Report of the Peer Consultation on Relationship between PAC Profile and Toxicity of Petroleum Substances

Volume I

October 8-9, 2007
Northern Kentucky University METS Center
Erlanger, Kentucky

Peer Consultation Organized by:
Toxicology Excellence for Risk Assessment
(http://www.tera.org/peer/)

January 28, 2008

NOTE

This report was prepared by scientists of Toxicology Excellence for Risk Assessment (*TERA*) and reviewed by the panel members. The members of the panel served as individuals on this panel, representing their own personal scientific opinions. They did not represent their companies, agencies, funding organizations, or other entities with which they are associated. Their opinions should not be construed to represent the opinions of their employers or those with whom they are affiliated.

Table of Contents

(This page intentionally left blank)

Executive Summary

A panel of scientists with experience in petroleum chemistry, biostatistics, toxicology, risk assessment, structure activity relationships, and reproductive and developmental toxicology met October 8-9, 2007 to conduct an expert peer consultation on a report on the relationship between polycyclic aromatic compound (PAC) profiles and toxicity of petroleum substances. The meeting was open to the public and was webcast on the Internet for any interested person to observe. The PAC Analysis Task Group of the Petroleum High Production Volume (HPV) Testing Group prepared two documents and supporting materials for the panel review and discussion. The Petroleum HPV Testing Group is a consortium of manufacturers who are funding a voluntary data disclosure and toxicity testing program on certain petroleum substances in response to the U.S Environmental Protection Agency (EPA) HPV Chemical Challenge Program. The Petroleum HPV Testing Group has submitted test plans to EPA for 13 petroleum substances categories. For six of these (crude oil, gas oils, heavy fuel oils, lubricating oil base stocks, aromatic extracts, and waxes and related materials) a relationship between toxicity and PAC content had been asserted or implied in publicly available documents submitted to the United States Environment Protection Agency as part of the US EPA Chemical Challenge Program. This relationship has been further examined in the documents under consideration for this peer consultation. The purpose of the meeting was to have a diverse group of expert scientists review the approaches and conclusions and provide opinions on the merit and adequacy of the proposed methods, as well as suggestions for improvement.

The authors provided the panel with brief presentations to summarize the work they had done in three main areas – selection and evaluation of data sets and endpoints, identification and characterization of relationships between PAC content and toxicity to determine use for prediction of untested petroleum substances, and validation of the methods and use of results. The authors hypothesized that systemic toxicity and developmental and reproductive toxicity observed in repeated-dose dermal toxicity studies are associated with PAC content and that the PAC content can be used to predict the toxicity of an untested petroleum stream. The authors emphasized that this project is not investigating the mechanism of toxicity, but rather statistical relationships between PAC content and biological endpoints. The Task Group collected study reports from their member companies and analyzed these data in order to address two key questions:

- Are there quantitative relationships between PAC content of petroleum substances and their critical effects as identified in repeat-dose, developmental, and reproductive toxicity studies?
- Can the critical effects/levels of untested petroleum substances as would be identified in an OECD 422 study be predicted using their PAC content?

Panel members discussed available data, the proposed approach, models, results, and conclusions over the two-day meeting. The intention of the peer consultation was to solicit individual expert opinions and suggestions from the diverse group.

The panel discussed the selection of data sets, data points, and endpoints for the model development. They discussed the different analytical methods to characterize PAC content and data available for each. The panel members were generally satisfied that the analytical method chosen from those available was most appropriate. The panel recommended that the method, any relevant validation or testing data for the method, and the rationale for choosing it be better explained in the document.

Panelists discussed support for the assumption that PACs are the source of toxicity for these petroleum substances and the usefulness and adequacy of the available data to support this assumption. They discussed whether the wealth of existing mouse oral and other data on the toxicity of petroleum substances might be used and limitations in the range of the test materials used in the available dermal studies. They also discussed whether other components or characteristics of petroleum are part of the toxicity equation and if correlations could be explored with other compositional characteristics of these substances. The panel explored the possibilities and limitations of using other approaches to predict toxicity, such as relative potency, an additivity approach of the individual component's toxicity, or using parallel analysis to individual PAHs. Panelists noted that the analytical methods could not support component approaches and that these petroleum streams contain thousands of individual isomers of PACs, in comparison to the relatively simple mixtures of PAHs, for which relative potency factors have been estimated. Panelists also discussed that PACs interact in ways that are not fully understood and that an integrated approach is limited to demonstrating association and not causation since the complex toxicity mechanisms are unclear.

Panelists discussed the selection of data from the available studies that were used to build the models. In particular, they discussed the number of data points, the appropriateness of criteria used for excluding data, and the use of biological considerations in selection of the endpoints used for modeling.

Panel members discussed issues regarding maternal toxicity, and dermal toxicity (including dermal irritation). The need for more analysis and information regarding maternal toxicity was noted by panelists, to determine whether the effects seen were associated with maternal toxicity or if the substances are selective developmental toxicants. Panel members suggested that the authors address maternal toxicity as a potential endpoint and conduct analyses to determine if developmental toxicity was associated with or occurred at the same exposure levels that produced signs of maternal toxicity. The panelists discussed dermal irritation and its potential influence on maternal health status, reproduction, and fetal effects. They suggested that the authors further explore dermal toxicity and determine if direct effects on the skin are an important endpoint to include in evaluation of PAC-associated reproductive failure. In general, the panel found the data set selection procedure and range of endpoints included to be appropriate, although individual panel members recommended further evaluation of some endpoints.

The authors explained how they developed the model. They started with a simple linear model and attempted numerous transformations of independent and dependent variables to increase statistical fit of the model. A factor analysis was conducted in an attempt to optimize the

variables, but there was no gain in predictive ability. Ultimately, the selected variables were based on comparative goodness of fit and general considerations related to PAC content and biological variables that were expected to be important predictors of the response.

Panel members discussed the statistical methods used to build the model and the alternative methods that were considered. Panelists raised questions about the general model structure and recommended better documentation of the rationales for the decisions made in the modeling, as many important considerations and tests that were performed in building the models did not appear to be fully described in the document. Panelists also raised issues regarding handling of control response, and use of mean response data versus individual data. They noted that with so many parameters, multicolinearity is important and suggested that additional presentation of mixed model results and sensitivity analysis would be helpful. Panelists suggested that the authors better describe the various approaches and exercises they used to develop the model and discuss how alternative modeling approaches improve or worsen model fit. They also suggested that a non-mathematical explanation of the model be included to help some readers understand these difficult concepts.

The authors used the models to predict predefined levels of response (PDx), which were developed based on informed professional judgment. The panelists discussed the definition of "toxicologically meaningful degree of change" and the basis for the response percentages selected for the evaluation, questioning the support for selection of one percentage value over another. Panelists had concerns over the biological and statistical basis for the authors' selections and concluded that the choice of what degree of change that is biologically significant is beyond the scope of the authors' charge and gets into risk assessment applications beyond the intent of the HPV program. Panelists suggested that the authors need not specify specific toxicologically-meaningful values in their report; rather, they should demonstrate the approach with multiple or hypothetical values and let the user to decide what value is most appropriate for the situation of interest. The development of an approach based on control response variability also could be considered so as to define an abnormal, not necessarily adverse, range for effects that are measured on a continuous scale. Then, PDx could be defined as a specified proportion of individuals in the abnormal range. Panelists concluded that the selection of PDx values, use of the PDx rather than benchmark response, and interpretation of results will likely be controversial, and these issues should be presented separately so as to not detract from the models themselves.

Panelists discussed reproductive toxicity and noted that the data set was inadequate to support the conclusion that developmental toxicity is more sensitive than reproductive toxicity and to conclude that developmental toxicity is likely to be a reasonably good predictor of reproductive toxicity. The panel discussed the minimum data set that would be needed to predict potential reproductive toxicity and how studies might be best designed.

The model validation was carried out in three phases: data splitting techniques; randomized pairing of independent and dependent variable sets; and, application of similar toxicological endpoints between the models where such data were available. The authors found the model did

very well on all three validation steps, and worked well for interpolated substances, but not as well for extrapolated.

The panel discussed the validation methods used and explored other possibilities for further validation or confirmation of the models. Some thought validating or confirming the model with additional data that had not been used in the development of the model or its parameters, or with data from sources that did not use Method 2 was needed. Some thought that new data or studies are needed, while others suggested looking for confirmatory or existing published or proprietary data. Other panelists were comfortable with the existing validation work, but suggested the authors might build the developmental toxicity models with a portion of the available data and test it with the remaining data. A sensitivity analysis was specifically recommended by a number of panelists, while others did not think it would improve the model or change the results.

The concept of interpolation and extrapolation was discussed at length. The authors clarified that the 7 - ring profile for the new substance has to be totally inside one of the profiles for the substances used to build the model to be considered an interpolation. To be considered an extrapolation, any one component of the 7 - ring profile has to fall outside of all of the corresponding components of the existing profiles upon which the model was built. The authors noted that the interpolation concept also includes the consideration of dose. In addition, the authors noted that the models only applied to petroleum substances whose boiling ranges covered the boiling ranges of PAC. One panelist pointed out that use of the interpolation and extrapolation concepts contributes to the strength and validity of the model, and that the document identifies important limitations of the model based on testing of interpolated versus extrapolated data.

The panel discussed whether the models can be used to predict repeat-dose and developmental toxicity of PAC-containing petroleum substances for purposes of the HPV program. Panel members discussed the models uses for HPV and also what constitutes a screening level toxicity analysis for HPV purposes. Some thought that if this work is for screening and prioritization in HPV, the current effort may be sufficient as is for these endpoints. Others panel members noted that additional validation of the general toxicity models and further work related to assessing correlations between maternal toxicity, dermal irritation, and developmental effects would be needed before the models could be used for HPV screening purposes. Some panel members felt that additional data related to potential reproductive toxicity would also be needed before making conclusions regarding that endpoint.

The panelists cautioned that when presenting the modeled predictions in the context of HPV, it should be made explicit which values are calculated from the models, with the concept and meaning of interpolated and extrapolated carefully explained. Panel members strongly cautioned that the model and results are not appropriate to use in quantitative risk assessments and that the documentation should state this.

Throughout the meeting, panelists suggested areas where further explanation was needed to help the reader understand what the authors have done and to make their work more transparent. Panelists suggested that more information was needed to provide context for the model,

including more information about the HPV program and how the modeling results would be used in that program. The authors should better document the model form and make the domain of the model very clear, as well as limitations and inappropriate uses of the model. In addition, specific suggestions were made related to providing a roadmap in the document that links the related data sets to the results of the analyses and the conclusions derived from those analyses.

Panelists agreed with the authors that interpretation and use of the model must be caveated to prevent misuse of the model or results. A panelist noted that the model does a good job of estimating various biological effects based on PAC profile, but this does not mean the PAC profile is the cause of toxicity; there may be an unknown factor involved. The authors have justified an empirical approach and built a correlation model based on associations. This needs to be stated unambiguously and the authors need to be careful that they do not imply or claim causation in the text. The panel discussed presenting significance levels for the correlations and noted the inappropriateness of trying to identify the parameters that contribute the most. Because this work is a correlational study, mechanistic based conclusions should be avoided.

In general, the panelists were positive and comfortable with the model, the general approach, and the statistical analyses that were done. Many mentioned that they thought it showed great promise and some stated that they thought it was close to being ready for regulators, while others thought the model and confirming assays should be published as a scientific paper in the open literature first, to gain wider exposure and input.

(This page intentionally left blank)

1. Participants

Sponsor

American Petroleum Institute

Presenters

Mr. Barry J. Simpson, PAC Analysis Task Group

Dr. F. Jay Murray, PAC Analysis Task Group

Dr. Mark J. Nicolich, PAC Analysis Task Group

Dr. Randy Roth, PAC Analysis Task Group

Peer Consultation Panel Members[1]

Dr. Caroline Baier-Anderson
Environmental Defense

Dr. John DeSesso
Noblis, Inc.

Mr. Stephen D. Emsbo-Mattingly
New Fields Environmental Forensics Practice, LLC

Dr. David Gaylor
Gaylor and Associates, LLC

Dr. Andrew Maier
Toxicology Excellence for Risk Assessment

Dr. Sati Mazumdar
University of Pittsburgh

Dr. Andrew Nicholson
Geomega

[1] Affiliations listed for identification purposes only. Panel members served as individuals on this panel, representing their own personal scientific opinions. They did not represent their companies, agencies, funding organizations, or other entities with which they are associated. Their opinions should not be construed to represent the opinions of their employers or those with whom they are affiliated.

Dr. Robert Scala
Consultant

Dr. Calvin Willhite
State of California

Facilitator

Dr. Michael Dourson
Toxicology Excellence for Risk Assessment

Observers and Other Attendees

A list of observers and other attendees is found in Appendix A.

2. Background

This peer consultation meeting was organized by Toxicology Excellence for Risk Assessment (*TERA)*. *TERA* is an independent non-profit organization with a mission to protect public health through the best use of toxicity and exposure information in the development of human health risk assessments. *TERA* has organized and conducted peer review and peer consultation meetings for private and public sponsors since 1996.

This meeting was a peer consultation, organized for the purpose of providing expert input and advice. The objective of this peer consultation was for a diverse group of appropriate experts to review the approaches and conclusions presented in the sponsor's documents and provide opinions on the merit and adequacy of the proposed methods. The most important goal in organizing a peer consultation is to locate the scientists with the needed expertise on the issues, and *TERA* intentionally sought out knowledgeable individuals from a variety of organizations and those with a backgrounds in petroleum chemistry and toxicology, evaluating alternatives to animal toxicity testing, and the HPV Challenge Program, so that a broad range of perspectives would be included on the panel.

