General Comments

R1: My overall impression is that the Development Support Document (DSD) has made good use of the available science on 1,3-butadiene (BD), and has appropriately used accepted risk assessment procedures to analyze that science and to establish reasonable reference values. Given the complexities of risk assessment and the need to exercise expert judgment in a number of places in the risk assessment process, it is not surprising that different regulatory agencies will arrive at different values for the same compound. Therefore, it is important that the assumptions made and basis for each judgment be clearly presented. The DSD does this in a transparent way.

I applaud the use of benchmark dose methodology to estimate the point-of-departure for the assessments. I agree with the authors of the DSD that benchmark dose uses much more of the available information and provides a more reliable POD than does the NOAEL, which can change markedly from study to study simply as an artifact of dose selection. The potential downside to the use of benchmark dose methodology is that there are still relatively few risk assessments using it. As a consequence, the opinions of the risk assessment community are still coalescing around the details, such as using a 5% or 10% effect level, a central tendency vs. a lower confidence limit, and for continuous variables such as body weight changes, how much of a change constitutes an adverse effect. Until there is more consensus on how to handle these issues, any decision will be open to criticism. The DSD handles this potential pitfall in the best way, by clearly providing the basis for each decision in the process.

I was pleased to see the extensive statistical re-analysis of critical studies. I found the re-analysis to be rigorous, detailed and appropriate. I found the critical analysis of the recent Cheng et al study, and its reanalysis by Sielken and Flores to be especially useful and believe that it makes an important contribution to BD risk assessment.

The reference values presented in the DSD were derived using practices that are in line with the ESL guidelines, and are consistent with procedures used by other state and federal regulatory agencies. There are a few points in the assessment in which different procedures could have been used which would still have been consistent with established risk assessment guidelines and which may have made the assessments even more data-driven. The example that comes to mind most readily is that, in the dosimetry adjustments for animal-to-human, no adjustment appears to have been made for different respiratory rates/ minute volumes for mice vs. humans. Default values for murine and human respiratory parameters exist and could have been used here to reduce some of the uncertainty in extrapolation that is cited in the text (e.g., p. 21, line 21, which lists "relatively high respiratory rates" as one of many possible factors that may predispose mice to have a more significant response to a given exposure. I don't see this as a fatal flaw, but it could have reduced uncertainty.

The only other minor weakness that I noted was the heavy reliance on other sources to provide critical review of the hazard data. Two of the major uncertainties in the assessments were the relative sensitivity of mice to BD, and the range of genetic variability in the human population. Some additional critical analysis of the data pertaining to these two points would improve the DSD, particularly given that new information has been published since the 2002 EPA assessment on which the DSD relies. It is probable that the new data are still insufficient to support any additional modification in the uncertainty factors for animal-to-human or intrahuman extrapolation, but it would be an exercise in completeness to be state that the new data were considered in the sections that assign values to each uncertainty factor.

R2: I found the document to be clearly written and well-organized. I was impressed with the fact that the most recent literature had been included.

R3: The document is comprehensive and clearly written. The pertinent studies and results are well presented. Dosimetrics for extrapolations and uncertainty factors are clearly described and justified for the calculation of acute and chronic Reference Values (ReVs) and Effects Screening Levels (ESLs).

Page 5, Table 1. It would be extremely useful to state that ESL = 0.3 x ReV.

Page 7, Figure 1. Consider adding TCEQ cancer ESL 28 ppb.

R4: Overall, I am very impressed with the TCEQ document on butadiene, and commend the authors for the high quality of the assessment. It is generally well written, although at some points the text is rather terse. The risk assessment appears to have been conducted in a manner that is consistent with the relevant guidelines. For the most part, the approach is well documented, reasonable, and defensible. There are a few points, however, where additional explanation is needed for alternatives chosen (or not chosen). I will identify these in my comments below. One of the major strengths of the document is the first chapter, which provides a very informative summary of the risk assessment in Table 1 and Figure 1.

Were procedures outlined in the ESL Guidelines followed by the TCEQ to perform butadiene's toxicity assessment? If references to accepted procedures in federal, state, or other appropriate guidance documents were made in the ESL Guidelines, were those accepted procedures followed?

R2: Yes, the guidelines were followed. An exception might be the use of a POD that is below the range of experimental data (page 27). With a policy that suggests use of a 5% benchmark response level for severe effects, I would think you might often run into the

problem of having a POD outside the experimental range.

R6: Except where noted below, the procedures in the ESL guidelines were followed by TCEQ, as were accepted procedures from other guidance documents. The approaches used were clearly described, except as noted below. Overall, the document is well-written and key decision points were for the most part well-documented. The document generally presented thorough and sophisticated analyses supporting the various ESLs.

Does the butadiene DSD clearly describe the approaches used by TCEQ to perform the toxicity assessment (i.e., hazard identification and dose-response assessment).

R2: I found the document to be both clear and concise.

Health-Based Acute ReV and ^{acute}ESL

The 1,3-Butadiene DSD describes the approaches used to derive the health-based acute ReV and acute ESL in Section 3.1 (page 8). Appendix 1 (page 66) describes the statistical analysis of data from the critical study. Appendix 2 (page 101) describes the benchmark concentration modeling conducted for estimating the point of departure. Please review the key decisions made by TCEQ in deriving these values. For each decision, please comment the consistency of the decision with TCEQ's ESL guidelines, the scientific appropriateness of the decision, and any additional approaches or additional information that would improve that decision.

The key decisions and some specific issues to consider are listed below. Please indicate if there are other issues specific to developing acute toxicity factors that have not been adequately addressed in the document.

The choice of Hackett et al. (1987b) as the critical study

R1: As a rule of thumb, I don't like to rely on measures of adult toxicity in developmental toxicity studies as the critical effect for risk assessment. However, in this case I believe it to be reasonable, and that it provides a conservative starting point for the assessment.

Maternal toxicity in a developmental toxicity study is not always optimal as the basis for reference values. Because the primary purpose of these studies is to evaluate developmental effects, the suite of measurements on the dam are usually limited to generic parameters such as changes in body weight gain, cageside clinical signs evaluated once or twice per day, and significant behavioral alterations (e.g., convulsions, ataxia) or mortality. It is uncommon to conduct histopathology or measure organ weights in these studies. For this reason, it is generally better to rely on an endpoint from a subacute or subchronic study in which, if decreased weight gain were used as the critical effect, the

systemic basis for the change would be known and an assessment of its relevance could be made. That said, because pregnancy in the rodent involves such a massive change in body mass over a short period, even small effects on homeostasis can lead to significant effects on weight gain. Furthermore, it can be argued that body weight gain is an integrator of effects on a number of organ systems (digestive, endocrine, nervous and others), such that it is a rigorous and sensitive measure of toxicity. Therefore, I believe it is acceptable to use the maternal body weight data as a critical effect, with the acknowledgement that it is not always right to do so.

The study itself appears to be of good quality, and is concordant with regulatory guidelines for developmental toxicity studies. Because these studies use approximately 20 animals per dose group, it may have more statistical power to distinguish a weight change effect than a standard subacute or subchronic study.

R2: This choice is logical according to TCEQ and EPA guidelines that indicate one should choose the most sensitive endpoint in the most sensitive species. It is not logical in the sense that humans (the species of interest) have not shown any adverse reproductive or developmental effects from exposure to 1,3-butadiene.

R4: I agree with the choice of the Hackett et al. (1987b) study as the critical study, but I do not agree completely with the approach used to evaluate the dose-response in that study.

The statistical re-analyses of the data from the Hackett (1987b) conducted by Green (2003) and Sielken et al. (Appendix 1 Statistical Analyses of Developmental Endpoints).

- **Are these re-analyses a more appropriate basis for risk assessment than the statistical analyses conducted by Hackett et al. (1987b)?**

R1: I believe that they are. The re-analysis corrects for a number of important limitations in the original analysis, the most significant being a failure to correct for multiple tests.

R2: I agree that the re-analyses were more appropriate than the statistical analyses conducted by the authors.

- **Should the analyses of Hackett et al. data in Appendix 1 adjust for litter size and percent of males in litter?**

R1: I believe that the adjustment based on litter size is appropriate, although it probably has only a small effect except when litter size is extremely small or large. Adjustments based on percent of males in each litter is unnecessary because it is customary to evaluate fetal weight for males and females separately, in addition to total litter

weight. Therefore, an effect that is more pronounced in one sex will still be identified without the need for covariate analysis.

R2: The litter size and the percent of males should be considered as co-variates.

- **Should the analyses of Hackett et al. data in Appendix 1 use mean data or individual data to determine the NOAEL?**

R1: I interpret this as a question about litter mean vs. individual fetal data as the basis for analysis. I believe that litter means are the more appropriate statistical unit. Because the fetuses are exposed only indirectly, via the dam, it is not possible to assume that their responses will vary independently of their littermates. Therefore, analysis using the litter as the statistical unit is most appropriate. This is the accepted consensus in the field of developmental toxicology.

R2: Individual data.

R4: With regard to the TCEQ question about adjusting the Hackett data for litter size and number of males per litter, and whether individual or grouped data should be used in the NOAEL analysis, I do not think there is any merit in pursuing the NOAEL calculations any further. Benchmark analysis should be pursued instead.

R5: These decision points in the DSD (*TERA*: referring to the charge questions up to this point) appear appropriate to this non-statistician, although some of the details of the analysis are beyond my area of expertise. However, I will note that adjustment for litter size is important, particularly when considering fetal weight. Similarly, % males/litter can be important, since males tend to be heavier than females. Mean data are generally used for this sort of analysis, but consideration of the individual data may be informative if there is evidence of a population distribution that is not normal or lognormal. For example, if there appears to be a subpopulation of pups that is particularly sensitive, this could be informative as to factors that increase sensitivity.

The choice of maternal extra-gestational weight gain, which occurs at a NOAEL of 40 ppm, as the critical effect.

- **Is this endpoint relevant for human risk assessment? If not, what would be a more appropriate critical effect.**

R1: See my comment above regarding the pros and cons of selecting this as the critical effect. The bottom line is that I believe it to be scientifically defensible.

Because body weight can be considered to be an integrator of function of a number of

different physiological systems, and because the basic function of these systems is conserved from rodents to humans, the case can be made that the endpoint is human-relevant. However, this comes with a few caveats. The one noted in a previous comment is that there are no data from a developmental toxicity study that provide greater clarity on the underlying cause leading to the decreased weight gain (e.g., some specific organ toxicity). Therefore, the human relevance can be assumed, but without additional support from the experiment. Another is that the effect may be entirely non-specific, i.e., decreased food consumption due to the animal's malaise, and may have a very steep dose-response curve that overstates the risk of a similar response at human-relevant exposure levels. Overall, my opinion is that the assumption of human relevance is more compelling than the opposite.

