Regulatory Toxicology and Pharmacology 67 (2013) S30–S45

Contents lists available at SciVerse ScienceDirect

Regulatory Toxicology and Pharmacology

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

The relationship between repeat-dose toxicity and aromatic-ring class profile of high-boiling petroleum substances

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

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

ARTICLE INFO

Article history:
Available online 7 June 2013

Keywords:
Biological models
Dermal
High-boiling petroleum substances
HPV chemical challenge program
Mixtures toxicity
Polycyclic aromatic compounds
QCAR modeling
Rat
Repeat-dose toxicity

ABSTRACT

A study was undertaken within the context of the U.S. EPA HPV Chemical Challenge Program to (1) characterize relationships between PAC content and repeat-dose toxicities of high-boiling petroleum substances (HBPS) and (2) develop statistical models that could be used to predict the repeat-dose toxicity of similar untested substances. The study evaluated 47 repeat-dose dermal toxicity and 157 chemical compositional studies. The four most sensitive endpoints of repeat-dose toxicity were platelet count, hemoglobin concentration, relative liver weight and thymus weight. Predictive models were developed for the dose–response relationships between the wt.% concentration of each of seven ring classes of aromatic compounds (the "ARC profile") and specific effects, with high correlations ($r = 0.91–0.94$) between the observed and model-predicted data. The development of the mathematical models used to generate the results reported in this study is described by Nicolich et al. (2013). Model-generated dose–response curves permit the prediction of either the effect at a given dose or the dose that causes a given effect. The models generate values that are consistent with other standard measures. The models, using compositional data, can be used for predicting the repeat-dose toxicity of untested HBPS.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

High-boiling petroleum substances (HBPS), i.e., substances with final boiling points ⩾ approximately 650 °F (343 °C), include substances such as asphalt, aromatic extracts, crude oils, gas oils, heavy fuel oils, lubricating oil basestocks, waxes and related materials, and certain petroleum waste substances. HBPS have a high degree of complexity due to the large number of isomeric structurally-related individual compounds, including a wide variety of 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 repeat-dose toxicity studies of HBPS have been published. The studies include dermal exposure of Sprague–Dawley rats (Cruzan et al., 1986; Feuston et al., 1994, 1996, 1997b), inhalation exposure of Sprague–Dawley rats (Dalbey et al., 1982), inhalation exposure of Wistar rats (Skyberg et al., 1990), oral exposure of Fischer-344 rats (Firriolo et al., 1995), dermal and oral exposures in C3H mice (Feuston et al., 1997a), and oral studies in the American mink (Mustela vision) (Schwartz et al., 2004). A number of these studies have reported effects that included death, decreased body weights, aberrant serum chemistry and hematology values, altered organ weights, and histopathology findings in selected organs. A few individual PACs have also been evaluated for their potential to cause repeat-dose toxicity (ATSDR, 1995).

Because a single HBPS is typically composed of at least thousands of chemical compounds and the composition varies (Speight, 2007), it is not feasible either to test each individual component of a petroleum stream or to test all possible petroleum substances for repeat-dose toxicity. Even if it were feasible to test every component, the repeat-dose toxicity of such complex substances is unlikely to be defined by a simple, additive approach (i.e., summing the toxicities of the individual components).

In an effort to meet the requirements of the U.S. EPA HPV Challenge (US EPA, 2000), the authors, working on behalf of the Petroleum HPV Testing Group (PHPVTG), examined the relationship between the PAC content of selected refinery streams and their repeat-dose and developmental toxicities. This was done as a follow-up to a previous report (Feuston et al., 1994) showing that effects on certain endpoints of repeat-dose and developmental

* Corresponding author.
E-mail address: rroth@rothtox.com (R.N. Roth).

0273-2300/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.yrtph.2013.05.010

toxicity of petroleum substances were correlated with the test samples' levels of 3–7 ring PAC. With regard to repeat-dose toxicity, Feuston et al. (1994) relied on the results of toxicity studies of 13 different petroleum streams in which the test sample was applied dermally to Sprague–Dawley rats for 13 weeks. Twelve of the streams tested by Feuston et al. were included in this project and were used to build the final statistical models (a.k.a. Aromatic Ring Class [ARC] models). The basic experimental design of these studies has been described previously (Cruzan et al., 1986). Endpoints for systemic toxicity included skin irritation, body and organ weights, and hematology and serum chemistry. On the basis of Spearman rank-order tests, Feuston et al. (1994) reported that repeat-dose toxicity (i.e., decreased thymus weight, increased liver weight, and aberrant hematology and serum chemistry values) was correlated with the concentrations of PACs composed of 3, 4, 5, 6, and/or 7 fused aromatic rings. This earlier paper was based on a smaller data set, was qualitative in nature, and did not allow for the prediction of toxicity in untested petroleum substances.

The current study extends the initial evaluation of Feuston et al. (1994) by evaluating a greater number of repeat-dose toxicity studies involving a larger number of HBPS and by developing a more sophisticated, mathematically-based model to evaluate the relationship between repeat-dose toxicity and PAC content. It covers a range of PAC-containing HBPS (i.e., asphalt, aromatic extracts, crude oils, gas oils, heavy fuel oils, lubricating oil basestocks, waxes and related materials, and certain petroleum waste substances).

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 HPV Challenge program. 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; Simpson et al., 2007, 2008). Additional aspects of the larger investigation, which also included genetic toxicity, are described in the accompanying articles (McKee et al., 2013b; Murray et al., 2013a,b; Nicolich et al., 2013).

The objective of the current study was to evaluate the relationship between PAC content and selected endpoints (Screening Information Data Set [SIDS]) of repeat-dose toxicity. This paper describes the data definition, selection, and collection; the thought processes applied in the model development; and, the final model forms. The paper also provides examples of the repeat-dose toxicity predictions of the ARC models and suggests some possible applications of the ARC models.

2. Materials and methods

2.1. Terminology

Throughout this paper, the following terms are used with regard to the aromatic compound content of petroleum substances:

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 wt.% 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), e.g., the ARC 3 value would be the wt.% of the DMSO-soluble 3-ring aromatic compounds within the petroleum substance.

2.2. Repeat-dose toxicity data

PHPVTG member companies were asked to provide the original laboratory reports of any repeat-dose toxicity studies of HBPS that had accompanying PAC compositional data of the test sample. A boiling point criterion was added later in the project to clearly define the exact nature of the substances on which the models were based, i.e., substances with final boiling points $\geqslant$ approximately 650 °F (343 °C). Conversely, the boiling point criterion also clarified the types of petroleum substances that are not likely to be addressed by models based on PAC content, e.g., gasoline and kerosene. Substances whose final boiling point is $\geqslant$ approximately 650 °F (343 °C) contain fused aromatic-ring compounds with $\geqslant$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 $\geqslant$3 aromatic rings.

A total of 47 studies were reviewed by the authors during the model building phase of the study (Table 1). The 47 studies consisted of 46 studies submitted by PHPVTG member companies and a recently completed study sponsored by the American Petroleum Institute (API). The materials tested in the reviewed toxicity studies covered a range of PAC-containing petroleum substances including asphalt, aromatic extracts, crude oils, gas oils, heavy fuel oils, lubricating oil basestocks, waxes and related materials, and certain petroleum waste substances. A few of the provided studies had been published in the scientific literature, but for purposes of this project, the original laboratory reports (not the publications) were used.

The 47 toxicity studies consisted of nineteen 28-day and twenty-eight 90-day repeat-dose toxicity studies. All but one of the 47 studies had been carried out in rats, the exception being a 10-week repeat-dose study in mice on sample 86001. This study was not used in the current evaluation, but has been published in Feuston et al. (1997a). For HBPS, dermal contact is considered to be the most likely route of human exposure. All 46 rat studies exposed animals via the dermal route. One 13-week study (sample 86187) included two groups (males) that had been exposed orally, and four groups (males and females) that had been exposed dermally. Only data from the dermally exposed animals were used in this evaluation. The studies conducted in rats included at least one concurrent control group, and most included three dosed groups. Reports of the studies were judged to be either "reliable without restrictions" or "reliable with restrictions", i.e., Klimisch reliability scores of 1 or 2 (Klimisch et al., 1997).

All experimental observations and measurements that were considered likely to be useful in subsequent evaluations were captured from the reports of the 46 studies performed in rats. Data from both 90- and 28-day repeat-dose studies were used to assess the relationship between PAC content and toxicity, with the difference in duration of dosing being considered in the statistical analysis. Data from 45 of the studies were used in the preliminary modeling effort (Table 1). The study provided by API was used only in the final modeling, as it was not completed until after the preliminary modeling effort.

Additional criteria for including studies and dose groups in the final modeling effort were established by the authors, as detailed by Nicolich et al. (2013). These criteria included factors such as group size and the method of compositional analysis. For example, a preliminary evaluation of the utility for final modeling of three analytical data sets found that compositional data derived from the Method II chemical characterization procedure generally

Table 1
Availability of repeat-dose toxicity studies.

Studies	28-Day studies	90-Day studies	Total no. studies
Studies reviewed	19	28	47
Studies from which data were extracted	19	27	46
Studies used for preliminary modeling	19	26	45
Studies used for final modeling	1	17	18
Studies obtained after final models completed[a]	0	2	2

[a] Studies used to evaluate the final models, see Section 3.6.

produced models with the best fit (Nicolich et al., 2013). Therefore, the authors decided to use data only from toxicity studies in which the composition of the test material had been analyzed using the Method II chemical characterization procedure, described in Section 2.4. Consequently, data from 16 studies were excluded from the final modeling exercise due to a lack of appropriate compositional data. Data from an additional 12 studies were excluded because the test samples were not considered HBPS. After application of the criteria for exclusion, data from 18 of the 46 studies from which data were extracted were identified for use in the final analysis (Table 1). The majority of the studies involved test materials that fell into two broad categories: heavy fuel oils and gas oils. Of the 18 repeat-dose toxicity studies used to build the ARC models, there were eight and five studies of heavy fuel oils and gas oils, respectively. Two studies on lubricating oil basestocks, two on aromatic extracts (one distillate and one residual) and one study on petroleum wastes made up the remaining five studies used for final modeling.

