United States
Consumer Product Safety Commission

July 22, 2025

CPSC Staff Statement on: Guidance Document for Conducting Class-Based Dose-Response

The U.S. Consumer Product Safety Commission (CPSC or Commission) contracted with ICF (Contract BPA No. 61320622A0005, Order No. 61320623F2030) to identify approaches to class-based dose-response assessment and complete a guide that can be used to conduct class-based dose-response assessments. This statement was prepared by the CPSC staff. ICF produced for CPSC staff the accompanying report, "Guidance Document for Conducting Class-Based Dose-Response" (the Dose-Response Guide), also known as Subtask 4A in the contract. The statement and report have not been reviewed or approved by, and may not represent the views of, the Commission.

The purpose of the Dose-Response Guide is to outline technical guidance for class-based dose-response analyses to develop quantitative toxicity reference value (TRV)[1] estimates for various subclasses of chemicals. The guidance is based on lessons learned from the collation of methods (Subtask 2A) and application of these methods to the polyhalogenated organophosphates (PHOPs) subclass of the organohalogen flame retardants (Subtask 2B).[2]

The Dose-Response Guide outlines a tiered approach to deriving class-based TRVs for chemical subclasses based upon different levels of data availability and quality. The proposed tiered approach aims to use the best data available for each chemical, while considering consistency between chemicals in a subclass. It also leverages the existence of multiple types of data to evaluate overall confidence in the data and provides opportunities to incorporate new data to increase confidence for relatively data-poor chemicals. The aim of this guidance document is to highlight key decision points and criteria for making those decisions when developing quantitative TRV estimates.

Overall, the Dose-Response Guide provides a process for deriving TRVs with key considerations for the application to class-based assessments. The analyses in the related case study for the PHOPs subclass in Subtask 2B demonstrate the strength of this approach and provide an example of how it might be applied to other subclasses.[2]

[1] "Toxicity reference value" or TRV quantifies the potency of a substance or estimates the dose level to which people can be exposed for a specified duration of time without experiencing an adverse health effect.

[2] The Dose-Response Guide cites supporting documents such as Subtasks 1C, 2A, and 2B from this contract call order as well as deliverables from an earlier call order (Contract BPA No. 61320622A0005, Order No. 61320622F2011) all of which can be found at https://www.cpsc.gov/Business--Manufacturing/Organohalogen-Flame-Retardant-Chemicals-Assessment.

Guidance Document for Conducting Class-based Dose-response

Call Order 61320623F2030
CPSC BPA No. 61320622A0005
January 31, 2025

Submitted To:
U.S. Consumer Product Safety Commission
5 Research Place
Rockville, MD 20850

Submitted By:
ICF
1902 Reston Metro Plaza
Reston, VA 20190 USA

Table of Contents

Tables

Figures

1. Introduction

Under CO-9, Subtask 4A seeks to outline guidance to apply class-based dose-response analyses to develop quantitative toxicity reference value (TRV) estimates for various subclasses of chemicals based on lessons learned from the collation of methods and application of these methods to the polyhalogenated organophosphates (PHOPs) subclass of the organohalogen flame retardants (OFRs).

This report outlines a tiered approach to deriving class-based TRVs for chemical subclasses. The tiered approach describes recommended approaches to conduct class-based dose-response analyses based upon different levels of data availability and quality. The aim of this report is to highlight key decision points and criteria for making those decisions when developing quantitative TRV estimates.

Importantly, it should be noted that the detailed decision-making approach and results depend on the specifics of the input data, and one case study is insufficient for considering all possibilities and permutations of available data. Therefore, much of the guidance in this report details issues to consider in weighing the evidence, rather than providing a prescriptive approach to all decisions. In particular, suggestions are made for how comparisons between points of departure (PODs) developed by different methods relate to preferred approaches and confidence levels. However, specifics of chemical classes and additional case studies may result in modified recommendations.

2. Points of Departure (PODs) and Toxicity Reference Values (TRVs)

2.1. PODs for Single Chemicals

2.1.1. Noncancer Endpoints

This guide provides a brief overview of the methods for deriving TRVs for context. Of necessity, this summary does not address all of the details in the methods. Additional details are provided in Pecquet and Haber (2024) and U.S. Environmental Protection Agency (U.S. EPA) (2023). In traditional single chemical noncancer dose-response assessments, the first step is hazard characterization, in which the effects of the chemical are identified, adversity is evaluated, the conditions resulting in the effect are considered (e.g., route specificity), and the human relevance of the effect is considered. Although this guide focuses on dose-response assessment, proper conduct of the hazard characterization step is critical to ensuring that the correct endpoints are chosen for dose-response evaluation.

Evaluation of human relevance involves two steps. First, the mode of action (MOA) in rodents is evaluated and then the human relevance is evaluated based on the fundamental qualitative or quantitative differences that may exist between the key events in rodents and humans (Boobis et al., 2008). The current state of the science for evaluation of MOA is based on the evolved Bradford Hill considerations (Meek et al., 2014). These considerations address the following questions:

1. Biological concordance — Does the hypothesized MOA conflict with the broader biological knowledge? How well established is the MOA in the wider biological database?

2. Essentiality of key events — Is the sequence of events reversible if dosing is stopped or a key event prevented?

3. Concordance of empirical observations among key events (including dose-response, temporality, and incidence) — Are the key events observed at doses below or similar to those associated with the apical effect? Did the key events occur in the hypothesized order? Is the incidence of the apical effect less than that for preceding key events?

4. Consistency — Is the pattern of effects across species/strains/organs/test systems consistent with expectations based on the respective biology?

5. Analogy — Would the MOA be anticipated based on broader chemical specific knowledge?

Once the relevant studies and endpoints are identified based on MOA and human relevance, potential PODs are identified within each study. For any adverse endpoint within a study, the potential POD could be one of the following:

- Lowest adverse effect level (LOAEL) — The lowest dose with a degree of response deemed adverse for the endpoint of interest.
- No observed adverse effect level (NOAEL) — The dose immediately below the LOAEL for an endpoint (e.g., the highest dose without an adverse effect for that endpoint). Non-adverse effects may occur at the NOAEL.
- Benchmark dose (BMD) — The BMD is determined by fitting a flexible mathematical curve to the dose-response data for a single endpoint and determining the dose corresponding to a specified degree of response, the benchmark response (BMR). BMDs are not used directly for calculating TRVs, but they are used for calculating relative potency.
- Lower 95% confidence limit on the BMD (BMDL). BMDLs are used in place of NOAELs in calculated TRVs.

The overall POD for a given study is the lowest of the NOAELs (or LOAELs/10) and BMDLs calculated for relevant endpoints. The PODs for TRV derivation are then arrayed and the lowest relevant POD is identified. The effect associated with the lowest POD is the critical effect (i.e., the first adverse effect that occurs in the most sensitive species at the lowest dose of the substance (U.S. EPA, 2024)).

2.1.2. Cancer Endpoints

Cancer risk assessment focuses on conducting the hazard characterization and evaluating the dose-response for tumors. As for the noncancer assessment, the evolved Bradford Hill considerations (Section 2.1.1) provide a useful framework for evaluating MOA. Determination of the human relevance of the tumor MOA observed in rodents is also important (Boobis et al., 2008; U.S. Environmental Protection Agency (U.S. EPA), 2005). The key difference between carcinogenic chemicals and non-carcinogenic chemicals is that, for classical assessments, MOA plays a larger role in determining the approach for low-dose extrapolation for cancer, although MOA is important for both types of endpoints. Typically, dose-response data are modeled from all tumor types with a weight-of-evidence of “probable” or “likely” human carcinogen or higher. The same approach is used for modeling of the tumor data to determine the POD, regardless of the MOA, but the approach for extrapolation to low doses depends on the MOA. A cancer slope factor (CSF) is calculated if the chemical is known to cause tumors via DNA

interaction or if the MOA is not known (Section 2.2.2), while uncertainty factors (UF) are applied if the MOA is shown to be something other than interacting with DNA (Section 2.2.1).

2.2. Deriving TRVs from PODs

Dose-response assessment for single chemicals is conducted using different methods for threshold endpoints (i.e., non-cancer effects and cancer effects thought to be induced by a non-mutagenic MOA) and non-threshold effects (i.e., cancer effects thought to be induced by a mutagenic MOA). Threshold endpoints use an UF approach as it is assumed that there are doses below which no adverse health outcome occurs, while non-threshold endpoints use a linear extrapolation approach as no dose is assumed to be completely safe.

2.2.1. Noncancer Endpoints

For chemicals causing threshold effects, a TRV such as a reference dose (RfD) or Acceptable Daily Intake (ADI) is calculated by dividing a POD by the product of UFs representing interspecies differences (UFA), intraspecies variability (UFH), and use of a LOAEL instead of a NOAEL or BMDL (UFL). Some organizations, such as the US EPA, also include factors for database uncertainty (UFD), and extrapolation from subchronic to chronic duration (UFS). Different organizations may use different sets of UFs. While CPSC generally uses UFA, UFH, and UFL for traditional single-chemical assessments, we are also considering UFS for use in a class-based risk assessment.

$$TRV = \frac{POD}{UFA \times UFH \times UFL \times UFS}$$

2.2.2. Cancer Endpoints

For chemicals causing effects via a MOA with a threshold (including chemicals causing cancer via a non-mutagenic MOA), the above approach in Section 2.2.1 is applied. For chemicals causing cancer via a MOA without a threshold, a CSF can be calculated:

$$CSF = \frac{Response}{POD}$$

For example, when tumor data are modeled, the POD is typically the lower 95% confidence limit on the benchmark dose corresponding to a 10% extra risk (BMDL10) and the corresponding response is 0.1 (for the 10% extra risk at the BMD10). Additional details on cancer quantification and derivation of the CSF, including bodyweight scaling to extrapolate from animal to humans, is provided in the Guidelines for Carcinogen Risk Assessment (U.S. EPA, 2005).

