Position Paper
Proposed November 20, 2014

TCEQ Recommendations for Systematic Review and Evidence Integration

Prepared by

Neeraja K. Erraguntla, Ph.D., DABT

Heather R. Reddick, MPH

Office of the Executive Director

TEXAS COMMISSION ON ENVIRONMENTAL QUALITY

Document Description and Intended Use

This document provides guidelines on how to conduct systematic review and evidence integration when developing chemical-specific reference values (ReVs), unit risk factors (URFs), reference dose (RfD), slope factors (SF_o), and effect screening levels (ESLs). It is a supplement to the TCEQ Regulatory Guidance-442 (RG-442), *TCEQ Guidelines to Develop Toxicity Factors (TCEQ 2012).*

Since the TCEQ has developed guidance to develop chemical-specific toxicity factors, there has been an increased interest in developing systematic review guidelines that include clear study inclusion and exclusion criteria and explicit criteria for determining study quality prior to identifying a key study. Also, because data comes from diverse evidence streams (e.g., human clinical data, epidemiological data, animal toxicological studies, and mechanistic data), there is a need to evaluate and integrate the information from the multiple streams to improve the decision-making process, increase transparency, minimize bias, and improve consistency amongst different risk assessments. Systematic review and evidence integration can also help improve confidence in establishing causality, a critical component of risk assessments.

TABLE OF CONTENTS

TABLES

FIGURE

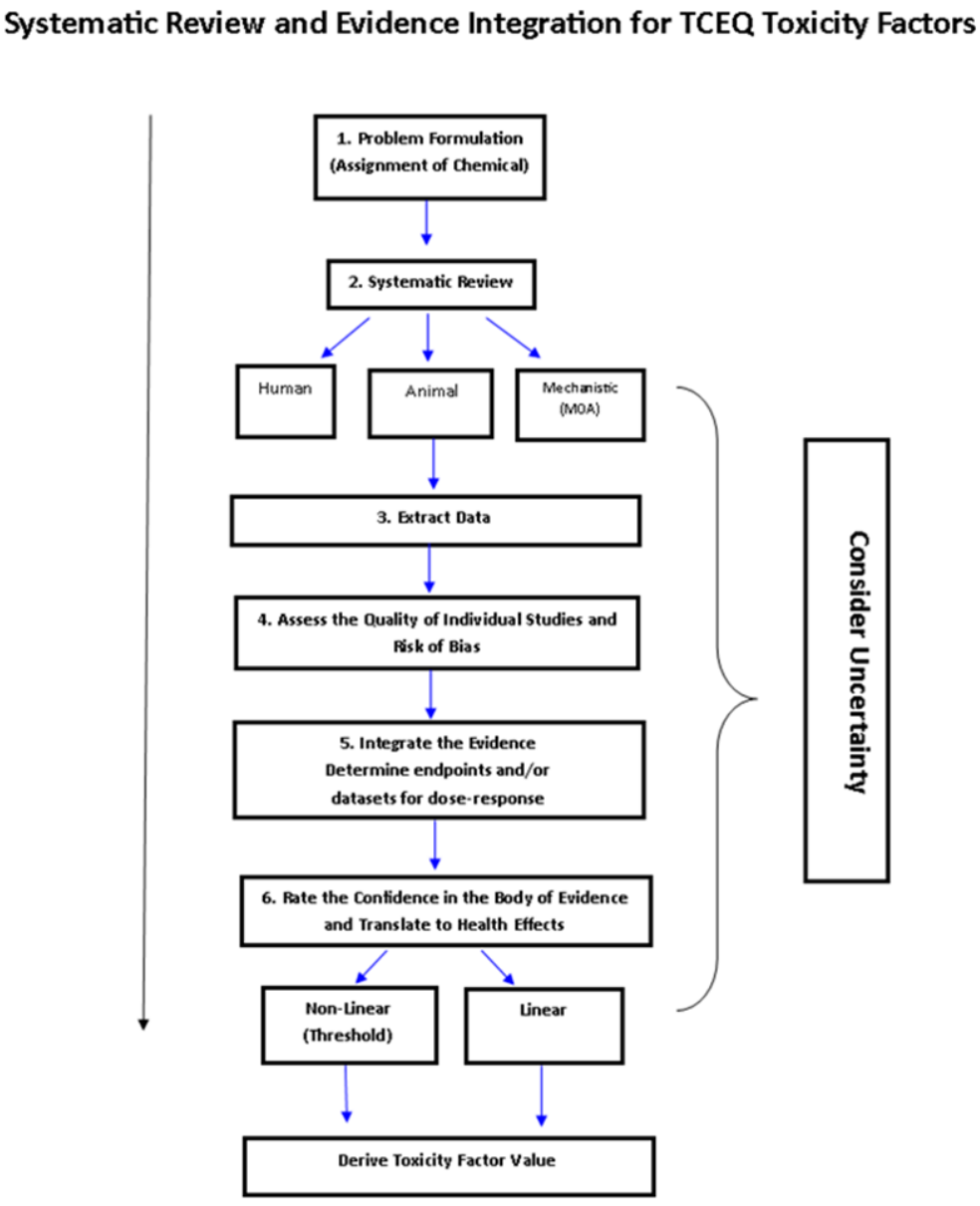
Systematic Review and Evidence Integration

Introduction

The National Research Council (NRC) recently released its evaluation of the United States Environmental Protection Agency's (USEPA) Integrated Risk Information System (IRIS) program (May 6, 2014) in which they suggested the term "Evidence Integration" as opposed to weight of evidence "WOE" (NRC 2014). The TCEQ agrees with this terminology and will use principals of evidence integration when conducting WOE analysis. The NRC also suggested the need to use better systematic literature review methodologies and reiterated the importance of using high-quality studies and evaluating risk of bias in their evaluations. Additionally, the Office of Health Assessment and Translation (OHAT), Division of the National Toxicology Program (NTP) of the National Institute of Environmental Health Services (NIEHS) recently published their method for conducting systematic review and evidence integration for reaching hazard identification conclusions (Rooney et al. 2014). The recent 2014 Society of Toxicology annual conference devoted a workshop session to the topic that sparked interest from a variety of stakeholders on how to integrate data from various data streams using a systematic review approach. In summary, there are several recent publications that have discussed the issues and have proposed best practices for conducting systematic reviews and WOE analysis (Rhomberg et al. 2013, NRC 2014, Rooney et al. 2014).

The present framework was developed to help TCEQ toxicologists conduct systematic reviews and evidence integration and will supplement TCEQ's 2012 published guidance on deriving toxicity factors (RG-442). One of the biggest strengths of the systematic literature review approach is that it will document clear inclusion and exclusion criteria. This step will help toxicologists document why particular studies were chosen as potential key studies and the reasons for excluding other studies (i.e., excluding them as key studies or completely excluding the studies from the review). This step can help with improving transparency, and subsequently help improve risk communication to a wide range of stakeholders. Figure 1 depicts the proposed systematic review and evidence integration process.

