External Peer Review

U. S. Environmental Protection Agency
Nitrobenzene Final Report

Final Compilation of Reviewer Comments
And Responses to Charge Questions

Prepared for
Integrated Risk Information System (IRIS) Program
Office of Research and Development
National Center for Environmental Assessment
U.S. Environmental Protection Agency

Prepared by
ORISE IRIS Technical Assistance Team
Oak Ridge Institute for Science and Education
Oak Ridge Associated Universities

June 2007

This document was prepared for the EPA by ORISE under interagency agreement No. DW-89939822-01-0 between EPA and the U.S. Department of Energy. ORISE is managed by Oak Ridge Associated Universities under a contract with DOE.

EPA Nitrobenzene Review

Table of Contents

EXTERNAL PEER REVIEWERS

Mark Miller, Ph.D., Chair
Wake Forest University School of Medicine

Bruce Allen
Bruce Allen Consulting

Rudolph Jaeger, Ph.D.
Environmental Medicine Incorporated

Martin Philbert, Ph.D.
University of Michigan

Richard Pleus, Ph.D.
Intertox

David Pyatt, Ph.D.
Summit Toxicology, LLP

Lorenz Rhomberg, Ph.D.
Gradient Corporation

The ORISE IRIS Technical Assistance Team has neither altered nor edited these comments for grammatical or other errors.

PEER REVIEW PROJECT

Todd Stedeford, Project Manager
U.S. Environmental Protection Agency

Leslie Shapard, Peer Review Manager
Oak Ridge Institute for Science and Education

CHARGE TO EXTERNAL REVIEWERS FOR THE IRIS ASSESSMENT OF NITROBENZENE

The U.S. EPA is conducting an external peer review of the human health risk assessment of nitrobenzene that will replace the assessment that currently appears on the Agency's online database, the Integrated Risk Information System (IRIS). The draft Toxicological Review contains derivations of the oral reference dose (RfD), inhalation reference concentration (RfC), cancer inhalation unit risk, and a cancer weight of evidence descriptor. Please provide detailed responses to the charge questions below.

General

Question 1 - Is the Toxicological Review logical, clear and concise? Has EPA objectively and transparently represented and synthesized the scientific evidence for noncancer and cancer hazard?

Mark Miller, Ph.D., Chair

The toxicological review of the literature appeared to be well documented and clearly presented. The authors did a good job of synthesizing the current literature and providing scientific evidence for the noncancer and cancer hazards. Several typographical errors were noted throughout the document, and the authors should edit and clean up the document. The readability of the document would be improved by providing summaries at the end of each section and adding graphical presentations of the data described in the tables – although this may seem to be redundant, it would aid in providing a visual representation of the data.

In addition, some additional discussion of the review criteria should be added. The authors should provide a checklist of what criteria were used to select specific studies for inclusion or further analyses. In a number of places throughout the text, the authors use the term significantly, which implies a quantitative measure – the authors should provide some explanation as to how they determine a particular finding is significant, or deserves more weight, than others which do not. This would enhance the transparency of the document and assure that the selection of specific studies was based on objective, easy to follow criteria.

Bruce Allen

The layout of the Toxicological Review is logical and concise. In general, the review is clear, although there are several issues specified below that I believe need to be addressed. The clarity would also be enhanced if corrections to several typographical errors could be made; at several places throughout the text and tables, some symbols (e.g., "<," "-," "x" as in "$x10^{-3}$") are not showing up correctly in the hardcopy we reviewed.

The representation and synthesis of the scientific evidence is for the most part transparent, but again there are some issues raised below that may need to be addressed in order to increase the transparency and to assure objectivity.

Rudolph Jaeger, Ph.D., DABT

This reviewer concludes that the Toxicological Review of Nitrobenzene is logical, clear and concise within the limits imposed by the form of the document and the amount of information that is available. There may be some editorial corrections that are required.

The material review appeared to be complete and well abstracted..

The EPA has objectively represented the data for the noncancer and cancer hazards. This reviewer and other panel members expressed a concern about the definition of the phrase "transparently represented". In my view, the data were enumerated in a linear fashion and they were interpreted as the current guidelines require.

Martin Philbert, Ph.D.

Yes: The authors are to be commended on the quality of this report. It is well-written, clear and concise, and makes the case for each of the categories examined.

Richard Pleus, Ph.D.

In general, the *Toxicological Review of Nitrobenzene* is logical in structure and approach. However, the document could be made clearer and more transparent.

While the document includes information on exposures to humans, these data primarily relate to accidental or intentional exposures (*e.g.*, suicide attempts). Although the document states that no epidemiological studies of exposures in the workplace or environment are available, a brief bibliographic search reveals that a number of studies appear to be available on workplace exposures, primarily biomonitoring studies and somewhat older foreign studies. The document should include a review of these studies and information from exposures or concentrations that have been measured in the workplace and in the environment. For example, what concentrations have been measured in contaminated soil, ground water, or in air in cities or around manufacturing facilities?

Given the relatively limited amount of toxicological data available for nitrobenzene, epidemiological data for other structurally similar compounds should be included (*e.g.*, other nitro aromatics). Later in the report, EPA provides a brief summary of toxicity data for four structurally similar nitroaromatic compounds, but the data discussed are almost exclusively from animal studies.

The document should have a thorough editorial review—I identified a number of typographical errors that made the document somewhat confusing to read.

David Pyatt, Ph.D.

For the most part, yes. However, an additional round of editorial review is needed, including some re-organization.

Also there were many 'qualitative' words that seemed a bit out of place in a highly technical document. For example, on p5… "revealed the presence of substantial amounts of…" It is not clear what is considered substantial. In the review of these various studies, if the quantification is available, then it should be provided. If not, the document should simply state what the data indicates, but that no quantitative information is available.

Lorenz Rhomberg, Ph.D.

For the most part, the Review is clear and logical. It tends to favor a study-by-study recounting of details that becomes a bit mind-numbing after a few dozen pages, and it would be good to interleave this with some summaries of key points and comparisons across studies. Some graphics helping to follow dose-response patterns within studies and relative doses across studies would be useful, and would help to foster more discussion of synthesis about consistency of potential endpoints and insights into potential modes of action.

Question 2 - Are you aware of additional studies that should be considered in the assessment of the noncancer and cancer health effects of nitrobenzene?

Mark Miller, Ph.D., Chair

Although this reviewer is unaware of any other studies that needed to be included in the assessment of the health effects of nitrobenzene, others at the review indicated that a more extensive literature describing workplace exposures and epidemiological studies could be found in the literature, particularly studies from foreign countries. The authors should check on this and include a discussion of these studies in the document.

Bruce Allen

I am not aware of any additional nitrobenzene studies.

Rudolph Jaeger, Ph.D., DABT

The reviewer is not aware of any additional studies that should be included or considered at this time.

Martin Philbert, Ph.D.

One point that is frequently asserted I the report revolves around the link between the enteric microfloral status and the ability to form methemoglobinemia. As indicated in the studies by Vasquez *et al*., (1995), nitroaromatic chemicals may induce methemoglobin formation directly and without activation by microsomes or in reducing/anaerobic environments. Also, the report does not take into consideration the potential for alterations in absorption due to changes in the morphology of the gut of germ-free or antibiotic-treated animals (Henegan, 1984). Moreover, it is assumed that the reactive intermediates required for the induction of MetHb formation are readily transported across the intestinal membranes rather than preferentially reacting with the contents of the gut. I am not certain that the assumptions made in the document are supported by the literature or that the authors may logically go beyond the assertion that the microfloral status influences susceptibility to MetHb formation.

Richard Pleus, Ph.D.

I found about 160 citations related to the key words nitrobenzene and toxicology. It appears that EPA cited only a small number of primary study references after 2004. Additional studies might be useful for this document. I mentioned them in Comment 1, above.

David Pyatt, Ph.D.

Not specifically for nitrobenzene, but some general references and background information on methemoglobin production might be useful. Same comment with regard to the use patterns and potential exposures of nitrobenzene.

Lorenz Rhomberg, Ph.D.

I would imagine there must be some relevant literature on methemoglobin and the effect of different percentages in the blood. That is, studies on potential modes of action when induced by chemicals other than nitrobenzene may also be relevant.

Oral reference dose (RfD) for Nitrobenzene

For the RfD, the draft reassessment uses a 90-day gavage study in rats by the National Toxicology Program (NTP, 1983) that was reviewed by the NTP Pathology Working Group. The critical effects used were splenic congestion, increased methemoglobin levels, and increased reticulocyte count. Alternate derivations for points of departure are presented in Appendix B-1 of the draft Toxicological Review.

Question 1 - Is the selection of the NTP (1983) study as the principal study scientifically justified? Is the rationale for selecting this study transparently and objectively described?

Mark Miller, Ph.D., Chair

Based on the review of the literature, the NTP study appears to be the most scientifically relevant study for use in the analyses although some concern was raised at the panel discussion regarding the use of a bolus dose of chemical. The authors provided a clear, well described justification and rationale for their decision to utilize this study, but should examine the dosing regimen more closely and compare this study to others that used lower doses over more extended time periods, which would be more likely to mimic real life exposure patterns. A better summary of how the route of exposure influences the metabolism and distribution of the compound is needed, and if possible, the authors should identify which forms of drug metabolic enzymes specifically metabolize nitrobenezene following different exposure routes. For example, are the enzymes that metabolize nitrobenezene in the lung the same as those that do so in the liver? How do organ-specific differences in enzyme composition influence site-specific metabolism?