2.3. Class-based PODs

Class-based dose-response analysis is a process for evaluating a group of structurally and biologically-related chemicals (a "class") by extrapolating key elements of the dose-response analysis from data-rich members of a chemical class to data-poor class members.

The goal of our work has been to use existing TRVs and additional hazard data to derive quantitative TRV values for PHOP chemicals that lack published TRVs. Broadly, the identified steps for a class-based dose-response analysis are:

1. Determination of chemical class (chemical grouping)
2. Identification of shared endpoints and toxicity data for chemical ranking and read-across, including new approaches/alternative methods (NAMs) data when available
3. Selection of an anchor or surrogate chemical(s)
4. Conducting dose-response evaluation through a variety of methods
5. Identification of appropriate UFs
6. Calculation of TRVs or CSFs

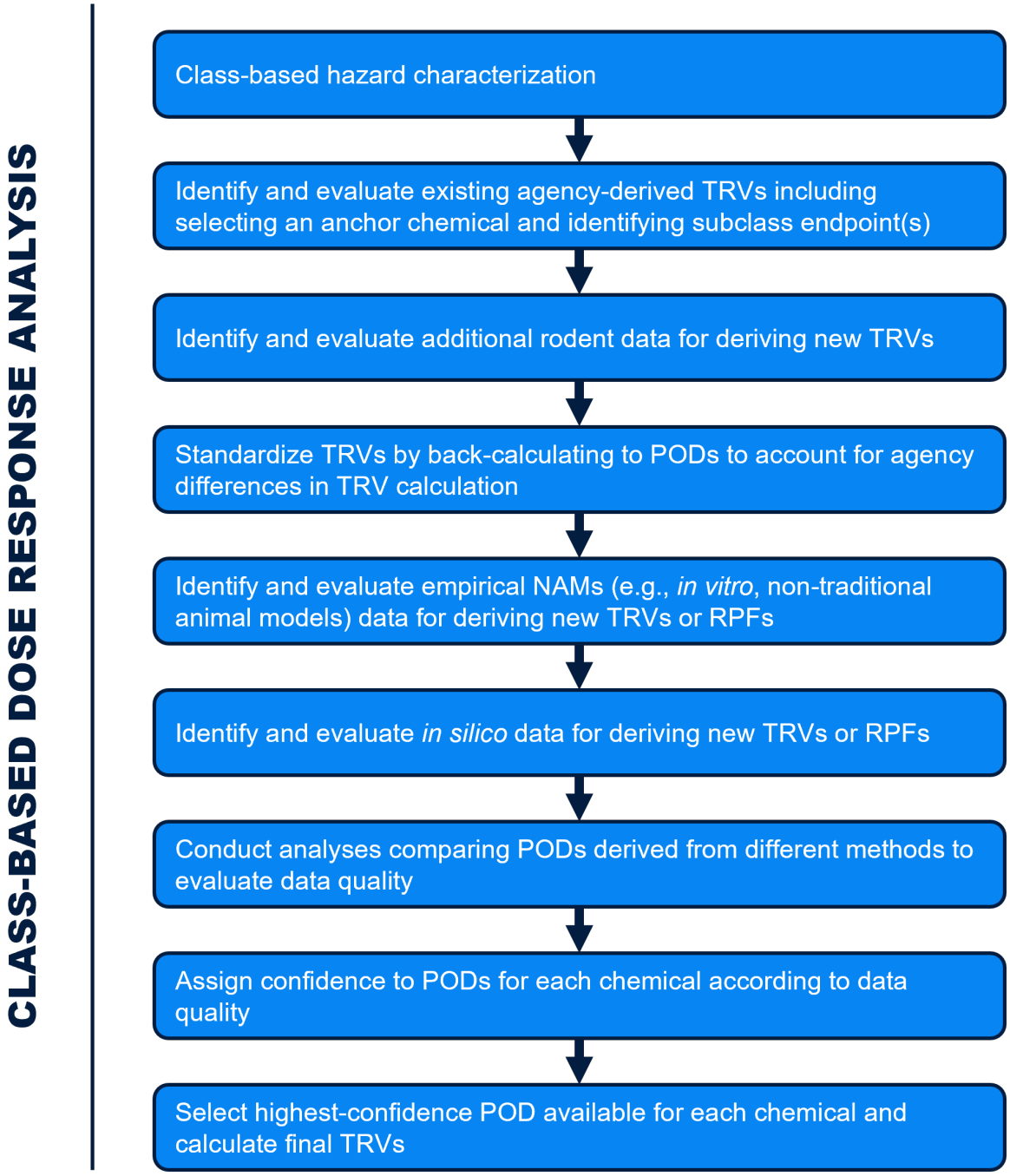
Some later steps in a class-based dose-response analysis may inform earlier steps, and so these steps may be better thought of as iterative rather than entirely sequential. For example, the choice of anchor/surrogate chemical may depend on the availability of data that serve as a basis for relative potency factors (RPFs), the endpoints/targets that have available data for other chemicals in the subclass, and/or the degree to which other chemicals in the subclass share endpoints/targets. Thus, an initial candidate(s) for anchor/surrogate may be selected early on but this choice should be continually evaluated in the context of later steps.

An important consideration in class-based dose-response assessment is a shared endpoint or target organ/system among class members. This is because one assumption of class-based assessments is that similar chemical structures (i.e., the first characteristic considered in defining a chemical class) result in similar hazards. At a minimum, to accurately extrapolate hazard and dose-response between class members, there must be commonality in adverse outcome pathway (AOP), mechanism, or MOA leading to health effects (particularly if the effects are in different tissues), or other commonality in health outcome or the manner in which it is elicited. This should be

considered during the first step, determination of chemical class, and again during the second step where shared endpoints and toxicity data are identified and ranked.

The remainder of this report details each of these steps and notes important lessons learned during their application. Figure 1 outlines the basic steps of a class-based dose-response analysis.

Figure 1. Basic Steps of Class-based Dose-response Analysis



TRVs = toxicity reference values; PODs = points of departure; NAMs = new approaches/alternative methods; RPFs = relative potency factors

3. Chemical Grouping

Chemical grouping involves identifying a universe of chemicals, evaluating structural fingerprints, and then building and evaluating the similarity of profiles related to the chemicals' potential interactions with relevant biology. We recommend five profiles: metabolite, physiologically based toxicokinetic (PBTK), physicochemical properties, mechanistic effects, and adverse effects. U.S. EPAs Generalized Read-Across (GenRA) tool can help define a class of chemicals by identifying surrogate chemicals and determining which of a group of chemicals are structurally related based upon NAMs data collated in the CompTox Dashboard. We recommend the following steps be taken during chemical grouping:

1. Define the universe of chemicals of interest. This may be based on aspects such as common uses, common occurrences, or other commonalities
2. Structural analysis to identify a class of similar chemicals
3. Analysis of metabolites of individual class members and identification of shared metabolites and parent-metabolite relationships within the class
4. Toxicokinetic properties and implications for internal dose (PBTK)
5. Physicochemical properties and implications for internal dose, bioavailability, clearance, and/or hazard
6. Biological effects, including data from both mechanistic (NAMs) and *in vivo* data sources on adverse effects

In the PHOPs case study, determination and evaluation of chemical grouping was a multi-tiered process. The OFRs were initially grouped into 14 subclasses based on physicochemical properties by National Academies of Sciences, Engineering, and Medicine (NASEM, 2019). The lists of OFRs included in these subclasses were modified and enhanced by CPSC staff (Bevington et al., 2022) based on market use and input from expert chemists. The modified and enhanced list for PHOPs was further evaluated in a class-based hazard assessment step using a multi-characteristic approach to grouping chemicals, developed in CO-1 (Call Order 61320622F2011 under CPSC Blanket Purchase Agreement (BPA 61320622A0005)).

4. Data Curation

4.1. Data Availability Sources

There are numerous sources of available data for hazard, chemical ranking, and PODs. The availability of data of different types will inform the methods available that can be used to derive PODs.

Generally, PODs used to derive TRVs in agency assessments will be found primarily via agency websites, although a POD used to derive a TRV in a high-quality journal publication may also be considered. Additional animal data from which to derive new PODs may be found through a literature search, and through databases such as U.S. EPA's ToxValDB, National Toxicology Program's (NTP) Integrated Chemical Environment (ICE), and Health Canada's automated workflow for prioritization (HAWPr). These databases may include PODs based on additional unpublished data that were not included in peer reviewed literature, or additional *in vivo* data useful for a class-based assessment.

Data sources such as ToxValDB, ICE, HAWPr, and peer reviewed literature may also be a source of *in vitro* PODs or other *in vitro* data useful for a class-based assessment. Literature searches may also be a source of NAMs data for derivation of PODs, though it may be simpler and more efficient to start with public repositories for NAMs data such as Gene Expression Omnibus (GEO), ICE, HAWPr, or U.S. EPA's CompTox Dashboard. Additional sources NAMs data are likely to become available as the field continues to mature, although coverage of less-studied class members may lag behind the availability of data sources.

The Danish QSAR Database is a good source of *in silico*–predicted LD_{50}s, though with the limitation that LD_{50}s are unlikely to be clearly related to critical effects associated with PODs. Aside from LD_{50}s, the Danish QSAR Database provides mostly qualitative predictions, and thus quantitative data for other endpoints are unlikely to be available from these sources. The OECD QSAR Toolbox or other QSAR tools may be alternative sources of *in silico* data. Additionally, the Chiu Dashboard (Aurisano et al., 2023; Kvasnicka et al., 2024), which is based upon collated animal data, contains PODs derived *in silico* for 804,572 chemicals (as of the writing of this report).