Figure 1 Systematic Review and Evidence Integration Flow Chart



Step 1: Problem Formulation

The first step in the proposed systematic review and evidence integration process (Figure 1) is Problem Formulation (PF). This step identifies and describes the causal question and describes the extent of the evaluation. The PF contains elements that promote transparency and consistency and accommodates different biologically plausible hypotheses (Rhomberg et al. 2013). In other words, the PF is a statement that includes specific questions pertinent to all of the steps of the systematic review process including the literature search, study selection, data extraction, and synthesis.

The PF is structured around causal questions: "does this chemical cause a hazard? And, if it does, what are the safe levels of exposure?". For example, Johnson et al. (2014) used the following query to define their research question: “Does fetal developmental exposure to perfluorooctanoic acid (PFOA) affect fetal growth in humans?”

Step 2: Systematic Review and Selecting Studies for Inclusion

2.1 Systematic Literature Review

Several months prior to the start of work on a Development Support Document (DSD), the TCEQ TD will announce via its email listserve that it is soliciting information for a particular chemical or class of chemicals. Refer to Chapter 1 of the RG-442 for more detailed information on the selection of chemicals and data solicitation for DSDs. Interested parties are encouraged to provide citations or toxicological information. Some studies are conducted following Good Laboratory practices (GLP); however, they may not be published in the scientifically peer-reviewed literature. Although studies published in peer-reviewed journals are preferred, the TD makes every attempt to critically review all toxicological studies submitted by interested parties prior to deciding to use them.

Systematic reviews are increasingly being recognized as important steps in establishing causality and therefore are becoming integral for risk assessment. Specifically, systematic reviews mandate precise pre-determined search criteria for the inclusion and exclusion of studies and a selection of key and supporting studies. The use of the pre-determined criteria will help promote transparency in the decisions to either include or exclude a study that in turn will help improve the transparency of the scientific judgment (Figure 1). Because of the clear and transparent nature of the search, this step can be easily reproduced by other researchers if needed, which in turn can improve confidence and reduce uncertainty in the derivation of the toxicity factors and in determining causality.

As a first step, publically available databases [PUBMED, TOXNET, ATSDR Toxicological Profiles, CalEPA’s OEHHA RELs and supporting documents, USEPA’s IRIS assessments, and Developmental and Reproductive Toxicology Database (DART)] are searched using explicitly

stated search criteria. Refer to Section 3.3.2.1 of the TCEQ Guidance RG-442 for a detailed list of database links to conduct database searches. This checklist is a living document and other sources, and databases will be added if deemed necessary and applicable to the PF. In addition, Lu (2011) published a useful review of the different web tools available for searching biomedical literature (http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/search/).

2.2 Inclusion and Exclusion Criteria

Clear and succinct inclusion and exclusion criteria need to be specified to identify the initial study database from which key and supporting studies are selected. The inclusion and exclusion criteria are to be decided based on the specific questions that form the basis of the PF. For example, the criteria are based on applicable outcomes, relevant exposures, durations, and the types of studies. Studies that contribute to identifying the critical effects relevant to humans are selected as studies for possible review. Using explicit criteria to select or omit studies helps to balance scientific judgment by providing clear and transparent documentation.

It is recommended to use Boolean operators such as: “AND” which will look for articles that include all the identified keywords;

“OR” which will look for articles that contain any of the defined keywords;

and “NOT” to exclude articles that contain a particular keyword.

For example, Bailey et al. (2009) conducted a literature search in PubMed and they limited the search to human studies. They used the following search terms: “manganese AND (neuro* OR neurotoxin*OR neurology OR neurologic*). They also specified that the studies had to be published after 1992; examined and reported manganese dust as the exposure of concern from personal air monitoring data; evaluated both exposed and unexposed populations; and evaluated neurological effects in relation to ongoing exposures to manganese in air.

Step 3: Data Extraction

Data extraction is the third step in the flow chart (Figure 1). Studies that meet the inclusion criteria are further critically reviewed, and extracted information from these studies are summarized into tables (evidence tables). These tables can be simple and can be created using MS Word, Excel, or Access, or can be commercially available databases that are discussed later in the section. The purpose of these tables or databases is to provide a summary table to display the data in order to identify potential trends and provide a basis to use the data as evidence. For example, the National Institute of Environmental Health Sciences (NIEHS) maintains several databases including the Chemical Effects in Biological Systems (CBES) database that can be useful resources for researchers (http://www.niehs.nih.gov/research/resources/databases/cebs/). According to NIEHS, the CEBS database houses data of interest to environmental health experts. CEBS is a public resource and has received depositions of data from academic, industrial, and

governmental laboratories. CEBS is designed to display data in the context of biology and study design, and to permit data integration across studies for novel meta-analysis.

Data extraction will differ for each data stream because of differences in study design, methodologies, and data quality. Epidemiology studies include experimental and observational (analytical and descriptive) studies. *In vivo* animal toxicity studies are conducted to determine dose-response and are usually conducted for particular durations (acute, subacute, subchronic, chronic), or to study a specific effect (carcinogencity, reproductive/developmental, or neurological). Mechanistic studies are useful to understand the mode-of-action (MOA). *In vitro* studies are often conducted to determine genotoxicity, cell transformation, cytotoxicity, or can be mechanistic in nature. It is important that toxicity factors are based on the most reliable information available so that the values reflect the most scientifically supported information on the potential hazards of the chemical.

Currently, there are several tools being developed to help inform decisions and transparently document the systematic review process. These tools will help maintain references in one place, and also group them based on the selection criteria. These tools can be powerful because they allow query of the databases and improve transparency of why certain studies were included and excluded. For example, ICF International created the DRAGON database to allow scientists to transparently record their decisions and rationales, and reduce document production time: http://www.icfi.com/insights/products-and-tools/dragon-dose-response. In addition, HAWC (Health Assessment Workspace Collaborative) is another software tool the OHAT is developing to help manage systematic review and data display. It is a modular web-based interface to facilitate the development of human health assessments of chemicals and is currently being developed: https://hawcproject.org/.

