed for ARC profile

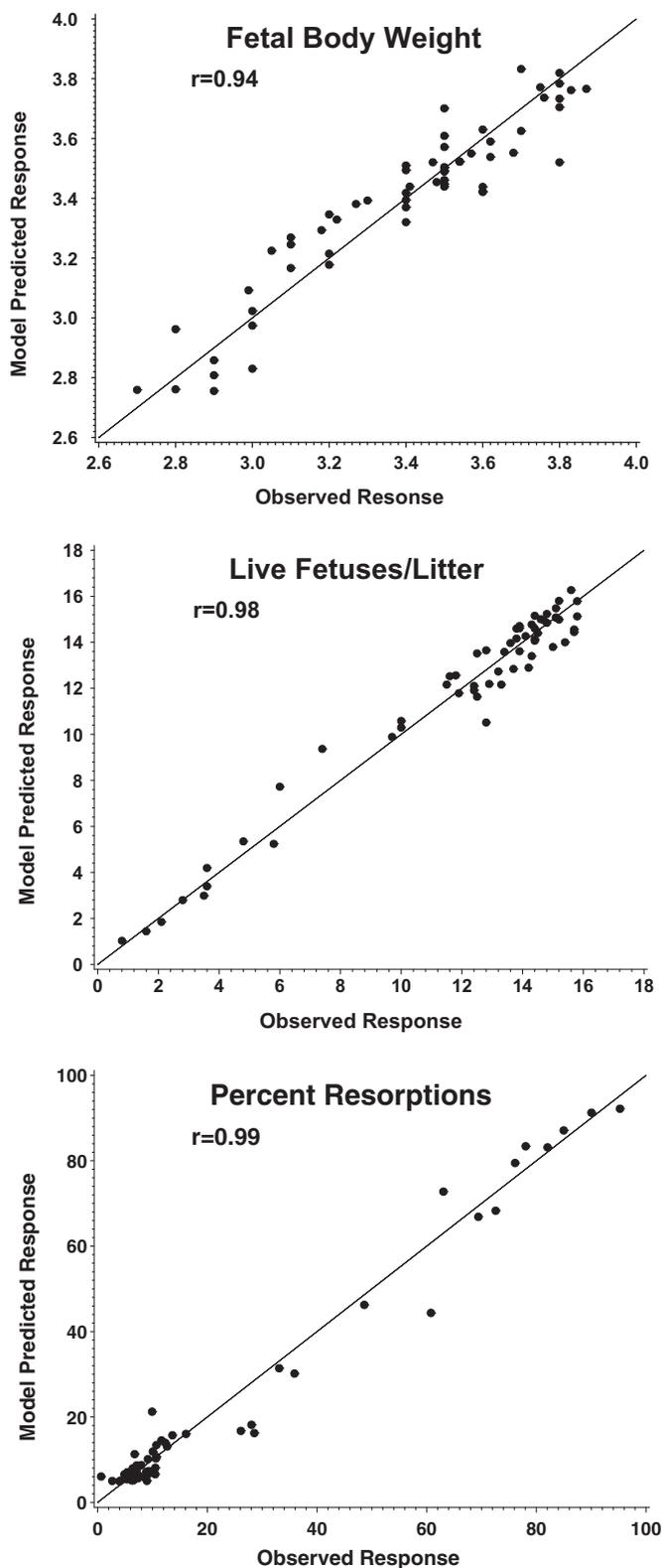


Fig. 1. Plots of observed vs. predicted values for percent resorptions, live fetuses per litter, and fetal body weight.

using the Method II chemical characterization procedure, and (4) for which both  $PDR_{10}$  and  $BMD_{10}$  values were available, as shown in Table 8. The  $PDR_{10}$  and  $BMD_{10}$  values were consistent, in five of six test samples. The exception, Sample F-220, had a  $PDR_{10}$  lower than the observed  $BMD_{10}$ ; thus, the model predicted that F-220 was more effective in reducing litter size than was actually ob-

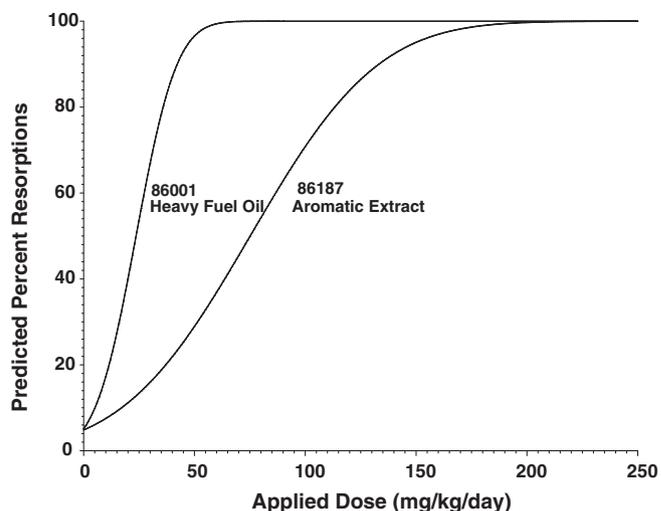


Fig. 2. Predicted dose–response curves for percent resorptions for two samples with different ARC profiles.

served. Overall, the results in Table 8 provide further corroboration of the model for the number of live fetuses per litter.

