Environmental Science & Technology

pubs.acs.org/est

Article

Quantifying Uncertainty in Ecotoxicological Risk Assessment: MUST, a Modular Uncertainty Scoring Tool

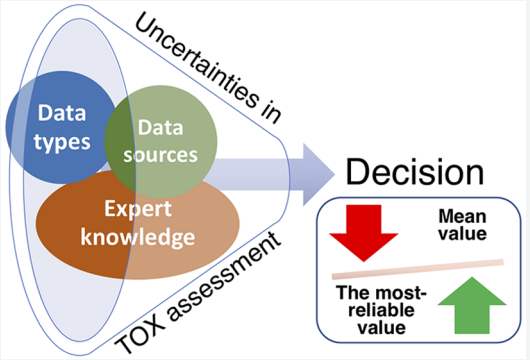
Jakub Kostal,* Hans Plugge, and Will Raderman

Cite This: *Environ. Sci. Technol.* 2020, 54, 12262−12270

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Whether conducting a risk, hazard, or alternatives assessment, one invariably struggles with the task of reconciling multiple available values of toxicological thresholds into a single outcome. When combining multiple pieces of evidence from many different sources, it is important to consider the role of data uncertainty. Uncertainty is inherent to all scientific data. However, in toxicological assessments, controversies and uncertainties are typically understated; they lack methodological transparency; or they poorly integrate qualitative and quantitative sources of information. Similarly, in model development, data curation is rarely performed with sufficient rigor, particularly when applying big data statistics. To overcome the hurdles of a decision process that must reconcile divergent data, we developed an uncertainty scoring tool that can be trained to reproduce specific decision-making paradigms and ensure consistency in the practitioner's judgment across complex scenarios. While designed to aid with ecotoxicological assessments and predictive model development, the tool's applicability extends to any decision-making process that calls for synthesis of incongruent data. Here, we highlight the development process, as well as demonstrate the method's utility in several prototypical ecotoxicological case studies.



INTRODUCTION

Over the years, considerable progress has been made toward developing hazard and risk assessment tools that better protect environmental health.[1,2] Some are (semi)quantitative; others use traffic-light-style scoring, that is red, yellow, and green. While some rely on highly curated data,[3] others use "big data" with little or no curation.[4] Despite methodological differences, these tools fall short of integrating uncertainty in a systematic and quantitative manner. Uncertainty is inherent to all scientific data, and its analysis is critical in human and environmental health assessments.[5] While some methods use the so-called uncertainty factors, these are not based on inherent data uncertainty of the model or test but on ballpark estimates of inherent extrapolation factors, for example, acute to chronic data.[6] Furthermore, there is little agreement on the magnitude of these factors and how to effectively combine them in a risk assessment.[7]

In view of the need to integrate diverse data types in 21st century risk and hazard assessments, multiple US EPA and National Research Council reports have called for greater understanding and transparency of uncertainty in ecotoxicological data, its sources and character.[8] Data are never as certain as a single value of toxicological threshold, such as lethal concentration at 50% mortality (LC_{50}), implies. Determination of an LC_{50} using a "standard" test has inherent uncertainties even if one adheres to the (legally proscribed) protocol. Inter- and intra-laboratory variables include age/size/source of organisms, water quality beyond toxicant, aeration and replacement protocols, and so forth, let alone measurement of actual toxicant concentrations.[9] Variability in test data for a single chemical can easily exceed 2 orders of magnitude, as shown in this report. While an in-depth evaluation of every single test could tease out these differences and how they affect outcomes and variability, such analyses often become impractical, multiyear efforts.

Beyond whole-animal models, additional layers of uncertainty must be considered for New Approach Methodologies, or NAMs, which are specifically designed for quick data generation and analyses.[10] Here, uncertainty arises both from assumptions and approximations of the applied method/theory and from model training and testing on the experimental data.[11] Though these uncertainties can be partly alleviated by either perfect knowledge of the mechanisms of chemical action, or by exhaustive testing of the chemical space of interest, neither can ever be completely achieved.[10,12] Thus, systematic, computer-aided quantification of uncertainty in NAMs data is especially important, given its implied additional

Received: April 9, 2020
Revised: July 27, 2020
Accepted: August 26, 2020
Published: August 26, 2020

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270

variability, the speed with which such data are produced, and the resulting quantity of such data.

Uncertainty embedded in toxicological data is carried into the uncertainty of the risk-based decision. Without a systematic approach to uncertainty analyses, risk assessments often fall back on expert opinions, which involve application of (biased) personal and professional judgments and assumptions.[13−15] To that end, it is not uncommon for different groups of experts to reach dissimilar conclusions based on the same data sets.[5] According to the 2009 NAS report, disagreements over quantity, quality, and source of scientific data were the most cogent reasons responsible for delays in regulatory responses.[16] Despite the absence of systematic consideration of uncertainty in risk-based decisions, there exist qualitative and quantitative approaches suitable for this purpose. In toxicology, they range from methods with narrow foci on variability in individual data quality to complex statistical frameworks that combine multiple lines of (diverse) evidence to inform a risk-based decision.[5,17−23] Complex statistical frameworks, such as those relying on Bayesian logic or Dempster−Shafer theory, have a proven track record in toxicology.[20,22,24] However, a simpler approach that captures magnitude and diversity of available data and easily integrates into existing workflows of the (non-statistician) risk assessor may be preferred in certain applications, for example, in fast processing of raw (big) data.

To that end, here we report the development and testing of a Modular Uncertainty Scoring Tool (MUST). MUST combines standard statistics and expert judgment of data quality in a transparent, easy-to-use tool that offers toxicologists a quantitative determination of uncertainty as a means to filter the available data. It can be used to select statistically representative value(s) for decision analysis, that is a risk, hazard, or alternatives assessment, as well as support predictive-model development, which is critical for filling data gaps.[1,12] Our main interest in developing this tool was to tackle the specific, yet increasingly more frequent case of multiple incongruent values originating from diverse data streams (*e.g.*, *in silico*, *in vitro*, or *in vivo* tests) and to understand how the distribution of these values and the practitioner's expert judgment impact uncertainty.

While MUST relies on user input to assess data quality, it allows the user to factor his or her confidence in assigned data quality into the overall uncertainty score. As a decision-support tool, MUST considers the value of expert judgment and the lack of consensus on the definitions of risk among assessors.[25,26] To that end, it is a modular tool that can be trained to reproduce specific decision-making paradigms and ensure consistency in the practitioner's judgment in complex scenarios. By allowing expert judgment affect computed uncertainties, MUST can be used to generate a range of scientifically plausible outcomes by having multiple users analyze the same data set. The combination of fast data-processing owing to standard statistics and the incorporation of expert knowledge outline the unique value proposition of MUST.