Of note, there is a unique opportunity presented when the same chemicals have data available from multiple data streams to compare these data to inform confidence (Section 6). That is, existing data can be used to refine relative confidence in the different data streams in a manner specific to the subclass being analyzed. When data from different data streams generally align, confidence in both data streams (though

particularly the lower quality/confidence data) may be increased. Conversely, when data do not generally align, confidence may be decreased, as discussed further in Section 4.2.

4.2. Data Quality Checks

When data are available from multiple sources, it is possible to assess them to determine how different data streams compare. Generally, during correlation analysis the directionality of the correlation should be confirmed to match between evidence streams (e.g., when one data type predicts higher PODs, so does another; inconsistent predictions indicate a problem with one or more of the data streams). Other methods to assess similarity in data types include interclass correlation coefficients (ICCs) that can be used to assess where groups are more (larger ICC up to ICC=1) or less (smaller ICC down to ICC=0) similar, and principal component analyses (PCA) that can be used assess sources of variation without *a priori* assumptions (Jolliffe & Cadima, 2016; Koo & Li, 2016). Specifically, ICCs and PCAs do not assume linearity of data and look for general groupings, while regression analysis can be used when linearity between data types has been assumed. Importantly, scores that indicate increased strength of similarity between different data streams increase confidence in the use of alternative data streams, such as *in vitro* or *in silico* experiments, when extrapolating PODs.

Lastly, one *post hoc* option to assess data quality is to rank PODs derived from different data streams in order of potency and evaluate whether the rank ordering and POD values obtained are similar. Ideally, PODs should be similar between data streams but given that some variability in the value may exist when evaluated from different data sources, rank orders should be similar at a minimum. For example, with PHOPs, data were available for *in vivo* LD_{50}s, *in vitro* LD_{50}s, and *in silico* LD_{50}s. We used correlation analyses, both parametric and rank-based, to compare data streams and found that the *in vivo* and *in silico* data were aligned only when certain criteria were met (i.e., empirical LD_{50}<2,000 mg/kg). This result limited our confidence in the *in silico* data.

5. Anchor Chemical Selection

Once data has been curated, an anchor chemical should be selected. This chemical is used as a basis or comparator for deriving PODs for chemicals that are not as data-rich. During a class-based assessment, data-rich chemicals are often termed “anchor chemicals,” “index chemicals,” or “surrogate chemicals" within this framework.

Selection of the anchor chemical depends on multiple factors such as:

- Availability of high-quality data for the relevant endpoint
- Structural similarity between the anchor chemical and the rest of the chemicals
- Similarity in physicochemical properties and metabolic pathways
- Similarity in toxicological effects (i.e., similar toxic endpoints) between the chemicals within a group

By convention, the anchor chemical is often the most toxic chemical in the class, but this is not necessarily required. Particularly when a subclass contains many chemicals that have data only in the low or very low confidence categories (Section 6), choosing an anchor chemical with a higher-confidence TRV is likely to strengthen the overall class-based dose-response analysis even if that chemical is not the most toxic member of the subclass. In addition to its use to define a chemical class (Section 3), U.S. EPA’s GenRA tool can also inform the choice of the anchor chemical.

It may be that a subclass needs different anchor chemicals for different endpoints such as the use of different anchor chemicals for the noncancer and cancer assessment of a subclass. For example, during our assessment of the PHOPs chemicals, TCIPP was selected as the anchor for the noncancer outcome based on availability and quality of data. However, available cancer dose-response data for TCIPP were based on tumors caused by an MOA deemed not relevant to humans. Thus, TCEP was initially selected as the anchor chemical for deriving RPFs for cancer. However, differences between the MOAs for TCEP and the other chemicals with positive tumor data (i.e., TDCIPP and TDBPP) meant that a class-based assessment was not possible for cancer effects of the PHOPs.

6. Deriving PODs

Overall, based upon the case study derivation of PODs for PHOPs, our recommendation is to use a tiered approach using the highest quality available data to derive a POD for each chemical. Different sources of information (Section 4.1) will be of different quality, and thus class-based PODs derived from them will be associated with different levels of confidence (Table 1). Note that Table 1 provides baseline confidence levels for the different data sources. This baseline confidence can be adjusted up or down based on the considerations noted in Table 1. This approach maximizes the use of the best quality data in the TRV derivation to make the best use of all available data. Relying on a single method of deriving PODs (e.g., only RPFs based on NAMs data) is likely to be limiting because a single method does not allow for variability in confidence in similar data streams obtained under different conditions. For example, there is much higher confidence in transcriptomic data obtained from the target organ or a related target where multiple doses were tested, compared with transcriptomic data from a nontarget species or organ that only tested one dose level. Thus, it is preferable to apply a variety of approaches to determining PODs for different chemicals within a single subclass. An additional benefit of deriving PODs using multiple methods is that a comparison of the PODs/TRVs derived for a given chemical using different methods can be used to benchmark the methods and perhaps provide insight as to their relative confidence in the final analysis. Thus, our recommendation is to derive PODs from all data streams with relevant data, and then conduct a post hoc comparison and select the most reliable POD (Section 7).

The following section describes an approximate hierarchy within which to consider different data streams and the level of confidence that can be assigned to PODs/TRVs derived from them. Suggested starting points for confidence in each data stream are discussed below and presented in Table 1, but confidence should be evaluated based on the data available for the specific chemicals in the subclass. Note that the relative confidence of PODs/TRVs derived using different categories of data, particularly lower confidence categories, may vary (e.g., TRVs based on NAMs data could be of higher or of lower confidence than TRVs based on *in vivo*–derived RPFs depending on the considerations noted in Table 1). Further, specific caveats that should be considered for tumor endpoints are presented, as these endpoints require special consideration.

Table 1. Mapping of Data Availability Category to Baseline Confidence Level and Factors Affecting Confidence Level

Data Availability Category	Suggested "Baseline" Confidence Level	Factors that Generally Increase Confidence[a,b]	Factors that Generally Decrease Confidence[a,b]
Available TRV(s) based on published assessment	High	• Multiple agencies agree on value of POD • Multiple agencies agree on critical effect • TRV derived based on most current methods • Human relevant MOA	• Conflicting agency reports on value of POD and/or critical effect • Agency reports that identify a POD but note insufficient confidence to derive a TRV • TRV derived based on outdated methods • Uncertainty related to human relevance of MOA
Newly derived TRV based on animal data	Medium	• Multiple (existing[c]) studies from different research groups are consistent in terms of POD value and critical effect • Principal studies are well conducted and well-documented, following test guidelines • Critical effect is consistent with toxicity observed in other studies of different design (e.g., one study observing histopathology changes, another observing related serum biochemistry changes) • Critical effect consistent with that of other class members	• Existing studies are inconsistent in terms of POD value and/or critical effect • Critical effect is unrelated to the critical effect for other class members. • There is no support showing a consistent critical effect across study designs (no studies or inconsistent results) • All existing studies are from a single research group • Only a single (or a small number) of studies are available
POD based on empirical NAMS data (e.g., tPOD, non-traditional animal model)	Low-Medium	• Similar to or lower than apical PODs for other class members but by a factor of 10 or less • Mechanistic connection between NAMs and critical effect *in vivo* (e.g., gene expression changes for pathways related to the critical effect *in vivo*)	• Higher than apical PODs for other class members • More than a factor of 10 lower than PODs for other class members • The NAMs data are unrelated to the critical effect *in vivo*
In silico-predicted POD	Low	• Data are based on a more accepted *in silico* model	• Data are based on a less accepted *in silico* model
RPF	Medium, Low, or Very Low if based on *in vivo, in vitro, or in silico* data, respectively	• Data used to derive RPF are clearly related to critical effect • Data are based on an accepted *in vitro* assay system or *in silico* where applicable • Data used to derive RPF are obtained under the same test conditions or *in silico* model for	• Data used to derive RPF are not clearly related (or are clearly not related) to critical effect • Data are based on a less well validated *in vitro* assay system or *in silico* where applicable • Data used to derive RPF are obtained under substantially different test conditions or *in silico*

Data Availability Category	Suggested "Baseline" Confidence Level	Factors that Generally Increase Confidence[a,b]	Factors that Generally Decrease Confidence[a,b]
		the anchor chemical and target chemical where applicable	model for the anchor chemical and target chemical where applicable
Read-across (RPF=1)	Medium, Low, or Very Low if based on *in vivo, in vitro, or in silico* data, respectively	• Anchor chemical is as toxic or more toxic than target chemical(s) • Data suggest anchor and target chemical(s) have the same or similar critical effect, target organ/system, and/or MOA	• Anchor chemical is less toxic than target chemical(s) • Data do not suggest anchor and target chemical(s) have the same or similar critical effect, target organ/system, and/or MOA
Most sensitive POD	Low to Very Low	• Greater degree of similarity to high-confidence PODs (especially anchor) • PODs predicted by this method are higher than the POD-equivalent of the TTC • Both considerations help to ensure PODs are not over-conservative	• Lower degree of similarity to high confidence PODs (especially anchor) • PODs predicted by this method are lower than the POD-equivalent of the TTC
TTC	Low to Very Low	• Very low confidence (i.e., precision) because the TTC is purposely a highly conservative value	• Chemical is not in the domain of applicability of TTC
Anchor chemical POD/10	Low to Very Low	• Greater degree of similarity to high-confidence PODs (especially anchor) • PODs predicted by this method are higher than the POD-equivalent of the TTC • Both considerations help to ensure PODs are not over-conservative	• Lower degree of similarity to high confidence PODs (especially anchor) • PODs predicted by this method are lower than the POD-equivalent of the TTC

TRV = toxicity reference value; POD = point of departure; MOA = mode of action; NAMs = new approaches/alternative methods; tPOD = transcriptomic point of departure; RPF = relative potency factor; TTC = threshold of toxicological concern.