The PAC Analysis Task Group of the Petroleum HPV Testing Group prepared two documents and supporting materials for the panel review and discussion. The Petroleum HPV Testing Group is a consortium of manufacturers who are funding a voluntary data disclosure and toxicity testing program on certain petroleum substances in response to EPA's HPV Chemical Challenge Program. The American Petroleum Institute (API) manages the consortium activities. The Petroleum HPV Testing Group has submitted test plans to EPA for 13 petroleum substances categories. For six of these, a relationship between toxicity and PAC content was asserted or implied and this relationship has been further examined in the documents under consideration for this peer consultation. The Task Group prepared several document and appendices for the panel's review during this peer consultation. These are available on the Internet at http://www.tera.org/peer/API/APIWelcome.htm .

The peer consultation panel included nine scientists who have expertise in the key disciplines necessary to evaluate the proposed approach. Each panelist is a well-respected scientist in his or her field. The panel members have training and experience in petroleum chemistry, biostatistics, toxicology, risk assessment, structure activity relationships, and reproductive and developmental toxicology. *TERA* asked the sponsors and several interested parties for suggestions of experts. *TERA* considered these suggestions and independently identified other candidates. It was from this larger pool of candidates, that the panel was selected. *TERA* was solely responsible for the selection of the panel members. Each panel member has disclosed information regarding potential conflicts of interest and biases related to the proposed approach and its sponsors. *TERA* carefully evaluated these disclosures when selecting panel members. Short biographical sketches and disclosure statements for panel members are provided in Appendix B.

Members of the public were invited to observe the panel discussions by attending the peer consultation meeting in person or by viewing a live web cast of it. They were also given the opportunity to provide brief oral and written technical comments on the assessment document for the panel's consideration. No written public comments were received prior to the meeting, but one written question was received from a web cast participant during the meeting, and it is discussed in this report.

TERA prepared this meeting report. The report summarizes the sponsor's presentations, the panel discussions, the sponsor's comments during the discussions, and comments from the public. The meeting report is a summary, not a transcript. Opinions and recommendations of the panel members are noted (although panelists are not identified by name). Panel members have reviewed the draft report, and their comments and corrections have been incorporated into this final version. The sponsors also were given the opportunity to review the draft report to confirm the accuracy of their presentations and remarks. This report is available on the Internet at http://www.tera.org/peer/API/APIWelcome.htm.

This report is organized into three parts that correspond with the charge to the panel: selection and evaluation of data sets and endpoints, relationship between PAC content and biological endpoints, and validation of methods and use of results. Listed below are the charge questions that formed the basis for the panel's discussions. Note that the panelists discussed these questions over a period of two days and some topics were discussed at multiple points during the meeting. In this meeting report we attempted to group the discussions by topic and placed the summary of the discussions under the charge question that seemed to fit the best.

A. Selection and Evaluation of Data Sets and Endpoints

1. Is the Method 2 analytical procedure described in this project a reliable and accurate method of determining the PAC profile of the categories of petroleum substances referenced in this report?

2. The authors made an assumption that PACs are the source of toxicity for petroleum substances that contain PACs. Did the authors have adequate sampling to test and support

this? Was the decision to conduct the statistical analyses based on PAC analysis method 2 supported by the data? Could data sets from studies using other analytical methods contribute significantly to the analyses? Does the description of the chemical composition of petroleum substances support the conclusion that toxicity evaluation methods based on individual PAH content cannot be used?

3. Discuss the criteria and procedures used for identification, inclusion, and exclusion of toxicological data sets for the modeling. Were the criteria and procedures fully described and are they defensible? Is it reasonable to assume that all the relevant data have been collected and accurately compiled and analyzed? Are there other data sets that should have been considered? Did the exclusions make statistical and biological sense and were the impacts of such exclusions adequately explored? Would the procedures used in data set selection generate any bias in the results?

4. Were the procedures for selecting biological endpoints for modeling adequately communicated? Were the methods adequate to identify all key relevant endpoints? Was the final selection of endpoints fully justified by the data including the analyses conducted? Should other endpoints be considered based on the current state of knowledge of PAH and petroleum stream toxicity?

5. Are there other important issues to discuss regarding the selection and evaluation of data sets?

B. Identifying and characterizing the relationships between PAC content and mammalian toxicity (SIDS endpoints) and determining if they could be used to predict the toxicity of untested petroleum substances.

6. Discuss whether the statistical methods were appropriate and adequate and if the procedures were implemented correctly. Would other valid statistical approaches yield different results? Would alternative model development approaches have improved the models and results? Were the authors' conclusions regarding the models clearly articulated and justified by the results?

7. The authors identified the toxicologically meaningful degree of change for each general toxicological endpoint (body weight and liver weight, hematology changes, thymus weight change). For each of the endpoints discuss whether these values are toxicologically and physiologically meaningful and if the best value was chosen. Are the observations consistent with what is known about petroleum toxicity? Are there other important issues regarding the relationship between PAC content and general systemic toxicity?

8. The authors identified the toxicologically meaningful degree of change for each developmental toxicity endpoints. For each of the endpoints discuss whether these values are toxicologically and physiologically meaningful and if the best value was chosen. Are the observations consistent with what is known about petroleum toxicity?

9. The authors conclude that the pre-defined change (PDx) for developmental toxicity will be a reasonably good predictor of the PDx for reproductive toxicity. Is this conclusion valid?

C. Validation of Methods and Use of the Results

10. The authors present predicted dose-response curves and compared these to actual results of the study from which the information had been derived. How accurately do the predicted dose-response curves fit the observed data, and how do the predicted PDx effect levels compare with the endpoint Lowest Observed Adverse Effect Levels/Lowest Observed Effect Levels (LOAELs/LOELs) observed in the actual studies?

11. Discuss the model validation methods. Could additional validation approaches be used to enhance confidence in the model?

12. Are the conclusions reached by the authors regarding utility of the models for interpolation versus extrapolation justified? Are the presented definitions and procedures adequate to identify data sets that can be accurately predicted by the proposed models?

13. Are the conclusions in Volume 2 Section 5 biologically plausible, supported by the data, and do they reflect sound statistical analysis?

14. Discuss the models' strengths and limitations. Were they clearly identified and the implications well described? Are there ways to ameliorate the weaknesses? Is the documentation transparent and complete? Are uncertainties in the approach fully articulated? Are there suggestions for improving the presentation of the analyses or information?

15. Can the models that were developed be used to predict repeated-dose and developmental toxicity of PAC-containing petroleum substances for purposes of the HPV program?

16. Are there other important issues regarding model development, validation, and use?

3. Introductions, Conflict of Interest, and Meeting Process

The meeting opened with a welcome by Ms. Jacqueline Patterson of *TERA*. She described the background and purpose of the peer consultation and the agenda for the meeting. Ms. Patterson noted that copies of panel members' biographical sketches and conflict of interest (COI) and bias disclosure statements were provided to all attendees (see Appendix B). The panel members then introduced themselves and noted whether they had additions or changes in their disclosure statements. None of the panel members had any changes to their statements.

Dr. Dourson, the panel facilitator, then described how the meeting would be conducted. He explained that discussions would be based on the items found in the Charge to the Panel (located in Appendix B). He noted that all panelists would have the opportunity to state their own positions on the charge items, to ask one another clarifying questions, and to further discuss the issues. Because this was a peer consultation, no formal attempt would be made to reach panel consensus positions on the charge items.

4. Introduction

4.1 Author Presentation

Mr. Barry Simpson of the Task Group provided background on the High Production Volume (HPV) Chemical Challenge Program and the Petroleum HPV Testing Group's efforts for petroleum substances. He noted that HPV requires producers/importers to identify data lacking from the SIDS (Screening Information Data Set) data set and provide a test plan to fill the data gaps. The U.S. Environmental Protection Agency (EPA) encourages categorizing materials that are related in some regular fashion to maximize usefulness of available data and to reduce the numbers of animal studies that may otherwise be required. With over 400 petroleum substances, the producers have grouped them into 13 categories; six are the subject of the present analysis and proposed method (crude oil, gas oils, heavy fuel oils, lubricating oil base stocks, aromatic extracts, and waxes and related materials). These categories follow those used by the European Union and contain polynuclear aromatics at increasing concentrations as the boiling temperature increases from about 300-400 degrees Fahrenheit. SIDS data are not available for all of the 400 sponsored petroleum substances. Substances described by the same CAS number may differ compositionally and may differ in toxicity. Therefore, a range of toxicity values may be required to capture the variability for the range of mixtures covered by a specific CAS number.

In some of the test plans submitted by the Petroleum HPV Testing Group, the Group hypothesized that repeat dose, genetic, developmental, and reproductive toxicity were associated with polycyclic aromatic compound (PAC) content and that the PAC content could be used to predict the toxicity of an untested petroleum stream. The Task Group addressed two questions:

- Are there quantitative relationships between PAC content of petroleum substances and their critical effects as identified in repeat-dose, developmental, and reproductive toxicity studies?
- Can the critical effects/levels of untested petroleum substances as would be identified in an OECD 422 study be predicted using their PAC content?

To answer these questions the Task Group issued a call for data from their member companies. Two member companies provided copies of study reports. The studies were conducted in the 1980s and 1990s and generally followed current guidelines and met EPA data quality requirements. Volume 1 of the review materials provides background on the HPV program and the project. Volume 2 is the report of the Task Group and the subject of the peer consultation. The authors also provided some written clarification on a number of points prior to the meeting

(see Appendix C). All the documents and appendices can all be found at http://www.tera.org/peer/API/APIWelcome.htm

4.1.1 Clarifying Questions from the Panel

In response to a number of clarifying questions from the panelists, Mr. Simpson explained that the petroleum substances covered by this approach consist of the petroleum streams that are the components of finished products and therefore do not apply to products that include additives. He noted that when the characterization of the petroleum streams is complete, further testing under HPV will not be required and they expect they will have captured the extremes of the composition and resulting toxicity. Mr. Simpson also explained the Volume 3 of the series (not the subject of this particular peer consultation) will address mutagenicity, but not carcinogenicity (carcinogenicity is not an endpoint for SIDS).

A panelist noted that crude oil differs somewhat with geography, particularly with metal content, and asked whether metal content was considered. Mr. Gray noted that metal content was considered in initial stages to answer whether it could contribute to the observed toxicity. The types of toxicity observed with petroleum products is more consistent with effects demonstrated by pure PAC. Metals in crude oil and fractions include nickel, vanadium, iron, and metalloid sulfur, but they are bound (in two types of compounds - metallo-petroporphyrins and metallo-nonporphyrins) and therefore may have limited impact on toxicity.

Another panelist asked whether the authors considered absorption kinetics, to see if there are matrix or vehicle effects that affect bioavailability of these diverse substances. Mr. Simpson explained that as they were seeing systemic toxicity, they felt reasonably comfortable that the substances were being absorbed. They also evaluated viscosity, but found that it did not have an effect on their results. Dr. Nicolich further noted that the authors were not investigating the mechanism of toxicity, but were more interested in the basic statistical relationship between PAC exposure and the biological response. The authors examined physical properties related to PAC composition as explanatory variables in preliminary model fitting, but found that the PAC composition itself was sufficient to yield significant correlations. Thus, the simpler models were carried forward.

5. Selection and Evaluation of Data Sets and Endpoints

5.1 Author Presentations

Drs. Randy Roth and Jay Murray of the Task Group provided information on selection and evaluation of data sets and endpoints. They sought to be transparent regarding selection of data and attempted to eliminate any potential bias resulting from data selection. The authors explained how they acquired the studies and narrowed the data set to rat dermal studies ultimately used in the analysis and model development. Drs. Roth and Murray noted that they used professional judgment to make decisions regarding endpoint selection and that the endpoints were identified without prior knowledge of PAC profile of the test samples and prior

to beginning the modeling efforts. The authors described their reasoning for eliminating studies from the data set. The authors explained their distinction between prenatal and postnatal developmental toxicity studies, noting that both types involved gestational exposure, but for the "prenatal" studies fetuses were removed prior to parturition and examined; while the dams in studies labeled postnatal were allowed to deliver and the neonates were observed. For the developmental toxicity studies, the most common endpoints at the LOAELs were indices of maternal toxicity and offspring survival or growth (e.g., body weight).

The authors discussed dermal irritation and how they addressed it in their work. They noted that the group of studies was from several laboratories over the course of two decades and there was not consistent reporting on dermal irritation. However, from their review of the data they did not see an obvious relationship in the study reports between dermal irritation and effects seen in the studies. They acknowledged that dermal irritation may play a role, but they did not have sufficiently robust data to investigate further.

The PAC content of a large number of the test samples used in the studies reviewed by the authors were chemically characterized by either Method 1 or Method 2 analyses (as referred to by the authors). In Method 1, individual refinery streams were fractionated by silica gel elution chromatography to determine gravimetrically the amount of nonaromatics and aromatics present. The aromatic fraction was further analyzed by electron impact mass spectrometry to quantitate and identify components based on the number of aromatic rings in their structures (1 to 5 ring PAC) or as S-PAC and unidentified aromatics. Method 2 involved analysis of a PAC-enriched dimethyl sulfoxide (DMSO) extracts of individual refinery streams. In essence, the fraction of aromatics isolated by Method 2 is a subset of what one would get from Method 1. The PACs isolated using Method 2 were quantified using procedures described below.