The 18 repeat-dose studies identified for use in the final modeling were evaluated to determine if any individual dose groups should be excluded from the modeling exercise and the generation of the $Estimate_{10}$ values. Four dose groups were subsequently excluded on the basis of small group size due to high mortality (50–90%). All of the excluded dose groups were the highest dose groups in the study. Because the modeling weighted each data point (dose group) equally, it was important to exclude data points that were based on an inadequate amount of data, i.e., small group size. Furthermore, the high mortality in these four groups is an indication that the MTD (maximum tolerated dose) in these animals had been reached or surpassed. The purpose of the selection criterion was to identify data appropriate for the analysis of the relationship of PAC content and sensitive repeat-dose endpoints, i.e., the effects that were observed at the lowest doses. For this purpose, data points in the range of <50% premature mortality were more useful than data points in the range of >50% premature mortality, which were at or above the MTD. In short, exclusion of data based on a small group size provided a more scientifically defensible basis for modeling the data at the lower end of the dose–response curve.

Subsequent to the completion of the ARC models, the authors were provided results from an additional two repeat-dermal studies on HBPS sponsored by the PHPVTG. These studies had been completed after the ARC models were finalized. In both studies, the test samples (20906 and 120801) had an ARC profile inside the model domains (Table 2). The results of these two studies were used to test the predictive accuracy of the models (see Section 3.6).

2.3. Identification of repeat-dose toxicity endpoints for mathematical modeling

The authors recognized that it would be difficult to attempt to characterize the PAC content – toxicity relationships for all the biological endpoints for which it had collected data. Consequently, it was decided to identify a smaller number of biological endpoints that would undergo preliminary quantitative assessment for dose–response relationship(s) between PAC content and an effect. The preliminary assessment served two purposes, (1) to identify a select number of biological endpoints that would undergo final modeling, and (2) to evaluate the utility for final modeling of the various analytical data sets. This subgroup of endpoints was selected based on three general considerations:

(1) The endpoints were among those that were most often statistically significantly affected in the studies from which data had been extracted;
(2) the endpoints were among those that were most often statistically significantly affected at the LOEL (lowest observable effect level) in the studies from which data had been extracted (i.e., those effects that would be predictive of a significant biological effect); and
(3) effects on an endpoint could be used to define/characterize a point of departure in the dose–response curve for a repeat-dose effect.

After completing the preliminary quantitative assessment of the dose–response relationship(s), the number of endpoints being characterized was reduced again considering the following:

(1) The overall degree of the reported statistical significance from all relevant individual study dose–response assessments (the significance was independent of the current modeling exercise);
(2) Whether related endpoints had also been characterized, thus making the analysis redundant (e.g., among hematocrit, hemoglobin, and erythrocyte count, only one endpoint was identified for final modeling); and
(3) Whether the effect on an endpoint was considered an adverse effect or predictive of an adverse effect.

Preference was given to selecting endpoints that the authors considered biologically significant. Biological significance was the determination that an effect on the endpoint could be considered either adverse, or a precursor to an adverse effect as defined by US EPA (2002).

2.4. Compositional analyses of petroleum substances evaluated for repeat-dose toxicity

As noted earlier, the individual petroleum substances considered in this paper are extremely complex, containing at least thousands of structurally-related individual substances, including a wide variety of polycyclic aromatic compounds (PACs) (Altgelt and Boduszynski, 1994; Gray et al., 2013; Speight, 2007; Potter and Simmons, 1998). Consequently, the precise composition of any given test substance is not known. As a result, all of the materials considered in this report are defined as TSCA Class II substances (Unknown or Variable Composition, Complex Reaction

Table 2
ARC profiles of samples used to build and evaluate final repeat-dose models.

Sample no.	ARC profile[a]							Sample used to:
	1-ring wt.%	2-ring wt.%	3-ring wt.%	4-ring wt.%	5-ring wt.%	6-ring wt.%	7-ring wt.%[b]	
60901	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Build final models[c]
82191	0.0	0.0	0.7	0.1	0.1	0.0	0.0	Build final models[c]
8281	2.0	29.5	14.7	0.0	0.5	0.5	0.0	Build final models[c]
83366	0.1	2.5	5.1	2.5	1.3	0.9	0.1	Build final models[c]
85244	0.0	0.1	2.5	1.9	1.2	0.5	0.0	Build final models[c]
86001	0.0	2.6	25.7	19.3	6.4	3.2	0.6	Build final models[c]
86181	0.2	2.5	12.4	7.4	2.5	0.5	0.0	Build final models[c]
86187	0.0	0.0	4.1	8.1	6.1	2.0	0.4	Build final models[c]
86193	0.8	2.9	0.4	0.0	0.0	0.0	0.0	Build final models[c]
86270	0.9	2.6	3.5	0.9	0.4	0.0	0.4	Build final models[c]
86271	0.1	0.8	5.3	3.2	0.4	0.2	0.1	Build final models[c]
86272	0.3	4.9	8.1	1.6	0.3	0.2	0.0	Build final models[c]
86484	0.0	1.0	9.8	19.5	9.8	4.9	1.0	Build final models[c]
87213	0.1	4.2	6.3	0.3	0.0	0.0	0.0	Build final models[c]
87476	0.0	0.0	0.0	0.1	0.3	0.5	1.6	Build final models[c]
89106	0.2	1.2	1.7	1.2	0.6	0.5	0.0	Build final models[c]
F-179	0.0	0.7	10.0	30.0	20.0	6.0	0.0	Build final models[c]
F-233	3.0	0.0	0.0	0.0	0.0	0.0	0.0	Build final models[c]
20906	0.0	0.0	5.4	6.8	1.4	0.3	0.0	Evaluate final models[d]
120801	0.1	2.2	0.6	0.0	0.0	0.0	0.0	Evaluate final models[d]

[a] ARC profile – the wt.% 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).
[b] The ARC 7 value is the wt.% of the 7 and larger aromatic-ring compounds within the petroleum substance as determined by the Method II chemical characterization procedure (see Section 2.4).
[c] See Sections 3.3 and 3.5.
[d] See Section 3.6.

Products and Biological Materials, referred to as UVCBs) (US EPA, 1995).

Compositional data on each test material in the repeat-dose toxicity studies were extracted from the corresponding analytical report. The analytical reports were judged to be either "reliable without restrictions" or "reliable with restrictions", i.e., Klimisch reliability scores of 1 or 2 (Klimisch et al., 1997). Among the analytical reports received, five different compositional analytical methods had been used. As reported in Nicolich et al. (2013), preliminary modeling found the analytical technique labeled "Method II" more highly related to the chosen biological endpoints than any of the four other techniques. As a result, the ARC models were developed using only Method II derived data, i.e., the wt.% of each of the seven classes of the 1–7-and larger aromatic-ring compounds in the test substance (the "ARC profile"). Table 2 presents the ARC Profiles for the 18 samples used to build and the two samples used to corroborate the repeat-dose models. As noted in Nicolich et al. (2013), it is not adequate to consider the total percent weight of the 1–7 and larger aromatic-ring compounds because the total percent weight is not related to the dose–response curve. Consequently, the models are based on the concentrations of each aromatic-ring class.

Method II, a rapid liquid–liquid chemical characterization procedure, was developed for routine isolation, classification and quantitation of complex polynuclear aromatic compounds (PAC) present in petroleum fractions with boiling points ranging from >300 °F (149 °C) to >1000 °F (600 °C). The DMSO-soluble aromatic compounds are first extracted into cyclohexane and then extracted with dimethyl sulphoxide (DMSO). This is then back-extracted into fresh pentane or cyclohexane by addition of water to the DMSO (Blackburn et al., 1996; Gray et al., 2013; Roy et al., 1985, 1988, 1994). For higher molecular weight, more viscous samples with initial boiling points of >1000 °F (600 °C), the cyclohexane in the back extraction is replaced with methylene chloride or carbon tetrachloride (Gray et al., 2013). The aromatic content (1–7 and larger aromatic-ring compounds) of these extracts is then determined by gas chromatography with mass spectrometry (GC–MS) or flame ionization detection (GC–FID).

2.5. Model development and evaluation

Preliminary modeling consisted of the development of linear regression models using data from 45 studies; incorporating sample compositional data from at least 4 different analytical methods (Simpson et al., 2007, 2008). The final modeling effort was restricted to compositional and response data from the studies (18) in which the test sample was defined as an HBPS and had been characterized by the Method II analytical procedure.

The ARC models were developed using linear regression analysis methods with biological endpoint (e.g., platelet count) as the dependent (i.e., predicted) variable. The independent (i.e., predicting) variables consisted of relevant study design features, biological parameters (e.g., control group response), and test substances variables (e.g., chemical classes based on wt.% of individual ARC rings) as shown in Table 3. The analyses were based on ordinary least squares (OLS) methods (Draper and Smith, 1998). The development of the preliminary models, and the mathematical forms and coefficients of the ARC models are described in detail by Nicolich et al. (2013).

The ARC models were evaluated using four different statistical methods: (1) using holdout sample data, (2) using nonsense data, (3) using new data, and (4) sensitivity analyses (screening and global analyses). The results of these techniques are described in detail by Nicolich et al. (2013).

2.6. Comparison of predicted values with estimates of toxicity derived using existing techniques

The ARC model predicted values were compared to estimates of the dose associated with a 10% change in response from the control group derived by existing methods, e.g., EPA's Benchmark Dose (BMD) method.

Table 3
Variables for final models of repeat-dose toxicity.

Dependent variable	Transformation on dependent variable	Covariate (independent biological variable)	Other independent biological variables	Additional Method II terms included[a]
Thymus weight (absolute)	None	Control group thymus weight (absolute)	Body weight, sex	Interaction term of the form $\sum_{j=1}^{5} \xi_j \cdot \text{dose} \cdot \text{ARC}_4 \cdot \text{ARC}_5 \cdot \text{ARC}_j$
Platelet count	None	Control group platelet count	Sex, duration of dosing	Interaction term of the form $\sum_{j=1}^{3} \xi_j \cdot \text{dose} \cdot \text{ARC}_4 \cdot \text{ARC}_5 \cdot \text{ARC}_j$
Hemoglobin concentration	None	Control group hemoglobin concentration	Sex, duration of dosing	Interaction term of the form $\sum_{j=1}^{5} \xi_j \cdot \text{dose} \cdot \text{ARC}_4 \cdot \text{ARC}_5 \cdot \text{ARC}_j$
Liver weight (relative)[b]	None	Control group liver weight (relative)[b]	Body weight, sex, duration of dosing	Interaction term of the form $\sum_{j=1}^{5} \xi_j \cdot \text{dose} \cdot \text{ARC}_4 \cdot \text{ARC}_5 \cdot \text{ARC}_j$

[a] Method II chemical characterization procedure (See Section 2.4).
[b] Relative to body weight.