Step 4: Assessing the Quality of Individual Studies and Risk of Bias (ROB)

Assessing data quality is a critical step in risk assessment (Figure 1). Studies that meet the inclusion criteria should be critically evaluated for study quality and risk of bias (ROB). Section 3.3.3.1 of the RG-442 guidance briefly discusses that data quality evaluation should consider method validity, reproducibility, study reliability, dose-response relationships, temporal associations between exposures and adverse health effects, and understanding if critical effects are relevant to humans. ROB is a concept that was defined by the Institute of Medicine (IOM), as the "extent to which flaws in the design and execution of a collection of studies could bias the estimate of effect for each outcome under the study" (IOM 2001 as described in NAS 2014). According to NAS (2014), bias is defined as an error that decreases validity, and ROB refers to the potential for bias to occur.

Although, study quality and ROB are interrelated to some extent, they are actually different concepts. There is a large amount of information on these topics. The goal of this position paper

is to provide guidance to the TCEQ staff based on the available best practices on how to evaluate study quality and ROB in both animal and human studies.

The NRC review of the IRIS assessment recommends treating the terms separately. The NTP OHAT review defines study quality broadly to include three main elements out of which ROB is one of the elements: 1) reporting quality which relates to the way the study was reported; 2) internal validity or ROB which refers to how plausible the results of the study are and will depend on how the study was designed and conducted; and 3) external validity or directness of applicability, which refers to evaluating if the study is pertinent and applicable for the particular issue being considered. The NTP OHAT review provides a detailed set of ROB questions that are discussed in Section 4.2 and can be useful for TCEQ staff (Rooney et al. 2014).

4.1 Study Quality

Risk assessments often include data from different streams of data (e.g., animal studies, human chamber studies, epidemiology studies). Each of these categories is different from the other in study design, study protocol, exposure, and species (e.g., laboratory animal studies vs. epidemiology studies). Although it is accepted that study quality is a critical component of risk assessment, there are no specific guidelines on how to assess study quality for the different data streams collectively. Additionally, it is very difficult to define a distinct set of rules across the different types of studies. Recently, there has been growing interest on the subject, and several journal articles have been published; however, many of the published studies have addressed only a few categories. For example, some of the available references evaluate study quality specifically for animal studies (Klimisch et al. 1997) or human studies (Money et al. 2013). However, there are only a few frameworks on how to evaluate and integrate evidence from both animal and human studies (Lavelle et al. 2012, Rhomberg et al. 2013, NRC 2014). This position paper will briefly summarize the highlights of these references.

4.1.1Animal Studies:

Klimisch et al. (1997) proposed a systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. The authors identified three categories (Reliability, Relevance, and Adequacy) to evaluate data quality in animal studies; however, they focused only on the Reliability category to determine the Klimisch score (Table 1). Relevance and Adequacy were not evaluated. Based on Klimisch et al. (1997), the three categories can be defined as:

- **Reliability** — assessing the inherent quality of the test report or publication relating to preferably standardized methodology and the way that the experimental procedure and results are described to give evidence of the clarity and plausibility of findings.
- **Relevance** — covering the extent to which data and/ or tests are appropriate for a particular hazard identification or risk characterization.

- **Adequacy** — defining the usefulness of data for risk assessment purposes. When there is more than one set of data for each effect, the greatest weight is attached to the most reliable and relevant.

Klimisch et al. (1997) state the obvious in that the more details provided on procedures, methodology and analytics, the more reliable and thorough the evaluation will be. In addition, the authors recommend that data reported in compliance with the principles of GLP should have the highest grade of reliability. By using Klimisch codes in evaluating study data, the information gathered is ordered so that the most reliable and relevant studies are assessed and then used. The following category of reliability codes are defined in Table 1.

Table 1. Reliability Codes and Categories (Klimisch 2007, Adapted from Bevan & Strother 2012)

Code	Category	Justification
1	Reliable without restriction	• Guideline Study • Comparable to guideline study • Test procedure according to national standards
2	Reliable with restrictions	• Acceptable, well-documented publication/study which meets basic scientific principles • Basic data given: comparable to guidelines/standards • Comparable to guideline study with acceptable restrictions
3	Not reliable	• Method not validated • Documentation insufficient for assessment • Does not meet important criteria of today standard methods • Relevant methodological deficiencies • Unsuitable test system
4	Not Assignable	• Only short abstract available • Only secondary literature

As mentioned previously, the Klimisch codes do not apply to "Relevance" and "Adequacy" categories; however, the TCEQ considers the following aspects to be important when evaluating data quality.

Relevance: As mentioned in Section 3.3.3.1 of RG-442, studies that contribute most significantly to the WOE (evidence integration) and that identify critical effects relevant to humans are selected as key studies. For example, inhalation exposure studies usually take precedence over oral exposure studies for determining inhalation toxicity factors and oral exposure studies typically take precedence over inhalation studies for determining oral toxicity factors. In addition, in the absence of adequate human data, animal studies and critical effects from the animal studies that are of known or

likely human relevance are preferred as key studies. Section 3.3.3.4, Section 3.4, and Figure 3-1 of the RG-442 guidance depicts the main steps in evaluating the human relevance of an animal MOA to humans.

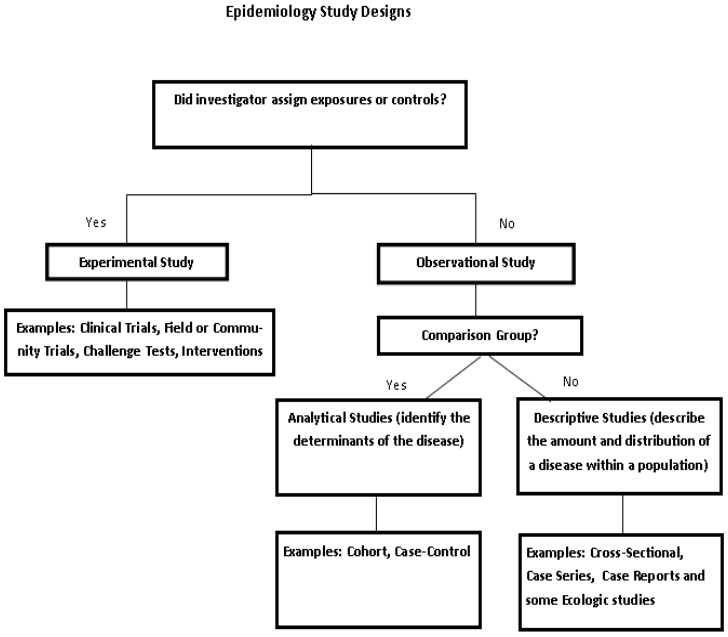
Adequacy: This review is pertinent to chemicals that are generally defined as data rich. For chemicals that are limited in data, please review Section 3.15 of RG-442.

4.1.2 Human Studies

There is an increased interest in incorporating human data in chemical risk assessments due to various initiates such as the World Health Organization's International Programme on Chemical Safety (IPCS) and European Union's Regulation on Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) initiative. Human studies are preferred over animal studies when developing toxicity factors as the need to conduct animal-to-human extrapolation (e.g., dose, effect) is unnecessary. However, while there is guidance on how to conduct human epidemiology studies, there is limited guidance on evaluating the integrity of the study designs and interpretation of the findings.