To further evaluate the ARC models, two samples that were not used to develop the ARC models were recently tested in standard rat developmental toxicity studies that meet the ARC model requirements. The sample profiles are interpolations, and the samples boil  $\geq$  approximately 650 °F. Sample 20906 is a light paraffinic distillate aromatic extract and Sample 120801 is an ultralow sulfur diesel oil. Table 9 provides the summary results from the experiment and the  $PDR_{10}$  prediction in a format similar to the data and comparisons in Table 7.

The light paraffinic distillate aromatic extract (Sample 20906) was toxic to the developing embryo as evidenced by statistically significant decreases in the number of live fetuses per litter and fetal body weight and increases in the percent resorptions at doses of 150 mg/kg<sub>bw</sub>/day or greater; a slight (5%), but a statistically significant, decrease in fetal body weight was also observed at 25 mg/kg<sub>bw</sub>/day. In comparison, the ultralow sulfur diesel oil sample (120801) was not associated with any statistically significant changes of any of the parameters identified in Table 9 at doses up to 600 mg/kg<sub>bw</sub>/day, the highest dose group. The ARC model for fetal body weight accurately predicted that Sample 120801 would have no effect on fetal body weight at to the highest dose tested, i.e., 600 mg/kg<sub>bw</sub>/day, but the model underestimated the effect of Sample 20906 on fetal body weight. The remaining estimates are listed as unreliable because the model predictions for both samples was the reverse of what was expected (increasing number of live fetuses per litter and decreasing percent resorptions with increasing dose).

Because of the degree of disagreement between the observed and predicted values, these sample predictions are not adequate for corroboration. The positive outcome is that they are not predicting false negatives; they are simply not providing reliable estimates. Both of the new test materials have ARC profiles that are interpolations. An interpolated profile has the characteristic that each of the seven ARC values is numerically between the corresponding ARC values of two substances used to develop the model. A more detailed discussion of the concepts is provided by Nicolich et al. (2013). The assumption is that if a profile is an interpolation, the model prediction will be accurate because the profile is surrounded by the profiles of samples used to develop the models. This assumption tacitly assumes the relations are linear, like the points on a line, and if a test number is greater than specific number and less than a third specific number the test number is “between” the other two values. However, the relations are not