METHODS

Broadly, uncertainty reflects variability in data, commonly represented by interquartile range, variance, or standard deviation, and the precision with which such data is measured. The underlying relationship in MUST builds on these basic statistical parameters. For a given data set, MUST computes uncertainty scores (US_x) based on variance of data weighted by its quality (eq 1)

$$US_{x_i} = \left(\frac{\sigma_g}{E \cdot R_{f_i}}\right) + y\frac{\left|x_i - \frac{\sum_i^N R_{f_i} \cdot x_i}{\sum_i^N R_{f_i}}\right|}{E} \quad (1)$$

From eq 1, x_i is the toxicity threshold value (*e.g.*, LC_{50}); E is the experimental, that is random, error associated with the test type; R_f is a data-reliability factor, which reflects perceived data quality associated with a given study; and σ_g is the sample standard deviation, that is $\sigma_g = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$, describing variability of the response, where $\bar{x}$ is the sample mean, and N is the sample size. The above relationship assumes that all available data, which can be considered equivalent in the analysis (*e.g.*, repeated determinations of LC_{50} values for acute aquatic toxicity for chemical X), form a random sample within a larger (unknown) population of (non-normally distributed) values. The extent of equivalency within any group of values is determined by the user, who may include or exclude data points on the basis of relevance, such as due to differences in testing/modeling protocols or target species. The data-reliability factor, R_f, is assigned by the user and translated into indices ranging from 0 (poor quality) to 1 (high quality). To estimate R_f, the user should ideally assess relevant criteria in the test or model (*e.g.*, per Organisation for Economic Co-operation and Development/good laboratory practice guidelines) and consider discarding any unverifiable data. Existing reliability metrics, such as Klimisch scores,[17] which are frequently reported for registered substances under ECHA/REACH, can be used to generate R_f values. In the absence of such metrics, data reliability can be quantified using the many available methods, for example, the ToxRTool, TRAM (Toxicological data Reliability Assessment Method), or fuzzy expert systems based on the Klimisch scoring approach.[27−29] For predicted data, both data quality of the training set, that is the underlying experimental data on which the model was built, and performance metrics of the statistical model itself,[30] should factor into R_f. In practice, a less rigorous and less time-consuming estimation of data reliability can be applied, and the user may simply choose to score analyzed values based on his or her expert judgment. This is made largely possible by incorporating a "scaling factor," y, into eq 1. The scaling factor's mathematical function is outlined below, while its practical utility is demonstrated in the Results and Discussion sections. Briefly, the scaling factor reflects the user's own uncertainty in the assignment of reliability scores; its magnitude drives the preference for either an average or the most-reliable value in the data set. To that end, the scaling factor is a variable that can be uniquely modified to suit a particular decision-making paradigm, as discussed in the following sections.

As formulated in eq 1, MUST comprises two terms, $\left(\frac{\sigma_g}{E \cdot R_{f_i}}\right)$ and $\frac{\left|x_i - \frac{\sum_i^N R_{f_i} \cdot x_i}{\sum_i^N R_{f_i}}\right|}{E}$, which are summed to determine the final uncertainty score, where the latter term is weighted by the scaling factor, y. The first term reflects the overall quality of the group of equivalent values; the assumption here is that general concerns regarding a particular test method or a chemical should impact the confidence in any individual value, measured

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270

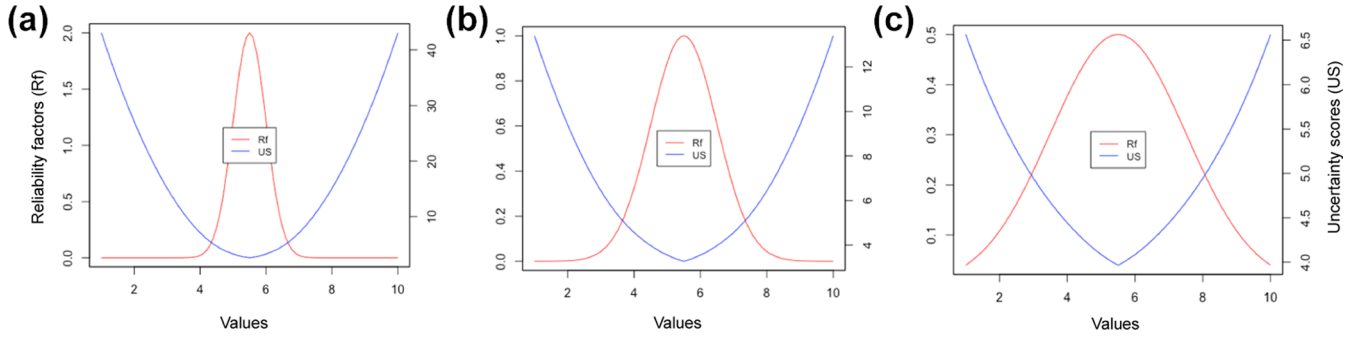


Figure 1. Effect of increasing and decreasing the standard deviation of normally distributed reliability factors, R_f, on computed uncertainty scores, US. Left (a): $\overline{x}$ = 5.5, σ_g = 0.5; center (b): $\overline{x}$ = 5.5, σ_g = 1; right (c): $\overline{x}$ = 5.5, σ_g = 2; N = 100, y = 1, and E = 0.1. Note: reliability scores distributed over (arbitrary) values 1−10 exceed the 0−1 range in (a) in order to conserve the total area under the curve.

