Estimating variabilities of toxicological endpoints of concerns to the Agency for Toxic Substances and Disease Registry (ATSDR)

By

Ram Siwakoti
Degree to be awarded: Master of Science in Public Health

Department of Biostatistics and Bioinformatics

____
Faculty Adviser

____
Reader

Estimating variabilities of toxicological endpoints of concerns to the Agency for Toxic Substances and Disease Registry (ATSDR)

By

Ram Siwakoti

Bachelor of Science
Georgie Institute of Technology
2016

Thesis Faculty Adviser:

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2019

Abstract

Estimating variabilities of toxicological endpoints of concerns to the Agency for Toxic Substances and Disease Registry (ATSDR)

By Ram Siwakoti

Background: A significant number of chemicals in governmental databases lack health guidance values (HGVs) that are essential to protect public during chemical emergencies. In public health risk assessments, computational models could be used to fill these gaps when better toxicological information are not available. HGVs are commonly derived from toxicological endpoints such as benchmark dose (BMD) or a no-observed-adverse-effect level (NOAEL) and in absence of them, a lowest-observed-adverse-effect level (LOAEL) or median lethal dosage (LD_{50}). Because these endpoints are measured quantities, they are expected to carry a certain level of natural variability. However, the magnitude of such variability in each endpoint is currently unknown.

Objective: In the present study, we assessed variabilities of LOAELs, NOAELs, and LD_{50}s using data from ATSDR toxicological profiles of chemicals and other publicly available databases.

Methods: We estimated variability of each toxicological endpoint using distribution of sample variances of endpoints per chemical in acute, intermediate, and chronic exposure durations. The variability estimates were then used to obtain scaling factors to derive lower bounds on respective endpoints. Additionally, we assessed the influence of experimental test species, target organ systems, and availability of Minimum Risk Levels (MRLs) on variability.

Results: Variability of LD_{50}s was approximately half that of LOAELs whereas difference in variability of LOAELs and NOAELs was smaller. Matching endpoints by test species had no significant impact (except in intermediate duration endpoints) whereas matching target systems lowered NOAELs variability across all three exposure durations. Additionally, the availability of MRLs did not significantly affect variabilities of corresponding endpoints.

Conclusions: The findings from this study provide insight into variabilities of toxicological endpoints in ATSDR and other governmental databases, which could be useful in public health guidance and risk assessments.

Estimating variabilities of toxicological endpoints of concerns to the Agency for Toxic Substances and Disease Registry (ATSDR)

By

Ram Siwakoti

Bachelor of Science
Georgie Institute of Technology
2016

Thesis Faculty Adviser:

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2019

Acknowledgements

Everything I have accomplished so far, I have done so with the great support of my parents and family, especially my late grandfather, Kamalapati Ghimire. I am grateful to my family, friends and mentors who have supported me through difficult times, always encouraging me to see the light at the end of a tunnel. I am especially indebted to Dr. Ravi Sharma and Mrs. Leslie Sokolow who steered me towards the world of public health when I was an undergraduate student.

I am grateful to my supervisor and mentor Dr. Eugene Demchuk at the Computational Toxicology and Methods Development Lab (CompTox) at ATSDR. I developed interests in statistics and data science while I was training under him during my undergraduate and gap years. His tough style of questioning everything and challenging me to my limits while still being patient and understanding has helped me become a better researcher. He has been a great asset to me for last four years and few sentences here do not fully cover the extent of his influence on my professional/personal growth and development.

I will do injustice if I do not express my gratitude to our CompTox Team Lead, Dr. Clement Welsh, for his continued support of my ORISE fellowship. Additionally, I am extremely thankful to all my CompTox team members, especially Dr. Andrew Prussia, who has helped me from my first day with the team.

I am grateful to Dr. Jeffrey Switchenko for agreeing be my thesis adviser even though I was working on a project that was initially unfamiliar to him. His guidance and comments on academic and statistical matters have been of great assistance.

Finally, I am thankful to all members of the Emory Rollins School of Public Health Department of Biostatistics and Bioinformatics for everything they have done to make my Rollins experience the one to remembe

Table of Contents

Introduction

The Agency for Toxic Substances and Disease Registry (ATSDR) is mandated by the United States (US) Congress to prepare toxicological profiles and ascertain safe human exposure levels (SHELs) for hazardous substances included in the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) National Priorities List (NPL) [1]. As a response to the mandate, ATSDR regularly develops and publishes Minimum Risk Levels (MRLs) for acute (0–14 days), intermediate (14–365 days), and chronic (365+ days) exposure durations [1]. An MRL is an estimate of the daily human exposure to a hazardous substance over a specified period of time that is likely to be without appreciable risk of adverse non-cancer health effects [1]. These substance-specific estimates are intended to serve as screening levels "to identify contaminants and potential health effects that may be of concern at hazardous waste sites" [1]. ATSDR uses a point of departure (PoD)/uncertainty factor (UF) approach to derive an MRL [1]. A PoD is a point on the toxicological dose-response curve that is chosen as a starting point for low dose extrapolation, whereas UFs are adjustment factors used to account for various types of uncertainties, including variations in inter- or intra-species sensitivity [1][2]. ATSDR commonly uses a benchmark dose (BMD) or no observed adverse effect level (NOAEL) or, in absence of them, a lowest observed adverse effect level (LOAEL) as a PoD [1].

The derivation of MRLs involves a comprehensive literature search and review of toxicological information [1]. At times, such information is unavailable and exposure-specific MRLs are not derived. In fact, of 154 chemicals on the current MRL list, only 27 (18%) have the complete set of oral MRLs for all exposure durations [1]. For acute, intermediate, and chronic exposure durations, respectively, only 78, 107, and 65 oral MRLs are on the list [1]. This disparity represents a challenge for ATSDR community health investigations, whenever risk assessors are dealing with a chemical with limited toxicological information, i.e. when an MRL, LOAEL, or NOAEL for the chemical is not readily available.

Recently, it was shown that median oral lethal dosage (LD_{50}) is strongly correlated to MRL, LOAEL, or NOAEL [3]. Thus, potentially, an LD_{50} could be used as a PoD, when better toxicological information is not available. However, if several LD_{50}s are reported for a chemical, a question emerges: which one is correct? Traditional risk assessment assumes the lowest dose that causes an adverse health effect as a starting point for risk assessment, thus conducting the most conservative approach to public health guidance [1][2]. On the other hand, each measured physical quantity has a measurement error associated with it. If only one LD_{50} is available, the experimental measurement error is uncertain. Although, that value itself represents the best-known estimate of the average (i.e. the statistical mean), the confidence interval (CI) on it is indeterminate [4]. Therefore, it is impossible to assert a desired level of risk assessment confidence (e.g. the 95th percentile). Similar disconcertment is expected with respect to LOAEL, NOAEL, or any other toxicological endpoint. To that extent, even the current practice using the no or lowest adverse observed endpoint (i.e. NOAEL or LOAEL) as a PoD begs re-examination, especially since not all chemicals have equal available quantities of toxicological data (i.e. different empirical percentiles may correspond to the PoD). In such scenario, the percentiles from which risk assessment stems are different for chemicals with less L(N)OAELs versus chemicals with more L(N)OAELs [1][2].

With increasing emphasis on the use of non-animal alternative models for toxicological endpoints derivation, governmental agencies are starting to explore the feasibility of using LD_{50}s based computational models for risk assessment purposes [5]. Hoffmann et al. compiled a database of in-vivo LD_{50}s and conducted a statistical analysis to assess their variability and reliability [6]. The median standard deviation (SD) of a log transformed LD_{50} was estimated to be 0.20 [6]. Additionally, Karmaus et al. compiled a comprehensive inventory of rat LD_{50}s and studied their variability [7]. The "global standard deviation" of log transformed LD_{50}s as defined in that study was estimated to be 0.83 whereas ±0.31 log10(LD_{50} in mg/kg) was recommended for interval estimation of median of lower quantile [7]. In both studies, it was shown that many chemicals had reported LD_{50}s that span an orders-of-magnitude range that defies a simplistic range-bound approach to LD_{50} classification, i.e., measured chemical's LD_{50}s may span several categorization bounds (e.g. toxic/non-toxic) [6][7]. Much recently, in April 2018, the Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) helped to organize a collaborative workshop to discuss and build LD_{50}s based computational models for acute oral systemic toxicity [8]. These developments highlight the interest as well as a movement towards making sense of publicly available oral LD_{50} data to utilize them in risk assessment purposes.

