TERA

Report of the Expert Peer Review of Sulfolane Reference Doses for the Alaska Department of Environmental Conservation

Volume One

Expert Panel:

Dr. Andrew Maier (Panel Chair)

Dr. Susan Griffin

Dr. Richard Hertzberg

Dr. Michael Luster

Dr. Deborah Oudiz

Dr. Stephen Roberts

Independent

Non-Profit

Science

For Public Health Protection

Toxicology Excellence for Risk Assessment

December 18, 2014

Contact: Jacqueline Patterson
(patterson@tera.org)

This page intentionally left blank.

Volume One

Volume Two

Appendix A. Meeting Handouts (panel biographical sketches, conflict of interest information, charge questions, agenda, and additional handouts)

Appendix B. Slides from ADEC presentation

Appendix C. Slides from panel discussion on model selection

Volume Three

RfD documentation and technical comments

NOTE

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

LIST OF ACRONYMS AND ABBREVIATIONS

ADEC - Alaska Department of Environmental Conservation

ADHSS - Alaska Department of Health and Social Services

AIC - Akaike Information Criterion

AIC_C - Akaike Information Criterion Corrected

AIDS - Acquired Immune Deficiency Syndrome

ALP - Alkaline Phosphatase

ATSDR - Agency for Toxic Substances and Disease Registry

BMC – Benchmark Concentration

BMD - Benchmark Dose

BMDL - Benchmark Dose Lower Confidence Limit

$BMDL_{10}$ – Lower 95% Confidence Interval on Dose Giving a 10% Response

$BMDL_{1SD}$ – Lower 95% Confidence Interval on Dose Giving a 1 Standard Deviation Response

BMDS - Benchmark Dose Software

BMR - Benchmark Response

$BW^{3/4}$ – Body Weight Scaled to ¾

CCME - Canadian Council of Ministers of the Environment

DABT – Diplomate of the American Board of Toxicology

FATS - Fellow, Academy of Toxicological Sciences

FHRA - Flint Hills Resources Alaska

GLP - Good Laboratory Practices

HED - Human Equivalent Dose

HHRA – Human Health Risk Assessment

HLS – Huntingdon Life Sciences

IPCS – International Programme on Chemical Safety

IRIS – Integrated Risk Information System

LOAEL - Lowest Observed Adverse Effect Level

LUC – Large Unstained Cell

mg/kg – milligrams per kilogram

mg/kg-day – milligrams per kilogram per day

MOA - Mode Of Action

MRL – Minimal Risk Level

NCEA - National Center for Environmental Assessment

NHANES - National Health and Nutrition Examination Survey

NHL - Non-Hodgkin’s Lymphoma

NIH - National Institutes of Health

NOAEL - No Observed Adverse Effect Level

NTP – National Toxicology Program

PBPK – Physiologically-Based Pharmacokinetic

POD – Point Of Departure

PPRTV - Provisional Peer Reviewed Toxicity Values

p-value – probability value

RBC – Red Blood Cell

RfD - Reference Dose

SAB – Science Advisory Board

SD – Standard Deviation

TCEQ - Texas Commission on Environmental Quality

TERA - Toxicology Excellence for Risk Assessment

TOX21 – Toxicology Testing in the 21st Century

UF - Uncertainty Factor

UF_A – Uncertainty Factor for Animal to Human Extrapolation

UF_C - Composite Uncertainty Factor

UF_D - Uncertainty Factor for Database Uncertainties

UF_H - Uncertainty Factor for Intrahuman Variability

UF_L - Uncertainty Factor for LOAEL to NOAEL Extrapolation

UF_S - Uncertainty Factor for Extrapolation from a Subchronic Study

U.S. EPA – United States Environmental Protection Agency

WBC - White Blood Cell

WHO – World Health Organization

WOE - Weight of Evidence

EXECUTIVE SUMMARY

An independent expert panel met on September 16 and 17, 2014 in Fairbanks, Alaska to review and discuss human health toxicity values for sulfolane. A sulfolane reference dose (RfD)[1] will be used by the Alaska Department of Environmental Conservation (ADEC) to develop cleanup levels for sulfolane found in groundwater in North Pole, Alaska. ADEC tasked Toxicology Excellence for Risk Assessment (TERA) to organize the expert panel review. TERA selected a panel of six experts in toxicology, immunology, human health risk assessment, RfD methods and derivation, contaminated site assessments, biostatistics, and benchmark dose (BMD) modeling. The panel was asked to use their independent professional scientific judgment to evaluate the available toxicity values and to identify the most adequate RfD for consideration by ADEC.

The panel drew upon U.S. EPA risk methods and guidance for BMD modeling, as these are the commonly accepted methods used in the United States to derive RfDs for development of protective cleanup levels for contaminated sites. The panel reviewed and considered the following sulfolane toxicity values: Canadian Council of Ministers of the Environment (CCME, 2006); the Agency for Toxic Substances and Disease Registry (ATSDR, 2010, 2011); the Texas Commission on Environmental Quality (TCEQ) (Haney, 2011); the U.S. Environmental Protection Agency (EPA, 2012a); ARCADIS U.S., Inc. (Magee, 2012); Thompson et al. (2013); and Health Canada (Health Canada, 2014). In preparation for the meeting, the expert panel reviewed the documentation for each RfD, along with key references, technical comments submitted by the public, and other pertinent reports and information. Authors of several of the RfD documents attended the meeting in person or via teleconference and authors from all the organizations who had derived a toxicity value offered to answer panel questions.

At the meeting, the expert panel systematically reviewed and discussed the data, methods, and decisions for derivation of an RfD for sulfolane. They used a series of charge questions to focus their discussions and applied best science within the framework of the EPA's risk assessment methods to determine the most adequate RfD.

- Discuss the strengths and weaknesses of the key studies and the available toxicity data on sulfolane.
- Discuss the endpoints and effects seen in the toxicity studies and potential mode(s) of action.
- What are the most appropriate values for the standard uncertainty factors commonly used?
- Identify any additional scientific issues or questions that the panel should discuss.

[1] Throughout the peer review meeting and in this report all of the various toxicity values for sulfolane are referred to simply as "RfDs."

- Discuss which of the RfDs reflects the best use of the currently available data, and why.
- Discuss the overall confidence in the selected RfD(s) and what additional studies or analyses, if any, would help reduce uncertainty or increase confidence.

Available Toxicology Data

The panel discussed the available toxicology data on sulfolane to select the best study for RfD development and provide general understanding and confidence in the body of knowledge for identifying the target organ, critical effect, uncertainty factors, and overall confidence in the results. Panel members reviewed the strengths and weaknesses of the two key studies (Zhu et al., 1987; HLS, 2001) as well as inhalation, developmental, and reproductive studies for sulfolane. The panel concluded that HLS (2001) is the best study to use for quantitative dose response and calculation of an RfD. Because of limitations in the Zhu et al. (1987) study's documentation, they determined that it is not adequate for quantitative dose response purposes; however, the Zhu et al. (1987) study, as well as other systemic toxicity studies provide support and contribute confidence for the identification of the immune system as an important target for sulfolane toxicity.

Endpoints and Effects

Effects on the immune system were a common sensitive finding among the available toxicity studies and the panel focused its discussion on the adversity and relevance of the measures of immune system effect for risk assessment. The panel agreed that decreased white blood cell (WBC) count (leukopenia) and related metrics are the most appropriate effects to use as the basis for deriving the RfD for sulfolane. These endpoints meet the RfD concept of an adverse effect although there are no direct immune function testing data available for sulfolane to know the functional impacts of the WBC count changes. They discussed that there is a general correlation between decreased WBC counts and greater risk of infection and cited precedent for using this endpoint (i.e., EPA IRIS benzene assessment).

While the current RfDs took different approaches regarding dosimetric adjustments to the experimental dose, the panel noted that calculating a human equivalent dose (HED) based on allometric scaling is the current default approach used by the EPA in deriving RfDs. The panel did not find any data to suggest the HED method is inappropriate for sulfolane.

The panel discussed the use of a no (or lowest) observed adverse effect level (NOAEL or LOAEL) or a benchmark dose approach for dose response. The panel agreed that the scientific consensus in risk assessment is that the BMD approach offers advantages over using a NOAEL or LOAEL, and that the HLS (2001) data set meets the requirement for a statistically significant dose response trend and is amenable to modeling. The panel recommended that BMD modeling would better inform on the nature of the dose response curve than would use of a NOAEL or LOAEL.

Benchmark Dose Modeling

The panel discussed a number of issues related to the BMD modeling. They agreed that log transformation of dose is an appropriate approach to use based on the data sets for sulfolane. For the benchmark response (BMR) (the degree of deviation from control that is considered adverse), the panel recommended use of 1 standard deviation (SD), which is consistent with standard practice in applying EPA risk methods when there is an absence of specific data to define adverse response. The panel also preferred use of the concurrent control data as the best representation of variability under the specific experimental conditions of the study, rather than using historical control data.

The panel discussed various criteria that are used for model selection including visual fits, scaled residuals, goodness of fit p-value, the Akaike Information Criterion (AIC), and the ratio of BMD to its lower confidence bound (i.e. the BMD/BMDL ratio). EPA BMD guidance recommends choosing the model with the lowest AIC among the subset of adequately fitting models. In this case, the AIC values for the subset of well-fitting models for WBC counts ranged from 109.06 to 109.17 and panel members questioned whether this difference is meaningful. The panel was not comfortable with basing the selection of the model on the lowest AIC as an exclusive final selection criterion given the small differences in AIC values. They noted that EPA guidance is not precise regarding the use of AIC and BMD/BMDL ratios as model selection criteria and agreed that deviating from EPA's regimented rules for model selection was appropriate for sulfolane. They recommended starting with the EPA guidance, but weighing all the model fit considerations, and recommended an approach that includes taking into account additional information pertinent to model selection. The panel set forth the following steps for removing models from consideration:

1. Numerical problems. When there are clear numerical problems that cause the models to fail, this is one way to weed out data sets. An example would be data that lack any dose-response pattern.

2. Unacceptable visual and statistical fit. EPA's poor-fit criterion of a p-value of less than 0.1 is very lenient. Preferably one would like to remove additional poor fitting models (with a p-value less than 0.1), with further removals based on visual appraisal and scaled residuals.

3. Over-specified models, for example, with higher order polynomials that are otherwise statistically identical in fit to a linear model. In this case, one is not gaining anything by adding parameters. In some cases, the models reduce to a simpler model and so several models will show up as having identical fits – these should be considered the same model with selection of the simplest form. In other cases, the AIC can be used to evaluate the role that model complexity is playing in improving overall fit. Generally, if the AIC values are close in absolute value they are all of similar reliability.

4. Scaled residuals too high. The scaled residuals provide additional information on the fit in the range of most interest and can be used to discriminate among models that have similar overall global fits (i.e., similar AIC values).

The panel discussed that other secondary considerations such as the BMD/BMDL ratios can also be considered. The panel thought that the BMD/BMDL ratio is a less important factor than the other selection criteria (i.e., it is unlikely to change model selection beyond the other criteria that are routinely considered such as visual fit and scaled residuals).

The panel applied their model selection criteria to the HLS (2001) immune endpoints using the modeling results from the BMD evaluation report (Gradient, 2014) prepared for ADEC. Applying the panel's criteria to the sulfolane data resulted in selection of the linear model using log(dose) for both WBC count and lymphocyte counts. The panel recommended that modeling both endpoints is an appropriate initial step for the sulfolane data, followed by identifying which are the appropriate models and metric selections. Differences in results for the two endpoints might help inform model selection based on biology. Overall, the panel favored the selection of decrease in WBC count as the critical effect for sulfolane since it better captures potential impacts on health (it covers more of the WBC types that were affected). Based on this consideration and the model selection procedure, the panel identified the BMDL of 12.66 mg/kg-day based on the concurrent control data set for WBC decreases in the HLS (2001) study as the most appropriate point of departure (POD).

Uncertainty Factors

For uncertainty factors, the panel considered the standard five areas of uncertainty as outlined in EPA RfD methods. The panel did not think there is a compelling reason to recommend moving away from EPA's guidance regarding uncertainty factors, in light of the lack of understanding of the MOA for sulfolane. The panel recommended:

- UF_A – Allometric scaling to calculate an HED with a UF of 3 for animal to human extrapolation to account for uncertainties in toxicodynamics.
- UF_H – A factor of 10 for intrahuman variability. Data were not considered adequate to move from the default.
- UF_L –A factor of 1 is most appropriate with use of a BMDL.
- UF_S – Either 3 or 10 for extrapolation from a subchronic study. The panel members did not agree on this UF. Some thought a full 10-fold factor is appropriate, while others thought this UF should be reduced to 3 because there are data that suggest severity does not progress with duration of exposure and effects may be reversible.

- UF_D – The panel recommended using a value of 3 for database uncertainties. They noted that the quality of the available studies varied and a two-generation reproductive study is not available, but the HLS (2001) study is of high quality.

Combining the individual UF values results in a composite UF of either 300 or 1000. Panel members all agreed that a composite factor of 300 (with use of the HED) would be most appropriate. Individual panelists noted concerns with compounding precautionary decisions (e.g., the use of a BMD based on 1 SD given the unclear functional impact of the degree of observed effect and the use of the BMDL) with a composite UF of 1000. On the other hand, the panel was also concerned with the level of protection afforded by the RfD if the UF went below 300, given studies that are lacking for the sulfolane database.

RfD that Best Reflects the Currently Available Data

The panel evaluated how each of the six sulfolane RfDs aligned with the panel's conclusions and recommendations. Panel members were careful to point out that their discussion and decisions reflect the current state of risk assessment science and the available data in 2014. They recognized that the sulfolane RfDs being reviewed were developed by different organizations using their own particular methods, that each was developed for a specific purpose that may or may not match with the RfD needs of Alaska, and that each reflects the data and scientific thinking at a particular point in time. The panel emphasized that their process of evaluating these existing RfDs is not meant to infer that any of the sulfolane assessments are inadequate or inappropriate; the panel members are layering their scientific judgments on others' previously developed values for the current needs.

The panel reviewed their conclusions and preferences for key decisions to derive a chronic sulfolane RfD and compared these with the available RfDs.

- Principal Study – The panel agreed that the HLS (2001) study was the best choice for principal study and would not recommend basing an RfD upon the Zhu et al. (1987) study.
- Endpoint – The panel agreed that reduced WBC count is the biologically most appropriate endpoint to use.
- Modeling Approach – The panel agreed that BMD modeling is preferred over using a NOAEL or LOAEL. The panel recommended log transformation of the doses to model the endpoints.
- Point of Departure – The panel agreed that allometric scaling to calculate an HED is most appropriate.

Of the six RfDs only the Thompson et al. (2013), and TCEQ (Haney, 2011) RfDs matched the panel's preferences on the above parameters. However, Thompson et al. and TCEQ used historical control data in the BMD modeling and the panel preferred use of concurrent controls to calculate 1 SD from the mean for the benchmark response.

The panel observed that the composite uncertainty factor for the six RfDs ranged dramatically, from 100 to 3000. The expert panel recommended a composite UF of 300 (with HED adjustment). While only two of the RfDs used an HED with a UF of 300, several others reached fairly equivalent RfDs with larger UFs and non-adjusted PODs.