As mentioned in Section 3.3.3.3 of RG-442, epidemiology studies provide data regarding associations between exposure and health effects that are useful in hazard identification, and if accompanied by sufficient, accurate and reliable exposure data, may be useful in the dose-response assessment for a toxicant. Epidemiological studies may be descriptive, analytical, or experimental in design. Descriptive studies can involve populations (ecological studies) or individuals (case reports and cross-sectional studies). Analytical designs, where individuals are also the units of observation, include observational studies (cross-sectional, case-control, and cohort studies), and experimental designs include randomized clinical trials, field or community trials, challenge tests (i.e. human chamber studies), and interventions (see Figure 2). Typically, observational study designs are the most common human studies used when determining environmental impacts on health outcomes (Rushton and Elliot 2003). A brief summary of the strengths and limitations of epidemiology studies are included in Table 2 of this document. In addition, Section 3.3.3.3 of RG-442 provides a brief summary of the different studies. The following

Figure 2 – Epidemiology Study Designs (adapted from Rushton & Elliot 2003 and Grimes & Schulz 2002).



information is provided as supplemental information to help staff compliment their toxicological expertise with epidemiology data.

Epidemiology studies indirectly evaluate causality through varying exposures; therefore, it is imperative to select useful, well-designed studies (Künzli & Tager 2007). Study designs can differ based on sample size and availability of subjects, units of observation, data collection methods, and directionality of exposure. Table 3 of this document provides a general sequence of research efforts in epidemiology and a hierarchy based on the overall strengths, limitations, and validity of study designs. For example, ecological studies and case-reports are in the lowest tier of the hierarchy because they lack controlled exposure, there is less confidence that exposure occurs prior to the outcome, individual data may or may not be available, and they are of little use for etiologic inference (Künzli & Tager 2007).

Epidemiology data can complement and enhance the evidence from toxicological studies; however, epidemiological data often lack exposure information and may have confounding issues and bias. Critical issues in the lack of exposure data include the type of assessment method used, patterns of exposure over time, and the metric used to represent exposure data (Rushton & Elliot 2003.) These issues can reduce confidence due to more uncertainty. Also, controlled experimental exposures rarely occur in epidemiology studies; therefore, gathering exposure data is often not feasible. Controlled exposures that occur in experimental human studies can be extremely useful and are preferred over observational epidemiology studies as they provide evidence of exposure and effect (i.e., cause-and-effect) while potential confounding can be identified and controlled for; however, there are also limitations. For example, human controlled exposure studies generally involve small sample sizes. Also, due to the nature of noninvasive methods and ethical considerations, exposures are limited to low exposure levels and only minor and reversible effects are studied (Rushton & Elliot 2003).

Strengths, weaknesses, and ROB should be weighed prior to making a causal association based on epidemiology studies. Further, it is important to note that statistically significant results should not be automatically deemed as evidence of a causal association (e.g., confounders may not be adequately controlled/adjusted for). A positive association does not necessarily imply causation (Phillips et al. 2004). A detailed review of the Bradford Hill criteria should be conducted and will be discussed in Step 5 of this document. Other types of epidemiological data include disease registries, surveillance data, data from poison centers, and information from the health department

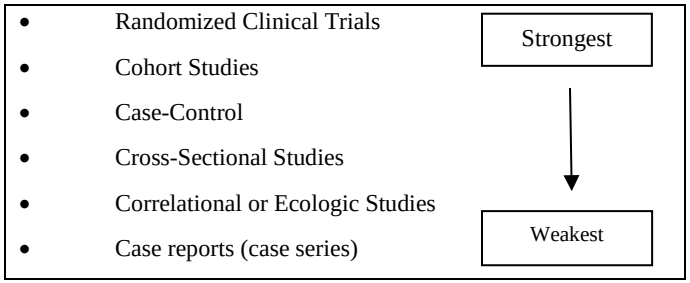
Table 2: Strengths and Weakness of Human Studies (exposure rating in body of evidence schematic adapted from Table S2 of the OHAT Approach, Rooney et al.2014).

Type of Study	Definition	Strengths	Weaknesses	Controlled Exposure	Exposure Prior To Outcome	Individual Outcome data	Comparison Group Used
Case Report	A case report is a descriptive study that describes and interprets single individual (case report) or small group (case series) cases based on detailed clinical evaluations and histories of the individual(s).	Provides rare information. Can help identify new diseases. Simple count.	Subject to selection bias. No control group. Findings may not be generalizable to a larger population.	Unlikely	Unlikely	May or May not	Unlikely
Ecologic (Szklo & Nieto 2007)	In an ecologic study, correlations are obtained between exposure rates and disease rates among different groups or populations.	Existing population data. Useful if exposure frequency varies between populations.	Ecologic Fallacy – Bias that may occur because an association observed between variables on an aggregate level does not necessarily represent the association that exists on an individual level. Imprecise measurement of exposure and disease.	Unlikely	Unlikely	Unlikely	Unlikely
Cross Sectional (Szklo & Nieto 2007)	A cross-sectional study design examines the relationship between disease and other variables of interest as they exist in a sample of (or the total) reference population at a given point in time.	Measures disease prevalence. Odds Ratio can be determined. Can study non-fatal diseases and effects on physiologic variables. Studies are often population based. Studies can be incorporated into	Less appropriate for investigating causal associations than cohort or case-control studies. Temporal sequence problematic (exposure and effect data is collected simultaneously). Problems with misclassification of disease.	Unlikely	Unlikely	Likely	Likely

		ongoing surveillance programs.					
Case-Control **(Szklo & Nieto 2007)**	**A case-control study compares diseased individuals-cases and non-diseased individuals-controls) with respect to their level of exposure to a suspected risk factor.**	**Can study rare outcomes.** **Can examine more than one risk factor.** **Odds Ratio can be determined.**	**Inefficient for rare exposures.** **May not have information on confounding factors.** **Recall bias.** **Problems with temporal sequence.** **Difficulties in identifying representative case/control groups (Selection Bias).**	**Unlikely**	**May or May not**	**Likely**	**Likely**
Cohort **(Szklo & Nieto 2007)**	**Two or more groups of people, who are free of disease and differ according to extent of exposure to a potential cause of disease, are compared with respect to incidence of disease in each group. The objective of a cohort study is to investigate whether the incidence of an event is related to a suspected exposure. Cohort studies can be prospective and retrospective in nature.**	**Lack of bias in determining exposures.** **Yields incidence rates as well as relative risks.** **Multiple outcomes can be investigated.** **Next best design to experimental studies.** **Correct temporal sequence maintained.**	**Loss of follow up.** **Large numbers of subjects are required.** **Problem of attrition.** **Changes over time of criteria and methods due to long follow up.** **Inefficient for rare diseases.**	**Unlikely**	**May or May not**	**Likely**	**Likely**
Human Controlled Studies	**Investigator intentionally alters one or more exposures to study outcome effects.**	**Provides control over the dose and frequency of exposure and the period of observation.**	**Small sample sizes.** **Artificial setting.**	**Likely**	**Likely**	**Likely**	**Likely**

		Considered the gold standard.	**Limited scope of potential impact.** **Sensitive sub populations are often not included.** **Ethical dilemmas.**				

Table 3: General sequence of research efforts in Epidemiology (Adapted from Künzli & Tager 1997).