In the current study, we expand on the findings from Hoffmann et al. and Karmaus et al. to access uncertainty inherent in any LD_{50} by estimating its SD. With the use of estimated SD, the variability or lower bound on any given endpoint can be ascribed. We used carefully curated LD_{50}s sourced from public databases to estimate LD_{50} variability. Additionally, we extended the estimation of variability to LOAELs and NOAELs from ATSDR's toxicological profiles. This will allow us to understand the diversity of endpoints within each chemical and, potentially, lead to a new model of probabilistic chemical risk assessment. Finally, whenever possible, we assessed the effects of experimental test species, health effects, availability of MRLs, and number of available studies on variability.

Methods

Data Compilation

All toxicological endpoints data used in the current study were publicly available. The data compilation procedures for LD_{50}s and L(N)OAELs are described below:

LD_{50}s: LD_{50}s and associated toxicological information for chemicals were extracted from the ATSDR's toxicological profiles, ChemIDplus TOXicology Data NETwork (ChemIDplus TOXNET), Registry of Toxic Effects of Chemical Substances (RTECS), and Drug Dosage in Laboratory Animals Handbook [11]. LD_{50}s from different sources were then integrated in a custom-designed Microsoft (MS) Access® database for additional data processing. The total number of LD_{50}s and unique CAS Registry Number (CASRN) corresponding to each database are presented in Table 1a.

L(N)OAELs: L(N)OAELs and associated toxicological information were extracted from the ATSDR's Levels of Significant Exposure (LSE) tables housed in a Sybase database system. They were then imported to MS Access® database and grouped by exposure durations. The total number of L(N)OAELs and unique CASRN corresponding to each exposure durations are presented in Table 1b.

Data Processing: In a preliminarily quality control assessment, the quality of the compiled endpoints was deemed. Following selection criteria were used to filter out unreliable records from our databases:

Route of administration: Only endpoints stemming from oral administration experiments were included. Endpoints derived using various kinds of extrapolation (for example, route-to-route extrapolation using physiologically-based pharmacokinetic (PBPK) modeling or an acute oral NOAEL for a chemical derived from an acute inhalation NOAEL) were excluded.

Unit of measurements: Milligram per kilogram of body weight per day (mg/kg-day) is a common standard for reporting oral mammalian toxicity. Endpoints reported in other units were excluded (except for oral endpoints reported in parts per million (ppm) unit, which has a 1-1 conversion with mg/kg).

Limit dose: Only LD_{50}s listed as a "point estimate" were included. Censored LD_{50}s (such as >, < 500 mg/kg-day) were excluded.

Elemental metals: Usually, the toxicity of a metal is tested using its salt. In absence of a complete information related to a compound, the correct molecular weight (MW) cannot be assigned. Hence, all endpoints recorded for CASRN corresponding to elemental metals were excluded.

Availability of at least 3 endpoints: It is not desirable to derive a variance using a very small sample. In any dataset, chemicals with less than 3 observations for a given endpoint were excluded from analysis for that endpoint.

Duplicate records: If multiple identical doses were identified for a chemical with same experimental species, they were assumed to be duplicates (meaning the same dose from the same study is available across different databases) and was recorded only once (only relevant to LD_{50}s).

Unverifiable records: Statistical outliers and other suspicious records were verified using the primary literature source. Records with no verifiable primary sources were excluded (applies only to LD_{50}s). L(N)OAELs were extracted from the LSE tables and were expected to be sufficiently verified during the MRL review process.

Stratification of the data: Stratification of endpoints by multiple factors was considered:

Species and target organ systems: Both LD_{50}s and L(N)OAELs (from acute, intermediate, and chronic exposure duration studies) were stratified by the experimental species (species matched) used to derive them (Table 2, Table 3a-3b). In addition, L(N)OAELs were also stratified by the target system (systems matched) (Table 3c-3d) used to derive them. A significant number of endpoints had "NA" or "NR" listed as their target systems. Hence, we consulted with a senior ATSDR scientist and based on her recommendation, replaced NAs or NRs in the system column with non-NA entries from the corresponding category column (LSE tables include columns for both target systems and categories associated with a chemical intake. See ATSDR's *Guidance for the Preparation of Toxicological Profiles* for details [12]). This step of merging health effects was necessary because of the way ATSDR classifies health effects in its toxicological profiles. For full (species or target organ mixed) data analysis, we excluded duplicate observations, i.e., if an identical dose was listed for multiple target systems, it was included only once in our analysis. Based on the availability of data, the impact of test species and target organs on variabilities were explored.

Availability of MRLs: The scientific rigor and quality of data requirements often limits whether or not an MRL can be derived for a chemical [1]. Hence, L(N)OAELs were stratified based on the availability of MRLs (Table 3e). This allowed us to explore the effect of availability of MRLs on variabilities of corresponding endpoints.

Unit conversion and data preparation: In order to account for the biochemical processes involving chemical metabolism and receptor binding at the molecular level, all doses were converted from mg/kg-day to mole/kg-day unit. MWs of chemical entities matched by their CASRNs were extracted from the TOXNET database. Furthermore, all mole-transformed values were log-transformed and negated ($-\log_{10}$) by analogy with the pH scale of acidity in chemistry. Therefore, further in the text they are denoted in a similar fashion, i.e. pLD_{50}, pNOAEL, pLOAEL. For instance, for an LD_{50} reported in mg/kg-day:

$$\mathrm{pLD}_{50} = -\log_{10}(\frac{\mathrm{LD}_{50}}{\mathrm{MW} \times 1000})$$

Log-transformation generally achieves normality of right-skewed data, which is important in statistical analysis [13]. Mathematical negation of log-molar doses was implemented for reader's convenience, as all studied values became positive. However, based on the implemented unit conversion, a smaller dose in mg/kg unit translates to a larger log-molar-converted value, i.e. an endpoint of a more toxic substance (small LD_{50}) converts to a large pLD_{50}. In the current study, all analyses, unless otherwise specified, were performed on negative-log molar transformed values.

Data analysis

Descriptive summary: The frequency distributions of experimental species and target organs for all relevant endpoints and exposure durations were summarized. Previously, it was shown that toxicological endpoints such as LD_{50}s follow a lognormal distribution [6][7]. Hence, the distributions of log-transformed endpoints were tested using the normal Q-Q plots and Shapiro-Wilk normality test (SW test). Points with outlier characteristics were identified based on Tukey fences $\left(\mathrm{Q}_{(1,3)} \pm\ 1.5(\mathrm{Q}_3 - \mathrm{Q}_1); \mathrm{Q} = \text{quantile}\right)$ and investigated for any discrepancies [14].

Estimation of variability: In the current study, we explored two different methods for the derivation of variances or SDs.

Method 1 (pooled variance): The derivation of pooled variance of endpoints was explored using a method described by Li et al. [15]. This method relies on two important assumptions:

Assumption 1: k samples are drawn from a normal distribution with a common mean, i.e., each sample is drawn from $\mathrm{N}(\mu, \sigma_i^2)$ where i = 1, 2, ... k.