The panel found that none of the six RfDs aligned perfectly with the thought processes and recommendations of the expert panel. The panel concluded that after a close and systematic evaluation of the steps and decisions for calculation of an RfD, using EPA methods as a backdrop and the panel's scientific expert judgments, the Thompson et al. (2013) RfD most closely aligns with the panel's conclusions. The Thompson et al. (2013) RfD does not match in all aspects, in particular the panel favored use of concurrent controls over historical control data, the selection of some individual UFs differed (e.g., the panel recommends a UF of 10 for human variability), and the process for selecting the best BMD model differed (e.g., the panel explicitly weighed biology in choosing WBC counts over lymphocyte counts, while Thompson et al. based their model selection on AIC). However, from a practical standpoint, the RfDs derived by TCEQ (2011), Magee (2012), and Thompson et al. (2013) are all similar at 0.01 mg/kg-day. In addition, those RfDs that did not match as closely are within the same range when one considers the definition of an RfD (i.e., spanning an order of magnitude).

Confidence in RfD and Data Gaps

Panel members assigned low to medium confidence in the database, medium or high confidence in the principal study, and medium confidence in the RfD.

The panel discussed what additional studies or analyses, if any, would help reduce uncertainty or increase confidence in a sulfolane RfD. Individual panelists noted data gaps, including examination of bone marrow to help identify the mode of action, developmental immunotoxicity studies, hematology studies, a longer well-conducted study, data on reversibility of the effects on immune cells, toxicokinetics, and progression of the effects. The panel noted that the NTP research plans address much of what the panel identified as gaps and that additional data will help resolve some of the risk assessment interpretation issues and questions regarding the currently available data. Panel members considered that their recommendations for deriving the RfD were health-protective, but did not think they had sufficient information to judge whether new data would result in a RfD value that was lower or higher.

The panel evaluated the data, methods, and decisions for a sulfolane RfD derivation and made recommendations regarding the most scientifically appropriate decisions. The panel based its conclusions on application of the best science within the framework of EPA RfD derivation methods.

Report of Expert Peer Review of Sulfolane Reference Doses for the Alaska Department of Environmental Conservation

PARTICIPANTS

Peer Review Panel[2]

Dr. Andrew Maier, DABT, Panel Chair
University of Cincinnati, College of Medicine
Dr. Susan Griffin, DABT
U.S. Environmental Protection Agency (EPA), Denver, Colorado
Dr. Richard Hertzberg
Argonne National Laboratory and Emory University
Dr. Michael Luster
West Virginia University, School of Public Health
Dr. Deborah Oudiz
California Environmental Protection Agency, Retired
Dr. Stephen Roberts, FATS
University of Florida, Center for Environmental & Human Toxicology

TERA Staff

Ms. Jacqueline Patterson
Ms. Alison Willis

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

INTRODUCTION

The Alaska Department of Environmental Conservation (ADEC) tasked Toxicology Excellence for Risk Assessment (TERA) with conducting an independent, expert peer review of the available human health toxicity values for sulfolane[3]. A sulfolane reference dose (RfD) will be used by ADEC to develop cleanup levels for groundwater in North Pole, Alaska. TERA is an independent non-profit organization whose mission is to support the protection of public health by developing, reviewing, and communicating risk assessment values and analyses; improving risk methods through research; and educating risk assessors, managers and the public on risk assessment issues. TERA organized and conducted the expert review under contract to ERM, Alaska. TERA followed its standard peer review approach and principles for this peer review (see http://www.tera.org/Peer/index.html).

TERA independently selected and convened a panel of six experts to evaluate the scientific basis for the available RfDs for sulfolane. The independent expert panel included scientists with expertise in the key disciplines for derivation of RfDs. Each panelist is a well-respected scientist in his or her field. The sulfolane peer reviewers are recognized technical experts who were selected for their relevant scientific technical knowledge and independence. Collectively, the panel has expertise in toxicology, immunology, human health risk assessment, RfD methods and derivation, contaminated site assessments, biostatistics, and benchmark dose (BMD) modeling. The experts have background and experience with government, university, industry, and non-profit sectors. TERA screened each expert for potential conflicts of interest and every effort was made to avoid conflicts of interest and biases that would prevent a panel member from giving an independent opinion on the subject. TERA was solely responsible for the selection of the panel members. The experts served as individual scientists and represented their own personal scientific opinions. They did not represent their companies, agencies, funding organizations, or other entities with whom they are associated. Short biographical sketches and conflict of interest information for panel members are provided in Appendix A.

The peer reviewers were asked to use their independent professional scientific judgment to evaluate the reference doses and to identify the most adequate RfD for consideration by ADEC. The peer review involved an in-depth assessment of the assumptions, calculations, alternate interpretations, methodology, and conclusions of the material under review. Review materials include the toxicity values (i.e., RfDs) that were derived by a number of organizations; documentation for these RfDs is available on the meeting website (http://www.tera.org/Peer/sulfolane/index.html). The panel drew upon U.S. EPA risk methods and guidance for BMD modeling, as these are the commonly accepted methods used in the United States to derive RfDs for development of protective cleanup levels for contaminated sites.

[3] Throughout the peer review meeting and in this report all of the various toxicity values for sulfolane are referred to simply as "RfDs."

The panel reviewed and considered the following RfDs:

- Canadian Council of Ministers of the Environment (CCME). 2006. "Canadian Environmental Quality Guidelines for Sulfolane: Water and Soil (Scientific Supporting Document)." PN 1368.

- Agency for Toxic Substances and Disease Registry (ATSDR). 2010. "Health Consultation: Sulfolane." February 3.

- ATSDR. 2011. "Health Consultation: Sulfolane." May 2.

- Haney, J. [Texas Commission on Environmental Quality (TCEQ)]. 2011. Sulfolane (CASRN 126-33-0) [re: Update of March 9, 2011 toxicity factor documentation with a slightly revised benchmark dose (BMD)]. September 6.

- United States Environmental Protection Agency (EPA). 2012. "Provisional Peer Reviewed Toxicity Values for Sulfolane (CAS No. 126-33-0)." National Center for Environmental Assessment (NCEA), Superfund Health Risk Technical Support Center, January 30.

- Magee, B. [ARCADIS U.S., Inc.]. 2012. Memorandum to Flint Hills Resources Alaska re: Assessment of dose response information for sulfolane. May 21.

- Thompson, CM; Gaylor, DW; Tachovsky, JA; Perry, C; Carakostas, MC; Haws, LC. 2013. "Development of a chronic noncancer oral reference dose and drinking water screening level for sulfolane using benchmark dose modeling." J. Appl. Toxicol. 33(12):1395-1406.

- Health Canada. 2014. "Drinking Water Guidance Value for Sulfolane." March 17.

In preparation for the meeting, the expert panel reviewed the documentation for the RfDs, key references (e.g., Zhu et al., 1987; HLS, 2001), and other pertinent information. This included the background document prepared by ADEC (ADEC, 2014) and a BMD report prepared for ADEC by Gradient Corporation (Gradient). In addition, interested parties were invited to submit technical comments for the panel's consideration. The following technical comments were received by TERA and are available at http://www.tera.org/Peer/sulfolane/index.html.

- Submitted by Dr. Laurie Haws on 9/08/14
  - ToxStrategies Comments TERA Sulfolane Peer Review_090814_2_submitted
- Submitted by Dr. David Gaylor on 9/08/14
  - Gaylor_Sulfolane_Gradient Report_2_submitted
- Submitted by Dr. Brian Magee on 9/08/14

    - ARCADIS (2012a) FHRA_NPR_Revised Final Draft HHRA Posted
    - ARCADIS Memorandum_Response to Gradient Report_20140908
    - ARCADIS (2012a) Revised Draft Final Human Health Risk Assessment (Excerpts from main document and relevant portion of Appendix H)
- Submitted by Dr. William Farland on 9/08/14
    - Farland_Coupling Exposure to the RfD_9814
- Submitted by Chad Thompson on 9/08/14
    - Supplemental materials for the Thompson et al. (2013) manuscript were provided to panel members. These are available online from the publisher's website.
- Submitted by Dr. William Farland on 8/25/14
    - Sulfolane Hazard Characterization – Considerations, William H. Farland, Ph.D., ATS, April 5, 2012
    - Perspectives on the Journal of Applied Toxicology Article entitled "Development of a chronic non-cancer oral reference dose and drinking water screening level for sulfolane using benchmark dose modeling", William H. Farland, PhD, ATS, May 30, 2014
- Submitted by Mr. David Smith on 8/26/14
    - ARCADIS (2014) Supplement to the Revised Draft Final Human Health Risk Assessment

Copies of the listed RfD documentation and submitted technical comments are found in Volume 3 of this report.

TERA provided the panel with a list of key questions (the "charge to peer reviewers") to help focus their review and discussions. The charge questions are briefly described below. A copy of the full charge is found in Appendix A, along with other meeting handouts.

- Discuss the strengths and weaknesses of the key studies and the available toxicity data on sulfolane.

- Discuss the endpoints and effects seen in the toxicity studies and potential mode(s) of action.

- What are the most appropriate values for the standard uncertainty factors commonly used?

- Identify any additional scientific issues or questions that the panel should discuss.

- Discuss which of the RfDs reflects the best use of the currently available data, and why.

- Discuss the overall confidence in the selected RfD(s) and what additional studies or analyses, if any, would help reduce uncertainty or increase confidence.

The meeting opened with a welcome by Ms. Jacqueline Patterson of TERA. She described the background and purpose of the expert review and reviewed the agenda. The panel members then introduced themselves and noted whether they had additions or changes in the biographies or conflict of interest information. None of the panel members had any questions regarding another's conflict of interest information or changes to their own statements.

Dr. Andrew Maier, the panel chair, then described how the meeting would be conducted. He explained that discussions would be organized around the charge questions and would follow the order in the agenda (see Appendix A). He noted that panelists were expected to share their scientific opinions on the discussion questions and panel members were encouraged to question one another to make sure that they understand the scientific basis for one another's opinions. The panel was asked to seek consensus, but if agreement was not possible, the meeting report would note this. He explained that the ADEC representatives would make a brief presentation on the sulfolane contamination and answer clarifying questions from the panel. The attending sulfolane RfD authors would then be invited to make brief comments or clarifications regarding their RfDs and answer clarifying questions from the panel. Dr. Maier noted that the panel members may ask clarifying questions of the RfD authors present during the meeting and that the authors would be given periodic opportunities to ask their own clarifying questions of the panel.

TERA drafted this meeting report to provide a summary of the expert panel's discussions and conclusions, and to serve as the official record of the expert review. The draft report was reviewed and revised by the panel members and the final report was approved by the panel. The meeting report is a summary, not a transcript of the discussions. Opinions and comments of panel members are summarized to describe the scope and breadth of the discussions. Individual panelist comments are not identified by name, as it is the conclusions of the panel as a whole that is the value of a peer review meeting. This final report reflects the panel's final opinion and conclusions.

PRESENTATIONS

Alaska Department of Environmental Conservation

Dr. Tamara Cardona of the ADEC, Division of Spill Prevention and Response, Contaminated Sites Program and Ms. Stephanie Buss, an ADEC contractor, provided a brief introduction on the sulfolane contamination situation and available RfDs. Slides of the presentation can be found in Appendix B. Their presentation summarized information contained in the background document prepared by ADEC and distributed to the panel (ADEC, 2014). Dr. Cardona began the presentation with an overview of sulfolane chemical physical properties and noted that sulfolane is an industrial solvent used during gasoline production to separate aromatic compounds from hydrocarbon mixtures and to purify natural gas. Sulfolane has low vapor pressure, is highly soluble in water, and is not well absorbed through the skin. It was discovered off the refinery boundary in private drinking water wells in North Pole, Alaska in 2009; the current sulfolane plume is approximately 2.5 miles wide by 3 miles long. Alternative water supplies have been provided to affected area residents. Dr. Cardona explained that an RfD is a key component in ADEC's calculations to determine a cleanup level. Because there are a number of RfDs available, and to ensure the most scientifically sound groundwater cleanup level for sulfolane is used, ADEC is seeking the panel's expert, independent recommendation on the most appropriate chronic oral RfD for use in the cleanup level.

Ms. Buss briefly described the available key studies for use in RfD derivation. Zhu et al. (1987) includes a 6-month study in guinea pigs that identified hepatic effects, changes in bone marrow cell counts, and dispersion in the white pulp of the spleen. The identified no effect level was 0.25 mg/kg-day. Huntingdon Life Sciences (HLS, 2001) conducted a 13-week study in rats that identified reduction in lymphocytes, monocytes, and LUC counts in females. The identified no observed effect level was 2.9 mg/kg-day. More detailed information on each study was provided in the ADEC report (2014). Ms. Buss presented key information on the toxicity values that have been developed by various agencies and authors. The values span about an order of magnitude and differ in choice of principal study, choice of critical effect, modeling approach, and application of uncertainty factors. She explained that ADEC tasked Gradient Corporation (Gradient) with verification of the benchmark dose modeling that some of the groups used for their values. Their report (Gradient, 2014) was provided to the panel.

In response to clarifying questions from the panel, the ADEC representatives explained that no clean up goal is currently in place and at the current time the refinery is providing alternative water supplies when there is any detection of sulfolane in a private well. They also noted that the State usually follows EPA's hierarchy for selection of toxicity values, with IRIS values at the top of the list, followed by PPRTVs (Provisional Peer Reviewed Toxicity Values) and then other peer-reviewed sources.

RfD Authors

ADEC invited authors of the RfD documents to attend the meeting in person or via teleconference. Several authors were able to attend and the other authors and organizations who had derived a toxicity values offered to answer panel questions. Following the presentation from the ADEC, attending authors for the RfDs were invited to share brief clarifying comments or questions. The panel members were also given the opportunity to ask the RfD authors clarifying questions. Scientists representing four of the RfDs attended the meeting and provided the following comments.

Agency for Toxic Substance and Disease Registry

Mr. James Durant of the Agency for Toxic Substances and Disease Registry (ATSDR) explained that ATSDR prepared an initial health consultation in 2010 (ATSDR, 2010) and a revised document in 2011 (ATSDR, 2011) for the Alaska Department of Health and Social Services

(ADHSS) for use at the North Pole site. The main purpose of the health consultation was to provide ADHSS with human exposure limits (provisional minimal risk level or MRL) for drinking water consumption as part of the public health assessment. Mr. Durant noted that the process for public health assessments diverges in key and important aspects from risk assessment.

A panel member asked if ATSDR were to revise this assessment today, whether they would change any of their decisions. Mr. Durant responded that notwithstanding any new information in the toxicology database, their decisions would not have changed and their conclusions (2011) would remain the same. Another panel member asked why the guinea pig was chosen over the rat, and the authors responded that the point of departure (POD) was lower for the guinea pig both based on the no observed adverse effect level (NOAEL) and lowest observed adverse effect level (LOAEL) for the 2010 document, and based on the BMD analysis in 2011. Mr. Durant explained that the 2011 document is the more current Agency recommendation; it has been reviewed both inside and outside of ATSDR and reflects more information than was available in 2010. Another panel member asked why the selection of uncertainty factors (UF) went from 100 in the 2010 document to 1000 in 2011. Mr. Durant explained that the initial 2010 document was put together as an emergency response request and was in essence a preliminary report with a provisional value. For the 2011 report, ATSDR brought together a larger workgroup that decided the newly available HLS (2001) study was a better choice of study and determined that a UF of 1000 was most appropriate with the use of that study.[4]

U. S. Environmental Protection Agency

Dr. Scott Wesselkamper of the U.S. Environmental Protection Agency briefly described the PPRTV process and development of the sulfolane PPRTV. Through EPA’s Superfund Program, the National Center for Environmental Assessment (NCEA) derives PPRTVs for use when no toxicity values are available in the EPA’s Integrated Risk Information System (IRIS). IRIS is the first tier choice for toxicity values in the Superfund program and PPRTVs are the second tier. PPRTVs are derived similarly to risk values on IRIS based on available data and IRIS assessment methods. They are internally peer reviewed by a panel of NCEA scientists, and subsequently undergo an external letter peer review by three independent experts. In contrast to IRIS, the PPRTVs do not have an EPA multi-program consensus review because of the typical need for faster turnaround for use in the field. The sulfolane PPRTV was initiated in the fall of 2010, and was finalized in the winter of 2012 (January).