As mentioned previously, a consensus among the scientific community on how to evaluate and rate different types of epidemiology studies is needed. Money et al. (2013) have proposed a systematic review for evaluating and scoring human data that builds on previously published information as proposed by Klimisch et al. (1997) for animal studies. The authors adapted the reliability scores to human studies to provide a comparable categorization in addressing evidence integration. However, the authors note that the interpretation of human data is not as straightforward as animal data due to variability in study designs, human genetic variation and the importance of accounting for confounding and bias. Therefore, assigning quality scores to human data is a challenge and professional judgment is a key factor in the process although it is subjective in nature. Table 2 of Money et al. (2013) focuses on assessing and assigning the reliability scores to human data and can be used by toxicologists as a guide. In addition, the Critical Appraisal Skills Programme (CASP) offers a set of downloadable lists that can be used as a guide when reviewing the various types of epidemiology studies. The CASP checklists are designed to assist the reader in thinking about the studies systematically while considering the broad issues of the study designs (CASP 2013).

4.1.3 *In Vitro* Studies

Traditional risk assessments that rely primarily on *in vivo* testing have several limitations. For example, *in vivo* testing typically focuses on apical endpoint testing that makes the whole toxicity testing process very resource intensive and expensive. The time and expense needed for *in vivo* toxicity testing are often inadequate for testing the vast influx of chemicals in commerce. *In vitro* testing has steadily gained popularity because *in vitro* assays in theory can generate molecular, biochemical, or histological data and can provide information on perturbations of critical pathways that can provide additional information on the toxicity of a specific chemical. These perturbed pathways are also called Adverse Outcome Pathways (AOPs). *In vitro* assays can also be easily scaled to high-throughput systems and therefore can potentially be used to screen a large number of chemicals in a short period of time.

In summary, *in vitro* assays can provide useful mechanistic information. However, there is insufficient evidence regarding translation of pathway perturbations to quantifiable adverse effects. According to the Office of Research and Development (ORD) of the US EPA, a critical challenge to using this type of mechanistic information is to translate it to outcomes relevant to risk assessment and risk management objectives (i.e., protection of individuals or populations) (http://www.epa.gov/nheerl/articles/2011/Chemical_Safety_Assessments.html).
Amongst the many publications and guidance documents regarding *in vitro* assays is the 2007 report from the National Academy of Sciences titled *Toxicity Testing in the 21st Century: A Vision and a Strategy, advances in molecular biology* (e.g., genomics) (NAS 2007). In addition, a collaborative effort amongst many agencies including the European Commission, the US EPA, the US Army Corps of Engineers, the Office of Economic Cooperation and Development (OECD) resulted in the birth of a WIKI called the “Adverse Outcome Pathway “(AOP) WIKI. The AOP is a conceptual framework that provides a scientific approach to linking mechanistic

information to responses considered relevant to risk assessment and management and can prove to be a useful supplemental tool for TCEQ toxicologists.

4.2 ROB

The Rooney et al. (2014) review provided a comprehensive set of questions to address ROB for the different streams of data including experimental animal studies, human chamber studies, and epidemiology studies (Table 4). These questions are part of a framework that underwent extensive peer-review and are also very pertinent to the TCEQ's chemical risk assessment program. These questions are therefore reproduced here to help staff evaluate ROB. The TCEQ guidance will adopt these recommendations for ROB as they allow for comparison of the different streams of data and will allow a quick overview of evaluating the strengths and limitations of the different types of data.

Table 4: Risk of Bias Questions (Table 1 from Rooney et al. 2014)

Bias categories and questions	**Applicable study designs**
Selection bias	
Was administered dose or exposure level adequately randomized? Randomization requires that each human subject or animal had an equal chance of being assigned to any study group including controls (e.g., use of random number table or	**Experimental Animal, Human Controlled Trial**
Was allocation to study groups adequately concealed? Allocation concealment requires that research personnel do not know which administered dose or exposure level is assigned at the start of a study. Human studies also require that allocation be concealed from human subjects prior to entering the study. *Note: 1) a question under performance bias addresses blinding of personnel and human subjects to treatment during the study; 2) a question under detection bias addresses blinding of outcome assessors.*	**Experimental Animal, Human Controlled Trial**
Were the comparison groups appropriate? Comparison group appropriateness refers to having similar baseline characteristics between the groups aside from the exposures and outcomes under study.	**Cohort, Case Control Cross Sectional**
Confounding bias	
Did the study design or analysis account for important confounding and modifying variables? *Note: a parallel question under detection bias addresses reliability of the measurement of confounding variables.*	**All**
Did researchers adjust or control for other exposures that are anticipated to bias results?	**All**
Performance bias	
Were experimental conditions identical across study groups?	**Experimental Animal**

Did researchers adhere to the study protocol?	**All**
Were the research personnel and human subjects blinded to the study group during the study? Blinding requires that study scientists do not know which administered dose or exposure level the human subject or animal is being given (i.e., study group). Human studies require blinding of the human subjects when possible.	**Experimental Animal, Human Controlled Trial**
Bias categories and questions	**Applicable study designs**
Attrition/exclusion bias	
Were outcome data complete without attrition or exclusion from analysis? Attrition rates are required to be similar and uniformly low across groups with respect to withdrawal or exclusion from analysis.	**Experimental Animal, Human Controlled Trial, Cohort, Case Control, Cross Sectional**
Detection bias	
Were the outcome assessors blinded to study group or exposure level? Blinding requires that outcome assessors do not know the study group or exposure level of the human subject or animal when the outcome was assessed.	**All**
Were confounding variables assessed consistently across groups using valid and reliable measures? Consistent application of valid, reliable, and sensitive methods of assessing important confounding or modifying variables is required across study groups. *Note, a parallel question under selection bias addresses whether design or analysis account for confounding.*	**All**
Can we be confident in the exposure characterization? Confidence requires valid, reliable, and sensitive methods to measure exposure applied consistently across groups.	**All**
Can we be confident in the outcome assessment? Confidence requires valid, reliable, and sensitive methods to assess the outcome and the methods should be applied consistently across groups.	**All**
Selective reporting bias	
Were all measured outcomes reported?	**All**
Other	
Were there no other potential threats to internal validity (e.g., statistical methods were appropriate)? On a project specific basis, additional questions for other potential threats to internal validity can be added and applied to study designs as appropriate.	**Additional items as applicable by study design**