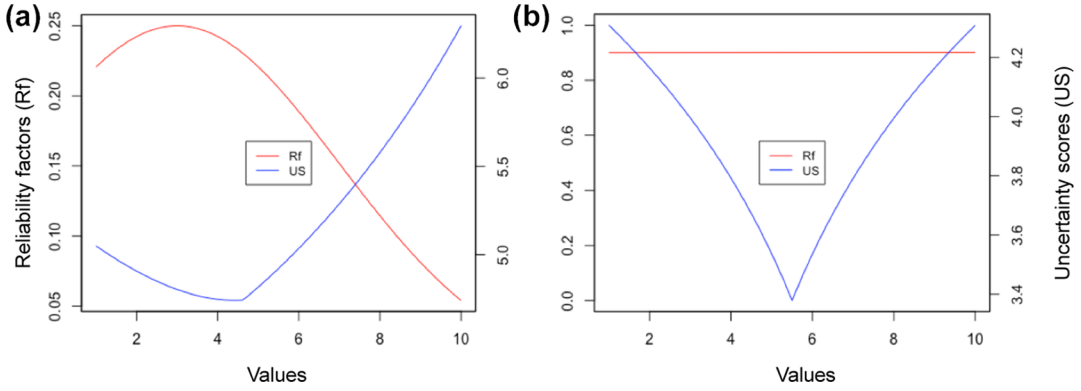


Figure 2. Left (a): Skewed distribution of reliability factors, R_f, across arbitrary values 1−10 ($\overline{x}$ = 3, σ_g = 4). Right (b): Near-flat (linear) distribution of reliability factors, R_f, across arbitrary values 1−10; N = 100, E = 0.1, and y = 1.

or predicted. The second term is the normalized deviation of any given value from the reliability-weighted average.

In the case studies described here, the toxicity values are log-transformed, where the log base corresponds to the spread of the toxicity categories for a given endpoint. For example, given thresholds for acute aquatic toxicity set by the US EPA Safer Choice Program,[31] a log base of 10 would be used to transform analyzed values. Unless noted otherwise, computed uncertainty scores using eq 1 are also log-transformed, that is $US_x = \log_{10} US_x$. Through the log-transformation, we account for non-parametric statistics; it is expected that data are log-normally distributed.[32] All statistical analyses in this report were performed using the R Program.[33] Equation 1 and its parametrization procedure were coded into a user-friendly interface with the Perl programming language.

RESULTS

Internal Validation. Validation of theoretical models is essential to ensure satisfactory behavior across anticipated user scenarios. Equation 1 was developed based on inductive reasoning and intuition, recognizing the key factors that drive uncertainty in a given value within a family of equivalent observables. The specific definition was based on axioms and logical relations developed to extract reasonable propositions and predictions. To broadly test these propositions and predictions was the basis for validating the model's underlying hypothesis.

We tested the behavior of eq 1 across different distributions of value reliabilities, starting with a normal distribution (Figure 1b). For the purpose of these exercises, both terms in eq 1 were weighted equally, that is y = 1; 100 equally spaced, discrete points between 1 and 10 were used on the x-axis to generate a smooth-fit curve; and a random error, E = 0.1, was applied. The selection of values, which may represent raw or log-transformed toxicity thresholds, and their error is irrelevant here as the following analysis aims at deriving qualitative trends in the behavior of eq 1. As expected, we observed that when reliabilities are distributed normally around a mean value, the corresponding distribution of computed uncertainty scores is parabolic with the lowest uncertainty score being assigned to the value with the highest reliability (Figure 1).

The trend noted in proceeding from Figure 1a−c, that is in broadening the normal distribution, is that of increasing uncertainty assigned to the most-reliable data point. Concurrently, when the distribution is contracted (Figure 1a) or expanded (Figure 1c) by decreasing or increasing the standard deviation of reliability factors, outliers are penalized by computed uncertainties to a greater or lesser extent, respectively. This behavior can be observed on the second y-axis as the range of computed uncertainties increases (Figure 1a) and decreases (Figure 1c). This effect is expected as one's confidence in a given value should be affected by the distribution of reliable outcomes across the population. For example, if a tight cluster of highly reliable values exists, one's relative confidence (which is defined by the range of the computed uncertainties) in selecting the most-reliable value should increase, while on the other end of the spectrum, confidence in the least-reliable data should proportionally decrease.

Given that magnitude of a value and its reliability are fundamentally unrelated, it is interesting to consider the

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270

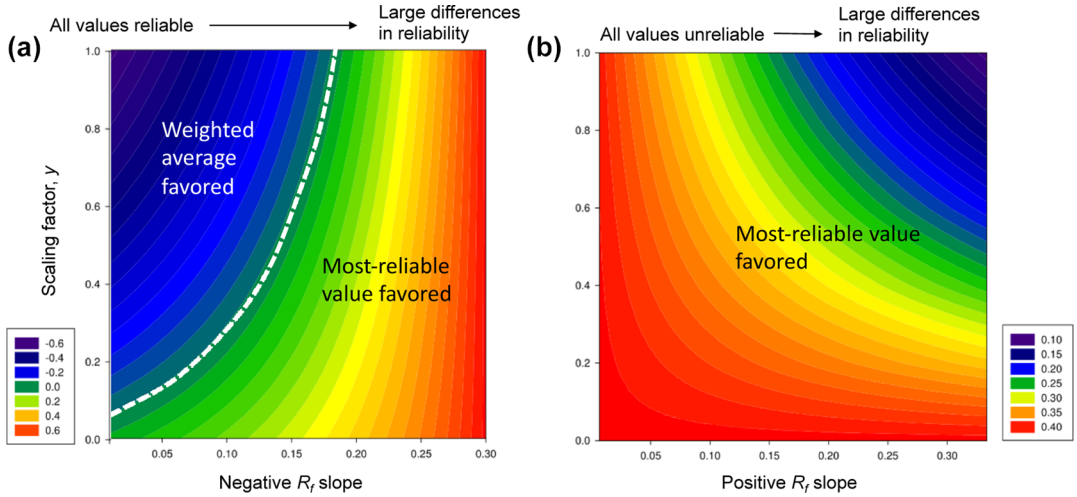


Figure 3. Preference for average (negative values) versus most-reliable data (positive values) in 3-point linear distributions as a function of the scaling factor (y-axis) and value reliabilities (reported as a slope on x-axis); N = 3: 1, 2, and 3. Left (a): All data start with equal and high reliabilities. Right (b): All data start with equal and low reliabilities. Dotted white line is used to separate regions where the average (*i.e.* middle) value is favored from regions where the most-reliable value is favored [absent in (b)]. Linear models: $R_{f_G} = \left\{(-1)\left(\frac{0.02}{3} + m_G\right)x_G + 0.9\right\}$ (left) and $R_{f_G} = \left(\frac{0.02}{3} + m_G\right)x_G$ (right), where $m_G = \left(\frac{0.02}{3} + m_{G-1}\right)$, starting from 0.007. E = 0.5.

behavior of eq 1 with skewed (asymmetric) distributions. This is especially relevant here as most toxicological data are not normally distributed. An example is provided in Figure 2a, showing a distribution similar to those in Figure 1 but with a sample mean of 3 and a standard deviation of 4.