2.6.1. Model predicted values (PDR_{10})

The ARC models can be used to predict the effect on a modeled endpoint at a given dose or the dose that causes a given effect. In this regard, the models can be used to predict the dose level that produces a predicted change from the controls, herein termed the Predicted Dose Response$_x$ (PDR_x), with "x" being the percentage change from control. The ARC models can also be used to predict dose–response curves for the four sensitive endpoints of repeat-dose toxicity. A 10% change from controls (PDR_{10}) was arbitrarily selected for illustrative purposes and PDR_{10} values were calculated using the appropriate ARC model and the ARC profile for the study material being evaluated.

2.6.2. Estimates of toxicity derived using existing techniques (Estimate$_{10}$ values)

In order to check the reasonableness of the values predicted by our models (i.e., PDR_{10}s), we provided a corresponding Estimate$_{10}$ value or outcome for each model prediction. This check was only one of several that we did to check the models' reasonableness. The Estimate$_{10}$ is defined as the dose estimated to produce a 10% response based on the *observed responses* from the study. The purpose of calculating the Estimate$_{10}$ values was to have a point of comparison to the PDR_{10} values predicted for multiple endpoints. In developing the Estimate$_{10}$ values, we took a hierarchical approach based on recognized methods of evaluating empirical data and deriving dose response values.

To develop an Estimate$_{10}$, we first attempted to calculate a BMD_{10}, which is often regarded as a standard (US EPA, 2012a). When the BMD_{10} could be calculated, it was set as the Estimate$_{10}$. In most cases in this paper, the Estimate$_{10}$ was the BMD_{10}. There were a few data sets where a BMD_{10} could not be calculated because (1) available models could not be adequately fit to the data or (2) we did not have the standard deviation (SD) of the observed data. In these cases, we used a linear or quadratic regression to estimate the equivalent of a BMD_{10}, and used this value as the Estimate$_{10}$. If a linear or quadratic regression could not be used to estimate the equivalent of a BMD_{10}, we relied on professional judgment to derive an Estimate$_{10}$ value. In summary, the Estimate$_{10}$ is a value whose method of estimation cascades down from the calculated BMD_{10} value, to a calculated regression value, to a professional judgment of the response range.

BMDs were calculated using the Benchmark Dose Software (BMDS) Version 2.2. A linear or quadratic model was used depending on sample size and which model had the better fit to the data based on established criteria, such as minimizing the residual sum of squares. The method and criteria used to calculate the BMD_{10} has been described in detail elsewhere (Crump, 1984; Gift et al., 2011).

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. Because of the small number of dose groups in each of the studies that were used in this analysis, we limited the models for current comparisons to either a linear or quadratic regression model of the form

$$y = \beta_0 + \beta_1 x$$

or

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

respectively, where y is the response and x is the dose.

The EPA has provided a 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 3 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 or quadratic regression equation of the form noted above based on ordinary least squares estimation, and this value was reported for the "Estimate$_{10}$" 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., hemoglobin concentration increases 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$_{10}$" value was reported and the reason noted. In all cases, if the "Estimate$_{10}$" value was greater than 2000 mg/kg$_{bw}$/day it was reported as ">2000 mg/kg$_{bw}$/day" to avoid overly precise estimates on materials that are judged to be "non-toxic".

There were several cases where the estimated BMD$_{10}$ or Estimate$_{10}$ was beyond the observed data, or where the PDR$_{10}$ was an extrapolated dose. These estimates are generally not recommended for regulatory purposes because they involve estimates beyond the corresponding model characterizations. However, we included these values so we could compare the predictions. To avoid this problem we could have used a BMD$_{05}$ or BMD$_{01}$ (and corresponding PDR) so the values would be within the relevant range, but this would add an unnecessary complication to an already complex discussion.

2.6.3. Assessment of the consistency of predicted and estimated values

If either a PDR$_{10}$ or Estimate$_{10}$ value for an endpoint was missing, no comparison was made. We judged the two values as "consistent" if their relative percent difference is less than 100, i.e., less than a threefold 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, 2012b). For example, 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 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.

3. Results

3.1. Identification of sensitive endpoints of repeat-dose toxicity

From among all the biological endpoints for which data had been captured, seven repeat-dose endpoints were identified for a preliminary mathematical characterization of potential dose–response relationship(s) between PAC content and endpoint-specific effects (See Table 4). Of these, four endpoints were subsequently selected for final mathematical characterization (See Table 5). The four endpoints of repeat-dose toxicity chosen for final modeling were among those endpoints most often statistically significantly affected in the studies and affected most often at the studies' LOELs (i.e., those effects that would be predictive of a significant biological effect). They included: thymus weight (absolute), hemoglobin concentration, platelet count, and liver weight relative to body weight (a.k.a. "relative liver weight").

3.1.1. Absolute thymus weight

Absolute thymus weight was statistically significantly decreased in more than half of the 90-day studies and in the one 28-day study in which thymus weight was recorded. Furthermore, the thymus was frequently identified in the study reports as a "target" organ. Feuston et al. (1994) also reported decreases in thymus weight as being related to the levels of 3–7 ring PAC.

Decreased thymus weight can be a general indication of potential adverse effects on the immune system, specifically a specific cell line (T lymphocytes). Regulatory guidance documents suggest a significant decrease in thymus weight might be considered an adverse effect by some regulatory authorities (Abadin et al., 2007; ATSDR, 1996, 2006).

While the identification of decreased absolute thymus weight for final modeling was based on biological considerations, when conducted, the final modeling indicated there is a high correlation between the observed and model predicted thymus weight, $r = 0.91$ (Table 6).

3.1.2. Hemoglobin concentration

Data were available on three indicators of RBC mass, erythrocyte count, hematocrit, and hemoglobin concentration. All three parameters provide information concerning the oxygen-carrying capacity of the blood and bone marrow erythropoietic activity. Given that these three measurements are indicative of RBC mass, and therefore, probably inter-related, it was decided that only one should be selected for detailed statistical modeling. Preliminary modeling of hemoglobin concentration indicated it could be modeled with a higher degree of confidence than the other two endpoints (see Table 4).

Hemoglobin concentration was statistically significantly decreased in more than half of the 90-day studies and in approximately 20% of the 28-day studies. The related parameters, erythrocyte count and hematocrit, had similar, highly affected incidence rates.

Available regulatory guidance documents suggest a significant decrease in hemoglobin concentration would in all probability be considered an adverse effect (Abadin et al., 1998; ATSDR, 1996, 2006).

While the identification of hemoglobin concentration for final modeling was based on biological considerations, when conducted, the final modeling indicated there is a high correlation between the observed and model predicted hemoglobin concentration, $r = 0.94$ (Table 6).

3.1.3. Platelet count

Platelet count was statistically significantly decreased in 50% of the 90-day repeat-dose toxicity studies and approximately 5% of the 28-day studies in which platelets were counted.

Table 4
Preliminary analysis: repeat-dose toxicity endpoints using linear regression models with compositional data determined by Method II.[a]

Endpoint	Number of dose groups	Multiple correlation coefficient (r)	Standard error (SE)[b]
Liver weight. (relative)[c]	124	0.94	0.07
Thymus weight. (absolute)	92	0.90	0.11
Erythrocyte count	128	0.54	0.13
Platelet count	112	0.91	0.09
Hemoglobin concentration	128	0.75	0.07
Hematocrit	128	0.60	0.17

[a] Analysis based on 45 studies, see Table 1.
[b] Calculated as the square root of the error mean square.
[c] Relative to body weight.

Table 5
Sensitive endpoints in repeat-dose toxicity studies.

Endpoint [a]	Sensitive endpoint [b]	Used in preliminary model development[c]	Good correlation in preliminary modeling[c]	Used in final model development[c]
Liver weight (absolute)	√			
Liver weight (relative)[d]	√	√	√	√
Thymus weight (absolute)	√	√	√	√
Erythrocyte count	√	√		
Hemoglobin concentration	√	√	√	√
Hematocrit	√	√		
Platelet count	√	√	√	√

[a] Key endpoints evaluated in the repeat-dose toxicity studies.
[b] 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).
[c] Blank cells represent endpoints that were judged "not sensitive" or not evaluated at some point in model development.
[d] Relative to terminal body weight.

Table 6
Final models: correlation between observed and predicted values for repeat-dose toxicity endpoints.[a]

	Number data points	Number of studies	Model degrees freedom	Multiple correlation coefficient (r)	Standard error (SE)[b]	p Normality test[c]
Thymus weight (absolute)	84	16	16	0.91	0.03	0.32
Platelet count	85	16	14	0.91	0.12	0.09
Hemoglobin concentration	98	18	16	0.94	0.60	0.02
Liver weight (relative)[d]	90	17	17	0.94	0.19	0.15

[a] Based on the 18 samples used to build final models, see Table 1.
[b] Calculated as the square root of the error mean square.
[c] p values less than 0.01 indicate the residuals are not distributed normally.
[d] Relative to body weight.

In addition to prothrombin time and activated partial thromboplastin, platelet count is one of the core recommended tests for assessment of hemostasis. Along with the hemoglobin measurements and thymus weights, the authors thought platelet count gave an indication of potential effects on a third line of blood cells, megakaryocytes.

The effects seen in the studies reviewed by the authors were quite substantial, both in magnitude and frequency of occurrence. Regulatory guidance documents suggest a significant decrease in platelet count would likely be considered an adverse effect (Abadin et al., 1998; ATSDR, 1996, 2006).

While the identification of decreased platelet count for final modeling was based on biological considerations, when conducted, the final modeling indicated there is a high correlation between the observed and model predicted platelet count, $r = 0.91$ (Table 6).

3.1.4. Relative liver weight

Relative liver weight was the endpoint statistically significantly affected most frequently in the repeat-dose toxicity studies. It was increased in approximately 20% and 80%, respectively of the 28- and 90-day repeat-dose studies in which relative liver weights were recorded. Furthermore, the liver was frequently identified in the study reports as a "target" organ. Feuston et al. (1994) also reported increased liver weight as being related to the levels of 3–7 ring PAC.

Available publications and regulatory guidance documents suggest an increase in liver weight without corresponding histopathological and/or marker enzyme changes would most likely be considered an adaptive rather than a toxicological response (Amacher et al., 1998; ATSDR, 1996, 2006; Pohl and Chou, 2005).