Bruce Allen

Based on the list of studies available, it appears that the choice of the NTP study is scientifically justified. However, I would not say that the rationale for its selection is transparently described. In fact, it is difficult to tease out the complete rationale for the selection and, in instances where a more difficult choice among potential studies might occur, it is not clear how the selection would be accomplished. I would suggest that there be some sort of list of the criteria and that the criteria be listed, perhaps in bullet item format, showing why the chosen study was superior to others. For example, if it could be shown explicitly that it

- is long enough (or the longest of all available)
- tested several doses in an appropriate range
- tested multiple species
- included an extensive examination to reveal a variety of endpoints
- etc.

then this would provide an objective and transparent basis for selecting the study. This would improve the ability of the reader to understand why the principal study was selected as such.

Rudolph Jaeger, Ph.D., DABT

In the opinion of this reviewer, the NTP (1983) study is valid and justified as the principal study that will fulfill the criterion of utility and validity. The rationale for selection is objectively described. The term transparent needs to be defined in this context.

Martin Philbert, Ph.D.

The choice of the 1983 NTP study appears to be justified and adequate for the purposes of this assessment.

Richard Pleus, Ph.D.

The NTP (1983) study is clearly important, representing exposure to rats and mice of both sexes at nitrobenzene doses ranging from 18.75 to 300 mg/kg-day for 90 days. However, EPA cites data from another study in rats (Fisher 344) exposed to lower doses for 28 days (Shimo *et al.*, 1994); EPA should include a discussion of the merits or deficits of this study in Section 5.1.

David Pyatt, Ph.D.

There is always a concern with gavage studies that a one-time bolus dose may not provide relevant information about human exposures (environmental or occupational). Additionally, it is well known that the pharmacokinetic behavior (and subsequent toxicity) of a single oral dose can be significantly different than what is observed via continuous exposures. Therefore, the 2 year inhalation bioassay would likely provide more relevant data. However, there was fairly good agreement between various routes of exposures and the observed toxicity of nitrobenzene. Additionally, a route to route extrapolation would be very difficult to do in the absence of well characterized PBPK model for nitrobenzene. Therefore, using the NTP (1983) study was scientifically justified.

Lorenz Rhomberg, Ph.D.

The selection is well justified but it would be good to have some comparison of this study with the alternatives, since there is a lot of corroboration and the overview would help understand dose-response.

Question 2 - Splenic congestion (increased by 10%), methemoglobin levels (increased by 1 SD), and reticulocyte count (increased by 1 SD) relative to the control values serves as the basis for the RfD. Is the selection of splenic congestion, methemoglobin levels, and reticulocyte count as the co-critical effects for deriving the RfD scientifically justified? Has the rationale for selection of these critical effects been transparently and objectively described? Is it appropriate to derive the point of departure by averaging BMDLs across sexes and co-critical effects?

Mark Miller, Ph.D., Chair

The 3 endpoints that were selected appear to be the most relevant for assessing the toxic effects of nitrobenzene. The rationale for utilizing these endpoints was well described by the authors and appears to be scientifically justified based on the literature. The authors provided a clear and objective explanation for their decision to utilize these endpoints. This reviewer lacks sufficient expertise with these types of modeling analyses to comment on the use of the BMDL approach.

Bruce Allen

There are several concerns and problems with the modeling and treatment of the endpoints from the NTP study.

First, the data used in the analysis are never presented in a clear and transparent manner. Neither Section 5.1 (Oral Reference Dose) nor the associated appendix (B-1) show the data used in the modeling. Although the response values are (partially) presented earlier (Section 4.2.1.1), it would be useful to have the data modeled shown, at least in the appendix where the details of the analysis should all be available. I say that the response values are "partially" shown because even in the tables of Section 4.2.1.1, the sample sizes associated with the response means and standard deviations (for the continuous endpoints such as metHb and reticulocyte count) are never presented. It is not sufficient to present in the text the initial dose-group sizes, because those often do not correspond to the number of animals examined for any given endpoint.

In fact, that is an issue for the high-dose rat groups (male and female), where survival reduced the number examined (for some of the endpoints at least). One might infer from Tables 4-3 and 4-5 that there was only one observation in the high-dose males for those continuous parameters ("means" but no standard deviations are shown), but it is less apparent that the sample size for the continuous endpoints in the female rat high-dose group is only 7. Moreover, it is never stated that the modeling of continuous endpoints for the male rat data set was done without the high dose group; this needs to be explicitly stated and reiterated by showing the input data sets used for the modeling (in the Appendix if nowhere else).

The fact that there was only one observation in the male rat high-dose group (and the exclusion of that observation from the modeling) has led to a number of concerns. It should be possible to include that observation in the modeling. I have tried it with BMDS and the fitted model results do appear to appropriately account for that observation in the maximum likelihood fitting of the dose-response models, although this needs to be confirmed with additional analyses (see also below). What BMDS cannot do correctly is to estimate the likelihoods for the models used to determine if the dose-response model being fit is adequate. This is a limitation of BMDS, not of the method. What should have been done is to use or develop alternative tools for calculating the likelihoods needed so that a complete analysis could be presented. I have a tool developed in Excel that is capable of doing the analyses required, so it is certainly not a high hurdle to jump.

Ignoring the high-dose male rat observation has other negative implications for the analysis presented. For 2 of the 3 "co-critical" effects, the BMDs and BMDLs used to get the point of departure are derived from models that have "unrestricted betas," which entails that the dose response is not constrained to be monotonic. The document suggests that those curves were examined and found to give "plausible" curve shapes. Yet, because the high-dose male rat observation was not considered in the modeling, it did not have any effect on the predicted curve shapes. And apparently the evaluation of "plausible" curve shapes did not consider that left-out observation either. My independent analyses have shown that the model selected to give the BMD and BMDL for metHb in male rats (unrestricted 2nd degree polynomial) predicts a mean metHb of

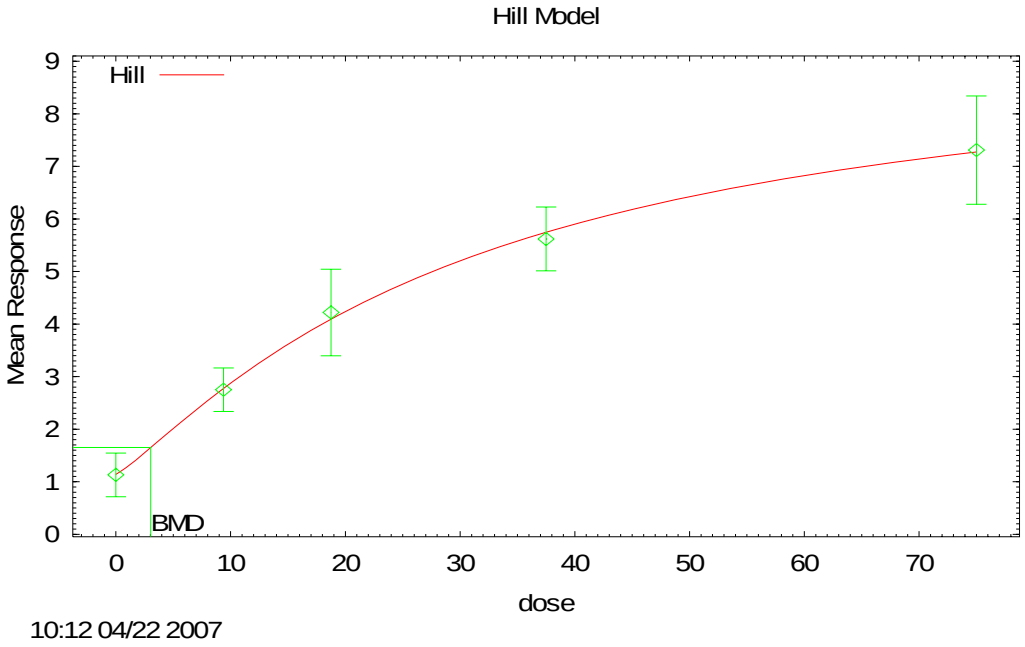
-0.19 at the high dose of 150 mg/kg, while the observation from the NTP study was 12.22. Similarly, for the unrestricted 3rd degree polynomial, the predicted response at 150 mg/kg was 25.5, a deviation as extreme as that obtained from the unrestricted 2nd degree polynomial, but in the opposite direction. These do not appear to be plausible curve shapes, especially considering that there is an *existing* observation that is seriously at odds with either model's predictions. The appendix (B-1) should have plots of the dose-response results, especially when some qualitative judgments about the adequacy of curve shapes are being made.

On the issue of survival problems for the high-dose rats, note that the splenic congestion endpoint (a quantal response) includes all 10 animals put on test in the high-dose groups (both males and females). Is it appropriate to count all animals (even those that died early) in that determination? I note that all 10 animals in each of the high-dose groups were indicated to have splenic congestion (Tables 4-7 and 4-8) so perhaps that is not an issue with that endpoint per se, but I would think that there should be some discussion of the point (at least acknowledging that the presence/absence of such congestion was not ascertained at the same time or age across animals in some groups).

There are other limitations of the analysis in support of the oral RfD. The male rat reticulocyte response does not appear to have been fit by anything other than a polynomial model (the linear is a special case of the polynomial[1]). Why not the power and Hill models?

The table of model results for the female rat reticulocyte response states that the Hill model crashed and the male rat metHb results table indicates that the Hill model could not compute the BMDL. These are not satisfactory or sufficient results. As indicated above in relation to the limitations of BMDS in the calculation of the likelihoods for some models, when BMDS fails then other tools need to be explored. This is particularly important for the male rat metHb response, because the fit of the Hill model to the data (at least when the high-dose group is ignored, as EPA has done) is one of the best ones, especially among models that are monotonic – see BMDS output below:

[1] The fact that a 1st degree polynomial is the same as a linear model appears to have been missed in the quantal analyses of spleen congestion, since the tables for that endpoint list both 1st degree polynomial and linear results every time. It is my opinion that the linear and quadratic models do *not* need to be listed at all, since they are both special cases of the polynomial and Weibull models; linear or quadratic results will be obtained if the data suggest that such curve shapes are the most likely, especially for the linear model which is the special case associated with common model constraints (restricted betas or Weibull power $\geq$1).