Experimental Animal studies are controlled exposure studies; **Human Controlled Trials** are carried out in humans using a controlled exposure, including randomized controlled trials and non-randomized experimental studies. **Cohort** studies include prospective studies that follow subjects free of disease over time or retrospective studies of subjects with prior information available. **Case Control** studies enroll subjects based on their disease status and compare exposures across the groups. **Cross Sectional** studies are conducted at one point in time and include population surveys with individual data [e.g., National Health and Nutrition Examination Survey (NHANES)] and population surveys with aggregate data (i.e., air pollution exposure estimated by ZIP code). All applies to Experimental Animal, Human Controlled Trial, Cohort, Case Control, Cross Sectional studies, as well as other study design types such as Case Reports or Case Series studies that lack a comparison group within the study.

Step 5: Evidence Integration

Evidence integration is a two-step process. In the first step, evidence from each stream of data (quality, relevance, and reliability) is identified. In the second step, the evidence from the individual streams is combined with the other streams of data (animal studies, human studies, and mechanistic). Please see the formaldehyde carcinogenic DSD as an example (TCEQ 2008).

Because chemicals differ vastly in the amount and quality of each stream of data it is almost impossible to have one set of rigid rules for evidence integration. Additionally, the different types of data also have different strengths and weakness. The NAS (2014) review provided a table (Table 5) that compares the common strengths and weakness of the different streams of data including human epidemiological, experimental animal, and mechanistic studies. This table is reproduced here to aid staff in evaluating the different types of studies. The challenge for regulatory agencies is to determine their objectives *a priori* so that evidence integration can be conducted in a transparent and consistent manner. Properly conducted evidence integration of all the available and different streams of data allows staff to rate the confidence in the body of evidence as a whole and reduces uncertainty when making causal determinations.

Table 5 : Common Strengths and Weaknesses of Human Epidemiologic (HE), Experimental Animal (EA), and Mechanistic (MECH) Studies for Hazard Identification (Table 6-1 from NAS 2014)

Source of Uncertainty	Strength	Weakness
Interspecies extrapolations	HE: Not applicable, because not needed. EA: Can use multiple species, and this provides a broad understanding of species differences. MECH: Can identify cellular, biochemical, and molecular pathways that are similar or different in humans and the test species and thus lend strength to the veracity of the extrapolation.	HE: Not applicable, because not needed. EA: Inherent weakness when interspecies extrapolation from animals to humans is required. MECH: For a given chemical, multiple mechanisms might be involved in a given end point, and it might not be evident how different mechanisms interact in different species to cause the adverse outcome.
Intraspecies extrapolation	HE: Often able to study effects in heterogeneous populations. EA: Effects seen during different life stages (such as pregnancy and lactation) can be evaluated. Use of transgenic animals can provide important mechanistic data. MECH: Observed differences between strains of a common test species (such as Fisher 344 rats and Sprague-Dawley rats) might be readily explained by different pathways. Comparison with human in vitro mechanistic data might allow better selection of the most appropriate animal model for predicting human response.	HE: Many studies involve occupational cohorts, which do not reflect the general population. EA: Often rely on a few strains in which animal genetics, life stage, diet, and initial health state are controlled. MECH: Putative mechanism of the adverse outcome might not be known, and mechanistic data might not reveal the basis of differences within a species.

Source of Uncertainty	Strength	Weakness
High-dose to low-dose extrapolation	HE: Often better suited for considering actual range of population exposures. EA: Wide range of exposures is possible, and this allows better estimation of quantitative dose-response relationships. MECH: Dose-related differences in ADME properties and pharmacodynamic processes might be used to adjust for differences in rate of response between high and low doses.	HE: Occupational exposure is often higher than that seen in the general human population. EA: Exposures used are often orders of magnitude higher than those seen in the general human population. MECH: The ultimate molecular target for toxicity might not be known at low or high doses, so mechanism might not accurately predict high-dose to low-dose extrapolations.
Acute to chronic extrapolation (temporal considerations)	HE: Might closely mimic exposure durations seen in the general population. EA: Wide range of exposure durations is possible. MECH: Provides invaluable information on whether a product or effect can accumulate on repeated exposure and whether repair pathways or adaptive responses can lead to outcomes that are significantly different between single and repeated exposures.	HE: Occupational exposure durations are often shorter (years vs lifetime; 8 hr/day vs 24 hr/day) than those seen in the general human population. EA: Highly dependent on study design. MECH: If mechanism differs between acute or chronic response, the information on one might not be informative of the other.
Route-to-route extrapolation	HE: Often involve route of exposure relevant to the general human population. EA: Can involve route of exposure relevant to the general human population. MECH: Pharmacokinetic differences (ADME, PBPK) might facilitate more accurate identification of target-tissue dose from different exposure pathways.	HE: Data might be available on only one route of exposure. EA: Often uses an exposure method that requires extrapolation of data (for example, diet to drinking water). MECH: Mechanism might be tissue-specific and therefore route dependent as the route determines the initially exposed tissue.
Other considerations	HE: Can evaluate cumulative exposures and health effects. EA: Shorter animal lifespans allow for more rapid evaluation of hazards. Reduced misclassification of exposures and outcomes. Allows examination of full spectrum of toxic effects. MECH: Conservation of fundamental biologic pathways (such as cellcycle regulations, apoptosis, and basic organ-system physiology) might allow quick and inexpensive identification of potential adverse effects of a new chemical in the absence of human or animal in vivo data.	HE: Long lag time to identify some effects. Increased potential for exposure and outcome misclassification and confounding. Variable cost. EA: Multiple extrapolations required. Variable cost. MECH: Identification of relevant pathways in producing the toxic response might be difficult because of the lack of understanding of pathobiologic processes.

The Rooney et al. (2014) review provides a comparison of the main elements of the different streams of data and can be used as a template to evaluate the evidence and determine the confidence in each stream of data. For example, based on Table S2 (Table 2) from Rooney et al.

(2104) it appears that the human controlled trials and experimental animal studies have all of the elements of a good study and therefore confidence in using these data streams will result in higher confidence compared to using observational epidemiological studies. However, it has to be remembered that human controlled trials and animal studies have their own strengths and weakness that must also be evaluated.