**Table 6**  
ARC profiles of samples used to build or evaluate final developmental toxicity models.

Sample No.	ARC profile <sup>a</sup>							Sample used to
	1-ring, Wt%	2-ring, Wt%	3-ring, Wt%	4-ring, Wt%	5-ring, Wt%	6-ring, Wt%	7-ring, Wt% <sup>b</sup>	
60901	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Build final models <sup>c</sup>
8281	2.0	29.5	14.7	0.0	0.5	0.5	0.0	Build final models <sup>c</sup>
83366	0.1	2.5	5.1	2.5	1.3	0.9	0.1	Build final models <sup>c</sup>
85244	0.0	0.1	2.5	1.9	1.2	0.5	0.0	Build final models <sup>c</sup>
86001	0.0	2.6	25.7	19.3	6.4	3.2	0.6	Build final models <sup>c</sup>
86181	0.2	2.5	12.4	7.4	2.5	0.5	0.0	Build final models <sup>c</sup>
86187	0.0	0.0	4.1	8.1	6.1	2.0	0.4	Build final models <sup>c</sup>
86193	0.8	2.9	0.4	0.0	0.0	0.0	0.0	Build final models <sup>c</sup>
86270	0.9	2.6	3.5	0.9	0.4	0.0	0.4	Build final models <sup>c</sup>
86271	0.1	0.8	5.3	3.2	0.4	0.2	0.1	Build final models <sup>c</sup>
86484	0.0	1.0	9.8	19.5	9.8	4.9	1.0	Build final models <sup>c</sup>
87213	0.1	4.2	6.3	0.3	0.0	0.0	0.0	Build final models <sup>c</sup>
87476	0.0	0.0	0.0	0.1	0.3	0.5	1.6	Build final models <sup>c</sup>
89106	0.2	1.2	1.7	1.2	0.6	0.5	0.0	Build final models <sup>c</sup>
F-179	0.0	0.7	10.0	30.0	20.0	6.0	0.0	Build final models <sup>c</sup>
F-193	0.0	2.0	4.0	2.0	0.7	0.2	0.0	Build final models <sup>c</sup>
F-195	0.1	3.0	4.0	0.1	0.1	0.1	0.0	Build final models <sup>c</sup>
F-196	0.1	0.3	3.0	2.0	2.0	0.7	0.0	Build final models <sup>c</sup>
F-197	0.0	0.4	4.0	2.0	0.6	0.2	0.0	Build final models <sup>c</sup>
F-199	0.1	4.0	10.0	0.0	0.0	0.0	0.0	Build final models <sup>c</sup>
F-215	0.2	4.0	4.0	0.0	0.0	0.0	0.0	Build final models <sup>c</sup>
F-194	0.1	4.0	4.0	0.0	0.0	0.0	0.0	Evaluate final models <sup>d</sup>
F-200	0.0	0.9	20.0	5.0	0.0	0.0	0.0	Evaluate final models <sup>d</sup>
F-201	0.1	0.4	4.0	3.0	0.9	0.4	0.0	Evaluate final models <sup>d</sup>
F-220	0.9	2.0	0.9	0.2	0.1	0.0	0.0	Evaluate final models <sup>d</sup>
F-225	0.0	0.4	4.0	1.0	0.4	0.1	0.0	Evaluate final models <sup>d</sup>
F-227	0.1	0.7	3.0	2.0	1.0	0.3	0.0	Evaluate final models <sup>d</sup>
20906	0.0	0.0	5.4	6.8	1.4	0.3	0.0	Evaluate final models <sup>d</sup>
120801	0.1	2.2	0.6	0.0	0.0	0.0	0.0	Evaluate final models <sup>d</sup>

<sup>a</sup> 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 (see Section 2.4).

<sup>b</sup> The ARC 7 value is the weight percent of the seven and larger aromatic-ring compounds within the petroleum substance as determined by the Method II chemical characterization procedure (see Section 2.4).

<sup>c</sup> See Section 3.4.

<sup>d</sup> See Section 3.6.

quite linear because there are seven ARC values in the profile, and the space behaves more like a bent sheet of paper than as a straight line; so the concept of “between” is not easy to define.

The ARC 6 values for the two new materials are low (0.3 and 0.0 for 20906 and 12080, respectively). The samples used to develop the developmental toxicity ARC models, and that are the basis of the interpolation characteristic, have ARC 6 values that are larger (0.5, 3.2, 4.9, and 6.0). Consequently, it is hypothesized that the current ARC models do not predict well when the ARC 6 value is low. If the ARC value for either sample is numerically increased to 1.0 then the resulting ARC model has the predictions in the expected direction as the dose increases.

The evaluation samples may not have provided useable results because the ARC profile was in an area that was poorly represented by the samples used to develop the ARC models for developmental toxicity. At a later time, if the ARC models are updated with these, and other, samples it is expected that the predictions will improve. In contrast, these samples were used to corroborate the four repeat-dose ARC models (Roth et al., 2013) and the samples used to develop the repeat-dose ARC models had ARC profiles similar to these two new substances. In comparison, the repeat-dose predictions seen in Roth et al. (2013) were very good, supporting the idea that the current ARC model is not adequate to predict the developmental toxicity of some of the HBPS because of the limited number of sample patterns of the materials used to develop the developmental toxicity ARC models.

#### 4. Discussion

This study confirms and extends the findings of Feuston et al. (1994), who reported a correlation between the PAC concentra-

tions of petroleum streams and endpoints of developmental toxicity using a Spearman rank-order test. Using a larger data set and a more sophisticated statistical approach, the present study also showed an association between the ARC profile of HBPS and several sensitive endpoints of developmental toxicity. As noted by Feuston et al. (1994) and confirmed in the present investigation, HBPS affect mainly endpoints of fetal survival and growth in developmental toxicity studies in rats.

Feuston and Mackerer (1996a) reported increased malformations among the offspring of pregnant rats given a single, large, oral (gavage) dose of several HBPS. By administering the test material at a higher dose for a shorter period of time, it was possible to reduce the embryo-lethality and demonstrate the potential of these substances to produce malformations. In contrast, daily dermal administration of HBPS to pregnant rats throughout the critical period of gestation produced no conclusive evidence of malformations, even at high dose levels (but not as high as in the gavage studies). However, developmental effects were seen at considerably lower doses in the dermal studies compared to the single dose gavage studies. Based on the dermal developmental toxicity studies of HBPS, embryo-lethality and decreased fetal body weight are more likely outcomes than malformations for these substances in conventional developmental toxicity studies, particularly at the LOAELs.