Modeling results for female rat metHb are not even presented. The reason given is that the BMDS variance model does not fit. Again, this shows a lack of persistence in the conduct of the analysis. One straight-forward "fix" to this problem is to add a constant to the mean of each dose group so that they "match-up" better with the observed variances. Note that when a constant is added to a random variable, the variance of the transformed variable is the same as the original variable, so that BMDs based on standard deviations from the control mean will not be affected by that transformation. A quick analysis that took me about 20 minutes was to determine that if one subtracts 0.96 from the reported mean metHb values and reruns BMDS, then one can obtain an adequate representation of the results, both for the variance model and the fitted curve (which in my example was the power model) – see the following BMDS output:

Table of Data and Estimated Values of Interest

Dose	N	Obs Mean	Est Mean	Obs Std Dev	Est Std Dev	Scaled Res.
0	10	0.01	0.0107	0.03	0.0282	-0.0802
9.38	10	1.1	1.35	0.45	0.624	-1.25
18.75	10	2.66	2.29	1.09	0.878	1.32
37.5	10	4.31	3.91	0.76	1.24	1.01
75	10	5.89	6.69	2.25	1.74	-1.44
150	7	11.8	11.4	1.83	2.46	0.414

Model Descriptions for likelihoods calculated

Model A1 $Yij = Mu(i) + e(ij)$
$Var\{e(ij)\} = Sigma^2$

Model A2 $Yij = Mu(i) + e(ij)$
$Var\{e(ij) = Sigma(i)^2$

Model A3 $Yij = Mu(i) + e(ij)$
$Var\{e(ij)\} = exp(lalpha + rho*ln(Mu(i)))$

Model A3 uses any fixed variance parameters
that were specified by the user

Model R $Yi = Mu + e(i)$
$Var\{e(i) = Sigma^2$

Likelihoods of Interest

Model	Log(likelihood)	# Param's	AIC
A1	-39.339428	7	92.678857
A2	7.267274	12	9. 465452
A3	2.768877	8	10.462246
fitted	-0.796492	5	11.592983
R	-104.058653	2	212.117306

Explanation of Tests

Test 1: Do responses and/or variances differ among Dose levels? (A2 vs. R)
Test 2: Are Variances Homogeneous? (A1 vs A2)
Test 3: Are variances adequately modeled? (A2 vs. A3)
Test 4: Does the Model for the Mean Fit? (A3 vs. fitted)

(Note: When rho = 0 the results of Test 3 and Test 2 will be the same.)

Tests of Interest

Test	-2*log(Likelihood Ratio)	Test df	p-value
Test 1	222.652	10	<.0001
Test 2	93.2134	5	<.0001
Test 3	8.99679	4	0.06118
Test 4	7.13074	3	0.06785

Some additional analysis could probably identify choices for a transformation that are even better than the one I obtained. In fact, some careful programming could automate the process to find an "optimal" additive constant.

The point of the above demonstration is that, as in some of the other cases highlighted above, the analysis presented for the nitrobenzene RfD determination is not thorough or rigorous enough to satisfy this reviewer that an adequate basis for that RfD has been obtained. Thus, the questions about the scientific basis for selecting specific endpoints as "co-critical" may not be answerable until the analyses are complete enough. I anticipate, however, that there will need to be some careful definition of what "co-critical" means, how and when endpoints become co-critical and why is it advisable or necessary to do something like average BMDs and BMDLs when co-criticality is observed. For instance, is there any greater stability or optimality or health-protectiveness that is obtained by averaging BMDs from co-critical endpoints? There is no discussion of these issues in the document now but I think there needs to be.

Rudolph Jaeger, Ph.D., DABT

[Splenic congestion (increased by 10%), methemoglobin levels (increased by 1 SD), and reticulocyte count (increased by 1 SD) relative to the control values serves as the basis for the RfD.] The three end points as a basis for the RfD are described in a manner that is clear and based on the prevailing Agency policy, they are presented scientifically.

(Is the selection of splenic congestion, methemoglobin levels, and reticulocyte count as the co-critical effects for deriving the RfD scientifically justified?) Yes.

(Has the rationale for selection of these critical effects been transparently and objectively described?) Yes.

(Is it appropriate to derive the point of departure by averaging BMDLs across sexes and co-critical effects?) Yes.

Martin Philbert, Ph.D.

Yes to all questions.

Richard Pleus, Ph.D.

Scientific transparency requires that subjective descriptors be well defined.. The co-critical effects were chosen due to statistically significant differences between treated and control animals in the NTP Study and because splenic congestion and increased reticulocyte count are histopathologic responses that are "potentially associated" with methemoglobinemia. It would be useful if EPA provided a clear and transparent definition and description of the adverse consequence of splenic congestion. For example, on page 40, EPA describes splenic congestion as "…increased brown pigmentation in red pulp, and increased extramedullary hematopoiesis…" However, it is not clear what the consequences of "brown pigmentation" and "red pulp" are from a pathophysiological perspective. Explanation of how increases in methemoglobin and reticulocyte count are adverse would also be important for scientific transparency.

David Pyatt, Ph.D.

These appear to be appropriate toxicological endpoints to use and the rationale was sufficiently described. However, there is ample evidence presented in the recovery experiments (as well as the clinical literature) that most, if not all, of these endpoints are completely reversible upon elimination of or decrease in exposure. This could be an important consideration in the interpretation of this potential toxic endpoint and would be a useful discussion to include in this document. The compensatory response observed in the long-term bioassay with regard to this end point would also be of interest to some readers.

Lorenz Rhomberg, Ph.D.

The choice of critical effects seems appropriate. The basis for choosing the BMR for the continuous endpoints is just recourse to guidance; there should be some discussion about what is known about how much change in the continuous measures might be expected to lead to adverse consequences (i.e., a biological justification of the BMR).

The rationale for averaging the BMDLs is not very well explained. Yes, the endpoints are mechanistically related, but it is not clear why they should share a BMD.

Question 3 - Are the uncertainty factors applied to the point of departure for the derivation of the RfD scientifically justified and transparently and objectively described?

Mark Miller, Ph.D., Chair

No. It would improve the transparency of the application if the authors documented how they arrived at the use of a specific uncertainty factor. It would also be useful and appropriate for them to provide the criteria they used in picking a specific uncertainty factor and should include a description of EPA guidelines on the use of uncertainty factors – this could be done as an additional appendix.

Bruce Allen

Although they may be transparently described, I have issues with the values for the uncertainty factors associated with subchronic to chronic exposure and with data base deficiencies as described in the responses to the next 2 questions.

Rudolph Jaeger, Ph.D., DABT

Yes.

Martin Philbert, Ph.D.

Requires discussion at the face-to-face meeting.

Richard Pleus, Ph.D.

The type and value of uncertainty factors EPA applied (*e.g.*, 1, 3, or 10) for the derivation of the RfD appear to be transparently and objectively described. Uncertainty factors are policy, not scientific facts. The total UF value of 1000 is composed of four parts: intraspecies uncertainty factor of 10; interspecies uncertainty factor of 10; subchronic to chronic uncertainty factor of 3; and database deficiency uncertainty factor of 3[10*10*3*3 ~1000]. 1,000 is a relatively large UF, and reflects a relatively uncertain database for extrapolating an RfD. An investigation of other types of data, *e.g.*, data on relative animal-to-human toxicity for structurally similar compounds, might provide additional information that could mitigate some of these uncertainties.

David Pyatt, Ph.D.

It would be helpful if the accidental human ingestion studies were quantified (to the extent possible). From the information on amounts ingested and nitrobenzene content, it seems plausible that the ingested dose of nitrobenzene could be calculated. This would allow for the actual dose to be correlated with the reported toxicity. As there are a wide range of ages involved in these various reports (including small children), this may be useful. A more quantitative analysis of this literature base may provide a better understanding of any potential differences in response due to age.

The G6PD deficiency is a legitimate concern as far as sensitive sub-populations are concerned, although it is not clear on a quantitative basis how much more susceptible G6PD deficient individuals actually are. There are ample clinical data from a wide variety of oxidatively damaging drugs on this endpoint that might provide some relevant insight into this question. Quantitative information of the differential susceptibility of fetal vs. adult Hb is also available.

The 10 fold safety factor to account for animal to human extrapolation was not fully justified. What quantitative evidence supports the position that rodents have a ten fold decreased sensitivity to oxidatively damaged RBC and compensatory changes? Again, there is likely clinical data with a

variety of drugs on this endpoint in humans and likely in experimental animals as well. A direct dose comparison is potenmtially possible. Does this literature support that experimental animals are really 10 times less sensitive than humans with regard to MetHb formation?

Lorenz Rhomberg, Ph.D.

The application of a factor of 10 for animal-to-human for the oral RfD and only 3 for the inhalation RfC may be frequent practice, but it is not clear what the rationale for this is. One would have liked to see some discussion of the likelihood that humans might have different met-Hb levels for a given intake of nitrobenzene than rats or mice, given the species differences seen among rodents and the dependence on gut flora.

Question 4 - An uncertainty factor of 3 was selected to account for less-than-lifetime exposure in the principal oral study. Is the choice of this UF scientifically justified and transparently and objectively described?

Mark Miller, Ph.D., Chair

I am not familiar enough with the use of uncertainty factors to be able to comment on the most appropriate number, however this again raises the issue of the authors providing a clearer discussion of the guidelines for using and assigning uncertainty factors and their criteria for selecting the uncertainty factors documented in this review.