While the identification of "relative liver weight" for final modeling was based on biological considerations, when conducted, the final modeling indicated there is a high correlation between the observed and model predicted relative liver weight, $r = 0.94$ (Table 6).

3.1.5. Other repeat-dose toxicity endpoints

A number of other endpoints were initially identified as candidates for model development and underwent preliminary statistical modeling (Simpson et al., 2007, 2008); but these endpoints were not used in the final statistical modeling. For example, significant, treatment-related decreases in terminal body weight were recorded in 11 of 46 studies in males and 7 of 44 in females. However, it was clear that body weight changes occurred only in animals in which other sensitive endpoints had also been affected. The authors concluded that terminal body weight was not a sensitive endpoint in the context of the evaluation being undertaken.

Although not identified as a sensitive endpoint, the authors did consider and decided not to include dermal effects in the modeling exercise. The authors identified three potential hypotheses associated with the dermal effects.

(1) *A hypothesis that effects on the skin are a critical endpoint and as such need to be characterized or modeled.*
The authors do not think the data available support this hypothesis. A high proportion of the studies reported dermal effects seen only at the site of test material application. This is a common and well-documented finding in studies utilizing dermal application as the route of administration. The authors concluded that these dermal findings were local effects that were not associated with any systemic effects. Evidence arguing against skin effects being a critical endpoint included several instances of materials that produced high skin irritation scores, but had no internal toxicity. Conversely, several materials with low skin scores produced internal toxicity.

While the authors recognize some PAHs are known skin carcinogens, and dermal irritation is believed to be involved in the dermal carcinogenesis seen with selected middle distillate materials (Nessel et al., 1999), the focus of this study is on non-carcinogenic endpoints. Other than local inflammatory responses at the site of application, no indication of more serious skin effects, e.g., carcinogenicity, would be expected to be observed in studies of the design that are included in this project.

(2) *A hypothesis that independent of any systemic toxic effects produced by PACs, dermal irritation alone could produce the pattern of effects seen in the reviewed studies, leading to an erroneous correlation between irritation-produced effects and PAC concentration.*

The authors do not think the available data support this hypothesis. There did not appear to be a consistent correlation between the degree of dermal effects and statistically significant effects on any of the other endpoints. For instance, in several studies "none to minimal" irritation was reported, yet there were statistically significant effects on thymus weight and hemoglobin concentration. Conversely, there were also instances in which "severe" skin irritation was reported, yet there were no statistically significant changes in either thymus weights or hemoglobin content.

(3) *A hypothesis that irritation could have led to alterations in the barrier properties of the stratum corneum, allowing increased PAC absorption.*

While this hypothesis is supportable, the effects seen from this increased absorption of PACs would be captured in the current set of endpoints. Furthermore, if true, this simply means that the outcomes represent a "worst case" relative to the consequences of exposures at lower, non-irritating levels. Consequently, the authors do not believe this would alter the accuracy of the predictive models. Finally, to attempt to define the mechanism of how the PACs are producing their adverse effects is beyond the scope of the current exercise.

3.2. Preliminary modeling of repeat-dose toxicity

Results of the preliminary analyses (Table 4) 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, as detailed in Nicolich et al. (2013).

Among the six repeat-dose endpoints modeled, the magnitude of the correlations (r) between the values predicted by the preliminary models and the values observed in the studies ranged from 0.54 to 0.94. The results of the preliminary analysis strongly suggested that for HBPS a relationship existed between ARC profile, as determined by the Method II chemical characterization procedure, and the sensitive endpoints of repeat-dose toxicity.

3.3. Final modeling of repeat-dose toxicity

Based on the results of the preliminary modeling, four sensitive endpoints of repeat-dose toxicity were selected for final mathematical characterization. The four endpoints were: thymus weight (absolute), platelet count, hemoglobin concentration, and relative liver weight (relative to body weight). In the preliminary modeling, all four of these endpoints were strongly related to the corresponding model's predictions (Table 4). While both the erythrocyte count and hematocrit also related to the ARC profile in the preliminary modeling, these endpoints were excluded from the final modeling because both had a lower correlation coefficient (r) than a similar measure (i.e., hemoglobin concentration).

Table 6 shows the values of the multiple correlation coefficients (r) and residual standard errors (se) for the ARC models. The multiple correlation coefficients (r) between the values predicted by the ARC models and the values observed in the studies for the four models range from 0.91 to 0.94, indicative of a very good model fit. These results indicate that models accurately predict, based on their ARC profiles (derived by the Method II chemical characterization procedure), the effects on selected repeat-dose toxicity endpoints of the test materials used to build the models. The degrees of freedom associated with each model are an indication of the complexity of the individual models (Table 6). 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 6) cannot be directly compared with those generated under the preliminary models (Table 4).

Plots of the observed data point vs. the predicted values for the four endpoints (Fig. 1) provide the most useful form for assessing model adequacy. The optimum model would have all points along the forty-five degree line representing equal values of the observed and predicted data. Nicolich et al. (2013) provides additional details on how the predictive ability of the ARC models was tested using three different techniques.

3.4. Use of models to predict repeat-dose toxicity

The ARC models can be used to generate dose–response predictions. Fig. 2 shows the results of using a model to generate dose–response curves for relative liver weights for two different HBPS with different ARC profiles. The curves are generated by using the equation for relative liver weight, along with the ARC profile (derived using the Method II chemical characterization procedure), the coefficients for relative liver weight, and the average relative liver weight among the control groups. Assuming a dose of 500 mg/kg_{bw}/day, and substituting the control values and the values of the coefficients from the equation for the relative liver weight, the mean relative liver weight at 500 mg/kg_{bw}/day is predicted to be 5.0 for sample 86187 and 10.0 for sample 86001. Repeating this calculation for a range of dose values would produce the two dose–response curves seen in Fig. 2. To determine the relative liver weight relative to control, each of the values from Fig. 2 would be divided by the corresponding predicted control value, and then multiplied by 100.

Fig. 3 presents an example of how the model may be used to predict the dose level that produces a 10% increase in the relative liver weight relative to controls (PDR_{10}), using the same sample of distillate aromatic extract that appears in Fig. 2. To determine the predicted relative liver weight 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 relative liver weight value (3.09%), adding 10% of this value gives a critical response of 3.40%. The dose associated with a response of 3.40% is 56 mg/kg_{bw}/day. Since Fig. 3 provides the ratio of the response at a dose divided by the response at zero, it can be seen that the dose associated with a response that is 110% of the control value is 56 mg/kg_{bw}/day, with the associated 95% confidence interval being approximately 41 and 72 mg/kg_{bw}/day (Fig. 3).

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

For the 18 studies used to build the final repeat-dose models, Table 7 provides comparisons of the dose associated with a 10%

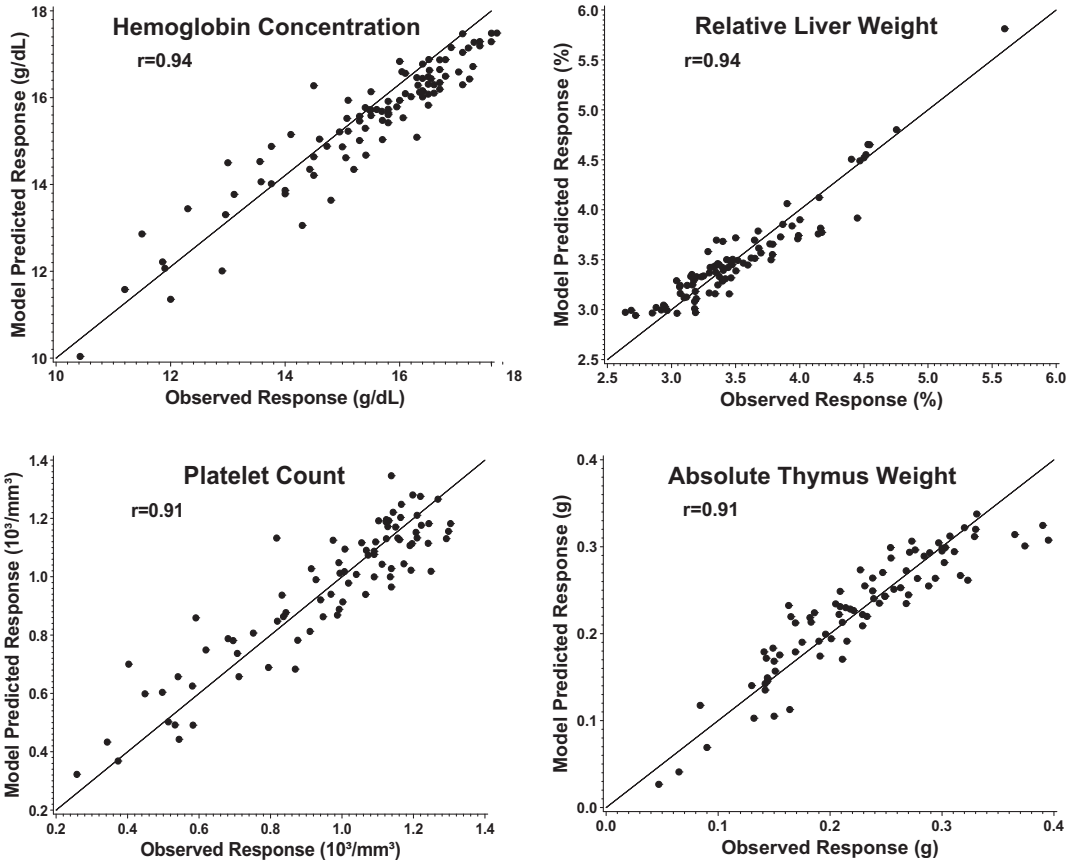

Fig. 1. Plots of observed vs. predicted values for hemoglobin concentration, relative liver weight, platelet count and absolute thymus weight.

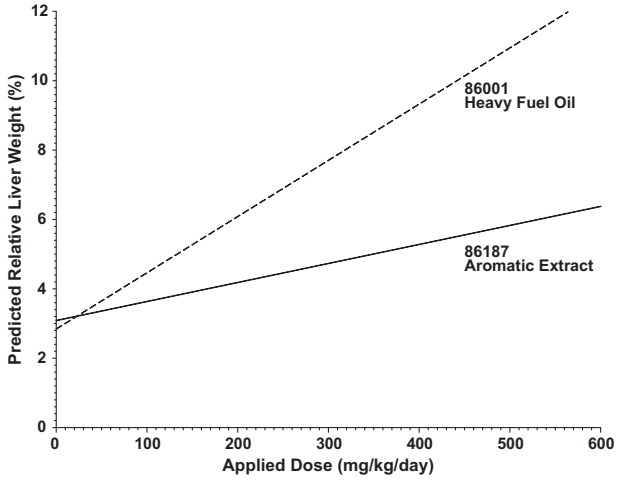

Fig. 2. Predicted dose–response curves for mean relative liver weights for two samples with different ARC profiles.

