Attachment D - Study Quality Guidance

Default study quality evaluation guidance is available in HAWC. The default study quality evaluation guidance was updated in HAWC with the following refinements for the vanadium systematic review. Reviewers only selected from "Good", "Adequate", "Deficient" or "Critically Deficient" judgements (i.e., "Not Reported" and "Not Applicable" were not used). Additionally, reviewers left "Direction of Bias" determinations as "not entered/unknown" for all study quality criteria.

Table C1. Refined EPA HAWC Study Evaluation Metrics and Guidance for Animal Studies

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
Chemical administration and characterization Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods?	For each study: Are there concerns [specific to this chemical] regarding the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? Was independent analytical verification of the test article (e.g., composition, homogeneity, and purity) performed? Were nominal exposure levels verified analytically? ○ For inhalation studies: were target concentrations confirmed using reliable analytical measurements in chamber air? For particles, was the particle size distribution measured using reliable analytical measurements in chamber air? ○ For oral studies: if necessary, based on consideration of chemical-specific knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed?	It is essential that these considerations are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (e.g., stability may be an issue for one chemical but not another). A judgment and rationale for this domain should be given for each relevant cohort or experiment in the study. • **Good**: Chemical administration and characterization are complete (i.e., source, purity, and analytical verification of the test article are provided). There are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. Exposure levels are verified using reliable analytical methods. For particles, the particle size distribution (preferably mass median aerodynamic diameter and geometric standard deviation) was determined using reliable analytical methods. • **Adequate**: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
	Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume, etc.)? *Notes:* • *Consideration of the appropriateness of the route of exposure is not evaluated at the individual study level.* • *Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.* • *Relatedly, consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and is not a risk of bias consideration.*	interpretation of the results (e.g., purity of the test article is sub-optimal but interpreted as unlikely to have a significant impact; analytical verification of exposure levels is not reported or verified with non-preferred methods; particle size distribution can be inferred from a study from the same laboratory using the same chemical and aerosol generation system). • **Deficient**: Uncertainties in the exposure characterization are identified and expected to substantially impact the results (e.g., source of the test article is not reported, and composition is not independently verified; impurities are substantial or concerning; administration methods are considered likely to introduce confounders, such as use of static inhalation chambers, or a gavage volume considered too large for the species and/or life stage at exposure). • **Critically Deficient**: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the results are largely attributable to factors other than exposure to the chemical of interest (e.g., identified impurities are expected to be a primary driver of the results).
<u>Allocation</u> Were animals assigned to experimental groups using a method that minimizes selection bias?	For each study: • Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)? • Is the allocation method described? Aside from randomization, were any steps taken to balance variables across experimental groups during allocation?	These considerations typically do not need to be refined by assessment teams. <u>A judgment and rationale for this domain should be given for each relevant cohort or</u> <u>experiment in the study.</u> • ***Good***: Experimental groups were randomized, and any specific randomization procedure was described or inferable (e.g., computer-generated scheme). [Note that normalization is not the same as randomization (see response for 'Adequate').] • ***Adequate***: Authors report that groups were randomized

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		but do not describe the specific procedure used (e.g., 'animals were randomized'). Alternatively, authors used a non-random method to control for important modifying factors across experimental groups (e.g., body weight normalization). • ***Not Reported*** **(interpreted as Deficient)**: No indication of randomization of groups or other methods (e.g., normalization) to control for important modifying factors across experimental groups. • ***Deficient*****:** Bias in the animal allocations was reported or inferable but is not expected to be severe. ***Critically Deficient***: Severe bias in the animal allocations was reported or inferable.
<u>Sensitivity</u> Are there concerns that sensitivity in the study is not adequate to detect an effect?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Was the exposure period, timing (e.g., life stage), frequency, and duration sensitive for the outcome(s) of interest? Based on knowledge of the health hazard of concern, did the selection of species, strain, and/or sex of the animal model reduce study sensitivity? Are there concerns regarding the timing (e.g., life stage) of the outcome evaluation? Are there aspects related to risk of bias domains that raise concerns about insensitivity (e.g., selection of protocols that are known to be insensitive or nonspecific for the outcome(s) of interest)? *Note: Consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and is not a study sensitivity consideration.*	These considerations may require customization to the specific exposure and outcomes. Some study design features that affect study sensitivity may have already been included in the other evaluation domains; these should be noted in this domain, along with any features that have not been addressed elsewhere. Some considerations include: • ***Good***: ○ The experimental design (considering exposure period, timing, frequency, and duration) is appropriate and sensitive for evaluating the outcome(s) of interest. ○ The selected animal model (considering species, strain, sex, or life stage) is known or assumed to be appropriate and sensitive for evaluating the outcome(s) of interest. ○ No significant concerns with the ability of the experimental design to detect the specific outcome(s) of interest. (e.g., outcomes evaluated at the appropriate life stage; study designed to address