Assumption 2: The SDs for each sample are identical, i.e. $\sigma_i = \ldots = \sigma_k$

For our problem, each chemical represents an individual sample. The mean centering of endpoints (i.e., $\mathrm{pLD_{50}s}$, pLOAELs, pNOAELs) for each chemical was expected to result in similar per chemical means. The Levene's - and F-tests can be used to test for the homogeneity of variances across different chemicals. If the homogeneity of variances assumption holds, the unbiased pooled variance estimate ($S_{p_i}^2$) is given by:

$$S_{p_k}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \ldots + (n_k - 1)S_k^2}{n_1 + n_2 + \ldots + n_k - K}$$

where $n_1, n_2, \ldots, n_k$ and $S_1, S_2, \ldots, S_k$ represent the sample size and sample SD of each chemical, respectively. However, if the assumption of homogeneity of variances does not hold, then the derivation of pooled variances using method 1 is not suitable.

Method 2 (distribution of variances): While method 1 relies on the distribution of mean-centered endpoints as well as the assumptions of identical means and SDs across samples to provide a single pooled variance estimate, method 2 relies on the distribution of sample variances. Using this method, we first estimated sample variances of endpoints (S_i^2), where $i = 1,2,, \ldots, k$ represent each chemical in our dataset:

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}_i\right)^2 \text{ where}$$

n_i = number of data points available per chemical

X_{ij} = pLD$_{50}$s ($j = 1,2,\ldots,m$ = individual pLD$_{50}$ for a chemical i)

$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n} X_{ij}$ = sample mean of pLD$_{50}$s for a chemical i

If $X_{ij} = X_{i1}, X_{i2}, \ldots, X_{im}$ are observations of a random sample of size n_i from a normal distribution $N(\mu_i, \sigma_i^2)$, as we are assuming in our study, then the distribution for each S_i^2 can be approximated using the following relationship [16]:

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2(n_i - 1)$$

However, the value of σ_i^2 is currently unknown for all endpoints. In fact, the main objective of the current study is to approximate this unknown parameter. Thus, it is not possible to ascertain the theoretical distribution of sample variances of endpoints at this time. Instead, an empirical distribution of S_i^2 for each endpoint was used to estimate the best measure of central tendency.

Selection of the best measure of central tendency: In a symmetric distribution, either of the sample mean or median could be used to express the central tendency. However, a distribution of sample variances (S_i^2) tends to be right skewed, [16] thus, making a choice of the variance measure of central tendency obscure. The distribution of sample SDs (S_i) is less skewed, but it is an inherently biased estimate [17], making the numerical estimation from variances (rather than SDs) preferable. Thus, on the one hand, estimation from the variance is preferred, but on the other hand, expression of the central tendency from the variance is indeterminate, because valid arguments can be raised in favor of either the mean or median.

Because:

(1) the goal of present work is to estimate a public-health protective lower boundary on the endpoint, which involves S_i, not S_i^2;

(2) in the combined distribution of SDs, the difference between mean and median is small (as compared to the difference between mean and median in the combined distribution of variances); and

(3) of the available measures of central tendency, only the median is robust in respect to monotonous-function transformations (2nd power);

we chose the interval estimation of the median from the distribution of variances and converted them to SDs as the measure of central tendency in the current study. The 95% CIs were approximated using the Bias-corrected & accelerated (BCa) method with 10,000 bootstrap replications as implemented in the "boot" package in R [18]. For the sake of comparison, we also derived the interval estimates of mean, lower, and upper quantiles of SDs.

Assessing test species or target organ specific effects: For both LD_{50}s and L(N)OAELs, we explored the effect of test species on variability. For L(N)OAELs, we also explored the effect of target systems. The distributions of sample variances across different species or systems were compared using the Kolmogorov Smirnov test (KS test) and the medians were compared using the Wilcoxon Signed Rank Sum test (Wilcox test) as implemented in R (α=0.05).

Derivation of scaling factors: If log-molar endpoints are normally distributed, the derived SDs could be back transformed to obtain scaling factors in original unit of endpoints. Based on these factors, the lower one-sided 95% bound (for example, $\mathrm{LD}_{50_\mathrm{L}}$, $\mathrm{LOAEL_L}$, or $\mathrm{NOAEL_L}$) on a single endpoint or geometric mean (GM) of endpoints can be calculated. For example:

$$-\log10\left(\mathrm{LD}_{50_\mathrm{L}}\right) = -\log10\left(\mathrm{GM}\left(\mathrm{LD}_{50_\mathrm{i}}\right)\right) + 1.64\mathrm{SD}$$

$$\log10\left(\frac{\mathrm{GM}\left(\mathrm{LD}_{50_\mathrm{i}}\right)}{\mathrm{LD}_{50_\mathrm{L}}}\right) = 1.64\mathrm{SD} \rightarrow \mathrm{LD}_{50_\mathrm{L}} = \frac{\mathrm{GM}\left(\mathrm{LD}_{50_\mathrm{i}}\right)}{10^{1.64\mathrm{SD}}}$$

In the current study, when feasible, we derived estimated scaling factors to derive lower bounds on the studied endpoints.

Results

Data description

LD_{50}s: After the application of selection criteria stated in methods section, 306 chemicals with 1420 LD_{50}s were available for analysis (Table 1a). The most common test species used to derive LD_{50}s were rat and mouse (Table 2). The log-molar distribution of LD_{50}s were approximately symmetric (Figure 1a) but failed the normality test (SW test: $p < 0.0001$; Table 3f). Based on Tukey-fence analysis, we identified extreme points possibly contributing to non-normality of data. Excluding these points improved symmetry (Figure 1b), however, the distribution was still non-normal (SW test: $p < 0.05$) and were further screened for accuracy.

Immediately, no defects were observed, and these points were not excluded. When the experimental species used to derive LD_{50}s were matched, 70 chemicals with 404 LD_{50}s were available (Table 1a). Similar to species mixed dataset, the log-transformed endpoints failed the normality test (SW test: $p < 0.0001$) due to outliers at tails (Figure 1a, 1b).

The normality assumption of log-molar endpoints is necessary while deriving a lower bound on an endpoint using a recommended SD based scaling factor (refer to *Derivation of factors based on SDs* section for details). Because previous studies have shown the distribution of log-molar endpoints to be normal [6], and Q-Q plots improved with the exclusion of few outliers (Figure 1b), we will proceed with the normal distribution assumption of log-molar endpoints in the current study.

L(N)OAELs: Based on the LOAELs datasets, 93, 121, and 58 unique chemicals corresponding to 740, 1209, and 368 data points for acute, intermediate, and chronic exposure durations, respectively, were available. The NOAELs datasets were slightly larger: 104, 132, and 75 chemicals corresponding to 848, 1529, 547 data points for acute, intermediate, and chronic exposure durations, respectively, were available (Table 1b). Similar to LD_{50}s, rat and mouse were the most common experimental species (Table 3a, 3b). Among the identified target systems, hepatic, neurological, and body weight change were the most common (Table 3c, 3d). Additionally, the stratification by the presence of an MRL showed that approximately 40 - 50% of endpoints in each exposure duration corresponded to chemicals that lacked MRLs (Table 3e).

The distributions of both endpoints followed trends similar to LD_{50}s (Table 3f), i.e., approximately symmetric with slight deviation from the normal lines at tails (Figure 1c, 1d). All L(N)OAELs used in the current study were extracted from the ATSDR toxicological profiles that were expected to be vetted before their publications. Hence, none of the extreme points were excluded.

Data analysis

Accessing homogeneity of variances assumption

In the methods section, we presented two approaches to estimate variability of an endpoint, of which method 1 relies on the assumption of homogeneity of variances across all samples. Hence, we tested that assumption using the Levene's test (Table 4a). The sample variances across different chemicals in all datasets (LD_{50}s, LOAELs, and NOAELs) were significantly different from each other (Levene's test: $p < 0.05$), indicating that the assumption was not valid. We further explored the homogeneity of variance assumption by conducting a test case: we selected all chemicals with more than 15 LD_{50}s from species mixed dataset (n = 4; Malathion, Dioxin, Disulfoton, and Parathion-methyl) and conducted 6 $\left(\frac{4!}{2!2!}\right)$ simultaneous multiple F-tests between chemical pairs. The LD_{50}s dataset was chosen because it was previously curated and the sample size greater than 15 was chosen so that we would have a reasonable balance of sample size and number of chemicals available for analyses. Adjusting for the multiple comparison of variances using the Bonferroni correction, the sample variances between any pair of chemicals were significantly different from each other ($p < \alpha = 0.0083$) 3 out of 6 times (Table

4a). The observed differences based on these tests indicated that the chemical specific differences arising either from their physical and chemical properties or experimental designs may exist. Given these results, we did not pursue the estimation of a pooled variance based on method 1. Instead, only method 2, i.e., the empirical distributions of sample variances, were used to estimate uncertainties of both LD_{50}s and L(N)OAELs in the current study.