R2: No, it is not. To follow standard guidelines, one must choose the most sensitive endpoint, RELEVANT TO HUMANS, in the most sensitive species. In this case, there is no indication that this endpoint has anything to do with the hazard to humans of exposure to 1,3-butadiene. My suggestion would be to examine endpoints in mice (the most sensitive species) having to do with adverse effects on the lymphohematopoietic system**.**

R4: I believe the re-analyses by Green and Sielken are more appropriate than the initial analysis by Hackett et al.. The results of these analyses demonstrate that there are three endpoints with a NOAEL at 40 ppm and LOAEL at 200 ppm: reduced male fetal body weight, reduced maternal body weight gain, and reduced extragestational weight gain. However, only two of these endpoints were carried forward by TCEQ into the Benchmark analysis. It is argued (on p.14) that the lack of effect on gravid uterine weight at the LOAEL for the maternal effects indicates that they are not intrauterine in origin. This argument does support a conclusion that the maternal endpoints represent direct maternal toxicity and are not secondary to fetal effects. However, this argument does not support any conclusion regarding the converse, i.e., whether the fetal body weight effects are secondary to maternal toxicity. In particular, it does not support the TCEQ conclusion that a point of departure (POD) based on the maternal effects would prevent any fetal effects, particularly the observed fetal body weight effects, which appear to have a similar dose-response and could actually occur at somewhat lower concentrations than the maternal effects. Given the imprecision of NOAEL/LOAEL estimates, it is not possible to determine whether the fetal or maternal effects occur at lower concentrations in this study, but Benchmark analysis would allow the endpoint with the lowest BMCL to be identified. Therefore, since the NOAEL and LOAEL for the male fetal body weight reduction are the same as those for the two maternal endpoints, all three endpoints should be included in the Benchmark analysis.

R6: The endpoint is relevant.

The choice of point of departure based on a 5% reduction in extragestational weight gain and reduction in maternal weight gain (GD 11-16) (i.e., $BMCL_{05}$) (see Appendix 2 BMC Modeling for Acute ReV).

- **Was the output from the most appropriate model selected? Should these models be monotone?**

R1: I believe so. I especially like the fact that the report provides graphs of the dose-response curves, and that the authors used visual inspection of the curves as a basis for selection after it was determined that there were no differences between the best models in goodness of fit or AIC. This is a good use of statistics for support, not illumination. When there were no differences between the best models, the BMCs were averaged, which is according to guidance published on benchmark dose methodology.

If the data fit a monotone model well and there are no biologically plausible hypotheses to suggest that a more complicated model would be better, then Ockham's razor supports the use of the simplest acceptable model.

R2: I agree with the choice of model used.

R3: Pages 17-18, Sec. 3.1.4. A good statistical fit of a dose response model is not an adequate reason for selection of a model. The model must also be biologically plausible. The non-monotonic polynomial model based on 4 dose groups cannot be justified biologically and should be discarded.

Extragestational weight gain (EWG), as presented in Table 7, appears to be the appropriate sensitive effect for calculating the point of departure (POD) to derive an acute ReV. A 5% reduction in a continuous value, such as EWG, does not represent a 5% risk and hence is not akin to a no observed adverse effect level (NOAEL). For continuous data, in the absence of a specified value associated with an adverse biological effect, an extreme percentile, e.g., estimated 1st and/or 99th percentile, of the values in the controls may be used to define abnormal (not necessarily adverse) values. For example, with normally distributed data, values outside the range defined by the control mean ± 2.33 x (standard deviation) would be considered abnormal. From the dose response model, the dose associated with a specified excess risk, e.g., 5%, of animals with abnormal values of the selected continuous endpoint can be estimated (Gaylor and Slikker, 1990; Crump, 1995). Allen *et al.* (1994) show for reproductive and developmental effects that the lower 95% confidence limit for an excess risk of 5% is often similar to the NOAEL. For normally distributed data, an excess risk of 5% of animals with abnormal levels (beyond the 1st or 99th percentile) occurs at the dose where the mean value changes from baseline (controls) by an amount equal to (0.77 x standard deviation). Here, the Hill equation provides an excellent fit to the EWG data. BMDS Version 1.4.1 gives a lower 95% confidence limit on the estimate of 1,3-butadiene

concentration associated with an excess risk of 5% for animals with abnormally low EWG of $BMCL_{05}$ = 8.74 ppm for a 6-hour exposure.

R4: As mentioned above, the male fetal body weight reduction should be included in the Benchmark analysis. The analysis of this endpoint should use a nested model to account for litter effects, but I do not think it is necessary to adjust for number of males in a litter. For all endpoints, the use of individual data rather than grouped is preferred.

I can't think of any reason to question the human relevance of any of the three endpoints; that is, whether such effects could be produced in humans. However, it should be recognized that the maternal weight effects are of questionable adversity. More importantly in the case of setting an acute (1 hour) health guideline, these effects are cumulative and require several days of exposure to be observed.

The TCEQ approach for conducting the Benchmark analysis appears to be appropriate, except that I do not agree with the use of the Hill model. If I remember correctly, there was general agreement at the external peer review for the EPA Benchmark Dose Technical Guidance document that the Hill model should not be used for estimating the dose-response in the lower end of the data. The problem is that the Hill equation inherently gives too much weight to the higher doses, compromising the fit to the lower doses. There is a tendency to apply the Hill model when the data suggests saturation of an effect at higher doses, but the point is that the nature of the response at high doses is not of interest for Benchmark modeling. Only the behavior at low effect levels is of interest. Therefore, the fit with standard models, such as the polynomial or model, with the high dose(s) eliminated is preferred. In this case, the polynomial model result for the lower 3 doses is clearly the best model in terms of AIC for both maternal endpoints and that is the result that should be used.

The TCEQ use of a Benchmark risk of 5% is a reasonable policy judgment. I also concur with the use of the 95% lower bound as the POD. I am pleased that TCEQ recognizes there is no need for the use of an "Effect Level Extrapolation Factor" when a $BMCL_{05}$ is used as the POD. It is ironic that EPA finds it necessary to consider a 10% risk level less protective than a NOAEL, which can be associated with an increased risk as high as 30% or more, depending on the study design. The $BMDL_{05}$ is, on the average, more conservative than the traditional NOAEL for typical study designs and provides a more consistent characterization of risk across studies with different designs than the NOAEL.

R5: On p. 18 the summary of modeling results is not adequate for describing the choices made. For example, why would a 3-dose analysis be done if the models available can fit all four dose groups in the study? In fact, a quick check of the modeling results shows that several of the statements regarding the fitting of the BMDS continuous models to the data are erroneous. For example, it is stated that a power model fit to the data did not fit well. As shown in the graph and BMDS output file in Attachment A below, a power model

does indeed provide a good fit to all four data points. There is no need, either, to allow a nonmonotonic polynomial model; the monotonic polynomial (parameters constrained to be non-positive) also fits satisfactorily. In fact, I would give preference to either the power or monotone polynomial results over the Hill model results – the Hill model is really for dose-response relationships with a clear saturation of effect, and that is not the case here (especially with only four doses).

The document also shows a lack of understanding of the use of the AIC. In the first and second paragraphs of p. 18, there are comparisons of AICs from different models. But the comparison is erroneous first because the 3-dose and 4-dose results cannot be compared on the basis of the AIC; the AIC must be used to compare models fit to the same data, and the 3-dose and 4-dose analyses are not of the same data. Secondly, it is stated that differences in AIC were not great, when the values were 132 and 165 in one instance and 155 and 201 for the other endpoint. These in fact are *huge* differences; a rule of thumb is that a difference of 2 suggests some difference in model fit.

The model fitting results and interpretation are not adequate as they stand now. Only 4-dose analyses need be used and they should all employ monotone dose-response models. Selection of the best fit by AIC would be fine in that instance, but do consider all the models and do correctly evaluate the fit of the models to the data (all of which appear to be satisfactory from my preliminary analysis).

R6: Yes, the models should be monotonic, particularly since the data are monotonic. TCEQ should have used the restricted polynomial model before dropping the high dose, since the general trend at the high dose was the same as at lower doses. However, there does appear to be a flattening of the dose-response curve at the high dose, suggesting that it may be necessary to drop the high dose to obtain adequate fit, even when using a restricted model. This would need to be confirmed by doing the modeling, to ensure that the appropriate BMCL was identified. Similarly, since one of the coefficients in the 3-dose version of the polynomial was negative, it would be useful to run that model with the coefficients restricted to non-negative.

There are several errors in the comparison of the polynomial and Hill model AICs. First, recall that the AIC is a function of the likelihood and the number of parameters estimated (AIC = -2L +2k, where L is the likelihood and k is the number of parameters). Since the likelihood is a function of the actual data, one cannot compare AICs for different data sets (different endpoints), or for different numbers of data points (i.e., 3 vs. 4 data points for the same endpoint). The Hill model AIC would be expected to be larger simply because it has an additional data point. Second, the AIC values for the Hill and polynomial models are not "similar." Similar for AIC comparisons is often defined as a difference of 2 or less (e.g., 155 vs. 157). Since adding a parameter increases the AIC by 2, that sort of comparison provides information on whether the fit improved by more than the decrement in AIC due to adding a parameter. *Nonetheless, the fit from the two*

models is comparable. Both models (Hill and polynomial with 3 data points) fit the means perfectly. The polynomial had a somewhat better fit to the control SD, but this is not material, since the BMR was not related to the control SD.
Overall, it appears that the correct final model choice was reached, although I would prefer the use of the BMCL based on a 10% change in weight gain. In addition, I would need to see the results of the polynomial model with 4 data points, and the polynomial model with 3 data points and restricted coefficients before reaching a final determination regarding the modeling.

- **Should the POD be based on the maximum likelihood estimate or the 95% lower confidence limit of the reduction of weight gain?**

R1: Good arguments can be made for either. I believe that the current guidance from US EPA is to use the lower confidence limit. I believe that the DSD has made a good case for using the LCL: go with it. That said, for my own peace of mind let me present a case for using the central estimate, at least in this instance. I need to acknowledge that this is a minority opinion, but here it is. One of the principal reasons for using a lower confidence limit is that it rewards robust study designs, because, all things being equal, a larger, more statistically powerful experiment should have tighter confidence intervals than a small study. Developmental toxicity studies already have a robust statistical design in which the number of animals per dose group has been selected based on a consideration of statistical resolving power. Therefore, a case can be made that these studies already have a robust statistical design and should not be subject to the reward/punishment paradigm implied in the use of the LCL.