In this broad distribution of relatively low reliabilities, the most-reliable data point is no longer assigned the lowest uncertainty score; the uncertainty minimum is skewed toward the average value. This change of preference is due to the second term in eq 1. It is reasonable to propose that when faced with values of similar reliability, one may lean toward selecting an average value to rely on (*vs* selecting the most-reliable data point when the relative differences in assigned reliabilities are large). Figure 2b demonstrates an extreme case of a near-flat distribution of reliabilities, which indicates a strong preference for an average value. In eq 1, the relative weight of the two terms, that is the magnitude of the scaling factor, y, can be altered to influence the relative preference for the distribution average versus the most-reliable value in the data set. A more detailed analysis of the scaling factor's role is presented in subsequent sections.

Role of the Scaling Factor. Our interpretation of the scaling factor is that it represents the uncertainty in the assignment of the reliability factors, and ultimately, the uncertainty in the computation of uncertainty scores. This manifests in analyzed data as changes of the relative preference for an average value versus the most-reliable value in the data set.

We investigated the role of the scaling factor that would apply across various toxicity value distributions by considering behavior on simple linear models of variable slopes, which can be used to reconstruct any distribution. To analyze this visually, we plotted the preference for the most-reliable *versus* average value (computed as the difference between the two corresponding uncertainty scores, $\Delta US_x = US_{\text{average } x} - US_{\text{most-reliable } x}$) as a function of both the scaling factor and the gradient of the linear distribution of reliability factors. Three data points, equally spaced in terms of magnitude of response with standard deviation of 1, were used for this analysis. Figure 3 outlines two prototypical cases. In Figure 3a, all values start with the highest reliability (R_f = 1) and the function gradually decreases (as shown on the x-axis); in Figure 3b, all values start with near-zero reliability and the function gradually increases. As the gradient increases (moving left to right on the x-axis in both plots), differences in assigned reliability factors increase. The regions where either the average or the most-reliable value are favored by computed uncertainty scores are denoted as such in the plots.

We offer the following interpretation of Figure 3a,b. With highly reliable data, eq 1 is more sensitive to the average of that data over small differences in assigned reliabilities. However, this preference for an average value can be "tuned out" by decreasing the scaling factor, y. Conversely, if we assign all values comparable but low reliability, lowest uncertainty favors the most-reliable value; this is true unless values get very close in which case the average value is favored, *viz.* Figure S1. In other words, there is greater sensitivity for smaller differences in reliabilities when all values are generally unreliable than when they are very reliable. It should be noted that the gradient changes in figures on the left and right are subtle and governed by the linear models $R_{f_G} = \left\{(-1)\left(\frac{0.02}{3} + m_G\right)x_G + 0.9\right\}$ and $R_{f_G} = \left(\frac{0.02}{3} + m_G\right)x_G$, respectively, with $m_G = \left(\frac{0.02}{3} + m_{G-1}\right)$. These models were selected to cover the effective (0−1) range of reliability factors, R_f, starting with m_G = 0.007. Thus, a slight increase in preference for the average, that is middle, value shown in Figure 3b can be attributed to the gradual rise in reliabilities in all three values as the gradient increases.

The above analysis can be readily extended to larger data sets with one important caveat: as the number of values increases, so does the relative preference for an average versus the most-reliable data point. We briefly illustrate this in Figure 4 by conducting the same analysis as in Figure 3b but for five discrete values instead of three. Thus, $R_{f_G} = \left(\frac{0.02}{5} + m_G\right)x_G$

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270

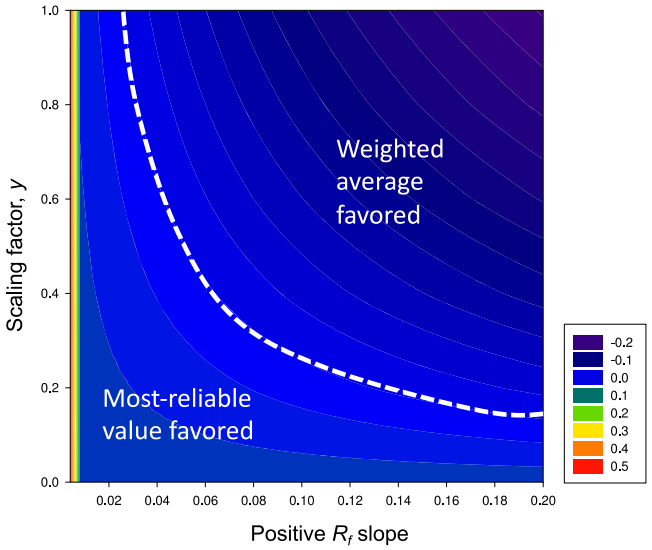


Figure 4. Preference for average (negative values) *vs* most-reliable data (positive values) in a 5-point linear rise distribution as a function of scaling factor (y-axis) and value reliabilities (reported as a slope on x-axis). Dashed white line represents the boundary. Linear model: N = 5: 1, 2, 3, 4, 5; $R_{f_G} = \left(\frac{0.02}{5} + m_G\right)x_G$ where $m_G = \left(\frac{0.02}{5} + m_{G-1}\right)$, starting with m_G = 0.004; E = 0.5.

where $m_G = \left(\frac{0.02}{5} + m_{G-1}\right)$, starting with m_G = 0.004. In this case, the average (middle) value is favored over the most-reliable value at reliability slopes greater than 0.028 for all values of the scaling factor, y. This trend is expected as the relevance of an average value increases with the sample size.

For the sake of completeness, it should be noted that because of the absolute-difference expression in the second term of eq 1, there is no change in computed uncertainties when the increase/decrease in the linear function is toward a value of greater or smaller magnitude. In other words, extending the x-axis beyond the intersect with y-axis in Figures 3 and 4 yields symmetrical outcomes. The relationship between the scaling factor, assigned reliability, and the computed uncertainty score suggests that when one has high confidence in his or her assigned reliabilities, the scaling factor can be decreased or completely turned off (*i.e.*, y = 0) to always favor the most-reliable outcome. Conversely, high uncertainty in assigned reliabilities may be best offset by a larger y that skews lowest uncertainty scores toward the average value.

Experimental Error. By placing the experimental error (E) variable into the denominator of eq 1, the larger the assigned experimental error is for a given data set, the smaller the computed uncertainty scores are as values become harder to distinguish from each other. In our implementation, E does not affect the relative trend of computed uncertainty scores for any given data set, for example, by switching preference from the highest-reliability value to the average when values effectively become identical within the experimental error. However, absolute uncertainty scores decrease as E increases and vice versa, consistent with the notion that our confidence in selecting a statistically representative value increases when values are similar to each other.