Variance estimation of LD_{50}s

Using log-molar LD_{50}s from each chemical, the empirical distributions of sample variances were drawn. As expected, the distributions of variances were right skewed (Figure 2a, 2b). The distributions of SDs were less skewed, yet still not symmetric.

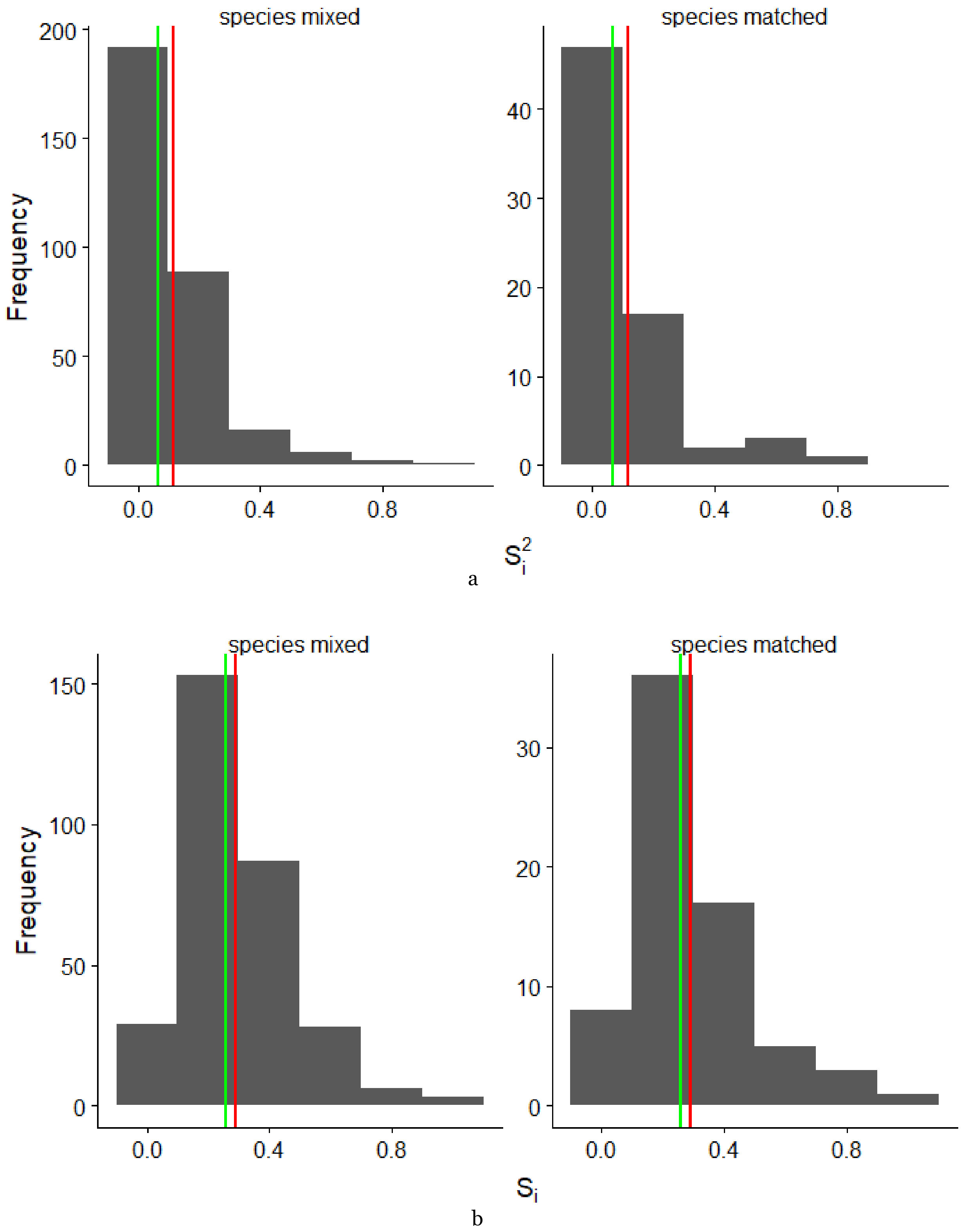


Figure 2. Distributions of **a.** sample variances (S^2) and **b.** SDs (S) of log-molar LD_{50}s. Green and red lines represent the median and mean of distributions, respectively.

Based on these distributions, several variance estimates were derived and converted to SDs (Table 4b). The median SD estimate of log-molar LD_{50}s was 0.26 (0.24-0.28). Matching test species did not reduce uncertainty, as the median SD remained statistically unchanged at 0.27 (0.23-0.29). Similar trend was observed at the 25th and 75th percentiles of distributions. A formal hypothesis test to compare SDs from full versus the species matched subset was not performed at this time because the species matched dataset is a subset of full dataset, thus violating an independent samples assumption required of any such test [19]. Instead, the 95% CIs of relevant SDs were used to make necessary inferences.

Table 4b. SD estimates for LD_{50}s for species mixed and species matched datasets. The 95% CIs were approximated using the BCa method with 10,000 bootstrap replications as implemented in the "boot" package in R.

Description	**species mixed (full)**	**species matched**
Number of LD_{50}s (n)	1420	404
Number of unique chemicals (N)	306	70
Standard Deviation (SD)		
25th percentile (95% CI)	0.17 (0.15 - 0.18)	0.17 (0.11 - 0.20)
Median (95% CI)	0.26 (0.23 - 0.28)	0.27 (0.22 - 0.29)
Mean (95% CI)	0.34 (0.32 - 0.36)	0.35 (0.30 - 0.41)
75th percentile (95% CI)	0.37 (0.31 - 0.44)	0.36 (0.31 - 0.45)

We further explored species-specific influence on variability by comparing the distributions of sample variances between rat and mouse. These two species were chosen because they had the largest sample size available for comparison. The distribution of variances between these species were not significantly different from each other (KS test: $p > 0.05$) (Table 6a, Figure 2c), consistent with our earlier findings from species mixed versus species matched datasets.

Variance estimation of L(N)OAELs

The distributions of sample variances and SDs of L(N)OAELs were similar to that of LD_{50}s, i.e., non-symmetric right skewed (Figure 3a, 3b). Hence, the median SDs of endpoints in acute, intermediate, and chronic exposure durations were also estimated using the empirical distributions of sample variances.