Dr. Tim Roy of Port Royal Research, LLC provided some background information on Method 2 on behalf of the Task Group. He explained that he was involved with the development of Method 2, which had its genesis with work on the modified Ames assay and the initial correlations between mutagenicity and the modified Ames test. Over time they developed and validated a variation of the Institute of Petroleum method 346 (IP 346) which involves a DMSO extraction of the petroleum product to concentrate the aromatic fraction followed by profiling of the aromatic extract by gas chromatography with flame ionization detection (GC/FID). They built Method 2 around that procedure and instituted basic analytical controls.

Method 2 fits into the general category of liquid-liquid fractionation procedures designed to enrich the aromatic fraction of petroleum products. The petroleum product is first dissolved in a hydrocarbon solvent (e.g., pentane, hexane, cyclohexane) and extracted with a polar solvent (e.g., furfural, dimethylsulfoxide). Water or saline (~2:1 by volume) is then added to the isolated polar solvent fraction which is back-extracted with a hydrocarbon solvent. Aromatics are recovered in the hydrocarbon layer, while polar compounds (e.g., acids, aldehydes, phthalates) are retained in the polar solvent-water layer. Removal of the hydrocarbon solvent and a gravimetric determination of the aromatic residue is the basis of IP-346 currently used in the EU for labeling of certain petroleum products. Method 2 or the PAC2 procedure analyzes the aromatic residue by gas chromatography using either a flame ionization detector or a mass

spectrometer to determine the distribution of aromatics according to ring number, 1-ring (benzenes) and 2-7 fused-ring polycyclic aromatic compounds

5.1.1 Clarifying Questions from the Panel

A panelist asked how many data points the models were built on after elimination of studies and data. Dr. Murray pointed to Tables 5 and 7 of Volume 2 where the n equals the number data points for each endpoint model (ranged from 34 to 128 in final models). Dr. Murray also confirmed that the criteria for study exclusion differed between the repeat dose and developmental toxicity studies. For the repeat dose studies they excluded the single mouse study and studies with only Method 1 analytical data. For developmental toxicity they used the same criteria, but also considered some aspects of study design, excluding studies that did not dose the animals for the entire duration of pregnancy (days 0-19) and dose groups with dams having very small litters (3 or fewer pups).

A panelist asked whether the elimination of data from dams with small numbers of litters resulted in loss of valuable information, perhaps on the most sensitive endpoint or maternal toxicity. An author explained that they were most interested in identifying the effects at doses near the no effect level and modeling for reduction of litter size with data at doses with high levels of resorptions would not allow them to accurately model data in the low dose range. The authors did not discount the possibility of maternal toxicity, but were attempting to model data in the low dose response range and therefore eliminated the high doses where there were high percentages of resorptions (high doses eliminated had 100% and 97% resorptions). A panelist noted that the typical procedure would be to model the data with and without these data sets, removing dose groups sequentially until the fit in the low-dose region no longer improves. This procedure ensures model fit, while not discarding data unnecessarily. The reviewer commented that this approach is important since removing the high effect groups can change the slope of the dose-response curve in a non-conservative direction.

A panelist sought further clarification on selection of endpoints and use of biological plausibility as a criterion. An author stated that this work is not a mechanism study; their goal was to see whether there was a statistical relationship between PAC content and biological endpoints. The endpoints the authors saw in this group of studies were not dissimilar to historical data with individual PACs, and therefore support a PAC-driven mode of action.

A panelist asked whether the authors checked to see how the eliminated endpoints fit with the model. The authors indicated that they did not do that exercise because their purpose was to see if a statistical relationship could be demonstrated for the data that met the selection criteria. Furthermore, regarding the excluded endpoints from the repeat dose studies, an author indicated that Table A4-1 in Appendix 4 lists all the biological endpoints captured from the study reports and Table 4 lists the endpoints affected most often (statistically) and those carried forward for the final model development. The author noted that there were no exclusions of data from the data extraction process for the repeat dose studies, although as noted in the report some subsets were not used in the ultimate development of the models.

The authors confirmed that they did not model skeletal ossification because reduced fetal weight captured the same underlying effect on fetal growth, and body weight was more robustly reported in terms of providing a quantitative continuous variable. However, for litter size and percent resorptions both of these related endpoints were used to make sure the results were consistent.

Several panelists sought clarification on the laboratories that conducted the studies. One asked what laboratories conducted the studies and how many studies were from each. A representative from the Task Group noted that the studies were submitted by the member companies and they would have to go back to the companies to get permission to identify the names of the laboratories. An author noted that the developmental toxicity studies came from two laboratories and about half from each, and that both laboratories used the same rat strains.

A panelist asked about completeness of the authors' literature searches prior to carrying out the analysis and whether this literature was captured in the project. An author responded that they intended to capture literature on petroleum substances and not on single PACs. They believe they captured the majority of that literature.

A panelist clarified that the group that developed the OECD 422 protocol never intended it to be applied to other than the oral route. The only published validation of the protocol was from this panelist's laboratory, and it was based on oral data. However, it was noted by the authors that the current OECD language indicates that the protocol can include other routes of exposure, including the dermal route.

5.2 Panel Discussion on Selection and Evaluation of Data Sets and Endpoints

The panel discussion on selection and evaluation of data sets and endpoints addressed five charge questions.

5.2.1 Charge Question 1

1. Is the Method 2 analytical procedure described in this project a reliable and accurate method of determining the PAC profile of the categories of petroleum substances referenced in this report?

Panel members discussed the relative merits of the various analytical methods and in particular Methods 1 and 2. A panelist noted that analytical Method 2 involves the extraction of PACs from the different petroleum products with DMSO. The question is whether that procedure is accurate and reliable. The panelist noted that the method has been published and peer reviewed, but as this report will be going to EPA, that Agency has criteria and standards for validation of analytical methods, and the authors should consider whether Method 2 meets those.

A panel member noted that Method 1 was eliminated because it focused on fewer PACs, but that it might be interesting to run the models with the Method 1 data to compare results. There is a question whether toxicity is disproportionally caused by alkylated PAC or parent PAC. Another panelist commented that for some endpoints, methods based on total sulfur or nitrogen-containing ring content appeared to be a reasonable predictor of the endpoints. It was suggested that comparing fit across different methods might provide insights into the actual toxic moieties. An author responded that after they selected the biological endpoints they used several different analytical techniques, but had to consider whether there were enough data available for the different methods, as well as whether they had good model fit. Method 1 did not always provide a good fit, the carbazoles and sulfur PACs did not always fit as well, and there was a smaller amount of data for Method 1 relative to Method 2. The authors made a subjective judgment on model fit across the range and did not rely only on summary statistics. Dr. Roy clarified that Method 2 makes no attempt to differentiate between non-heterocyclic and heterocyclic PACs as it just looks at the 1-7 ring distributions. Ring groups (1-7 ring) are defined with standard indices and published methods.

A panelist noted that traditional geochemical methods are more directly analogous to what EPA does, than the DMSO method. The panel member thought that Method 1 may resonate more with EPA as it is more consistent with typical PAC profile determinations used for many site investigations. The paper by Feuston et al. (1994) shows the two methods side-by-side and notes that the DMSO extraction method (Method 2) has an inherent loss of some constituents. Compared with Method 1, the DMSO method (Method 2) has a low bias for the 1 ring category, with bias increasing with the number of rings and size of compound, so that 4 and 5 ring categories are higher. If one compares by petroleum product, one sees a significant amount of variability between DMSO percent weights and what one would expect to see with a benchmark method (Method 1). For the 1, 2, and 3 rings, Method 2 will under represent some compounds and over represent some others. There is an issue with benchmarking and comparing results from Method 2 and more traditional methods. For accuracy and precision, the literature demonstrates considerable amount of variability; a panelist thought that it would be helpful to have a side-by-side comparison between Method 1 data and Method 2 data. This comparison would help benchmark results with what EPA looks at in their environmental toxicity studies.

A panelist asked whether Method 2 is more robust given it includes the 1-5 rings from Method 1 as well as rings 6 and 7. Another panelist clarified that Method 1 looks at a specific set of isomers and picks out discrete compounds, while Method 2 measures the percentage of the sample mass accounted for by each ring. Method 2 involves the use of flame ionization detection, which is the gold standard for hydrocarbons and petroleum. This panelist thought it was remarkable that the two methods had comparable results as they are very different. Another panelist noted that Method 1 has a smaller set of compounds than Method 2 and the absolute concentrations will not be the same between methods and so they cannot be compared directly and the underlying data sets cannot be combined.

Another panelist pointed out that what is really of interest are relative concentrations. Method 2 gives a relative scale, but may not be totally accurate, but as long as it is not way off, and if the

scale is reproducible and relatively consistent, then it is a place to start. The goal is not to try to nail down specific toxicity.

A panelist pointed out that Table 5 on page 18 shows there was little difference in standard error and correlation coefficients between Methods 1 and 2. Another panelist noted that while Methods 1 and 2 may be equal with regard to statistical fit, the slopes may differ and asked which method showed more potency. A panelist cautioned that the issue of slope is difficult to deal with because the two methods do not directly correlate and the difference in slope reflects differences in the analytical methods. Panelists recognized that there are clear analytical differences in the methods and that comparison across methods might be explored further to help understand toxic mechanisms. This might be useful to help predict streams that do not fit well with the model and as a tool for sensitivity analysis to compare the dependency of outcomes on the method that is used.

The panelists generally were satisfied with use of the Method 2 procedure but recommended that the authors better explain and defend their selection of Method 2. The document clearly focused on results of Method 2, without little on development of Method 1. The authors should provide a more complete description of the methods and the quality assurance and quality control measures used in the analyses. Panelists found Method 2 promising and the relationships with risk are compelling, but it would be helpful to know more about validation procedures and the performance of the overall method. One panelist remarked that the effort is very exciting, in that the authors are on track to find the bioavailable fraction of petroleum, which is something many have been trying to do for some time.

5.2.1.1 Public Comment

A web cast participant, Dr. Jeri W. Higginbotham of the Kentucky Environmental and Public Protection Cabinet, sent the following question:

> I have a question concerning the analytical methods. My understanding is that, if these models are eventually accepted, industry will generate PAC data for other petroleum categories (crude oil, for example) which can be used as input into the models to generate measures of toxicity. The analytical method, then, is an important consideration. Will someone be addressing the relative strengths and weaknesses of these analytical methods in more detail than what has been done to this point? How are they comparable in what they extract in each ring number structural group? If industry already has PAC information for a petroleum category that was not obtained using method 2, will that information be usable as input data? Dr. Nicolich seemed reluctant to validate the models (constructed with data obtained using Method 2 to quantify PACs) by using data obtained with Method 1 quantifying the PACs. What does this portend for its future application?

In response to this question, Dr. Tim Roy provided further explanation of Methods 1 and 2. He noted that Methods 1 and 2 are very different from each other, and also very different from EPA methods. Analyzing petroleum streams is very difficult and accuracy in measurement is hard.

Method 1 is very complex and involves initial separation of saturates and aromatics and further separation of the aromatic fraction by solid liquid chromatography, with the aromatic fractions being analyzed by mass spectroscopy. It would be hard to apply the rigors of EPA methods (e.g., SW-846) to either of these methods. Method 2 starts with a variation of IP 346 – a DMSO-cyclohexane partitioning of the petroleum product. Due to the complexity of these materials one cannot affix accuracy constraints to the procedure, although there is a precision standard with limits for repeatability and reproducibility available for Method 2. Although the method was (peer-reviewed) published and used in their laboratory for over 10 years, it was never evaluated for inter-laboratory reproducibility.

[Post meeting, Dr. Roy further explained the quality control differences between Methods 1 and 2 and standard EPA methods. He noted that many petroleum product compositional analysis procedures such as Method 1 are intended to provide information to facilitate either refining or post-distillation treatment or blending. The 'mentality' surrounding the design and implementation of these methods is quite different from that of, e.g., EPA solid waste procedures (SW-846) where precision and accuracy determinations are an integral part of the procedure…("This [(SW-846] manual presents the state-of-the-art in routine analytical tested adapted for the RCRA program. It contains procedures for field and laboratory quality control, sampling, determining hazardous constituents in wastes…"). Rigorous quality control is designed into the SW-846 methods since thousands of laboratories use the methods to generate data on hazardous materials that may be used in human and environmental exposure and risk assessments. On the other hand, Method 1 is a highly specialized procedure used by comparatively few laboratories as part of an overall refinery/blending plant quality control program. Although Method 1 does not contain definitive accuracy/precision procedures (i.e., as in SW-846), quality control measures are imbedded in the standard operating procedures (SOPs) for the analytical equipment used in the method. Method 2 as well contains imbedded quality control measures in the form of instrument SOPs.]

Dr. Roy noted that Method 2 tries to mimic a biological method – a modification of the Ames Salmonella Assay specific for petroleum products. In essence, the fraction of aromatics isolated by Method 2 is a subset of what one would get from Method 1. However, since the sample preparation procedures for the two methods are very different, he did not think that one could take Method 1 data and apply to a model built on Method 2 data. Dr. Roy went on to say that Method 2 is simpler, and provides more specific or selective information. Method 2 is a simpler procedure than Method 1 requiring less sample preparation, less sample analysis and less data analysis. By virtue of the model-building process described, Method 2 data have been shown to be the more highly correlated with the chosen biological endpoints. Assuming the observed toxicity is due to PAC, Method 2 is the more 'specific' or 'selective' procedure as regards the isolation/concentration of those PAC most responsible for the observed toxicity.

In response to panelist questions, Dr. Roy explained that Method 1 is an ASTM method and so has gone through some rigorous evaluation and validation. It is however, very expensive and not rugged (i.e., not easily adaptable to inter-laboratory use). Its purpose is to characterize the aromatic fraction. It is a cantankerous method that requires lots of brain power to work it and he was not sure if it is still being used. Method 2 is robust and simple and if one wanted to go on

round robin for validation, one could find enough laboratories to conduct the work. He did not see any systematic way to translate data from Method 1 to Method 2.