This study demonstrates that mathematical models provide a promising tool to predict the developmental toxicity of HBPS based on ARC profile. The results of this investigation are consistent with the hypothesis that the ARC profile is the primary determinant of developmental toxicity for this class of petroleum substances. However, further refinement of the developmental toxicity models is recommended before the models can be applied with confidence

**Table 7**  
Comparison of PDR<sub>10</sub> and Estimate<sub>10</sub> values for developmental endpoints for samples used to build the final models.

Sample No.	Decreased fetal body weight			Decreased live fetuses/litter			Increased percent resorptions		
	PDR <sub>10</sub> <sup>a</sup> (mg/kg <sub>bw</sub> /day)	Estimate <sub>10</sub> <sup>b</sup> (mg/kg <sub>bw</sub> /day)	PDR <sub>10</sub> & Estimate <sub>10</sub> consistent <sup>c</sup>	PDR <sub>10</sub> <sup>a</sup> (mg/kg <sub>bw</sub> /day)	Estimate <sub>10</sub> <sup>b</sup> (mg/kg <sub>bw</sub> /day)	PDR <sub>10</sub> & Estimate <sub>10</sub> consistent <sup>c</sup>	PDR <sub>10</sub> <sup>a</sup> (mg/kg <sub>bw</sub> /day)	Estimate <sub>10</sub> <sup>b</sup> (mg/kg <sub>bw</sub> /day)	PDR <sub>10</sub> & Estimate <sub>10</sub> consistent <sup>c</sup>
60901	>2000 <sup>d</sup>	>1000 <sup>e</sup>	Yes	>2000 <sup>d</sup>	>1000 <sup>e</sup>	Yes	– <sup>f</sup>	<1000 <sup>e</sup>	–
8281	579 <sup>g</sup>	720	Yes	420	333 <sup>h</sup>	Yes	665 <sup>g</sup>	698	Yes
83366	126	127	Yes	26	26 <sup>h</sup>	Yes	48	26	Yes
85244	518	622	Yes	147	168 <sup>h</sup>	Yes	257	170	Yes
86001	14	13	Yes	5	4 <sup>h</sup>	Yes	8	4	Yes
86181	51	66	Yes	15	31	Yes	28	35	Yes
86187	82	85	Yes	15	15 <sup>h</sup>	Yes	26	17	Yes
86193	430	– <sup>i</sup>	–	102	250	Yes	176	682	No
86270	372	421	Yes	109	10 <sup>h</sup>	Yes	194	67	Yes
86271	222	233	Yes	80	67 <sup>h</sup>	Yes	137	87	Yes
86484	16	19	Yes	4	4 <sup>h</sup>	Yes	7	4	Yes
87213	>2000 <sup>g</sup>	– <sup>i</sup>	–	416 <sup>g</sup>	– <sup>i</sup>	–	>2000 <sup>g</sup>	– <sup>i</sup>	–
87476	>2000	>2000	Yes	1924	– <sup>i</sup>	–	>2000	– <sup>i</sup>	–
89106	271	223	Yes	59	64 <sup>h</sup>	Yes	111	62	Yes
F-179	2 <sup>g</sup>	– <sup>i</sup>	–	<1 <sup>g</sup>	– <sup>i</sup>	–	1 <sup>g</sup>	– <sup>i</sup>	–
F-193	447	336	Yes	169	148 <sup>h</sup>	Yes	253	184	Yes
F-195	>2000 <sup>g</sup>	1668 <sup>l</sup>	Yes	259	– <sup>i</sup>	–	723 <sup>g</sup>	– <sup>i</sup>	–
F-196	518 <sup>g</sup>	295	Yes	125	106 <sup>h</sup>	Yes	238	222	Yes
F-197	478	301	Yes	162	224 <sup>h</sup>	Yes	283	352 <sup>j</sup>	Yes
F-199	>2000 <sup>g</sup>	– <sup>i</sup>	–	108	– <sup>i</sup>	–	357	1111	No
F-215	>2000 <sup>d</sup>	>2000 <sup>l</sup>	Yes	1413 <sup>g</sup>	>2000 <sup>l,h</sup>	Yes	– <sup>f</sup>	– <sup>i</sup>	–

<sup>a</sup> ARC model values; all results greater than 2000 are reported as >2000.

<sup>b</sup> Dose estimated to cause a 10% change from control value, derived using the observed data from existing toxicity study. Unless otherwise noted, the value represents a BMD<sub>10</sub> calculated using the EPA method (see Nicolich et al., 2013). All results greater than 2000 are reported as >2000.