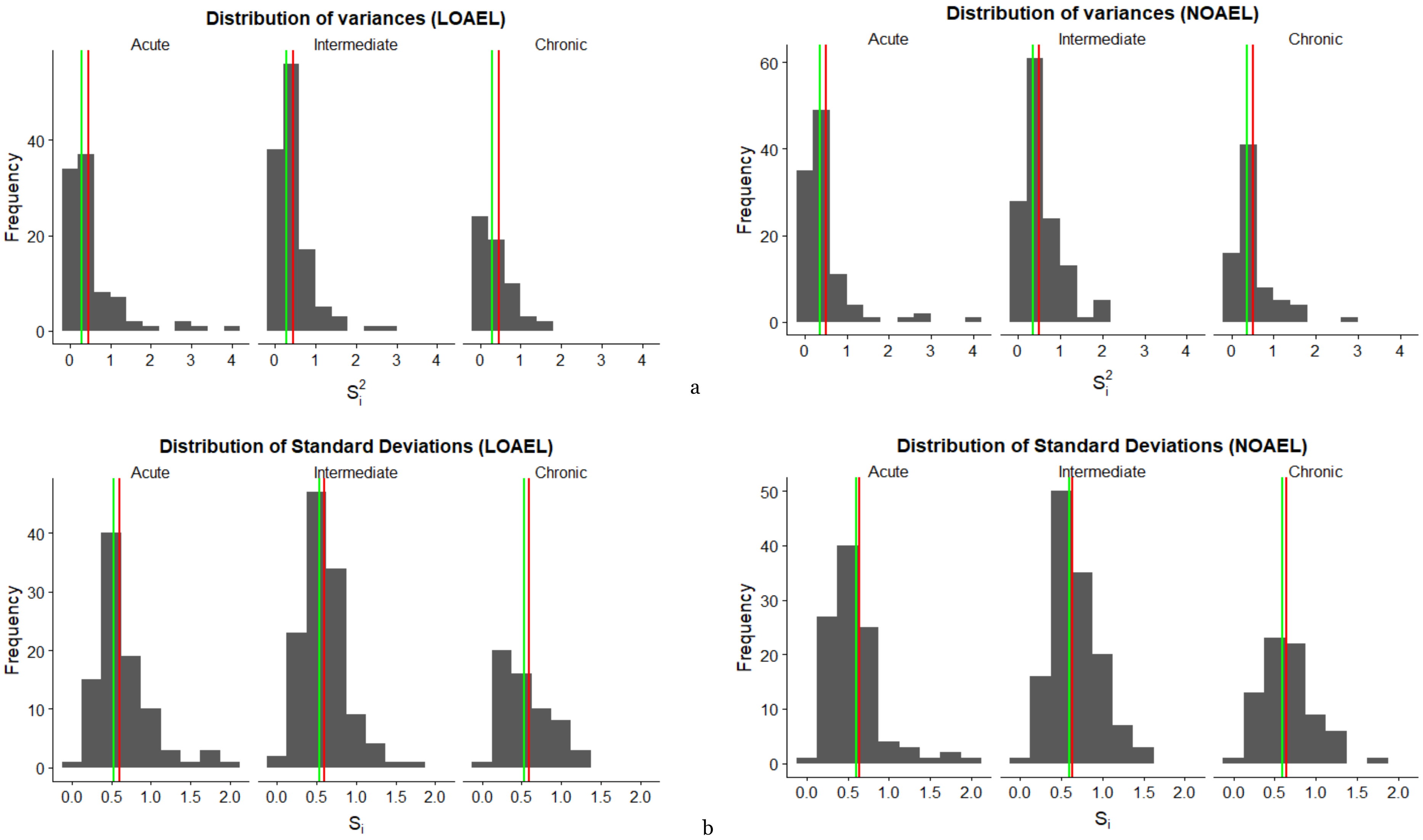


Figure 3. Distributions of **a.** sample variances (S^2) and **b.** SDs (S) of log-molar LOAELs and NOAELs. Green and red lines represent the median and mean of distributions, respectively.

The median SDs of log-molar LOAELs and NOAELs using full datasets were not significantly different from each other (LOAELs: 0.53 (0.46 - 0.60); NOAELs: 0.52 (0.48 - 0.60)) in acute exposure duration ((Wilcox test: $p > 0.05$); Table 5a). In contrast, the median SDs of NOAELs were higher as compared to the median SDs of LOAELs in intermediate (LOAELs: 0.55 (0.50-0.62); NOAELs: 0.62 (0.58-0.68)) and chronic (LOAELs: 0.49 (0.37-0.58); NOAELs: 0.63 (0.58-0.68)) exposure durations (Wilcox test: $p < 0.05$). These relations were consistent at the 25th and 75th percentiles of distributions. A specific trend in percentile SDs relative to exposure duration (for example, acute duration SD < intermediate duration SD < chronic duration SD) was not observed.

Table 5a. SD estimates for LOAELs and NOAELs. All available endpoints were used, i.e., distributions were based on full datasets. However, if an identical dose was recorded for multiple systems, it was included only once.

L(N)OAELs – full datasets						
	Acute		**Intermediate**		**Chronic**	
Description	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**
Number of endpoints (n)	740	848	1209	1529	368	547
Number of unique chemicals (N)	93	104	121	132	58	75
Standard deviation						
25th percentile (95% CI)	0.43 (0.33-0.44)	0.35 (0.32-0.44)	0.40 (0.33-0.45)	0.47 (0.43-0.51)	0.32 (0.2-0.38)	0.49 (0.35-0.55)
Median (95% CI)	0.53 (0.46-0.60)	0.52 0.48-0.60)	0.55 (0.50-0.62)	0.62 (0.58-0.68)	0.49 (0.37-0.58)	0.63 (0.58-0.68)
Mean (95% CI)	0.73 (0.65-0.84)	0.68 (0.60-0.78)	0.66 (0.61-0.72)	0.73 (0.68-0.78)	0.63 (0.55-0.71)	0.73 (0.66-0.81)
75th percentile (95% CI)	0.77 (0.65-0.89)	0.71 (0.63-0.84)	0.72 (0.68-0.84)	0.84 (0.77-0.93)	0.77 (0.60-0.91)	0.76 (0.71-0.94)

In addition to deriving SD estimates based on the species matched subsets, for L(N)OAELs, we also derived SD estimates based on the target systems matched subsets. Overall, the median SDs of both endpoints derived using the species matched subsets were lower as compared to the median SDs derived using full datasets, although the differences were not statistically significant in acute and chronic exposure durations (Table 5b). When the target systems were matched, the median SDs of LOAELs were still similar to those from full datasets whereas the median SDs of NOAELs were lower (Table 5c).

In order to further investigate the test species or target system-specific effect on variabilities, we stratified all 6 L(N)OAELs datasets by the test species or target systems. From each full dataset, we chose 3 strata with the largest sample size (i.e., 3 species strata and 3 system strata) and compared their empirical distributions of sample variances (Table 6a). The distributions of variances between the compared species were not significantly different from each other (KS test: $p > 0.05$). Similar results were observed when the distributions of variances were compared between different target systems, i.e., only 3 comparisons (in chronic LOAELs and NOAELs datasets) resulted in significantly different distributions of variances between the compared systems (KS test: $p < 0.05$; Table 6b). These 3 comparisons involved body weight change (-versus hepatic or renal system in LOAELs; -versus hematological system in NOAELs).

The species or target system specific comparisons can also be visualized using histograms, scatter, or density plots (Figure 4a - 4c). The rat and mouse distributions of variances were approximately similar, i.e., the density curves were overlapping. However, based on both chronic LOAELs and NOAELs datasets, the distributions of variances based on body weight change system were more skewed. Finally, we compared variabilities of endpoints based on the availability of corresponding MRLs. The distributions of variances based on the "MRLs present" subsets were not significantly different from the distributions of variances derived using the "MRLs not present" subsets across all endpoints and exposure durations ((Table 7a, Figure 4d). As such, the median SDs between two subsets were also approximately similar (Table 7b, 7c). These similarities in the distributions of SDs, however, do not indicate that the quality of data in compared subsets were similar as several factors influence the selection of chemicals for MRLs derivation [12].

Variability versus sample size

Usually, in experimental studies, the precision of an effect is expected to increase as the number of repeated measurements increase, i.e. the standard error of the mean $\left(\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{\mathrm{n-1}}}\right)$ decreases as n increases. However, in the current study, we observed that an increase in number of LD_{50}s or L(N)OAELs per chemical did not necessarily translate to corresponding decrease in uncertainty (Figure 5a,5b). In fact, SDs exhibited a subtle upward trend with an increase in sample size. The reason for this behavior, at present, is unclear although several factors such as experimental design, purity of chemicals, strains and age of species might play a role. As the number of independent experiments increase, variability in these factors may also increase, which disrupts uniformity of the stochastic process of endpoint derivation.