Aquatic Toxicity Case Studies. To demonstrate the utility of our scoring approach on real-world cases, we present analysis of the acute fish toxicity of two compounds, nickel sulfide (CAS 16812-54-7), a metal salt, and ethylbenzene (CAS 100-41-4), an organic compound. Two sets of 123 and 60 independent LC_{50} acute aquatic toxicity values were compiled for these chemicals from the ECHA REACH database (Tables S1−S3). Unless otherwise noted, Klimisch scores of 1, 2, 3, and 4 were normalized to reliability factors, R_f, of 1, 0.7, 0.3, and 0.1 in eq 1, respectively. This assignment reflects the general notion that values with Klimisch scores of 1 and 2 are largely dependable, while those with scores of 3 and 4 are not.[17] In both case studies, we examined the effect of varying the magnitude of the scaling factor on computed uncertainties and, consequently, the selection of a representative LC_{50} value.

Acute Aquatic Toxicity of Nickel Sulfide. For nickel sulfide (CAS 16812-54-7), reported toxicity thresholds (LC_{50}) span all Klimisch scores of data reliability (1−4) and all categories of concern, as defined by the US EPA (Figure 5),

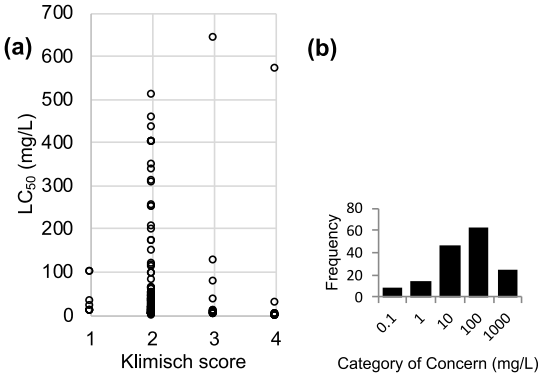


Figure 5. Left (a): Distribution of LC_{50} values for nickel sulfide (CAS 16812-54-7) as a function of corresponding Klimisch scores. Values were extracted from the ECHA database. Right (b): Histogram of LC_{50} values based on US EPA's categories of concern for acute aquatic toxicity, ≤0.1, >1 and ≤10, >10 and <100, and ≥100 mg/L.[31]

with the lowest and highest LC_{50} values of 4.2×10^{-3} and 509 mg/L, respectively. This data set illustrates the limitations of traditional decision-making approaches in toxicological assessments. Computed geometric mean, 13.7 mg/L, is close to the cutoff between two regulatory categories, and if relying on the most conservative value, one ignores the rest of the data while using a value, LC_{50} = 4.2×10^{-3} mg/L, that is not the most reliable. Importantly, unreliable values represent less than 5% of all data; thus, curation based on data quality is not particularly helpful.

Our full uncertainty analysis is shown in Table S1; qualitative trends in computed scores and the effect of the scaling factor, y, can be gauged from Figure 6. For small values of y, LC_{50} values with low Klimisch scores correspond to low uncertainty scores and vice versa. However, variability in computed uncertainty increases as the scaling factor, that is the impact of the reliability-weighted mean, increases. Thus, for larger y's, the correspondence between Klimisch and uncertainty scores breaks down as values further from the reliability-weighted mean are penalized more by greater uncertainty.

With the default scaling factor equal to 1, the lowest uncertainty score, 0.69, is assigned to LC_{50} = 10.9 mg/L, which has a Klimisch score of 1 (*viz.* point #107 in Table S1). This value is close to the computed geometric mean of the distribution. In this particular data set, decreasing the scaling factor (from 1 to 0.2) does not change the uncertainty scoring

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270

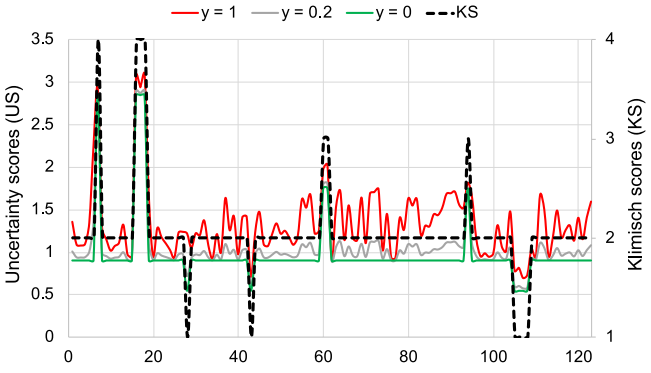


Figure 6. Computed uncertainty scores (left *y*-axis) and assigned Klimisch scores (right *y*-axis) for 123 acute aquatic toxicity values (LC_{50}'s) corresponding to nickel sulfide (CAS 16812-54-7). *X*-axis represents the test count (1−123). Uncertainty analysis was carried out using scaling factors y = 1, 0.2, and 0.

with regard to the selection of the least-uncertain value. We believe this is due to the (fortuitous) proximity of the most-reliable values to the geometric mean, which is 11.1 mg/L for only the most-reliable LC_{50}'s (Klimisch scores of 1). Zeroing the scaling factor cancels the second term in eq 1, leading to equal uncertainties for data points with the same Klimisch score, that is all values with Klimisch score of 1 are assigned the same lowest uncertainty score (0.55).

It should be noted that while the least-uncertain value remains the same for y between 0 and 1 (LC_{50} = 10.9 mg/L), the ranking of all values in the set by their computed uncertainty scores changes. For example, in selecting the top 10 values with the lowest scores (Table S1, shaded cells), data point #25 (LC_{50} = 13.6 mg/L, Klimisch score = 2) is substituted for data point #28 (LC_{50} = 100 mg/L, Klimisch score = 1) when decreasing the scaling factor from 1 to 0.2. This finding is consistent with our previous analysis, showing preference for the most-reliable (*vs* average) values when the scaling factor is lowered (*viz.* discussion of Figures 3 and 4).