Bruce Allen

I find the rationale for selection of this factor to be a bit problematic. As pointed out, there are very good subchronic and chronic inhalation studies. Those studies did not show that the severity or frequency of responses increased as exposure increased from subchronic to chronic duration. In fact, elsewhere in the Toxicological Review document (Appendix B-2), one of the main noncancer effects (increasing metHb) that may account for the splenic congestion and reticulocyte endpoints also favored as co-critical effects in the document, was said to be subject to some sort of compensatory mechanism, when considering the difference between interim (or subchronic) and chronic exposures. If that compensatory mechanism is not related to the stated differences between inhalation and oral routes of exposure with respect to metabolic details (3-step vs. 6-step), then why would those same compensatory responses not be operative for oral exposures as well?

With respect to the details of the metabolism, the rationale presented also appears to be rather selective. As an example, the document states (p. 118) that nitrosobenzene is formed from oral exposure “whereas inhalation exposure leads to nitroanion radical formation.” But, earlier in the document (Figure 3-18), metabolism of inhaled nitrobenzene is also shown to yield nitrosobenzene. The role of the radical formation, given what the document repeatedly refers to as at best weak evidence of genotoxicity, and in light of the fact that the “noncancer effects of inhalation exposure

to nitrobenzene were generally similar to those observed following oral exposure" (p. 119) may be highly suspect.
Thus it is difficult for me to agree that route-specific differences in metabolism could lead to a strong possibility for other toxic endpoints occurring over long exposure periods that are not already accounted for in the subchronic study.

Rudolph Jaeger, Ph.D., DABT

Yes. The various factors employed are clearly enumerated and when compared to the available data, they are consistent with the usual Agency policy as regards such estimates. Transparent in this context may mean that the description of the basis for the choices is clear and this does appear to be the case.

Martin Philbert, Ph.D.

Requires discussion at the face-to-face meeting.

Richard Pleus, Ph.D.

The principal study is a 90-day gavage study in rats and mice, with the rat data used as the basis for the RfD. Because this is not a chronic study, use of an uncertainty factor for less-than-lifetime exposure is a reasonable consideration, unless clear evidence indicates that chronic exposures are not likely to yield more significant adverse effects than represented by the 90-day study. EPA presents a number of reasons why they chose a UF of 3, which appears to be a reasonable approach to weighing scientific information in selecting a UF. However, EPA could clarify their choice by describing the guidance they used to evaluate the scientific studies and select UFs.

David Pyatt, Ph.D.

There is very good consistency across studies with regards to the most sensitive endpoint. Therefore, is doesn't seem likely that additional toxicity not identified in the sub-chronic studies would occur in the chronic bioassays. Therefore, it is not clear that this uncertainly factor is required.

Lorenz Rhomberg, Ph.D.

I think serious consideration should be given to whether one needs a factor for sub-chronic to chronic extrapolation. There is the experience with the inhalation studies that suggests no such factor is needed. Moreover, more consideration should be given to the nature of the endpoints and what it is about them that might or might not get worse with longer durations of exposure. It would seem that, especially for methemoglobinemia (but similarly for the other effects) the key is the balance between the rate of new generation and the repair of met-Hb. This should reach a plateau rather quickly, and so toxicity would seem to be a question of whether the resulting degree of methemoglobinemia is tolerable or not. The results of much shorter-term studies tend to show just

the same kind of results as the longer-term studies, and if the dose-response patterns for these were compared, it may well show that a dose rate has just about the same effect no matter how long it is sustained.

Question 5 - An oral database uncertainty factor of 3 was applied. The database of oral studies includes the principal study (NTP, 1983b), a 90-day gavage study in two species and both sexes; high quality reproductive/developmental studies (Mitsumori et al., 1994; Morrissey et. al., 1988; Bond et al., 1981); structure-activity relationship studies comparing nitrobenzene to dinitro- and trinitrobenzene; and a multidose immunological study in mice (Burns et al, 1994). However, due to a lack of an oral multigeneration reproductive toxicity study and in light of evidence of male reproductive toxicity, a factor of 3 was applied. Is the choice of an UF of 3 scientifically defensible given the available oral and inhalation databases? Does the available data suggest that oral exposures may result in new adverse effects at oral doses equivalent to or lower than the inhalation concentrations used in the multigeneration reproductive and developmental study by Dodd et al. (1987)?

Mark Miller, Ph.D., Chair

This is an area fraught with tremendous uncertainty. As I noted above, I am not familiar enough with modeling analyses and uncertainty factors to comment on whether 3 is the most appropriate correction factor for this type of data. Again, a more detailed discussion of how specific uncertainty factors were selected would be very useful and aid in interpretation of the authors' rationale.

The literature for many other compounds has shown that the route of administration of a compound can affect its biological actions. In the case of nitrobenzene, the metabolic activation of this compound is very different following oral *vs.* inhalation exposure. The bacterial nitro reduction of nitrobenzene by the gut flora is the major metabolic pathway and produces higher concentrations of different active metabolites than metabolism following oral exposure. Thus, without hard scientific data and a comparison study, it would not be possible to project potential long term effects on reproductive function from chronic, multigeneration exposure to oral nitrobenzene relative to inhalation exposure. Use of an uncertainty factor appears to be well justified in this case, and the key question is whether 3 is sufficient to account for potential effects of oral exposures at lower or equivalent concentrations to that seen in the inhalation studies. Some additional discussion and justification for the use of an uncertainty factor of 3 is needed.

Bruce Allen

The discussion of a previous oral RfD determination summarized in the Toxicological Review on pp. 118-119 suggests that a route-to-route extrapolation was made, so that the previous assessment could be based on inhalation study results. So, given even a crude route-to-route extrapolation capability, why has the document itself not addressed this question of oral exposures equivalent to or lower than inhalation concentrations from Dodd et al. That is, it is something that should have been done in the document and not posed to the reviewers. At the external review meeting of May

15, 2007, other reviewers (Drs. Jaeger and Rhomberg) had some very reasonable suggestions for addressing the issue of route-to-route extrapolation; the document should explore those options for answering this question. As Dr. Rhomberg pointed out, even approaches that roughly estimated what the oral exposure equivalents would be might be sufficient to answer the question in the negative; if those oral exposures are substantially greater than those associated with any correctly derived oral point of departure from the selected study , then no other UFs would be needed.

In general, I have problems with assessments that assign different database deficiency UFs for the oral and inhalation routes. As suggested in the preceding paragraph, there are ways to address database deficiencies for one route with observations from the other route, so there does not appear to be the need to separately consider a database deficiency UF for those two routes.

The rationale given for the database UF is that "there are known differences in metabolism between oral and inhalation exposures that may produce uncertainty in the potential for transgenerational effects from longer term oral exposures" (p. 118). As discussed above, the metabolism differences appear not to be that great; they do not produce any observable difference in systemic response from oral or inhalation exposures.

In summary, with respect to the UFs for chronic to subchronic and for database deficiency, it appears that these are not unrelated concerns. The structure that forces them to be considered separately is an artifact that constrains the decision making in this case. At most, I would suggest a factor of 3 for these two UFs combined, pending more complete investigation of the route-to-route extrapolation that might allow a fuller integration of the oral and inhalation databases.

Rudolph Jaeger, Ph.D., DABT

Yes. The addition of inhalation as a route in this section in the question asked is not warranted as the oral route is the pathway being considered.

No, this is unlikely. It is difficult, if not impossible, to calculate dose equivalence between oral and inhalation routes without a precise basis and understanding for the mechanism of action for the biologic effect and for an equivalent metric for the measured changes between the routes.

Martin Philbert, Ph.D.

Requires discussion at the face-to-face meeting.

Richard Pleus, Ph.D.

The strength of EPA's discussion is that they present a number of reasons why they chose a UF of 3; this is a reasonable approach to weighing scientific information in order to derive an oral RfD. However, as in my previous comment, it would be useful for the document to identify the guidance EPA uses to evaluate scientific studies and how this influenced their choice of UFs.

David Pyatt, Ph.D.

Available evidence suggests that the reproductive toxicity occurred at doses greater than doses required to induce the critical effects used to establish a POD. Therefore, it was not clear why this additional safety factor was applied.

Additionally, the multi-generation inhalation study does not support the position that these reproductive effects would likely occur at doses below those associated with the critical effects.

Lorenz Rhomberg, Ph.D.

The need for this should be discussed at the meeting, but the case is well laid out in the document. The question is how much the experience of the inhalation study should obviate this factor in the oral case.

Inhalation reference concentration (RfC) for Nitrobenzene

The draft reassessment of nitrobenzene uses a 2-year inhalation study for deriving the RfC. Several endpoints were identified as potential critical effects, including bronchiolization of the alveoli (mice), olfactory degeneration (mice), methemoglobin levels (rats), and splenic congestion (rats). Bronchiolization of the alveoli was chosen as the critical effect for the following reasons: 1) bronchiolization of the alveoli, a metaplastic lesion, occurred in ≥87% of male and female mice at the lowest exposure concentration and none of the controls (olfactory degeneration occurred in 1.5% of males and 32% of females at the lowest concentration; methemoglobin levels were ~3% in both male and female rats at the lowest concentration tested); 2) the severity of bronchiolization of the alveoli was consistent in both male and female mice; 3) bronchiolization of the alveoli is a portal of entry effect that is relevant to oronasal breathers (e.g., humans); and 4) this endpoint was obtained from a chronic inhalation study in which 43% of male mice developed bronchio-alveolar adenomas or carcinomas at the highest concentration tested. Alternate derivations of the RfC are presented in Appendix B-2 of the draft Toxicological Review.

Question 1 - Is bronchiolization of the alveoli the most scientifically justifiable endpoint on which to base the RfC? Have the rationale and justification for this selection been transparently and objectively described? Are there any other studies that you believe would be justified scientifically as the basis for the RfC?