Derivation of factors based on SDs

The median SD estimates of LD_{50}s and L(N)OAELs derived in the current study were used to obtain scaling factors based on the original unit of endpoints (Table 8). Based on these factors, lower bound on an endpoint or a geometric mean of endpoints can be deduced. For LD_{50}s, the median SDs based on either species mixed or matched dataset transformed to a scaling factor of 2.75 (approx.).

Table 8. Scaling factors based on estimated median SDs to derive 95% lower bound on a geometric mean of endpoints or a single endpoint. For example, to obtain a lower bound on an acute LOAEL, a division factor of 7.40 could be applied.

Endpoint	Duration	Species mixed (full)	Species matched	Target systems matched
LD_{50}s	**Acute**	2.67 (2.38 - 2.88)	2.77 (2.38 - 2.99)	na
LOAELs	**Acute**	7.40 (5.68 - 9.64)	6.13 (5.27 - 7.68)	6.86 (5.07 - 8.29)
	intermediate	7.98 (6.61 - 10.39)	5.68 (4.88 - 6.86)	6.61 (5.47 - 7.40)
	Chronic	6.36 (4.04 - 8.94)	5.47 (3.61 - 8.94)	5.27 (4.04 - 7.68)
NOAELs	**Acute**	7.13 (6.13 - 9.64)	6.36 (5.27 - 7.68)	4.70 (3.89 - 5.90)
	Intermediate	10.39 (8.94 - 13.04)	7.40 (6.61 - 8.94)	6.36 (5.68 - 7.68)
	Chronic	10.79 (8.94 - 13.04)	8.94 (6.61 - 10.79)	6.36 (5.47 - 7.13)

Preliminary factors were also derived for LOAELs and NOAELs. For LOAELs, the median SDs based on full datasets transformed to a scaling factor of 7 (approx.), irrespective of exposure durations. In contrast, when the experimental species or target organs were matched, it decreased to 6 (approx.). On the other hand, for NOAELs, scaling factors of 7 (approx.) for acute and 10 (approx.) for intermediate or chronic exposure durations were obtained. When the experimental species were matched, the factors decreased to 6 (approx.) for acute and 8 (approx.) for intermediate or chronic exposure durations. Similarly, when the target systems were matched, the factors further decreased to 5 (approx.) for acute and 6 (approx.) for intermediate or chronic exposure durations.

Discussion

Variability of LD_{50}s and comparison to previous studies

The median SD of a log-molar LD_{50} based on full dataset was 0.26. This estimate was consistent with the range of SDs derived in an earlier study by Hoffmann et al [6]. After excluding multiple outliers, Hoffman et al. reported a median SD of 0.20 with SDs for majority of chemicals smaller than 0.5. In their study, LD_{50}s were only log-transformed and not log-molar transformed, which is necessary for standardizing effects when chemicals with different MWs are studied together. Additionally, some of the data points used in their study could not be independently verified. On the other hand, the median SD derived in the current study was lower than the "global SD" of 0.81 derived by Karmaus et al, but closer to the SD they proposed to derive a 95% CI range (±0.31 log10(mg/kg) units) [7]. Similar to Hoffman et al., Karmaus et al. did not standardize LD_{50}s of individual chemicals by their MWs. Although, a significantly larger number of chemicals or LD_{50}s were used in the Karmaus study, we did not supplement the current study with the addition of new chemicals from their database. The data composition in the Karmaus study was not clearly understandable as several LD_{50}s indicated as "limit values" were present in the database uploaded to their website. In addition, the primary sources of LD_{50}s were not provided, making impossible the independent verification of endpoints when necessary. Absent clear knowledge regarding the data extraction or reduction processes, we decided to complete our analysis based on the in-house LD_{50}s database that we previously developed and curated. Furthermore, no species-specific effect on variability of an LD_{50} was observed, consistent with the findings from Hoffman at al. Test species may uniquely respond to a chemical with differing levels of sensitivities. However, in the present study, the interspecies variation of response did not exceed the intraspecies one. Hence, the SDs or the scaling factors estimated in the current study can be applied to any LD_{50} regardless of the test species used to derive it.

Variability of L(N)OAELs and comparison to LD_{50}s

The median SDs of log-molar LOAELs and NOAELs were similar to each other in acute exposure duration whereas the NOAELs SDs were significantly higher in intermediate and chronic exposure durations. The observation of higher variability in NOAELs is consistent with the toxicological properties based on which these endpoints are commonly determined. For instance, both NOAELs and LOAELs stem from the same dose-response curve (Figure 6), however, LOAELs are only determined when the predetermined level of adverse health effects are observed whereas NOAELs are determined when no adverse health effects are observed [12].

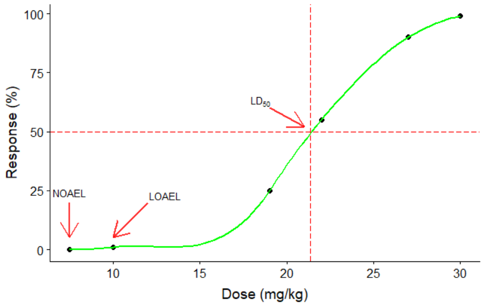


Figure 6. An illustration of a dose response curve used in toxicological endpoint derivation studies. Plot is not drawn to scale.

The study design factors such as sample size (i.e., animals used per dose group) and dose spacing play important roles in the determination of a LOAEL or NOAEL [20]. When a study is conducted with a larger sample size, the researchers are likelier to detect significantly different health effects between the treatment and the control group at a given dose (thus resulting in a LOAEL). However, with a smaller sample size, the differences in health effects between the compared groups may not be detected (thus resulting in a NOAEL). Similarly, the dose selection, i.e., a priori specific dose selected to test toxicological effects of a chemical plays an important role in the endpoint determination [20]. If a study is conducted with a large dose spacing, the true LOAEL might lie between the tested doses whereas the true NOAEL might lie farther below the lowest tested dose, resulting in an increased uncertainty of the reported values.

In contrast, the issues related to dose spacing and sample size are less severe in LD_{50} studies. The cumulative lethal effect observed for an LD_{50} is definite; either a species is dead or not at a tested dose and no comparison between the treatment and control groups is made. These differences possibly explain much lower variability of an LD_{50}. In fact, based on full datasets, the median SD of an LD_{50} was approximately half the median SD of either a LOAEL or NOAEL.

Another important distinction between LD_{50}s and LOAELs or NOAELs is the systems affected by a dose. L(N)OAELs are derived for different target organs or health effects depending on their sensitivities. In the current study, when the distributions of L(N)OAELs variances across target systems with the largest sample sizes were compared, only statistically significant differences were observed between the body weight change and hepatic or neurological system in chronic LOAELs and between the body weight change and the hematological system in chronic NOAELs dataset. The observed difference could be a random fluctuation, especially in chronic LOAELs dataset where the sample sizes of the compared systems are unbalanced and small. However, the difference could also be systematic. Measuring the body weight effect endpoint is a straightforward and uniform process, i.e., it is based on the significant difference between the terminal body weight of treatment versus control group [12]. The researcher could simply weigh the animals and based on the percentage change in weight, decide whether an effect exist at a given dose. On the other hand, the measurement of other systems effects is not obvious. For example, hematological effect endpoints are based on several effects such as anemia, cyanosis, erythrocytopenia, Leukopenia and so on [12]. The measurement and analysis procedures for multiple effects could vary among different studies, leading to an increased variability.