While we proposed a default translation of the Klimisch scores into reliability factors, the latter are considered to be user-defined variables in eq 1 (just as they would have to be in the absence of a formalized scoring system). These factors can be adjusted based on user's expert judgment, reflecting professional experience and the context of the study. To that end, we briefly considered how the translation of Klimisch scores impacts the final outcome of computed uncertainties in this case study. Translating Klimisch scores to R_f values more evenly, that is assigning reliability factors of 1, 0.9, 0.6, and 0.3 to Klimisch scores of 1, 2, 3, and 4, respectively, prompts a selection of LC_{50} = 15.3 mg/L (Klimisch score of 2, point #62 in Table S1) as the least-uncertain data point with a score of 0.67. This result is consistent with greater relative reliability assigned to values with lower Klimisch scores in this alternate scheme. When y is reduced to 0.2, LC_{50} = 10.9 mg/L (Klimisch score of 1, point #107) is assigned the lowest uncertainty score (0.58). Thus, reducing the impact of the second term of eq 1 leads to concordance between the two translation schemes.

Last, we examined how a reduced set of values for nickel sulfide, which only includes Rainbow trout (*Oncorhynchus mykiss*) test results, would fare compared to the full set. Such initial curation is not unreasonable given that Rainbow trout was shown to exhibit greater general sensitivity to toxic substances, particularly metals, than the other species included in this data set.[34] While the geometric means for the full and subset are relatively close, 13.7 and 17.7 mg/L, the computed lowest uncertainty scores (using the default scaling factor of 1) show a greater difference: LC_{50} = 10.9 mg/L *versus* LC_{50} = 21.2 mg/L (Table S2). Additionally, because the reduced set only contains values with Klimisch scores of 1 and 2, the lowest-computed uncertainty score (0.404) is below that of the full set (0.694), indicating greater relative confidence in the assessment.

Acute Aquatic Toxicity of Ethylbenzene. In considering ethylbenzene (CAS 100-41-4), we noted a considerable impact of the scaling factor on computed uncertainties (Table S3). In this smaller data set, values with Klimisch score 1 (LC_{50} = 2.4, 5.1, 5.8, and 7.0 mg/L) are well below, as well as in different category of concern,[31] than the geometric mean, 57.2 mg/L. Consequently, equal weighting of both terms in eq 1 leads to assigning the lowest-computed uncertainty score (0.72) to a value much smaller than the geometric mean, LC_{50} = 15 mg/L (Klimisch score 2, point #32 in Table S3). Decreasing the scaling factor to 0.5 changes this selection to LC_{50} = 7.0 mg/L (Klimisch score 1, point #22), which is consistent with our previous explanations of the role the scaling factor plays in the selection of average *versus* the most-reliable data. It is interesting to note that in this case (*vs* the previous case study), our selected value of LC_{50} = 7.0 mg/L would result in the same decision outcome as selecting the most conservative value in the data set, LC_{50} = 2.4 mg/L, based on US EPA's guidelines.[31] In contrast, taking an average (105.8 mg/L), or better a geometric mean (57.2 mg/L), of this data set would lead to a very different regulatory outcome.

DISCUSSION

MUST is a robust tool for ranking ecotoxicological data by computed uncertainty scores, which then can be compared within and across different data sets to inform decisions. In practice, the user can use MUST to select either a single-value or subset of values. We view the latter as being beneficial when dealing with mixed data streams, for example, combining *in silico*, *in vitro*, and *in vivo* data, or different test species, experimental conditions, etc. The user may want to include all available data (with appropriately assigned reliabilities) in the uncertainty analysis; however, he or she may prefer to subsequently "hand-pick" a certain data type from a reliable subset to support a decision. We should emphasize that there is a fundamental difference between a curation that eliminates certain data prior to the uncertainty analysis and post-analysis curation. This can be readily demonstrated on the nickel sulfide case study using the test species as the curation criterion. Considering the entire data set, reliable subset of top 10 LC_{50} measurements with the lowest uncertainty scores consisted of eight test results for rainbow trout (*O. mykiss*), one result for bluegill (*Lepomis macrochirus*), and one result for white perch (*Morone Americana*) (Table S1). Relying only on the Rainbow trout test results, given their greater sensitivity for metal salts, the user would subsequently select LC_{50} = 10.9 mg/L, which corresponds to the lowest uncertainty score (0.69, y = 1). If, however, curation was carried out prior to the uncertainty analysis, the selected value would be LC_{50} = 21.2 mg/L (Table S2). The two values are only *ca.* 10 mg/L apart; however, because the first one is close to the L/E/IC_{50} cutoff of 10 mg/L provided by the US EPA's Safer Choice program,[31] it might lead to a different regulatory decision. Overall, we

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270

recommend that any manual curation, beyond issues with erroneous data reporting and data interdependence, is applied after MUST assessment. MUST incorporates both data reliability and user's confidence in assigned reliability factors in its algorithm; thus, it is a "safer" choice to give less weight to less reliable data than to ignore it completely, especially in cases where low-quality data constitute majority of the available data and/or points to greater risk/hazard.

The ability to generate a subset of dependable values can also facilitate development of more accurate and robust predictive models by providing reliable training and test sets. While a plethora of databases exist online to aid in predictive-model development,[35] the quality of experimental results and quality of its online storage and reporting (*e.g.*, accuracy, completeness, and integrity) are of increasing concern,[36] particularly for statistical, big-data models, for which detailed data curation is impractical.[37] We should note that the extent of data set curation using MUST depends on the type of a predictive model: for a highly mechanistic model, a smaller subset of data with very low uncertainty scores may be preferred, while for statistical models that require larger training sets, eliminating only data with the highest uncertainty might be useful.[12]

Parametrization. The unique value proposition of eq 1 is in skewing the preference, that is the assignment of the lowest uncertainty score, toward either an average value or the most-reliable value in the data set, and that this "balance" can be altered by the scaling factor. While the default equal weighting of both terms in eq 1 generates reasonable propositions and predictions, it is nonetheless important to recognize the breath of standard practices and preferences involved in risk, hazard, and alternatives assessments. To that end, we implemented eq 1 into a computer algorithm using the Perl programming language, which can be used to input data and generate corresponding uncertainty scores. We also incorporated the ability to train and parametrize eq 1 via the scaling factor in order to reproduce specific decision-making paradigms. To facilitate this process, we developed a questionnaire based on 27 case studies that combinatorially explore unique assignments of high, medium, and low reliability to values of high, medium, and low toxicological concern (Table S4). We further added 12 cases where reliability is scored in the range of 1−10, offering insights into the user's sensitivity to relative reliabilities. While we do not specify data type(s), or what an assigned reliability value or a category means, we expect both would be viewed and characterized through the lens of the end-user's professional setting to facilitate effective customization.