<sup>c</sup> PDR<sub>10</sub> and Estimate<sub>10</sub> values are considered consistent if the relative percent difference between the PDR<sub>10</sub> and the Estimate<sub>10</sub> is less than 100 (see Section 2.7).

<sup>d</sup> Model predicted dose–response slope is not in the appropriate direction, but the slope is less than or equal to the absolute value of control value divided by 10,000 (see Section 3.6) >2000 reported.

<sup>e</sup> Only two dose groups (control and dosed group); Estimate<sub>10</sub> value reported as the dose range in which a 10% is likely to occur.

<sup>f</sup> Unreliable prediction, no value reported because the model predicted dose–response slope is not in the appropriate direction, and the slope is not negligible (i.e., slope is greater than the absolute value of the control value divided by 10,000) (see Section 3.6).

<sup>g</sup> The PDR<sub>10</sub> is greater than the highest observed dose used to develop the ARC model; PDR<sub>10</sub> reported.

<sup>h</sup> Poor model fit or no SD available; Estimate<sub>10</sub> is from a simple linear regression from the toxicity study data (see Section 2.7).

<sup>i</sup> Observed data response is in a direction inconsistent with the expected biological effect for the specific endpoint (see Section 3.6); No value reported.

<sup>j</sup> No statistically significant change was seen in any dose group of the study; Estimate<sub>10</sub> is from a simple linear regression from the toxicity study data.

**Table 8**  
Comparison of PDR<sub>10</sub> for live fetuses per litter on GD 20 vs. BMD<sub>10</sub> for live pups per litter observed at birth (PND 0) in Type II developmental toxicity studies of substances not used to build the final model.

Sample No.	Live Fetuses per litter PDR <sub>10</sub> mg/kg-d	Live Pups per litter BMD <sub>10</sub> mg/kg-d	PDR <sub>10</sub> and BMD <sub>10</sub> consistent <sup>a</sup>
F-200	27	31	Yes
F-220	73	>250 <sup>b</sup>	No
F-201	91	52	Yes
F-227	207	77	Yes
F-225	224	240	Yes
F-194	>2000	>250 <sup>b</sup>	Yes

<sup>a</sup> The results are considered to be consistent if the BMD<sub>10</sub> value on the day of birth is not markedly higher than the PDR<sub>10</sub> value on GD 20. In comparison, the BMD<sub>10</sub> value could be lower than the PDR<sub>10</sub> value because the model predicts litter size at GD 20 and additional offspring could have died after GD 20 through the day of birth.

<sup>b</sup> This dose was the highest dose level in the study, and no statistically significant effect on fetal body weight was observed. Therefore, the BMD<sub>10</sub> is greater than this value. It may be substantially higher than this value.

to all HBPS. 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 as indicating only that there is an association between developmental toxicity and ARC profile. The models should not be used to draw conclusions about whether any of the specific aromatic ring structures, individually or collectively, is the cause of the observed developmental toxicity.

Efforts to corroborate the models produced mixed results. Corroboration with six Type II developmental toxicity studies, which were not used to build the ARC models, showed consistency between predicted and observed values. However, efforts to corroborate the ARC models using recently-conducted developmental

toxicity studies of two samples of HBPS that were not used to build the ARC models proved disappointing. The poorer predictive value for these two HBPS might be related to their PAC profiles (particularly the relatively low ARC 6 concentrations), which was in an area poorly represented by the samples used to build the ARC models. If the ARC models are updated with these and other sample results, the predictions may improve. In addition, because the ARC models were mathematically-derived without biological input, it was possible for the ARC models to predict a response in the wrong (and biologically improbable) direction (e.g., increases in fetal body weight and the number of live fetuses per litter). In the future, the ARC models could be improved by providing biological input that informs the ARC models that responses in the wrong direction are not biologically plausible. In the meantime, caution should be