[a] Factors may increase or decrease confidence by a "full" level or a "partial" level (e.g. from Medium to Low or from Medium to Low-Medium). By necessity, such decisions require some level of professional judgement. Consider the number of factors and the magnitude of their effect. High may be promoted to Very High, though Very High is not suggested as a baseline confidence level.

[b] For all data categories, consistency with other members of the subclass in terms of POD/TRV value and critical effect (either specific effect or more general target) will increase confidence, while inconsistency will decrease confidence.

[c] A single high-quality, and especially guideline compliant, traditional animal study is usually sufficient to establish a POD/TRV, so the absence of multiple existing traditional animal studies should not necessarily decrease confidence in a POD/TRV derived from a single such study. While multiple existing traditional animal studies may increase confidence in PODs/TRVs, for the purposes of establishing a medium "baseline confidence level", it is usually not necessary (and may indeed be a waste of resources) to conduct additional high-quality traditional animal studies of the same design when one already exists.

6.1. Noncancer Endpoints

6.1.1. Agency Assessments (Baseline High Confidence PODs)

The most desirable POD to use is one based on an existing high-confidence assessment previously selected by an agency to derive a TRV. Agency assessments may show the derivation of TRVs from PODs using UFs or the POD may be used more directly with implied UFs to determine a minimum margin of exposure (MOE). Additionally, TRVs should be converted to "standardized PODs" roughly equivalent to a NOAEL or a BMDL from a rodent study to facilitate comparison to PODs from other evidence streams later (Section 7). For example, a reported LOAEL would be divided by a factor of 10 to approximate a NOAEL. To calculate a TRV from these standardized PODs, two UFs are needed: UFH and UFA. Note that these adjustments are made to the POD to the degree necessary, rather than back-calculating PODs from TRVs. This approach avoids the need for the POD identification step to consider modifications to default UFs, such as the use of chemical-specific adjustment factors (CSAFs) (IPCS, 2005) or data-derived extrapolation factors (DDEFs) (U.S. EPA, 2014).

Although agency assessments tend to be high quality, it is important to critically evaluate even agency TRVs to make sure that they reflect the current science, particularly if the assessment is more than ~10 years old. For example, there has been substantial progress in the past decade in understanding of MOA and human relevance of certain tumor MOAs (Corton et al., 2014; Dellarco et al., 2006; Felter et al., 2018). Finding agency assessments that differ by more than a factor of 3 in their standardized POD would also indicate that a closer look is required to compare the critical effect and derivation of the TRV. If the PODs are of similar quality, the general approach is to choose the TRV from the most recent assessment (based on the presumption that it reflects the most recent data and most recent science). In the rare case where there are two TRVs of comparable quality from the most recent year, the more sensitive one can be chosen as a health-protective approach. An exception would be if the standardized POD from an earlier assessment was lower than the more recent one, *and* the principal study in the earlier assessment was not cited by the later assessment.

In short, PODs used by agencies to derive PODs are likely to be high confidence, based on the factors that go into deriving a TRV. However, the underlying data and approach used do need to be considered.

6.1.2. Additional Animal Data (Baseline Medium Confidence PODs)

If no agency assessment exists, the next approach would be to identify studies (published or grey literature) that could be used to derive PODs. Any existing TRVs (as described above) should inform the choice of key targets.

Given that a systematic literature review for any data on chemicals of interest can be costly and time consuming, utilizing databases with supplemented targeted searching may be a faster method to identify data available for deriving a POD. For example, databases such as ToxValDB are efficient sources of POD information. This database is batch searchable and downloadable via the EPA CompTox Dashboard as a spreadsheet that is easily edited to be filterable to quickly identify potential PODs for species of interest. However, ToxValDB may not include all of the relevant studies of interest. A few targeted literature searches may be useful to identify additional studies suitable for deriving a POD beyond those included in databases. Conducting a limited literature search focused on review articles may be the most efficient option. Another option would be a targeted literature search, perhaps informed by endpoints from existing agency TRVs for other class members or other existing information.

If agency assessments used to identify existing PODs (Section 6.1.1) are greater than ~10 years old, it may be useful to conduct a targeted literature search even for chemicals with existing TRVs. For example, if the existing TRV is based on a subchronic study, it could be useful to search for a chronic study both on the NTP website and in a literature database such as PubMed. This is especially true if the TRV was based on a subchronic NTP study. Similarly, if there is an indication that the class targets the reproductive system or development, it could be useful to do a targeted literature search for those endpoints. The goal of such targeted updates would be to identify relevant new studies that could affect the TRV without needing to do all of the work for developing a new TRV.

Once relevant studies are identified, they are reviewed to identify relevant adverse effects and PODs. Ideally, this POD would be a BMDL, but a NOAEL or LOAEL may be used if the data are not suitable for BMD modeling or if a meaningful BMDL cannot be obtained (Section 2). In this step, priority should be placed on traditional animal models such as rats or mice, though other mammalian models may be included if the endpoint is relevant to humans.

For the PHOPs, initial work towards identifying relevant studies was done as part of a literature survey under this call order and further supplemented by a Division of Translational Toxicology (DTT) literature search. Although neither the literature survey nor the DTT literature search identified key studies, the results of these tasks, together with an evaluation of mechanistic effects and adverse effects in the sort of class-based

hazard assessment conducted in CO-1, helped in identifying potential key targets and studies.

Generally, PODs independently derived by the assessors from animal data would be considered medium confidence. While these PODs would have been derived using traditional methods, they are lower confidence than the high confidence PODs because they have not undergone the level of peer review associated with agency TRVs or published TRVs. The absence of agency assessments may be because no organization has prioritized resources to develop a TRV, there are not enough high-quality studies to conduct a high-quality assessment, high-quality studies became available only recently, or some other reason.

6.1.3. NAMs Data (Baseline Confidence of PODs Varies by Data Stream — Low-Medium for Empirical, Low for *In Silico*)

If no suitable traditional animal data exists for derivation of a POD for a given chemical, the next step is to identify empirical NAMs data from which PODs could be derived, as opposed to computational approaches (i.e., *in silico* NAMs). Empirical NAMs data can also be useful for chemicals with existing or independently-derived TRVs to aid in benchmarking PODs derived for chemicals without TRVs.

Examples of potentially useful empirical NAMs include transcriptomic data, other *in vitro* data, and zebrafish assays. At this time, one of the more common approaches is to derive PODs based on transcriptomic data (tPODs). Data from other NAMs may become available and useful as more data are collected in ToxCast and other databases. If data other than lethality (e.g., zebrafish) or transcriptomic data are used, it is essential that the data from different chemicals are directly comparable (i.e., same endpoints with comparable conditions for testing). For example, they must have been tested under identical test conditions and results need to be reproducible within and between laboratories. Quantitative use of receptor binding assays for identification of PODs is currently limited because of large inter-test differences in the background variability making it problematic to define the BMD in terms of a one standard deviation change in the control mean. PODs based on *in silico* models may also be another available data stream. For example, the Chiu Dashboard (Aurisano et al., 2023; Kvasnicka et al., 2024) provides *in silico* PODs based on either general noncancer or reproductive/developmental endpoints.

For the PHOPs, the selected empirical NAMs data stream, transcriptomic data, were available only for chemicals that also had traditional *in vivo* test data. Therefore, the transcriptomic data were not used in place of the available high quality data. However,

the chemical coverage of transcriptomic data may be broader for other chemical subclasses or in future years.

Generally, PODs derived from empirical NAMs data would be considered low-medium confidence because of the following reason. While they are based on empirical data, these methods are generally less well accepted for risk assessment purposes than more traditional methods. A number of factors could contribute to increasing or decreasing confidence in PODs based on NAMs. In general, results consistent with higher quality data streams would increase confidence. A mechanistic connection between data from NAMS and the critical effect observed *in vivo* is key. Numerical values of NAM PODs that are similar to or lower than (i.e., more protective than) apical PODs are desirable. However, if PODs derived from different data streams differ by more than one order of magnitude, there is lower confidence in the POD. This order of magnitude estimate is based on both the definition of the RfD as an estimate within an order of magnitude, and an analysis by Health Canada (2021) comparing assessments based on *in vivo* and *in vitro* data. That comparison concluded that use of *in vitro* data would require a standard factor of 10 for human variability, plus an additional factor of 10 related to the use of *in vitro* data. Health Canada broke this additional factor of 10 into a factor of 3 for uncertainty regarding whether all possible effects have been tested (i.e., incomplete biological space, similar to EPA's database UF), and a factor of 3 for limitations in cell-based assays. This Health Canada study is described in additional detail in the Subtask 2A report under the current call order. For transcriptomic data, a greater number of relevant responding genes and data generated from a target tissue would increase confidence.

PODs derived from *in silico* NAMs would generally be considered low confidence. *In silico* predictions would usually be lower confidence than empirical data.

6.1.4. RPF and Read-across Approaches (Baseline Confidence of PODs Varies by Data Stream – Medium for *In Vivo,* Low for *In Vitro,* Very Low for *In Silico*)

RPFs involve a direct comparison of the target chemical's potency with that of an anchor chemical using the same measurement technique for the same endpoint. RPFs are class-based by definition and have the advantage of using data from a (usually) well-defined endpoint from a data-rich class member for benchmarking other class members. RPFs are frequently applied in evaluating classes of chemicals, although more typically as part of a cumulative assessment than in class-based evaluations of individual chemicals. For example, CPSC and other groups have applied RPF-based approaches in cumulative assessments of phthalates (CHAP, 2014; EFSA, 2019; U.S. EPA, 2023). RPFs may be derived using *in vivo*, *in vitro,* or *in silico* data, with confidence generally decreasing in the

following order: *in vivo* data > *in vitro* data > *in silico* data. Ideally, the RPF is calculated by comparing the potency of the class members using the same assay and one related to the critical effect.