The NAS (2014) review of the IRIS process recommends a combination of qualitative and quantitative approaches for integrating evidence. The qualitative approaches of integrating evidence include both informal (unstructured processes) and formal (structured) processes. The informal unstructured approaches rely on subject matter expertise (scientific judgment) to evaluate and integrate evidence. The structured qualitative approaches on the other hand rely less on scientific judgment and more on systematic documentation of the evidence which potentially makes the evidence evaluation and integration a more transparent process, albeit a resource and time intensive process. Examples of structured approaches include the following: Hill Criteria, Navigation Guide, Grading of Recommendations, Assessments, Development and Evaluation (GRADE) system, and the NTP criteria.

The Bradford Hill Criteria has been widely accepted to infer “causation” from statistical association and help differentiate “causation” from “association” (Phillips et al. 2004). Broadly, the Hill criteria include strength of association, consistency, specificity, temporality, biological gradient, plausibility, coherence, experimental evidence, and analogy. However, the Bradford Hill criteria have been reported to be of limited use in situations where the mechanism is unknown (Rotham and Greeland 2005 as cited in NRC 2014). Rhomberg et al. (2013) also advise caution in using the Hill criteria strictly as a checklist and recommend using them as guides for evaluating alternative explanations of the patterns of results.

The GRADE approach and the Bradford Hill criteria have been described to be analogous for determining the quality of evidence to justify causation (Schunemann et al. 2011). The NRC (2014) review defines GRADE as a system for rating the quality of evidence and the strength of recommendation. In the GRADE approach, the quality of evidence available for each outcome is evaluated based on several elements (e.g., ROB, inconsistency, indirectness etc) and a decision is made either to upgrade or downgrade the confidence in the evidence.

The strength of association as defined in the Hill criteria, for example, can be upgraded or downgraded based on the GRADE. However, the GRADE approach has often been used in evidence based medicine and in clinical-trials and has been proposed to be of limited use in chemical hazard assessments. The main reason for GRADE 's limited use in chemical risk assessment is because of the difference in the evidence streams between clinical medicine and environmental health.

Recently, a GRADE like approach (Navigation Guide) has been proposed for chemical hazard assessments that addresses the limitations present in the GRADE approach. The Navigation Guide was developed to assess reproductive and developmental risk on exposure to environmental chemicals (Woodruff and Sutton 2011 and 2014).

The NTP OHAT group also proposed similar criteria (NTP criteria) for integrating evidence from different streams (NTP 2013). Table 6-4 of the NAS (2014) report provides a comparison of the qualitative structured approaches that might be useful to TCEQ staff interested in learning more about the approaches.

In summary, the systematic documentation of evidence in the structured approaches can improve transparency and balance scientific judgment. However, the systematic documentation of evidence is a resource and time intensive step that cannot completely replace professional judgment in evaluating and integrating evidence.

In addition to the qualitative approaches, there are several quantitative approaches for integrating evidence and include: meta-analysis, probabilistic bias analysis, and the Bayesian approach. The quantitative approaches as the name implies will provide quantitative estimates of an effect size and can be used to estimate the magnitude of harm potentially caused by a chemical and the uncertainty of the estimate (NAS 2014). Quantitative estimates generated by using such methods can then be described in a narrative form that might be of use in communicating risk estimates. A summary of the available quantitative approaches is included here:

The quantitative approaches (meta-analysis, probabilistic bias analysis, and the Bayesian approach) are often recommended as part of the evidence integration analysis. The meta-analysis and type of meta-analysis approaches have been discussed in Section 7.11 of the RG-442 guidance and therefore discussion on meta-analysis will not be repeated in this document. USEPA recently finalized a white paper on probabilistic risk assessment methods and case studies and is available at http://epa.gov/raf/prawhitepaper/pdf/raf-pra-white-paper-final.pdf.

The TCEQ has used a meta-analysis approach in its derivation of the final URF for inorganic arsenic (Erraguntla et al.2012) and hexavalent chromium (Haney et al. 2014). The RG-442 also recommends a simpler but effective alternate meta-analysis type of approach to combine evidence from several cohorts using a weighting factor. The procedure recommends first determining the individual URFs from individual cohort's studies and then weighing the evidence from different cohorts using "inverse variance" prior to combining the evidence to derive a single quantitative estimate. This kind of evidence integration results in a single estimate based on all the data instead of relying only on one estimate (e.g., the most conservative) and effectively discarding all other data.

The main disadvantages of the quantitative approaches are that they are resource intensive and require specialized expertise. A brief summary of the probabilistic bias analysis and the Bayesian

approach as discussed in the NRC (2014) is provided here. The meta-analysis approach is limited as it requires combining statistical evidence from similar studies. However, in reality, one may need to integrate vastly different streams of data (i.e., animal, human, and mechanistic). The Bayesian approach can help integrate diverse streams of data. The probabilistic bias analysis is another type of quantitative approach for integrating evidence that involves producing intervals around the effect estimate that integrates uncertainty that is due to random and systematic sources. It is beyond the scope of this paper to provide a detailed discussion of the probabilistic bias analysis and the Bayesian approach. Staff are encouraged to review the NRC (2014) for a detailed review of these approaches.

Additionally, Rhomberg et al. (2013) provide an excellent discussion on this topic and provide a list of best practices for data integration that can serve as guidelines to TCEQ staff and are reproduced here as cited in their publication. These will be provided as guidance to staff to systematically integrate all the available evidence:

- Assess all animal and human data relevant to MOA, their human relevance (HR MOA), and dose-response;
- Evaluate what types of data have been considered (i.e., human, animal, *in vitro*, MOA, or adverse effects or biological perturbations that may be markers of the apical effect);
- Trace the reasoning by which these data bear on evaluation of the assessment question;
- Consider alternative modes of action and develop a biological story for each plausible MOA/endpoint combination – that is, articulate the reasoning behind how the endpoint illustrates an underlying causal process of potential concern;
- Consider relevance, response, and predictivity of the outcomes being treated as potential evidence, and use knowledge from other information (e.g., understanding of biological processes and pathways) and relationships on other chemicals to inform the relevance determinations, such as counterfactuals (observations that outcome changes when a hypothesized key causal element is removed);
- State professional judgments explicitly, noting the role of epistemic values and standards for declaration of sufficiency of evidence;
- Focus on exposures of concern for a chemical rather than solely on hazards that might be posed by that chemical at some sufficiently high, but unlikely real-world, exposure – that is, consider the range of exposures relevant to the problem formulation in drawing conclusions about possible toxicity-generating processes;
- Use diagrams to help articulate and communicate causal hypotheses;
- Evaluate and integrate negative and null results in addition to positive results;
- Integrate data across all lines of evidence so that interpretation of one will inform interpretation of the others. For example:
    Ask, if the proposed causative process and MOA were true, what other observable consequences should it have, and are these in fact seen;

Note assumptions, especially when they are ad hoc in that they are introduced to explain some phenomenon already seen;
Evaluate, compare, and contrast alternative explanations of the same sets of results.
Compelling hypotheses not only "are consistent" with particular pieces of data, but actually explain the array of results at hand much better than competing, contrary hypotheses;

- Present conclusions (in text, tables, and figures) not just as the result of judgments, but in the context of how they were derived and chosen over competitors, including sensitivity analysis of dependence of conclusions on specific data or assumptions;
- Recognize that applying specific study results to address a more general causation question is an exercise in generalization.
- Based on results of the WOE evaluation, identify data gaps and needs and propose next steps;
- Clearly present and communicate the WOE results and explore ways to measure and communicate different magnitudes of "weight" of evidence and different degrees of plausibility of explanations and their risk assessment consequences.