5.2.2 Charge Question 2

2. The authors made an assumption that PACs are the source of toxicity for petroleum substances that contain PACs. Did the authors have adequate sampling to test and support this? Was the decision to conduct the statistical analyses based on PAC analysis method 2 supported by the data? Could data sets from studies using other analytical methods contribute significantly to the analyses? Does the description of the chemical composition of petroleum substances support the conclusion that toxicity evaluation methods based on individual PAH content cannot be used?

One panelist started the discussion by voicing a concern that while it seems there are a lot of data to suggest PACs are the most toxic components of the petroleum substances, there is not much presentation in the report on other components or aspects of petroleum that could be part of the toxicity equation. The panelist also expressed concern that given the variability among various petroleum streams, it was not clear what specific petroleum substances were used in the analysis. The panelist suggested it would be useful to present the range of physical properties for the substances represented in the analysis and look for data that are outliers. For example, if one had lubricating oil and diesel range hydrocarbons, where would these fall and are the data used to build the model biased for different compound categories? One of the authors explained that when they initially developed the models, they used HPV categories, but they were not comfortable with that approach, as these are discrete categories, while the petroleum substances are on a continuum. Terms were added to the model to take care of this. They also looked at plots of the observed data and model predicted data, and the points were not grouped in any way that would suggest differences due to the HPV grouping. The panelist noted that this issue should be more fully explained in the report

A panel member expressed concern with sample limitations and representativeness of the substances used to build the model. The panelist looked at the test materials for the studies listed in Table 2A-5 and compared that list with the CAS numbers for the HPV categories as provided in Appendix A to Volume 1. The panelist questioned the representativeness of the materials used in this analysis, relative to the universe of substances in these categories and suggested that the report should be re-titled to identify the specific fractions covered, and it should be made clear that the CAS number representativeness of the analysis was not that broad. The panelist thought this a significant limitation on the utility of the model. The panelist did not think that the critical question of why the PACs are considered the most toxic constituent in the petroleum stream was adequately addressed. An author noted that between HPV categories, it did not appear that the model fit better or more poorly for different categories. The panelist clarified that the concern was not between categories, but within each category and how one answers the question of the range of compositions within all the CAS numbers. The author thought that identification of interpolation and extrapolation would address this concern.

Another panelist pointed out the difficulty in gaining a complete picture for a substance and category. Specifically, the panelist had difficulty in cross-referencing and linking information about particular substances through the chapters and appendices. For example, dose-response plots in Appendix 9 are labeled by substance number, but the reader needs to go to Appendix 2 to find the study, look in Appendix A to find out what substance it is, and consult Appendix 10 to determine the PAC composition. A table or other device to cross reference the information is needed.

A panelist asked why the wealth of existing mouse oral and other data cannot be used for HPV and what is needed to fill the HPV mandate? An author responded that for HPV they need to provide a toxicity value or indication for all CAS numbers. However, even with the same CAS number, the variation of composition of these petroleum materials can be great.
Another panelist noted that there are compositional issues from these other data, the studies often do not identify the extraction methods used and thus PAC content may not represent those that would be obtained using Method 2. The authors confirmed that the relevant route of exposure for petroleum substances is dermal and the model they designed is for dermal exposures only.

A panelist noted that the there is a putative assumption that total aromatics are strongly linked to toxicity. The panelist explained that this hypothesis was tested with regard to petroleum-based solvents many years ago in response to a proposed approach for setting occupational exposure limits based on boiling point and total aromatics content. After extensive investigation and many publications, the hypothesis was not demonstrated. This suggests that for the fractions tested in those experiments something other than PACs may be driving the toxicity. A second panelist commented that the result of the earlier work would be consistent with the current investigation since as shown in Figure 1 of the main report total PAC content was not adequately predictive of toxicity. In addition, since the earlier investigation used primarily lower boiling point substances the study results are not fully comparable. One would expect the lower molecular weight non-aromatic hydrocarbons to play a greater role for low boiling point streams as compared to the large chain length non-aromatics in the heavier boiling point fractions covered in the current report. Other panelists also cautioned on drawing too many parallels from the studies. One panelist noted that the assumption that long-chain non-aromatics in heavy streams have limited toxicity is not consistent with the work of Grasso et al. (1988) who showed effects for a stream lacking aromatics. Another panelist asked whether the methods from the earlier study were based on samples prepared using DMSO extraction, noting that methods should be considered in examining the implications of the earlier work to the current effort.

A panelist expressed concern about the limitations of the rat dermal studies, which has not been the traditional model for dermal investigations and may be unique to the two laboratories whose studies were used in this analysis. The panelist thought that the authors need to provide justification for their use of this type of protocol. The authors noted that some of these rat dermal studies have been published and while it is not the most common protocol, there have been a significant number of rat dermal developmental toxicity studies over the last decade. Another panelist agreed that rat dermal developmental studies are more common and accepted today.

The panel discussed the possibility of a using a relative potency approach for these mixtures, which has been done with PAHs using benzo[a]pyrene (BaP) as an index chemical. This approach involves selecting a toxic component and basing the toxicity prediction on the concentration and relative potency of other compounds in the mixture relative to the index chemical. The Method 2 analytical method could not support either the relative potency or an additive approach, because it does not identify individual components, rather it provides concentrations of the 1-7 rings hydrocarbons, for which there are no corresponding toxicity studies. Method 1 does provide individual component measurements, which could be used in a relative potency scheme. Nevertheless, a panelist concluded that there is good understanding based on carcinogenicity evaluations that an individual PAH surrogate will not fully explain the data. As described in Volume 1, petroleum streams contain thousands of individual isomers of PACs, rather than relatively simple mixtures of the standard PAHs for which relative potency factors have been estimated. Several panelists suggested the authors explain this further in the document, to explain to readers who are not familiar with those data and analyses.

Another panelist suggested looking at parallel analysis, compare the results of this analysis to individual PAH or individual methods. This may answer questions such as whether a single compound (e.g., benzene) alone could predict toxicity as well as this model. An author noted that qualitatively, the developmental effects seen with these petroleum substances are the same as those seen in benzene studies; however, benzene requires much higher doses to produce the same developmental effects seen with these petroleum mixtures. A panelist agreed that if one is trying to hypothesize as to what is causing the toxicity of these mixtures, then a single chemical approach could be used; however, given what is known about PAH in general, the authors' highly interactive term better describes what the studies show. However, several panelists reminded the others that the purpose of this method is not to look for causation; one cannot draw mechanistic conclusions of causative agent or agents with the analysis presented.

The panel also explored an additivity approach (adding the individual components' toxicity), and comparison to the models' results. One panelist noted that if there were an easy way to do additivity, it would have been already done, but that this lack of success was frustrating because it is intuitive that the PAHs are the most toxic components. A panelist noted that while it is attractive to identify a single marker PAC compound, in a European analysis of skin cancer data they found that BaP alone was not able to predict carcinogenicity, and that the total of PAHs also failed to predict cancer outcome. However, the weight of DMSO extract was predictive, while the individual PAC content alone was not enough. Another panelist noted that the data on developmental toxicity for petroleum substances is so much more limited than that for dermal carcinogenicity.

A panelist explained that it is known that 1-ring aromatics are more potent than 2-ring, which are more potent than 3-ring, etc., but the PACs interact in ways that are not understood, noting that some are carcinogenic promoters, while other inhibit carcinogenicity. The skin of the mouse appears to integrate these various underlying mechanisms and so a predictive method that integrates contributions of various components is preferred. However, an integrated approach is limited to demonstrating association and not causation since the complex mechanisms that drive the observations remain unclear. Carcinogenicity and mutagenicity are an integrated response

and the FDA prefers the DMSO-based analysis for these endpoints. Based on the carcinogenicity and mutagenicity work, the assumption of additivity would not work for these substances as not all are additive, a few act synergistically, and many inhibit toxicity of others; thus, the sum would be less than the total of the parts. In clarifying with the authors another panelist noted that this point is demonstrated in the figures on pages 22 and 23 of the report, which show very different dose-response for two substances with similar total PAC extract weights.

Although some panel members thought that the traditional relative potency approach would have limited application for predicting toxicity for the diversity of petroleum streams, one panelist further suggested that that the multiple regression approach used to build the models might answer the questions about relative toxicity. In viewing models with good fits the "coefficient times dose" could provide information on the overall contribution to the dependent variables (i.e., toxicity outcome). This panelist also noted that PAC rings 1-7 look highly correlated and another panelist noted that co-linearity is a problem in using this approach for interpreting the coefficients, with the high number of co-linear variables in the models. The panelist noted that the models include an interaction term.

A panelist referring to Table 5, page 18, found some information to be gleaned from the comparison between methods and noted that Method 5's results are good even though there is a small sample size. For some endpoints, methods based on total sulfur or nitrogen containing ring content appear to be a reasonable predictor of the endpoints. The panelist asked if there is a way to build a model that has a covariant for nitrogen content since for some data sets there is information available. An author replied that what they have is a correlational study and he is loathe to put any meaning on these coefficients. He suggested that the model is best considered as a mathematically well-described black box with no biological meaning. While the model works well, he is not sure how well correlated the results are with Methods 1 and 2.

Another panelist asked whether there is a sense of relative toxicity between rings and/or the S-PACs and N-PACs, or if there is some logical progression in toxicity that could be used in model building. The panelist reviewed dose-response plots from Appendix 9 to see what additional information might support the current black box model and the hypothesis that the PAC content can predict toxicity. The plots suggest there may be a trend or relationship between toxicity and the predominance of certain ring structures. The dose response curves go from flat to steep slope as the number of rings increases. To illustrate the point, the panelist pointed to dose-response plots in Appendix 9 and contrasted the slopes of a number of samples: on page 10 is a light crude with a lot of naphthalene and 1 ring PAH; on page 11 is a heavy oil predominated with 2-3 ring PAC; on page 8 is syntower bottoms, which has been cooked and has a lot more petrogenic material -- it is even heavier and contains 4, 5, and 6 ring PAH; and on page 15, visbreaker gas oil a mid-distillate with lots of 2 ring PAH. The panelist detected a logical construct through these data and noted that these results are consistent with what is known about petroleum toxicity, with 1 ring more toxic than 2-3 ring PAH.

Other panelist found these insights and observations very useful and supportive to the hypothesis. One suggested reorganizing Appendix 9 to identify each of the materials by type and

then group them from light to heavy and discuss the patterns seen. Another panelist asked what kind of compositional data besides PAC is available for each material and noting that the treatment of the materials in cracking and reforming can be severe and perhaps there is a correlation with treatment. The first panelist speculated that the correlation might not be with PAC, but with a new species being formed. The panelists suggested the authors explore further potential correlations with other compositional characteristics.

5.2.3 Charge Question 3

3. Discuss the criteria and procedures used for identification, inclusion, and exclusion of toxicological data sets for the modeling. Were the criteria and procedures fully described and are they defensible? Is it reasonable to assume that all the relevant data have been collected and accurately compiled and analyzed? Are there other data sets that should have been considered? Did the exclusions make statistical and biological sense and were the impacts of such exclusions adequately explored? Would the procedures used in data set selection generate any bias in the results?

Panelists covered some of this charge question in previous discussions. A panelist complimented the authors for a good job of gathering and culling available data to use reports of comparable quality and noted that the laboratories look equally competent. The data sets used are compatible. To reduce the number of endpoints to a reasonable number for modeling the authors looked for factors that were responsible for toxicity. While the process and selection of endpoints were reasonable, each time one reduces the number of endpoints, one potentially is discarding information that may be useful.

Panelists discussed the exclusion of dose groups and data; in particular the exclusion of reproductive dose groups with complete resorptions. The panelists agreed with the logic that to see effects one needs fetuses, and that with the goal to capture the low dose behavior excluding the extremes avoids skewing the model. Several panelists suggested the authors might test whether this excluded data makes a difference by modeling with and without the excluded data. An author noted that they eliminated relatively few data and did not think including those data would make the model fit better in the low dose region, which is the region of interest. Another panelist agreed with the authors, noting that at the high doses one is outside the linear range, the data will be linear before saturation of the systems and curvilinear at higher doses. Eliminating the high doses is the easiest and most common way to deal with this and is a good approach in this case, as the authors have plenty of data in the region of interest.

For transparency and understanding, several panelists suggested that the authors plot all data points to show the regression with all the data and then show what they eliminated or excluded and explain why they excluded those data points. This would be similar to EPA guidance on benchmark dose (BMD) modeling which allows dropping the highest dose, but requires transparency in demonstrating the impact. The panelists thought this analysis could be useful and potentially supportive, providing greater confidence in results. Another panelist cautioned the authors not to conduct more statistical analyses without consideration of biological significance and focus on adverse effects, not just statistically significant findings. Yet another

panelist however, was not enthusiastic about a sensitivity analysis, noting that it is well known that for linear regression low and high doses have the most influence on the slope. If data are curvilinear, and one uses high dose data off the linear line, this will change the slope and the correlation coefficient. The panelist was not sure what is learned from the exercise and did not think it would improve the models or change the results. An author agreed that they could conduct a sensitivity analysis and that it might give further confidence to their work.