It may also be possible to conduct read-across. Read-across is analogous to applying an RPF = 1 and shares the key requirements of applying RPFs. However, it also requires that the anchor chemical is as toxic or more toxic than the target chemical(s).

Ideally, an RPF analysis should be based on data from studies conducted using an exposure length and route of exposure consistent with the problem formulation, and from a common species and target organ for a group of chemicals in that class. When this is not possible, RPFs can be extrapolated to other routes, although this increases the uncertainty. For example, RPFs for polycyclic aromatic hydrocarbons (PAHs) are derived from dermal data, but typically applied to oral exposures (Haber et al., 2022). It is worth noting that risk assessment of chemicals as a group usually adds complexity to the evaluation methods and can make interpretation of the results difficult for individual compounds (Savitz & Hattersley, 2023).

Additionally, the RPF is ideally calculated by comparing the potency of the class members using an assay related to the critical effect (Section 2.1.1). When data are not available for an assay related to the critical effect or its MOA, other data can be used as long as a generally comparable approach (e.g., exposure duration, assay type) was used for evaluating all of the chemicals. For example, using LD_{50}s to derive RPFs is likely to improve chemical coverage at the expense of some level of confidence (Section 4). ToxValDB is one potential source of *in vivo* data to calculate RPFs, though many others exist. *In vitro* data (e.g., AC_{50}s) from ICE for assays related to the endpoint/target of interest may also be suitable for RPFs. However, *in vitro* data may be lacking for more data-poor subclass members potentially limiting the potential for calculating RPFs based on *in vitro* data. Predicted LD_{50}s from the Danish QSAR Database may also be a source of data to derive RPFs, though this approach would require careful evaluation of quality compared to other data streams to ensure the plausibility of the predicted values (Section 4.2).

RPFs derived from different data streams would themselves have varying levels of quality, and thus, PODs derived from them would also vary in confidence. Generally, RPFs derived from traditional animal data related to the critical effect would be the highest quality, those from *in vitro* NAMs or nontraditional animal models such as zebrafish would be intermediate quality, and those based on *in silico* predictions would be lowest quality. In addition to these data streams, the endpoint evaluated is important in determining the confidence in the POD. For example, data consistent with the endpoint (or at least

broader target) of existing or newly derived PODs (particularly that of the anchor/surrogate chemical(s)) would be of higher confidence than data that do not share this consistency. Inconsistency with the endpoint/target for the broader class may not be disqualifying but should be carefully considered. Further, as with PODs for the same chemical derived from different methods, RPFs derived from different methods can be compared to each other when the same chemical has data sufficient for more than one method (Section 4.2). While a full quantitative RPF-based analysis is preferable, when that is not possible (or is possible but of relatively low confidence), a qualitative or semi-quantitative approach may be suitable. In this case, instead of deriving RPFs, the potencies may be ranked (either individually or categorized into "buckets") and used to indicate that some chemicals are more potent or less potent than others without specifying by how much. However, such ranking should also be evaluated for reliability.

6.1.5. Alternative Approaches for PODs Based on Lower Quality Data

In some cases where very limited data are available for data-poor class members both in terms of potency and hazard, alternative approaches can be applied to determine an estimate of the POD. In these alternative approaches, the POD comes from data provided by other chemicals that are in the same class and does not require any chemical specific data on the target chemical. Depending on the data available for a given chemical class, one or more of these approaches may be appropriate. Considerations for evaluating relative confidence and choosing among these approaches are discussed in Section 7.1.

6.1.5.1. Most Sensitive POD (Baseline Low to Very Low Confidence PODs)

One possible approach when no data or limited quality hazard data exists is to assign the most sensitive (i.e., lowest numerical value) POD in the class to the target data-poor chemicals. This approach is not recommended in general when data are available to inform the POD because using the most sensitive POD in the presence of adequate quality data on the target chemical may result in an over-protective POD. However, this is a health-protective approach when there is substantial uncertainty about the potency of data-poor chemicals.

6.1.5.2. Threshold of Toxicological Concern (Baseline Low to Very Low Confidence PODs)

The threshold of toxicological concern (TTC) is defined as "a level of exposure for all chemicals below which there would be no appreciable risk to human health" (Patlewicz et al., 2008). As for the most sensitive POD approach, this approach can be applied in the absence of an empirical or *in silico* basis for estimating PODs (or in the case of lower confidence in such data). The POD-equivalent of the TTC is defined as the noncancer TTC x 100. TTCs can be identified in batch by determining Cramer class in ToxTree (Patlewicz

et al., 2008). Table 2 outlines the different Cramer classes and their associated TTCs in different units. Note that the original distributional analysis of NOAELs to determine the TTCs for the different Cramer classes was conducted using mg/kg/day and then converted to µg/person/day using a body weight of 60 kg (Kroes et al., 2000). This means that no additional correction to a 70 kg body weight is needed in converting to mg/kg/day.

Table 2. Values for Threshold of Toxicological Concern for Different Cramer Classifications

Classification	TTC (µg/person/day)	TTC (µg/kg/day based on 60 kg)	TTC (mg/kg/day)
Potential DNA-reactive mutagens and/or carcinogens	0.15	0.0025	0.0000025
Organophosphates and carbamates	18	0.3	0.0003
Cramer Class III	90	1.5	0.0015
Cramer Class II	540	9	0.009
Cramer Class I	1,800	30	0.03

TTC = threshold of toxicological concern.

6.1.5.3. Anchor Chemical POD/10 (Baseline Low to Very Low Confidence PODs)

A final possible approach for developing PODs for chemicals with limited quality hazard data or no data is to assign to these chemicals the POD for the anchor chemical divided by an additional factor of 10. This factor of 10 is based on the phrase in the reference dose (RfD) definition that the RfD is an estimate with "uncertainty spanning perhaps an order of magnitude". This phrase has been interpreted to mean that the RfD is at the upper end, the lower end, or the middle of the range of an order of magnitude (U.S. EPA, 2002). Based on a desire to express the TRVs relative to the anchor chemical, but be more conservative, a factor of 10 seems reasonable based on the assumption that the potency of the subclass members is within an order of magnitude. However, it is important to note that if subclass members have substantial differences in chemical functional groups (Section 3), this factor of 10 may not be sufficient. An example of a substantial difference is when key features that are related to toxicity (e.g., the presence of bay regions in PAHs) are present in only some class members and not others.

6.2. Cancer Endpoints

Class-based POD derivation approaches for cancer are generally similar to those for noncancer but also include some additional considerations. First, a dose-response analysis for cancer endpoints should exclusively use BMDLs as PODs instead of NOAELs or LOAELs (U.S. EPA, 2005). Second, there is a greater emphasis on MOA when

considering cancer endpoints. As with noncancer endpoints, the weight-of-evidence analysis should consider whether the MOA for the cancer endpoint is relevant to humans. Additionally, whether the MOA is genotoxic or nongenotoxic has implications for the low-dose extrapolation approach (Section 2.2). Third, NAMs appropriate for cancer dose-response are not available at this time. Genotoxicity assays such as the Comet assay, Ames assay, and DNA binding (adduct formation) assays have been used in the qualitative prediction of whether a chemical is likely to be a carcinogen. Similarly, although dose-response analyses have been conducted on genotoxicity data, the goal of such analyses is to predict the risk of genotoxicity endpoints, rather than predicting cancer risk (MacGregor, Frötschl, White, Crump, Eastmond, Fukushima, Guérard, Hayashi, Soeteman-Hernández, Johnson, et al., 2015; MacGregor, Frötschl, White, Crump, Eastmond, Fukushima, Guérard, Hayashi, Soeteman-Hernández, Kasamatsu, et al., 2015; Thienpont et al., 2024). DNA adducts may or may not result in a mutation, and a mutation may or may not result in cancer. The relationships among adduct formation, mutation, and tumorigenesis depends on the nature of the adduct, its rate of repair, and whether the repair is accurate (Bates et al., 2023). Therefore, there is not an established quantitative relationship among adduct formation, mutation incidence, and tumor formation, limiting the use of genotoxicity data for the quantitative prediction of cancer.

7. POD Selection and Deriving Final TRVs

7.1. POD Selection

The proposed approach for deriving TRVs considers all available data. This approach also includes derivation of PODs for multiple data streams for the same chemical when such data exist to aid in evaluating relative confidence of the different approaches and determining the best POD for each chemical. Once individual PODs have been derived for each chemical by any of the available methods, each POD needs to be evaluated in the context of other PODs available for that chemical and in the context of the entire subclass of chemicals being evaluated. Selection of the most appropriate POD for each chemical will depend on the level of confidence in the data used to derive it. As a general rule, the selected POD should be the one derived from the highest quality data. Data quality factors considered in the choice of POD include study quality, consistency of evidence across data streams, and endpoint/target/MOA (including both consistency across chemicals and human relevance).

Similarly, at this stage it would be useful to evaluate the relative appropriateness of presenting a separate POD for each chemical in the subclass (and thus a separate TRV) or a single POD applied across the entire subclass (or smaller group of chemicals within the subclass). For example, if derived PODs are based on a lower quality data stream and are substantially (perhaps one order of magnitude or greater) higher (i.e., less protective) than those for the rest of the class (especially the anchor chemical), consider direct read-across from the anchor chemical. If the data stream is of at least medium confidence, using a more conservative approach (such as the most sensitive POD, TTC, or anchor chemical POD/10) is not recommended since these latter approaches would result in PODs substantially lower than one derived based on the medium confidence data. However, if the data stream with the higher POD is of low confidence, it may be appropriate to use these more conservative approaches.