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		known endpoint variability that is unrelated to treatment, such as estrous cyclicity or time of day). ○ Timing of endpoint measurement in relation to the chemical exposure is appropriate and sensitive (e.g., behavioral testing is not performed during a transient period of test chemical-induced depressant or irritant effects; endpoint testing does not occur only after a prolonged period, such as weeks or months, of non-exposure). ○ Potential sources of bias towards the null are not a substantial concern. • ***Adequate***: Same considerations as *Good*, except: ○ The duration and frequency of the exposure was appropriate, and the exposure covered most of the critical window (if known) for the outcome(s) of interest. ○ Potential issues are identified that could reduce sensitivity, but they are unlikely to impact the overall findings of the study. • **Deficient**: Concerns were raised about the considerations described for *Good* or *Adequate* that are expected to notably decrease the sensitivity of the study to detect a response in the exposed group(s). **Critically deficient**: Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely to be explained by bias. The rationale should indicate the specific concern(s).
<u>Observational bias/ blinding</u> Did the study implement measures to reduce observational bias?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Does the study report blinding or other procedures for reducing observational bias? If not, did the study use a design or approach for which such procedures can be inferred?	These considerations typically do not need to be refined by the assessment teams. [Note that it can be useful for teams to identify highly subjective measures of endpoints/outcomes where observational bias may strongly influence results prior to performing evaluations.] <u>A judgment and rationale for this</u>

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**

	What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?	domain should be given for each relevant cohort or experiment in the study.
		
		• ***Good***: Measures to reduce observational bias were described (e.g., blinding to conceal treatment groups during endpoint evaluation; consensus-based evaluations of histopathology lesions[1]).
		• ***Adequate***: Methods for reducing observational bias (e.g., blinding) were not explicitly stated but can be inferred.
		• ***Not Reported***: Measures to reduce observational bias were not described.
		  ○ ***(interpreted as Adequate)***: The potential concern for bias was mitigated based on use of automated/computer driven systems, standard laboratory kits, relatively simple, objective measures (e.g., body or tissue weight), or screening-level evaluations of histopathology.
		  ○ ***(interpreted as Deficient)***: The potential impact on the results is major (e.g., outcome measures are highly subjective).
		• ***Critically Deficient***: Strong evidence for observational bias that could have impacted results
		
		[1] For non-targeted or screening-level histopathology outcomes often used in guideline studies, blinding during the initial evaluation of tissues is generally not recommended as masked evaluation can make 'the task of separating treatment-related changes from normal variation more difficult' and 'there is concern that masked review during the initial evaluation may result in missing subtle lesions.' Generally, blinded evaluations are recommended for targeted secondary review of specific tissues or in instances when there is a pre-defined set of outcomes that is known or predicted to occur (Crissman, 2004).