From Table S4, using expert knowledge, the user ranks toxicity values for every case from the most-likely to the least-likely to inform his or her decision. Corresponding scaling factor values between 0 and 1 that satisfy the user's selection based on eq 1 are calculated. MUST carries out this analysis by iteratively examining scaling factors that reproduce the user's rank in terms of relative uncertainty scores and then determines the final scaling factor to be the value that is most-frequently featured across all 39 case studies. If there is more than one winner, MUST reports the relevant range of scaling factors. A % concordance metric is provided to the user that reflects goodness of fit. In our testing of MUST, pilot users reported an improved decision logic based on the questionnaire and subsequent MUST parametrization.

Applicability. In this study, we showcased MUST's utility on two ecotoxicological data sets, which are sufficiently large to benefit from our approach. In principle, MUST's applicability extends beyond ecotoxicology, owing to its general definition of uncertainty in terms of reliability-biased data variability. It should be noted that the practical usefulness of MUST diminishes for very small data sets. Limited toxicological data is a frequent problem, particularly for tests of mammalian (chronic) toxic endpoints, which are economically and ethically expensive. Thus, the end-user may encounter chemical-endpoint combinations where data are scarce and the use of MUST is impractical even when available *in vitro* or *in silico* data are included in analysis. In those cases, expert review based on data reliability or strategies using Bayesian logic or Dempster−Shafer theory, which were referenced previously, may be more suitable.

In the present analysis, we assumed data were independent within each data set. In the two case studies, this assumption was based on our initial curation of the ECHA/REACH data to ensure independent test results for a single endpoint (LC_{50}). When aggregating different data types (*e.g.*, *in vivo*, *in vitro*, and *in silico*), the user should be cognizant of data interdependence. For example, because *in silico* models are trained on experimental data, unwanted bias toward a particular value (or set of values) may arise. In those cases, we recommend closer inspection of data to remove any values that overlap between models and may be part of a training set, prior to MUST analysis.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.0c02224.

> Study of uncertainties as a function of linear rise in relative reliabilities across three data points; list of acute aquatic toxicity (LC_{50}) values along with Klimisch scores and computed uncertainties for nickel sulfide and ethylbenzene; and a questionnaire for MUST parametrization (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jakub Kostal − *Department of Chemistry, George Washington University, Washington, District of Columbia 20052, United States;* orcid.org/0000-0001-9727-0477; Phone: (202) 994-7320; Email: jkostal@gwu.edu

Authors

Hans Plugge − *Safer Chemical Analytics, Verisk 3E, Bethesda, Maryland 20814, United States*

Will Raderman − *Department of Chemistry, George Washington University, Washington, District of Columbia 20052, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.0c02224

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by The George Washington University and was instrumental to projects funded by the National Science Foundation (NSF1805080 and NSF1943127).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Drs. Steve DeVito (US EPA), Emma Lavoie (US EPA), Lauren Heine (Northwest Green Chemistry), Lauren Brown (Abt Associates), Nihar Das (Verisk 3E), and Adnan Korkutovic (Verisk 3E) for their contributions.

■ ABBREVIATIONS

AOP	adverse outcomes pathway
REACH	registration, evaluation, authorization and restriction of chemicals
TSCA	toxic substances control act
US EPA	United States Environmental Protection Agency
NAS	National Academy of Sciences
MUST	modular uncertainty scoring tool
ECHA	European Chemicals Agency
OECD	Organisation for Economic Co-operation and Development
GLP	good laboratory practice
CAS	chemical abstracts service

■ REFERENCES

(1) Maertens, A.; Plugge, H. Better Metrics for "Sustainable Design": Toward an in Silico Green Toxicology for Green(er) Chemistry. *ACS Sustainable Chem. Eng.* **2018**, *6*, 1999−2003.

(2) Panko, J. M.; Hitchcock, K.; Fung, M.; Spencer, P.; Kingsbury, T.; Mason, A. A comparative evaluation of five hazard screening tools. *Integr. Environ. Assess. Manage.* **2017**, *13*, 139−154.

(3) Melnikov, F.; Kostal, J.; Voutchkova-Kostal, A.; Zimmerman, J. B.; Anastas, P. Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals. *Green Chem.* **2016**, *18*, 4432−4445.

(4) Luechtefeld, T.; Maertens, A.; Russo, D. P.; Rovida, C.; Zhu, H.; Hartung, T. Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008−2014. *ALTEX* **2016**, *33*, 95−109.

(5) Beck, N. B.; Becker, R. A.; Erraguntla, N.; Farland, W. H.; Grant, R. L.; Gray, G.; Kirman, C.; LaKind, J. S.; Jeffrey Lewis, R.; Nance, P.; Pottenger, L. H.; Santos, S. L.; Shirley, S.; Simon, T.; Dourson, M. L. Approaches for describing and communicating overall uncertainty in toxicity characterizations: US Environmental Protection Agency's Integrated Risk Information System (IRIS) as a case study. *Environ. Int.* **2016**, *89*−*90*, 110−128.

(6) Stedeford, T.; Hsu, C.; Zhao, Q. J.; Dourson, M. L.; Banasik, M. The Application of Non-Default Uncertainty Factors in the U.S. EPA's Integrated Risk Information System (IRIS). Part I: UF_L, UF_s and "Other Uncertainty Factors". *J. Environ. Sci. Health, Part C: Environ. Carcinog. Ecotoxicol. Rev.* **2007**, *25*, 245−279.

(7) Escher, S. E.; Mangelsdorf, I.; Hoffmann-Doerr, S.; Partosch, F.; Karwath, A.; Schroeder, K.; Zapf, A.; Batke, M. Time extrapolation in regulatory risk assessment: The impact of study differences on the extrapolation factors. *Regul. Toxicol. Pharmacol.* **2020**, *112*, 104584.

(8) NAS, Committee on Decision Making Under Uncertainty; Board on Population Health and Public Health Practice; Institute of Medicine. *Environmental Decisions in the Face of Uncertainty*; National Academies Press (US): Washington (DC), 2013; available from: https://www.ncbi.nlm.nih.gov/books/NBK200848/.

(9) Degraeve, G. M.; Cooney, J. D.; McIntyre, D. O.; Pollock, T. L.; Reichenbach, N. G.; Dean, J. H.; Marcus, M. D. Variability in the performance of the seven-day fathead minnow (pimephales promelas) larval survival and growth test: An intra- and interlaboratory study. *Environ. Toxicol. Chem.* **1991**, *10*, 1189−1203.