R2: I think the 95% lower confidence limit is a conservative choice that should be protective of public health.

R6: Use of the 95% lower confidence limit of the reduction of weight gain is most appropriate, based on existing guidance.

- **Was the appropriate benchmark response selected (5% vs 10% reduction of weight gain)**

R1: My opinion is that the choice of benchmark response of a 5% reduction may be overly conservative. Allen and coworkers, in their exhaustive comparison of BMDs and NOAELs for developmental toxicity studies (Fund. Appl. Toxicol. 23: 487-495 (1994)) found that a 5% effect level for BMD correlated reasonably well with NOAELs for fetal weight. It should be noted, however, that a 5% decrease is typically the minimum statistically detectable change in fetal weight in a guideline-compliant developmental toxicity study. The same is not necessarily the case for maternal weight gain, in which it

generally takes about a 10% change to be statistically significant. This was the case in the Hackett study, in which an 8% decrease in the 40 ppm was not statistically significant but a >10% change at the next dose level was. My opinion is that the BMC should represent something that is close to the minimal detectable change in well-designed studies, at least for a continuous variable like weight for which there is no sharply defined division between adverse and non-adverse. For this study, the BMCLs for the 10% weight reduction are much closer to the actual NOAEC for the experiment, but still conservative (on average) compared to the NOAEC.

R2: I understood that this choice was a policy one. I agree that the 95% lower confidence limit for the 5% response is approximately a NOAEL, while a 10% response is approximately a LOAEL.

R5: The rationale for selecting a 5% change in the mean as the BMR used to derive BMCs and BMCLs is not explained. In fact, used for POD estimates, it appears to be nonstandard. Though perhaps not an EPA "default," the use of a BMR defined in terms of a change in the mean equal to one standard deviation is more common and can be justified on the basis of an implied increase in the associated risk of response. If one thinks about the tail of the distribution of the endpoint (decreased weight gain in this case) representing abnormal response, then the dose at which the change in the mean equals one standard deviation can be interpreted as the dose that gives slightly less than 10% additional risk (assuming a background response rate of 1%). This way of defining BMRs and the associated BMCs is consistent with the manner in which one is accustomed to defining BMCs for quantal endpoints. Use of a BMR defined in terms of percent change in the mean does not consider the variability in the endpoint and can not be defined *a priori* in terms of a risk of response. I strongly recommend use of the BMCs defined in terms of standard deviation changes, or a presentation of a strong argument for selecting the 5% change in the mean.

R6: It is not entirely clear why the BMCL05 was chosen as the BMR, but based on the available information, the BMCL10 (i.e., 10% change in the mean weight gain) appears to be more appropriate. The ESL guidelines note that severity of effect is taken into account in choosing the BMR. Small decreases in maternal weight gain would be considered a mild effect. In addition, U.S. EPA (2000) recommends basing the BMR on consensus definitions of adversity when possible. Such definitions are rarely available, but decreases in body weight of <10% are generally not considered adverse. (Note that this is typically expressed as 10% change in final body weight; a 10% change in weight **gain** is generally more sensitive.) Finally, in the DSD discussion regarding publications on the choice of BMR, it was not clear how the analyses were done. Note that the BMCL05 and BMCL10 for maternal body weight refer to a 10% change in the mean, not a 10% change in the percent of animals affected. (For a normally distributed population, 50% of the animals would have a 10% change or larger if the mean is changed by 10%.)

- **Should the POD be considered to be a NOAEL or a LOAEL? Explain your reasoning**.

R1: NOAEL, definitely. The level selected was below that which would be statistically detectable, and I would argue that in this particular animal model is of negligible biological consequence.

R2: That depends on what you choose as a POD. If you choose the 5% response as the POD, I think the 95% lower confidence limit for that value should be considered a NOAEL.

R6: The appropriate UF was used to extrapolate from the BMCL05, based on the comparisons noted in the DSD regarding BMCL05 being equivalent to NOAELs on average. **However**, the BMCL05 should not be considered a NOAEL or LOAEL – it is a BMCL, which is different. One can, however, term it a NOAEL surrogate.

The choice of dosimetric adjustments

R1: These were appropriate and follow established guidelines. As noted in an earlier comment, (*TERA*: comment copied again below) it is also possible to adjust for respiratory rate/ minute volume between mice and humans, which would improve the extrapolation. Doing this necessitates calculating a dosage instead of an atmospheric concentration, and therefore back calculating to an acceptable human concentration; however, given the concern expressed that the high respiratory rate of mice may be unduly influencing the reference value, it is worth considering.

The reference values presented in the DSD were derived using practices that are in line with the ESL guidelines, and are consistent with procedures used by other state and federal regulatory agencies. There are a few points in the assessment in which different procedures could have been used which would still have been consistent with established risk assessment guidelines and which may have made the assessments even more data-driven. The example that comes to mind most readily is that, in the dosimetry adjustments for animal-to-human, no adjustment appears to have been made for different respiratory rates/ minute volumes for mice vs. humans. Default values for murine and human respiratory parameters exist and could have been used here to reduce some of the uncertainty in extrapolation that is cited in the text (e.g., p. 21, line 21, which lists "relatively high respiratory rates" as one of many possible factors that may predispose mice to have a more significant response to a given exposure. I don't see this as a fatal flaw, but it could have reduced uncertainty.

R2: The statement is made that "default duration exposure and dosimetric adjustments from animal to human exposure were used." This is a logical thing to do except for one major

point. The most sensitive species, the mouse, makes much more (up to about 100 fold) of the most toxic metabolite for both the cancer and noncancer adverse health effects, i.e, .the diepoxide, than do humans. The default dosimetrics are based only on the parent compound, which is not the compound of interest. In fact, the existing PBTK models were rejected because that did not include the tissue doses of the key metabolites, but the default dosimetrics also do not take into account the dosimetrics of the reactive metabolites.

Throughout the document (examples on page 14, lines 30-31;page 16, lines 38-39), it is suggested that "Uptake of BD in mice is faster than rats and may account for the increased susceptibility of mice compared to rats." This is absolutely not true and should be deleted. It is true that smaller animals have a greater ratio of minute volume to body weight than do larger animals, so that the smaller animals will always receive a greater internal dose per unit body weight than the larger animals in an inhalation exposure. However, if one normalizes to body surface area, instead of body weight, much of that difference goes away. Granted, the mouse metabolizes the BD faster than the rat, but the differences in the degree of response to BD is far greater than what one would expect based on the observed differences in uptake.

The difference between the response of mice and rats to BD has been shown to be largely based on the greater metabolism of BD to the diepoxide in the mouse versus the rat. Humans metabolize even less of the BD to the diepoxide than do rats, as documented by the recent data from Swenberg et al., 2007. The pattern of urinary metabolites (M1 vs M2) also demonstrates that human metabolism follows the hydrolytic pathways much more than rodents and thus excrete almost exclusively M1 and very little M2. The mouse has a high ratio of oxidative enzymes relative to hydrolysis enzymes while human have the reverse. Thus, all species metabolize BD to the monoepoxide, which can then be further metabolized by three main pathways (further oxidation to the diepoxide, conjugation with GSH to form M2, or hydrolysis to the epoxydiol). [The document describes this well.] These pathways have been studied in mice, rats and humans and it is clear that the mouse metabolizes the monoepoxide mainly via the conjugation and oxidation pathways, while humans use the hydrolytic pathway to convert most of the monepoxide to the butenediol and then further to the M1 urinary metabolite or to the epoxydiol. [The latter metabolite is produced in almost equivalent amounts by mice, rats, and humans, based on hemoglobin adducts, and thus would not be expected to be on the pathway to toxicity.] The upshot of all this is that the drastic differences in sensitivity to inhaled BD between mice and rats cannot be accounted for by greater uptake of BD in the mice and it should not be so stated in the document**.**

R3: BMDS Version 1.4.1 gives a lower 95% confidence limit on the estimate of 1,3-butadiene concentration associated with an excess risk of 5% for animals with abnormally low EWG of $BMCL_{05}$ = 8.74 ppm for a 6-hour exposure. Extrapolation of concentration to a shorter exposure duration is estimated by the ten Berge procedure

(ten Berge *et al.*, 1986), not Haber's Rule as incorrectly stated on Page 20, Line 4. The ten Berge procedure has been adopted by the EPA for calculation of acute exposure guidelines (AEGLs), see Krewski *et al.* (2004). Extrapolation is based on the expectation of equal effects for equal values of $C^n t$, where C is concentration and t is duration. Using the default value of n=3 for extrapolation to a shorter duration results in a 1-hour $POD_{adj} = 6^{1/3} \times 8.74 = 15.9$ ppm. Incorporating the combined uncertainty factor of 30 gives ReV = 15.9 / 30 = 0.53 ppm = 530 ppb and ESL = 0.3 x 530 = 159 (rounded to 160 ppb).

It is **highly recommended** that these values be used for the Acute ReV (530 ppb) and ^{acute}ESL (160 ppb). Change Summary Table 1, Page 5, accordingly.

R4: Due to the uncertainties associated with predicting tissue concentrations of the diepoxide in the human, I agree with the TCEQ decision not to use a PBPK model to perform cross-species dosimetry.

With regard to the duration adjustment, it is hard to believe that a single, one-hour exposure at 10.65 ppm could cause the same effect on weight gain as 10 daily exposures for 6 hours at 5.86 ppm. I think that for a cumulative effect such as decreased weight gain following several days of exposure, a linear adjustment could be used (i.e., Haber's rule with n=1). The use of n=3 is based primarily on short-term irritant effects and toxicities with a rapid onset. The application of this extremely conservative approach in this case of a cumulative effect requires more justification. If the n=3 assumption is used, T_1 should be 60 hours, not 6.

I agree with the use of the Category 3 default dosimetry adjustment for cross-species extrapolation. The default expectation for the cross-species dosimetry of a metabolite that is both produced and cleared by metabolism is that it would be similar to that of the parent chemical.

I am in complete agreement with the choice of uncertainty factors by TCEQ. The use of the default UF of 3 for animal-to-human pharmacodynamics, as opposed to the use of a UF of 1 for the ovarian atrophy, is justified because of the lack of mode-of-action evidence tying the weight gain effects specifically to DEB.