A panel member asked why EPA did not transform the dose to a log scale for use in BMD modeling. Dr. Wesselkamper explained that log transformation of dose was not a common practice for PPRTVs or within EPA at the time, and if they were to redo the PPRTV today, they

[4] ATSDR staff clarified after the meeting that the 2010 provisional value was for subchronic/intermediate exposure, and the 2011 provisional MRL was for chronic exposure.

would still not use log transformation. Another panelist asked about lack of adjustment to a Human Equivalent Dose (HED) and Dr. Wesselkamper explained a programmatic decision was made to begin implementation of EPA's HED recommendations within PPRTV assessments with those submitted for internal clearance after September 30, 2011. The HED recommendations were not implemented in the sulfolane PPRTV assessment because it was submitted for internal clearance prior to that date. If EPA were to do the sulfolane PPRTV today, he said they would have applied an HED.

One panel member asked why the Zhu et al. (1987) study was excluded. Dr. Dan Petersen of EPA replied that the Zhu study contains some useful data for weight of evidence (WOE), but that all the available studies were deficient in some way. Both Zhu et al. (1987) and Ministry of Health and Welfare Japan (1999) had reporting deficiencies and were less than ideal for risk assessment. The HLS (2001) study had more complete reporting, lower doses, and a more complete endpoint analysis. Zhu et al. (1987) was used to support the effects found in the HLS (2001) study (splenic, WBC suppression in second species), and corroborated the data found in the HLS (2001) study. But due to missing details, the Zhu et al. (1987) study was not considered the best study to identify health effects.

ARCADIS

Dr. Brian Magee explained that ARCADIS is a contractor for Flint Hills Resources and that he developed the RfD for sulfolane that was included in the Human Health Risk Assessment (HHRA) ARCADIS prepared for Flint Hills (ARCADIS, 2012) and the supplement to the HHRA in 2014(a). Dr. Magee clarified several points that may have been misunderstood based on the Gradient (2014) report. Dr. Magee explained that he independently identified key studies and the relevant data and did his own BMD modeling on multiple data sets to derive the RfD included in the HHRA. He also explained that in his 2010 BMD modeling only four models gave acceptable fits, but that with BMD modifications and software updates, the current version would result in five models.

Dr. Magee also noted that he thought that the sulfolane database is robust, and that the HLS (2001) study was well planned, executed, and documented – which is why he (he also noted EPA as well) chose it as the study upon which to base the RfD. He noted that the HLS (2001) study was superior to the Zhu et al. (1987) study, which was essentially an "extended abstract," with no necessary documentation provided. He used BMD modeling as this is the current state of the science for RfD derivation (as evidenced by 68 chronic RfDs on IRIS derived with BMD rather than NOAELs or LOAELs). He noted that log dose transformation is explicitly recommended by EPA, has precedence in published literature (see example of EPA's benzene assessment), and other sulfolane RfD authors also chose to transform the doses. Lastly, he noted the question of biological significance of the statistically significant decrease in circulating white blood cells (WBC) observed in female rats with no evidence of an adverse effect on immune cells. He noted that WBC counts are naturally variable and this may be a reversible adaptive response. When deriving an RfD for benzene based on the same endpoint (decreases in the numbers of circulating white blood cells), EPA (2003) reached the same conclusions that the numbers of circulating white blood cells are naturally variable and that decreases in white blood cells are not necessarily adverse effects. Specifically, EPA (2003) stated with regard to benzene that "there is no evidence that it [decrease in absolute lymphocyte count] is related to any functional impairment at levels of decrement near the benchmark response." Despite no evidence that the decreased WBC counts are adverse, he and other sulfolane RfD authors identified 1 standard deviation (SD) from the mean for the benchmark response (BMR) used in the modeling. He considered this a prudent approach, which he took into account when assigning UFs.

The panel sought clarification on a number of questions. The ARCADIS RfD did not use an HED and Dr. Magee was asked if he would use allometric scaling if he were to redo the RfD today. Dr. Magee explained that at the time he derived his RfD he tried to keep it simple, and followed the PPRTV lead, but if he were to redo it today, he would calculate an HED. A panelist noted that HLS (2001) study authors indicated no evidence of impaired immune function and asked if Dr. Magee found immune function test results in the HLS (2001) report (the panelist did not find evidence that the HLS (2001) study conducted immune function tests). Dr. Magee responded that he relied on the conclusion of the HLS (2001) report, and did not drill down into the results for that information.

In response to a panel question on uncertainty factors, Dr. Magee explained that in applying uncertainty factors, he recommended that the panel focus on the total uncertainty factor and make sure the total composite uncertainty factor covered all areas, while considering the potential for double counting. He was comfortable with a total UF of 1000 and noted that 91% of RfDs on IRIS derived since the 1990s have UFs of 1000 or less, and that it is rare to have a total UF over 1000 (EPA, 2014). He also explained that they felt a total UF of 1000 was appropriate given there is no clear evidence that the leukopenia is adverse and therefore the total UF should account for this, perhaps dropping a 10 to 3 for one of the factors. He noted that the EPA (2003) called leukopenia a "sentinel effect" for benzene and concluded that there was no evidence that the leukopenia was adverse and also that leukopenia "was not very serious in and of itself."

When asked what he would change if he redid this assessment today, Dr. Magee said that they would: use the current BMD model, including new guidance on the ratio of Benchmark Dose (BMD) to Benchmark Dose Lower Confidence Limit (BMDL) (BMD/BMDL); they may not do BMDL averaging based on the new guidance; and they would calculate an HED.

Thompson et al.

Dr. Chad Thompson of ToxStrategies noted that Flint Hills supported his and Dr. Haws attendance at the peer review meeting, as well as the work to prepare the Thompson et al. publication, but Flint Hills had no role in study design, data collection, data analysis, data interpretation, decision to publish, or preparation of the manuscript. To follow up on the presentation by ADEC, he emphasized that the question of route of administration in the Zhu et al. (1987) study is significant as it appears they used gavage dosing, but because sulfolane is highly soluble in water, drinking water is a more relevant route of exposure.

Dr. Thompson thought that the BMD modeling analysis done by Gradient was very thorough and that they laid out a reasonable approach that was generally consistent with EPA guidance and current practice, including use of the benchmark dose software (BMDS) Wizard. Dr. Thompson observed, as the Gradient report notes, the Wizard computed the BMD/BMDL ratios for all models and found all were less than 5; however, the Wizard did not consider that the BMD and BMDL values were in log space and should be transformed back to an arithmetic dose and space prior to computing the ratios. He noted that if Gradient had used the BMD/BMDL criteria (in arithmetic space), the BMDLs ranged from 10.77 to 16.1. Dr. Thompson also noted that he and his co-authors used allometric scaling with a corresponding reduction of the uncertainty factor for extrapolation from animals to humans (UF_A) (resulting in a 1200-fold adjustment) and that this is best practice when no physiologically based pharmacokinetic (PBPK) model is available, and is the reason their total UF is smaller than most of the others, although the overall reduction is similar. He also noted that the questions regarding relevance and adversity of the endpoint influenced their decision on uncertainty factors. And lastly he noted that EPA's BMD guidance indicates use of historical control SD data can be used and they did so because they thought the historical control data provides a more robust estimate of variability. The historical control hematology data were from over 300 rats of the same sex, strain, laboratory, and timeframe, making them very comparable to the study animals. Dr. Thompson provided the historical control data to TERA prior to the peer review meeting for the panel's consideration.

A panel member asked Dr. Thompson to further explain the rationale for the UF for intrahuman variability (UF_H) of 3 in the Thompson et al. paper, noting that none of the other RfDs used a similar value for this UF. Dr. Thompson explained that they looked at the data from different perspectives and felt that they essentially had a sensitive subpopulation and an effect that the HLS (2001) study authors did not think was clearly adverse. With the allometric scaling, the Thompson et al. authors thought it appropriate to reduce this factor, otherwise one ends up with an extremely large UF. A larger UF did not seem reasonable given the use of BMD modeling, allometric scaling, and the use of best practices in risk assessment. In response to a panel question, the Thompson et al. authors indicated that they are not aware of any human data on sulfolane.

A panel member noted that allometric scaling is used for chemicals where the parent chemical is toxic and accounts for the pharmacokinetics, but not pharmacodynamics, and hence a UF of 3 is generally applied. Dr. Thompson explained that his understanding is that the scaling accounts for systemic versus portal of entry effects and said that the issue is currently under debate, noting that the TCEQ guidance that underwent a peer review organized by TERA, recommends a factor of 1. He said that not just pharmacokinetics, but overall toxicity in general scales across species, and that EPA cancer assessments do allometric scaling before cancer slope factor derivation. He noted that EPA in their guidance (EPA, 2011) considered harmonization of cancer and non-cancer risk methods and indicated that the $BW^{3/4}$ scaling accounts for species differences in both dynamics and kinetics. Although EPA considered values ranging from 1 to 10, EPA ultimately decided that a default value of 3 should be used for the interspecies UF_A.

In response to additional panel questions Dr. Thompson clarified that they thought historical control data were better to use for the standard deviation because this effect was not defined as adverse and there was a need for additional data on normal variability of this endpoint. He also explained that they considered the BMD/BMDL ratio to be an aid in final selection of the POD and noted it is part of EPA's Wizard software.

PANEL DISCUSSION

The expert panel systematically reviewed and discussed the data, methods, and decisions for derivation of an RfD for sulfolane. They used a series of charge questions to focus their discussions and applied best science within the framework of the EPA's risk assessment methods. Below is a summary of their discussions organized around the charge questions.

Charge Question 1: Strengths and weaknesses of the key studies and the available toxicity data

Charge Question 1: The subject RfDs selected HLS (2001) or Zhu et al. (1987) as the principal study. Discuss the strengths and weaknesses of the key studies and the available toxicity data on sulfolane. Are there additional relevant references that should be considered for the RfD and if so, explain the reasoning for considering them.

The panel discussed the available toxicology data on sulfolane to select the best study for RfD development and provide general understanding and confidence in the body of knowledge for identifying the target organ, critical effect, uncertainty factor, and overall confidence in the results. Panel members reviewed the strengths and weaknesses of the two key studies (Zhu et al., 1987; HLS, 2001) as well as discussed several inhalation studies and developmental and reproductive studies for sulfolane.

A panel member thought that the ADEC background document captured well the strengths and weaknesses of the two key studies. Panel members noted that the HLS (2001) study was compliant with Good Laboratory Practices (GLP), used a standard design for a subchronic systemic toxicity study, and used the most relevant route of exposure (drinking water). While noting that the HLS (2001) was a reliable and well-conducted study, another panel member noted limitations in the study design for assessing effects relevant to sulfolane, including that the study was of subchronic duration, evaluated effects only in rats, and did not appear to include a full evaluation of immunological effects. Overall, the panel agreed that the study was sufficiently well conducted to serve as a reliable study for developing an RfD.

The panel also considered the Zhu et al. (1987) study as a candidate study for derivation of an RfD. Strengths of the study included multiple species and durations of sulfolane administration by the oral route. The panel considered the reliability of the study to be limited because basic information on experimental design and statistical analysis of results was lacking. The deficiencies were such, that one expert would have considered the paper unacceptable for publication. In addition to the lack of information on the study design, the presentation of the results was not sufficiently clear to be able to use this as a critical study. The panel agreed that because of limitations in documentation, the Zhu et al. (1987) study is not sufficiently reliable for quantitative dose-response purposes. The panel did note that the Zhu et al. (1987) study adds to the weight of evidence that the immune system effects from sulfolane exposure are treatment related (i.e., it included spleen histopathology and bone marrow cellularity). The study also highlighted questions related to species sensitivity (guinea pigs vs. rats) that need additional consideration in the evaluation of uncertainty factors.

The panel discussed other available studies, which can provide insight for mode of action (MOA) and species sensitivity, as well as understanding and confidence for target organ and critical effect selection. No studies beyond those listed in the RfD documents were identified by the panel. The panel noted that given the limitations in the data set one strategy is to use data on other related compounds for read across [5]to sulfolane. However, it was noted that data for similar chemicals were limited and such an approach was discussed for assessment of carcinogenicity in Thompson et al. (2013). The panel noted they did not find any relevant studies for related chemicals with a leukopenia endpoint that could be used for read across. One panelist indicated that in doing a search for chemicals with risk assessments based on immunotoxicity, only benzene was identified as potentially informative for sulfolane. Dinitrotoluenes were also identified, but given the differences in chemical structure, dinitrotoluenes would not provide a useful analogy.

In reviewing other sulfolane toxicology studies, panel members noted that the developmental and reproductive studies are limited. Zhu et al. (1987) showed effects at relatively high doses but did not provide documentation of exact dosing methods that were used. Doses used in the one reproductive study (Ministry of Health and Welfare Japan, 1999), were higher than those doses showing systemic effects in the subchronic studies. The panel noted that although BMD modeling has been done for reproductive and developmental effects, the resulting values were all significantly higher than POD candidate values from the systemic toxicity studies. Overall, the

[5] EPA defines “read across” as “a technique of filling data gaps. To ‘read across’ is to apply data from a tested chemical for a particular property or effect (cancer, reproductive toxicity, etc.) to a similar untested chemical.” (www.epa.gov/pesticides/science/comptox-glossary.html#r).

panel concluded that the available reproductive and developmental toxicity studies would not provide the most appropriate POD for an RfD derivation.

The panel also considered the inhalation study by Andersen et al. (1977). The panel charge was to consider the derivation of the oral RfD, but one panel member asked whether inhalation of sulfolane during showering would be a concern. Another panelist noted that the showering exposure pathway is limited for chemicals with low volatility; however, the inhalation studies can provide insights into how a chemical is acting in the body and thus inform the conclusions drawn from the available oral dosing data. Panel members noted that the Andersen study has limitations for use in deriving an RfD due to limitations in details of methodology, absence of differential cell counts - only RBC and WBC counts were determined, and lack of information on animal surveillance (especially since non-human primates were involved). Others noted that the Andersen study is better documented than the Zhu et al. (1987) study, but without toxicokinetic models to develop an internal dose, it would not be appropriate to extrapolate to an oral exposure for a quantitative dose-response assessment. The panel agreed that the inhalation study would not be an appropriate choice for derivation of a sulfolane oral RfD, but the findings provide confidence that the immune system is a target for sulfolane toxicity.

Based on the discussion of the available studies, the panel agreed that the HLS (2001) study was the most appropriate choice for developing the dose-response assessment. The other available systemic toxicity studies provide support for the identification of the immune system as an important target for sulfolane toxicity. Several panel members noted that the current RfD derivations could have more fully used the whole of the database to highlight this observation.