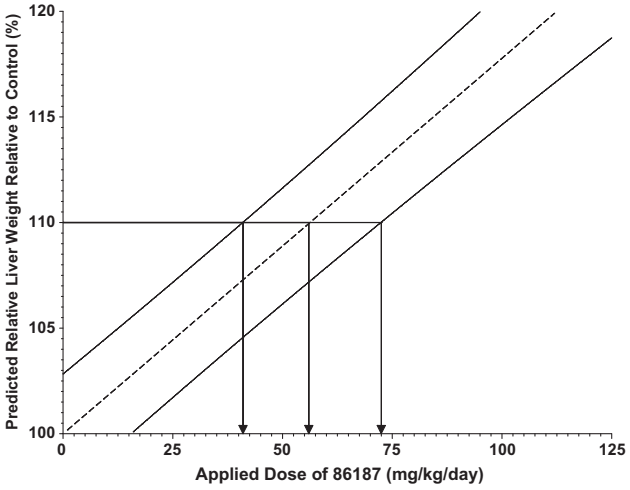

Fig. 3. Prediction of the PDR_{10} (dashed lines) and 95% CI (solid lines) for mean relative liver weight for a sample of a distillate aromatic extract.

change in response from the control value derived using either the ARC statistical model (PDR_{10}) or the observed data ($Estimate_{10}$) for studies that have the appropriate observed data. As seen in Table 7, the ARC statistical models for the four repeat-dose endpoints generate values that are consistent with other standard measures.

To avoid overly precise predictions on materials that are judged to be relatively inactive, PDR_{10} values that are greater than 2000 mg/kg_{bw}/day are shown in the tables as ">2000 mg/kg_{bw}/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 whose dose–response slope is (1) not in the appropriate direction, i.e., a direction inconsistent with the expected treatment effect on the specific endpoint, or (2) nearly flat, i.e., the magnitude of the slope is small (less than or equal to the absolute value of the control value divided by 10,000). The choice of the control value divided by 10,000 is somewhat arbitrary and corresponds to an approximate 20% change from control at a dose of 2000 mg/kg_{bw}/day. Based on the results of the repeat-dose toxicity studies of HBPS reviewed for this paper, the appropriate

Table 7
Comparison of PDR_{10} and $Estimate_{10}$ values for repeat-dose endpoints for samples used to build the final models.

Sample no.	Sex	Decreased absolute thymus weight			Decreased platelet count			Decreased hemoglobin concentration			Increased relative liver weight		
		PDR_{10}[a] (mg/kg$_{bw}$/day)	$Estimate_{10}$[b] (mg/kg$_{bw}$/day)	PDR_{10} & $Estimate_{10}$ consistent[c]	PDR_{10}[a] (mg/kg$_{bw}$/day)	$Estimate_{10}$[b] (mg/kg$_{bw}$/day)	PDR_{10} & $Estimate_{10}$ consistent[c]	PDR_{10}[a] (mg/kg$_{bw}$/day)	$Estimate_{10}$[b](mg/kg$_{bw}$/day)	PDR_{10} & $Estimate_{10}$ consistent[c]	PDR_{10}[a] (mg/kg$_{bw}$/day)	$Estimate_{10}$[b](mg/kg$_{bw}$/day)	PDR_{10} & $Estimate_{10}$ consistent[c]
60901	Female	>2000[g]	–[j]	–	>2000[g]	>1000[h]	Yes	>2000[g]	>1000[h]	Yes	>2000[g]	<1000[h]	Yes
60901	Male	>2000[g]	<1000[h]	Yes	>2000[g]	<1000[h]	Yes	>2000[g]	–[j]	–	>2000[g]	>1000[h]	Yes
82191	Female	1259	–[j]	–	–[k]	–[j]	–	>2000[g]	>1720[h]	Yes	759	<1720[h]	Yes
82191	Male	1428	<1720[h]	Yes	–[k]	–[j]	–	>2000[g]	>1720[h]	Yes	767	<1720[h]	Yes
8281	Female	177	147[d]	Yes	–[e]	–[l]	–	948[f]	776[i]	Yes	182	189	Yes
8281	Male	200	82	Yes	–[e]	–[l]	–	950[f]	904[i]	Yes	184	150	Yes
83366	Female	37	27	Yes	78	47[i]	Yes	>2000[g]	120[i]	No	71	39	Yes
83366	Male	42	26	Yes	77	102[i]	Yes	>2000[g]	129[i]	No	72	49	Yes
85244	Female	87	–[l]	–	365	431[i]	Yes	925	1162[i]	Yes	120	338[i]	Yes
85244	Male	99	–[l]	–	359	391[i]	Yes	927	708[i]	Yes	121	403[i]	No
86001	Female	19	15	Yes	–[e]	–[l]	–	47	46[i]	Yes	20	10	Yes
86001	Male	22	7	No	–[e]	–[l]	–	47	22[i]	Yes	20	15	Yes
86181	Female	23	23[d]	Yes	38	43	Yes	83	83	Yes	34	33	Yes
86181	Male	26	20[m]	Yes	38	91	Yes	83	78	Yes	35	18	Yes
86187	Female	59	21	Yes	73	66[i]	Yes	172	170[i]	Yes	58	27	Yes
86187	Male	67	29	Yes	72	27[i]	Yes	173	102[i]	Yes	58	27	Yes
86193	Female	452	–[j]	–	95	463[d]	No	>2000[f]	387[d]	No	>2000[g]	416[d]	No
86193	Male	513[f]	183[d]	Yes	93	–[j]	–	>2000[f]	932[d]	Yes	>2000[g]	318[d]	No
86270	Female	248	157[i]	Yes	210	361[d,i]	Yes	481	475[i]	Yes	226	127[i]	Yes
86270	Male	281	135[i]	Yes	205	401[d,i]	Yes	482	511[i]	Yes	229	185[i]	Yes
86271	Female	84	44	Yes	132	228	Yes	257	397	Yes	110	94	Yes
86271	Male	95	83	Yes	130	320	Yes	257	162	Yes	111	53	Yes
86272	Female	109	35	No	162[f]	72	Yes	744[f]	138	No	96	52	Yes
86272	Male	124	35	No	160[f]	38	No	746[f]	122	No	97	57	Yes
86484	Female	10	7[i]	Yes	10	8[i]	Yes	18	26[i]	Yes	14	10[i]	Yes
86484	Male	11	7[i]	Yes	10	5[i]	Yes	18	12[i]	Yes	14	12[i]	Yes
87213	Female	>2000[g]	63[i]	No	>2000[g]	174[d,i]	No	1503[f]	364[d,i]	No	167[f]	61[i]	Yes
87213	Male	>2000[g]	40[i]	No	>2000[g]	321[d,i]	No	1506[f]	649[d,i]	Yes	168[f]	86[i]	Yes
87476	Female	>2000[g]	–[j]	–	>2000	>2000[d,i]	Yes	>2000	>2000[d,i]	Yes	>2000	>2000[d]	Yes
87476	Male	>2000[g]	1121[d,m]	Yes	>2000	>2000[d,i]	Yes	>2000	>2000[d,i]	Yes	>2000	1755	Yes
89106	Female	179	194	Yes	167	176	Yes	534	838	Yes	223	198	Yes
89106	Male	203	144	Yes	164	143	Yes	535	356	Yes	225	95	Yes
F-179	Female	249	38	No	165	32	No	253	99	Yes	200	128[m]	Yes
F-179	Male	290	148[m]	Yes	162	29	No	253	67	No	202	31	No
F-233	Female	–[e]	–[l]	–	–[k]	–[j]	–	>2000[f]	>2000[d]	Yes	>2000[f]	1052	Yes
F-233	Male	–[e]	–[l]	–	–[k]	–[j]	–	>2000[f]	1908[d]	Yes	>2000[f]	–[j]	–

[a] ARC model values; all results greater than 2000 are reported as >2000.
[b] 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_{10} calculated using the EPA method (see Section 2.6.2). All results greater than 2000 are reported as >2000.
[c] PDR_{10} and $Estimate_{10}$ values are considered consistent if the relative percent difference between the PDR_{10} and the $Estimate_{10}$ value is less than 100 (see Section 2.6.3).
[d] No statistically significant change was seen in any dose group of the study; $Estimate_{10}$ value is from a simple linear regression from the toxicity study data (see Section 2.6.2).
[e] The ARC profile is beyond the profiles used to develop this ARC model (an extrapolation); no PDR_{10} value reported.
[f] The PDR_{10} value is greater than the highest observed dose used to develop the ARC model; PDR_{10} value reported.
[g] Model predicted dose–response slope is not in the appropriate direction, but the slope is less than or equal to the absolute value of the control value divided by 10,000 (see Section 3.5); >2000 reported.
[h] Only 2 dose groups (control and dosed group); $Estimate_{10}$ value reported as the dose range in which a 10% change is likely to occur.
[i] No SD available; $Estimate_{10}$ value is from a simple linear regression from the toxicity study data (see Section 2.6.2).
[j] Observed data response is in a direction inconsistent with the expected biological effect for the specific endpoint (see Section 3.5): No value reported.
[k] 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 3.5).
[l] No data on this endpoint was reported in the laboratory toxicity report.
[m] Poor model fit; $Estimate_{10}$ value is from a simple linear regression from the toxicity study data (see Section 2.6.2).

directions of the dose–response slopes for the four repeat-dose endpoints would be: an increase for relative liver weight, and decreases for thymus weight, hemoglobin concentration, and platelet count. PDR_{10} values were considered unreliable and are not reported when the model predicted dose–response slope is not in the appropriate direction, and the slope is not negligible (i.e., the slope is greater than the absolute value of the control value divided by 10,000).

There were 125 instances in which a comparison could be made of the PDR_{10} and the $Estimate_{10}$ values of the four repeat-dose endpoints for individual samples. Of the 125 determinations, 102 (82%) were judged "consistent" (Table 8), indicating very good general agreement between the two measures of relative toxicity.