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
<u>Confounding</u> Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups?	For each study: • Are there differences across the treatment groups, considering both differences related to the exposure (e.g., co-exposures, vehicle, diet, palatability) and other aspects of the study design or animal groups (e.g., animal source, husbandry, or health status), that could bias the results? If differences are identified, to what extent are they expected, based on a specific scientific understanding, to impact the results?	These considerations may need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical. <u>A judgment and rationale for this domain should be given for each cohort or experiment in the study, noting when the potential for confounding is restricted to specific endpoints/outcomes.</u> • ***Good***: Outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups. • ***Adequate***: Some concern that variables that were likely to confound or modify results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results. • ***Deficient***: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups and are expected to substantially impact the results. ***Critically deficient***: Confounding variables were presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
<u>Attrition</u> Did the study report results for all tested animals?	each study: Are all animals accounted for in the results? If there is attrition, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)? If unexplained attrition of animals for outcome assessment are identified, what is the expected impact on the interpretation of the results? *Note: This domain does **not** consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated under Results Presentation.*	These considerations typically do not need to be refined by assessment teams. <u>A judgment and rationale for this domain should be given for each cohort or</u> <u>experiment in the study.</u> ***Good***: Results were reported for all animals. If attrition is identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results. ***Adequate***: Results are reported for most animals. Attrition is not explained, but this is not expected to significantly impact the interpretation of the results. ***Deficient/Critically Deficient***: Moderate to high level of animal attrition that is not explained and may significantly impact the interpretation of the results
<u>Endpoint Measurement</u> Are the selected procedures, protocols, and animal models adequately described and appropriate for the endpoint(s)/ outcome(s) of interest?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Are the evaluation methods and the animal model adequately described and appropriate? Are there concerns regarding the specificity and validity of the protocols? Are there concerns about the specificity of experimental design? Are there serious concerns regarding the sample size or how endpoints were sampled? Are appropriate control groups for the study/assay type included? *Notes:* *Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under Sensitivity.*	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and typically must be refined by assessment teams. <u>A judgment and rationale for this domain should be given for each relevant cohort or</u> <u>experiment in the study.</u> Some considerations include the following: • ***Good***: o Adequate description of methods and animal models. o Use of generally accepted and reliable endpoint methods. o Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the endpoint protocol (e.g., sampling procedures for histological analysis). o Includes appropriate control groups and any use of nonconcurrent or historical control data (e.g., for evaluation of rare tumors) is justified (e.g., authors

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
	Considerations related to adjustments/corrections to endpoint measurements (e.g., organ weight corrected for body weight) are addressed under results presentation.	or evaluators considered the similarity between current experimental animals and laboratory conditions to historical controls). • ***Adequate***: Issues are identified that may affect endpoint measurement but are considered unlikely to substantially impact on the overall findings or the ability to reliably interpret those findings. • ***Deficient***: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings. • ***Critically Deficient***: Severe concerns are raised about endpoint measurement, and any findings are likely to be largely explained by these limitations. • The following specific examples of relevant concerns are typically associated with a **Deficient** rating but **Adequate** or **Critically Deficient** might be applied depending on the expected impact of limitations on the reliability and interpretation of the results: o Study report lacks important details that are necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols, information on the species, strain, sex, or life stage of the animals) o Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols.* o Overt toxicity (e.g., mortality, extreme weight loss) is observed or expected based on findings from similarly designed studies and may mask interpretation of outcome(s) of interest.

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		○ Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal)** ○ Control groups are not included, considered inappropriate, or comparisons to non-concurrent or historical controls are not adequately justified *These limitations typically also raise a concern for insensitivity ** Sample size alone is not a reason to conclude an individual study is critically deficient
Results Presentation Are the results presented and compared in a way that is appropriate and transparent?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Does the level of detail allow for an informed interpretation of the results? Are the data analyzed, compared, or presented in a way that is inappropriate or misleading?	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and typically must be refined by assessment teams. A judgment and rationale for this domain should be given for each relevant cohort or experiment in the study. Some considerations include the following: • ***Good***: ○ No concerns with how the data are presented. ○ Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). ○ No concerns with completeness of the results reporting.* • ***Adequate***: Concerns are identified that could affect results presentation but are considered unlikely to substantially impact on the overall findings or the ability to reliably interpret those findings. • ***Deficient***: Concerns with results presentation are