Although the HLS (2001) study was preferred as the critical study, one panelist noted that the study report was not explicitly clear about what histopathology was done on immune system organs. The panelist could find only one mention of spleen cell histopathology (page 67 of Volume 1), where it notes that one animal at the high dose had appearance of hematopoiesis occurring. The panelist explained that the spleen is a compensatory organ for hematopoiesis in the rodent and when toxicity is seen in bone marrow, one should look for increased hematopoiesis in the spleen. The panelist explained that there have been changes since 2000 in how the immune system is assessed histopathologically in these types of studies; today, the usual practice is to conduct extended immune system pathology and include more examination of lymphoid organs. When the HLS (2001) study was conducted, very few laboratories were doing this additional analysis as part of a standard study. If the laboratory had done immune organ histopathology, it could have confirmed if effects were at the bone marrow level – the most plausible explanation of the reported findings. Some of the questions around the decreased WBC count as an adverse effect might also have been addressed with this additional information.

The panel considered the available data and concluded that HLS (2001) is the best study to use for quantitative dose response and calculation of an RfD. Because of limitations in the Zhu et al. (1987) study documentation, it is not adequate for quantitative dose response purposes;

however, the Zhu et al. (1987) study, as well as the study by Andersen et al. (1977), contribute confidence in the immune system effects as the critical effect for sulfolane. There are data gaps in the available studies for sulfolane that were considered in the uncertainty factor discussion.

Charge Question 2a. Endpoints and effects, potential mode(s) of action

Charge Question 2a. Discuss the endpoints and effects seen in the toxicity studies and potential mode(s) of action.

Since effects on the immune system appeared to be a common sensitive finding among several available toxicity studies, the panel discussed the relevance of the measures of immune system effect for risk assessment. The discussion opened with a description of the basis for such effects and their relevance to human health. A panelist discussed the decreased WBC counts in the toxicology studies and whether the effects seen are adverse. The panelist believes that leukopenia (decrease in number of white blood cells) is an adverse effect; noting that while leukopenia is not a disease *per se*, it is a risk factor for clinically-relevant diseases, most often associated with infectious disease as well as certain types of cancers such as Non-Hodgkin's Lymphoma (NHL). In humans, a decrease in WBC count is associated with increased infections and the National Institutes of Health (NIH) AIDS multicenter study is the best quality data to inform this question (see review by Luster et al., 2005; Shearer et al., 2000). In one study, it was found that when WBC counts dropped to 25% of normal (75% reduction), patients had increased incidence of opportunistic infections like tuberculosis. If the patients' WBC counts did not increase, their survival rates were low.

The data are less certain about risk of infection with lesser reductions of WBC counts. Long-term monitoring of patients with stem cell transplantation show that those patients get more infections from common pathogens when WBC counts are reduced by more than 50% (Storek et al., 2000). For example, when neutrophil counts drop below 50% in patients undergoing chemotherapy they are usually also considered to have an increased risk for developing infections.

Data on risk of infection from smaller decreases of WBC counts (in the 20-25% reduction range) are not as clear. NHANES (National Health and Nutrition Examination Survey) data show that in the general population WBC counts range considerably (Cheng et al., 2004). People at the low end of the range (approximately 25% reduction from the mean) are more susceptible to reinfection from latent viral infections (e.g., herpes) (Parks et al., 2007) or when under high stress (e.g., Alzheimer caregivers) (Yang and Glaser, 2000; Kiecolt-Glaser et al., 1987). These data are not robust, but do support the conclusion that smaller decrements in WBC count can increase the incidence of adverse clinical effects. The panelist also noted that there are studies showing increased incidence of certain types of tumors (e.g., non-Hodgkin's Lymphoma) and leukopenia occurring in individuals with moderate levels if immunosuppression (see review by Luster et al., 2005; Penn, 2000).

Quantifying what percent reduction of WBC count is adverse is difficult with the available data. A panelist noted that the HLS (2001) study with its maximum 25-30% reduced WBC count did not report significant incidences of animals becoming ill with infections, but the laboratory was a very clean facility and it is likely that infections would only be seen if severely immunosuppressed animals were used under those circumstances. The panelist also noted that there is some precedence for considering leukopenia as an appropriate effect for risk assessment in the U.S. EPA's IRIS file for benzene (EPA, 2003).

Panelists discussed how the definition of "adverse" for purposes of RfD development has been the subject of debate for many years. In the context of RfD derivation "adverse" has been expanded to include immediate precursor effects; a precursor is directly linked to an adverse effect, but may not be clinically adverse itself. Another panelist noted that adverse effects can be identified as effects that inhibit the animal's ability to withstand additional challenge. These aspects of "adversity" support leukopenia as an appropriate basis as a critical effect.

Since the panel considered the role of precursor effects in the determination of adversity a panelist asked whether the pattern of effects seen among the studies provided mode of action understanding that might clarify the relationship among immune system effects. The panel briefly discussed possible modes of action for the immunotoxicology effects from sulfolane. One panel member suggested that sulfolane is targeting both myloid and lymphoid precursor cells, which are usually seen at the level of the bone marrow stem cell. The panelist explained that the bone marrow is the primary site for the formation of all blood cells in adults, including white blood cells. WBCs enter the spleen (a secondary lymphoid organ) where lymphocytes can proliferate in response to foreign materials such as pathogenic bacteria. Most WBCs live for only 20-30 days or less but they are continuously renewed from the bone marrow. Any loss in the ability to produce WBCs has the potential to decrease the body's ability to respond to foreign materials including infectious diseases. The Zhu et al. (1987) study showed decreased white cell dispersion in the spleen, which would be consistent with the decrease in circulating WBC counts. The spleen effects and decreased bone marrow cellularity reported by Zhu et al. (1987) are concordant with sulfolane targeting bone marrow and the observed decrease in WBCs. Another panel member asked if the rapid nature of the effects reported in the Andersen et al. (1977) study suggested the possibility of a direct effect on circulating WBCs. However, the first panelist indicated that there is no direct biological basis for such unique targeting, other than at the level of the bone marrow or WBC maturation processes. This panelist further clarified that although the effect is rapid, this would not suggest a direct effect on circulating WBCs. The panelist noted that WBCs are immediately released from the bone marrow after being formed and that sulfolane being directly cytolytic to WBCs in circulation is not very plausible. Based on the mode of action hypotheses one panel member thought that the decrease in WBCs would be expected to be observed at the same or lower doses than the spleen effects, and therefore use of the decrease in WBC count as a critical effect is appropriately sensitive to protect from other downstream effects, for example the spleen effects seen n the Zhu et al. (1987) study. Other panelists thought that this line of reasoning gives the suggestion that the leukopenia effect would be an appropriate early effect for risk assessment. Another panelist explained that if HLS had looked at bone marrow histology, one would have expected effects related to bone marrow cellularityto be seen.

One panelist noted that leukopenia has been seen in multiple sulfolane studies and with multiple routes of administration. What is not demonstrated is a progression of effects with increasing dose, due to the limited experimental data on doses above 35 mg/kg-day and greater than 250 mg/kg-day. Decreases in WBCs (which are not considered to be serious adverse effects) started to occur around the higher dose groups in the Zhu et al. (1987) 6-month study and the HLS (2001) 13-week study. Filling this gap with more sensitive indicators such as bone marrow histopathology might provide a better understanding of how this effect is progressing and provide clarity regarding adversity. This panelist thought it is prudent to treat leukopenia as a precursor effect for immunotoxicity. Another panelist observed that the degree of change seen in the experimental animal studies (e.g., 54.5% reduction in leukocytes [HLS, 2001, page 41]) is roughly equivalent to the decrement in WBC counts in human populations that show an increase in infections. Another noted that the response seemed to get monotonic at higher doses; Andersen et al. (1977) saw a dramatic decrease (77.5% [Andersen et al., 1977, Table 2, page 467]) but only one dose was tested. The latter panelist did not give a lot of weight to this observation but thought it was worth noting. This panelist was not comfortable with a quantitative comparison between the animal results and human infection information, due to the difficulty in collecting reliable data on colds and severity of infections above background levels, but thought that on a qualitative basis the decreased WBC counts should be considered adverse. The expert also noted that after the age of 40, the human immune system deteriorates and resistance to infection is impaired. Another panelist also noted that frank effects were seen in the inhalation study at high doses, but reiterated that there is a substantial gap in dosing between that study and the HLS (2001) study. The panelist thought that further study in this gap in dosing could answer many of the current questions about sulfolane toxicity.

With regard to the determination of adversity, a panelist noted that terms such as adaptive and reversible have a different biological meaning than adverse. This panelist thought that when a precursor to a clearly adverse effect blocks the organism's ability to withstand an additional challenge, then the precursor is an appropriate basis for establishing the RfD. Another panelist thought that anything contributing to compromised immune function should be considered an adverse effect, while another pointed out that there is no evidence to say that leukopenia is not adverse, and thus there is no evidence to discount it as a critical effect. Other panelists also thought leukopenia should be considered adverse, noting that while the available data fall short of showing compromised immune function in the laboratory animals and that it would be desirable to get a better handle in terms of actual effect on a person's health, there is enough evidence in the laboratory studies to demonstrate the potential for affecting the immune system and therefore the effects should not be dismissed.

A panelist raised another consideration, that is, how reversibility of effect relates to differences in response for neonates. This panelist explained that in adult humans, bone marrow is the only WBC producing organ, while WBC cell production can also occur in the liver of neonates and prenatally in the spleen. If hematological effects arise from neonatal or prenatal targets, these tissues have no ability to recover, while adult bone marrow usually (but not always) has the ability to recover. These differences raise a concern regarding developmental immunotoxicology potential and this is a data gap that planned sulfolane research by the National Toxicology Program (NTP) could address.

The panel also discussed other endpoints, specifically the male kidney effects, and their relevance to humans. The EPA PPRTV document concluded that sulfolane meets some of the criteria for indicating the effect is based on an α2u-globulin mode of action, but not all the criteria were documented. If the kidney effects are due to an α2u-globulin mode of action, EPA would conclude that the male rat kidney effects are not relevant to humans. In this case, where the data are incomplete, the endpoint would be included for further consideration and modeling for an RfD. A panelist noted that over the years EPA has moved away from considering male rat hydrocarbon nephropathy as an adverse effect, but EPA does allow for BMD modeling for effects that are not ultimately considered adverse, thus it is appropriate to consider this effect in the overall selection of effects. A panel member asked whether the observation of effects in males was persuasive and another thought this depends on the mode of action involved. Another panelist thought that if the kidney endpoint were modeled, the results would indicate a higher effect level than the immune endpoints. Thus, even if the kidney endpoint were excluded these effects would not be the basis for the RfD.

Regarding other target organs, the panel also considered effects in the liver, and thought that liver effects would not be more sensitive than the effects on the immune system. An ATSDR representative asked about potential neurotoxicity as an effect for consideration. The panel confirmed that effects on the nervous system were reported in acute high concentration studies (e.g., Andersen et al., 1977). However, the panel felt that such effects would not likely be the primary concern for an oral RfD based on chronic drinking water scenario. This was supported by the absence of effects on neurobehavioral outcomes and neurological organs in the HLS (2001) study.

The panel agreed that leukopenia (reduced WBC count) should be considered an adverse effect for the purposes of, and within the context of, deriving an RfD. They cited precedent for using the endpoint in another assessment (i.e., benzene), and that there is a general correlation between decreased WBC counts and greater risk of infection, suggesting that the endpoint is appropriate for setting an RfD, despite the fact that the functional implications of the immune responses have not been tested for sulfolane. The panel agreed that WBC count decreases and *related metrics are the most appropriate effects to use as the basis for deriving the RfD for sulfolane. These endpoints meet the RfD concept of an adverse effect although there are no direct immune function testing data available for sulfolane to know the functional impacts of the WBC count changes.*

Charge Question 2b: Dosimetric adjustments

Charge Question 2b: What dosimetric adjustments should be made for the relevant endpoints?

The panel discussed whether dosimetric adjustments should be made to the experimental dose, noting that the current RfDs took different approaches regarding this issue. The assessments by TCEQ and Thompson et al. calculated an HED using the allometric scaling method. Others did not adjust the dose, although authors of the EPA PPRTV and ARCADIS assessment, which did not include dosimetric adjustments, stated at the meeting that they would use allometric scaling to calculate an HED, if they were to redo their assessments today.

A panelist explained that EPA currently uses standard dosimetric scaling techniques to adjust the experimental dose to one that reflects continuous exposure and a human equivalent dose or concentration. This is usually done by allometric scaling and these dosimetric adjustments have been applied to most of the recent EPA IRIS chemicals assessments. The underlying biological assumption behind this scaling is that there is a correlation between basal metabolism rate and body weight among species that can be described as a power function. Toxicity is scaled based on body weight ratios raised to a power to account for basal metabolism rate, which likely reflects overall clearance. The panel discussed whether this is an appropriate way to adjust the dose for sulfolane; several noted the toxicokinetic data reported by Andersen et al. (1977) suggest that sulfolane metabolism is saturable and largely excreted as a primary metabolite or as the parent compound. Others agreed with this logic and noted that there is considerable literature to support dosimetric adjustment based on allometric scaling for chemicals with this toxicokinetic profile.

Dosimetric scaling adjustments are done to adjust for toxicokinetic differences between animals and humans. Panel members noted that typically, unless there is information to indicate otherwise, the uncertainty factor for extrapolation from animals to humans will be reduced to the half-log of 10 (i.e., 3), with the remaining half-log of 10 accounting for the uncertainty in toxicodynamics. The panel deferred consideration of the appropriate residual factor to the uncertainty factor discussion (see below).

The panel noted that calculating a human equivalent dose, or HED, based on allometric scaling is the current default approach used by the EPA in deriving RfDs. The panel did not find any data to suggest the HED method is inappropriate for sulfolane.

Charge Question 2c: NOAEL, LOAEL, benchmark dose modeling

Charge Question 2c: Discuss the no and lowest observed adverse effect levels (NOAELs and LOAELs). Evaluate the endpoints for suitability for benchmark dose (BMD) modeling and discuss model fit.

Use of NOAEL/LOAEL approach

The panel discussed the use of a no (or lowest) observed adverse effect level (NOAEL or LOAEL) or a benchmark dose approach. Panelists noted that the scientific consensus is that the BMD approach offers advantages and is preferred, but the data have to meet minimum requirements to be modeled; for example, one needs to have a statistically significant dose-response trend. The HLS (2001) data set meets the requirements and using BMD modeling provides better information on the dose response curve. A panel member was impressed by the level of detail and attention paid to the sulfolane BMD modeling efforts. Based on review of the modeling performed, the panel did not see any errors in the performance of the modeling reported in the publication by Thompson et al. (2013) and as confirmed by the analysis conducted by Gradient on behalf of ADEC.

A panelist provided clarification on a common misconception - explaining that the BMDL is an approximation of the confidence limit lower bound, since it is based on response and not dose. Because it does not consider the choice of endpoint or uncertainty factors, it does not provide confidence limits on the RfD itself. The BMDL is a point of departure – a starting point to carry out the rest of the procedures to get to the reference dose. This issue was raised due to a generic argument the panelist had seen regarding use of the BMDL as a justification for reducing the uncertainty factors for variability in susceptibility. Another panelist observed that some of the BMDL values are greater than the LOAEL of 10.6 mg/kg-day in the HLS (2001) study and wondered if this is uncommon and should be a concern. Other panelists thought that this does happen on occasion and can reflect the nature of the dose-response curve; the BMDL represents a lower bound on dose based on the variability of the modeled curve. This is in contrast to LOAEL identification, which is based on statistical significance (difference in means) that is impacted by variability in the data for the comparator dose groups. Thus, the measures of variability upon which the BMDL and the LOAEL are identified are not the same.