Step 6: Rate the Confidence in the Body of Evidence

In this step the confidence in the whole body of evidence is evaluated. The confidence in the body of evidence is determined by evaluating all the elements including type of data, study design, study quality, sample size, human relevance, and ROB that are discussed in detail in the previous steps. For example, good quality studies and lower ROB can translate to higher ratings that in turn indicate greater confidence and lower uncertainty that the key study findings accurately depict the true association of exposure and effect. Staff can rate the confidence in the body of evidence and use it as guidelines to assign uncertainty factors. The TCEQ Regulatory Guidance (RG-442) Section 7.13, Uncertainty Analysis, briefly describes the importance of recognizing and characterizing uncertainties. Higher confidence rating will result in lower uncertainty factors. Appropriately applying uncertainty factors is critical because the evidence integration approach requires some scientific judgment, use of assumptions and data extrapolations.

Conclusions

Systematic reviews and evidence integration are increasingly becoming recognized as important in chemical risk assessments (Rooney et al. 2014, NRC 2014, and Rhomberg et al. 2013). Each step of the systematic literature review and evidence integration plays an important role in improving confidence in the risk assessment process. For example, the PF step sets the stage to conduct a systematic review and compels staff to articulate the causal question. The systematic review requires staff to set clear inclusion and exclusion criteria that can greatly promote

transparency and limit subjective scientific judgment. Assessing data quality and ROB analysis will result in higher confidence and less uncertainty in the key studies. Weighing the evidence from different data streams prior to integrating the evidence will result in a final toxicity factor that is more predictive than if the most conservative estimate were simply chosen. The TCEQ will adopt this framework in its future chemical risk assessments. However, this document may be refined and updated as additional tools and resources become available.

References

Bevan, C., D. Strother. (2012). “Best Practices for Evaluating Method Validity, Data Quality and Study Reliability of Toxicity Studies for Chemical Hazard and Risk Assessments.” White Paper Developed for American Chemistry Councils Center for Advancing Risk Assessment Science and Policy. Accessed July 2014: http://arasp.americanchemistry.com/Data-Quality-Evaluation.

Critical Appraisal Skills Programme (CASP). 2013. CASP Checklists. Accessed Novmeber 2014: http://www.casp-uk.net/#!casp-tools-checklists/c18f8.

Erraguntla, NK, Sielken, RL, Valdez-Flores, C, Grant, RL. (2012). An updated inhalation unit risk factor for arsenic and inorganic arsenic compounds based on a combined analysis of epidemiology studies. Regulatory Toxicology and Pharmacology 64: 329–341.

Grimes, DA. (2002). "An overview of clinical research: the lay of the land." Lancet (London, England) 359(9300): 57-61.

Klimisch, H. J., M. Andreae, et al. (1997). "A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data." Regulatory Toxicology and Pharmacology 25(1): 1-5.

Haney, J.T., Erraguntla, N.K., Sielken, R,L., et al. (2014). Development of an Inhalation Unit Risk Factor for Hexavalent Chromium. Regul. Toxicol. Pharmacol. 68: 201-211.

Künzli, N. (1997). "The semi-individual study in air pollution epidemiology: A valid design as compared to ecologic studies." Environ Health Perspect 105(10): 1078-1083.

Lavelle, K. S., A. Robert Schnatter, et al. (2012). "Framework for integrating human and animal data in chemical risk assessment." Regulatory Toxicology and Pharmacology 62(2): 302-312.

Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford).;2011:baq036. doi: 10.1093/database/baq036. Print 2011. Review. PubMed PMID: 21245076; PubMed Central PMCID: PMC3025693.

Money, C. D., J. A. Tomenson, et al. (2013). "A systematic approach for evaluating and scoring human data." Regulatory Toxicology and Pharmacology 66(2): 241-247.

National Research Council. *Review of EPA's Integrated Risk Information System (IRIS) Process*. Washington, DC: The National Academies Press, 2014.

Phillips, C. and K. Goodman (2004). "The missed lessons of Sir Austin Bradford Hill." Epidemiologic Perspectives & Innovations 1(1): 1-5.

Rhomberg, L.R., J.E. Goodman, et al. (2013). “A survey of frameworks for best practices in weight-of-evidence analyses.” Crit Rev Toxicology 43(9): 753-84.

Rooney, A. A., A. L. Boyles, et al. (2014). "Systematic review and evidence integration for literature-based environmental health science assessments." Environ Health Perspect 122(7): 711-718.

Rushton L, Elliott P. 2003. Evaluating evidence on environmental health risks. Br Med Bull. 2003;68:113-28

Med Bull. 2003;68:113-28. Szklo, M. and F. J. Nieto (2007). Epidemiology : beyond the basics. Sudbury, Mass., Jones and Bartlett Publishers.

Woodruff T. J, P. Sutton (2011). Navigation Guide Work Group. An evidence-based medicine methodology to bridge the gap between clinical and environmental health science Health Aff (Millwood). 2011 May;30(5):931-7.

Woodruff T. J, P. Sutton (2014). The Navigation Guide Systematic Review Methodology: A Rigorous and Transparent Method for Translating Environmental Health Science into Better Health Outcomes. Environ Health Perspect. 2014 Jun 25.

Schünemann H, S. Hill S,et al. (2011) The GRADE approach and Bradford Hill's criteria for causation. J Epidemiol Community Health. (5):392-5.

TCEQ. 2008. Texas Commission on Environmental Quality. Development Support Document for Formaldehyde. Chief Engineer’s Office. http://www.tceq.texas.gov/assets/public/implementation/tox/dsd/final/formaldehyde_50-00-0_final.pdf

TCEQ. 2012. Texas Commission on Environmental Quality. Guidelines to develop effects screening levels, reference values, and unit risk factors. Chief Engineer’s Office.