R5: The extrapolation to a 1-hour based acute value appears to have some serious issues associated with it. First, why would an assumption that n=3 be used for the Haber's Law-based extrapolation from 6 hour exposures? From my experience analyzing a variety of acute inhalation data sets, n=3 would appear to be a rather high-end value. Typically, values from fitting a probit model to the available data suggest that a range from slightly less than 1 to slightly less than 3 would be more usual. Is there some specific reason why the value of n=3 was used?

More importantly, how is the extrapolation from the reproductive/developmental study that exposed pregnant animals for 6 hours per day *for 10 days* using this Haber's Law procedure justified? In actuality, the dams who had the effect were exposed for a total of 60 hours, so if this simplistic approach is to be used, would not the extrapolation be from 60 hours to 1 hour, making the adjusted PODs much larger than those calculated (see top of p. 20). This use of repro/developmental results to estimate acute values is fraught with problems that do not appear to be adequately considered or addressed.

R6: At the least, more explanation is needed to explain the rationale for the exposure duration adjustment. First, it is not clear whether the authors considered the study to be a subacute study or a developmental toxicity study for the purposes of identifying the approach for extrapolation, although it appears that it was considered to be a subacute study, since the endpoint is systemic maternal toxicity. For such exposures, the guidelines appear to recommend that the DSD authors should compare ESLs developed based on the data with ESLs developed using the approach(es) for chemicals with minimal data, to ensure that the value derived is not over-conservative; it appears that such a comparison was not done for butadiene. **This consideration may have a quantitative impact on the acute ESL.**

The choice of uncertainty factors.

- **Have all of the appropriate uncertainty factors been considered and are the values assigned to the uncertainty factors clearly justified and defensible?**

R1: Yes. I believe that the DSD is transparent in its use of uncertainty factors and in selecting values for these factors that are within guidelines and representative of the available information on BD. The use of a 10-fold factor for human variability is a default value. There are some data on the variability of BD metabolism in the human population, but unless there are extensive studies to support moving from the default, I believe the value of 10 is justified. The value of 3 for animal-to-human acknowledges that the kinetics of BD favor the sensitivity of the murine model over humans, but that dynamic data are lacking. The other UFs are appropriately set at 1.

R2: I consider the UF_A of 3 to be highly conservative for the reasons stated on page 21, lines 19-22. It is well known that humans are much less sensitive to BD than mice and a much smaller UF or even a fractional UF should be considered.

R5: The choice of uncertainty factors is appropriate. In particular, the decision to not include an uncertainty factor different from 1 for the NOAEL/LOAEL component is consistent with the history of use of BMCs and BMCLs for POD determination. The goal for defining BMCs and BMCLs has always been to find a replacement for the NOAEL (not LOAEL) that has better properties (see the preceding paragraph for some aspects of that determination). There is nothing in this assessment to suggest that the BMCLs used for

defining the POD should be considered LOAELs from the standpoint of uncertainty factor selection.

R6: I agree with the choice of uncertainty factors, with one specific comment. I agree with a database UF of 1, but recommend a clarification of the justification. The acute database for butadiene meets the minimum database requirements for a high confidence acute ReV, not the minimal database for an acute ReV.

- **Would you make recommendations for a different approach to select uncertainty factors to calculate the acute ReV?**

R1: No. The approach taken is the standard one used by regulatory agencies. There is no basis for selecting an alternative.

R2: One of the largest uncertainties is making use of an endpoint that is not relevant to humans, i.e., a toxicity not known to occur in humans. The approach used is highly conservative and protective of human health, but I am not convinced it needs to be that conservative. I get a mental picture of the ReV protecting all the little mice in Texas and making sure they can reproduce—not exactly the goal of the TCEQ. Humans are not nearly as sensitive as mice to the effects of BD.

Other Comments on the Health-Based Acute ESL

R5: With respect to the charge questions relating to the statistical analyses and the appropriate way of doing them: it is important to get the statistical methodology correct and to use the best statistical methods available. However, in the context of this risk assessment, the discussion of the identification of NOAELs and LOAELs is irrelevant. As soon as one has decided that an endpoint in question exhibits some dose response (e.g., with a trend test), then there need be no further statistical determinations of whether or not certain dose groups differ from controls. That is so because the dose-response modeling that is used to define BMCs and BMCLs does not depend in any way on those determinations. The dose-response analysis is all that is needed; NOAEL/LOAEL designations are immaterial.

Note however, that the selection of endpoints for modeling should probably not be restricted to those that have the lowest NOAEL or LOAEL. Since the overall dose-response pattern is what determines where the BMC is (based on the definition of the BMR), it is possible that an endpoint that yielded the lowest NOAEL or LOAEL does not yield the lowest BMC or BMCL. Thus, in the context of this butadiene risk assessment, TCEQ should include further discussion of the other endpoints that were not chosen to model; it may suffice to show that the responses at 40 or 200 ppm were much less (relative to their own controls) than those for the weight gain reduction endpoints that

were modeled.

In connection with the comments above about the statistical analyses and the fact that they are irrelevant to the definition of the ReV and ESL, note that the statistical determination of an "effect level" is highly dependent on the sample sizes used in an experiment. And, as has been discussed repeatedly, the effect of sample size is the opposite of what one would hope; smaller sample sizes (associated with more uncertainty) tend to give larger NOAELs and therefore less health-protective PODs. Use of the lower bounds on a benchmark dose or concentration, conversely, penalizes experiments with greater uncertainty because those bounds tend to be wider (lower bounds tend to be lower) with smaller sample sizes. For this reason, here and throughout the remainder of this assessment, the use of the lower bounds for POD determination is appropriate.

R6: P. 12, second paragraph: Early resorptions were lower in the exposed groups. Thus, regardless of the statistical approach used to analyze the data, no adverse effect was observed, and there should be no LOAEL for this endpoint. The identification of the LOAEL appears to be an error by TCEQ, but if TCEQ believes that this effect was adverse, additional explanation should be provided. However, this consideration does not affect the ESL.

Welfare-Based Acute ESL

The key decisions and some specific issues to consider are listed below. Please indicate if there are other issues specific to developing welfare-based ESLs that have not been adequately addressed in the document.

The choice of the Nagata (2003) study as the basis of the acute ESL for odor.

R1: I agree with the choice of the study. It appears to be current and well conducted.

R2: I agree with your choice of the Nagata study to set an acute ESL for odor.

The decision to not derive an acute ESL for vegetative effects.

R1: I agree that this is not a priority, given that the reference level will be orders of magnitude above that calculated for health effects.

R2: I agree with the reasoning the led you not to derive an acute ESL for vegetative effects.

R3: Agree with the Development Support Document as written.

R4: I agree with the choice of the Nagata study as the basis of the odor ESL, and the decision not to derive an ESL for vegetative effects.

R5: The section on a welfare-based acute ESL is very brief. It appears to be suitably so, if the comments about the relative values for odor threshold and vegetative effects are correct. I have no comments on those determinations.

R6: These decisions were appropriate and followed the ESL guidelines.

Health-Based Chronic ReV and $^{chronic}ESL_{noncancer}$

The key decisions and some specific issues to consider are listed below. Please indicate if there are other issues specific to developing chronic toxicity factors that have not been adequately addressed in the document.

R3: Section 4.1, Noncarcinogenic Potential. This section is comprehensive, clearly written and supports the derivation of the Chronic ReV = 15 ppb and $^{chronic}ESL_{nonlinear\ (nc)}$ = 4.5 ppb.

The choice of NTP (1993) as the critical study.

R1: I believe that this is the most appropriate study for setting a chronic non-cancer reference value. It is a high quality study evaluating the effects of lifetime exposure to BD in a sensitive model, the mouse.

R2: This seems to be the appropriate study if one wants to use the most sensitive endpoint in the most sensitive species. It was certainly a well-done study. However, as stated in the document on the same page, neither rats nor humans exhibit this toxic endpoint after chronic exposure to BD.

R4: I agree with the choice of the NTP (1993) study as the critical study and ovarian atrophy as the critical effect. I can think of no reason why this endpoint should not be considered relevant to humans. I also agree with the use of parent chemical concentration as the dose metric, the $BMCL_{05}$ as a NOAEL equivalent, and the use of data from all doses in the time-to-response modeling. I don't believe the $BMCL_{05}$ is sufficiently far below the data to introduce excessive uncertainty into the POD.

R5: The NTP (1993) study appears to be an adequate basis for performing the evaluation, although no explicit comparison with other candidates is provided. And the use of ovarian atrophy appears to be justified, although I cannot speak to its relevance to humans. The modeling and the choice of dosimetric adjustments as well as of the uncertainty factors all appear to be satisfactory. In particular, I believe the time-to-response analysis is appropriate and does make it feasible to include the high-dose observations.

R6: The NTP study is appropriate, and the endpoint is relevant, to the best of my knowledge.

The choice of ovarian atrophy as the critical effect.

- **Is the selected endpoint relevant for humans?**

R1: I believe that this is a reasonable choice. The effect was observed even at the lowest dose level. It has relevance to human health. Investigative studies indicate that ovarian toxicity is dependent on metabolic generation of the diepoxide metabolite, which mice generate at much higher levels than other species. However, there is some data in the literature indicating that humans can generate this metabolite, albeit at lower levels. Therefore, at least in a qualitative sense, the mouse is a relevant model for humans.

Yes, the endpoint is relevant for humans. The human ovary is susceptible to agents that are reactive (either as parent or metabolites) and toxicity can be manifested in a number of different ways, including infertility, subfertility, or accelerated reproductive senescence, among others. The effects observed in the mouse studies should be considered human-relevant.

R2: I think it is appropriate as a sensitive endpoint in the most sensitive species. Is it relevant to humans? No, as you have stated, this endpoint does not appear to occur in humans. Mice are known to have hematopoietic toxicities after chronic exposures (rats do not). I think you should consider using those data.

The choice of point of departure based on a 5% increase incidence of ovarian atrophy in female mice (i.e., $BMCL_{05}$) (Appendix 3 Statistical Analyses of Reproductive Endpoints)?

- **Were the ovarian atrophy data in mice correctly modeled?**

R1: I believe that the benchmark dose methodology was applied according to guidance that has been provided on its use. I believe that the DSD takes a pragmatic approach to model selection that is appropriate.

R2: I agree with the modeling of the ovarian atrophy data in mice.

R5: My biggest concern is the use of 5% risk instead of 10% risk as the BMR. The U.S. EPA tends to use the 10% risk BMR. As in the case of butadiene, they sometimes use an effect level-related uncertainty factor, as a result of that choice. In general I am not in favor of that uncertainty factor; there does not appear to be any strong evidence to support the contention that a 10% risk level is associated more closely with a LOAEL than a NOAEL. In fact, evidence from a quantal (dam-based) analysis of developmental toxicity studies (having typically around 25 subjects per dose) suggested that the

$BMCL_{10}$'s would tend to be less than the corresponding NOAELs (Allen et al., 1994). Thus, not only should the $BMCL_{05}$'s be considered NOAEL replacements, the $BMCL_{10}$'s should be as well and probably should be the basis for the POD calculations.