The choice among available approaches for PODs based on low quality data may also depend on the degree of similarity of the high confidence PODs or between the anchor chemical and the low confidence chemical. For example, if the anchor chemical POD/10 is lower than the POD for the most sensitive chemical, that could be a reason for using the anchor chemical POD/10, since the potency range of the class may not be adequately captured by the chemicals with available PODs. On the other hand, if there are several chemicals with PODs close to each other and the anchor chemical POD is among the lower ones, using the anchor chemical POD/10 may be overly conservative and it may be more appropriate to apply the POD for the chemical with the lowest high confidence

POD. Note that these values are best applied to chemicals that lack empirical data that can be used to derive a POD. While these approaches would be health protective, they would be overly conservative when high or medium quality PODs exist.

The TTC provides a bounding to these estimates in the face of uncertainty. Since the TTC is based on a low percentile of the distribution of NOAELs for a given Cramer class, it is reasonable to consider the value equivalent of the TTC a reasonable lower bound for the POD. If the anchor chemical TRV/10 and the most sensitive high quality TRV (both converted to POD equivalents) are lower than the TTC, it would be reasonable to apply the TTC to the data-poor class members.

POD selection is also an appropriate step for reevaluating the coherence of the chemical subclass. While a chemical subclass may have continued to "make sense" up through a hazard assessment-style analysis, the available dose-response data may lead to a different conclusion. For example, it may be that one subset of a subclass has dose-response data for reproductive/developmental targets, and a separate subset of the same subclass has dose-response data for liver effects in adults (and neither subset has dose-response data for the other target). Support for splitting the subclass by endpoint would be particularly strong if the division by endpoint was consistent with similar groupings suggested by the analyses described in Section 3. Additional analyses to determine whether the two apparent groups should be analyzed separately or together may be needed if there are some hazard data indicating shared endpoints but also some data indicating different targets or critical effects.

Alternatively, it may also be possible to conduct an "endpoint-agnostic" analysis where the most sensitive POD for each chemical is chosen without regard to shared endpoints/targets across subclass members. The "general noncancer" PODs in the Chiu database are one example of such an endpoint-agnostic approach. This may be suitable only when no clear consistency of endpoint/target exists neither for the entire subclass nor for two or more subsets within the subclass. However, justification for choosing to conduct an endpoint-agnostic analysis, and interpretation of such an analysis, may prove challenging.

7.2. Final TRV Derivation

The overall goal is to derive health-protective TRVs that reflect both the available data and the level of confidence for each category of chemicals. When relatively higher confidence data exist, using those PODs would be best. On the other hand, when existing data used to identify PODs or derive RPFs are themselves lower confidence, the greatest confidence may be based on applying a conservative adjustment to the anchor chemical

TRV. Particularly, when there is lower confidence in TRVs derived using the RPF method, an additional factor may need to be applied to the TRV from the anchor chemical to be sufficiently protective.

Once the final approach to POD selection is determined, the focus switches from the POD for the anchor chemical to the associated TRV or CSF. This means that the RPFs are applied to the TRV or the CSF rather than to the POD. This approach avoids the need to reconsider UFs and instead directly expresses the target chemical TRV relative to the anchor chemical TRV. In the case where new animal data or NAMs data of sufficient quality are identified and chosen for the POD, UFs appropriate to the individual chemical would be applied to calculate the TRV. For chemicals where the chosen POD was the anchor chemical POD/10 or the most sensitive POD, the target chemical TRV would be based on the anchor chemical TRV/10 and the most sensitive TRV, respectively. The TTC is designed to serve as a TRV substitute, thus allowing the TTC to be used as the TRV if that method is chosen. Finally, for generic *in silico*–derived PODs such as those from the Chiu database, appropriate UFs need to be applied to calculate a TRV. It is not clear at this time which UFs should be applied to *in silico*–derived PODs to derive a TRV. The PODs present in the Chiu database can illustrate some issues to consider when selecting UFs. *In silico* PODs (Aurisano et al., 2023; Kvasnicka et al., 2024) are based on *in vivo* data from ToxValDB so UFA=10 and UFH=10 are logical to include but additional research is needed to determine whether additional UFs are needed and what areas of uncertainty they should cover.

7.3. Cancer Considerations

The same general process should be considered for deriving TRVs and evaluating confidence in a class-based dose-response analysis for cancer endpoints as for noncancer endpoints. However, some additional considerations are necessary for cancer endpoints.

Due to the absence of cancer-related NAMs that can be used quantitatively, options for developing cancer PODs are more limited than those for noncancer PODs (Section 6.2). PODs from existing assessments or new ones derived from animal data are preferred after considering issues noted regarding MOA and human relevance. CSFs based on the same target tissues would be preferable. However, if a chemical induces tumors in two different target tissues, and only the less sensitive target tissue is shared with the class, it would be necessary to carefully consider the relative benefits of a shared target compared with the additional health protectiveness of using the CSF for a different target tissue.

It is likely that most class-based cancer assessments would be based on RPFs, which may be calculated based on tumor data (as done for polycyclic aromatic hydrocarbons (PAHs)) or, in rare cases for well-studied classes, on a key event related to carcinogenesis (e.g., aryl hydrocarbon (Ah) receptor binding for dioxins). The anchor chemical would generally be the most well-studied chemical in the subclass although a more potent chemical that is also well-studied would be a reasonable alternative. Precedent for classes where the most potent chemical is not the most well-studied is unavailable to provide more specific guidance. However, if the available RPFs are of low confidence and are insufficient to support the conclusion that the target chemical is less potent than the anchor chemical, the anchor chemical CSF (i.e., potency) could be multiplied by 10. This approach is analogous to the anchor chemical POD/10 approach used for TRVs.

Carcinogens acting via mutagenic MOAs will also have their own TTC value while those acting via non-mutagenic MOAs will have the same TTC as non-carcinogens. When considering weight-of-evidence for carcinogenicity, qualitative categorizations by agency assessments may be informative. For example, "some evidence" and "possible human carcinogen" do not generally lead to cancer quantification. Additional available data may override this consideration.

8. PHOP Class-based POD Selection and Class-based TRV Derivation

The Subtask 2B report details the full process of data identification and evaluation to select PODs and derive class-based TRVs for PHOPs. This section focuses on how final TRVs were developed. As noted earlier in this report, the process, recommendations, and lessons learned reflect the current state of data and may need to be updated to keep pace with rapidly evolving science.

For PHOPs, we used a variety of methods, including class-based TRVs from agency assessments, direct read-across from the anchor chemical, RPFs based on *in vivo* rat LD_{50}s, and applying the anchor chemical TRV/10. This last approach (i.e., anchor chemical TRV/10) was applied to chemicals with three data types: *in vivo* LD_{50}s deemed unsuitable for deriving RPFs, *in silico* LD_{50}s (all of which were deemed unsuitable for deriving RPFs), and no available empirical or *in silico* data. Two of these three (i.e., *in silico* LD_{50}s and no available data) share the same confidence level. Additional information could allow for further distinction in the methods and/or confidence levels applied to these data types. For example, better concordance between *in silico*–predicted and empirical LD_{50}s could have resulted in applying different TRV derivation methods or different confidence levels to the chemicals that have *in silico*–predicted LD_{50}s. The specifics of the approach to select and derive the final PODs are outlined below.

8.1. Noncancer

Four PHOPs were assigned high confidence TRVs. These chemicals had existing high quality agency assessments that either directly derived a TRV (TCEP, TCIPP, and TDCIPP) or used an MOE approach from which a TRV was derived (V6). While using existing TRVs derived individually may not appear to be a class-based approach, the similar/related endpoints (reproductive or developmental effects) across existing TRVs for these four chemicals support their use on a class basis. From these existing TRVs, TCIPP was selected as the anchor chemical for deriving TRVs for other PHOPs when the selected approach required it. TCIPP was selected because its POD shares a developmental endpoint with most other PHOPs for which experimental data exist to derive a TRV, it had the most *in vivo* LD_{50} values (13), it had a relatively large number of existing risk assessments, and its *in silico*–predicted rat oral LD_{50} (1,200 mg/kg) was similar to the median *in vivo* rat oral LD_{50} (1,260 mg/kg).

Three PHOPs were assigned medium confidence TRVs. These PHOPs had additional PODs for developmental endpoints identified in ToxValDB. Though the endpoints identified

were similar to those for the four existing TRVs, the values of the PODs were 10-40 times higher than those of the four high quality TRVs. The relatively high value of the PODs, in combination with the relatively limited documentation of the associated toxicity studies for these three PHOPs, meant that there was substantial uncertainty that the most sensitive endpoint had been identified. Thus, a more conservative approach of direct read-across from the anchor chemical was selected to derive TRVs for these three PHOPs.

Four PHOPs were assigned low-medium confidence TRVs. Evaluation of *in vivo* rat LD_{50} values indicated that LD_{50}s for these four PHOPs were suitable for calculating an RPF. Thus, TRVs for these chemicals were calculated by applying the RPF to the TRV for the anchor chemical. The uncertain relationship between the LD_{50}-derived RPFs and the reproductive/developmental endpoints for the subclass is unclear, reducing confidence in these TRVs.