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		identified and expected to substantially impact results interpretation and reduce the reliability of the study findings. • ***Critically Deficient***: Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations or failure to report any results (qualitative or quantitative) for a prespecified outcome.* • The following specific examples of relevant concerns are typically associated with a **Deficient** rating but **Adequate** or **Critically Deficient** might be applied depending on the expected impact of limitations on the reliability and interpretation of the results: ○ Non-preferred presentation of data (e.g., developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate; presentation of only absolute organ weight data when relative weights are more appropriate). ○ Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages). ○ Incomplete presentation of the data* (e.g., presentation of mean without variance data; concurrent control data are not presented; dichotomizing or truncating continuous data; incidence/severity of histopathologic findings not included). *Failure to describe <u>*any*</u> findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) results in a critically deficient rating for the outcome(s) of interest for Results Presentation;

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		overall completeness of results reporting at the study level is addressed under Selective Reporting.
Selective Reporting Did the study report results for all prescribed outcomes?	For each study: Are results presented for all endpoints/outcomes described in the methods (see note)? If unexplained results omissions are identified, what is the expected impact on the interpretation of the results? *Note: This domain does not consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in Results Presentation.*	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each relevant cohort or experiment in the study. • ***Good***: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation time points. Data not reported in the primary article are available from supplemental material. If results omissions are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results. • ***Adequate***: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred) and evaluation time points. Omissions are not explained but are not expected to significantly impact the interpretation of the results. • ***Deficient***: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), omissions are not explained and may significantly impact the interpretation of the results. ***Critically Deficient***: Extensive results omission is identified and prevents comparisons of results across treatment groups.
Overall confidence (animal) Considering the identified strengths and limitations, what is the overall confidence	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified? If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study	The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias and sensitivity on the results. Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well-conducted and

Questions Used to Guide the Development of Criteria for Each Domain in Animal Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
rating for the endpoint(s)/ outcome(s) of interest?	results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects? *Note: Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well-conducted and an effect is observed, it may increase the strength of evidence judgement.*	an effect is observed, it may increase the strength of evidence judgement. A confidence rating and rationale should be given for each relevant cohort or experimental design in the study. ***High confidence***: No notable concerns are identified; the potential for bias is unlikely or minimal, and the study used sensitive methodology. *High* confidence studies generally reflect judgments of good across all or most evaluation domains. ***Medium confidence***: Possible deficiencies or concerns are identified, but the limitations are unlikely to have a significant impact on the study results or their interpretation. Generally, *medium* confidence studies include *adequate* or *good* judgments across most domains, with the impact of any identified limitation not being judged as severe. ***Low confidence***: Deficiencies or concerns are identified, and the potential for bias or inadequate sensitivity is expected to have a significant impact on the study results or their interpretation. Typically, *low* confidence studies have a *deficient* evaluation for one or more domains, although some *medium* confidence studies may have a *deficient* rating in domain(s) considered to have less influence on the magnitude or direction of effect estimates. ***Uninformative***: Serious flaw(s) are judged to make the study results uninterpretable for use in the assessment. Studies with Critically Deficient judgements in any evaluation domain are almost always rated Uninformative.

Table C2. Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
<u>Participant Selection</u> Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?	For longitudinal cohort: Did participants volunteer for the cohort based on knowledge of exposure and/or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome? For occupational cohort: Did entry into the cohort begin with the start of the exposure? Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status? Could exposure produce symptoms that would result in a change in work assignment/work status ("healthy worker survivor effect")? For case control study: Were controls represented of population and time periods from which cases were drawn? Are hospital controls selected from a group whose reason for admission is independent of exposure? Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure? For population-based survey: • Was recruitment based on advertisements to people with knowledge of exposure, outcome, and hypothesis? *Notes: Reviewers may have to seek out previous publications to get the full description of the cohort including recruitment details. Judgements for this domain can be modified slightly to include quantitative measures of loss to follow-up and response rates for volunteer populations because the lower these two are, the less likely the study will be impacted by selection of participants.*	These considerations may require customization to the outcome. This could include determining what study designs effectively allow analyses of associations appropriate to the outcome measures (e.g., design to capture incident vs. prevalent cases, design to capture early pregnancy loss). • **Good**: ○ Minimal concern for selection bias based on description of recruitment process and follow-up (e.g., selection of comparison population, population-based random sample selection, recruitment from sampling frame including current and previous employees). ○ Exclusion and inclusion criteria for participants specified and would not induce bias. ○ Participation rate is reported at all steps of study (e.g., initial enrollment, follow-up, selection into analysis sample). If rate is not high, there is appropriate rationale for why it is unlikely to be related to exposure (e.g., comparison between participants and nonparticipants or other available information indicates differential selection is not likely). • **Adequate**: ○ Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias. ○ Inclusion and exclusion criteria for participants specified and would not induce bias. ○ Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure.