PDR_{10} and $Estimate_{10}$ values are both representations of the same concept; the dose associated with a 10% change in response from a control group. The PDR_{10} values are based on models developed from a series of studies with similar protocols ($\geqslant$79 data points in the model) and represent an expected average value from a statistical model. In contrast, $Estimate_{10}$ values referred to in these tables are based on individual studies, each study having 2–6 data points (mean 3.8). Because it is based on a limited number of data points, one unusual response, such as an atypical control group value or an uncharacteristic response at a high-dose exposure, or a different choice of submodel, could have markedly affected the $Estimate_{10}$ value.

Of the 23 comparisons judged "inconsistent," two had a PDR_{10} value less than the $Estimate_{10}$ value, so the PDR_{10} values could be seen as being "conservative," i.e., they over-predict the effect at a given dose. An additional six comparisons had relative percent differences between 100 and 125.

Of the remaining 15 comparisons judged "inconsistent," at least 7 are due to unusual observed responses or poorly fitting models used to derive the $Estimate_{10}$. The limitations we imposed on the choice of which BMD_{10} model to use may have also have led to the inconsistency. Consider the platelet count response in males exposed to sample F-179. The observed data are shown in Table 9 and plotted in Fig. 4. The PDR_{10} value was 162 mg/kg_{bw}/day and the $Estimate_{10}$ value derived from the BMD_{10} calculation was 29 mg/kg_{bw}/day.

Fig. 4 shows that the modeled dose response curve is unusual in shape. While the BMD criterion of visual fit might have led to a different choice of model and resulting different BMD_{10} value, the limitation we imposed on the use of the BMD_{10} model choice to negate the possibility of bias and to ensure consistency in model use, limited us to the curve shown in the Fig. 4. The BMD is known to vary considerably with the choice of submodel used in its development. Thus, a different choice of model may have led to more agreement.

Many of the "inconsistent" comparisons of PDR_{10} and $Estimate_{10}$ values have a pattern of a non-linear response leading to a low $Estimate_{10}$ value with the corresponding PDR_{10} value being high. Consider the absolute thymus weight results for females, sample 87213. There were 3 data points from the observed data, shown in Table 10 and plotted in Fig. 5. As can be seen, the two dosed groups have almost the same response. However, the control value in this study is 0.368 mg, in contrast to the mean females absolute thymus weight of 0.291 mg derived from the control values of the other studies used in model development. The unusually high control mean value in this study would account for the low $Estimate_{10}$.

Table 9
Observed data sample F-179 (males).

Dose (mg/kg_{bw}/day)	Platelet count
0	0.936
1.1	0.910
10.6	0.752
53	0.695
106	0.619
530	0.533

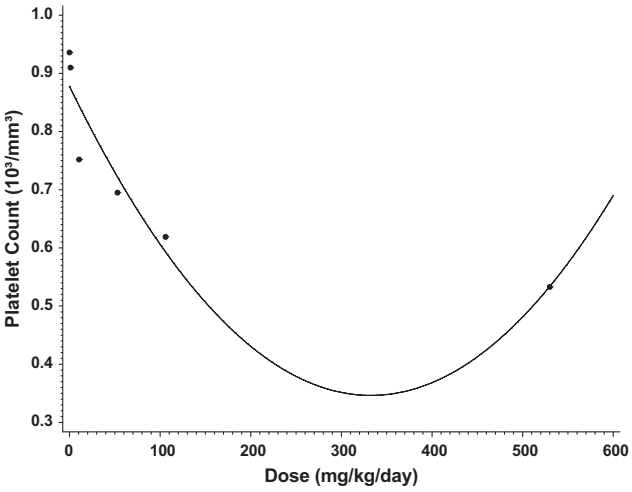

Fig. 4. Platelet response sample F-179 (males).

Table 10
Observed data sample 87213 (females).

Dose (mg/kg_{bw}/day)	Absolute thymus weight
0	0.368
30	0.300
125	0.284

Table 8
Consistency of PDR_{10} and $Estimate_{10}$ for repeat-dose endpoints of samples used to build the final models.

Endpoint	Number of PDR_{10}[a] and $Estimate_{10}$[b] comparisons	% Consistent[c] (PDR_{10} vs $Estimate_{10}$)
Decreased absolute thymus weight	28	79% (22/28)[d]
Decreased platelet count	27	78% (21/27)[d]
Decreased hemoglobin concentration	35	80% (28/35)[d]
Increased relative liver weight[e]	35	89% (31/35)[d]
All endpoints	125	82% (102/125)[d]

[a] ARC model values.
[b] Dose estimated to cause a 10% change from control value, derived using the observed data from existing toxicity study.
[c] PDR_{10} and $Estimate_{10}$ values are considered consistent if the relative percent difference between the PDR_{10} and the $Estimate_{10}$ value is less than 100 (see Section 2.6.3).
[d] Number of comparisons judged "consistent" vs number of total comparisons.
[e] Relative to body weight.

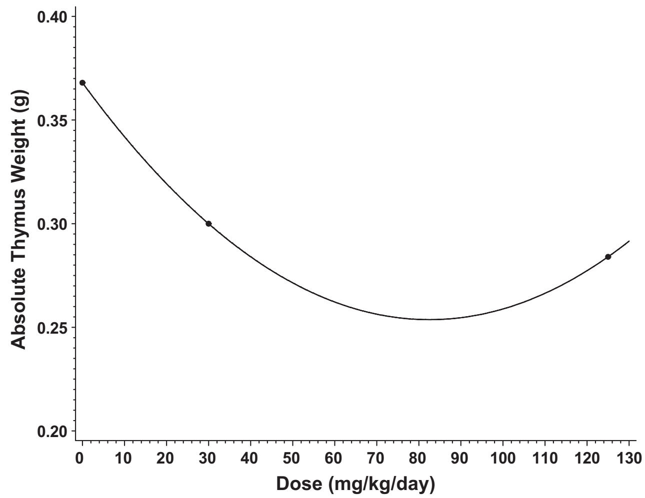

Fig. 5. Absolute thymus weight response sample 87213 (females).

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

As a part of the evaluation process, two samples of HBPS that were not used to develop the ARC models were tested in dermal, 90-day repeat-dose experiments in rats. Table 11 provides the summary results from the toxicity studies and the corresponding PDR_{10} predictions for each of the four repeat-dose endpoints for sample 20906 (a distillate aromatic extract) and sample 120801 (an ultra-low sulfur diesel oil).

As shown in Table 11, modeled PDR_{10} values for sample 20906 (males) are very good, with values for all four repeat-dose endpoints being judged "consistent" with the corresponding $Estimate_{10}$. For sample 20906 (females), only one of four of the PDR_{10} values were judged "consistent" with its corresponding $Estimate_{10}$. However the three PDR_{10} values listed as "inconsistent" were close to being labeled "consistent", with relative differences between the PDR_{10} and $Estimate_{10}$ values ranging from 101% to 132%.

For sample 120801 (females), $Estimate_{10}$ values could only be generated for two of the four endpoints. The comparisons of these two $Estimate_{10}$ values with the corresponding PDR_{10} values were judged "consistent." For males, an $Estimate_{10}$ value could not be derived for the endpoint "platelet count." PDR_{10} and $Estimate_{10}$ values for the three remaining endpoints were judged "consistent." The authors think it is highly likely the PDR_{10} values for "platelet count" are incorrect for both males and females from sample 120801 given (1) the observed data dose–response for platelet counts were in an inappropriate direction (they increased with increasing dose), and (2) were an order of magnitude lower than the PDR_{10}s for the other three endpoints.

4. Discussion

This study describes an association between the ARC profile of HBPS and several sensitive endpoints of repeat-dose toxicity. In doing so, it confirms the findings reported by Feuston et al. (1994). However, the larger data set available for this study and the use of a more sophisticated statistical approach extends the Feuston et al. (1994) findings by allowing for the development of predictive models for selected endpoints of repeat-dose toxicity.

The large number of data points used to develop the models is a particular strength of the current evaluation. The models are relatively simple linear models, all with a similar mathematical form

Table 11
Comparison of PDR_{10} and $Estimate_{10}$ values for two evaluation samples.

Sample no.	Sex	Decreased absolute thymus weight			Decreased platelet count			Decreased hemoglobin concentration			Increased relative liver weight		
		PDR_{10}[a] (mg/kg_{bw}/day)	$Estimate_{10}$[b] (mg/kg_{bw}/day)	PDR_{10} & $Estimate_{10}$ consistent[c]	PDR_{10}[a] (mg/kg_{bw}/day)	$Estimate_{10}$[b] (mg/kg_{bw}/day)	PDR_{10} & $Estimate_{10}$ consistent[c]	PDR_{10}[a] (mg/kg_{bw}/day)	$Estimate_{10}$[b] (mg/kg_{bw}/day)	PDR_{10} & $Estimate_{10}$ consistent[c]	PDR_{10}[a] (mg/kg_{bw}/day)	$Estimate_{10}$[b] (mg/kg_{bw}/day)	PDR_{10} & $Estimate_{10}$ consistent[c]
20906	F	63	13	No	43	14[g]	No	80	188[g]	Yes	78	360[g]	No
	M	71	24	Yes	42	31[g]	Yes	80	77[g]	Yes	79	36[d]	Yes
120801	F	1618[e]	–[h]	–	136	–[h]	–	>2000[e]	>2000[d,g]	Yes	>2000[e,f]	1590[d,g]	Yes
	M	1834[e]	741[d]	Yes	134	–[h]	–	>2000[e]	>2000[d,g]	Yes	>2000[e,f]	1559[d,g]	Yes

[a] ARC model values (PDR_{10} values) greater than 2000 are reported as >2000.
[b] 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_{10} calculated using the EPA method (see Section 2.6.2). All results greater than 2000 are reported as >2000.
[c] PDR_{10} and $Estimate_{10}$ values are considered consistent if the relative percent difference between the PDR_{10} and the $Estimate_{10}$ value is less than 100 (see Section 2.6.3).
[d] No statistically significant change was seen in any dose group of the study; $Estimate_{10}$ value is from a simple linear regression from the toxicity study data (see Section 2.6.2).
[e] The PDR_{10} value is greater than the highest observed dose used to develop the ARC model; PDR_{10} value reported.
[f] 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 3.5); >2000 reported.
[g] Poor model fit or no SD available; $Estimate_{10}$ value is from a simple linear regression from the toxicity study data (see Section 2.6.2).
[h] Observed data response is in a direction inconsistent with the expected biological effect for the specific endpoint (see Section 3.5); no value reported.

across the endpoints, which provides a measure of the consistency of the models (Nicolich et al., 2013). The plots of the observed vs. predicted points shown in Fig. 1 demonstrate that the models are accurate descriptors of the data and are accurate predictors if the ARC profile of the untested petroleum substance falls within the ARC profiles that had been used for model development (i.e., the prediction would be an interpolation). A more detailed discussion of the concepts of interpolated and extrapolated data points and their interpretation and significance is provided by Nicolich et al. (2013).