The panel agreed that the scientific consensus in risk assessment is that the BMD approach offers advantages over using a NOAEL or LOAEL, but the data set has to meet minimum requirements, including a statistically significant dose response trend. The HLS (2001) data set *meets the requirements and is amenable to modeling. The panel recommended that BMD modeling would better inform on the nature of the dose response curve than would use of a NOAEL or LOAEL.*

Log transformation of data for benchmark dose modeling

The panel discussed the issue of transforming the dose data to a log scale for BMD modeling. Table 2.2 of the Gradient report provides a summary of the log-transformed dose-response data for BMD modeling for four endpoints: WBC count, lymphocyte count, LUC (large unstained cell) count, and monocyte count. In response to a question from the panel, the Gradient authors noted that the LUC data did not provide an adequate fit even after log-transformation and thus LUC BMD results were not included in the final summary table. One panelist noted that transforming the data to a log scale is not based on biological or mechanistic information; rather it is to fit the data. The EPA guidance is clear that one can log transform the data if the linear models do not fit. The panel asked the EPA authors if they considered doing the log-transformation in their assessment. EPA indicated that such an approach is not part of the standard method within the PPRTV program. A panelist asked Dr. Rhomberg if he supports the use of the log transformation approach for sulfolane. He said he does. He noted a caveat that has not been fully addressed in the field, that is, the mathematical conversion of the transformed data by the addition of a value of unity to the log dose to ensure transformed positive dose values. This approach can distort the doses on the log scale, but the impact of this factor was expected to be small in this case and would not preclude use of the log transformation procedure as done in the Thompson et al. (2013) assessment.

The panel agreed that log transformation is an appropriate approach based on these data sets for sulfolane and considered the modeling results appropriate for identification of candidate point of departure values.

Historical vs. concurrent controls

The Thompson et al. (2013) assessment used historical control data from the HLS laboratories to determine the standard deviation from control that they used to define the benchmark response in the dose-response modeling. Because the HLS (2001) study authors were not clear about the adversity of the effect, the Thompson et al. authors thought it reasonable to look at historical variability in WBC counts as a basis for defining the benchmark response for the BMD modeling approach. Panel members questioned this use of historical controls rather than concurrent controls from the study. A panel member noted that the EPA BMD guidance (EPA, 2012b) indicates that historical controls can be used, but the guidance is not clear on when or why one would use them. Some panelists expressed a strong preference to use concurrent controls, noting that the purpose of having concurrent controls is to control for both the known study variables as well as things that are not known that may impact the results. These panelists noted that historical controls are useful to compare with concurrent controls to identify inconsistent results that may point to a problem with the study. In this case, the concurrent control mean falls within two standard deviations of the historical control mean and it does not appear that there is anything abnormal with the concurrent controls. Another panelist noted however, that a common intent of using the historical controls to better capture the true variance response, in this case, the WBC parameters. This is concordant with the concept of preferring the larger sample size offered by the historical control data set. In response to a question from the panel, the authors of the Thompson et al. (2013) assessment confirmed that they used historical controls from the HLS laboratories from a relevant time period (June 1998 to June 2003) that were matched for species, sex, strain and age. A panelist observed that in this case the historical controls (with a larger number) had a greater variance than the concurrent controls, which might indicate some factor in the study that is causing the concurrent control response range to be tighter. While all the panelists thought the concurrent control data should be preferred over historical controls, they did not have a specific rationale to determine that the use of historical controls in this case (where the data are within two standard deviations of the concurrent controls) is scientifically inappropriate. However, the panel's strong preference was for using the concurrent control data based on standard risk assessment practice.

The Thompson et al. and TCEQ RfDs used a 1 SD change from control mean, rather than a response percentage for the BMR, and the panel chair asked the Thompson et al. authors to explain their reasoning. Dr. Thompson explained that when one does not know what is adverse for a continuous variable, using 1 SD for the BMR is appropriate. Panel members discussed the use of 1 SD as the metric for the BMR. A panelist noted that the choice of 1 SD is somewhat arbitrary as it is not clinically or biologically based. The justification for 1 SD this panelist has found in the literature is Crump (1995)[6]. The panel noted that use of 1 SD is consistent with standard practice in applying EPA risk methods when there is an absence of specific data to identify a degree of deviation from controls that is adverse.

An author of the ATSDR assessment asked whether anyone has done any time series analyses for trends for WBC counts over time. A panelist explained that NHANES has a large amount of data on WBC counts. These data are broken down by age, smoking, and sex and show significant variation within the U.S. population. Another panel member noted that in clinical practice, laboratory results are evaluated in the context of normal ranges which is often defined as within 2 SDs of the mean. The panelist questioned how this informs the selection of the BMR. Another panel member questioned whether 2 SDs should be used instead of 1 SD, based upon the larger variability in WBC counts in the human population.

[6] Post meeting, the panelist explained the rationale as follows. Consider a response where lower values indicate toxicity, and assume that a limit is set below which 1% of values lie in the control population (the left tail of the distribution), so any value below that is considered adverse. That 1% limit corresponds to 2.3 SDs below the control mean value. If the population values follow a normal distribution, and the mean shifts by 1.1 SDs to the left (lower values), then 10% of that shifted distribution now lies in this "adverse" effect region. Thus, if the mean response of the exposed group is 1.1 SDs below the control mean response, then the corresponding BMR is at a 10% response.

As noted above in earlier discussions, the panel agreed that quantitation of the degree of change in WBC counts that causes increased susceptibility to infection and other effects is a gap in the scientific literature. The panel acknowledged the arbitrary nature of the 1 SD approach, but recognized that this is within the BMD guidance. The biological underpinning for this choice for a specific endpoint is not clear, but it does not mean that it is inappropriate to use. However, the panel also noted that the issue of defining the clinical relevance of such a change, given the degree of human variability, would be helpful to include in the RfD assessment.

The panel strongly preferred use of the concurrent control data as the best representation of variability under the specific experimental conditions of the HLS (2001) study. They also recommended use of 1 SD for the BMR because it is consistent with standard practice in applying EPA risk methods when there is an absence of specific data to define adverse response.

Model selection procedures

The panel discussion of model selection opened with a brief review of the modeling validation effort completed by Gradient. The presentation of results was provided by Dr. Lorenz Rhomberg and Mr. David Mayfield. Dr. Rhomberg explained that Gradient was asked by ADEC to review and repeat the BMD modeling to verify the repeatability of the dose-response curve fitting that other groups had done in their derivation of RfDs. He noted that under the scope of their evaluation, they looked at the curve fitting and tried to repeat the results, but they did not evaluate adversity of effect, evaluate the POD, or make other judgments related to the basis of the RfDs under review. Dr. Rhomberg commented that they were able to repeat most of the results, and reported that the differences observed were based on the different approaches such as log transformation, different models, and the selection criteria used to choose the most appropriate BMD and BMDL. While many of the sulfolane RfDs used a BMDL for the POD, the BMDS Wizard was not available at the time and the RfD authors performed model selection by reviewing each model set. The Gradient evaluation of the BMD modeling utilized the Wizard and discussed results based on different model selection criteria. Key elements of this summary and panel and RfD authors' clarifying questions are captured below.

Based on this opening discussion, a panelist noted the preferred approach to BMD modeling is to model the data for each appropriate endpoint and then compare results to find the best model fit. The reviewer noted that the EPA's BMD guidance (EPA, 2012b) provides general criteria and recommendations, but the BMDS Wizard has incorporated elements of the model selection process. Where the written guidance and the EPA Wizard tool do not align, there is room for confusion.

The panel discussed a number of specific issues related to guidance on model interpretation and model selection raised by the RfD derivations and the questions and comments submitted by RfD authors. One of the panelists commented that the EPA BMD guidance is not firm and is not necessarily consistent with regard to model selection criteria. The panelist thought that the guidance is not definitive for selection of endpoints and if one has a biological preference for an endpoint, one should model that endpoint and judge whether the results look appropriate (i.e., provide an overall reasonable visual fit and meet the statistical criteria laid out in the BMD guidance). If the results are questionable, then one should model other endpoints. There are a number of criteria for model selection, including visual fits, residuals, and goodness of fit p-value. Akaike Information Criterion (AIC) and the ratio of BMD to BMDL (BMD/BMDL ratio) are two additional metrics to consider. The panel discussed several specific questions related to model selection, including:

- What is the relevance of the BMD/BMDL ratio and how does it fit into model selection process?
- What is the role of AIC as a model selection criterion?
- What are various approaches for model averaging procedures?

Use of BMD/BMDL ratios

One issue that had been raised in the comments from the authors of Thompson et al. (2013) assessment was the use of a BMD/BMDL ratio as a criterion for model selection. A concern was raised that the calculation of these ratios using the logarithmic BMD and BMDL values resulted in a lower ratio than if the BMD and BMDL had been converted to arithmetic values before calculating this ratio. This approach had the effect of including more models in consideration if a criterion of five were used to define an acceptable BMD/BMDL ratio. Dr. Rhomberg noted that in his view no mistake was made in not retransforming doses to arithmetic scale. He felt that since the fitting was done using the log transformed dose, it was appropriate to leave it in the log format for the ratio analysis. It might have been of interest to view them arithmetically as well, but it did not change the final conclusions. There is no guidance that says one has to transform them back from log, and this would be inconsistent to use when comparing models in the Wizard.

The use of a value of five for the BMD/BMDL ratio as a selection criterion was also discussed. Dr. Rhomberg explained that their evaluation considered the way the BMD/BMDL ratio is used in the BMDS Wizard and a ratio of five was not preselected as a specific criterion for choosing or rejecting models. Rather, this criterion relates to looking at the behavior of the curve in the region of the BMD, and serves as a measure to evaluate the ability of the data to describe the curve in that region. The approach used by Gradient was more about curve fits than model selection criteria. Dr. Rhomberg noted that in his view the BMDS Wizard uses the ratio only to warn about poor model fits, indicating when a user is stretching the limits of model reliability and the ratio is not a criterion for exclusion. The EPA guidance on choosing the best model states that a user should evaluate all models and eliminate those with a lack of fit first. Then, if the BMDLs are in same range, one chooses the model with the lowest AIC. If BMDLs are not within a narrow range, then one chooses the model with the lowest BMDL. The ratio of five does not come into this decision; it merely identifies a warning of ‘stretching’ the data. Based on this, both methods may be appropriate. Dr. Rhomberg also noted that the EPA BMD Wizard does not reject models until the ratio is over 20; this value was not reached for sulfolane even when the ratios were calculated using the logarithmic values.

Dr. Rhomberg also noted a paper published in Environmental Health Perspectives this year (Wignall et al., 2014) that evaluated BMD fitting to data and evaluated fit and the EPA BMDS Wizard. The authors of this paper found a few BMD/BMDL ratios above five, but did not drop any of those from their analysis, nor suggest it would be appropriate to do so. This suggests that retaining models that have a BMD/BMDL ratio above five is not a mistake and is in line with EPA guidance. Dr. Rhomberg concluded that the results found in the Gradient report reflect following of EPA guidance.

One panel member asked Dr. Rhomberg if he knew of publications in which EPA discusses this ratio, aside from the documentation of the EPA BMDS Wizard. Dr. Rhomberg thought that these types of references were cited in the Wignall et al. paper, but noted that EPA does not make this criterion explicit in their BMD guidance.

Another panel member asked if the ratio of five should not be used as a criterion for excluding a model, at what point would one consider it? Dr. Rhomberg responded that his job was to follow the existing practices and to evaluate how the different model results came about. Based on that assignment, Gradient followed the practice of excluding a model with a BMD/BMDL ratio of 20 as indicative of a poor fit. A ratio greater than five was considered worth noting, but indicated the need for a close visual evaluation; as the visual fits are not captured in an automated model screening tool like the EPA BMD Wizard. Overall, Dr. Rhomberg recommended following the guidance unless a user has evidence to move away from it.

The panelist asked Dr. Thompson to describe the criteria ToxStrategies used for model selection. Dr. Thompson indicated that he did not think there is a bright line in BMD/BMDL ratios for model selection and that goodness of fit of 0.1 is somewhat arbitrary. Dr. Haws added that when they developed their BMD modeling results, the BMDS Wizard was not widely used and they followed the procedure described by the panel - weighing a variety of criteria including using visual inspection and residuals. Dr. Thompson clarified that they did not use the BMD/BMDL ratio to discard models from their selection; he noted that he would defer to statistician's judgments regarding use of five or 20 as cut-off for this criterion.

One panelist questioned whether the BMD/BMDL ratio is adding much more than what one already sees in the graphs. In addition, one could look at integrated residuals across the entire dose-response curve -- to give an idea of distance; how many data points track well; trends seen in the low end, but not the high end; and, how sensitive the model is in the high end. If one thinks the model is still stable or robust then one makes a choice. This panelist did not think there is anything very definitive about any single criterion, but would personally put more stock in residuals and visual fit in order to get a sense of what is going on in the low dose region. One could also look at box and whisker plots for control groups to evaluate consistency and see if there are any outliers. This panelist prefers visuals as the first step and then looks at the statistical results. The panel thought that the BMD/BMDL ratio is a less important factor than the other selection criteria (i.e., it is unlikely to change model selection beyond the other criteria that are routinely considered such as visual fit and scaled residuals).

Use of AIC for model selection

The selection of the model with the lowest AIC among the adequately fitting model was also discussed by the panel. A panel member asked Dr. Rhomberg his view on the issue of AIC, and what AIC value is considered significantly different for differentiating between models. Dr. Rhomberg replied that significance in AIC is a challenging consideration for identifying a difference that is practically and consequentially different, rather than statistically different. Dr. Rhomberg explained that the goal is to choose the most parsimonious model. The AIC penalizes the goodness of fit for number of parameters and thus the lowest AIC is in fact is the most parsimonious model, not the model with the lowest number of parameters. Therefore, EPA guidance to choose the lowest AIC is a parsimonious choice by default.

Another panel member asked Dr. Rhomberg if he would retain two models that have AIC values that are close (e.g., two or less apart). Dr. Rhomberg replied that when the models are close, both models should move forward since significance is related to a practically meaningful criterion. In that small range (a value of less than two between models), for all practical purposes, the models are equivalently good, and a tiny margin does not help in choosing the best model. Dr. Rhomberg continued that an absolute value of two is a good rule of thumb for selecting based on AIC, and this recognition is built into the methodology and the Wizard. To clarify how EPA would select a model, a panelist walked through the standard EPA model selection steps using the WBC count data (log transformed). The first step would be to look at model fit. Any model with a p-value below 0.1 would be discarded. Then one would look at AIC and scaled residuals. For WBC this leaves five models (or three condensed models) that fit. Using standard EPA guidelines - if the BMDLs are within a narrow range (i.e., within a factor of three), one would choose the model with the lowest AIC. If the range is outside a factor of three, one would choose the lowest BMDL. Another panelist questioned using the lowest AIC in this case because they are so similar (AIC values range from 109.06 to 109.17). The panelist suggested looking at other indices (e.g., scaled residuals) rather than strictly following the EPA guidance around selection of the model with the lowest AIC.