**Table 9**  
Comparison of estimate<sub>10</sub> and PDR<sub>10</sub> values for evaluation samples.

Sample	Decreased fetal body weight			Decreased live fetuses/litter			Increased percent resorptions		
	PDR <sub>10</sub> <sup>a</sup> (mg/kg <sub>bw</sub> /day)	Estimate <sub>10</sub> <sup>b</sup> (mg/kg <sub>bw</sub> /day)	PDR <sub>10</sub> & Estimate <sub>10</sub> consistent <sup>c</sup>	PDR <sub>10</sub> <sup>a</sup> (mg/kg <sub>bw</sub> /day)	Estimate <sub>10</sub> <sup>b</sup> (mg/kg <sub>bw</sub> /day)	PDR <sub>10</sub> & Estimate <sub>10</sub> consistent <sup>c</sup>	PDR <sub>10</sub> <sup>a</sup> (mg/kg <sub>bw</sub> /day)	Estimate <sub>10</sub> <sup>b</sup> (mg/kg <sub>bw</sub> /day)	PDR <sub>10</sub> & Estimate <sub>10</sub> consistent <sup>c</sup>
20906	420 <sup>d</sup>	117 <sup>e</sup>	No	– <sup>f</sup>	42 <sup>e</sup>	–	– <sup>f</sup>	47 <sup>e</sup>	–
120801	>2000 <sup>d</sup>	>2000 <sup>g,e</sup>	Yes	– <sup>f</sup>	– <sup>h</sup>	–	– <sup>f</sup>	– <sup>h</sup>	–

<sup>a</sup> ARC model values (PDR<sub>10</sub> values); all results greater than 2000 are reported as >2000.

<sup>b</sup> Dose estimated to cause a 10% change from control value, derived using the observed data from existing toxicity study. Unless otherwise noted, the value represents a BMD<sub>10</sub> calculated using the EPA method (see Section 2.7). All results greater than 2000 are reported as >2000.

<sup>c</sup> PDR<sub>10</sub> and Estimate<sub>10</sub> are considered consistent if the relative percent difference between the PDR<sub>10</sub> and the Estimate<sub>10</sub> is less than 100 (see Section 2.7).

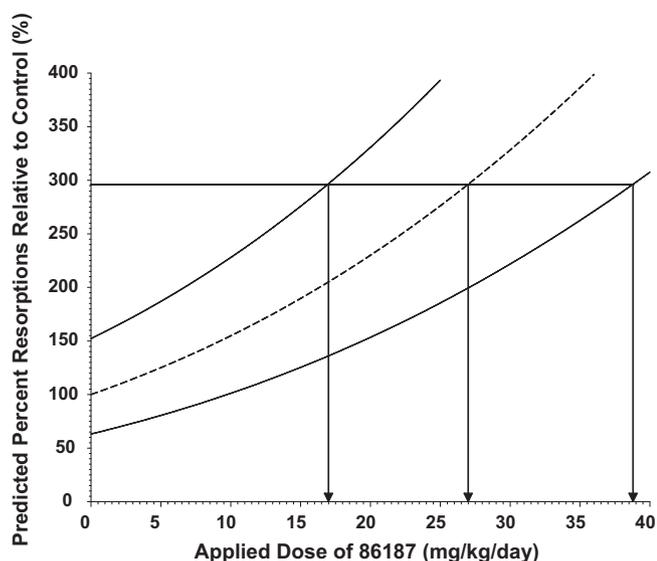
<sup>d</sup> The PDR<sub>10</sub> value is greater than the highest observed dose used to develop the ARC model; PDR<sub>10</sub> value reported.

<sup>e</sup> Poor model fit or no SD available; Estimate<sub>10</sub> is from a simple linear regression from the toxicity study data (see Section 2.7).

<sup>f</sup> Unreliable prediction, no value reported because the model predicted dose–response slope is not in the appropriate direction, and the slope is not negligible (i.e., slope is greater than the absolute value of control value divided by 10,000) (see Section 3.6).

<sup>g</sup> No statistically significant change was seen in any dose group of the study; Estimate<sub>10</sub> is from a simple linear regression from the toxicity study data (see Section 2.7).

<sup>h</sup> Observed data response is in a direction inconsistent with the expected biological effect for the specific endpoint (see Section 3.6); no value reported.



**Fig. 3.** Prediction of the PDR<sub>10</sub> and 95% CI for percent resorptions for a sample of a distillate aromatic extract (dashed response line is prediction line, solid response lines are 95% CI).

exercised in applying the ARC models for developmental toxicity to all HBPS, particularly those with generally high ARC profiles, but low ARC 6 levels.

With additional refinement, the ARC models may provide a more viable alternative to developmental toxicity testing in animals. The 110 HBPS sponsored by the petroleum industry are complex substances, containing at least thousands of components. In addition, the composition of HBPS can vary substantially, even among substances with the same CAS number. It is not feasible to conduct developmental toxicity testing on such a large number of substances.

The ARC models may prove valuable in characterizing the developmental toxicity of sponsored substances under the US EPA HPV Challenge Program, obviating the need for extensive testing for developmental toxicity. The ARC models also show promise as a tool that might also be used to maximize limited resources for toxicity testing. For example, possible uses might include prioritization of substances for developmental toxicity testing and the prediction of appropriate dose levels for study protocols. Roth et al. (2013) provide a discussion of how the ARC models might be used to fulfill regulatory needs.