The panel discussed the issue of reproductive toxicity and whether the authors should or could have included endpoints related to reproductive competence. A panelist noted that the developmental toxicity studies cannot address male fertility but can provide some limited information regarding loss of fetuses and reduced litter sizes, which might provide indications of impaired fertility. However, the developmental toxicity studies only included exposure during gestation, and would not assess impacts on mating or fertilization (i.e., reproductive toxicity). A panelist also noted that the document refers to the availability of several developmental studies that included exposure during a pre-mating period. These studies were not further described in the rationale for excluding the reproductive toxicity endpoint. The panelist also cautioned that signs of maternal intoxication are likely at the dose levels where severe developmental effects (e.g., embryonic death, congenital malformations) are seen, and the authors should keep alert to the possibility of maternal toxicity.

An author explained the rationale for the conclusion that the data were not sufficient to allow for building a correlation model for reproductive toxicity endpoints. He stated that only two reproductive studies (one male and one female) were submitted and both focused on carbon black oil. The developmental studies were not of adequate design for addressing effects on male or female reproduction. The authors looked at histology and reproductive organ weights in the repeat dose studies, and concluded that reproductive organs are not a sensitive target organ system. When the authors compared the effects on reproductive organs to the developmental toxicity studies, they observed developmental toxicity at much lower doses than changes in reproductive organ weights or histology. One panelist suggested the authors evaluate the carbon black studies and provide additional discussion of how those data might be used, perhaps looking at PAC content and systemic uptake, and making comparisons, to predict reproductive toxicity because regulatory agencies may ask why the data cannot be used in this way.

5.2.4 Charge Question 4

4. Were the procedures for selecting biological endpoints for modeling adequately communicated? Were the methods adequate to identify all key relevant endpoints? Was the final selection of endpoints fully justified by the data including the analyses conducted? Should other endpoints be considered based on the current state of knowledge of PAH and petroleum stream toxicity?

A panelist noted that the procedures for selecting the biological endpoints for modeling were adequately communicated and there was adequate discussion of the relevant endpoints, particularly reproductive and developmental. The authors have noted that they inadvertently omitted renal toxicity (a well-known light hydrocarbon effect) from the tables listing the

endpoints for which data were captured. However, the authors explained that kidney effects were considered and not selected for modeling as they were not statistically significant. The panelist noted that this project is an empirical exercise and the authors have done a superb job with empirical associations. But, because it is empirical, and not mechanistic, one cannot draw any mechanistic conclusions, which is what everyone would like to do.

Another panelist did not think the reasons for excluding endpoints were explained as fully as they might have been. Looking at Appendix 3, one sees that liver weight changes frequently and it makes sense that this endpoint was included. But for a few others, one wonders why they were not included; for example, blood urea nitrogen or white blood cell count were affected in similar proportions of the studies as other endpoints that were included. The panelist suggested that the authors should start with their biological rationale for endpoint inclusion or exclusion, and then the statistical reasoning should follow, guiding the reader through the tables in Appendix 3. Another panelist suggested statistical significance should be considered regardless. In the case of developmental toxicity, it is possible effects may be missed because of the relatively small sample sizes due to few litters per dose group. For example, a change in liver weight may be statistically significant, perhaps this is not biologically significant, but it indicates a physiological change is happening, as in the case of microsomal enzyme induction. The panelist thought that all associations should be considered, and by using the most sensitive adverse endpoint, there will be protection for other less sensitive endpoints.

One panelist stressed that the work is empirical and not biologically-based, and therefore, removing endpoints for biological reasons made the panelist uncomfortable. In addition, it appears that there may be uneven application of the criteria for eliminating endpoints. The panelists and authors recognize the screening purpose of the proposed method, but panelists cautioned that strong caveats are needed because it would be naive to think others will not apply this work inappropriately to other situations, (such as Proposition 65 in California) or use it as a basis for human health risk assessment. Panelists noted that the information provided was not sufficient to do a risk assessment on petroleum substances.

The panel discussed dermal irritation and its implications for toxicity. One panelist noted that Feuston et al. (1994) observed that in most cases severe skin irritation precluded administration of higher doses. An author commented on the animal care aspects of the studies conducted, noting that the animals in the studies they used were cared for according to guidelines. In several cases high doses were discontinued because of severe skin irritation. A panelist agreed that if the laboratory saw obvious skin damage, animal care considerations dictate discontinuation of the study. However, not all destruction of tissue is visible, and one would need histological sections of an area to detect if tight junctions were intact, because if they were not, there would be greater absorption.

A panelist noted that irritation was obviously important and suggested the authors further explore dermal toxicity and determine if the direct effects on skin, such as irritation are an important endpoint. An author reiterated that their goal was not to understand the underlying mechanism, but to see if they could demonstrate a statistical relationship between PAC and endpoints seen in the studies. Therefore, they did not intentionally try to factor in rates of dermal intake, but the

fact that they saw statistical correlations between dermal exposure and effects clearly indicates that the test materials are being absorbed. The panelist rejoined that this is an assumption; data showing that these materials are actually absorbed are not available in the materials submitted for review. One of the authors then commented on bioavailability of PAC compounds in the dermal studies. There are studies in which individual PACs have been measured for skin penetration *in vitro*. The authors are aware of several studies, all performed at the laboratory that did many of the developmental toxicity studies used in this analysis, that included radio labeled marker compounds (^{14}C-carbazole and ^{3}H-benzo[a]pyrene) in the test material. These marker compounds were subsequently measured in maternal (e.g., blood) and embryonic (e.g., amniotic fluid, and yolk sacs) tissues. The presence of these marker compounds indicates that PACs are bioavailable systemically after dermal application. Panelists suggested this information be included in the report.

The authors explained that skin irritation was seen frequently in the studies, but the details reported were inconsistent, with many studies simply indicating the presence or absence of irritation and some others providing some qualitative descriptors. They suggested that they could add dermal irritation to the table of endpoints Table 3A-1) and indicate whether dermal irritation/toxicity was seen so the information is captured. However, quantifying dermal irritation could not be done with the data available and, thus, correlations could not be built.

Panelists thought adding the qualitative information would be helpful, but one panelist noted that dermal irritation does not cause systemic effects; rather it reflects the skin breaking down, increased penetration, and increased systemic uptake. Panelists suggested dermal irritation be further explained and discussed in the text because if there is enough dermal irritation, the absorption might change, resulting in different or greater potency. Other corollary literature (e.g., dermal carcinogenicity studies on petroleum substances) should be consulted to further explore this issue. Panelists also thought that dermal toxicity also needs to be sorted out for its influence on maternal health status (body weight gain, food consumption) and reproduction and fetal effects. Understanding the effect of dermal irritation/toxicity on the mother is important for the continuum of effects to the offspring.

The panelists discussed maternal toxicity and the need for the authors to be alert to the possibility of maternal toxicity when exploring developmental effect. One panelist cautioned that maternal effects are likely at the dose levels where severe developmental effects (e.g., malformations) are seen. Panel members suggested that the authors address maternal toxicity as a potential endpoint and conduct analyses to determine if developmental toxicity was secondary to maternal toxicity. One panelist cautioned that maternal toxicity does not cause birth defects; rather it is a condition of the mother. Some fetal toxicity endpoints may be associated with maternal toxicity. Some panelists suggested that the authors carefully analyze maternal endpoints. A panelist also pointed out that an effect seen in the offspring without maternal toxicity would potentially be a much bigger issue for risk assessments, than effects noted in both and therefore, identification of such cases is important.

One panelist noted that the appropriate dosage unit for reproductive and developmental studies is the litter and if the mothers were sick, the first thing affected will be fetal weight and survival,

often as a result of overt maternal toxicity. Therefore, it was suggested that the authors look at maternal body weight loss and reduced fetal weight. If the correlation between the two is high and the dams showed obvious signs of systemic intoxication one could then look closely at the association between the mothers' condition and systemic effects. A panelist also noted that the authors need to look at other clinical signs in addition to maternal weight. The authors might look for available PAH kinetic information (e.g., ATSDR Toxicological Profiles) as these might help shed some insight into the developmental effects observed, and alleviate concerns that dismissing developmental effects in the presence of maternal effects might miss developmental effects that are occurring concurrently.

Panelists raised concerns about the usefulness of the maternal toxicity data because irritation may have affected animal behavior. One panelist questioned whether adequate maternal body weight data were available and was concerned that feed or water data may not be good indicators of maternal health. If the animals are not well and/or the skin is severely damaged, animals waste feed or water. In studies where a collar is used, that may impede access to the feed cup or annoy the animal and cause it to scatter its feed. The panelist thought that if there was a suggestion of any of these things, then one must attribute much of what is seen to maternal effects.

An author noted that they had not looked at maternal effects specifically; however, in listening to the panel discussion, he sees that it makes sense to look at some maternal endpoints and noted that maternal toxicity was seen in most of the developmental studies, including decreased maternal weight gain and decreased food consumption.

5.2.5 Charge Question 5

5. Are there other important issues to discuss regarding the selection and evaluation of data sets?

No further comments that were not otherwise captured above.

6. Identifying and Characterizing Relationships between PAC Content and Toxicity to Determine Use for Prediction of Untested Petroleum Substances

6.1 Author Presentation

Dr. Mark Nicolich of ExxonMobil Biomedical Sciences, Inc. described the relationship between PAC content and biological endpoints and noted that the goal of the model building process was to predict the relationship between PAC content and biological response. The authors sought to create a simple model. They did not base the model solely on statistical significance, but no attempt was made to identify causal relationships, as that requires mechanistic understanding. Dr. Nicolich pointed out that the models show high correlations between PAC profile and toxicity outcomes in the low dose area of interest.

The authors started with a simple linear model and attempted numerous transformations of independent and dependent variables to increase statistical fit of the model. A factor analysis was conducted in an attempt to optimize the slate of variables in each model, but this approach was discarded as overly complicated with no gain in predictive ability. Ultimately the selected variables were based on comparative goodness of fit and general considerations related to PAC content and biological variables that were expected to be important predictors of the response. For example, the authors also tried variations of the control group response, such as ratio to response or as a covariate, or not at all. They considered non-linear models and mixed models, but neither approach significantly improved the fits. The final models used all seven ring concentrations without regard to statistical significance and they based the model adequacy on the relation of observed versus predicted (correlation) and accuracy (residual standard error). Dr. Nicolich explained that the models were built conceptually as a black box, where mechanism of toxicity was not considered; rather, the authors considered overall reasonableness, outliers, influence points, and model validation in selecting the final model structures.

Dr. Nicolich explained the models for predicting predefined levels of response (PDx). For application of the models the authors suggested a "critical level of change" for each of the modeled endpoints, but only for demonstration purposes. The selected response rates were developed based on informed professional judgment and the rationale was described in an appendix to the report. However, the authors recognized that other values could be selected based on the needs of the user. The authors noted that the models are useful over the observed range of data and that any degree of change could be used. The authors explained that they also considered the BMD and LOAEL as alternative approaches for describing the critical dose, but preferred the PDx for a number of reasons. Information in Appendix 7 shows the relationship between PDx and the LOAEL and the models with dose-response fit well. The models were considered by the authors as "good descriptors and good predictors" within the data range and are useful when applied to a concept like the PDx or another method to summarize the biological response after exposure to a PAC-containing substance.

6.1.1 Clarifying Questions from the Panel

In response to a question regarding use of individual data points, rather than litter averages, Dr. Nicolich responded that the study reports presented mean responses and individual data were not readily available. The panelist pointed out that when one wants to predict for an untested substance, it will be an independent future observation. The author noted that using individual data would result in the same regression line, but with greater variance. Predictions based on the best estimate should be about the same. He noted that, in a sense, one wants to predict mean response.

Another panelist asked the author to explain how they considered outliers and influence points to pick model parameters. Dr. Nicolich explained that he visually examined plots (e.g., Figure 4), and looked for outliers in the center of the data and also looked at low doses. He tried to enhance the terms in the model that could account for the outliers by using various techniques (e.g., cube root of the dose). They did not do a formal sensitivity analysis, but tried many

different ways to break the model, such as adding extraneous terms, and dropping out terms. They found it was hard to perturb the model, giving them confidence that the models were stable.

A panelist asked whether they tested the model against a kerosene or another petroleum substance that had no PAH content to see what would be predicted. Dr. Nicolich indicated that he did not do that formally, he only used the data that were in the provided studies, but some of those studies were low in PAH content. A panelist asked whether these models would work for other organic matrices and Dr. Nicolich said he did not know.

In response to a question regarding Figure 4 and the hemoglobin concentration plot, the author discussed the outlier point in the top left half of the plot that was far above the prediction line. Hemoglobin concentration of controls was 17-18 gm/dl and so a response between 12 and 18 was within the area of interest. The outlier points, as well as a second, much lower point, appear to be in a range where things "may not be right." The author was not sure if this was a recording error in the original report, or something else. He could not explain it, but noted that it is one point out of 80 or so data points. He could have tried to model it, but noted there is the low value of 7 or 8 that does fit the model well. Another panelist asked whether the authors considered which side of the predicted line the outliers fell. Dr. Nicolich answered that they did not consider over or under prediction, as their goal was to maximize the correlation. They looked at the pattern of residuals and saw no real pattern, but did not try to account for this in the modeling procedure.

In response to a question regarding how the authors developed the predicted response, the author referred the panel to the equation on the top of page 25. A panelist asked about the PAC 4 x PAC 5 term and the author explained that the 1-7 PACs were considered variables. This term represents an interaction between PAC 4 and PAC 5.

Another panelist asked how they came up with the parameters. Dr. Nicolich indicated that he used a mixture of professional judgment as well as trial and error. He started with simple terms. The PAC 4/5 interaction term came as a substitute for HPV groupings to describe that. A panelist asked whether the same set of parameters were used for each model. Dr. Nicolich explained that each equation is slightly different, but the second and third lines of the equations listed in A6.6 are generally the same, the control group with β_1 is always there, but β_2 uses the other independent biological variables found in Table 8 on page 26 of the report.