One panelist also noted that an improvement in the AIC metric is the version that is corrected for small sample size (i.e., AIC_C). Simulation studies for small samples (n<40) have shown that the correct model is selected more often when based on AIC_C than when based on AIC (Burnham and Anderson, 2004). A panelist asked if there was a role for using the AIC_C for evaluating the sulfolane WBC data sets. The first panelist thought that AIC_C can be most helpful when there are larger differences in AIC values (and so more correction with the AIC_C) and when it is otherwise difficult to choose from among the models.

The panel did not think that selecting the final model solely based on the lowest AIC was appropriate when the AIC values were so similar for the adequately fitting models. If one picked the lowest AIC, then a different model would have been favored, but when one does not differentiate between the close AICs, the linear model was favored based on consideration of scaled residuals. The panel thought that weighing of residuals is an important step to include in the model selection process.

Averaging model results

The authors of the Thompson et al. (2013) RfD noted that several of the models yield the same results. A panel member commented that this reflects that such models all collapse to the same mathematical form because parameters for the higher order models were not adding to the overall fit. Dr. Thompson stated that his team recognized this and, in fact, did not double count models that condensed to the same mathematical form in the analyses described in their publication.

A panelist noted that the issue of averaging of model results was raised in some of the RfDs reviewed by the panel and in the Gradient evaluation. The panelist commented that EPA guidance does not provide clear procedures regarding methods for averaging. For example, there are differences between arithmetic and geometric means but little guidance on selecting the best approach. In addition, there are other averaging approaches that are not discussed in the EPA methods. For example, the field of model averaging has been rapidly developing in last 10 years and deterministic and Bayesian approaches are used in the statistical sciences that are not being used by EPA. While recognizing the availability of newer approaches, the panelist noted that for the current evaluation the focus of the panel is on consistency with EPA and current risk assessment methods, and not on recently developing statistical methods. The panel agreed that for the scope of the current peer review only methods in the BMDS guidance for averaging model results would be considered. In this case, for the sulfolane RfD, a single preferred model was identified when weighing the various criteria, and an average of the BMDL values was not needed.

Panel recommended model selection procedure

The panel agreed that for model selection they are not comfortable with basing the selection of the model on the lowest AIC as an exclusive final selection criterion given the small differences in AIC values.. The panel noted that EPA guidance is not precise regarding the use of AIC and BMD/BMDL ratios as model selection criteria. The panel agreed that there is scientific rationale to deviate from EPA's regimented rules for model selection. The panel recommended starting with the EPA guidance and considering deviations from EPA guidance based upon weighing all the model fit considerations while providing a clear rationale for the model(s) ultimately selected.

The panel developed a list of criteria for model selection for use with the data sets of interest from the HLS (2001) study. The panel's recommended approach starts with the EPA BMD guidance (EPA, 2012b), but includes taking into account additional information pertinent to model selection. The panel's criteria and steps for removing models from consideration are as follows:

1. Numerical problems. When there are clear numerical problems that cause the models to fail, this is one way to weed out data sets. An example would be data that lack any dose-response pattern.

2. Unacceptable visual and statistical fit. EPA's poor-fit criterion of a p-value of less than 0.1 is very lenient. Preferably one would like to remove additional poor fitting models (with a p-value less than 0.1), with further removals based on visual appraisal and scaled residuals.

3. Over-specified models, for example, with higher order polynomials that are otherwise statistically identical in fit to a linear model. In this case, one is not gaining anything by adding parameters. In some cases, the models reduce to a simpler model and so several models will show up as having identical fits – these should be considered the same model with selection of the simplest form. In other cases, the AIC can be used to evaluate the role that model complexity is playing in improving overall fit. Generally, if the AIC values are close in absolute value they are all of similar reliability.

4. Scaled residuals too high. The scaled residuals provide additional information on the fit in the range of most interest and can be used to discriminate among models that have similar overall global fits (i.e., similar AIC values).

RfD authors from ATSDR asked several clarifying questions related to the model selection process. One question was whether trends in residuals would be evaluated as a criterion for selection of model. A panelist clarified the trend analysis for residual was not included in the panel's assessment evaluation. RfD authors who had done BMD modeling also indicated they had not included such an analysis in their assessments. The authors also asked whether evaluating residuals in terms of the logarithms would have any impacts on the selection process. A panelist noted that should not be a concern because the residuals are for the response, while it is the dose that is log transformed.

Charge Question 2d: Point of departure

Charge Question 2d. Which is the most scientifically defensible point of departure (POD) for a sulfolane RfD?

A panel member opened the discussion noting that the usual approach for BMD modeling is to choose relevant biological endpoints, model each of them, and then select the best fitting model. The panel discussed whether there is a biological basis for preferring one immune system metric over another for selecting candidate endpoints for POD determination.

Previously the panel discussed which endpoint would be the best point of departure and narrowed down to WBC or lymphocyte counts. An ATSDR author asked if conversion of monocytes to macrophages was consistent with the results on cell-type specific effects and whether that should be considered in selecting the endpoint. One panelist thought it would be difficult to construct a biological rationale for selecting one cell type over another (based on adversity or other biology). The LUC are early lymphocytes and all these measurements are involved in the immune system and are linked together; they all reflect early progenitor cells. This panelist recommended using WBC count because most of the WBC count is driven by the lymphocytes, not by neutrophils, basophils, or monocytes. The panelist added that the LUC cells are probably precursors of lymphocytes and changes in LUC, which were not modeled, would also be captured in the WBC measure. It is likely that most of the effect on WBC was the decrease in lymphocytes. Therefore, the most conservative biologically explainable way to address this is to use the WBC count. Even though LUC did not fit the BMD models, it is intuitive that the source of the effect is the stem cell and bone marrow. Based on these considerations this panelist questioned the value of modeling both WBC and lymphocytes, noting that both originate from the stem cells and a physician would treat a patient for the reduced WBC count.

A panel member asked if certain cell types were more sensitive indicators of immune system effects, and if this might be a reason to select a subpopulation of WBCs. Another responded that there is no clear basis for one cell type in this case. If only the lymphocytes were affected, then using the lymphocytes as a POD would be preferred, but since all but the neutrophils were affected in the rats in the HLS study, it makes more sense to use WBC. The panelist did note that it is not clear why the neutrophils were not affected; based on their bone marrow lineage; one would have expected them to also be affected. Some panelists agreed with this line of reasoning to focus on WBC count for POD selection.

However, other panelists expressed reservations about selecting WBC count data at the beginning of the POD process, and suggested that without any clear biological basis for a choice, all the endpoints (WBC, lymphocytes, LUC, and neutrophils) should be modeled and then biological considerations could be used to inform the best choice. One panelist expressed hesitation in setting aside the lymphocytes because of the lack of knowledge about MOA and therefore agreed that modeling all related immunological endpoints would be best approach. If the results differ, then one can decide which endpoint is more biologically relevant or more important. Others noted that the Gradient report (Table 4.1) indicates a significant difference between lymphocyte and WBC results for some individual models, but when the models are averaged they come out to a similar BMDL value. A panelist noted that results for an individual model may differ for various reasons, and the average of the models shows a consistent result. Another panelist thought that if the two endpoints have reasonably consistent answers, then the consistency adds to overall confidence in the result. If they are different, then one would question if there is a modeling issue that needs to be addressed.

The panel walked through application of their model selection criteria using the modeling results from Tables 3.2 and 3.6 of the Gradient (see slides in Appendix C). The panel's preference for concurrent controls was included. Applying the panel's criteria to the sulfolane data for WBC count using concurrent controls results in selection of the linear model using log(dose). When the process was used for lymphocyte counts with concurrent controls, the same model was selected. Two alternatives remained:

- HLS (2001),WBC count, concurrent controls, linear model with a BMDL of 12.66 mg/kg-day
- HLS (2001), Lymphocyte count, concurrent controls, linear model with a BMDL of l4.45 mg/kg-day

The panel agreed that deciding which of these is the most appropriate POD is not a statistical modeling choice; rather it should be based on biology. In a biological sense the two are relatively concordant and so the determining criterion would be the most relevant choice in the context of health. Based on this consideration the panel agreed that leukopenia (decreased WBC count) was the preferred endpoint for developing the POD.

The panel recommended that modeling both endpoints is an appropriate initial step for the sulfolane data followed by identifying which are the appropriate models and metric selections. Differences in results for the two endpoints might help inform model selection based on biology. Overall, the panel favored the selection of decrease in WBC count as the critical effect, on a toxicological basis since it better captures potential impacts on health effects (it covers more of the WBC types that were affected). Based on this consideration and the model selection procedure, the panel identified the BMDL of 12.66 mg/kg-day, based on the concurrent control data set for WBC decreases in the HLS (2001) study as the most appropriate POD.

Charge Question 3: Selection of uncertainty factors

Charge Question 3. Discuss the basis for selection of uncertainty factors. What are the most appropriate values for the standard factors commonly used?

The panel discussed what uncertainty factors would be appropriate to use in deriving a sulfolane RfD. They discussed each of the standard five uncertainty factors individually, and then considered the most appropriate value for the composite uncertainty factor.

Interspecies Differences - UF_A

As discussed earlier, the panel agreed that using the HED was appropriate and that there is nothing to suggest that calculating an HED would be inappropriate in this case. The panel agreed that the general risk assessment guidance is to reduce the uncertainty factor for extrapolation from animals to humans when an HED has been applied. In response to a clarifying question from the panel, an EPA author noted that current EPA guidance(EPA 2011) recommends reducing the UF to 3 (half-log of 10) unless there are data on toxicodynamic differences that can be used to further reduce the UF. Another RfD author (Dr. Magee) added that following the IPCS methodology (IPCS, 2005) a residual factor of 2.5 would be applied.

The panel noted that the Thompson et al. (2013) assessment calculated an HED and applied a factor of 3 for this UF. The TCEQ assessment also used the HED approach, but applied a factor of 1. TCEQ guidance (TCEQ, 2012, page 174) indicates that an UF_A of 1 be used "unless species- and chemical-specific data adequately support an alternative UF_A value (e.g., data indicate a UF_A greater than 1 is needed to account for toxicodynamic differences or that $BW^{3/4}$ scaling is inappropriate and an alternate UF_A is scientifically justified)". Authors of the Thompson et al. (2013) assessment asked the panel whether the move by EPA to harmonize dosimetry methods for cancer and non-cancer assessments provides a rationale for not including the remaining factor of 3 for toxicodynamics. A panel member commented that the EPA language regarding harmonization is aimed more at general approaches such as using mode of action considerations rather than the specific consideration of a sub-factor for toxicodynamics. Moreover, the absence of this factor in the cancer risk assessment reflects the use of low-dose linear extrapolation; uncertainty factors are not applied in such assessments.

Panel members expressed concerns that there is not enough known about the MOA of sulfolane, and the toxicodynamics of sulfolane in humans is not sufficiently understood, to move away from a factor of 3. While they recognize that TCEQ presents support for a factor of 1 in their guidance and applies this to their sulfolane RfD, panel members have not seen EPA use a factor of 1 in IRIS or other assessments.

The panel did not think there is compelling reason to recommend moving away from EPA's guidance in light of the lack of understanding of the MOA for sulfolane. The panel recommended use of allometric scaling to calculate an HED, and then application of a UF of 3 for animal to human extrapolation, which is consistent with EPA guidance.

Human Variability in Susceptibility - UF_H

A number of issues were considered regarding the most appropriate value for the UF_H. Panel members discussed whether there are compelling data or reasons for reducing this uncertainty factor and if there are potentially susceptible sub-populations for sulfolane toxicity. One panelist explained that the purpose of the UF_H is to account for variability in sensitivity within the human population. EPA has reduced this factor when the critical effect is based upon data in a sensitive population. Although for sulfolane the limited animal data suggest that developmental effects occur at much higher doses than the immune effects, several panelists were concerned that there are not data available on the full progression of effects, the developmental and reproductive studies did not look at functional endpoints, and stem cells have different roles in neonatal or early life. In addition, even if the developing individual were not particularly sensitive, there are other subpopulations that may be more sensitive. In particular, these may include those with compromised immune function due to disease or drug therapy, and the elderly who have lower WBC counts than other age groups. These potential sensitive subpopulations have not been studied.

Dr. Chad Thompson and Dr. Laurie Haws of ToxStrategies asked for the panel's thoughts regarding differences in effects seen between males and females and whether the females are more sensitive. A panel member noted that that in the Andersen study (Andersen et al., 1977) male rats showed a marked decrease in WBC. Furthermore, studies conducted with benzene have shown only slight, if any, sex differences in the effects on WBCs in rodents. One panel member could not envision a biologically plausible reason for the difference in response between male and female rats and thought that without reproduced studies, firm conclusions cannot be drawn. Another panel member agreed that not enough is known; noting that the available data are limited to make sex comparisons and toxicologists see sex differences in rodent species frequently. In the case of sulfolane, the panel agreed that it is not clear that this is only a female phenomenon. An observer, Dr. William Farland, asked whether the panel thought that a full 10 is needed given the precautionary choices regarding uncertainty factors and toxicodynamics, and the use of a sensitive endpoint in the sensitive sex for the point of departure, suggesting that, because of the lack of toxicity across the sulfolane database, he would lean toward a factor of 3. A panelist thought the argument interesting, but the male/female differences were not as important in this case because there are potentially susceptible populations that have not been addressed, most notable the very old, the very young and those already on immunosuppressive drugs. Other panelists noted that animal data regarding sex specificity do not necessarily translate to human populations, and there are not risk assessment precedents for making this assumption.

A panelist mentioned that EPA is developing more guidance for selection of uncertainty factors and how to apply quantitative data to derive choices other than 1, 3, and 10. However, there do not appear to be adequate data for sulfolane to apply such methods and none of the proposed RfDs considered a data-derived factor. When data are available, such an approach is definitely preferable.

The panel agreed that an uncertainty factor of 10 is appropriate for human variability; given the immune endpoint and that the potential sensitive populations have not been investigated in the studies to date. The panel did not see a compelling scientific rationale to deviate from the EPA default of 10.

Extrapolation from a LOAEL to a NOAEL - UF_L

The panel noted that the usual practice is to use a UF_L of 1 when a BMDL is used as the POD.

The panel agreed that a factor of 1 is appropriate and saw no scientific rationale to move away from the usual practice.

Extrapolation from a Subchronic Study - UF_S

The panel noted that most of the RfDs for sulfolane applied a factor of 10 for this consideration. This is consistent with the typical default used in EPA risk assessment when deriving a chronic RfD based on a study of less than chronic duration. Lines of evidence that can be used to depart from the default value of 10 were discussed. Three important considerations include:

- Toxicokinetics – does dose accumulate and increase the target tissue dose with longer duration of exposure? This consideration can be evaluated using toxicokinetic studies.
- Toxicodynamics – is the effect likely to progress or will damage accumulate due to inadequate repair periods with longer exposure durations? This consideration can be evaluated based on data with studies of different durations.
- Reversibility of effect – does the adverse effect resolve following periods of no exposure? This consideration can be evaluated based on data from studies with observations for recovery periods.

The panel discussed the availability of data related to each of these considerations.

One panel member asked whether any toxicokinetic data are available that provide evidence that sulfolane accumulates over time. This panelist felt that the absence of such accumulation might allow consideration of a reduction in the UF_S. Another panel member noted that data regarding toxicokinetics were limited with findings summarized in a single inhalation study (Andersen et al., 1977), which reported a relatively short elimination half-life of 3 to 5 days and the potential for metabolism. Based on such data it appears that sulfolane would reach steady state relatively quickly and would not have significant potential for accumulation.