Mark Miller, Ph.D., Chair

This reviewer had concern that the bronchiolization of the alveoli was not the most justifiable endpoint. First of all, it is not clear from the text exactly what "bronchiolization" of the airway epithelium means, and how the authors can distinguish this effect as a non-cancer endpoint given the likelihood that enhanced proliferation of bronchial epithelial cells and damage to lung tissue often precedes the onset of frank tumor formation. Although this is described somewhat in footnote 4 at the bottom of page 53, some additional discussion and a figure demonstrating the pathological lesion are needed. If possible, the authors should identify the exact cell types involved in this process or indicate that this is unknown. This reviewer is very concerned that this process is likely a pre-neoplastic event and classification of bronchiolization as a non-cancer hazard is likely inaccurate and inappropriate, especially given the later documentation of lung tumors in the same species.

It is also not clear what the mode of action of nitrobenezene is in lung tissue – does this effect occur as a result of metabolism or a direct effect of the compound on lung tissue? The metabolic competence of the specific cell types in the lung and a correlation between metabolic activity and damage to the airway tissue was not discussed; this gets back to exactly what bronchiolization means – which cell type(s) are involved, are they metabolically competent, is there equivalent metabolic activity between mouse, rat, and human lung cells for nitrobenezene, what are the isoform(s) of metabolic enzymes that mediate lung specific metabolism of nitrobenezene? There is some concern that, in this case, the lung damage could be a very species-specific effect and limited to the mouse – some discussion of the relevance of this endpoint to human exposure is warranted.

It appears to this reviewer that the occurrence of methemoglobinemia would be the more sensitive biomarker/endpoint. While I understand the concerns for using a biomarker for toxicity that displays a dose-response effect, it is possible in this case that the lowest dose used in the cited studies may already have provided near maximal effects. As a result, the authors of the study may be ignoring the most sensitive biomarker of nitrobenezene exposure, especially given the data seen with the oral dosing studies and the concerns noted in the previous two paragraphs regarding the relevance of bronchiolization to humans. This reviewer believes that the lower RfC obtained for the methemoglobinemia should be employed as the more relevant endpoint due to its demonstrated high sensitivity in every exposure pathway analyzed. One should keep in mind the fact that none of the doses employed exhibited a no-effect response for all of the examined endpoints.

Bruce Allen

I do not see that the document presents any scientific justification of the choice of bronchiolization of the alveoli, except perhaps justifications (presented later in Section 6.1.5, p. 138) related to the fact that that endpoint may be relevant to both facultative and obligatory nose-breathers (in contradistinction to olfactory degeneration). In fact, in that same paragraph, it is stated that there is no hypothesis concerning the development of bronchiolization. If, as stated earlier in Section 6.1.5, p. 137), metabolic activation is required for any of the toxic effects associated with nitrobenzene,

then how does that metabolic requirement fit in with the effects in the alveoli? That issue should be addressed.

The choice of bronchiolization appears to be solely driven by the fact that it gives the lowest point of departure. That may be fine, lacking other information relevant to picking a critical endpoint, but it should be stated clearly and prominently as the deciding factor.

Rudolph Jaeger, Ph.D., DABT

Based on the material presented and the unique nature of this respiratory outcome in reference to the inhalation route (versus oral or dermal where a similar outcome not seen), the application of these data are valid.

Yes, subject to a better definition of transparency as noted previously.

Not studies that are known to this reviewer.

Martin Philbert, Ph.D.

NO. This endpoint is highly species-specific and does not follow a classical (or otherwise) dose dependent response. No supporting scientific evidence is provided for metabolic activation of nitrobenzene by pneumocytes (type I or II) or by alveolar macrophages sufficient for such a change in morphology. Moreover, it is not clear that the mouse pulmonary metabolic machinery (as it has been described in the literature) is present, required or sufficient for induction of similar changes in the human or other species.

Richard Pleus, Ph.D.

CIIT (1993; published as Cattley et al., 1994) is the key study used in the development of the inhalation RfC. This study used B6C3F1 mice and F344 rats of both sexes and CD male rats. Rats were exposed to 0, 1, 5, or 25-ppm nitrobenzene and mice to 0, 5, 25, or 50-ppm nitrobenzene for 6 hours/day, 5 days/week for 2 years. Bronchiolization appears to have only been reported in studies conducted in mice and was reported at 5 ppm, the lowest dose tested in the study. The rationale and justification of selection of this study as the basis for the RfC would be improved if EPA provided a discussion on the possible anatomical and metabolic similarities or differences between mice and humans related to this compound and effect.

David Pyatt, Ph.D.

It appears from the 2 year rodent inhalation study, that alveoli bronchiolization was an appropriate endpoint to use. However, the relevance of this endpoint to humans was not discussed nor was the toxicological significance of this endpoint presented. What other chemicals and what exposure conditions have been shown to result in this endpoint in humans? Further, there is good evidence that this is highly species specific, which could be important in determining the likelihood that this

would occur in humans following nitrobenzene exposure. The species specificity could also impact the interpretation of the animal data. Therefore, the species specificity as well as the potential relevance to humans should be fully addressed in the document

Lorenz Rhomberg, Ph.D.

There is a good discussion of the reasoning for choosing bronchiolization as the critical endpoint. It is hard, however, to completely dismiss the methemoglobinemia that is concordant with effects seen in the oral studies.

Question 2 - If bronchiolization of the alveoli is the most scientifically justifiable endpoint on which to base the RfC, is the LOAEL-to-NOAEL approach the best method for deriving the RfC?

Mark Miller, Ph.D., Chair

This reviewer lacks sufficient expertise in modeling analysis to comment on this, although as noted above, this reviewer felt bronchiolization was not the best endpoint for these analyses.

Bruce Allen

Absolutely not. While it is true that the responses observed at the lowest dose (87 to 92%) are much greater than the typical choices for BMRs (typically around 10% above background), this is not a case where all the positive doses gave 100% response, which is the only case where a BMDL can not be defined. Yes, there will be some extrapolation downward to 10%, but calculating lower bounds, BMDLs, will reflect that uncertainty. Even if a straight line was fit between the responses at the 0 ppm dose and the 5 ppm dose (accounting for uncertainty in the observed response rates), this would be much better than some generic factor of 10 that gets tacked on to LOAELs regardless of the response that defines the LOAEL.

Using the NOAEL/LOAEL approach for this one endpoint makes the results for it not comparable to the results for any of the other endpoints considered in the Toxicological Review.

Rudolph Jaeger, Ph.D., DABT

Yes, this approach is reasonable based on the data and Agency policy.

Martin Philbert, Ph.D.

No - I have no idea how one would compare bronchilization to other endpoints used in this analysis.

Richard Pleus, Ph.D.

Based on the information provided in the Toxicological Review, "bronchiolization" is a consistent outcome for male and female mice at all doses tested. The document does not provide information on similar variables in rats exposed to nitrobenzene via inhalation. It would be preferable to have a study that included a dose that demonstrated a NOAEL for bronchiolization; however, I am not aware of such a study and the external peer review did not identify any. Several additional points should be included in the document. First, the loss of animals in the experiment and the impact on EPA's assessment should be discussed. Second, "bronchiolization" should be better defined. The CIIT 1993 original document provides adequate information on this in Appendix R. Third, since this affect is only seen in mice, the document should discuss- any anatomical or species difference regarding bronchiolization between the human and the mouse.

David Pyatt, Ph.D.

The only comment would be that the dose response for this endpoint was pretty quirky (plateau). As such, the extrapolation might involve even more uncertainty than usual.

Lorenz Rhomberg, Ph.D.

Use of the LOAEL is not ideal, but given the odd dose-response, it is hard to see how the BMD could be used. The odd dose-response—a substantial but similar response across doses—warrants some discussion, however. Moreover, some discussion is needed about treating bronchiolization as a portal-of-entry endpoint. This seems reasonable but there are cases (e.g., naphthalene) of respiratory epithelial effects that appear to be systemic, even upon inhalation exposure.

Question 3 - A database UF of 1 was applied in deriving the RfC because the database includes a two-year (lifetime) chronic inhalation study with an interim (15-month sacrifice), two-generation reproductive and developmental inhalation studies, a subchronic (10-week) inhalation neurotoxicity study, and two 90-day inhalation studies. Is the application of a database UF of 1 scientifically defensible and transparently and objectively described given the available data for nitrobenzene?

Mark Miller, Ph.D., Chair

Given the data available for nitrobenzene in the literature, the use of 1 is probably appropriate. However, the authors should provide some additional discussion and justification for their assertion that the data in the literature is complete - these analyses are still based on a few limited studies. While the uncertainty factors will most likely adjust for some of the age-related effects and the lack of data available for lower doses, some additional discussion and, as noted above, more thorough justification of the use of the uncertainty factors is needed.

Bruce Allen

I believe that this UF is reasonable. My only comments relate to the inter-connectedness of the database UF for inhalation and for oral routes of exposure. To the extent that these can be integrated, I think the UF choices will be improved.

Rudolph Jaeger, Ph.D., DABT

Yes, this valid based on the length of the study and Agency policy for such data sets is justified.

Martin Philbert, Ph.D.

No basis for judgment.

Richard Pleus, Ph.D.

The Toxicological Review discusses the rationale for a UF of 1. On page 126, it is stated that "The inhalation database is considered complete." The document should include discussion of the confidence in choosing a UF of 1.

David Pyatt, Ph.D.

This seems like a reasonable conclusion based on the available data.

Lorenz Rhomberg, Ph.D.

The factor of 1 seems appropriate and is appropriately justified.