A panelist sought clarification on whether the number of implants for a particular dose group or for the control group was used and was concerned that dissimilar items were being compared. Dr. Nicolich responded that they used number of implants for a particular dose group, but found that the number of implants was not useful. He explained that they had data for many variables, including implants, control implants, corpora lutea, etc., and they included some of these (as described in the report) in the model. In developing the models, he put the variables in for those that were correlated or related, and he looked for combinations to help reduce the number of variables, as too many variables could make the models unstable. He then took out data at random to see if it changed the model substantially and looked at summary statistics to see what made the biggest differences.

A panelist noted the dose is in mg/kg bw-day and asked whether surface area was considered. Dr. Nicolich responded that the data were reported in the studies as mg/kg bw-day and the dose as surface area was not modeled.

Another panelist noted that standard techniques vary by laboratory, and some of these studies used occlusion and others did not. This is a large variable and the techniques impact the results. This information and consistency or lack of consistency should be noted.

A panelist asked whether the authors evaluated status of the dams (e.g., body weight gain and maternal food consumption). An author indicated they had decided to focus on the developmental endpoints and did not try to correlate the maternal body weight gain with the developmental endpoints. However, in many cases the substances appeared not to be selective developmental toxicants. The panelist agreed that would be the case, based on the nature of these materials, but suggested presenting this evaluation would be helpful.

A panelist sought clarification regarding Slide 2 of the Part B presentation. The stated goal on this slide is to "accurately describe and predict the relationship between PAC content and biological endpoints." This language implies the authors considered it a given that there was a relationship. Dr. Nicolich acknowledged that the slide was not worded well, and clarified the authors' 2-part charge: first, to determine if there is a relationship and second, if there is a relationship, what is the statistical nature of that relationship and whether it can be used for predictive purposes.

Panelists referring to statements on Slide 8 noted that the PDx is essentially a BMD, as one has to identify a response level of interest in calculating the corresponding dose. They asked why the authors did not use BMD or benchmark response (BMR) to minimize confusions and noted that PDx sounds like administered dose. An author explained that they used a new term because the BMD is often a choice among multiple models and the PDx is tied to a specific equation and model.

6.2 Panel Discussion on Identifying and Characterizing Relationships

6.2.1 Charge Question 6

6. Discuss whether the statistical methods were appropriate and adequate and if the procedures were implemented correctly. Would other valid statistical approaches yield different results? Would alternative model development approaches have improved the models and results? Were the authors' conclusions regarding the models clearly articulated and justified by the results?

A panelist commented that the very good correlations may be a statistical artifact due to the large number of parameters (close to 20) and the relatively small number of data. More parameters, compared to data points, provide better predictive ability. An author noted that this is generally correct and that over parameterization is a concern in building models. The panelist suggested

the authors might look into how many variables they have and more fully document efforts to eliminate extraneous parameters. In particular, with many parameters multicolinearity becomes very important. A panel member noted that the document states that mixed models were also attempted to identify which variables are most important, but that the authors had reverted back to modeling with an ordinary least squares (OLS) approach as a simpler means to identify models that had similar overall fits.

A panelist also raised several questions about the general model structure and suggested enhancements to the document to provide the rationale for decisions made in modeling, noting as an example that for count data there is a need to justify the use of normal distributions. The inclusion of responses such as "control live fetuses" or number of implants on the right side of the equation was also noted as a potential concern, since such parameters have their own inherent biological variability that will not be reflected in the prediction if they are treated as true independent variables. The balance of this consideration versus the need to provide some normalization of the response to control responses should be made clearer. In addition, the use of mean response data versus individual data and the use of the OLS methods further reduce the meaningful interpretation of the confidence limits presented for the dose response curves. Use of mixed models would address this concern to some degree.

In response to a number of questions and comments, Dr. Nicolich clarified that the control group is treated as an independent variable and in the models they only plot response of the dose group and do not predict the control group. The authors tried looking at deviations and looked at many different forms to account for control response. They wanted stability and good descriptions and found that this model form worked best.

A panel member also commented on the range of usability for the models, indicating that the model building should include a wide spectrum of data for application to real dose-response applications. An author commented that the data points in the studies go fairly low. For example, the range goes from average or typical to very low response as indicated in Figure 4.

Several panelists recommended better documentation of the modeling procedures. One suggested that a non-mathematical explanation and example of multiple regression would be helpful to include as this is a concept difficult for most people to visualize. The panelist suggested a simple example, perhaps explaining the PAC 4 and 5 interaction by saying that PAC 4 and PAC 5 are special because they do not operate independently of each other. Therefore, if 4 and 5 are both high, then the term will be large because they have a big effect. A second panelist pointed out that how alternative modeling approaches improve or worsen model fit should be fully described. In particular, the authors should discuss the multiple parameter fit and the co-linearity issues and problems. The authors should describe the various exercises they went through and what they tried to provide a more complete understanding of the strength of the results. Panelists suggested that the authors consider making available the documentation of the exercises to interested parties. This type of documentation would show the sensitivity of the model to various choices and would provide more support for what was ultimately selected.

A panelist suggested that significance levels of the correlations in the multiple regressions should be presented and explored, noting that while the prediction of the models is good, one cannot tease out which PAC component is operative or important without understanding the significance of the regression coefficients. An author clarified that they did not think that significance was important, because this is simply a correlational study and they want to guard against people making mechanistic statements such as the 3-ring PAC is significant and it is the driver; they want to avoid those types of conclusions. Another panelist suggested strong cautionary language be added to guard against readers doing this, as it will be inevitable readers will try to identify which parameters contribute the most.

6.2.2 Charge Questions 7 and 8

7. The authors identified the toxicologically meaningful degree of change for each general toxicological endpoint (body weight and liver weight, hematology changes, thymus weight change). For each of the endpoints discuss whether these values are toxicologically and physiologically meaningful and if the best value was chosen. Are the observations consistent with what is known about petroleum toxicity? Are there other important issues regarding the relationship between PAC content and general systemic toxicity?

8. The authors identified the toxicologically meaningful degree of change for each developmental toxicity endpoints. For each of the endpoints discuss whether these values are toxicologically and physiologically meaningful and if the best value was chosen. Are the observations consistent with what is known about petroleum toxicity?

The panelists discussed the definition of "toxicologically meaningful degree of change (PDx)" for repeat dose and developmental toxicity endpoints, and the basis for the percents selected. The panel questioned the support for selection of one percentage over another. Panelists had concerns over the biological and statistical basis for the authors' selections and suggested that the authors need not specify specific toxicologically-meaningful values in this report.

A panelist pointed out that the authors provided support for their recommended PDx values (for the repeat dose studies) in Table A7-1 on page 2 of Appendix 7. The Table provides information on the power of the study to detect change, whether the degree the change is clinically relevant, any precedent by federal agencies (e.g., EPA, ATSDR), and any corollary to how one would define a BMR. The panelist liked the idea of noting the BMR corollary to the PDx and noted that in U.S. EPA risk assessments for some endpoints a specified percent change in response is used. When such a percentage change cannot be established for a response measured on a continuous scale, then a change in the response equal to one standard deviation based on the variability in the control response is often used.

One panelist noted that for continuous (non-quantal) data, the percent change that is considered toxicologically meaningful needs to be considered relative to the standard deviation among animals. For example, if the coefficient of variation (standard deviation expressed as a percent of the mean) is 10%, a change in the mean of 10% due to exposure to PACs will result in shifting

approximately an additional 10% of animals to abnormal levels. On the other hand, if the coefficient of variation were 30%, a shift of the mean by 10% would result in an undetectable increase in the proportion of animals with abnormal levels. In the absence of a universally accepted change in an endpoint that is considered biologically meaningful (significant), Gaylor and Slikker (1990) and Crump (1995) proposed a procedure for estimating the proportion of animals with abnormal values as a function of dose. Then the dose, PDx, can be estimated with a risk of an excess of x% of animals with abnormal levels. Abnormal (not necessarily adverse) levels for an endpoint are defined by low or high percentiles (e.g., 1^{st} or 99^{th} percentile) of levels determined for unexposed control animals. For example, for a normally distributed endpoint, a shift in the mean equal to 1.1 times the standard deviation will result in an estimated excess risk of 10% of the animals exposed at this dose to reach abnormal levels. Confidence limits and uncertainty factors can then be applied to this dose to arrive at a reference dose.

The panelist noted that other issues could be considered in this analysis. For example, for thymus weight, the authors might consult the work by Luster et al. (1992; 1993) on immunotoxicity points and how they relate to each other. In addition, the thymus weight discussion should include white blood cell count for immunotoxicity and the kidney effects should be discussed. With regard to the PDx values for developmental endpoints, another panelist pointed out that as few as one less pup per litter can be biologically significant and this would be before a 15% change was seen. Similarly, for fetal body weight a drop of as little as 0.2 g/pup (probably less than 10% of weight) can indicate a problem, depending on the procedures being followed by the laboratories. Furthermore, in quantifying the PDx values for developmental effects, a panelist clarified that one needs individual litter data to calculate the abnormal levels for an endpoint and it cannot be done with the means. For litters, one would need the standard deviation among litters to base analysis on the more appropriate control standard deviation.

Several panelists concluded that the choice of what degree of change that is biologically significant is beyond the scope of the authors' charge and gets into risk assessment applications beyond the intent of the HPV program. A panelist pointed out that the actual percentage is not needed for predictive purposes. However, one panelist while agreeing with this sentiment noted the HPV program does require denotation of a potency estimate of some kind – and thus predictive models would need to generate such an estimate. Panelists suggested that the authors not identify or try to support any particular toxicologically meaningful degree of change for specific endpoints. They suggested that they should present several different percentages for a few endpoints (with no judgments on toxicity), or use hypothetical values to serve as examples to demonstrate that how the model can be used to predict toxicity. The user should decide what value is most appropriate for the situation of interest. A panelist suggested calling these hypothetical cases. Another panelist suggested picking hydrocarbons in various categories to illustrate. Several panelists thought that this concept should be presented separately from the models themselves. Suggestions included presenting this in an appendix, a separate document, or publishing the analysis separately in the peer-reviewed literature. Another panelist suggested that if specific values were needed for the HPV program, then further development of the methodology to incorporate the use of the control standard deviation approach in defining the PDx should be explored. Panelists concluded that the selection of PDx values, use of the PDx

rather than BMR, and interpretation of results will likely be controversial, and these issues should be presented separately so as to not detract from the models themselves.

Statistical approaches for accounting for maternal toxicity were also discussed. One panelist suggested using maternal weight as a covariable in the model, but was not sure how other maternal effects could be added. Another panel member felt uncomfortable about adding too much to the model if it gets so specific that problems are missed and false negative problems are created. It may be more useful to think of pups and the dams in a bivariate way because they are correlated. A panelist suggested that the statistical technique of mediation analysis might help with this analysis to determine whether effects on the pup are mediated through effects on the mother. However, panelists did not think that adding another parameter will necessarily help if one considers the model a "black box." There is a temptation to add more variables to the model, but the goal should be to construct the model in way that describes the system well and as simply as possible. An author said that he needs to look more carefully how maternal toxicity can be addressed in the model. Looking at change in mother's weight as an independent variable may not be good enough; they may need to build two equations and link them by dose group.

6.2.3 Charge Question 9

9. The authors conclude that the pre-defined change (PDx) for developmental toxicity will be a reasonably good predictor of the PDx for reproductive toxicity. Is this conclusion valid?

A panelist summarized what was presented in the documents regarding reproductive toxicity. Two reproductive studies (one male and one female reproductive toxicity screen) were available in this data set and both were with carbon black oil. No adverse effects on reproduction or fertility were identified in either screen and the No Observered Effect Level (NOEL) for fertility and reproductive effects was 250 mg/kg-day, the highest dose tested in both males and females. The male reproductive endpoints evaluated included mating and fertility parameters, as well as sperm count, motility and morphology. The female reproductive endpoints evaluated included mating and fertility parameters, as well as estrous cycling. The same laboratory also conducted a developmental study of carbon black oil using a range of doses that were similar, and the results of these studies were used as the basis of the assertion that developmental endpoints are more sensitive than reproductive endpoints. The panelist had concerns regarding this conclusion, noting that reproductive toxicity is very complex, with many pathways, involving numerous organs and target cells, as well as biochemical pathways and hormonal effects that could modulate reproduction. The panelist was concerned about making decisions based on just two studies with a single material, given the wide variety of modes of action for reproductive toxicity and the wide range of petroleum products considered. Another panel member agreed that the data set was inadequate; even with data from the repeat dose studies there is not enough to make or support rigorous conclusions. While the repeat dose studies looked at organ weights and some histopathology, none of those studies looked at reproduction function. Furthermore, the comment in the document that it is only a reduction in testes weight that would be a concern does not account for all mechanisms – for example increases in testes weight due to edema.

One author agreed that the data set was limited with just the one compound, but noted this compound is high in PAC content and that there is a dramatic difference between reproductive and developmental toxicity seen with this material. The no effect level for developmental toxicity was 10 mg/kg-day, much lower than the reproductive effect level. The author did not know how representative this one compound is for all petroleum substances, but for carbon black, reproductive toxicity is not a sensitive endpoint. He also noted that in the HPV program reproductive toxicity studies are not always required.

A panelist cautioned that if the model was not built with the right kind of data it cannot predict responses. Another agreed that the available data were not adequate to support the hypothesis that developmental toxicity is more sensitive than reproductive toxicity. Because of this, another panelist suggested that the statement at the bottom of page 6 of Appendix 8 be deleted ("Therefore, the PDx for developmental toxicity is likely to be a reasonably good predictor of the PDx for reproductive toxicity, as well.")