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		• **Deficient**: Little information on recruitment processes, selection strategy, sampling framework or participation OR aspects of these processes raise the potential for bias (e.g., healthy worker effect, survivor bias). • **Critically Deficient**: Aspects of the processes for recruitment, selection strategy, sampling framework, or participation result in concern that selection bias is likely to have had a large impact on effect estimates (e.g., convenience sample with no information about recruitment and selection, cases and controls are recruited from different sources with different likelihood of exposure, recruitment materials stated outcome of interest and potential participants are aware of or are concerned about specific exposures).
<u>Exposure Measurement</u> Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome?	For all study types: Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure? Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably? Was the exposure measurement likely to be affected by a knowledge of the outcome? Was the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)? For case-control studies of occupational exposures: Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials? For biomarkers of exposure, general population: Is a standard assay used? What are the intra- and inter-assay coefficients of variation? Is the assay likely to be affected by contamination? Are values less than the limit of detection dealt with adequately?	These considerations require customization to the exposure and outcome (relevant timing of exposure). • **Good**: ○ Timing ▪ Measurement is appropriately captured in consideration of temporality (exposure occurs prior to outcome measurement). ▪ Measurement represents the etiologically relevant time period of interest. ○ Type/Analysis ▪ Valid exposure assessment methods used ▪ Exposure misclassification is expected to be minimal. ▪ Measurement quantitatively captures information on frequency/duration/dose to support quantitative exposure-response ▪ Repeated measures are captured to support evaluation of measurement variability.

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**

	What exposure time-period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure?
	
	Notes: In order to fully and accurately evaluate the role of bias in the measurement of exposure, consider splitting up this domain into three sub-parts to be able to evaluate each of the components separately – measurement type, timing of measure, and analytics.
		• **Adequate**:
		  ○ Timing
		  ▪ Measurement is appropriately captured in consideration of temporality (exposure occurs prior to outcome measurement).
		  ▪ Measurement represents the etiologically relevant time period of interest.
		  ○ Type/Analysis
		  ▪ Valid exposure assessment methods used
		  ▪ Exposure misclassification might exist but is not expected to greatly change the effect estimate.
		  ▪ Measurement quantitatively captures information on exposure, but information on frequency/duration/individual dose may be limited.
		• **Deficient**:
		  ○ Timing
		  ▪ Measurement is collected at the same time as exposure may not represent etiologically relevant period.
		  ▪ Specific knowledge about exposure and outcome raises concerns about reverse causality, but there is uncertainty whether it is influencing the effect estimate.
		  ○ Type/Analysis
		  • Exposed groups are expected to contain a notable proportion of unexposed or minimally exposed individuals.
		  • The method did not capture important temporal or spatial variation, or there is other evidence of exposure misclassification that would be expected to notably change the

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		effect estimate. • Categorical exposure estimation or there is no information on frequency or duration of exposure. • **Critically Deficient**: o Timing ▪ Exposure measurement does not characterize the etiologically relevant time period of exposure or is not valid. Temporality is a major concern. ▪ There is evidence that reverse causality is very likely to account for the observed association. o Type/Analysis ▪ Exposure measurement was not independent of outcome status. ▪ Categorical exposure, limited to capturing ever/never exposure only.
<u>Outcome ascertainment</u> Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?	For case-control studies: • Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the disease? For mortality measures: • How well does cause of death data reflect occurrence of the disease in an individual? How well do mortality data reflect incidence of the disease? For diagnosis of disease measures: • Is diagnosis based on standard clinical criteria? If based on self-report of diagnosis, what is the validity of this	These considerations require customization to the outcome. • **Good**: o High certainty in the outcome definition (i.e., specificity and sensitivity), minimal concerns with respect to misclassification. o Assessment instrument was validated in a population comparable to the one from which the study group was selected. • **Adequate**: o Moderate confidence that outcome definition was specific and sensitive, some uncertainty with respect to misclassification but not expected to greatly change the effect estimate.