Identification of the repeat-dose toxicity endpoints that were modeled was carried out with considerable care. Confirmation that it is biologically plausible that changes in the endpoints identified for modeling were related to exposure to PACs is provided in several reviews of the toxicity of PAHs (ATSDR, 1995; IPCS, 1998; SCF, 2002; US Dept Energy, 2007; US EPA, 2007). In these reviews, the spectrum of effects attributable to PAHs was similar to the endpoints that were selected for modeling. Further support that the selected endpoints are reasonable is found in the robust summaries and test plans for HBPS prepared by API in their activities to fulfill the requirements of the HPV challenge program, in which the spectrum of effects of PAC-containing substances is similar to the endpoints selected for modeling (API, 2002, 2003a,b,c,d,e, 2004, 2008, 2009, 2010, 2011a,b,c,d,e, 2012a,b).

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. The models use the concentration of each ring-class rather than the total wt.% of PAC or any subset of ring classes, e.g., 4–6 or 3–7-ring PACs. This approach was found to be essential as many substances with similar total wt.% of PAC may be predicted to have significantly different toxicities.

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 better definition of the various constituents that contribute to toxicity, and a detailed understanding of the mechanisms of PAC toxicity, or at least a general understanding of the underlying mode of toxic action. Both exercises were beyond the scope of the current evaluation. Since the mechanism of action of PAC is not understood, the data should be viewed as indicating only that there is an observed relationship, and should not be used to assess whether any of the specific aromatic-ring class values are causal for the response.

A number of constraints were identified regarding the current versions of the predictive models. As with most linear regression models of this form, the models were found to be good predictors when the ARC profile of the petroleum substance fell within the ARC profiles that had been used for model development (i.e., the prediction would be an interpolation). Not surprisingly, the models were sometimes less accurate predictors if the ARC profile of the petroleum substance fell outside the ARC profiles that had been used for model development (i.e., the prediction would be an extrapolation). The current effort has defined the boundaries of the models' domains to the best of our ability, based on the available samples. In the future, if new test data become available, they could be incorporated into the current models, further increasing their accuracy or expanding the models' domains of applicability, and thereby increasing the models' usefulness (Nicolich et al., 2013).

The models described in this paper were developed using data from dermal toxicity studies. An earlier publication by Feuston et al. (1997a) reported clarified slurry oil, a heavy fuel oil, to be more toxic to mice when administered dermally versus orally. In a study of a distillate aromatic extract that had been administered by both the dermal and oral routes to rats there was greater mortality amongst the dermally treated than the orally dosed rats (API, 2012b). However, the predictive capacity and applicability of the current models to routes other than dermal is currently unknown. While beyond the scope of this investigation, a better understanding of the systemic dose following exposure by different routes might provide a basis for determining the utility of the models to predict effects from other routes of exposure.

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

Because the compositional component of the models is based only on the ARC profile and not on specific category membership, the models are applicable to a wide range of petroleum substances in which PAC content is thought to be responsible for the toxicological effects. Although the various models were built using experimental data developed on samples from across a range of petroleum categories, a large proportion of the samples were from the gas oils and heavy fuel oils categories. If further information becomes available from studies conducted on substances from HPV petroleum categories other than gas oils and heavy fuel oils, this would provide additional support for their use across all petroleum substances in which PAC content may define toxicity.

A comparison of model predicted and estimates of response derived using traditional methods, shows a high degree of consistency, 82% of 125 comparisons were judged "consistent". However, for an individual sample, there may not be agreement between the $Estimate_{10}$ and PDR_{10} values for all four individual endpoints of toxicity. The authors do not think these inconsistencies for individual endpoints values should prevent the use of the models as a screening tool for untested substances. In the screening of an untested substance, each endpoint would not likely be evaluated separately, but the lowest PDR_{10} value from among all four endpoints would be used to characterize the sample.

There are two circumstances where the ARC 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 platelet count for every 100 mg/kg_{bw}/day increase in dose, in this case the prediction is contrary to what is expected (i.e. a *decrease* in platelets is expected) and the predicted effect is large. Because the ARC models are complex and have been built with a relatively small number of materials (individual PAC profiles), there may be areas within the PAC 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 for the data poor regions. In the future, as new test data become available, the results can be incorporated into the current ARC models, further corroborating them and expanding the domain of applicability.

The authors envision that possible uses for the ARC models described in this paper, as well as for the ARC models for developmental toxicity endpoints (Murray et al., 2013a; Nicolich et al., 2013) include:

Assigning an overall PDR_{10} value to a sample

For each sample, the repeat-dose ARC models provide PDR_{10} values for four sensitive endpoints of repeat-dose toxicity. Among the four values for each sample, the lowest PDR_{10} value could be designated as the sample PDR_{10}s. This concept can be expanded to include the PDR_{10} values estimated from the three developmental toxicity ARC models (Murray et al., 2013a).

Placing untested substances in broad categories of toxicity

Although the PDR_{10} does not necessarily represent an adverse effect level, it could nevertheless be used to place petroleum substances of similar biological activity into broad categories. For example, if the PDR_{10}s of a range of previously untested petroleum substances were determined, a range of PDR_{10}s would result. Such a range would allow the substances to be sorted into groups with lower or higher PDR_{10} ranges.

Identifying potential variability in toxicity among samples with the same CAS RN or in the same category

As noted earlier, 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). These differences in composition between samples may in turn produce variations in toxicity among samples of HBPS with either the same CAS number or in the same category. For example, the predicted sample PDR_{10} values for repeat-dose toxicity endpoints for 46 crude oil samples (CAS 8002–05-9) ranged from 80 to 560 (McKee et al., 2013a). The use of PDR_{10} values allowed comparisons to be made across the 46 samples without the unnecessary use of animals and resources.

Predicting that an untested substance will have toxicity similar to a specific tested substance

Read across is recognized as a way to predict the toxicity of an untested substance, provided there is an established relationship between the source and destination of the information. In the case of untested HBPSs, the PDR_{10}s for the sensitive endpoints could be matched to the BMD_{10}s of a tested substance. In such cases, it is not unreasonable to assume that the untested substance would have biological properties that are similar to the material for which data are available. This approach could be used to support the use of "read across" to predict the likely effects on those endpoints that have not been modeled (e.g. histopathological changes).

A variation of this has already been done for crude oil (McKee et al., 2013a). Model predicted PDR_{10} values for repeat-dose and developmental toxicity endpoints were estimated for 46 samples of crude oil. The model predicted PDR_{10}s indicated that the empirical data from a previously tested crude oil approximated a "worst case" situation, and that test data on that sample could be used to characterize the systemic and developmental toxicological hazards of crude oils in general.

Identifying and prioritizing those HBPSs that require further evaluation

During the evaluation of the toxicity of a single category of petroleum substances, it is necessary to identify the boundaries of the category in terms of toxicity. The PDR_{10} values could be used to identify those category members that are likely to be the most or least active, based on low and high PDR_{10}s. Such information would allow the selection of samples for further testing in order to adequately define the boundaries of the category. Similarly, if multiple samples of the same HBPS were available for toxicity testing, PDR_{10}s could be employed to identify the sample chosen for use in the toxicity tests.

Selecting doses for use in toxicity testing

The ARC models allow the prediction of the dose–response curve for a petroleum substance for which an ARC profile is available. This will assist toxicologists to design experimental 90-day or developmental toxicity studies on petroleum substances that fall within the model domains more efficiently. Since an available predicted dose response curve may allow dose selection to be more efficient, this may reduce the need for preliminary dose range-finding studies before embarking on a full definitive study.

5. Conclusions

The current review and evaluation shows there was an association between a substance's DMSO-extracted wt.% of each ring-class of the 1- through 7-ring compounds (the "ARC profile") and effects on selected repeat-dose endpoints. Predictive models based on these associations were developed for effects on four repeat-dose toxicity endpoints (absolute thymus weight, relative liver weight, hemoglobin concentration and platelet count). The models generate values that are consistent with other standard measures. The authors think the models described in this paper may have use in the prediction of repeat-dose toxicity for untested HBPS and the selection of such substances for biological testing.

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) was, at the time this work was carried out, 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 U.S. EPA HPV Challenge Program. The American Petroleum Institute (API) manages the PHPVTG's activities.

References

Abadin, H.G., Murray, H.E., Wheeler, J.S., 1998. The use of hematological effects in the development of minimal risk levels. Regul. Toxicol. Pharmacol. 28, 61–66.

Abadin, H.G., Chou, C.-H.S.J., Llados, F.T., 2007. Health effects classification and its role in the derivation of minimal risk levels: immunological effects. Regul. Toxicol. Pharmacol. 47, 249–256.

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

Amacher, D.E., Schomaker, S.J., Burkhardt, J.E., 1998. The relationship among microsomal enzyme induction, liver weight and histological change in rat toxicology studies. Food Chem. Toxicol. 36 (9–10), 831–839.