Scaling factors to derive a lower bound on an endpoint

The median SDs derived in the current study were transformed to scaling factors to obtain the lower bound on an endpoint in the original unit. The factor for LD_{50}s was smaller than the factors for LOAELs which were equal or smaller than the factors for NOAELs. An illustration of an application of a scaling factor is shown in Table 9. The geometric mean of all currently available Acrylamide LD_{50}s

Table 9. LD_{50}s for Acrylamide and the proposed lower bound based on the scaling factor of **2.7**

LD_{50}	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
species	**rat**	**rat**	**rat**	**rat**	**rat**	**rat**	**rat**	**mouse**	**mouse**	**gn. pig**	**rabbit**
Dose (mg/kg-day)	124	175	180	203	294	316	413	107	195	180	150
GM	**196.79**										
LD_{50_L}	**72.89**										

regardless of test species is 196.8. When the scaling factor of 2.7 based on the median SD of LD_{50}s is applied, the lower bound is estimated at 73 mg/kg-day, which is lower, thus more health protective, than the current minimum LD_{50} of 107 mg/kg-day. Based on the risk assessment need, an SD estimate at a higher percentile can be applied to derive a higher and more health protective scaling factor using a similar protocol.

The variances estimated in the current study could also be useful in computational toxicological studies that seek to derive cross-duration (i.e., acute -> intermediate, acute -> chronic, intermediate -> chronic) or cross-endpoints (i.e., LD_{50} -> LOAEL, LD_{50} -> NOAEL, etc) extrapolation factors. For example, when regression models are used to study a possible association between cross-duration endpoints, an application of the ordinary least squares (OLS) regression approach is not valid since both endpoints are subjected to uncertainty [21]. Hence, an error-in-variable model approach of regression analysis is warranted, which requires the estimates of sample variances of both endpoints such as the ones derived in this study [21].

Further study

In this study, we extracted toxicological endpoints from the ATSDR and other governmental databases to estimate their variabilities. The LD_{50}s variability was estimated using previously curated in-house database. While the quality of data used was better, the sample size could be improved upon. Recently, ICCVAM has published LD_{50}s for thousands of chemicals. However, a careful curation and data processing is required before these LD_{50}s could be used. Hence, in the future, attention should be given to curation and possible utilization of these LD_{50}s to improve study power and precision. Similarly, for L(N)OAELs, additional data may be collected from the Environmental Protection Agency (EPA)'s Integrated Risk Information System (IRIS) database.

Another possible area for further exploration is the use of a pooled variance estimation under the assumption of non-homogeneous variances across samples. One such method is provided by Li et al.; however, we were not able to use that method in the current study. A pooled variance estimate based on theoretical distribution of variances, when valid, could provide a useful comparison or validation of variances estimated using a non-parametric approach.

Furthermore, we observed that endpoints variability did not necessarily decrease with an increase in number of observations or studies. As mentioned above, several factors might affect this behavior and deserves more rigorous exploration. Such exploration could be carried out either through increased toxicological and experimental data collection, simulation studies, or more robust statistical techniques to account for possible confounders.

Finally, while we estimated scaling factors to derive lower bounds on endpoints, we did not formally validate our models due to smaller sample size. For example, we did not divide our datasets into testing and training sets to evaluate the reliabilities of derived factors. In the future, additional efforts towards systematic validation of these factors could be explored.

Conclusions

In this study, we characterized and quantified variabilities of publicly available oral LD_{50s} as well as LOAELs and NOAELs from the ATSDR's toxicological profiles. The effects of experimental species, target organs, or availability of MRLs were also explored. The derived SDs were back-transformed to obtain scaling factors to derive lower bounds on respective endpoints.

The major findings from this study are listed in the form of "take home" points below:

- The variability of LD_{50}s was approximately half the variability of LOAELs or NOAELs. The variability of LOAELs was lower than or similar to that of NOAELs.
- Approximately, the scaling factors to derive lower bounds on an endpoint were: 2.7 for LD_{50}s, 7 for LOAELs (all three exposure durations), and 7 (acute exposure duration) or 10 (intermediate and chronic exposure durations) for NOAELs.
- Test species had no major influence on variability (except intermediate duration endpoints) whereas matching target systems lowered NOAELs variability.
- The variability of endpoints did not necessarily decrease with the increase in number of studies.

References

1. ATSDR. (2018). MRL Information for the General Public. Retrieved from https://www.atsdr.cdc.gov/mrls/index.asp
2. EPA. (2012). Benchmark Dose Technical Guidance. Retrieved from https://www.epa.gov/sites/production/files/2015-01/documents/benchmark_dose_guidance.pdf
3. Siwakoti, R. Prussia., A; Demchuk, G. (2017). Estimating Variability of a chemical LD50. Society of Toxicology.
4. Rosner, B. (2015). Fundamentals of biostatistics: Nelson Education.
5. Rusyn, I., & Daston, G. P. (2010). Computational toxicology: realizing the promise of the toxicity testing in the 21st century. Environ Health Perspect, 118(8), 1047-1050. doi:10.1289/ehp.1001925
6. Hoffmann, S., Kinsner-Ovaskainen, A., Prieto, P., Mangelsdorf, I., Bieler, C., & Cole, T. (2010). Acute oral toxicity: variability, reliability, relevance and interspecies comparison of rodent LD50 data from literature surveyed for the ACuteTox project. Regul Toxicol Pharmacol, 58(3), 395-407. doi:10.1016/j.yrtph.2010.08.004
7. Karmaus, A. F., J; Allen, D; Patlewicz, G; Kleinstreuer, N; Casey, W. (2018). Variability of LD50 Values from Rat Oral Acute Toxicity Studies: Implications for Alternative Model Development. Retrieved from https://ntp.niehs.nih.gov/iccvam/meetings/sot18/karmaus-poster.pdf
8. Kleinstreuer, N. C., Karmaus, A., Mansouri, K., Allen, D. G., Fitzpatrick, J. M., & Patlewicz, G. (2018). Predictive Models for Acute Oral Systemic

Toxicity: A Workshop to Bridge the Gap from Research to Regulation. Comput Toxicol, 8(11), 21-24. doi:10.1016/j.comtox.2018.08.002TOXNET Toxicology Data Network. Available from National Library of Medicine ChemIDplus TOXNET Retrieved 1/28/2019.

9. TOXNET Toxicology Data Network. Available from National Library of Medicine ChemIDplus TOXNET Retrieved 1/28/2019 https://chem.nlm.nih.gov/chemidplus/
10. RTECS® Search. Available from Canadian Centre for Occupational Health and Safety RTECS® Search http://ccinfoweb.ccohs.ca/rtecs/search.html
11. Borchard, R. E., et al. (1990). Drug dosage in laboratory animals: a handbook. Caldwell, N.J., Telford Press.
12. Guidance for the Preparation of Toxicological Profiles. (2018). ATSDR Retrieved from https://www.atsdr.cdc.gov/toxprofiles/guidance/profile_development_guidance.pdf
13. Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. (2014). Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry, 105-109.
14. Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass., Addison-Wesley Pub. Co.
15. Shi, L., & Daniel Roth, H. (1994). The bias of the commonly-used estimate of variance in meta-analysis AU - Li, Yuanzhang. *Communications in Statistics - Theory and Methods*, 23(4), 1063-1085. doi:10.1080/03610929408831305

16. Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2): Duxbury Pacific Grove, CA.
17. John Gurland & Ram C. Tripathi (1971) A Simple Approximation for Unbiased Estimation of the Standard Deviation, The American Statistician, 25:4, 30-32, DOI: 10.1080/00031305.1971.10477279
18. Canty, A. J. J. T. N. o. t. R. P. V. (2002). Resampling methods in R: the boot package. 2, 3.
19. Haynes, W. J. E. o. s. b. (2013). Wilcoxon rank sum test. 2354-2355.
20. Zarn, J. A., & O'Brien, C. D. J. A. o. T. (2018). Current pesticide dietary risk assessment in light of comparable animal study NOAELs after chronic and short-termed exposure durations. 92(1), 157-167. doi:10.1007/s00204-017-2052-4
21. Cornbleet, P. J., & Gochman, N. (1979). Incorrect least-squares regression coefficients in method-comparison analysis. Clin Chem, 25(3), 432-438.