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
	measure? For laboratory-based measures (e.g., hormone levels): • Is a standard assay used? Does the assay have an acceptable level of inter-assay variability? Is the sensitivity of the assay appropriate for the outcome measure in this study population? *Notes: Similar to the Exposure Measurement domain, assessing bias involves evaluation of the type of measurement as well as timing. Consider separating this domain into sub-parts if these concepts are important to the outcome of interest.*	○ Assessment instrument was validated but not necessarily in a population comparable to the study group. • **Deficient**: ○ Outcome definition was not specific or sensitive. ○ Uncertainty regarding validity of assessment instrument. • **Critically Deficient**: ○ Invalid/insensitive marker of outcome (e.g., lack of adjustment of pulmonary function testing for age/sex). ○ Outcome ascertainment is very likely to be affected by knowledge of, or presence of, exposure. *Note: Lack of blinding should not be automatically construed to be critically deficient.*
<u>Confounding</u> Is confounding of the effect of the exposure likely?	Is confounding adequately addressed by considerations in… a. … participant selection (matching or restriction)? b. … accurate information on potential confounders, and statistical adjustment procedures? c. … lack of association between confounder and outcome, or confounder and exposure in the study? d. … information from other sources? Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)? *Notes: Confounding variables are those that influence both the exposure and outcome. It is important to do targeted literature*	These considerations require customization to the exposure and outcome, but this could be limited to identifying key covariates. Key confounders for studies of general populations include PM2.5 co-exposures and smoking status. Key confounders for occupational studies include exposure to "dust" and smoking status. Key covariates include co-exposures to other respiratory irritants or carcinogens. Other endpoint-specific covariates (confounders or effect modifiers) may be identified. • **Good**: ○ Conveys strategy for identifying key confounders, including co-exposures. This could include a priori biological considerations, published literature, causal diagrams, or statistical analyses, with recognition that not all "risk factors" are confounders. ○ Inclusion of potential confounders in statistical

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
	searches to determine what confounding factors are relevant. Identify key confounders, including co-exposures that must be considered. Consider modifying the judgement criteria for this domain to reflect the authors' process for identifying and considering important confounders for the vaccine of interest in their study.	models not based solely on statistical significance criteria (e.g., $p < 0.05$ from stepwise regression). ○ Does not include variables in the models likely to be influential colliders or intermediates on the causal pathway. ○ Key confounders are evaluated appropriately and considered unlikely sources of substantial confounding. This often will include: ▪ Presenting the distribution of potential confounders by levels of the exposure of interest or the outcomes of interest (with amount of missing data noted); ▪ Consideration that potential confounders were rare among the study population, or were expected to be poorly correlated with exposure of interest; ▪ Consideration of the most relevant functional forms of potential confounders; ▪ Examination of the potential impact of measurement error or missing data on confounder adjustment; or ▪ Presenting a progression of model results with adjustments for different potential confounders, if warranted. ○ Assessment instrument was validated in a population comparable to the one from which the study group was selected. • **Adequate**: Similar to good but might not include all key confounders, or less detail might be available on the evaluation of confounders (e.g., sub-bullets in good). It is possible that residual confounding could explain part of the observed effect, but concern is minimal. • **Deficient**:

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		○ Does not include variables in the models likely to be influential colliders or intermediates on the causal pathway. -And any of the following- ○ The potential for bias to explain some of the results is high based on an inability to rule out residual confounding, such as a lack of demonstration that key confounders of the exposure-outcome relationships were considered; ○ Descriptive information on key confounders (e.g., their relationship relative to the outcomes and exposure levels) are not presented; or ○ Strategy of evaluating confounding is unclear or is not recommended (e.g., only based on statistical significance criteria or stepwise regression [forward or backward elimination]). • **Critically Deficient**: ○ Includes variables in the models that are colliders or intermediates in the causal pathway, indicating that substantial bias is likely from this adjustment; or ○ Confounding is likely present and not accounted for, indicating that all results were most likely due to bias.
<u>Analysis</u> Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?	• Are missing outcome, exposure, and covariate data recognized, and if necessary, accounted for in the analysis? • Does the analysis appropriately consider variable distributions and modeling assumptions? • Does the analysis appropriately consider subgroups of interest (e.g., based on variability in exposure level or duration, or susceptibility)? • Is an appropriate analysis used for the study design? • Is effect modification considered, based on	These considerations may require customization to the outcome. This could include the optimal characterization of the outcome variable and ideal statistical test (e.g., Cox regression). • **Good**: ○ Use of an optimal characterization of the outcome variable. ○ Quantitative results presented (effect estimates and confidence limits or variability in estimates; i.e., not presented only as a p value or “significant”/”not