The panel also considered whether there is empirical evidence that the threshold dose for toxicity decreases with longer durations of exposure. The panel discussed studies that evaluated effects at a given dose, but different durations of dosing. A panel member noted that decreases in WBC counts from sulfolane in guinea pigs were similar for exposure durations ranging from 20 to 90 days (Anderson et al., 1977), suggesting that the POD did not decrease with longer exposure durations in that study. In addition, doses required for spleen effects from the 30-day and 90-day exposure durations in the Zhu et al. (1987) study did not appear to be substantially different across these time periods. A panelist concluded that there are limited studies for different durations, but these studies do not show evidence of decrease in the point of departure with longer duration exposures within the time ranges of the studies. The RfD authors present at the meeting, when queried, also noted and confirmed the same studies and results described by the panel. Panel members asked if effects such as decreases in WBC would plateau or continue to decline with longer duration of exposure. One of the panel members noted that with chemotherapy treatment at a given dose, WBC counts may decrease and stay depressed with ongoing exposure, but might not worsen. However, there are cases where bone marrow failure has occurred from myelotoxic chemicals and the WBC decreases are not always reversible and could be due to dose, length of exposure and individual sensitivity. From the sulfolane data available for subchronic exposures, this panelist did not think sulfolane would cause bone marrow failure, but noted that it is not known what would happen with chronic exposure.

The panel also examined if data were available on effect reversibility. A panelist noted that if recovery occurs after stopping dosing, this is evidence of reversibility. EPA has used reversibility of effect as one of the considerations in determining the size of UF_S. The panel discussed the available data on stop exposure and recovery studies. In the Andersen et al. (1977) study the WBC count in the 200 mg/m^3 groups showed a decrement with recovery with continuous dosing over time. Referring to Table B9 on page 50 of EPA PPRTV document, challenges were noted with evaluating the exposure and recovery period data from the developmental toxicity study (Ministry of Health and Welfare Japan, 1996). These included that the effects on WBC count were not identified in the study as adverse and that the group means were not statistically different. In addition, units are different among studies and evaluating if the decreases in the treated group and increases in the recovery group are within the normal range of historical controls is not clear. A panelist questioned the biological basis for this response, noting that it appears that sulfolane had no effect, but when dosing stopped the WBC count increased. Another panelist added that since none of the results were statistically significant and the variability was large, it is hard to conclude whether there was recovery – there is not strong evidence of reversibility in this study since it is not known if there was a decrement in the first place. Although the studies for sulfolane were limited, a panelist thought that based on other chemicals causing this type of effect, one would expect recovery.

The panel considered the data and above and discussed their confidence in the data and whether a reduction in the UF_S could be justified. One panelist felt that a more thorough review of the evidence might be useful and might justify reduction to 3. This panelist indicated that further reducing the factor to a value of 1 would not be recommended given the limited data. This panelist further noted that EPA reduced this factor to 3 in the RfD assessment for trimethylbenzenes, where the data were also limited. However, for trimethylbenzenes EPA had a physiologically-based pharmacokinetic (PBPK) model to extrapolate dose to longer durations. In the case of sulfolane the toxicokinetic data suggest lack of dose accumulation as well. The panelist noted that if this approach were used, the limitations in the data and caveats in interpretation would need to be clearly documented. A second panelist agreed with the idea of reducing this factor based on the reasons noted, with the additional consideration that other elements of the assessment (POD selection, other uncertainty factors) had already reflected precautionary judgments. A third panel member favored a value of 3 noting that while data are not robust for each point, the available data all point in the same direction supporting less concern for a progression of effect.

Several other panel members noted that the rationale for reducing the factor was plausible, but had reservations about the quality of the data. One panelist was concerned that it is not known that chronic dosing would not identify a lower POD and therefore the panelist did not feel comfortable with reducing the UF_S. A second panel member favored the factor of 10 - because sulfolane does not have a robust database and there is significant conjecture regarding recovery and the potential for longer term effects. Several panelists thought that the planned NTP study will answer many of the open questions, but that it is most appropriate to err on more protective basis until more data are available. A third panelist agreed with the more protective approach and retaining the full factor of 10, noting that based on lack of statistically significant effects (partly due to high variance) the recovery studies were not persuasive enough; more data would be needed to make the arguments compelling.

The panel members did not all agree on the most appropriate value for UF_S. Three panel members indicated they would be comfortable with a value of 3 for this factor as long as shortcomings in the data were noted. Three other panel members did not feel that the data were sufficient to move away from the default. Key considerations for selecting a value of 3 or 10 are summarized below. All the panel members agreed that new data would help inform uncertainty factor selection.

Rationale for reducing the factor to 3

- There are limited data suggestive that the degree of effect does not increase with time at a given dose (Andersen et al., 1977 and Zhu et al., 1987).
- Based on limited toxicokinetic data (Andersen et al., 1977) the half-life is such that sulfolane is not likely to bioaccumulate, and would reach a steady state concentration relatively quickly.
- One data set evaluated effect reversibility (Ministry of Health and Welfare Japan, 1996). However, the WBC count data did not show a significant effect at any dose, although the mean values showed a downward trend. As a result, the interpretation of the decrease in the recovery phase high dose females is not clear.
- The draft EPA IRIS review for trimethylbenzenes used similar arguments on reversibility and toxicokinetics. This is a draft assessment under U.S. EPA Science Advisory Board (SAB) review.
- For clinical immunosuppressing drugs, it is not uncommon to see recovery following termination of exposure.
- Level of effect severity for the POD is more consistent with potential for recovery. Since the POD is based on an effect that is not severe, there is a greater likelihood of recovery.

Reasoning for maintaining the default value of 10

- The above lines of evidence are suggestive, but the underlying data sets are too limited to move away from the default.

Insufficiency of the Database - UF_D

For the database uncertainty factor the panel reviewed the available studies and what is missing. With oral administration, there are subchronic studies in two species (Zhu et al., 1987; HLS, 2001; Ministry of Health and Welfare Japan, 1996), a six month chronic study (Zhu et al., 1987), a screening-level reproductive study (Ministry of Health and Welfare Japan, 1996) and a developmental study (Zhu et al., 1987). There is no two-generation reproductive study. Panel members all agreed the UF_D should be 3. In support of this conclusion, they noted that the quality of the available studies varied and a two-generation reproductive study is missing, but the HLS (2001) study is of high quality. *The panel agreed the UF_D should be 3.*

Composite Uncertainty Factor

Looking at each uncertainty factor individually, the total UF would be either 300 or 1000:

- UF_A – 3 with use of the HED
- UF_H - 10
- UF_L - 1
- UF_S – some panel members favored 3, others 10
- UF_D - 3

The panel discussed whether the 300 or 1000 represents an appropriate overall value to use for sulfolane. In particular, the panel discussed their individual opinions regarding the overall level of precaution (i.e., more health protective) of a resulting RfD with the composite UF. Panel members all agreed that 300 would be most appropriate composite UF, but there were differing reasons for this conclusion. One panelist considered that a total UF of 1000 reflects precautionary choices for individual UFs and this should be balanced with the decision regarding using a full factor for interindividual variability; therefore, a composite UF of 300 is preferred. Several others voiced agreement with a composite UF of 300, noting that there is an issue of compounding precautionary judgments that started before UF selection, in particular, the use of a BMD based on 1 SD (given the unclear functional impact of the degree of observed effect) and the use of the lower confidence bound on the BMD to get the BMDL. In addition, looking more holistically at the overall RfD process, if one included the quantitative value of the HED adjustment, the total range of adjustment to the POD would be 1200 to 4000[7]; one panelist thought that adjusting the POD by 4000 is too much. Others thought the composite of 300 is

[7] This reflects the composite UF times the HED for rats of a factor of 4 (based on allometric scaling).

appropriate but voiced concern that a composite UF of less than 300 would not be appropriate given studies that are lacking.

The panel agreed that a composite uncertainty factor 300 is appropriate for sulfolane.

Charge Question 4: Additional issues

Charge Question 4. Please identify any additional scientific issues or questions that the panel should discuss.

Panel members did not identify additional issues beyond those discussed in the other charge questions.

Charge Question 5: Recommended RfD

Charge Question 5: Please discuss which of the RfDs reflects the best use of the currently available data, and why.

The panel walked through the steps of deriving an RfD and evaluated how each sulfolane RfD aligned with the panel's conclusions and recommendations. Panel members were careful to point out that their discussion and decisions reflect the current state of risk assessment science and the available data in 2014. They recognized that the sulfolane RfDs being reviewed were developed by different organizations using their own particular methods, that each was developed for a specific purpose that may or may not match with the RfD needs of Alaska, and that each reflects the data and scientific thinking at a particular point in time. The panel emphasized that their process of evaluating these existing RfDs is not meant to infer that any of the sulfolane assessment are inadequate or inappropriate; the panel members are layering their scientific judgments on others' previously developed values for the current needs.

Table 1 below is copied from the ADEC background document (ADEC, 2014) and summarizes the key decision points for each of the RfDs.

Table 1. Available Chronic Oral Reference Doses for Sulfolane (ADEC, 2014).

Source	Principal Study	Test Species	Endpoint	Modeling Approach	Point of Departure (mg/kg-day)	Composite Uncertainty Factor	Reference Dose (mg/kg-day)
CCME, 2006	HLS 2001	Rat (female)	WBC counts	NOAEL	NOAEL = 2.9	300	0.0097
ATSDR, 2010	Zhu *et al.* 1987	Guinea pig	Hepatic effects, changes in serum ALP, WBC counts	NOAEL	NOAEL = 0.25	100	0.0025
ATSDR, 2011	Zhu *et al.* 1987	Guinea pig	Dispersion of spleen white pulp	BMD	$BMDL_{10}$ = 1.5	1,000	0.002
TCEQ (Haney), 2011	HLS 2001	Rat (female)	WBC counts	BMD	$BMDL_{1SD}$ = 16.1 $BMDL_{HED}^{a}$ = 3.9	300	0.013
US EPA, 2012a	HLS 2001	Rat (female)	WBC counts	NOAEL	NOAEL = 2.9	3,000	0.001
Magee, 2012	HLS 2001	Rat (female)	WBC counts	BMD	BMDL = 11.64	1,000	0.01
Thompson *et al.*, 2013	HLS 2001	Rat (female)	WBC counts	BMD	$BMDL_{1SD}$ = 16 $BMDL_{HED}^{a}$ = 3.9	300	0.01
Health Canada, 2014	HLS 2001	Rat (female)	Lymphocytes	BMD	$BMDL_{1SD}$ = 4.12	1,000	0.00412

Notes:

BMD	= benchmark dose
BMDL	= benchmark dose limit
NOAEL	= no observed adverse effects level
WBC	= white blood cell

(a) – The HEDs (human equivalent doses) were derived using a species scaling adjustment factor of ¾ body weight.

The panel's key conclusions and preferences for key decisions to derive a chronic sulfolane RfD are as follows:

- **Principal Stud**y **–** The panel agreed that the HLS (2001) study was the best choice for principal study and would not recommend basing an RfD upon the Zhu et al. (1987) study.
- **Endpoint** – The panel agreed that reduced WBC count is the biologically most appropriate endpoint to use.
- **Modeling Approach –** The panel agreed that BMD modeling is preferred over using a NOAEL or LOAEL. The panel recommended log transformation of the doses to model the endpoints.
- **Point of Departure –** The panel agreed that allometric scaling to calculate an HED is most appropriate.

Of the six RfDs only the Thompson et al. (2013) and TCEQ (Haney, 2011) RfDs matched the panel's preferences on the above parameters. However, Thompson et al. and TCEQ used historical control data in the BMD modeling and the panel preferred use of concurrent controls to calculate 1 SD from the mean for the benchmark response.

Using Table 2 below, which is taken from ADEC (2014), the panel reviewed their preferences for selection of individual uncertainty factor values and compared these with the six RfDs. The individual uncertainty factors for the panel's evaluation are based on the categorization used by EPA (EPA, 2002). As a result, in some cases direct comparisons are not appropriate due to differences in methodologies used. For example, Health Canada and CCME tend to combine overall data insufficiency, including subchronic to chronic extrapolation and severity of effect into one UF rather than addressing them separately.

Table 2. Summary of Uncertainty Factors[a] (ADEC, 2014).

Source (By Date)	UF_A	UF_D	UF_H	UF_L	UF_S	UF_C
CCME, 2006	10	3[b]	10	--	--	300
ATSDR, 2010	10	--	10	--	--	100
ATSDR, 2011	10	--	10	--	10	1,000
TCEQ (Haney), 2011	--	3	10	--	10	300
US EPA, 2012	10	3	10	1	10	3,000
Magee, 2012	10	--	10	--	10	1,000
Thompson *et al.*, 2013	3	3	3	--	10	300
Health Canada, 2014	10	10	10	--	--	1,000

Notes:

--	= no value provided
UF_A	= animal to human uncertainty factor
UF_D	= incomplete-to-complete database uncertainty factor
UF_H	= intrahuman uncertainty factor
UF_L	= LOAEL-to-NOAEL uncertainty factor
UF_S	= subchronic-to-chronic uncertainty factor
UF_C	= composite uncertainty factor

(a) Uncertainty factors are not defined by all groups using the same nomenclature. The uncertainty factors were categorized based on most appropriate descriptor based on the intent of each factor under US EPA method.

(b) Based on the CCME application of uncertainty factors, this value was used to account for adequate, but not extensive dataset; subchronic- chronic extrapolation; and serious effects concerns (CCME 2006).

The panel's key conclusions and preferences for uncertainty factors are as follows:

- **UF_A –** The panel recommended a value of 3 (for uncertainties in toxicodynamics) to accompany use of an HED. All of the RfDs except for Thompson et al. assigned a value of 10 for this UF, but these did not use an HED. A panel member noted that the actual difference in these choices is small (HED with UF_A of 3 is about 12, compared to using a UF of 10 without an HED).
- **UF_H –** The panel favored a full factor of 10 for intrahuman variability. All the RfDs with the exception of Thompson et al. used 10. Thompson et al. made a case for females being more sensitive, but the panel did not agree that the data were sufficient to reduce the uncertainty factor.
- **UF_L –** Consistent with common risk assessment practice, the panel agreed that a factor of 1 is most appropriate with use of a BMDL. All the RfDs had a value of 1 or did not include this uncertainty factor.
- **UF_S –** The panel members did not agree on the individual UF for extrapolation from a subchronic study. Some thought a full 10-fold factor is appropriate, while others thought this UF should be reduced to 3 because there are data that suggest severity does not progress with duration of exposure and effects may be reversible. For those RfDs that listed a factor for subchronic to chronic extrapolation, all used a factor of 10. CCME, ATSDR (2010) and Health Canada did not include a value for this individual uncertainty factor.
- **UF_D –** The panel recommended using a value of 3 for database uncertainties. Those RfDs that included a value for UF_D also chose 3, with the exception of Health Canada (see note above).

The panel observed that the composite uncertainty factor for the six RfDs ranged dramatically, from 100 to 3000. The expert panel recommended a composite UF of 300 (with HED adjustment). While only two of the RfDs used an HED (TCEQ and Thompson et al.) with a UF of 300, several others reached fairly equivalent RfDs with larger UFs and non-adjusted PODs (e.g., CCME and ARCADIS).