(10) Parish, S. T.; Aschner, M.; Casey, W.; Corvaro, M.; Embry, M. R.; Fitzpatrick, S.; Kidd, D.; Kleinstreuer, N. C.; Lima, B. S.; Settivari, R. S.; Wolf, D. C.; Yamazaki, D.; Boobis, A. An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. *Regul. Toxicol. Pharmacol.* **2020**, *112*, 104592.

(11) Cronin, M. T. D.; Madden, J. C.; Yang, C.; Worth, A. P. Unlocking the potential of in silico chemical safety assessment - A report on a cross-sector symposium on current opportunities and future challenges. *Comput. Toxicol.* **2019**, *10*, 38−43.

(12) Kostal, J.; Voutchkova-Kostal, A. Going All In: A Strategic Investment in In Silico Toxicology. *Chem. Res. Toxicol.* **2020**, *33*, 880−888.

(13) Kasperson, R. E.; Renn, O.; Slovic, P.; Brown, H. S.; Emel, J.; Goble, R.; Kasperson, J. X.; Ratick, S. The Social Amplification of Risk - a Conceptual-Framework. *Risk Anal.* **1988**, *8*, 177−187.

(14) Hoberg, G. Risk, Science and Politics - Alachlor Regulation in Canada and the United-States. *Can. J. Polit. Sci.* **1990**, *23*, 257−277.

(15) Slovic, P. Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield (Reprinted from Environment, ethics, and behavior, pg 277-313, 1997). *Risk Anal.* **1999**, *19*, 689−701.

(16) NAS. *Science and Decisions Advancing Risk Assessment. Committee on Improving Risk Analysis Approaches Used by the USEPA Board on Environmental Studies and Toxicology*; U.S. National Academy of Sciences, National Academies Press: Washington, DC, 2009.

(17) Klimisch, H.-J.; Andreae, M.; Tillmann, U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmacol.* **1997**, *25*, 1−5.

(18) Pollard, S. J. T.; Davies, G. J.; Coley, F.; Lemon, M. Better environmental decision making - Recent progress and future trends. *Sci. Total Environ.* **2008**, *400*, 20−31.

(19) Lepper, P. *Towards the Derivation of Quality Standards for Priority Substances in the Context of the Water Framework Directive*. Contract No. B4-30401/20001111/30637/MA/E1; Schmallenberg: Germany, 2002.

(20) Park, S. J.; Ogunseitan, O. A.; Lejano, R. P. Dempster-Shafer Theory Applied to Regulatory Decision Process for Selecting Safer Alternatives to Toxic Chemicals in Consumer Products. *Integr. Environ. Assess.* **2014**, *10*, 12−21.

(21) Martin, P.; Bladier, C.; Meek, B.; Bruyere, O.; Feinblatt, E.; Touvier, M.; Watier, L.; Makowski, D. Weight of Evidence for Hazard Identification: A Critical Review of the Literature. *Environ. Health Perspect.* **2018**, *126*, 076001.

(22) Rathman, J. F.; Yang, C.; Zhou, H. Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment. *Comput. Toxicol.* **2018**, *6*, 16−31.

(23) Schleier, J. J., III; Marshall, L. A.; Davis, R. S.; Peterson, R. K. A quantitative approach for integrating multiple lines of evidence for the evaluation of environmental health risks. *PeerJ* **2015**, *3*, No. e730.

(24) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and ucnertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−1375.

(25) Pollard, S. Toxicity and risk. Context, principles and practice. *J. Risk Res.* **2005**, *8*, 453.

(26) Dourson, M.; Charnley, G.; Scheuplein, R. Differential sensitivity of children and adults to chemical toxicity - II. Risk and regulation. *Regul. Toxicol. Pharmacol.* **2002**, *35*, 448−467.

(27) Bian, Q.; Ping, Y.; Jun, W.; Lyu, Z.; Song, Y.; Zhang, L.; Liu, Z. A new method to evaluate toxicological data reliability in risk assessments. *Toxicol. Lett.* **2019**, *311*, 125−132.

(28) Schneider, K.; Schwarz, M.; Burkholder, I.; Kopp-Schneider, A.; Edler, L.; Kinsner-Ovaskainen, A.; Hartung, T.; Hoffmann, S.

"ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol. Lett.* **2009**, *189*, 138−144.
(29) Yang, L.; Neagu, D.; Cronin, M. T. D.; Hewitt, M.; Enoch, S. J.; Madden, J. C.; Przybylak, K. Towards a Fuzzy Expert System on Toxicological Data Quality Assessment. *Mol. Inf.* **2013**, *32*, 65−78.
(30) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(31) U. S. EPA. *Office of Pollution Prevention & Toxics, EPA's Safer Choice Program Master Criteria for Safer Ingredients*, version 2.2, 2015. https://www.epa.gov/sites/production/files/2013-12/documents/dfe_master_criteria_safer_ingredients_v2_1.pdf.
(32) Connors, K. A.; Beasley, A.; Barron, M. G.; Belanger, S. E.; Bonnell, M.; Brill, J. L.; de Zwart, D.; Kienzler, A.; Krailler, J.; Otter, R.; Phillips, J. L.; Embry, M. R. Creation of a curated aquatic toxicology database: EnviroTox. *Environ. Toxicol. Chem.* **2019**, *38*, 1062−1073.
(33) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2009.
(34) Teather, K.; Parrott, J. Assessing the chemical sensitivity of freshwater fish commonly used in toxicological studies. *Water Qual. Res. J. Can.* **2006**, *41*, 100−105.
(35) Pawar, G.; Madden, J. C.; Ebbrell, D.; Firman, J. W.; Cronin, M. T. D. In Silico Toxicology Data Resources to Support Read-Across and (Q)SAR. *Front. Pharmacol.* **2019**, *10*, 561.
(36) Fu, X.; Wojak, A.; Neagu, D.; Ridley, M.; Travis, K. Data governance in predictive toxicology: A review. *J. Cheminf.* **2011**, *3*, 24.
(37) Alves, V. M.; Borba, J.; Capuzzi, S. J.; Muratov, E.; Andrade, C. H.; Rusyn, I.; Tropsha, A. Oy Vey! A Comment on "Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships Outperforming Animal Test Reproducibility". *Toxicol. Sci.* **2019**, *167*, 3−4.

https://dx.doi.org/10.1021/acs.est.0c02224
Environ. Sci. Technol. 2020, 54, 12262−12270