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
	considerations developed a priori? • Does the study include additional analyses addressing potential biases or limitations (i.e., sensitivity analyses)?	significant"). ○ Descriptive information about outcome and exposure provided (where applicable). ○ Amount of missing data noted and addressed appropriately (discussion of selection issues—missing at random vs. differential). ○ Where applicable, for exposure, includes limit of detection (LOD, and percentage below the LOD), and decision to use log transformation. ○ Includes analysis that address robustness of findings, e.g., examination of exposure-response (explicit consideration of nonlinear possibilities, quadratic, spline, or threshold/ceiling effects included, when feasible); relevant sensitivity analyses; effect modification examined only on the basis of a priori rationale with sufficient numbers. ○ No deficiencies in analysis are evident. Discussion of some details might be absent (e.g., examination of outliers). • **Adequate**: Same as good, except: ○ Descriptive information about exposure provided (where applicable) but could be incomplete; might not have discussed missing data, cut points, or shape of distribution. ○ Includes analyses that address robustness of findings (examples in good), but some important analyses are not performed. • **Deficient**: ○ Does not conduct analysis using optimal characterization of the outcome variable. ○ Descriptive information about exposure levels not provided (where applicable). ○ Effect estimates and p-value presented without

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		standard error or confidence interval. ○ Results presented as statistically "significant"/"not significant." • **Critically Deficient**: ○ Analysis methods are not appropriate for design or data of the study.
<u>Overall Confidence (human)</u> Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/ outcome(s) of interest?	*[Note: IRIS/HAWC does not provide prompting questions]* *Notes:* • *The overall confidence in the study considered all domains equally. However, some domains for epidemiology studies are more critical in assessing bias than others.* • *Similar to the OHAT study quality evaluation guidance, there are three "key" elements: outcome ascertainment, exposure measurement, and confounding.* • *To obtain an overall confidence rating of "High", all key domains are rated as good or adequate, with a majority of the other domains also rated as "Good" or "Adequate".* • *For an overall confidence rating of "Low", all key domains are rated "Deficient", with a majority of the other domains also rated as "Deficient".* • *Studies with an overall confidence rating of "Medium" did not meet the criteria for "High" or "Low" confidence ratings.* • *Any ratings of "Critically Deficient" would result in an overall confidence of "Uninformative".*	The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias and sensitivity on the results. Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well-conducted and an effect is observed, it may increase the strength of evidence judgement. <u>A confidence rating and rationale should be given for each relevant cohort or experiment in the study.</u> ***High confidence***: No notable concerns are identified; the potential for bias is unlikely or minimal, and the study used sensitive methodology. *High* confidence studies generally reflect judgments of good across all or most evaluation domains. ***Medium confidence***: Possible deficiencies or concerns are identified, but the limitations are unlikely to have a significant impact on the study results or their interpretation. Generally, *medium* confidence studies include *adequate* or *good* judgments across most domains, with the impact of any identified limitation not being judged as severe. ***Low confidence***: Deficiencies or concerns are identified, and the potential for bias or inadequate sensitivity is expected to have a significant impact on the study results or their interpretation. Typically, *low* confidence studies have a *deficient* evaluation for one or more domains,

Questions Used to Guide the Development of Criteria for Each Domain in Epidemiology Studies		
Core Question	**Prompting Questions**	**Basic Considerations**
		although some *medium* confidence studies may have a *deficient* rating in domain(s) considered to have less influence on the magnitude or direction of effect estimates. ***Uninformative***: Serious flaw(s) are judged to make the study results uninterpretable for use in the assessment. Studies with Critically Deficient judgements in any evaluation domain are almost always rated Uninformative.