- **Should the highest dose group be included or excluded in the time-to-tumor model?**

R1: (I am interpreting that time-to-tumor was intended to be time-to-response.) I agree with the decision to include the highest dose group in the model. As modeled, the BMCs are close to each other with or without the high dose group, and as a matter of practice I like the idea of including all the data unless there is a good reason not to.

R2: The results did not differ very much whether the highest dose was excluded or not. I think it was appropriate to include the highest dose for benchmark modeling because it uses all of the data to establish a POD.

- **Should the POD be based on the maximum likelihood estimate or the 95% lower confidence limit of the benchmark exposure concentration for an extra risk of 0.05?**

R1: As noted above in my response for the acute reference value, most scientists in the field prefer using the LCL. I believe that there is support for doing so, and I can live with the decision to do so here. However, as noted above, there are legitimate reasons to choose the central estimate as the POD, particularly when it is based on a study that has been designed to provide robust statistical power. As with the regulatory-compliant developmental toxicity study, the NTP two-year bioassay is an extremely robust study that has taken into account the need for a large sample size. Therefore, it may be reasonable to choose the central estimate of the dose-response curve for these types of studies.

R2: I understand why the TS chose to use the 95% lower confidence limit, but it does force the agency to break the "rule " for benchmark modeling of setting the POD within the range of the experimental data. This is a policy choice. How protective do you want to be? If policy says you must use the 95% lower confidence limit for severe effects, I think there will be many instances when you will be outside the range of the experimental data.

R6: The mathematical details of the modeling are beyond my expertise. But I do recommend the use of the 95% lower limit. Further, in light of the potential severity of the endpoint, the 5% response appears more appropriate than 10% extra risk, according to the ESL guidelines.

- **Should the POD be based on the benchmark exposure concentration for an extra risk of 0.05 or some other extra risk level, e.g., 0.10?**

R1: As noted above, my opinion about BMD is that it be close to the lowest detectable level for the study design being used. This not only promotes consistency in risk assessment between BMD and NOAEL-based approaches, but also ensures that there isn't too much extrapolation below the experimental data points on the dose-response curve. The DSD deals explicitly with the choice between a 5 and 10% effect level. It acknowledges that the BMDL05s are below the experimental range. However, it cites both the severity of the observed effect and regulatory precedent as rationale for using the 5% level. This is one of those difficult decisions that might be made differently by others, but is clearly described and defensible.

R2: I think the use of an extra risk of 0.1 might allow you to use data in the experimental range (see comments above).

- **Is the POD considered to be a NOAEL or a LOAEL?**

R1: NOAEL, definitely. The DSD acknowledges that the BMCL05 is below the experimental data range and that one of the reasons for selecting this effect level is that there is regulatory precedent for considering it to be a NOAEL equivalent.

R2: That depends on the data. I agree with your approach that the BMCL05 can be associated with a NOAEL.

The choice of dosimetric adjustments

R1: These were appropriate and follow established guidelines. As noted in an earlier comment, it is also possible to adjust for respiratory rate/ minute volume between mice and humans, which would improve the extrapolation. Doing this necessitates calculating a dosage instead of an atmospheric concentration, and therefore back calculating to an acceptable human concentration; however, given the concern expressed that the high respiratory rate of mice may be unduly influencing the reference value, it is worth considering.

R2: It is unfortunate that not enough data are available to allow you to use the appropriate dose metrics such as diepoxide to the tissue. Could you not extrapolate from the data published on the diepoxide in the blood of mice exposed to varying levels of BD (See Thornton-Manning et al., 1995, 1998)

R4: I agree with the TCEQ decisions regarding duration and cross-species adjustments, as well as with their selection of UFs. In particular, I agree with the decision to lower the UF for

animal-to-human toxicodynamics due to the strong evidence that the human produces less of the proximal toxic species, DEB. I believe the MOA evidence associating the ovarian atrophy with DEB is adequate to support this decision. I also agree that, while the available data supports a reduced UF, it is not adequate for the calculation of a CSAF. The use of the default UF of 3 for animal-to-human pharmacodynamics for the acute ReV and ESL, as opposed to the use of a UF of 1 for the ovarian atrophy, is appropriate because of the lack of mode-of-action evidence tying the weight gain effects specifically to DEB.

R6: In general, I agree with the choices made by TCEQ, but see the comments regarding UF_A.

The choice of uncertainty factors.

- **Have all of the appropriate uncertainty factors been considered. Are the values assigned to the uncertainty factors clearly justified and defensible?**

R1: I believe that the appropriate factors were all considered and that the values selected were appropriate. The intrahuman uncertainty factor was kept at the default value of 10 to acknowledge that not enough is known about variability across the population to move from the default. The animal-to-human factor was reduced to 1 to acknowledge not only that from a kinetics standpoint is the mouse more sensitive, but from a dynamics standpoint enough is known about the mode of action of the ovarian toxicity to support a conclusion that humans are less able to make the toxic metabolite. A database uncertainty factor of 3 was applied to account for the fact that there is no multigeneration reproductive study for BD. I believe that this can reasonably be considered a data gap given that the critical effect is on a reproductive organ.

R2: Yes, I think you did a good job on that, except that it would be good if there were some way to indicate the large uncertainty associated with the huge difference in the dose of the diepoxide to the tissue in humans versus mice. See answer to next query.

R6: I agree with the choice of UF_A, but recommend a refinement to the presentation. All of the considerations addressed in the DSD reflect toxicokinetics – the tissue dose resulting from a given external exposure concentration. Both the default dosimetric adjustments and the consideration of Hb adducts reflect the relationship between tissue dose and external exposure concentration. In contrast, a toxicodynamic metric would be some measure (e.g., an early precursor) related to the ovarian atrophy. Hb adducts have been used to calculate AUC in the blood, but information on the rate of reaction between the DEB and Hb would be needed to conduct such calculations. (See, for example, Fennell et al., 2005, for similar calculations conducted for acrylamide and glycidamide.) It is not clear whether such rate data exist for butadiene or the DEB metabolite, but if they do exist, they could be used to refine the choice of UF_A. In the absence of such data, this reviewer agrees that it appears that an appropriate CSAF can not be calculated, but the

data do support the conclusion that the interspecies kinetic factor would be less than 1. In the absence of information on interspecies differences on toxicodynamics, the interspecies toxicodynamic subfactor would still be 3. However, TCEQ could make a valid argument that the toxicokinetic factor would be less than 0.3, and so an overall interspecies factor of 1 is still appropriate. Use of a factor less than 1 would require additional quantitative analysis of toxicokinetic differences.

Note that the choice of a UF_A of 1 is based on the comparison of DEB levels (and surrogates for DEB levels. No data were presented implicating DEB for the other potential critical effects, suggesting that UF_A may be 3 for these other critical effects. However, the potential PODs for these other endpoints were more than 3-fold higher than the POD for ovarian atrophy, and so this reviewer agrees that further analysis for these other endpoints is not needed.

I agree with the choice of 3 for the UF_D. Even though the critical effect is a reproductive effect, a lower POD could still be identified in a multi-generation study. This is of particular concern in light of the genotoxicity of butadiene and the evidence that it does reach the female reproductive tract.

I agree with the documentation regarding the other uncertainty factors.

- **Was the Swenberg et al. (2007) data on differences between occupationally-exposed workers and mice and rats if the formation of hemoglobin adduct used properly to characterize the animal to human toxicodynamic uncertainty factor?**

R1: Yes, the Swenberg data provide evidence from humans that the conclusion that humans generate very little of the toxic diepoxide metabolite is supportable.

R2: Yes, the Swenberg data were used and I applaud you for that. I think the uncertainty factor should really be less than 1, but you stated there are not procedures to do that. Perhaps Texas could be at the forefront to set up such procedures. The fact that the adduct that is specific for the diepoxide was not detectable in the Czech workers indicate the high degree of uncertainty in using toxicity in mice as a model for humans.

- **Would you make recommendations for a different approach to select uncertainty factors to calculate the chronic ReV?**

R1: No. The approach taken is the standard one used by regulatory agencies. There is no basis for selecting an alternative.

R2: My suggestion is for Texas to "take the bull by the horns" and come up with a way to use uncertainty factors less than one, when such are warranted.

Other Comments on Noncancer Assessments

R2: Additional summary statement for noncancer endpoints: It is conventional, in the absence of human data, to base risk assessments on the most sensitive endpoint in the most sensitive animal model. We have taken this approach for several years with 1,3-butadiene, using the mouse as our model. However, now we DO have human data. We know that the most toxic metabolite, for cancer or noncancer endpoints is the diepoxide. We know that the diepoxide is associated with ovarian atrophy. We know that humans produce minimal amounts of this metabolite and humans have not shown evidence of reproductive toxicity in response to BD. Thus we know that using mouse data for a risk assessment will greatly overestimate human risk. I think that we have enough data now to no longer use the mouse reproductive toxicity data to estimate risk for humans, or if we use it , we should correct with a fractional uncertainty factor to reflect the fact that we know we are going to get an overestimate. For cancer endpoints, of course, we can use the human data.

R6: More of a critical analysis is needed of the molecular epidemiology study conducted by Albertini et al. (2007), particularly considering that it is the only human study that evaluated endpoints related to the critical effects for the acute and chronic noncancer ESLs. In particular, the power and sensitivity of the study should be considered, since no effect was seen on the reproductive endpoints. Is the study adequate to put a bound on the risk of reproductive effects?

Cancer Weight of Evidence and Unit Risk Factor (URF)

The key decisions and some specific issues to consider are listed below. Please discuss other issues specific to developing unit risk factors for carcinogenic effects that have not been adequately addressed in the document.

R3: Section 4.2, Carcinogenic Weight-of-Evidence and MOA. This is an excellent Section. The Section is very comprehensive, covers several options for selection of data and analyses, and is well-written. The derivation of the $^{chronic}ESL_{linear\ (c)} = 28$ ppb is fully supported.

R5: In general the cancer risk assessment appears to be appropriate and satisfactory. The analyses of the epidemiological data appear to be thorough and to have considered the two approaches (Cox and Poisson regression) typically applied to such retrospective cohort studies. Although there are advantages to each approach, it is not clear that one approach is superior to the other. The inclusion of the HITs covariate, and the discussion

of its impact on the extrapolation of the modeling results to the general population, is particularly appropriate; its inclusion in the modeling is a definite improvement over the more typical consideration only of a ppm-year metric.