Two PHOPs were assigned low confidence TRVs. While these PHOPs also had *in vivo* rat LD_{50}s, the LD_{50}s were not considered reliable for calculating RPFs. The TRVs for these two PHOPs were calculated by dividing the TRV of the anchor chemical by an additional factor of 10. These two TRVs were consistent with the most sensitive of the four TRVs derived using the RPF approach and were generally consistent with the *in silico*-predicted reproductive/developmental PODs from the Chiu Dashboard.

Sixteen PHOPs were assigned very low confidence TRVs. Of these, thirteen PHOPs had *in silico*-predicted rat LD_{50}s though these LD_{50}s were not considered reliable for calculating RPFs. The final three PHOPs had no empirical or *in silico* data available. For all of these PHOPs, the TRVs were set as the anchor chemical TRV/10. These TRVs were consistent with the most sensitive of the RPF-derived TRVs and the Chiu Dashboard. Confidence in these TRVs was further lowered by the general lack of data for these chemicals.

8.2. Cancer

Only three PHOPs (TCEP (115-96-8), TDCIPP (13674-87-8), and TDBPP (126-72-7)) had existing cancer dose-response data relevant to humans, with all three sharing an endpoint of kidney tumors. Of these three, one (TCEP) is thought to act via a nongenotoxic MOA, while the other two are thought to likely act via a genotoxic MOA. A fourth PHOP (TCIPP (13674-84-5)) induced some rodent tumors by an MOA that is not human relevant; other potentially human-relevant tumors were also observed with this PHOP but at a lower weight-of-evidence. Therefore, this fourth PHOP (TCIPP) was not included in the cancer quantification.

Deriving TRVs based on such a small database was not pursued because there was low confidence in the quality of data and the resulting RPF estimates. RPFs were derived for the cancer endpoints, as an illustrative example, using the same method as for noncancer assessments. TCEP was chosen as the anchor chemical for the cancer quantification because of the relatively large differences between the *in vivo* and QSAR-predicted LD_{50}s for the other two chemicals that caused kidney tumors. However, given that TCEP is the only one of the three PHOPS with human-relevant tumor data thought to act via a nongenotoxic MOA, choosing one of the PHOPs acting via a nongenotoxic mechanism as an anchor chemical might have been more appropriate for deriving TRVs. The application of RPFs for the cancer assessment of the PHOPs was considered as proof of concept only since there was no good choice for an anchor chemical.

9. Discussion

This report outlines a generalized tiered approach to conducting class-based dose-response analyses for chemical subclasses with lessons learned from the application of this approach to the PHOPs subclass. The proposed tiered approach aims to use the best data available for each chemical, while considering consistency between chemicals in a subclass. It also leverages the existence of multiple types of data to evaluate confidence in the different varieties of data.

Generally, the proposed analysis relies on a tiered approach to evaluating shared endpoint across a subclass for conducting a class-based dose-response analysis. While structural similarity alone is insufficient, other kinds of biological similarity may suffice, including but not limited to target organ/system, AOP, or MOA leading to similar health effects in different tissues. Additionally, the tiered approach also provides opportunities to readily incorporate new data to increase confidence for relatively data-poor chemicals (or entire chemical groups).

A major lesson from the PHOP assessment is the need to think broadly about available data from the beginning and throughout the process. When identifying existing TRVs, it is important to ensure that the assessment reflects the current science regarding the identification of adverse effects, the adversity of endpoints identified, and the evaluation of MOA and human relevance. Similarly, these issues should be considered when reviewing new studies rather than simply modeling every endpoint. These issues should also be considered as part of the comparative analyses of TRVs and their underlying data. This could potentially enable more targeted approaches (e.g., prioritizing data for a particular endpoint when looking to transcriptomics, *in silico* results, other NAMs to fill in gaps). It could also suggest specific databases to look for the most relevant NAMs data, though doing so may not ultimately increase chemical coverage within a subclass.

In considering how this approach can evolve in the future, some of the options for incorporating additional data may result in group realignment. Applying the full approach for re-evaluation of chemical groups (Section 3), particularly if new data are generated, may present alternative groupings that enable analysis of chemicals that otherwise might be excluded. For example, one of the OFRs excluded from the PHOP subclass in CO-1 contained a triazine. The Polyhalogenated Triazine OFR subclass itself is relatively data-poor but would be an obvious first candidate for this former PHOP and (re-)evaluating if any sort of assessment could be done. An additional data stream of interest is metabolism data. Specifically, it is possible that information on common metabolism of PHOPs from CO-1 could have been used to more directly inform chemical grouping for dose-response. However, there was substantial uncertainty in this information for PHOPs

as described as follows. Given that most of the PHOP metabolites identified were predicted and not empirically confirmed, there was no available confirmation of whether shared metabolites were also major metabolites or whether parents, shared metabolites, or non-shared metabolites were the toxic form. However, if such confirmation is possible for other chemical groups or becomes possible for PHOPs in the future, consideration of metabolites may aid in addressing dose-response data gaps.

In addition to class-realignment, new empirical NAMs data could be easily incorporated to generate high quality PODs and categorize hazard. There are several relatively easy ways to acquire new data that can be used to generate a dose-response curve or calculate RPFs. One key challenge noted in several earlier reports (Subtask 1C and Subtask 2B of the current call order, as well as in the deliverables from CO-1) is the existence of NAMs data (e.g., AC_{50}s in ICE) primarily/only for chemicals that are already relatively data-rich. The availability of such *in vitro* data for chemicals that are data-poor could have provided additional approaches to POD determination and potentially higher confidence TRVs. Similarly, common hazard endpoints shared among members of a group based on existing *in vivo* data can be used to limit the scope of additional data to be acquired or point to key additional data that could enhance the confidence in the assessment. For example, several endpoints were common across OFR subclasses in evidence maps produced during the literature survey, including developmental and reproductive (DART) endpoints and endocrine-related endpoints. DART was a relatively data-rich category of animal data, while endocrine-related endpoints were a relatively data-rich category of *in silico* data. Both categories could benefit from additional *in vitro* data. For DART, additional *in vitro* data could help to fill in the gaps by providing data for additional chemicals. For endocrine, *in vitro* data for a more limited set of chemicals could be used to bridge the gap between the available *in vivo* and *in silico* data to establish greater confidence in the *in silico* data that already cover most OFRs.

Currently, although having NAMs data for chemicals that have *in vivo* data is helpful for validation of NAMs, it is not helpful for expanding the number of chemicals that have data, in light of the limited amount of NAMs data for chemicals lacking *in vivo data*. In the future for OFRs, targeting DART and endocrine endpoints with NAMs assays may be a way to generate additional data for most if not all OFR subclasses while using a very limited number of assays. For example, DART-relevant NAMs such as those in zebrafish may be broadly applicable across OFR subclasses. Zebrafish developmental assays are being developed toward regulatory use in drug development (Weiner et al., 2024). Assays such as ToxCast endocrine assays may be a relatively low lift path to generate *in vitro* endocrine data for a greater number of OFR chemicals. Further, these assays can be scaled to easily provide data at multiple doses, allowing for assessment of dose-responsiveness and derivation of a POD. Using this existing assay pipeline may also be an

opportunity for further collaboration with DTT that can broadly be applied to other classes of chemicals beyond OFRs.

Lastly, the lessons learned from the noncancer assessment apply to cancer assessments with additional considerations. A key additional lesson from the attempted class-based cancer dose-response assessment for PHOPs is that the additional MOA requirements for cancer assessments may preclude the development of cancer assessments even when noncancer assessments are possible. Specifically, the lack of a shared cancer-causing MOA and lack of human relevance of identified MOAs meant that it was not possible to apply the assumption that all class members could be evaluated together based on a common hazard and MOA. For PHOPs, future work could look to identify cancers that share common target organs and a human relevant MOA to apply a class-based approach. To minimize the need for conducting new bioassays, a tiered approach could be applied. This might begin with genotoxicity assays, as well as targeted *in vitro* or short-term *in vivo* assays testing potential key events for nongenotoxic MOAs (e.g., increased cell proliferation, endocrine activity, immunosuppression) and proceed with targeted testing using the approach recommended by Cohen and colleagues (Cohen et al., 2019; Doe et al., 2022).

Overall, the work presented here provides a step-by-step guide with key considerations for the application of class-based dose-response. The approach relies on shared endpoints and a tiered approach to incorporate the highest quality data available for each chemical. The current paradigm is both responsive to new data, as it can be readily used to update existent TRVs with new TRVs, and informative of data gaps as it is easy to see where limited or no data exists. Lastly, the key decision points highlighted in our report, as well as presentation of data related to the PHOPs case study, demonstrate the strength of this approach, provide a concrete example of how it might be applied in the future to other subclasses, and how new data can be integrated to strengthen the estimated TRVs.