API (American Petroleum Institute), 2002. Petroleum HPV Testing Group; waxes and related materials category HPV test plan, August 6, 2002. Posted to EPA

Website Aug 22, 2002; <http://www.epa.gov/hpv/pubs/summaries/wxrelmat/c13902tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2003a. Petroleum HPV Testing Group; lubricating oil basestocks category HPV test plan, March 24, 2003. Posted to EPA Website April 4, 2003; <http://www.epa.gov/hpv/pubs/summaries/lubolbse/c14364tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2003b. Petroleum HPV Testing Group; gas oils category HPV test plan, November 3, 2003. Posted to EPA Website December 16, 2003; <http://www.epa.gov/hpv/pubs/summaries/gasoilct/c14835tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2003c. Petroleum HPV Testing Group; crude oil category HPV test plan, November 21, 2003. Posted to EPA Website December 19, 2003; <http://www.epa.gov/hpv/pubs/summaries/crdoilct/c14858tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2003d. Petroleum HPV Testing Group; reclaimed substances category HPV test plan, December 15, 2003. Posted to EPA Website January 20, 2004; <http://www.epa.gov/hpv/pubs/summaries/resbscat/c14906tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2003e. Petroleum HPV Testing Group; asphalt category HPV test plan, December 15, 2003. Posted to EPA Website January 20, 2004; <http://www.epa.gov/hpv/pubs/summaries/asphlcat/c14901tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2004. Petroleum HPV Testing Group; heavy fuel oils category HPV category test plan, June 17, 2004. Posted to EPA Website July 2, 2004; <http://www.epa.gov/hpv/pubs/summaries/heavyfos/c15368tp.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2008. Petroleum HPV Testing Group; aromatic extracts category HPV revised test plan, September 11, 2008. Posted to EPA Website March 11, 2003; <http://www.epa.gov/hpv/pubs/summaries/aroexcat/c14900rt.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2009. Petroleum HPV Testing Group; asphalt analysis and hazard characterization document, July 14, 2009. Posted to EPA Website January 6, 2011; <http://www.epa.gov/hpv/pubs/summaries/asphlcat/c14901ad2.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2010. Petroleum HPV Testing Group; category assessment document: reclaimed petroleum hydrocarbons: residual hydrocarbon wastes from petroleum refining, August 30, 2010. Posted to EPA Website November 15, 2010; <http://www.epa.gov/hpv/pubs/summaries/recpethy/c14755ca1.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2011a. Petroleum HPV Testing Group; crude oil category assessment document, January 14, 2011. Posted to EPA Website February 3, 2011; <http://www.epa.gov/hpv/pubs/summaries/crdoilct/c14858ca.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2011b. Petroleum HPV Testing Group; lubricating oil basestocks category assessment document, April 5, 2011. Posted to EPA Website May 3, 2011; <http://www.epa.gov/hpv/pubs/summaries/lubolbse/c14364ca.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2011c. Petroleum HPV Testing Group; waxes and related materials category assessment document, January 21, 2011. Posted to EPA Website February 4, 2011; <http://www.epa.gov/hpv/pubs/summaries/wxrelmat/c13902ca2.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2011d. Petroleum HPV Testing Group; gas oils category assessment document, August 31, 2011. Posted to EPA Website September 8, 2011; http://www.epa.gov/hpv/pubs/summaries/gasoilct/c14835cad.pdf (accessed June 6, 2012).
API (American Petroleum Institute), 2011e. Petroleum HPV Testing Group; heavy fuel oils category assessment document, July 12, 2011. Posted to EPA Website November 30, 2011; <http://www.epa.gov/hpv/pubs/summaries/heavyfos/c15368hc.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2012a. Petroleum HPV Testing Group; aromatic extracts category analysis and hazard characterization document, May 18, 2012; <http://www.petroleumhpv.org/docs/aromatic_extracts/2012_may21_Aromatic%20extracts %20category%20final%20May%2018 %202012.pdf> (accessed June 6, 2012).
API (American Petroleum Institute), 2012b. Petroleum HPV Testing Group; robust summary of information on aromatic extracts, May 21, 2012; <http://www.petroleumhpv.org/docs/aromatic_extracts/2012_may21_Aromatic%20Extracts_RS_2012_May_21_FINAL.pdf> (accessed June 6, 2012).
ATSDR (Agency for Toxic Substances and Disease Registry), 1995. Toxicological Profile for Polycyclic Aromatic Hydrocarbons (PAH). US Department of Health and Human Services, Public Health Services, Atlanta, GA.
ATSDR (Agency for Toxic Substances and Disease Registry), 1996. Minimal risk levels for priority substances and guidance for derivation. Federal Register 61, 25873–25882.
ATSDR (Agency for Toxic Substances and Disease Registry), 2006. Guidance for the preparation of a twelfth set toxicological profile (MRLs). Personal Communication from Dr. Selene Chou, April 6.
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.
Crump, K., 1984. A new method for determining allowable daily intakes. Fund. Appl. Toxicol. 4, 854–871.
Cruzan, G., Low, L.K., Cox, G.E., Meeks, J.R., Mackerer, C.R., Craig, P.H., Singer, E.J., Mehlman, M.A., 1986. Systemic toxicity from subchronic dermal exposure, chemical characterization, and dermal penetration of catalytically cracked clarified slurry oil. Toxicol. Ind. Health 2, 429–444.
Dalbey, W., Lock, S., Garfinkel, S., Jenkins, R., Holmberg, R., Guerin, M., 1982. Inhalation exposures of rats to aerosolized diesel fuel. In: MacFarland, H.N., Holdsworth, L.E., MacGregor, J.A., Call, R.W., Kane, M.L. (Eds.), Proceedings of the Symposium on the Toxicology of Petroleum Hydrocarbons. API, Washington, DC, pp. 13–25.
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.
Felter, S., Dourson, M., 1998. The inexact science of risk assessment (and implications for risk management). Human Ecol. Risk Assessment 4 (2), 245–251.
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., 1996. Systemic and developmental toxicity of dermally applied distillate aromatic extract in rats. Fund. Appl. Toxicol. 30, 276–284.
Feuston, M.H., Hamilton, C.E., Mackerer, C.R., 1997a. Oral and dermal administration of clarified slurry oil to male C3H mice. Int. J. Toxicol. 16, 561–570.
Feuston, M.H., Low, L.K., Hamilton, C.E., Mackerer, C.R., 1997b. Systemic and developmental toxicity of dermally applied syntower bottoms in rats. Fund. Appl. Toxicol. 35, 166–176.
Firriolo, J.M., Morris, C.F., Trimmer, G.W., Twitty, L.D., Smith, J.H., Freeman, J.J., 1995. Comparative 90-day feeding study with low-viscosity white mineral oil in Fischer-344 and Sprague–Dawley-derived CRL:CD rats. Toxicol. Pathol. 23, 26–33.
Gift, J., Howard, A., Zhao, J., 2011. Benchmark dose modeling and its use in risk assessment. <http://www.epa.gov/ncea/bmds/training/captivate/bmd-intro/bmd-intro.html> (updated on Tuesday, September 13, 2011. accessed September, 2011).
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.
IPCS (International Programme on Chemical Safety), 1998. Environmental Health Criteria: 202: Selected Non-Heterocyclic Polyaromatic Hydrocarbons. WHO, Geneva, <http://www.inchem.org> (accessed November, 2011).
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., Nicolich, M., Roy, T., White, R., Daughtrey, W., 2013a. Use of a statistical model to predict the potential for repeated-dose and developmental toxicity of dermally-administered crude oil and relation to reproductive toxicity. Int. J. Toxicol.
McKee, R.H., Schreiner, C.A., Nicolich, M.J., Gray, T.M., 2013b. The Relationship between Repeat-Dose Toxicity and Aromatic-Ring Class Profile of High-Boiling Petroleum Substances. Regul. Toxicol. Pharmacol. 67 (2S), S75–S85.
Murray, F.J., Roth, R.N., Nicolich, M.J., Gray, T.M., Simpson, B.J., 2013a. The relationship between developmental toxicity and aromatic-ring class profile of high-boiling petroleum substances. Regul. Toxicol. Pharmacol. 67 (2S), S46–S59.
Murray, F.J., Gray, T.M., Roberts, L.G., Roth, R.N., Nicolich, M.J., Simpson, B.J., 2013b. Evaluating the male and female reproductive toxicity of high-boiling petroleum substances. Regul. Toxicol. Pharmacol. 67 (2S), S60–S74.
Nessel, C.S., Freeman, J.J., Forgash, R.C., McKee, R.H., 1999. The role of dermal irritation in the skin tumor promoting activity of petroleum middle distillates. Toxicol. Sci. 49, 48–55.
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.
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.
Pohl, H.R., Chou, C.-H.S.J., 2005. Health effects classification and its role in the derivation of minimal risk levels: hepatic effects. Regul. Toxicol. Pharmacol. 42, 161–171.
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.
Roy, T.A., Blackburn, G.R., Deitch, R.A., Schreiner, C.A., Mackerer, C.R., 1985. In: Polynuclear Aromatics Hydrocarbons: A Decade of Progress. Estimation of Mutagenic Activity and Dermal Carcinogenic Activity of Petroleum Fractions Based on Polynuclear Hydrocarbon Content. pp. 809–824
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.
SCF (Scientific Committee on Foods), 2002. Opinion of the Scientific Committee on Food on the risks to human health of polycyclic aromatic hydrocarbons in food. Scientific Committee on Food, Brussels, Belgium. <http://www.efsa.europa.eu/en/efsajournal/doc/724.pdf> (accessed November, 2011).

Schwartz, J.A., Aldridge, B.M., Lasley, B.L., Snyder, P.W., Stott, J.L., Mohr, F.C., 2004. Chronic fuel oil toxicity in American mink *(Mustela vison)*: systemic and hematological effects of ingestion of low-concentration of bunker C fuel oil. Toxicol. Appl. Pharmacol. 200, 146–158.

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 July 23, 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/pac.html> (accessed 19 February 2013).

Skyberg, K., Skaug, V., Gylseth, B., Pedersen, J.R., Iversen, O.H., 1990. Subacute inhalation toxicity of mineral oils, C15–C20 alkylbenzenes, and polybutene in male rats. Environ. Res. 53, 48–61.

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

US Dept Energy (U.S. Department of Energy), 2007. RAIS (The Risk Assessment Information System). Office of Environmental Management, Oak Ridge Operations (ORO) Office. <http://rais.ornl.gov> (accessed November, 2011).

US EPA (U.S. Environmental Protection Agency), 1995. Toxic Substances Control Act inventory representation for certain chemical substances containing varying carbon chain lengths (Alkyl Ranges Using the Cx-y Notation) (March 29, 1995); <http://www.epa.gov/opptintr/existingchemicals/pubs/tscainventory/alkyl-rg.pdf> (accessed June 6, 2012).

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

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

US EPA (U.S. Environmental Protection Agency) 2007. IRIS (Integrated Risk Index System), Office of Research and Development, National Center for Environmental Assessment, <http://www.epa.gov/iris/index.html> (accessed November, 2011).

US EPA (U.S. Environmental Protection Agency) 2012a. Benchmark dose technical guidance. EPA/100/R-12/001. Risk Assessment Forum, June.

US EPA (U.S. Environmental Protection Agency) 2012b. Glossary of terms, http://www.epa.gov/radon/glossary.html, last updated 2, August, 2012 (accessed 29 January 2013).