Carcinogenicity of Nitrobenzene

Question 1 - Under EPA's 2005 *Guidelines for Carcinogen Risk Assessment* (www.epa.gov/iris/backgr-d.htm), nitrobenzene is classified as *likely to be carcinogenic to humans*. Have the rationale and justification for this designation been transparently and objectively described? Do the available data support the conclusion that nitrobenzene is a likely human carcinogen? If the weight of the evidence supports the descriptor *likely to be carcinogenic to humans*, is it appropriate to describe nitrobenzene as a case that lies on the low end of the range of this descriptor?

Mark Miller, Ph.D., Chair

While it is likely that the classification of nitrobenzene as *"likely to be carcinogenic to humans"* is appropriate, the authors need to provide further discussion and justification for this assessment. There should be a description of the criteria used to arrive at this classification and the reason why this classification was used and not *"suggestive evidence of carcinogenic potential"*. Given the uncertainties regarding the mode of action of nitrobenzene, the decision to classify nitrobenzene at the *"low end of the range"* is questionable given the results of the 2 year rodent bioassay and the fact that tumor formation was detected at multiple sites in two different species and exhibited a clear dose response for some tumor responses. It is not clear what this descriptor, *"low end of the range"*, actually means since this appears to require a quantitative judgment as to the potency of the compound, and no discussion as to how this descriptor was arrived at is provided.

Bruce Allen

In the section that describes the characterization of the human carcinogenic potential (Section 6.1.6), there is a brief mention of the basis for the *likely to be carcinogenic to humans* descriptor. All it really says is that the 2-year bioassay with 2 species resulted in tumors at multiple sites; that the genotoxicity tests suggest that nitrobenzene is at most a weakly genotoxic compound; that there are no data about species differences in metabolism; and that there is no reason to assume that a cancer mode of action exists in animals that might not be relevant to humans (presumably referring to the earlier dismissal of the α_{2u}-globulin hypothesis for kidney tumors and of the follicular cell activation hypothesis for thyroid tumors). If that, in conjunction with the lack of human cancer data, is all it takes to get that descriptor, then yes, the rationale and justification has been transparently and objectively described.

However, the statement that nitrobenzene lies at the low end of the range of this descriptor implies (to this reviewer) that there are gradations of the weight of the evidence that are not being presented here. I have no idea what being on the low end of the range means, because there is no discussion of the different weights that are given to different pieces of information. Some "scale" or contextual clues are needed to be able understand what this statement means. Examples and comparisons to other compounds that are or are not at the low end of the range would be useful, indicating what mix of data or evidentiary weights go into that determination.

Rudolph Jaeger, Ph.D., DABT

Based on this reviewer's knowledge and the description given in the document, the rationale and the justification for this designation have been described. The basis appears to be objective and if transparency implies that there is no attempt to hide facts in evidence or to withhold information not yet in evidence, then transparency can be a term that is used to describe the effort.

Within this reviewer's knowledge of the EPA guidelines for this designation, the data do support this conclusion (Yes).

It is not known to this reviewer, based on the data given in the document, how this judgment and the associated ranking fit with other substances that might be placed in this category. See the discussion in 6.1.6 for a discussion that appears without an effort to place the result on a continuum of other compounds and categories/potencies.

Martin Philbert, Ph.D.

(No answer was provided)

Richard Pleus, Ph.D.

The evidence cited for EPA's conclusion is that "Nitrobenzene has caused neoplasia in a 2-year chronic inhalation study (CIIT, 1993; published as Cattley et al., 1994) in a dose-related fashion in the livers of male F344 rats and the lungs of male B6C3F1 mice. Increased incidences of neoplasia with statistically significant, positive dose trends were also observed as kidney and thyroid adenomas and carcinomas in male F344 rats, endometrial polyps in female F344 rats, hepatocellular adenomas and carcinomas in male CD rats, and kidney neoplasia in male B6C3F1 mice." This appears to be the only animal study available that assessed ; I am not aware of other studies at this time.

The Toxicological Review should provide clear scientific support for EPA's assessment and conclusion. For example, the statement that "nitrobenzene is likely to be carcinogenic to humans" is somewhat confusing when the document also states that "No studies exist on the carcinogenicity of nitrobenzene in humans" and "The mode of carcinogenic action of nitrobenzene cannot be classified as either genotoxic or nongenotoxic. Nitrobenzene was inactive in all bacterial mutagenicity assays and gave equivocal results in both in vivo and in vitro mammalian assay systems. There is limited experimental evidence that nitrobenzene can form DNA adducts or cause oxidative DNA damage, but no evidence was seen that would support a threshold mechanism such as cytotoxicity followed by regenerative hyperplasia, the scientific assessment is not clear." It might be helpful to point out that the carcinogenicity descriptors reflect a policy decision, not necessarily strict scientific definitions.

The Toxicological Review would also benefit by describing in more detail the basis for the selected cancer descriptor "Likely to be carcinogenic to humans" as opposed to alternative choices such as "Suggestive evidence of carcinogenic potential," which are outlined in EPA's *Guidelines for Carcinogen Risk Assessment* (U.S. EPA, 2005).

For perspective, the International Agency for Research on Cancer (IARC) reached a somewhat different classification of nitrobenzene using what appears to be essentially the same data set: *e.g.*, IARC classified nitrobenzene as "possibly carcinogenic to humans" (Group 2B) (IARC 1989). The distinction between "likely" versus "possible" warrants greater consideration in the document.

David Pyatt, Ph.D.

Based on the selection criteria for the various classifications, there doesn't appear to be a better descriptor. However, it is not known how to interpret the description "on the low end of the range".

Lorenz Rhomberg, Ph.D.

Under the new Guidelines, it is hard to argue with any classification, since the criteria are all loose. It seems as though most compounds with any carcinogenic endpoint get called "likely" and I would concur that this one should be at the low end. One wishes for better ability to distinguish this from compounds with more compelling and consistent animal data.

For nitrobenzene, there is considerable reason to entertain nonlinear modes of action for most of the hypothesized endpoints, a topic that warrants a fuller discussion.

Question 2 - The two-year inhalation cancer bioassay (CIIT, 1993; published as Cattley et al., 1994) was used for development of an inhalation unit risk (IUR). Is this study the most appropriate selection for the principal study? Has the rationale for this choice been transparently and objectively described?

Mark Miller, Ph.D., Chair

Based on the available data in the literature, these two studies would be the most appropriate for the development of the inhalation unit risk (IUR). The rationale and justification for the use of these studies was well documented and objectively described.

Bruce Allen

I believe that the CIIT (1993) bioassay is the most appropriate choice in this case, because it appears to be the only 2-year bioassay available. However, as in my comments on the RfD and RfC calculations, I would in general prefer to see a bullet list of criteria used to select a principal study and why the selected study is superior to alternatives (a short list, since there appear to be no alternatives).

Rudolph Jaeger, Ph.D., DABT

Based on the data presented, the opinion of the reviewer is yes.

Yes, this would appear to be the case.

Martin Philbert, Ph.D.

(No answer was provided)

Richard Pleus, Ph.D.

The two-year inhalation cancer bioassay (CIIT, 1993; published as Cattley et al., 1994) was used for development of an inhalation unit risk (IUR). This is the only relevant study identified by the group of external reviewers. I found it useful to review the detail in the original study; providing additional information from this study in the Toxicological review, e.g., information on experimental design, animal loss, etc., would allow for a more transparent understanding by the reader.

David Pyatt, Ph.D.

This is really the only viable choice and yes, the justification was adequately described.

Lorenz Rhomberg, Ph.D.

This is the only viable choice. The lack of consistency with the oral studies, even for "systemic" endpoints, needs to be more fully discussed.

Question 3 - Data on hepatocellular tumors in F344 rats were used to estimate the IUR. Are the reasons for basing the quantitative assessment on hepatocellular tumors in male F344 rats scientifically justified and transparently described? For calculating the IUR, adenomas and carcinomas were combined. Has EPAs justification for this approach been objectively and transparently presented? Is combining adenomas and carcinomas the most scientifically justifiable approach for these tumors? Please suggest any other scientifically justifiable approaches for calculating the IUR.

Mark Miller, Ph.D., Chair

The choice to utilize the hepatocellular tumors in F344 rats was objectively described and for the most part, good scientific justification was provided. An issue could be raised regarding the choice of assessing liver tumor formation as the endpoint instead of lung tumors. Liver tumors were observed in both species (mouse and rat) and in both strains of rats studied. In addition, a clear dose response was observed with the F344 rats when assessing liver tumor incidence. For the inhalation route, the incidence of lung tumors at the point of entry exhibits a greater incidence than that for liver tumors (Tables 4-19 and 4-41, pages 48 and 107, respectively), while rats appear to be resistant to lung tumor formation. However, given that the background lung tumor incidence in the mice is relatively high and the lung tumor incidence was not tested at the 1 ppm concentration used in the rat, the authors appear to have selected the most representative and relative endpoint and provided clear scientific justification for their decision.

Combining adenoma and carcinoma incidence is appropriate and well justified in this case. While not all adenomas will necessarily progress to carcinomas, it is logical and appropriate to assume the

adenoma burden (number of adenomas present) is a direct reflection and will correlate with the likelihood of developing more carcinomas. The authors provided not only the combined analyses of adenomas and carcinomas, but also included the incidences of the two lesion types individually. This reviewer thought this to be very appropriate and provides further justification and enhanced transparency to the document.

Bruce Allen

The rationale for not selecting certain tumors other than the male rat hepatocellular tumors is fairly well described. There were reasons for eliminating from consideration some tumors; for example, there was a lack of data for mammary adenocarcinomas in female mice, certain other observations were not consistently elevated over controls or historical rates; and for some other tumors there was a lack of consistency across sex and strain.

But that left for consideration several endpoints in mice and rats. What is less clear from the documentation is why the male rat hepatocellular tumors were selected from among those remaining. If it was because they gave the greatest IUR, that should be explicitly stated as a criterion for selection early (not after the results were obtained) and can be justified as being the most health protective from among the candidates for which dose-response modeling was considered appropriate.