References

Aurisano, N., Jolliet, O., Chiu, W. A., Judson, R., Jang, S., Unnikrishnan, A., Kosnik, M. B., & Fantke, P. (2023). Probabilistic points of departure and reference doses for characterizing human noncancer and developmental/reproductive effects for 10,145 chemicals. *Environmental health perspectives*, *131*(3), 37016. https://doi.org/10.1289/ehp11524

Bates, C. A., Haber, L. T., Moore, M. M., Schoeny, R., & Maier, A. (2023). Development of a framework for risk assessment of dietary carcinogens. *Food and Chemical Toxicology*, *180*, Article 114022. https://doi.org/10.1016/j.fct.2023.114022

Bevington, C., Williams, A. J., Guider, C., Baker, N. C., Meyer, B., Babich, M. A., Robinson, S., Jones, A., & Phillips, K. A. (2022). Development of a flame retardant and an organohalogen flame retardant chemical inventory. *Scientific Data*, *9*(1), 295. https://doi.org/10.1038/s41597-022-01351-0

Boobis, A. R., Doe, J. E., Heinrich-Hirsch, B., Meek, M. E., Munn, S., Ruchirawat, M., Schlatter, J., Seed, J., & Vickers, C. (2008). IPCS framework for analyzing the relevance of a noncancer mode of action for humans. *Crit Rev Toxicol*, *38*(2), 87-96. https://doi.org/10.1080/10408440701749421

Chronic Hazard Advisory Panel on Phthalates and Phthalate Alternatives. (2014). *Report to the U.S. Consumer Product Safety Commission by the Chronic Hazard Advisory Panel on phthalates and phthalate alternatives*. https://www.cpsc.gov/s3fs-public/CHAP-REPORT-With-Appendices.pdf

Cohen, S. M., Boobis, A. R., Dellarco, V. L., Doe, J. E., Fenner-Crisp, P. A., Moretto, A., Pastoor, T. P., Schoeny, R. S., Seed, J. G., & Wolf, D. C. (2019). Chemical carcinogenicity revisited 3: Risk assessment of carcinogenic potential based on the current state of knowledge of carcinogenesis in humans. *Regul Toxicol Pharmacol*, *103*, 100-105. https://doi.org/10.1016/j.yrtph.2019.01.017

Corton, J. C., Cunningham, M. L., Hummer, B. T., Lau, C., Meek, B., Peters, J. M., Popp, J. A., Rhomberg, L., Seed, J., & Klaunig, J. E. (2014). Mode of action framework analysis for receptor-mediated toxicity: The peroxisome proliferator-activated receptor alpha (PPAR α) as a case study. *Critical Reviews in Toxicology*, *44*(1), 1-49. https://doi.org/10.3109/10408444.2013.835784

Dellarco, V. L., McGregor, D., Berry, S. C., Cohen, S. M., & Boobis, A. R. (2006). Thiazopyr and thyroid disruption: Case study within the context of the 2006 IPCS Human Relevance Framework for analysis of a cancer mode of action. *Crit Rev Toxicol*, *36*(10), 793-801. https://doi.org/10.1080/10408440600975242

Doe, J. E., Boobis, A. R., Cohen, S. M., Dellarco, V. L., Fenner-Crisp, P. A., Moretto, A., Pastoor, T. P., Schoeny, R. S., Seed, J. G., & Wolf, D. C. (2022). A new approach to the classification of carcinogenicity. *Arch Toxicol*, *96*(9), 2419-2428. https://doi.org/10.1007/s00204-022-03324-z

EFSA Panel on Food Contact Materials, E., Aids, P., Silano, V., Barat Baviera, J. M., Bolognesi, C., Chesson, A., Cocconcelli, P. S., Crebelli, R., Gott, D. M., Grob, K., Lampi, E., Mortensen, A., Rivière, G., Steffensen, I.-L., Tlustos, C., Van Loveren, H., Vernis, L., Zorn, H., Cravedi, J.-P., . . . Castle, L. (2019). Update of the risk assessment of di-

butylphthalate (DBP), butyl-benzyl-phthalate (BBP), bis(2-ethylhexyl)phthalate (DEHP), di-isononylphthalate (DINP) and di-isodecylphthalate (DIDP) for use in food contact materials. *EFSA Journal*, *17*(12), e05838. https://doi.org/10.2903/j.efsa.2019.5838

Felter, S. P., Foreman, J. E., Boobis, A., Corton, J. C., Doi, A. M., Flowers, L., Goodman, J., Haber, L. T., Jacobs, A., Klaunig, J. E., Lynch, A. M., Moggs, J., & Pandiri, A. (2018). Human relevance of rodent liver tumors: Key insights from a Toxicology Forum workshop on nongenotoxic modes of action. *Regul Toxicol Pharmacol*, *92*, 1-7. https://doi.org/10.1016/j.yrtph.2017.11.003

Haber, L. T., Pecquet, A. M., Vincent, M. J., & White, L. M. (2022). The long goodbye: Finally moving on from the relative potency approach to a mixtures approach for polycyclic aromatic Hhdrocarbons (PAHs). *International Journal of Environmental Research and Public Health*, *19*(15), 9490. https://www.mdpi.com/1660-4601/19/15/9490

Health Canada. (2021). *Science approach document - Bioactivity exposure ratio: Application in priority setting and risk assessment*. https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/science-approach-document-bioactivity-exposure-ratio-application-priority-setting-risk-assessment.html

International Programme on Chemical Safety (IPCS). (2005). *Chemical-specific adjustment factors for interspecies differences and human variability : Guidance document for use of data in dose/concentration-response assessment*. https://www.who.int/publications/i/item/9241546786

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kroes, R., Galli, C., Munro, I., Schilter, B., Tran, L., Walker, R., & Würtzen, G. (2000). Threshold of toxicological concern for chemical substances present in the diet: A practical tool for assessing the need for toxicity testing. *Food Chem Toxicol*, *38*(2-3), 255-312. https://doi.org/10.1016/s0278-6915(99)00120-9

Kvasnicka, J., Aurisano, N., von Borries, K., Lu, E. H., Fantke, P., Jolliet, O., Wright, F. A., & Chiu, W. A. (2024). Two-stage machine learning-based approach to predict points of departure for human noncancer and developmental/reproductive effects. *Environ Sci Technol*, *58*(35), 15638-15649. https://doi.org/10.1021/acs.est.4c00172

MacGregor, J. T., Frötschl, R., White, P. A., Crump, K. S., Eastmond, D. A., Fukushima, S., Guérard, M., Hayashi, M., Soeteman-Hernández, L. G., Johnson, G. E., Kasamatsu, T., Levy, D. D., Morita, T., Müller, L., Schoeny, R., Schuler, M. J., & Thybaud, V. (2015). IWGT report on quantitative approaches to genotoxicity risk assessment II. Use of point-of-departure (PoD) metrics in defining acceptable exposure limits and

assessing human risk. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *783*, 66-78. https://doi.org/10.1016/j.mrgentox.2014.10.008

MacGregor, J. T., Frötschl, R., White, P. A., Crump, K. S., Eastmond, D. A., Fukushima, S., Guérard, M., Hayashi, M., Soeteman-Hernández, L. G., Kasamatsu, T., Levy, D. D., Morita, T., Müller, L., Schoeny, R., Schuler, M. J., Thybaud, V., & Johnson, G. E. (2015). IWGT report on quantitative approaches to genotoxicity risk assessment I. Methods and metrics for defining exposure–response relationships and points of departure (PoDs). *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *783*, 55-65. https://doi.org/10.1016/j.mrgentox.2014.09.011

Meek, M. E., Palermo, C. M., Bachman, A. N., North, C. M., & Jeffrey Lewis, R. (2014). Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *J Appl Toxicol*, *34*(6), 595-606. https://doi.org/10.1002/jat.2984

National Academies of Sciences, E., & Medicine. (2019). *A class approach to hazard assessment of organohalogen flame retardants*. The National Academies Press. https://doi.org/10.17226/25412

Patlewicz, G., Jeliazkova, N., Safford, R. J., Worth, A. P., & Aleksiev, B. (2008). An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res*, *19*(5-6), 495-524. https://doi.org/10.1080/10629360802083871

Pecquet, A., & Haber, L. T. (2024). Noncancer risk assessment: Principles and practice in environmental and occupational settings. In D. Paustenbach (Ed.), *Patty's Toxicology* (Seventh ed.). Wiley and Sons, Inc. https://www.wiley.com/en-us/Patty's+Toxicology%2C+6+Volume+Set%2C+7th+Edition-p-9781119634164

Savitz, D. A., & Hattersley, A. M. (2023). Evaluating chemical mixtures in epidemiological studies to inform regulatory decisions. *Environmental health perspectives*, *131*(4), 045001. https://doi.org/10.1289/EHP11899

Thienpont, A., Cho, E., Williams, A., Meier, M. J., Yauk, C. L., Rogiers, V., Vanhaecke, T., & Mertens, B. (2024). Unlocking the power of transcriptomic biomarkers in qualitative and quantitative genotoxicity assessment of chemicals. *Chem Res Toxicol*, *37*(3), 465-475. https://doi.org/10.1021/acs.chemrestox.3c00318

U.S. Environmental Protection Agency (EPA). (2024). *Conducting a human health risk assessment*. https://www.epa.gov/risk/conducting-human-health-risk-assessment

U.S. Environmental Protection Agency (U.S. EPA). (2002). *A review of the reference dose and reference concentration processes* (EPA/630/P-02/002F).

U.S. Environmental Protection Agency (U.S. EPA). (2005). *Guidelines for carcinogen risk assessment* (EPA/630/P-03/001F). https://www.epa.gov/risk/guidelines-carcinogen-risk-assessment

U.S. Environmental Protection Agency (U.S. EPA). (2014). *Guidance for applying quantitative data to develop data-derived extrapolation factors for interspecies and intraspecies extrapolation* (EPA/R-14/002F). https://www.epa.gov/risk/guidance-applying-quantitative-data-develop-data-derived-extrapolation-factors-interspecies

U.S. Environmental Protection Agency (U.S. EPA). (2023). *Draft proposed approach for cumulative risk assessment of high-priority phthalates and a manufacturer-requested phthalate under the Toxic Substances Control Act* (EPA-740-P-23-002). https://www.epa.gov/system/files/documents/2023-02/Draft%20Phthalate%20CRA%20Approach.pdf

Weiner, A. M. J., Irijalba, I., Gallego, M. P., Ibarburu, I., Sainz, L., Goñi-de-Cerio, F., Quevedo, C., & Muriana, A. (2024). Validation of a zebrafish developmental defects assay as a qualified alternative test for its regulatory use following the ICH S5(R3) guideline. *Reproductive Toxicology*, *123*, 108513. https://doi.org/10.1016/j.reprotox.2023.108513