The panel discussed the minimum data set that would be needed to predict potential reproductive toxicity. One panelist suggested that 1-generation reproduction studies could be done on a moderately heavy compound and a moderately light one. If both are negative, then that provides some confidence that reproductive toxicity endpoints are not as sensitive as developmental toxicity endpoints. If they are not negative, then more investigation may be needed. In addition, dermal irritation is a factor that must be considered. If the female skin is irritated or damaged, the animal may avoid contact and mating. As the light-end compounds generally do not irritate the skin as much as heavy end compounds, the studies will not be entirely comparable due to this complication.

In further consideration of appropriate study design, a panelist suggested that infusion or subcutaneous delivery may be needed. The blood level of the compound is what is important for developmental toxicity as the maternal blood concentrations determine the delivered dose to the conceptus. For dermal exposure there is no first pass through the maternal liver and the absorbed and metabolized dose after topical application can be different from that seen after ingestion. Because some of these materials irritate the skin and may cause discomfort or pain for the animals, alternative ways to get the appropriate blood levels should be explored. If one could establish the blood levels of the active compound(s) at dermal exposure levels that showed adverse effects, then a study could be designed to leave the skin intact to control for the local irritation effects. A good laboratory might be able to place a series of mini pumps subcutaneously at low enough levels not to destroy tissues. The blood levels would need to be sustained over the length of the ovulatory cycle (30-60 days) for females, and for the males the spermatagenic cycle (68 days). This would also have to be carried through the mating period and through postnatal day 6 or 8. Skin painting would not work due to increased absorption from damaged skin and/or conditions could affect mating behavior of animals. An oral study would only work if there was no first pass metabolism and the gut lining is not impacted.

7. Validation of Methods and Use of the Results

7.1 Author Presentation

Dr. Mark Nicolich presented the Task Group's efforts on validation of the methods and use of the results. He pointed out that the plots of the observed versus predicted data points clearly demonstrate that the models describe the data very well. The model validation was carried out in three phases. First, the authors used data splitting techniques. They had used all the data to determine the form of the model, but just 70% of the data to determine the coefficient values for the final parameters in the selected model form. To test the model predictions, the hold out data (30% of the data points) were used, with random selection of the hold out subsets repeated 100 times. The results showed that the models predicted well for interpolated points (those within the bounds of the original data range), and more poorly for extrapolated points outside the data set. Second, to address whether the correlations were an artifact of the model building, they randomized paired independent and dependent variable sets to determine if random pairings would do as well as the real pairing. They did not. Third, they looked at the application of similar toxicological endpoints between the models where such data were available, for example the results from relevant endpoints in the repeat-dose studies were used as a test data set against the results of the developmental toxicity model. For these they found again that interpolated points were predictive, but the extrapolated were not. The authors found the model did very well on all three validation steps, and worked well for interpolated substances, but not always as well for extrapolated.

Dr. Nicolich also explained that as a new extrapolated data point is added to the model, the basic form of the model stays the same, but the coefficients will be re-estimated and thus the location of the regression line might shift. However, the predicted values in the region of the original points are not likely to change much. He characterized the model's strengths – they are accurate descriptors of the data, they accurately predict new data points for interpolated substances, the model is simple, there is similarity of model form across endpoints, and the models are based on a relatively large data base.

7.1.1 Clarifying Questions from the Panel

Panel members sought clarification on how interpolation and extrapolation were defined for new data points. They asked whether the new data point has to be totally encapsulated on a per sample basis or whether the new data point only has to fit in the bounds of the maximum ring contents for the data set as a whole. Dr. Nicolich confirmed that the procedure for determining whether a new data point is considered an extrapolation is to test the PAC content (for each of the 7 rings) for that new data point sequentially against the profile for all the other data points used to build the model. To be interpolated, the new data point would have to fit within all of the 7 - ring percent weights for at least one of the samples in the database and outside for at least one sample.

One panelist pointed out that slides 7 and 8 do not fully represent what would happen for different kinds of original data. The new points shown on these slides are within the scatter of other data and so will not shift the predictive line by much. If the new point were more distorted,

say a lower dose and higher response, the whole line would have changed. The author acknowledged that these graphs represent an idealized situation.

A panelist asked about similarity of models across endpoints and inclusion of non-significant terms. Dr. Nicolich referred to Table 8 on page 26 and explained that the general form of the models is the same, but the independent variables vary by endpoint. He said that non-significant terms that involved the PACs were retained because, while the terms were not significant for the current data sets, future use of the models might identify cases where these terms are important and including the standard terms for all PAC rings ensures the appropriate application of the models for such new samples.

Other panelists noted that linear regression is commonly driven by high and low points and decisions regarding inclusion and exclusion of data are important, especially for extreme data points. High data points can influence the fit of a curve at the low end. Exclusion of some high data points may improve the fit of the model in the more relevant lower dose range. Several panelists suggested demonstrating what happens with and without the extreme data points and presenting a rationale for why certain data points were excluded, as excluding such data can also yield changes in the predicted PDx values (since the regression slopes will differ).

7.2 Panel Discussion on Validation of Methods and Use of the Results

7.2.1 Charge Question 10

10. The authors present predicted dose-response curves and compared these to actual results of the study from which the information had been derived. How accurately do the predicted dose-response curves fit the observed data, and how do the predicted PDx effect levels compare with the endpoint LOAELs/LOELs observed in the actual studies?

The panel discussed the PDx and LOEL comparisons found in Appendix 9. One panelist did not think the comparison appropriate because the LOEL is dependent on sample size and the selection of dose spacing within a study. The panelist noted that the BMD is the preferred method because it is based on a defined effect level (control standard deviation) and not sample size. For developmental toxicity studies this is less of a concern because most studies use the same number of animals.

Panelists discussed the utility and feasibility of comparing the PDx to the BMD, noting that EPA BMD software only has simple regression models and not the multiple regression needed here. Individual study results could be compared to BMDs and some panelists thought this would be a comparison that people would make in seeking to understand how the PDx compares with what they usually would use. Moreover, they emphasized that the comparison of LOELs with PDx's could lead a reader to conclude the model underestimates toxicity, and encouraged comparisons with the BMD instead (for individual studies). Since the PDx is based on the selection of a value for the percent change – by adjusting that value up or down, one can "match" the LOEL. A better comparison would be to compare the PDx with the BMD using the same percentage change. An author pointed out that if they used the BMD, they would be comparing two models,

and they do not know if either is biologically real. By comparing PDx to LOEL, they are comparing modeled results to observed data.

The authors' purpose was to show that the PDx was not extremely different from the LOEL. An author asked if plots of individual study data points with the model shown as a dark line would be useful. They could add the BMD model results to see how their model compares with a BMD for the same set of data. Panelists however cautioned that as this model used multiple studies, one would have to calculate BMDs for all the studies and combine them to compare with this model. A panelist cautioned that this effort to predict effect levels is going beyond the charge to the Task Group and into risk assessment, and suggested that this type of analysis could be presented in hypothetical case studies, but not as part of this report.

7.2.2 Charge Question 11

11. Discuss the model validation methods. Could additional validation approaches be used to enhance confidence in the model?

The panel members voiced some confusion regarding the data used to build the model and that used for the hold-out data validation. The authors clarified that initially in building the model structure they used all the study data, and tried many different approaches and combinations of parameters to create the best models as shown on page 25 of the report. They then used 70% of the data to estimate the coefficients. The remaining 30% of the data (hold out sample) was used to test how well the model predicted similar data. The results of this validation are shown in Figures in Appendix 6 of the report.

A panelist noted that it is not surprising that if the authors took 30% of the data out at random, that the model built with all the data (including that 30%) will fit. This is not the standard approach wherein one would build the model with a portion of the data (training set) and test with the remaining portion (test set). If less data had been used, some terms might not have been included. But other panelists thought that the parameters were likely to remain the same, and there was very little difference between the models for the different endpoints. There were biological reasons to include parameters. In building the model the authors looked at all the data on a particular endpoint and identified which parameters made biological sense and put those terms in. They then looked at the model and removed what was not necessary and used trial and error to make the best model. Similarly, they looked at all the data to determine the interaction terms and looked at quadratic terms, linearity, 2-way interactions, some 3-way interactions, etc., to determine the best model form. If only 70% of the data had been used, some interaction terms might not have been included, but a panelist pointed out that too many terms would not matter as this is a black box model. An author agreed that as a general principle using entirely different data would be best, but noted that the data were very homogenous from study to study and each study behaved much like the others. He thought that if he were to build the model with just 70% of the data it would likely look the same. The author also noted that hold out sample validation is a method published in text book by Harrell (2001) and has some standing. A panelist indicated that with such homogenous data, cross-validation would not provide that much information, but agreed with the author that the difference between interpolated and extrapolated

results was an important determination of the validation process and should be elaborated upon in the report.

A panelist noted that the model does a good job of estimating various biological effects based on PAC profile. This does not mean the PAC profile is the cause of toxicity; there may be an unknown factor involved, as discussed, but the model is predicting well. The panelist went on to note that the hold-out testing approach the authors used is similar to a well-accepted statistical process called jackknifing, in which one takes one point out at a time to estimate how well one did with each data point missing. This panelist felt that the approaches used by the authors was even better than this traditional technique and suggested that this be explained more completely in the body of the text and that the authors include some examples up front to show how they used the 30% hold out data to test the models. An author suggested that they could prepare tables similar to those found on page 25 of Appendix 6, which shows the correlation of all the hold out points, extrapolated hold out points, and interpolated hold out points.

Panel members discussed what is meant by validation, with one panel member pointing out that the term carries many meanings and could be taken to the point of being onerous. A panelist noted that the model has not been validated in the same fashion as other alternative methods and specifically cited Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (1997) and also Scala and Springer (1997) and Goldberg (1987). These methods deal with sensitivity, specificity, and accuracy of the model. These methods focus on use of a new data set for validation. Another said for the regulatory community what is desired is similar results with the same chemical from multiple laboratories and sometimes blinded. The word "improved" might better describe what is needed.

The author reminded the panel that the third step in their validation efforts used completely different data on the same endpoints. For example, thymus weight data from repeat dose studies were used to predict prenatal thymus weight and the predictions were very similar. However, they found poor results in the opposite direction, using the prenatal to predict repeat dose; but the interpolated data predictions were quite good and the extrapolated values were very poor. This is confirmation that the model works well when interpolating, but not as well when extrapolating. The author asked the panel if that was not sufficient in the classical sense. Panelists commented that what the authors did went a long way toward validation, but it is novel to what ICCVAM and the other papers recommend. Another panel member thought that the third approach using the repeat dose data to validate the developmental study results was a nice way to validate, but it did not cover all the endpoints due to the nature of the available data sets.

Several panel members suggested validating the model with data not used for model development. One asked whether the wealth of existing mouse oral and other data might be used and if that is not possible then the authors need to get confirming test data. Another panelist thought that the approach did well at prediction for the limited data set used, but thought there needed to be more structural and physical diversity with new test articles. Several panelists suggested that additional validation using newly collected data be done to enhance the model confirmation approaches used by the authors. These new data would be more akin to the classical definition of method validation.

Panel members suggested that the text should be made clearer regarding the data used to build and test the model and the various model confirmation methods used. Panelists suggested a number of approaches to consider for further validation. These included finding a new data set on a petroleum substance within the universe of the model, perhaps from a different laboratory; build the model with just 70% of the data and test with the remaining 30%; use studies analyzed with Method 1; and use one of the laboratory's studies to build the model and the other to test it.

The panel discussed the various approaches and difficulties with each. They discussed that one difficulty with using one laboratory's data to build the model and another's to test the model, is that the proportion of types of studies was not equal between the two laboratories. Another suggested idea was to use the laboratory with the most data for each endpoint as the training set and test with the other. The authors thought that using just one laboratory would make a much smaller data set and they would prefer random selection.

Panelists had concerns that there may not be comparable data from other laboratories. One panelist thought there may be no other laboratories that have done rat dermal studies of petroleum substances, in which case new testing may be needed. Another panelist cautioned that rat strains have been rebred over time and there may be differences in weights and percent resorptions among strains. Since variations in weight is a key measurement and marker for effects, one needs to look within the rat strain to identify sources of the rats and make any needed and appropriate adjustments for differences in the rats between studies

A panelist thought that it would be important to use the Method 2 statistical methodology on data generated using Method 1 (selected 2- to 5-ring aromatics). The available data indicate that Method 2 corresponds to biological effects slightly better than Method 1. A preliminary review of the data indicates a negative bias for the light molecular weight 2- and 3-ring PACs in Method 2 relative to Method 1. These observations suggest that the 3- to 5-ring PACs may be responsible for more of the biological effects compared with the 2- and 3 - ring PACs. However, these findings would benefit from a more detailed characterization than presented in the report. A more complete side by side statistical analysis of the Methods 1 and 2 results would likely help demonstrate the strengths and weaknesses of Methods 1 and 2. In addition, the more complete presentation of the statistical relationship of the Method 1 results and the biological effects would help data users apply PAH compositional information generated for many other products, like diesel fuel, mineral oil, and lubricating oils.

Regarding the idea of using Method 1 data, a panelist referred to Table 5 on page 18 of the report, and pointed out that models based on Method 1 without DMSO seem to match the results from Method 2. The panelist asked whether data from Method 1 studies could be used as for validating the models. Another panel member clarified that the analytical methods result in very different PAC content percentages for the various ring classes for a given sample and thus such data could not be directly translated to results derived based on Method 2. An author confirmed that there is no clear way to convert the PAC profiles from Method 1 to equivalent profiles for Method 2.