With respect to the combination of adenomas and carcinomas, the document is a bit inconsistent. I take at face value the statement that the decision to model adenomas combined with carcinomas was based on the assumption that they represent stages along a continuum of carcinogenic effects resulting from the same mechanism, as recommended by EPA (2005). But why then do Tables 5-7, 5-8, and 5-9 show adenomas and carcinomas modeled separately? Those tables should be reduced to just showing results for adenomas combined with carcinomas. And why, when considering whether to model male rat kidney tubular cell tumors, is the fact that only one carcinoma (as opposed to 6 adenomas and carcinomas combined) called out as being somehow important (top of p. 129)? EPA should adopt one position on adenomas and carcinomas and stick with it for all aspects of every assessment.

In the meeting of the external peer reviewers, Dr. Miller suggested that there might be some reasons for presenting data on adenomas and carcinomas separately, because there may be some decisions affected by differences in the incidences or counts of those tumors. If that is the case, then the document should state what those decisions might be and how the adenoma/carcinoma mix would affect them. Only in that instance might it be appropriate to model and distinguish between the two tumor types, and if that were the case, then it would be an open question whether the combination of adenomas and carcinomas was scientifically justified (in fact, the modeling and discussion of them separately would presumably be addressing that question of scientific justification).

The model predictions for the BMDs differed across endpoints because different BMRs were selected for different endpoints, based on selecting BMRs “consistent with the lowest tumor incidences observed” (p. 130). This is not a good approach; it leads to numerous potential

problems. Does that mean non-zero incidence? Because several positive doses in this study (and other studies) have 0 observed incidence. If it is only non-zero incidence, then what about dose-response data like that for nitrobenzene, the kidney tubular adenomas and carcinomas in male F344 rats, where the only non-zero incidence is in the high dose and it is at about 13%. Why would the restriction be to BMRs around 13% when there are two doses that give point estimates at or close to zero? And what if that high-dose response had been around 50%; would that mean that the BMR would have to also be around 50%? And what if an experiment had only 10 animals per group, so that the lowest non-zero incidence that could be observed was 10%? Compared to a study with 100 animals per group, the smaller study would theoretically be constrained to have a much higher BMR than could be derived for the larger study, if the reasoning applied here was carried forward, leading to larger BMDs. This is not the type of behavior ("rewarding" of poorer studies) that we want to be associated with application of the BMD approach.

The true driving factor for BMR selection should be that the resulting BMD estimate not be too far below the range of the experimental *doses*; the range of the experimental *responses* should not be a determinant. But even here, there need not be slavish adherence to this rule of thumb. Much more important is that there be some consistency for the estimates across data sets. This is achieved by picking a BMR level that – *on average for the type of study being considered* – corresponds to a response level likely to be within the range of the experimental doses. Ad hoc changes based on the particular responses in the data set being analyzed detract from the desired consistency.

This is particularly important because the proposed decision to get BMDs corresponding to different BMRs is apparently made even before the method for low-dose extrapolation has been determined. While the choice of the BMR may not make very much difference when the low-dose-linear approach is to be applied (i.e., when slope factors are to be estimated), it could make a very big difference if a non-linear, point-of-departure approach were to be adopted. In the latter case, the BMDs are replacing the poorly conceived NOAELs and thus should be as consistent as possible across endpoints and indeed across compounds.

The preceding discussion leads to the question about alternatives to the IUR calculation. The case of nitrobenzene is a prime example of how information on mode of action (e.g., genotoxicity) is inadequately considered in cancer risk assessments. In many places throughout the Toxicological Review, statements are made that the "available evidence suggests that nitrobenzene is not, or is at most weakly, mutagenic" (p. 131 for this particular quote). Moreover, that same paragraph cites Section 4.6.3 as saying that the data are not complete enough to substantiate that possible DNA damage is responsible, i.e., that there are not strong indicators that DNA damage is linked to a MOA of nitrobenzene-induced tumors. Indeed, Section 4.6.3 again states that nitrobenzene is at most weakly genotoxic. And, the only study cited in detail (Ohkuma and Kawanishi, 1999, which, by the way, probably should be discussed in the section of chapter 4 that presents the genotoxicity data), describes a complex set of experimental conditions under which DNA damage was induced in calf thymus, involving the presence of Cu^{2+} (not other metal ions) and apparently not ameliorated by superoxide or free radical scavengers. This suggests to me that the conditions under which nitrobenzene would induce DNA damage in vivo are unlikely to occur and are apparently not the consequences of superoxide formation that has been cited elsewhere in the document (e.g., Section

6.1.6) as occurring because of the 6-step metabolism of nitrobenzene and potentially responsible for genotoxic insult.

Thus, it is not at all clear to me why the predominately negative genotoxicity evidence is ignored when it comes time to derive a cancer risk estimate. What would ever be sufficient evidence that a cancer risk estimate would not be based on calculation of a unit risk? Until and unless EPA can come up with some reasonable criteria that appear to make sense when applied to a variety of compounds, the entire exercise of "weight-of evidence" that takes into account genotoxicity data appears pointless. In the case of nitrobenzene specifically, a start could be made by doing dose-response analyses of the genotoxicity data to see for what dosing regimes (and under what conditions) a genotoxic response is changed from background; comparing those genotoxicity-inducing dose levels to the ones that induced tumors; and seeing if there is any concordance.

My conclusion at this point is that no IUR should be calculated and that a point-of-departure approach should be used for the nitrobenzene inhalation cancer risk assessment. This requires that consistent points of departure (10% extra risk BMRs) be used across all cancer endpoints modeled. My first impression is that an uncertainty factor on the order of 30 to 100 would be used in conjunction with a suitably chosen point of departure.

Rudolph Jaeger, Ph.D., DABT

The hepatocellular tumors in F344 male rats were stated as having the best dose response relationship. Thus, this data set is the most amenable to analysis.

The reasons for this selection are stated without further justification and thus, in the opinion of this reviewer (unless we missed it), the answer is NO, the reasons for this choice are not objectively and transparently presented in the fourth paragraph on page 129. Perhaps this is given elsewhere within the document or is a policy under EPA guidelines.

No other methodology of suitable robustness is known to this reviewer.

Martin Philbert, Ph.D.

(No answer was provided)

Richard Pleus, Ph.D.

A significant issue is that data are available from only one animal study. If more data were available, one would be able to compare and contrast the different studies and determine if the sites of adenomas and carcinomas are consistent. The document should provide more of the raw data (e.g., page 417, Appendix R is very useful). A review of the pathology report provides more information on the histology of hepatocellular tumors. However, my review of EPA's document did not provide me with an adequate understanding of whether it is scientifically feasible to combine

adenomas and carcinomas. This information, if included in the EPA document, would provide greater scientific transparency in EPA's assessment.

David Pyatt, Ph.D.

This approach seems reasonable and was adequately described.

Lorenz Rhomberg, Ph.D.

The use of hepatocellular tumors and their combination seems appropriate.

Question 4 - The IUR was calculated from hepatocellular tumors in male F344 rats. The recommended upper bound estimate on human extra cancer risk from continuous lifetime exposure to nitrobenzene was calculated to be 3 x 10^{-5} ($\mu g/m^3$)$^{-1}$. Is it scientifically defensible to base the IUR on liver tumors alone? Have the rationale and justification for this analysis been transparently and objectively described? Is it more appropriate to calculate the IUR using combined tumor incidence of liver, thyroid, and kidney tumors in male F344 rats as done in the alternate derivation of the IUR in the Appendix? If summing of tumors is scientifically justified, is the method used to sum the tumors supported by the science and the data? If not, what alternative method should be used?

Mark Miller, Ph.D., Chair

The IUR of 3 x 10^{-5} ug/m^3 was arrived at regardless of whether the analysis was done with liver tumors alone or a combination of liver, thyroid, and kidney tumors. In this reviewer's opinion, the best approach when multiple tumor sites are involved should be to use the endpoint and value for the most sensitive tumor site. In this case, given the data available and the results of the IUR calculations with liver tumors alone *vs.* combined data for liver, kidney and thyroid, the liver alone can be justified scientifically as the appropriate tumor site to use in this assessment. The rationale and justification for this approach were well written and clearly justified in the report. Including the IUR for the combined assessment in an appendix provided strong justification for this approach, and this appendix should be included in the final report. However, it is not entirely clear that the kidney and thyroid tumors are relevant for extrapolation of risk to humans, particularly given historical concerns over thyroid tumor incidences in rodent models. The justification for inclusion of these two cancer endpoints was superficial and should be discussed in much greater detail.

One concern this reviewer had is that uncertainty factors were not utilized to account for potential age related susceptibilities to cancer induction by nitrobenzene during early life exposures. The EPA has developed guidelines for *"Assessing Susceptibility from Early-Life Exposure to Carcinogens"*, but this analysis has made no attempt to account for the enhanced susceptibility of fetal or early juvenile exposures to nitrobenzene. Early exposure to single doses or chronic exposure starting at very early ages could increase susceptibility to induction of tumors, as has been shown for a number of chemicals (Rice, 1979;Anderson *et al.*, 2000;Miller *et al.*, 1996, 2004;Hattis

et al., 2004). The authors should consider the inclusion of age related factors in their assessment of cancer susceptibility.

Bruce Allen

As discussed above, I do not think that the calculation of an IUR is scientifically justified until and unless EPA can present a scheme for objectively considering and analyzing genotoxicity data and factoring that into the cancer risk estimation procedure.