The panel discussed which among the six RfDs has the closest alignment with the panel preferences, which were based on EPA methods and application of the best science. Panel members found that none of the six RfDs aligned perfectly with the thought processes and recommendations of the expert panel. However, all agreed that the Thompson et al. RfD most closely reflects the panel's scientific comments and preferences. One aspect of RfD derivation where the panel differs with Thompson et al. is over the use of historical control data for calculating the standard deviation in response and the panel preference for a 10-fold UF_A. However, from a practical standpoint, the RfDs derived by TCEQ, ARCADIS, and Thompson et al. are all similar at 0.01 mg/kg-day. In addition, those RfDs that did not match as closely are within the same range when one considers the definition of an RfD (i.e., spanning an order of magnitude).

The panel also noted the significant differences in approach and decisions with several other RfDs that do not align as well. However, in doing so, the panel again emphasized that its evaluation of these existing RfDs is not meant to infer that any of these assessment are inadequate or inappropriate; the panel members are layering their scientific judgments on others' RfDs that were developed based upon available data and methodology choices in the past. Specifically, the EPA PPRTV differs due to fundamental choice not to log transform the data or use BMD modeling. EPA also used a larger composite UF. ATSDR used a different study (Zhu et al., 1987), different species (guinea pig), different endpoint (dispersion of the white pulp of the spleen), and no HED.

In summary, the panel concluded that after a close and systematic evaluation of the steps and decisions for calculation of an RfD, using EPA methods as a backdrop and the panel's scientific expert judgments, the Thompson et al. RfD most closely aligns with the panel's conclusions. The Thompson et al. RfD does not match in all aspects, in particular the panel favored use of concurrent controls over historical control data: the selection of some individual UFs differed (e.g., the panel recommends a UF of 10 for intrahuman variability); and the process for selecting the best BMD model differed (e.g., the panel explicitly weighed biology in choosing WBC counts over lymphocyte counts, while Thompson et al. based their model selection on AIC).

Panel members commented that it was interesting to see the scientific input that was brought to bear over time as the various RfDs for sulfolane evolved. They thought that the six individual RfDs should not be judged as right or wrong, but seen as a progression of scientific thinking and methodology over time.

Charge Question 6: Overall confidence and additional studies

Charge Question 6: Discuss the overall confidence in the selected RfD(s) and what additional studies or analyses, if any, would help reduce uncertainty or increase confidence.

Data gaps and additional studies

Panel members thought that the panel's systematic approach to evaluating the steps and decisions in deriving an RfD resulted in the best scientifically-based recommendations with the available data. The HLS (2001) study was very good for hazard identification, but there are some remaining questions regarding the mode of action and role of bone marrow in that neither bone marrow cellularity nor bone marrow differentials were examined. A panelist thought that the panel did a credible job of being scientific and appropriately precautionary; and in the next few years, there will be additional information from the NTP research that will help clarify current uncertainties.

One panelist noted several data gaps, including examination of bone marrow to help identify the mode of action, developmental immunotoxicity studies, hematology studies, and data on reversibility of the effects on immune cells. Another panelist agreed with the first panelist's thoughts on data gaps, noting that if the effects from the available studies are precursors, it would be useful to have more information to inform the progression of the effects. Other panelists noted that it would be helpful to have a longer well-conducted study, perhaps in rats, although other species should be considered, and toxicokinetic data and modeling would be useful.

Several panelists discussed the need for further exploration of the neurotoxicity seen in the existing studies. One noted that HLS (2001) conducted an extensive array of functional neurotoxicity tests that did not identify a neurological concern, while the Andersen et al. (1977) inhalation study showed neurological effects, perhaps suggesting that neurotoxicity is happening at higher doses or it is an acute effect. Without more information on MOA, this is difficult to determine; but many chemicals have different MOAs for acute versus long term exposures.

The panel had concerns that there may be differences in immune response prenatally and in young animals, and suggested that these could be explored with a developmental inmmunotoxicology study. The developmental immunotoxicity study is important because the role of bone marrow differs between the neonate/prenatal stage and adulthood, and impacts on immune cell development and clonal expansion at the early stages can have permanent effects on numbers of immune cells in later life stages.

In addition to the discussion of animal toxicology studies it was also noted that at the current time there are no data on health effects of sulfolane in humans.

Dr. Haws commented that as part of TOX21 testing, NTP has tested sulfolane in a panel of 87 different *in vitro* assays focused on 14 different protein targets and found sulfolane to be inactive in 85 out of the 87 assays, with the results in the remaining two assays determined to be inconclusive. A panel member noted that the TOX21 assays did not include explicit measures of immune function, the primary effect of concern for sulfolane.

The panel noted that the NTP research plans address much of what the panel identified as gaps and additional data will help resolve some of the risk assessment interpretation issues and questions with the currently available data.

Confidence in the RfD

The panel discussed their level of confidence in the database, the principal study, and the RfD. A panel member explained that when EPA develops RfDs they consider the quality of the database, and specifically whether there are studies of the relevant duration in multiple species; whether there are reproductive and developmental studies; if the critical effect is clearly defined;

and if the studies examined endpoints in the most susceptible subpopulations; to assign an overall confidence in the database of low, medium or high. The panelist further clarified that the confidence level may not have a direct impact on risk management decisions, but it is considered. If one has low confidence, there may be impetus to do additional studies and the evaluation helps inform what studies would be most useful. The panel chair asked the ADEC representatives if it would be helpful to them to have the panel provide its opinion on confidence and an ADEC representative indicated that it would be helpful and important.

Panel members were in agreement that for a sulfolane chronic RfD that their confidence in the database is in the low end of medium. They noted there is only one good study and the data are somewhat limited. Panelists pointed out that the confidence in the database aligns with the uncertainty factor for database and the question of whether additional studies would result in a significant change to the RfD. In this case, one panelist commented that new studies would not likely identify a whole new category of effect, but more data may lead to changes in uncertainty factor selection. Panel members considered that their recommendations for deriving the RfD were health-protective, but did not think they had sufficient information to judge whether new data would result in a RfD value that was lower or higher.

The panel discussed the strengths and weaknesses of the principal study earlier in the meeting; several panel members assigned the principal study high confidence, while others assigned the study medium confidence. For the overall RfD, the panel agreed on medium confidence, following from the other confidence conclusions.

Panel members assigned low to medium confidence in the database, medium or high confidence in the principal study, and medium confidence in the RfD.

CLOSING

In closing, the panel chair explained that the panel evaluated the data, methods and decisions for a sulfolane RfD derivation. The panel based its conclusions on application of the best science within the framework of EPA RfD derivation methods. The chair asked the RfD authors whether they had any final questions. Dr. Haws noted that the panel went through the process of evaluating which RfD best aligned with the panel’s recommendations, and stated that the panel’s best science approach, though slightly different from that followed by Thompson et al., yielded the same RfD value as Thompson et al. The panel chair asked the ADEC representatives if the panel had answered the questions that they were asked and covered the issues that ADEC wanted discussed. An ADEC representative confirmed that the panel had addressed the charge and thanked the expert panelists for their work.

REFERENCES

ADEC (Alaska Department of Environmental Conservation). 2014. Review of Oral, Chronic Reference Doses for Sulfolane (CASRN 126-33-0). Available online at: http://www.tera.org/Peer/sulfolane/ADEC%20Background%20Document.pdf.

Andersen, M.E., Jones, R.A., Mehl, R.G., Hill, T.A., Kurlansik, L., and Jenkins Jr., L.J. 1977. The inhalation toxicity of sulfolane (tetrahydrothiophene-1,1-dioxide). Toxicol. Appl. Pharmacol. 40(3):463-472.

ARCADIS. 2012. Revised Draft Final Human Health Risk Assessment. Flint Hills North Pole Refinery, North Pole, Alaska. Project No. B0081981.0008.00005. May 2012.

ARCADIS. 2014a. Supplement to the Revised Draft Final Human Health Risk Assessment. Flint Hills North Pole Refinery. ARCADIS U.S., Inc. Chelmsford, MA. May 30.

ATSDR (Agency for Toxic Substances and Disease Registry). 2010. Health Consultation: Sulfolane. Division of Toxicology and Environmental Branch, Agency for Toxic Substances and Disease Registry.

ATSDR (Agency for Toxic Substances and Disease Registry). 2011. Health Consultation: Sulfolane. Division of Toxicology and Environmental Branch, Agency for Toxic Substances and Disease Registry.

Burnham, K.P., Anderson, D.R., 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol. Methods Res. 33:261-304

CCME (Canadian Council of Ministers of the Environment). 2006. Canadian Environmental Quality Guidelines for Sulfolane: Water and Soil. Scientific Supporting Document. Canadian Council of Ministers of the Environment, Winnipeg, Manitoba.

Cheng, C.K., Chan, J., Cembrowski, G.S., and van Assendelft, O.W. 2004. Complete blood count reference interval diagrams derived from NHANES III: stratification by age, sex, and race. Lab. Hematol. 10(1):42-53.

Crump, K.S., 1995. Calculation of Benchmark Doses from Continuous Data. Risk Analysis 15, 79-89.

EPA (United States Environmental Protection Agency). 2002. A Review of the Reference Dose and Reference Concentration Processes. Risk Assessment Forum, U.S. Environmental Protection Agency. EPA/630/P-02/002F. Available at: http://www.epa.gov/raf/publications/pdfs/rfd-final.pdf

EPA (United States Environmental Protection Agency). 2003. Integrated Risk Information System (IRIS). Benzene (CASRN 71-43-2). U.S. Environmental Protection Agency. Available at: http://www.epa.gov/iris/subst/0276.htm [last accessed on November 7, 2014]

EPA (United States Environmental Protection Agency). 2011. Recommended use of body weight 3/4 as the default method in derivation of the oral reference dose, U.S. Environmental Protection Agency. EPA/100/R11/0001. Available at: http://www.epa.gov/raf/publications/interspecies-extrapolation.htm [last accessed on December 14, 2014]

EPA (United States Environmental Protection Agency). 2012a. Provisional Peer Reviewed Toxicity Values for Sulfolane (CAS No. 126-33-0). National Center for Environmental Assessment (NCEA), Superfund Health Risk Technical Support Center, January 30.

EPA (United States Environmental Protection Agency). 2012. Benchmark Dose Guidance. Risk Assessment Forum, U.S. Environmental Protection Agency. EPA/100/R-12/001. Available at: http://www.epa.gov/raf/publications/pdfs/benchmark_dose_guidance.pdf

EPA (United States Environmental Protection Agency). 2014. Integrated Risk Information System (IRIS). Review of Reference Doses in IRIS database. U.S. Environmental Protection Agency. Available at: http://www.epa.gov/IRIS/

Gradient Corp. 2014. Review and Verification of Existing Sulfolane Dose-Response Assessments. Gradient Corp, Cambridge, MA.

Health Canada. 2014. Drinking Water Guideline Value for Sulfolane. March 17, 2014. Health Canada.

HLS (Huntingdon Life Sciences). 2001. Sulfolane toxicity study by oral administration via the drinking water to CD rats for 13 weeks. Huntingdon Life Sciences Ltd, Huntingdon, United Kingdom.

IPCS (International Programme on Chemical Safety). 2005. Chemical-Specific Adjustment Factors For Interspecies Differences And Human Variability: Guidance Document For Use Of Data In Dose/Concentration–Response Assessment. World Health Organization (WHO). ISBN 92 4 154678 6.

Kiecolt-Glaser, J.K., Glaser, R., Shuttleworth, E.C., Dyer, C.S., Ogrocki, P. and Speicher, C.E. 1987. Chronic stress and immunity in family caregivers of Alzheimer's disease victims. Psychosom. Med. 49(5):523-535.

Luster, M.I., Germolec, D.R., Parks, C., Blanciforti, L., Kashon, M., and Luebke, R. 2005. Are changes in the immune system predictive of clinical disease? Models for changes Immunotoxicology. In: *Investigative Immunotoxicology*. Tryphonas, H., Fournier, M., Blakley, B., Smits, J., and Brousseayey, P. (Eds.) CRC Press LLC, Boca Raton, FL. Chapter 11 pp. 165-182.

Magee, B. 2012. Memo to Flint Hills Resources Alaska. Assessment of Dose Response Information for Sulolane. Aracadis Project No. B0081981.0029. Sent May 12, 2012.

Ministry of Health and Welfare Japan. 1996. Sulfolane: 28 day repeat dose oral toxicity test. In: Toxicity testing reports of environmental chemicals. Tokyo, Japan. pp. 437−445.

Ministry of Health and Welfare Japan. 1999. Sulfolane. In: Toxicity testing reports of environmental chemicals. Tokyo, Japan. pp. 473−481.

Parks, C.G., Andrew, M.E., Blanciforti, L., and Luster, M.I. 2007. Variations in the white blood cell counts and other factors associated with reporting of herpes labialis: a population based study of adults. FEM Immunol. Med. Microbiol. 51(2):336-343.

Penn, I. 2000. Post-transplant malignancy: the role of immunosuppression. Drug. Saf. 23(2):101-113.

Shearer, W.T., Easley, K.A., Goldfarb, J., Rosenblatt, H.M., Jenson, H.B., Kovacs, A., and McIntosh, K. 2000. Prospective 5-year study of peripheral blood CD4+, CD8+, and CD19+/CD20+ lymphocytes and serum Igs in children born to HIV-1+ women. J. Allergy. Clin. Immunol. 106(3):559-566.

Storek, J., Espino, G., Dawson, M.A., Storer, B., Flowers, M.E., and Maloney, D.G. 2000. Low B-cell and monocyte counts on day 80 are associated with high infection rates between days 100 and 365 after allogeneic marrow transplantation. Blood. 96(9):3290-3293.

TCEQ (Texas Commission on Environmental Quality). 2011. Sulfolane. CASRN 126-33-0. September 6, 2011. Report by Joseph Haney, Senior Toxicology. Texas Commission on Environmental Quality.

TCEQ (Texas Commission on Environmental Quality). 2012. TCEQ Guidelines to Develop Toxicity Factors. Toxicology Division, Office of the Executive Director. RG-442. Available at: https://www.tceq.texas.gov/publications/rg/rg-442.html

Thompson, C.M., Gaylor, D.W., Tachovsky, J.A., Perry, C., Carakostas, M.C., and Haws, L.C. 2013. "Development of a chronic noncancer oral reference dose and drinking water screening level for sulfolane using benchmark dose modeling." J Appl Toxicol 33(12):1395-1406.

Wignall, J.A., Shapiro, A.J., Wright, F.A., Woodruff, T.J., Chiu, W.A., Guyton, K.Z., and Rusyn, I. 2014. Standardizing benchmark dose calculations to improve science-based decisions in human health assessments. Environ. Health. Perspect. 122:499–505.

World Health Organization (WHO). 2005. Chemical-Specific Adjustment Factors For Interspecies Differences And Human Variability: Guidance Document For Use Of Data In Dose/Concentration–Response Assessment. International Programme for Chemical Safety, World Health Organization. ISBN 92 4 154678 6.

Yang, E.V. and Glaser, R. 2000. Stress-induced immunomodulation: impact on immune defenses against infectious disease. Biomed. Pharmacother. 54(5):245-250.

Zhu, Z., Sun, M., Li, Z., Yang, Z., Zhang, T., Heng, Z., Xiao, B., Li, Q., Peng, Q., Dong, Y., Jiang, S., and Jiang, J. 1987. An investigation of maximum allowable concentration of sulfolane in surface water. J West China Univ. Med. Sci. 18(4):376-380.