In general, the utility of a computational model may expand as the dataset upon which it is based enlarges. Further, it may be possible to improve the ARC models as understanding of the underlying cause(s) and mechanism(s) of action are elucidated and applied to the ARC models. It may also prove feasible to simplify the ARC models with further data, since the current models are relatively complex, requiring multiple variables and coefficients.

Based on the results of the ARC models, it does not appear necessary to identify specific PAC substances in a HBPS in order to predict the developmental toxicity of the HBPS. The ARC models use the concentration of each ring class (i.e., the ARC profile), rather than the total weight% of ARC, PAC, or any subset of ring classes (e.g., 4–6 or 3–7 ring ARCs). However, the information on the weight% of ring classes is not chemical-specific. For example, two test materials may have exactly the same percentage of 5-ring substances, but the specific structures of the 5-ring substances may vary between the two test materials. Therefore, the ARC models do not rely on the specific identity of individual PAC substances to predict the developmental toxicity of HBPS. Since all of these materials are derived from crude oil, the key determinants of chemical composition are the feedstock, temperature, and process. Accordingly, the specific PAC ring structures may well be related to the weight percentage of 1–7 and larger aromatic-ring substances. In other words, the weight% of 1–7 and larger aromatic-ring compounds (ARC profile) may be a marker for specific PAC compounds (or other compounds) that are responsible for the developmental toxicity of HBPS.

These ARC models are based on the results of developmental toxicity studies in which the HBPS were applied dermally to pregnant rats. The results are considered to be relevant to humans since dermal contact is the most likely route of human exposure to these types of petroleum substances. It should not be presumed that the ARC models based on dermal contact would be directly applicable to other routes of exposure. However, to the extent that it is possible to quantify differences in systemic dose from dermal application versus other routes of exposure, it may be possible to determine route-specific adjustment factors and to use the ARC models in the future for other routes of exposure.

Although the petroleum substances were applied dermally in all of the developmental toxicity studies used for preliminary and final ARC modeling, the method of dermal application varied among studies. For example, in some studies, the test material was applied undiluted; in others, acetone was used as a vehicle. The test material was wiped from the skin after 6 h in some studies, whereas in others, no attempt was made to remove the test material. These differences in the method of dermal application are a potential source of variability among the study results. During the preliminary

model development, differences in the method of dermal administration were evaluated, and they did not appear to be a significant factor. However, it is possible that physical factors, such as the form or viscosity of the material, may limit the applicability of the ARC models for certain HBPS.

The ARC models are designed specifically for predicting the developmental toxicity of HBPS in rats administered the test material throughout most of gestation. The ARC models are not expected to accurately predict the dose level at which a specific level of response occurs when the test material is given for a shorter duration of exposure, such as a single day or a few days during pregnancy. Also, the ARC models are not expected to predict the developmental toxicity of HBPS in other species of laboratory animals, such as mice or rabbits, since the ARC models were developed based exclusively on data from developmental toxicity studies in rats.

The modeled response level for a developmental effect chosen for this paper was a 10% change from controls. A 10% response was chosen for illustrative purposes to demonstrate the utility of the models, and it was not assumed that a 10% response rate was the toxicologically appropriate degree of response for the selected endpoints. The ARC models are designed to predict any magnitude of response along the dose–response curve, and the ARC models are expected to have similar utility for other selected magnitudes of response.

## 5. Conclusion

A relationship between sensitive endpoints of developmental toxicity and the substance's weight percent of each of the 1–7 and larger aromatic-ring compounds (the "ARC profile") was demonstrated for HBPS. Predictive ARC models based on these associations were developed for effects on three developmental toxicity endpoints (percent resorptions, live fetuses per litter, and fetal body weight). Such associations provide a promising approach for predicting the developmental toxicity of untested HBPS. However, further development and refinement of the ARC models is recommended before they can be reliably applied to all HBPS.