7.2.3 Charge Question 12

12. Are the conclusions reached by the authors regarding utility of the models for interpolation versus extrapolation justified? Are the presented definitions and procedures adequate to identify data sets that can be accurately predicted by the proposed models?

The panelists discussed the complex concept of interpolation and extrapolation and how each was determined. Through discussion with the authors, the concept was fully explained. It was clarified that the 7 - ring profile for the new substance has to be totally inside one of the profiles for the substances used to build the model. If it falls within any one of the existing profiles then it is interpolated; to be considered extrapolated, one component of the 7 - ring profile has to fall outside of all of the corresponding components of the existing profiles upon which the model was built. Panel members encouraged the authors to provide a more clear explanation of the procedure in their report.

A panel member asked whether the definition of extrapolation and interpolation includes other considerations and if it includes limiting the lower bound of values. The panelist expressed concern that the current definition is a one-tailed procedure, so a pure compound without any PAC, like hexane, would technically be "interpolated" even though it is vastly different from the types of petroleum compounds for which the model was built. To address this concern the panelist recommended adding another criterion to the definition based on the stream being predicted falling within the intended boiling point range of the data sets used to build the model.

The author clarified that the interpolation concept includes the consideration of dose. The models contained data sets such that all the PACs doses as low as zero were included for each ring size. The controls with dose 0 are an independent variable and it would be an extrapolation to predict a dose below the lowest dose used. However, another panelist rejoined that using the value of the control in the model, the dose is 0, and this is a problem for compounds whose PAC content is less than the dose range that is a minimum of those compounds in the data set.

A panelist expressed satisfaction with the work on interpolation and extrapolation, and commented that it is well known that if you work with linear models, it is dangerous to extend outside the range of your data. The panelist suggested these PAC profile data be put in a table that facilitates ready determination of whether a new data point is interpolated or extrapolated, with maximum allowable values for each ring, and the percentages for each ring for each sample. An author indicated that he had developed a spreadsheet using inputs for PAC concentration and dose to make the determinations of whether each point was interpolated or extrapolated.

A panelist pointed out that interpolation and extrapolation is an empirical observation to account for variability in the predictions and use of the concept contributes to the strength and validity of the model. The panelist further noted that the document identifies important limitations of the model based on testing of interpolated versus extrapolated data. The limitations of the model need to be clearly defined because people will use this in all sorts of ways the authors never conceived.

A panelist asked how the information on extrapolation would be used within HPV. The authors indicated that in completing the HPV matrix indicating data availability, new products that are extrapolated would be examined carefully to see where they fall relative to the prediction line and judge reasonableness. Some common sense would be used.

7.2.4 Charge Question 13

13. Are the conclusions in Volume 2 Section 5 biologically plausible, supported by the data, and do they reflect sound statistical analysis?

The panel discussed some of the conclusions previously, as noted above. Some panelists revisited some of their earlier issues and identified several specific statements of concern. Several panel members reiterated their concerns regarding reproductive toxicity conclusions and stated that not enough data were available to reach the conclusion on page 44 that "postnatal survival and development are likely to be the most sensitive endpoints of reproductive toxicity." The section on reproduction needs to be revised as indicated in earlier discussion.

A panelist questioned what is meant in the charge question by "sound statistical analysis." The panelist thought that the methods used were acceptable, although other methods could also have been used. There are some problems with cross validation that have been discussed and additional presentation of mixed modeling and sensitivity analysis has been suggested. All of this should be looked at together to come to an integrated conclusion as to the robustness of the results.

Another panelist was troubled by the last line on page 45, "The TG does not think the application of the model to other routes of exposure or species is justified at this time." The panelist understood what the authors intended, but was concerned others would take this out of context and say that the work does not apply to humans, which is not the case as animal data are extrapolated to humans in risk assessment routinely.

7.2.5 Charge Question 14

14. Discuss the models' strengths and limitations. Were they clearly identified and the implications well described? Are there ways to ameliorate the weaknesses? Is the documentation transparent and complete? Are uncertainties in the approach fully articulated? Are there suggestions for improving the presentation of the analyses or information?

Many of the models' strengths and limitations have been discussed above. Panelists reemphasized a number of areas and identified additional concerns. Many thought that the document's transparency could be improved with further explanation and description, and noted that transparency is important to establish credibility. Panelists offered suggestions regarding transparency throughout the meeting. Panelists suggested that the document would benefit from an overall road map to help the reader navigate the chapters and appendices. Particularly

important would be to provide cross-referencing and linking of information about particular substances through the chapters and appendices. Hyperlinks were suggested as one possible avenue to facilitate this. A panelist concerned with the representativeness of the materials used in the analysis suggested the title of the document be made more specific and that the limitation of representativeness should be made more clear and explicit in the report.

Several panelists suggested a more complete explanation of the model and the concepts of interpolation and extrapolation would be helpful. Panel members suggested the authors first describe the materials and what they know about them, explaining the chemistry and biology. The authors can lead the reader through an example compound and how it fits into the universe, and provide examples of compounds that do not fit. The authors should then explain how the model was built, the different steps, and why they decided on the final model form. One panelist suggested that the authors provide narrative descriptions of the model in plain words understandable to those who are not familiar with the mathematics. The panelist suggested taking the reader through the model step-by-step explaining what each term is for. Another panelist suggested they clearly define the types of petroleum substances this model applies to and clearly state that it applies to refined streams, and not additives.

One panelist noted that the authors have justified an empirical approach and built a correlation model based on associations. The authors should state this unambiguously and be more careful that they do not imply or claim causation in the text. The panelist was also concerned that in some places the document takes an advocacy perspective, rather than reading like a scientific research document. For example, the panelist noted that on page 12 of the report, the reader is referred to Appendix 3 for the criteria used to identify and exclude studies. The panelist thought that Appendix 3 presented opinions and conclusions, but not documented criteria. The panelist cited another example on page 19, Section 3.3, saying that this discussion is argumentative and not substantive.

7.2.6 Charge Question 15

15. Can the models that were developed be used to predict repeated-dose and developmental toxicity of PAC-containing petroleum substances for purposes of the HPV program?

Panel members discussed the models uses for HPV and also what constitutes a screening level toxicity analysis. They noted there are varying degrees and approaches to address toxicity. Some thought that if this work is for screening and prioritization in HPV, the current effort may be sufficient as is for these endpoints. Others panel members noted that additional validation of the general toxicity models and further work related to assessing correlations between maternal toxicity, dermal irritation, and developmental effects would be needed before the models could be used for HPV screening purposes. Some panel members felt that additional data related to potential reproductive toxicity would also be needed before making conclusions regarding that endpoint. All the panelists thought that if the intent is to get at the mechanisms underlying toxicity, more testing and validation of the model are needed to reach that next step.

Panel members made suggestions throughout the meeting for the authors to strengthen and improve the analyses, validation, models, and documentation. The panelists cautioned that when presenting the modeled predictions in the context of HPV, it should be made explicit which values are calculated from the models, with the concept and meaning of interpolated and extrapolated carefully explained. Panel members strongly cautioned that the model and results are not appropriate to use in quantitative risk assessments and that the documentation should so state this.

7.2.7 Charge Question 16:

16. Are there other important issues regarding model development, validation, and use?

Panelists were asked at the end of the meeting to identify their key concerns, suggestions, and recommendations. The following is a summary of the panelist's final statements and thoughts.

Several panelists reiterated the need to confirm the model with additional data that had not been used in the development of the model or parameters for it. Some thought that new data or studies are needed, while others suggested looking for confirmatory or existing published or proprietary data to help provide increased confidence about the model and its limitations. Others were comfortable with the repeat dose model and its validation with the prenatal test data, and suggested the authors might build the developmental toxicity models with just 70% of the data and test it with the remaining 30% of the data.

A sensitivity analysis was specifically recommended by a number of panelists. One noted that the analysis has potential, but all statistical analyses can use more documentation to identify robustness in results and because of the uncertainties inherent in the model. The panelist suggested three areas to include in the sensitivity analysis. First, model assumptions; for example, consider including results using mixed model analysis. Such an analysis will yield coefficients that are the same as the OLS method, but standard error will not be the same. These will determine confidence intervals and a lower confidence interval. The authors could consider presenting model results using alternative assumptions and using categorical count data as normal. Second, use some variations of the model to determine sensitivity of the model to different designs. The authors should demonstrate the impacts of selection of different variables, rather than just stating that multiple approaches were tried and had no impact. Specifically, exploration of interaction terms should be presented. The rationale for interaction term inclusion and exclusion also needs to be demonstrated. Third, the panelist suggested that the authors try to reduce the number of coefficients to see if they get the same results.

Another panelist suggested that a sensitivity analysis look for reasons for the outliers, for example bioavailability issues may be affecting hemoglobin results. The sensitivity analysis should be conducted with and without censored data to clearly show that data at extremes in the linear model. The panelist also suggested authors might show linear and nonlinear regressions to see if results improve. This panelist suggested the authors could evaluate the probabilistic variation of the parameters and distributions could be developed, but recognized that this would

be a lot of work. The authors could explore sensitivity of parameters; for example, whether the interaction terms affect results and what drives the model the most. However, the panelist cautioned that because this is a statistical correlation model and not mechanistic, there is a fine line to walk.

A panelist noted that the authors tried many different models and different approaches, without much difference in results. The panelist noted that the approach used here is used in toxicology and medicine, for example data to predict heart risk uses multiple factors to predict probability. While it is not perfect, it does a reasonable job of prediction. The panelist noted that the model has some internal validation and thought it correct to do this with the data for which they modeled. While more could be done, if the authors wait for the nearly perfect model and validation, science will not progress very rapidly.

Throughout the meeting, panelists suggested areas where further explanation was needed to help the reader understand what had been done and to make the work more transparent. Panelists suggested that more information was needed to provide context for the work, including more information about the HPV program and how the modeling results would be used in that program. The authors need to make sure that they respond to the different types of readers' needs, even if that means some redundancy of information or repetition of tables. A panelist also suggested that the tables need more explanations. Some suggested the authors conduct a thorough literature review and present this information up-front, discussing the compounds and what is known about their toxicity and chemistry. Throughout their description of methods and approaches, the authors should make it very clear whether decisions were made for statistical or biological reasons. They should better document the model form and make the domain of the model very clear and they should further explore and document various variables and conduct a sensitivity analysis to help understand potential failure modes in the models. A panelist also stressed that the authors need to define the limitations of the model very carefully and clearly and identify what would be inappropriate uses for the model or results.

Panelists cited the need for more information regarding maternal toxicity to determine whether the effects seen were caused by maternal toxicity or if the substances are developmental toxicants. One panelist emphasized that the authors should err on the side of caution when interpreting the data because developmental and maternal toxicity could be concurrent. Another reemphasized the need to explore irritation's role. While one panelist stressed that for risk assessment, it does not matter if the toxicity is to the conceptus or the mother, what is being looked for is the critical effect, or the effect at the lowest exposure. Another panelist noted that identifying the lowest effect level and effect is crucial, but developmental endpoints resonate with the public and so the authors should show that they carefully considered the developmental endpoints.

Panelists also reemphasized that the authors should compare PDx modeling with BMD modeling to see how the results compare, and that reproductive toxicity needs more exploration as the data argue both ways with regard to it being more or less sensitive than developmental toxicity.

Several panelists complimented the authors for their receptivity and helpfulness during the peer consultation, noting that the authors' patience and constructive listening would be very helpful in the further development of this method.

In general, the panelists were positive and comfortable with the model, the general approach, and the statistical analyses that were done. Many mentioned that they thought it showed great promise and some stated that they thought it was close to being ready for regulators, while others thought the model and confirming assays should be published in the open literature first to gain wider exposure and input. Individual panelists commented that it was good to mine the data, which so often is not done, and appreciated that the authors created a relatively simple construct. Several panelists suggested the authors publish this work in the peer reviewed literature to share the results. Others noted that once the model is validated and refined they thought it would be very useful for screening and development of additional hypotheses.

8. References

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

Grasso, P., Sharratt, M. and A.J. Ingram. (1988) Early changes produced in mouse skin by the application of three middle distillates. Cancer Letters, 42:147-155.

Gaylor, D.W. and Slikker, W. (1990) Risk assessment for neurotoxic effects.
NeuroToxicology 11: 211-218

Crump, K.S. (1995) Calculation of benchmark doses from continuous data.
Risk Analysis 15: 79-89 (1995).

Luster, M.I., Portier, C., Pait, D.G., White, K.L., Gennings, C., Munson, A.E. and Rosenthal, G.J. (1992) Risk assessment in immunotoxicology. I. Sensitivity and predictability of immune tests. Fund. Appl. Toxicol. 18:200-210.
Luster, M.I., Portier, C., Pait, D.G., Rosenthal,G.J., Germolec, D.R., Corsini, E. Blaylock, B.L., Pollack, P., Kouchi, Y., Craig, W., White, K.L., Munson, A.E. and Comment, C.E. (1993) Risk assessment in immunotoxicology. II. Relationships between immune and host resistance tests. Fund. Appl. Toxicol. 21:71-82.

Harrell, F.E., Jr. Regression Modeling Strategies, (2001) Springer-Verlag, NY,

ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods) (1997) Validation and regulatory acceptance of toxicological test methods: A report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods. NIH publication no. 97-3981. Research Triangle Park: NIEHS

Scala RA. and Springer J. (1997) Guidelines for the evaluation of eye irritation alternative tests: Criteria for data submission. Fd Chm Toxicol 35:13-22

Goldberg, A.M. (1987) Alternative Methods in Toxicology, Vol. 5. In Vitro Toxicology - Approaches to Validation. Mary Ann Liebert, Inc., New York,