But, in addition, the calculation of unit risks based on summing unit risks from two separate tumor types is ill-advised. The difficulties with the ad-hoc method for combining the unit risks are based on the following observations. It is assumed that the estimates of the BMC are normally distributed around the maximum likelihood estimate with, for example, the 95% LCL risk being equal to the MLE (mean) minus 1.645 times the standard error. The estimation of the BMCL in software such as BMDS does *not* make such simplistic assumptions (e.g., that the BMC estimates are normally distributed about an MLE); it uses a profile likelihood procedure that identifies the likelihood of various BMC values and selects the smallest value that gives a likelihood that could not be rejected with 95% confidence. The bounds on the BMC are not symmetric, providing the first clue that the simple normal assumption or approximation is not appropriate.

In fact, a likelihood profile method can be defined that finds the BMC for any combination of tumors, making the important assumption (perhaps needing some separate justification or discussion) that the tumors arise independently. The maximum likelihood estimates will be directly based on the maximum likelihood estimates for each individual tumor, but the lower bound estimate will be a more complicated expression of the likelihood that needs to be optimized for various choices of potential BMDLs and which accounts for the fact that we are looking for a dose that , for all the tumors combined, yields the correct reduction in the log-likelihood and still can give, for the parameters corresponding to the maximized likelihood, a response equal to the BMR.

In any case, the contention starting the second paragraph of p. 58 of the appendix – that a "statistically valid upper bound on aggregate potency" has been used – is wrong. While it is appropriate to add MLE estimates to get a combined MLE, the method for IUR derivation is not valid. The implication that this is statistically reasonable needs to be removed from the document.

Aside from the overarching incorrectness of the approach, there are a number of smaller issues that should be addressed in any revision.

- In the main text, the BMRs varied across endpoint (either 0.05 or 0.10 extra risk), and this was a concern as discussed above. In the appendix section on the aggregation of the male rat tumor site results (Section B-3.4), the low-response BMRs were again varied, but this time by an order of magnitude, being either 10^{-6} or 10^{-5} extra risk. There is no justification for choosing different BMRs here, even less so than when 5% and 10% extra risks were selected depending on endpoint.

- A statement is made about numerical stability as a basis for selecting the low-response BMRs of 10^{-5} or 10^{-6} extra risk (last line of p. 58 of the appendix). Nowhere is an evaluation of numerical stability presented or discussed. In fact, the entire footnote associated with this statement is wrong, since it is manifestly true that one can numerically combine the predictions of the separate tumors even when they have different dose-response relationships. A statistically valid approach must be used, so part of the problem revealed by the wrong-headedness of the footnote is undoubtedly that the method proposed in this Toxicological Review is not valid.

- The penultimate paragraph of the appendix starts by stating that "extrapolation of data from animals to estimate potential risks to human populations has generated some uncertainty in the results." (p. 61 of the appendix). Such extrapolation does not generate uncertainty in the results; results are results and the only uncertainties associated with them are those associated with measurement, recording, or experimental design concerns. Rather, it is the extrapolation of results that generates uncertainties. The second sentence of the paragraph in question implies that model and parameter uncertainties are associated with that animal-to-human extrapolation. But, the remainder of that paragraph and the following paragraph relate to the application of the multistage model *to the rat data*, and the estimation of parameters (and bounds on those parameters) for that model (with one short exception where the question of which species to use to extrapolate to humans is mentioned). This entire discussion of "uncertainty" fails to reveal any understanding of its various components and the importance of each one to the overall uncertainty associated with this cancer risk assessment. This is a significant difficulty in the sense that a thorough understanding of, and careful communication of, the uncertainties associated with any risk assessment is a key issue when it comes time to apply the results of such an assessment.

Rudolph Jaeger, Ph.D., DABT

The calculated IUR is based on total hepatocellular tumors in male F344 rats. In order to support this estimate, additional data elements were considered within the calculations performed and presented in the appendix. Based on their summation, the additional tumor data did NOT greatly increase the risk estimate (when rounded to one decimal place) and thus, the majority of the risk estimate appears to be correctly computed from the liver data alone.

Martin Philbert, Ph.D.

(No answer was provided)

Richard Pleus, Ph.D.

See answer to Question 3 above.

David Pyatt, Ph.D.

If the different tumors potentially have different mechanisms and/or differing relevance with respect to human exposures, then combining them doesn't make much sense. In this case, the risk estimate seems to be driven primarily by the liver tumor data, so it may not matter much. In any case, the rationale and/or justification were inadequate.

Lorenz Rhomberg, Ph.D.

What is less clear is whether this along with all the other endpoints (also kidney and thyroid) should be used. Doing so means accepting the likely relevance of all types for humans, and given the lack of consistency in these effects across rodent species and sexes, this seems marginal.

If the risks from three tumor types are to be combined, there are some questions about the method for doing so. First, if there are individual animal data, it is possible to determine whether the different tumors are indeed independent in their appearance in the bioassay.

Second, the method for generating a joint uncertainty can be questioned. One uses maximum likelihood methods so as not to have to make assumptions about the distributions of model parameter values, and so the assumption of normality can be questioned, especially in view of the non-normal (indeed, usually bimodal) distribution that is often found.

Third, the method supposes linearity of the BMD (and not just the BMDL) at low risk levels, and this is probably not true, especially for the kidney tumors. This makes the answer dependent on the particular risk value used as a benchmark. (In practice, the kidney contribution is minor in this case, but the principle is questionable.)

MISCELLANEOUS COMMENTS

Mark Miller, Ph.D., Chair

References

1. Anderson, L. M., Diwan, B. A., Fear, N. T., and Roman, E. (2000). Critical windows of exposure for children's health: cancer in human epidemiological studies and neoplasms in experimental animal models. Environ. Health Perspect. 108 Suppl 3, 573-594.

2. Miller, M. S. (2004). Transplacental lung carcinogenesis: molecular mechanisms and pathogenesis. Toxicol.Appl.Pharmacol. 198, 95-110.

3. Miller, M. S., Juchau, M. R., Guengerich, F. P., Nebert, D. W., and Raucy, J. L. (1996). Drug metabolic enzymes in developmental toxicology. Fundamental & Applied Toxicology 34, 165-175.

4. Hattis, D., Goble, R., Russ, A., Chu, M., and Ericson, J. (2004). Age-related differences in susceptibility to carcinogenesis: a quantitative analysis of empirical animal bioassay data. Environ. Health Perspect. 112, 1152-1158.

5. Rice, J. M. (1979). Perinatal period and pregnancy: intervals of high risk for chemical carcinogens. Environ. Health Perspect. 29, 23-27.

Rudolph Jaeger, Ph.D., DABT

Reviewer's Concluding Remarks

The Peer Review Panel was in agreement that the document was valid and objective in most respects. Specific comments about style, organization and format were offered. Varying quantitive considerations were discussed with some being provided by Reviewer's in greater detail than others. This Reviewer (RJJ) found no point of disagreement with the Review Panel in general and accepts most other comments as either supplementary to those given here or points not noticed or remarked upon in the current review.

Lorenz Rhomberg, Ph.D.

General Comments:

This is a good review of the relevant technical literature. The choices made during the risk assessment process are described and, for the most part, the key particulars of calculations are clearly shown. As noted below, one could hope for more synthesis of the toxicological literature (considering possible generalizations and interpretations rather than just documenting study facts), more narrative description of hypothesized modes of action, more discussion of how toxicity may

depend on metabolism (especially bacterial versus hepatic reductive metabolic activation) and on duration of exposure. Some of the risk assessment choices could be more thoroughly justified. It would be good to see more attention to documenting the consequences of the choices made and a deeper examination of how the default approaches used compare to conceivable chemical-specific approaches. These questions can be discussed at the review meeting.

Richard Pleus, Ph.D.

The following outlines my final responses to the Charge Questions submitted to me on March 27, 2007. I have reviewed the document *Toxicological Review of Nitrobenzene: In Support of Summary Information* on the Integrated Risk Information System (IRIS) (U.S. Environmental Protection Agency Washington, DC, EPA/635/R-07/001) and its associated Appendix B, attended the external peer review meeting (held at the American Geophysical Union, 2000 Florida Avenue, Washington, DC) on May 15, 2007, and reviewed a key primary study (CIIT, 1993; published as Cattley et al., 1994). My comments rely upon these sources and my experience in toxicology and risk assessment. I've summarized my overall comments first, followed by more detailed responses in the table.

As noted in my initial responses, obtaining key studies as a component of this review would be useful in determining whether their interpretation is appropriate for deriving noncancer and cancer criteria. On the day of our external peer review meeting, I was provided a copy of CIIT, 1993[2], which I have subsequently reviewed. However, the document was not complete— missing sections included Appendices A-F, H, J, L, T-W, and Y.

Overall, I found the comments of external peer reviewers at the meeting to be scientifically sound and logical and would therefore recommend that US EPA use the comments expressed in the meeting to strengthen *Toxicological Review of Nitrobenzene: In Support of Summary Information*.

At that meeting, it was stated this toxicological review is the first to implement the "new" cancer guidelines (*Guidelines for Carcinogen Risk Assessment* (U.S. EPA, 2005)). I have a number of concerns about the outcomes of the following the guidelines and the options presented by the guidelines. These are discussed below.

Martin Philbert, Ph.D.

Literature Cited

Henegan, J.B. (1984) Physiology of the alimentary tract. In, *The Germ-Free Animal in Biomedical Research*. ME Coates and BE Gustafsson (Eds.) Laboratory Animals, Ltd., London, UK., pp 169-191.

[2] CIIT 1993. A chronic inhalation toxicity study of nitrobenxene in B6C3F1 mice and Fischer 344 Rats and Sprague-Dawley (CD) rats, two vols. 1993.

Vasquez GB. Reddy G. Gilliland GL. Stevens WJ. Dinitrobenzene induces methemoglobin formation from deoxyhemoglobin in vitro. [In Vitro. Journal Article. Research Support, U.S. Gov't, Non-P.H.S.] Chemico-Biological Interactions. 96(2):157-71, 1995