R7: Overall I believe the Texas risk assessment is reasonably conducted and is based on the best epidemiologic data available. The epidemiologic data is furthermore rather strong, and is a sounder basis for human risk assessment than animal data, in my opinion. Excess risk calculations based on observed epidemiologic exposure-response relationships appear to be correctly done using standard methods. I was able to reproduce the basic excess risk calculations using BIER life table methods.

The weight of evidence statement

- **Is the epidemiological evidence in Albertini et al. (2007) properly used in the characterization of chronic cancer risks?**

R1: This statement summarizes the opinions of EPA, IARC, OSHA, NIOSH and ACGIH, all of which have concluded that BD is a potential or known human carcinogen. These conclusions are based on epidemiological studies in occupationally-exposed populations in which lymhohematopoietic cancer rates were increased, animal studies showing tumors at multiple sites, and genotoxicity studies indicating that the carcinogenic process is initiated by direct DNA reactivity of BD metabolites that are generated in humans and animals (albeit at different rates). I believe the weight of evidence statement to be an accurate synopsis of current national and international regulatory opinion.

The data from this paper does not appear to have been considered in the weight of evidence section. This section deals exclusively with the potential for BD to be a human carcinogenic hazard, whereas the Albertini work addressed the question of risk in humans. The Albertini research appears to have only been used to compare the NOAEL level for a biomarker of BD effect (800 ppb in a Czech population) with the air concentration estimated to convey a 10^{-5} risk. I believe that this is appropriate given that the cancer risk estimates are based on human data. The Albertini results add support that these estimates are protective.

R2: I think it is excellent that you used the Albertini data. It is up-to-date and confirms earlier findings that humans are much less sensitive to BD than are mice or even rats.

R6: The weight of evidence statement is appropriate. The Albertini study is a useful one, but it is not clear what conclusions TCEQ is deriving from this study in the overall context of the cancer evaluation. In particular, consideration of the "clear NOAEL" for gene mutation in the context of the calculated cancer risk would be useful (considering such issues as sample size and sensitivity). In addition, the finding that DEB levels were

lower than those in mice or rats exposed to 1.0 ppm is not surprising, considering the worker average exposures were several fold below 1.0 ppm (although some individual 8-hour TWA values were higher). Human DEB values would be expected to be lower than the rodent ones, even without the observed differences in metabolism. Finally, a key bit of information in the Albertini study was buried in other discussion. The finding that urinary BD metabolite levels were lower in females was not surprising, since females had lower exposures. But Albertini also noted that the metabolite levels were lower for the same butadiene exposure concentration compared to males. This latter finding is more significant, and does not appear to be mentioned in the DSD.

The statistical and modeling approaches used for selecting different butadiene cancer potency estimates: Cheng et al. (2007): Sielken et al. (2007); and Sielken et al. (Appendix 4 Additional Cox Proportional Hazards Models).

- **Was the dose metric selected, cumulative ppm-years, the most relevant and appropriate choice?**

R1: My general opinion is that the statistical and modeling approaches used by Sielken and Cheng are reasonable. They are consistent with accepted practice for dose-response modeling and with the apparent mode of action of BD. The epidemiological data being modeled appears to be of high quality. However, the uncertainties associated with estimating exposures over an entire career, as well as the potential for co-exposures to other agents, posed challenges to the modelers. I believe that these difficulties were addressed to the extent possible by both Cheng and Sielken. It is reassuring that their estimates of beta and the upper confidence limit on beta were all within an order of magnitude of each other, and generally within a factor of two or three.

I believe that this is an appropriate choice. Given the likely mechanism of action, direct DNA reactivity, and the accepted model of carcinogenesis being the accumulation of mutations in cellular control genes, ppm-years acknowledges that the accumulation of exposures over time is consistent with what is understood about carcinogenesis. Furthermore, ppm-years is a standard metric for these kinds of calculations, so there is precedence for using it. I am not ready to conclude that it is the most appropriate choice, however. Sielken's analysis that included both ppm-years and excursions above 100 ppm (HITs) appears to fit the data well and may be more biologically relevant. The biological relevance is based on the presumption that a single exposure of 100 ppm may be more damaging than, say 100 exposures of the same duration to 1 ppm, because there is a greater chance that detoxification pathways and DNA repair mechanisms would be saturated. Therefore, I believe that the ppm-years, adjusted for HITs, is the best metric (and the one that appears to have been used in the calculation of risk for the general population). I believe that it would be desirable to address the assumptions about saturation of metabolism and repair experimentally, to provide more support for using this metric. The Albertini data provide some support that these are reasonable

assumptions.

R2: I think it was an appropriate choice.

R4: The TCEQ approach for conducting the cancer dose-response modeling appears reasonable to me, but many of the technical issues are outside my area of expertise. I agree with theweight of evidence determination from Preston (2007) and believe the use of linear extrapolation to obtain low-concentration risk estimates is appropriate in this case.

I agree with the use of cumulative ppm-years as the dose metric and "all leukemia" as the endpoint. I also agree with the use of a response rate of 0.1% as the POD in this case, due to the large size of the cohort.

R5: Although ppm-years is the proper metric to use to relate population exposures to population risks, it may be more appropriate to consider a window of exposure (probably lagged by some number of years) as the measure to be used. Evidence from analyses of benzene and from observations of the occurrence of leukemias in Japanese populations exposed to radiation from the dropping of the atomic bombs (Crump and Allen, 1987) suggests to me that a window of exposure (or a weighting of past exposures) may give a better characterization of the risks associated with butadiene exposure.

R6: Most of the details of the modeling and data choices are beyond my level of expertise. However, I will note the following:
-cumulative exposure appears appropriate, and adjustment for the number of HITs is desirable.
--use of confidence limits (rather than MLEs) appears appropriate, despite the use of epidemiology data, in light of the uncertainties in the analysis. However, the EPA (2005) guidelines are ambiguous on this issue.

R7: The dose-metric selected , cumulative ppm-years, is standard for chronic disease and is appropriate.The epidemiologic support for the cancer risk assessment appropriately comes from the large epidemiologic study most recently updated by Delzell (HEI 2006), Sielkin et al. (2007), Mancaluso et al. (2004), and Cheng et al. (2007). Albertini et al (2007) addresses biomarkers can is not directly relevant to this risk assessment. Mancaluso et al. (2004) provides a validation of the estimate exposures and not any exposure-response data, which is clearly important since no prior validation was available. It appears that the exposure estimates are on the whole reasonable, but nonetheless involve a considerable amount of mismeasurement. The key studies for risk assessment, which models the exposure-response for butadiene and leukemia, are appropriately the basis for the risk assessment, are Cheng et al. (2007) and Sielkin et al. (2007).

- **Are there reasons to prefer Cox regression modeling over Poisson regression modeling or vice versa?**

R1: I am a toxicologist, not a statistician, so the fine points of the two models is lost on me. Given that the two models provide roughly the same answer (similar betas), I see no reason to prefer one over the other in this instance.

R7: Poisson regression is inherently categorical and the shape of the exposure-response in inherently determined by the choice of cutpoints for categories, which is in itself ultimately somewhat arbitrary. Cox regression avoids these issues and is to be preferred. Adoption of a parametric model in Cox regression should be done after examination of categorical and spline results, preferably conducted within the Cox model (not within Poisson regression) so have as much as possible comparability between categorical and purely parametric models. Categorical and spline results will provide minimally parametric graphical representation of the exposure-response shape and will guide choice of a simpler parametric model. Such results are provided in Cheng et al. (2007). It might be useful to incorporate into the Texas risk assessment as a graph with the spline and the categorical points.

- **Comment on the relevance of using penalized spline regression and restricting the data to the lower 95% of the exposure range of all subjects.**

R1: The general question in my mind is whether it is reasonable to restrict the data set. Given the high uncertainty in the exposure assessment at the upper end of the distribution, I think it is reasonable to restrict the data set. Many of the model runs took other approaches to limit the impact of the high exposure estimates, particularly the models that clumped the exposures into deciles. Doing so resulted in potency estimates that appear to be more robust.

R5: One specific charge question asked us to comment on the relevance of using penalized spline regression and restricting the data to the lower 95% of the exposure range of all subjects. I did not see that these approaches were used in the analyses in the current DSD and so I have no comment on this point.

R7: Penalized spline regression do not appear in section 4.2 of the risk assessment. However, penalized splines are used appropriately in Cheng et al. (2007), and could usefully be part of the risk assessment. Regarding analyses without the top 5% of the data, I believe such analyses are appropriate as a sensitivity analysis, particularly as the highest exposures are likely to involve greater mis-measurement. Such exposures can sometimes have a strong influence on the shape of the exposure-response curve. High exposures do appear to

have a big influence here, as per page 19 of Chang et al. (2007), section 3.4. Texas might consider basing its risk assessment on the data excluding the top 5% of exposures, or some kind of average between the full data and the restricted data. These data appear to be linear in the low dose region as per Figure b on page 19 of Chang et al.

- **Are exposures in the distant past and the immediate past equally biologically relevant? Are lags or windows of exposure biologically relevant?**

R1: I don't believe that we have sufficient data to address these questions specifically for BD. In general, because cancer is a disease that has a long latency, earlier exposures could be considered to be more likely to lead to a carcinogenic outcome than later exposures. However, in the context of occupational exposures, I don't think that this is terribly relevant.

R2: This is a good question, but I do not think we have the data to address it. The use of the cumulative ppm-years as a dose metric is an appropriate choice in the absence of data to the contrary.

R5: Although ppm-years is the proper metric to use to relate population exposures to population risks, it may be more appropriate to consider a window of exposure (probably lagged by some number of years) as the measure to be used. Evidence from analyses of benzene and from observations of the occurrence of leukemias in Japanese populations exposed to radiation from the dropping of the atomic bombs (Crump and Allen, 1987) suggests to me that a window of exposure (or a weighting of past exposures) may give a better characterization of the risks associated with butadiene exposure.

R7: Lags were appropriately explored by Cheng et al. (2007) and as noted there and in the risk assessment did not provide a significantly improved fit to the data nor much change in the exposure-response coefficient. It is therefore reasonable not to use lagged exposures in the risk assessment. Leukemia would be expected to have only a short lag in any case, and epidemiologic studies might not be sufficiently sensitive to estimate such a lag accurately.