Additional considerations for TCEQ reviewers regarding study quality judgements for epidemiology studies:

Based on comments in the public domain for previous IRIS drafts, EPA has not appropriately considered bias by correlated confounding co-exposures in epidemiology study quality, which should be a critical consideration that downgrades such mixture studies. ***TCEQ should ensure this important issue is appropriately considered for study quality moving forward in the TCEQ's assessments. Consider the following scenario:***

> *When conducting a dose-response assessment for a single chemical, it is difficult to envision that a regulatory agency would derive a toxicity factor for a single chemical based on a mixture animal study in which the animals had significant exposure to numerous related chemicals (e.g., chemically similar with correlated exposures), both quantified and unquantified confounding exposures. Such a mixture animal study would be excluded, and rightly so (e.g., such studies would be deficient/critically deficient per p. 4-33 of the IRIS Handbook; USEPA 2022), as the resulting toxicity factor for a single component of the mixture would be unreliable. The IRIS handbook (USEPA 2022) cites concerns about confounding as follows:*

"Co-exposures should also be considered as potential confounders. Some exposures tend to be found together in the environment or in occupational settings and are highly correlated. For example, it might be difficult to distinguish the independent effects from exposure to specific phthalate or per- and polyfluoroalkyl substances in drinking water, isomers of polychlorinated biphenyls in fish, or volatile organic compounds generated by a common source (e.g., benzene, toluene, ethylbenzene, xylene in traffic emissions) due to confounding by these co-exposures."

Examples to consider specific to the vanadium assessment include independent effects from exposure to $PM_{2.5}$ and its components, or general dustiness in occupational settings. Epidemiology mixture studies with significant and correlated co-exposures to chemicals with the same or similar MOAs (known or expected) and endpoints should be considered deficient under EPA or similar guidelines because the potential for bias to explain some of the results is high based on an inability to rule out residual confounding by key confounders of the exposure-outcome relationship (e.g., see p. 4-21 of the IRIS Handbook; USEPA 2022). This is particularly true when there is a lack of multi-pollutant modeling that can accurately isolate the magnitude of the contribution of the single component of interest of the mixture to the observed adverse effect and/or when there are known (or reasonably expected) but unquantified correlated co-exposures to similarly acting chemicals, and this should be reflected in how study quality is evaluated (e.g., as poor due to confounding) for such studies. Correlated co-exposures, both quantified and unquantified (e.g., PFAS both measured and unmeasured in serum), to multiple other chemicals with the same or similar MOAs and endpoints should be considered among study "attributes that would be likely to have a large effect, compared to a small effect, on confidence in the study results" (Section 4.2.1 of the IRIS Handbook; USEPA 2022). Such correlated co-exposures can positively bias study results and produce mixture effects that in the absence of an accurate multi-pollutant model and data on all such co-exposures cannot be adjusted for in a scientifically defensible manner. ***Consequently, epidemiology studies that are inherently mixture studies with significant and correlated co-exposures to chemicals with the same or similar MOAs (known or expected) and endpoints should not be used as the basis for quantitative dose-response assessment and/or toxicity factor derivation.*** *This would be consistent with the IRIS Handbook (USEPA*

2022)[a] and more consistent with fairly recent conclusions by the Australian government (FSANZ 2021)[b] and the U.S. Agency for Toxic Substances and Disease Registry (ATSDR 2021).

[a] USEPA (U.S. Environmental Protection Agency). (2022). ORD Staff Handbook for Developing IRIS Assessments. EPA/600/R-22/268. December 2022.

[b] FSANZ (Food Standards Australia and New Zealand). 2021. PFAS and Immunomodulation Review and Update. Available at: https://link.edgepilot.com/s/bfb8fffa/Y1mzwXsvYk20CpTbcyey3Q?u=http://www.foodstandards.gov.au/publications/Documents/PFAS%2520and%2520Immunomodulatory%2520Review%2520and%2520Update%25202021.pdf