## 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.2013.05.003>.

## References

- Altgelt, K.H., Boduszynski, M.M., 1994. *Composition and Analysis of Heavy Fractions*. Marcel Dekker, Inc.
- Archibong, A.E., Inyang, F., Ramesh, A., Greenwood, M., Nayyar, T., Kopsombut, P., Hood, D.B., Nyanda, A.M., 2002. Alteration of pregnancy related hormones and fetal survival in F-344 rats exposed by inhalation to benzo(a)pyrene. *Reprod. Toxicol.* 16, 801–808.
- 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. *Poly Arom. Comp.* 11, 201–210.
- Bui, Q.Q., Tran, M.B., West, W.L., 1986. A comparative study of the reproductive effects of methadone and benzo(a)pyrene in the pregnant and pseudopregnant rat. *Toxicology* 42 (2–3), 195–204.
- Choi, H., Jedrychowski, W., Spengler, J., Camann, D.E., Whyatt, R.M., Rauh, V., Tsai, W.Y., Perera, F.P., 2006. International studies of prenatal exposure to polycyclic aromatic hydrocarbons and fetal growth. *Environ. Health Perspect.* 114 (11), 1744–1750.
- Choi, H., Rauh, V., Garfinkel, R., Tu, Y., Perera, F.P., 2008. Prenatal exposure to airborne polycyclic aromatic hydrocarbons and risk of intrauterine growth restriction. *Environ. Health Perspect.* 116 (5), 658–665.
- 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 U.S. 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, NY.
- Felter, S., Dourson, M., 1998. The inexact science of risk assessment (and implications for risk management). *Hum. Ecol. Risk Assess.* 4 (2), 245–251.
- Feuston, M.H., Mackerer, C.R., 1996a. Developmental toxicity of clarified slurry oil, syntower bottoms, and distillate aromatic extract administered as a single oral dose to pregnant rats. *J. Toxicol. Environ. Health* 49, 45–66.
- Feuston, M.H., Mackerer, C.R., 1996b. Developmental toxicity study in rats exposed dermally to clarified slurry oil for a limited period of gestation. *J. Toxicol. Environ. Health* 49 (2), 207–220.
- Feuston, M.H., Kerstetter, S.L., Singer, E.J., Mehlman, M.A., 1989. Developmental toxicity of clarified slurry oil applied dermally to rats. *Toxicol. Ind. Health* 3, 587–599.
- 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., Schreiner, C.A., Mackerer, C.R., 1997a. Developmental toxicity of dermally applied crude oils in rats. *J. Toxicol. Environ. Health* 52, 79–93.
- Feuston, M.H., Hamilton, C.E., Mackerer, C.R., 1997b. Systemic and developmental toxicity of dermally applied syntower bottoms in rats. *Fund. Appl. Toxicol.* 35 (2), 166–176.
- 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), S4–S9.
- Jedrychowski, W., Perera, F.P., Tang, D., Stigter, L., Mroz, E., Flak, E., Spengler, J., Budzyn-Mrozek, D., Kaim, I., Jacek, R., 2012. Impact of barbecued meat consumed in pregnancy on birth outcomes accounting for personal prenatal exposure to airborne polycyclic aromatic hydrocarbons: birth cohort study in Poland. *Nutrition* 28 (4), 372–377.
- 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.
- McKee, R.H., Schreiner, C.A., Nicolich, M.J., Gray, T.M., 2013. Genetic toxicity of high-boiling petroleum substances. *Regul. Toxicol. Pharmacol.* 67 (2S), S75–S85.
- Murray, F.J., Gray, T.M., Roberts, L.G., Roth, R.N., Nicolich, M.J., Simpson, B.J., 2013. Evaluating the male and female reproductive toxicity of high-boiling petroleum substances. *Regul. Toxicol. Pharmacol.* 67 (2S), S60–S74.
- Nicolich, M.J., Simpson, B.J., Murray, F.J., Roth, R.N., Gray, T.M., 2013. 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. *Regul. Toxicol. Pharmacol.* 67 (2S), S10–S29.
- 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.
- Potter, T.L., Simmons, K.E., 1998. *Composition of Petroleum Mixtures Total Petroleum Hydrocarbon Criteria Working Group Series*, vol. 2. Amherst Scientific Publishers, Amherst.

- 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., Mackerer, C.R., 1988. Correlation of mutagenic and dermal carcinogenic activities of mineral oils with polycyclic aromatic compound content. *Fund. Appl. Toxicol.* 10, 466–476.
- Roy, T.A., Blackburn, G.R., Mackerer, C.R., 1994. Evaluation of analytical endpoints to predict carcinogenic potency of mineral oils. *Poly Arom. Comp.* 5 (1), 279–287.
- Speight, J., 2007. *The Chemistry and Technology of Petroleum*, fourth ed. CRC Press, Taylor & Francis Group, Boca Raton.
- US EPA (U.S. Environmental Protection Agency), 2000. Data Collection and Development on High Production Volume (HPV) Chemicals. Federal Register 65(248). Tuesday, December 26, 81686–81698.
- US EPA (U.S. 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](http://www.epa.gov/crem/library/cred_guidance_0309.pdf)> (accessed 21.04.12).
- US EPA (U.S. Environmental Protection Agency), 2012. Glossary of terms. (<http://www.epa.gov/radon/glossary.html>) (last updated 02.08.12; accessed 29.01.13).