- **Was endpoint selected as the basis of the potency estimates, "all leukemia", the most appropriate and relevant choice?**

R1: I believe that it was. The epidemiological data suggest that BD exposure has the potential to cause a variety of hematopoietic cancers. This is supported by animal data showing tumors at multiple sites. Therefore, it seems reasonable to lump all leukemias in modeling the dose-response.

R2: I have heard this discussed many times. To the extent that similar mechanisms of action are in play for "all types of leukemia", it is an appropriate choice. I do not think that the MOA is the same for all leukemias, but all the regulatory bodies you list on page 31 have accepted this as an appropriate choice.

R7: All leukemia appears to be the correct endpoint for risk assessment. Other outcomes were less related to butadiene exposure.

- **Should excess risk be calculated using leukemia incidence rates or leukemia mortality rates? Comment on Appendix 7 (page 145) Calculating Excess Risk when Specified Response is Mortality versus Incidence.**

R1: Ideally, one would want to calculate risk based on incidence rather than mortality, as it is the former that one wants to protect against. However, there are pragmatic limitations to this approach, which are well presented in Appendix 7 and have been published in the literature by others. Given these limitations and the inherent conservatism of the risk estimation procedure, it seems to me that a less-biased estimate of risk based on mortality is better than a more-biased estimate based on incidence.

R2: In this case, I think incidence rate is more appropriate, because we now have cures for some leukemias. We want to know the risk of developing leukemia.

R7: I do not see any major methodological difficulty in calculating excess risk from incidence as opposed to mortality data using life table methods. It would seem that one has only to make a minor adjustment to incorporate the subsequent mortality of those who get leukemia. As these will be few, such an adjustment will have little effect. A more important issue is how to extrapolate from RRs (ie, the exposure-response) calculated for leukemia mortality to RRs for leukemia incidence. Typically one assumes that these RRs are the same, which is a reasonable assumption, and one can of course do some sensitivity analyses about the effects if this assumption is varied.. I note that other risk assessors in Health Canada and US EPA have done this in past risk assessment for butadiene. For another example, with a full discussion of this issue, for a different agent, see the recent EPA risk assessment for ethylene oxide, where risk assessment based on epidemiology for hematopoetic cancer mortality was conducted for hematopoetic cancer incidence.

- **Does using the 95% UCL estimate instead of the central estimate somewhat account for the uncertainty that leukemia incidence rates are higher than leukemia**

mortality rates?

R1: No. The confidence intervals are indicators of the variability (and to some extent uncertainty) in the dose-response curve for mortality. They don't speak to the relative relationship between incidence and mortality. It can't be claimed that the UCL accounts for the difference. What can be said is that using the UCL adds conservatism to the estimate, and that by doing so it is probable that risk of incidence will be lowered.

R2: It is one way to try to account for the difference. Are there any hard data by which that might be tested?

R4: I am comfortable with the TCEQ decision to use the 95% UCL rather than the MLE, but don't think that decision should in any way be construed to take into account concerns about leukemia incidence vs. mortality. If there is a concern about the use of data on incidence vs. mortality, it should be accounted for directly.

R5: With respect to the use of the confidence limit calculations: I believe that these are the appropriate bases for deriving URFs; they cover a known uncertainty (that associated with estimating the parameters of interest given a finite sample size) and should be used to ensure health protective conservatism. However, I do not consider that their use covers or adjusts for the use of leukemia mortality instead of leukemia incidence, nor for the fact that the cohort analyzed consisted primarily of males. Those are separate concerns and ones for which separate uncertainty considerations need to be applied.

However, in relation to the mortality/incidence issue, my experience suggests that mortality analyses are the more standard ones used in regulatory settings with epidemiological data (especially occupational cohort data). Also in relation to the mortality/incidence question, I think the references to Appendix 7 are a bit misleading. While it is not appropriate to use parameter estimates (e.g., β from Cox or Poisson regressions) derived from mortality data as if they applied to incidence, the discussion implies that one could never do a lifetable (BEIR type) analysis given incidence rates or estimates of age- and dose-specific incidence. That is not true. If one considers the occurrence of a cancer (incidence) as a removal from the population (as one would do for any mortality), and adjusts the all-cause mortality rates by excluding leukemia mortality (because leukemia *incidence* is now counted as a removal), then the standard lifetable analysis using incidence rates can be done and one can derived lifetime estimates of cancer (incidence) risks. Of course, as stated above, if one can only estimate mortality rates from the studies with exposure and response data, then one should not attempt to calculate lifetime cancer incidence risks.

R7: Use of the 95% UCL estimate is a common conservative practice which I believe is standard in this type of risk assessment, to allow for uncertainty of all types, including for example estimation of RRs for incidence based on mortality RRs. (see (7) above).

- **Would best estimates (maximum likelihood estimates) of excess risks be more appropriate than estimates based on 95% upper confidence limits given that the estimates are based on human epidemiological data?**

R1: No, not in this case. One could make an argument that, because the cancer risk estimates are based on human data, all of the uncertainty around extrapolating from animal data is removed, and at least some of the uncertainty around human variability has been addressed. However, the human data comes with increased variability in exposure measurements and the possibility that not all covariates have been identified and controlled for. Using the MLE dismisses these uncertainties.

R2: Good idea! I would like to see both estimates considered. But the reason for using the 95% UCL is that there are some subpopulations that may be more sensitive to BD than others.

R7: I believe the use of the 95% UCL for exposure-response coefficients is appropriate.

- **Did the approach used adequately address the potential impacts of exposure misclassification? Would use of exposure deciles have been more appropriate than using continuous exposure?**

R1: The statistical analysis addressed the exposure misclassification issue in a number of ways, including the use of decile categories. I believe that the approaches used were reasonable. Importantly, the estimates of beta (central and UCL) were reasonably similar whether the actual exposure (ppm-year) or deciles were used.

R2: I think the approach used adequately addresses the problem of exposure misclassification.

R7: Exposure misclassification in log-linear models such as used here can bias estimates either towards or away from the null, and is not easily predictable without extensive work, and even then might be difficult to estimate. Use of the mid-points of exposure deciles does not necessarily mitigate any effects of exposure mismeasurement using estimated exposure levels for each individual. Exposure deciles are dependent on the choice of cutpoints (the best way is based on allocating leukemia cases equally into all categories, which I believe is what was done here , given that it was done in Cheng et al. (2007) from which Texas has taken its data. Furthermore, categorization essentially assumes everyone within the exposure category has the same risk of leukemia, regardless of their exposure. Essentially the result is a step function with 10 RRs. Assigning the midpoint of the category (what was done for the uppermost one, details not provided in Cheng et al. (2007)?) and then calculating a regression line as if these midpoints were a continuous variable probably introduces mismeasurement because it assumes all subjects within category (decile) were exposed equally. While this is a reasonable procedure among

many modeling strategies, I would not necessarily base the risk assessment on this model. I note that the log transformation and the square root transformation in Cheng et al. (2007) Table 4 provide as good a fit to the data as the decile analysis. The log transformation usually provides an unrealistically high slope in the low dose reason and for this reason is not preferred, but the square root transformation might be worth further exploration. That said, data presented suggests excess risks calculated from a variety of different exposure-response models do not differ much.

- **Were the appropriate covariates used in estimating cancer potency? Have the results using alternative covariates been properly weighted?**

R1: I believe the appropriate covariates have been used. As noted above, I found the use of the HIT metric especially compelling from the standpoint of biological relevance. I believe that the use of covariates has been appropriately considered by the DSD.

R2: I thought you did a nice job in this area. Well done.

R7: I do not agree with using models with both high intensity peaks and continuous exposure in the model at the same time. These are two different measurements of exposure and are certainly to some extent co-linear, thereby including them together will necessarily reduce the effects of each, as can be seen in Cheng et al (2007)'s result in the text on page 30 (section 3.5), bottom right. Furthermore, I have very little confidence that the number of peaks above 100 ppm can be estimated accurately. Unlike the exposure measurements in general, there is no validation of this exposure measure, which is too some degree really just an estimate of industrial hygienists of how often leaks might occur. I would de-emphasize any results based on analysis of peaks, and I would not include peaks in any models with more important exposure metrics.

Other than that, the non-use of other covarates is appropriate. For example, Years since hire is an effect modifier and not a confounder and should not be included.

- **Have the results considering the number of high intensity tasks (number of HITS) been properly weighted?**

R1: See Previous Answer.

R2: Yes, you provided a good analysis of this parameter and its effect on the risk assessment.

- **Would the consistency of the Cox regression results using continuous untransformed exposure data be reason to emphasize these results?**

R1: I believe that the models presented – Cox, Poisson, continuous exposure, exposure in deciles, models with covariates, etc. – in Tables 15 and 16 all adequately model the data. Whereas the use of the untransformed data appeals to my sense of parsimony, I also believe that there are reasonable cases to be made for the use of data lumped into deciles and other procedures that address the significant uncertainty in the exposure estimates.

R2: I think so.

The choice of response rate, 0.1% (in EC_{001} and LEC_{001}) for linear extrapolation to lower exposures.

R1: As noted above, the estimates at this level are reasonably consistent from model to model. Furthermore, this value is not far from the death rate from leukemia in the study cohort (81 in approx. 16,000). Therefore, I believe this is a reasonable starting point for extrapolation.

R2: This is a conservative approach that should be protective of public health.

R5: With respect to the uncertainty discussions: I believe that the discussion on a qualitative level is adequate. However, it might have been possible and desirable to quantitatively address some of the uncertainties, especially those related to the exposure estimation, which is always the most uncertain component of a retrospective epidemiological analysis. Although the text states that the approaches that used categorized cumulative exposure (rather than continuous cumulative exposure measures) may minimize the effects of misclassification or errors in exposure estimation, it would be nice to do some quantitative uncertainty analyses to demonstrate that. This is particularly important in light of the text descriptions of more recent data that suggest exposure estimates may not be correct (Section 4.2.5.3).

The use of a relatively low risk (.001) as the basis for the URF calculations is appropriate; when estimating a low-dose slope, one should attempt to start from a part of the curve where the risk is predominantly determined by that slope. The use of the ADAFs (though they themselves are default values) in an age-dependent, lifetable method to determine the lifetime risk associated with given exposure levels is appropriate and well described.

R7: The choice of the 0.1 % excess risk for a POD since appropriate. An alternative might be the 1% excess risk, which might be more within the range of the epidemiologic data used to estimate it.

The application of ADAFs to the slope factors using life table analysis and the BIER IV methodology (NRC 1988) to address susceptibility from early-life exposure to butadiene (Appendix 6 Calculating Excess Risk with Age-Dependent Adjustment Factors).

R1: The ADAFs appear to have bene applied appropriately, according to guidance provided by EPA and using the BIER IV methodology. That said, there was not a large difference in the potency estimate when the ADAFs were applied (3.466E-04/ppm vs. 3.505E-04).

R2: I found this to be appropriate.

R4: The application of ADAFs to the slope factors appears to have been done correctly, and is appropriate in the case of butadiene.

R7: I do not believe the application of the ADAFs to the excess risk calculation changes things much. I suppose this is standard in this type of risk assessment. I don't think there is any particularly good support for the use of quantitative ADAFs, although. I understand this is another attempt to be conservative to protect susceptible populations.

Other Issues Related to the Cancer assessment

R2: What I find missing in the document is a good discussion of the uncertainty associated with the numbers you derive for the various standards. A whole section on this subject should be added. The values determined are estimates only and there are many sources of uncertainty associated with each value. It is particularly disturbing to see values such as those listed in Table 16 with 4 significant figures! Knowing the uncertainties involved, the values might well be off by an order of magnitude. I am sure you are also aware of this problem, but there is no indication in the document of the degree of uncertainties involved. I would like to see a range of values rather than one point value, but I know that regulators want a single value. This gives the public the wrong impression of the degree of precision involved in risk assessments. If you have to give a single value, I recommend that you at least describe how certain or uncertain you are about the value.

R4: It is not clear what factors account for the much lower potency estimated by TCEQ (0.00036/ppm) as compared to the potency previously estimated by the USEPA (0.04/ppm before doubling). This difference of greater than 100-fold demands explanation. TCEQ should attempt to characterize the contribution to this discrepancy of the various differences between the two analyses:

- The use of the original Delzell cohort vs. the more recent UAB cohort update
- The use of exposure estimates from the original Delzell study vs. the more recent UAB revised exposure estimates

- The use of leukemia incidence rates for 85 years vs. the use of Texas-specific leukemia mortality and survival rates for 70 years
- The use of a linear model vs. an exponential Cox regression model

R6: The DSD includes a nice discussion in the uncertainty analysis (Section 4.2.5). However, it was sometimes difficult to determine what the bottom line was for each section regarding TCEQ's conclusions of the impact of the uncertainty discussed in each section on the overall assessment.

A discussion and consideration of the relative toxicodynamic sensitivity of children vs. adults to chemical-induced leukemia would be of value, particularly in light of the public comment stating that susceptibility to leukemia of unknown origin appears to increase between younger and older children, and between older children and adults. This database is beyond my knowledge base.

Note that the MOA is understood, and is identified as DNA reactivity and resulting mutagenicity. The **mechanism** of action is what is not known in sufficient detail to develop a BBDR.)

R7: It is not clear to me why Texas has chosen to calculate excess risk only through age 70. I note that US EPA has used 85 in its risk assessment for ethylene oxide, for example. Obviously a number of leukemia cases occur after age 70. Excess risk will increase with the use of age 85. The current life expectancy in the US is close to age 80. Note that the use of 85 at a limit for life table calculations of excess risk assumes a population which is followed through age 85, but has a life expectancy considerably less than, probably in the 70s.

References cited by Reviewers

Allen, BC, Kavlock, RJ, Kimmel, CA, and Faustman, EM. Dose-response assessment for developmental toxicology. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels. *Fundamental Applied Toxicology* 23: 487-495 (1994).

Crump, K. and Allen, B. (1987). Quantitative assessment of carcinogenic hazards using epidemiological data. Environmental Health Risks: Assessment and Management, R. Stephen McColl (ed.). University of Waterloo Press, pp. 133-158.

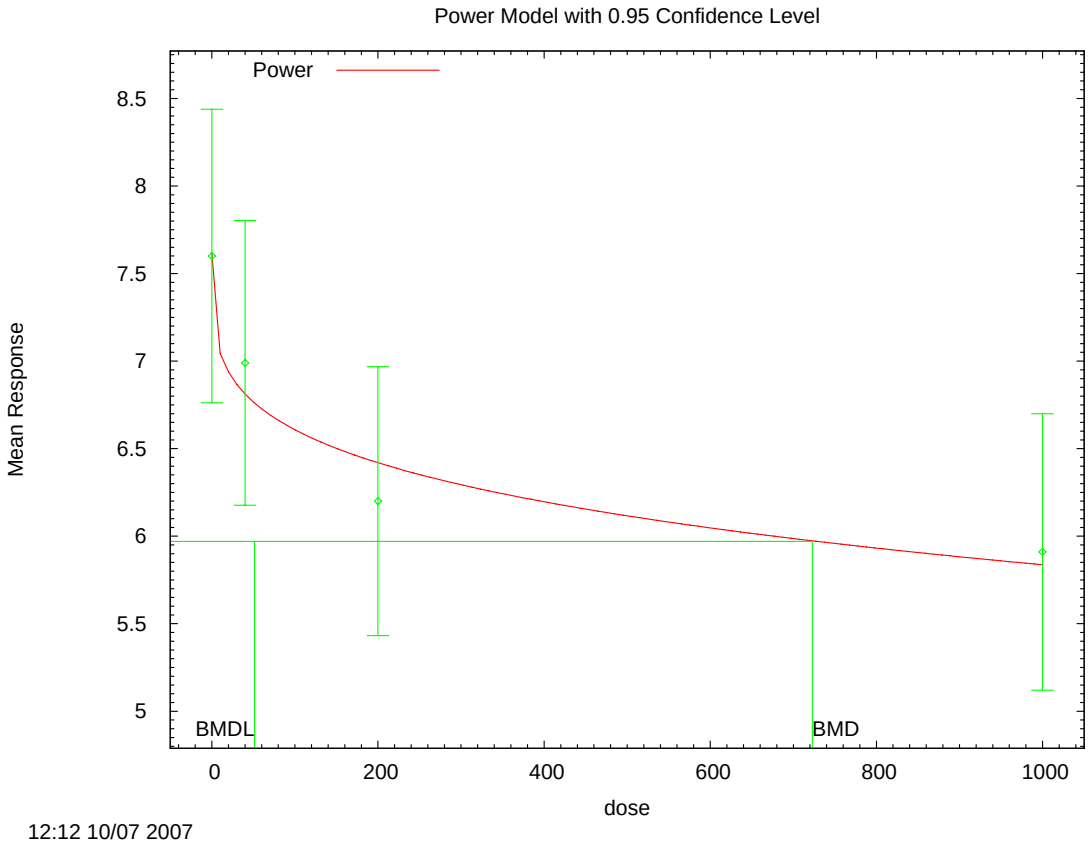
Crump, KS. Calculation of benchmark doses from continuous data. *Risk Analysis* 15: 79-89 (1995).

Fennel, T. R., Summer, S. C. J., Snyder, R. W., Burgess, J., Spicer, R., Bridson, W. E., Friedman, M. A., 2005. Metabolism and Hemoglobin Adduct Formation of Acrylamide in Humans. Toxicol. Sci. 85**,** 447-459.

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

Krewski, D, Bakshi, K, Garrett, R, Falke, E, Rusch, G, and Gaylor, D. Development of acute exposure guideline levels for airborne exposures to hazardous substances. *Regulatory Toxicology Pharmacology* 39: 184-201 (2004).

ten Berge, WF, Zwart, A, and Appelman, LM. Concentration-time mortality response relationship of irritant and systemically acting vapours and gases. *J Hazardous Materials* 13: 301-309 (1986).




```
 ======================================================================
        Power Model. (Version: 2.14;  Date: 02/20/2007)
        Input Data File: C:\USEPA\BMDS2Beta\Data\1Powtcetce.(d)
        Gnuplot Plotting File:  C:\USEPA\BMDS2Beta\Data\1Powtcetce.plt
                                              Sun Oct 07 12:12:58 2007
 ======================================================================

 BMDS Model Run
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

   The form of the response function is:

   Y[dose] = control + slope * dose^power


   Dependent variable = MEAN
   Independent variable = Dose
   rho is set to 0
   The power is not restricted
   A constant variance model is fit

   Total number of dose groups = 4
```

Total number of records with missing values = 0
Maximum number of iterations = 250
Relative Function Convergence has been set to: 1e-008
Parameter Convergence has been set to: 1e-008

Default Initial Parameter Values

alpha =	2.84578	
rho =	0	Specified
control =	7.6	
slope =	-0.189737	
power =	0.316578	

Asymptotic Correlation Matrix of Parameter Estimates

(*** The model parameter(s) -rho
have been estimated at a boundary point, or have been specified by the user,
and do not appear in the correlation matrix)

	alpha	control	slope	power
alpha	1	-1.7e-008	9.8e-009	9.5e-009
control	-1.7e-008	1	-0.64	-0.45
slope	9.8e-009	-0.64	1	0.96
power	9.5e-009	-0.45	0.96	1

Parameter Estimates

Variable	Estimate	Std. Err.	95.0% Wald Confidence Interval Lower Conf. Limit	Upper Conf. Limit
alpha	2.72201	0.43587	1.86772	3.5763
control	7.62021	0.382717	6.87009	8.37032
slope	-0.326249	0.354843	-1.02173	0.36923
power	0.246202	0.15213	-0.0519678	0.544372

Table of Data and Estimated Values of Interest

Dose	N	Obs Mean	Est Mean	Obs Std Dev	Est Std Dev	Scaled Res.
0	18	7.6	7.62	2.04	1.65	-0.052
40	19	6.99	6.81	1.66	1.65	0.473
200	21	6.2	6.42	1.74	1.65	-0.605

```
1000    20       5.91         5.83         1.25         1.65          0.209


 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)} = Sigma^2
     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          -77.734504            5     165.469009
             A2          -75.503795            8     167.007591
             A3          -77.734504            5     165.469009
         fitted          -78.053441            4     164.106882
              R          -83.514230            2     171.028460


                   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              16.0209          6         0.01364
   Test 2              4.46142          3          0.2158
   Test 3              4.46142          3          0.2158
   Test 4             0.637873          1          0.4245

The p-value for Test 1 is less than .05.  There appears to be a
difference between response and/or variances among the dose levels
It seems appropriate to model the data

The p-value for Test 2 is greater than .1.  A homogeneous variance
model appears to be appropriate here


The p-value for Test 3 is greater than .1.  The modeled variance appears
```

```
 to be appropriate here

The p-value for Test 4 is greater than .1.  The model chosen seems
to adequately describe the data


               Benchmark Dose Computation

Specified effect =             1

Risk Type        =     Estimated standard deviations from the control mean

Confidence level =          0.95

             BMD = 722.796


            BMDL = 51.3032
```