Appendix A: Tables

Table 1a. Total number of chemicals and corresponding LD_{50}s categorized by their database sources. Highlights in red represent the sample sizes used in this study.

	Total number of			
Source	**LD_{50}s originally extracted**	**Unique CASRN**	**Chemicals with >=3 LD_{50}s**	**LD_{50}s (used in the study)**
RTECS	959	349	161	
ChemID	1619	943	95	
LSE Table	689	182	79	
Chemical handbook	30	10	9	
Total (after applying selection criteria)				
Species mixed (full) dataset	**2849**	**1339**	**306**	**1420**
Species matched subset	**-**	**-**	**70**	**404**

Table 1b. Total number of unique chemicals and corresponding L(N)OAELs for acute, intermediate, and chronic exposure durations after applying selection criteria.

Exposure Duration	Endpoints	Unique CASRN	Number of endpoints
Acute	**LOAELs**	93	740
	NOAELs	104	848
Intermediate	**LOAELs**	121	1209
	NOAELs	132	1529
Chronic	**LOAELs**	58	368
	NOAELs	75	547

Table 2. Frequencies of experimental species used to derive $LD_{50}s$. [a]Although large number of rabbits are present in species mixed dataset, they are not present in species matched subset because >= 3 observations for same test species were necessary for a chemical to be included.

Species	**species mixed (n=1420)**	**species matched (n=404)**
Cat	30	-
Dog	77	-
Guinea pig	151	7
Hamster	16	-
Monkey	5	-
Mouse	379	76
Other	10	-
Pig	3	-
Rabbit[a]	158	-
Rat	591	321

Table 3a. Frequencies of experimental species used to derive L(NOAELs) for chemicals with >= 3 respective endpoints.

	Acute		**Intermediate**		**Chronic**	
Species	**LOAEL (n=587)**	**NOAEL (n=682)**	**LOAEL (n=1017)**	**NOAEL (n=1365)**	**LOAEL (n=26 5)**	**NOAEL (n=422)**
Dog	3	-	16	42	15	32
Human	40	28	10	3	3	3
Mouse	141	149	198	350	49	104
Other	3	8	20	49	-	7
Rabbit	5	26	3	5	3	-
Rat	395	462	751	884	179	265
Gn pig	-	3	-	3	-	-
Hamster	-	6	-	-	-	-
Monkey	-	-	19	29	16	11

Table 3b. Total number of unique CASRN with >= 3 LOAELs or NOAELs when experimental species were matched, i.e., a chemical was only included if >= 3 observations were available for a given species. It was possible for a chemical to be included more than once if multiple species with >= 3 observations were present.

	Acute		Intermediate		Chronic	
Species	**LOAEL (n=104)**	**NOAEL (n=123)**	**LOAEL (n=148)**	**NOAEL (n=197)**	**LOAEL (n=61)**	**NOAEL (n=93)**
Dog	1	-	4	9	5	9
Human	7	7	3	1	1	1
Mouse	29	30	36	60	13	25
Other	1	1	4	9	-	2
Rabbit	1	8	1	1	1	-
Rat	65	75	96	108	39	54
Gn pig	-	1	-	1	-	-
Hamster	-	1	-	-	-	-
Monkey	-	-	4	8	2	2

Table 3c. Frequency distribution of organ systems used to derive L(NOAELs) for chemicals with >= 3 L(N)OAELs. Systems listed as "NA" or "NS" were replaced with non-NA entries from the corresponding category column. If both systems and categories were listed as "NA", "NS", or "NR", then those records were excluded.

	Acute		Intermediate		Chronic	
System	**LOAEL (n=1010)**	**NOAEL (n=1575)**	**LOAEL (n=1799)**	**NOAEL (n=3847)**	**LOAEL (n=607)**	**NOAEL (n=2282)**
Body weight change	117	266	243	530	103	221
Cardiovascular	15	60	23	208	11	185
Dermal	14	33	31	111	19	133
Developmental	87	162	131	158	4	7
Endocrine	50	55	91	200	31	139
Gastrointestinal	58	67	66	206	35	174
Hematological	50	73	152	299	52	179
Hepatic	254	204	383	396	146	221
Immunological/lymphoreticular	55	54	130	183	23	93
Metabolic	8	8	19	26	1	7
Musculoskeletal	9	30	19	133	9	140
Neurological	134	183	163	307	46	145
Ocular	10	24	14	111	13	105
Renal	75	146	171	380	67	203
Reproductive	46	127	128	401	25	153
Respiratory	28	82	35	198	22	177
Genotoxic	-	1	-	-	-	-

Table 3d. Total number of unique CASRN with >= 3 LOAELs or NOAELs when target systems were matched, i.e., a chemical was only included if >= 3 observations were available for a given system. It was possible for a chemical to be included more than once if multiple systems with >= 3 observations were present.

	Acute		Intermediate		Chronic	
System	**LOAEL (n=103)**	**NOAEL (n=189)**	**LOAEL (n=219)**	**NOAEL (n=542)**	**LOAEL (n=65)**	**NOAEL (n=322)**
Body weight change	11	39	33	70	14	35
Dermal	1	2	2	13	1	17
Developmental	11	19	19	17	-	-
Endocrine	6	5	13	31	1	20
Gastrointestinal	4	7	4	28	3	21
Hematological	4	8	20	43	5	29
Hepatic	31	29	51	60	19	33
Immunological/lymphoreticular	4	3	12	30	2	9
Metabolic	1	-	-	3	-	1
Neurological	18	29	21	46	7	19
Renal	7	21	21	57	7	29
Reproductive	5	12	15	56	1	24
Cardiovascular	-	6	-	30	-	26
Respiratory	-	9	3	30	2	25
Musculoskeletal	-	-	3	15	2	18
Ocular	-	-	2	13	1	16

Table 3e. Total number chemicals categorized by the availability of MRLs (only chemicals with >=3 L(N)OAELs were included).

	Acute		Intermediate		Chronic	
MRL	**LOAEL (n=93)**	**NOAEL (n=104)**	**LOAEL (n=121)**	**NOAEL (n=132)**	**LOAEL (n=58)**	**NOAEL (n=75)**
present	45	49	71	78	35	43
not present	48	55	50	54	23	32

Table 3f. SW test of normality of endpoints (chemical mean centered log-molar endpoints). [a]When extreme points identified by Tukey fence analysis were excluded, p-value increased to 0.0051 for species mixed and 0.0175 for species matched datasets.

Dataset/Duration	Dataset/Endpoint	W Statistic	p-value
LD$_{50}$s	Species mixed	0.97	<0.0001[a]
	Species matched	0.95	<0.0001[a]
Acute (full)	LOAEL	0.97	<0.0001
	NOAEL	0.96	<0.0001
Intermediate (full)	LOAEL	0.97	<0.0001
	NOAEL	0.97	<0.0001
Chronic (full)	LOAEL	0.97	<0.0001
	NOAEL	0.99	0.0308

Table 4a. Output from the Levene's and F-tests for the homogeneity of variances across different samples.

Description	Df	F-value	Pr(>F)
Levene's test group			
LD$_{50}$s species mixed	305; 1112	1.285	0.00238
Acute LOAELs	92; 647	1.935	<0.0001
Acute NOAELs	103; 744	1.962	<0.0001
Intermediate LOAELs	120; 1088	2.098	<0.0001
Intermediate NOAELs	131; 1397	1.768	<0.0001
Chronic LOAELs	57; 310	1.505	0.01599
Chronic NOAELs	74; 472	1.4157	0.01831
F-test groups (LD$_{50}$s)			
Malathion and Dioxin	24 (num), 18 (denom)	0.369	0.02393
Malathion and Parathion	24 (num), 16 (denom)	0.724	0.46190
Malathion and Disulfoton	24 (num), 18 (denom)	4.202	0.00285
Dioxin and Parathion	18 (num), 16 (denom)	1.958	0.18280
Dioxin and Disulfoton	18 (num), 18 (denom)	11.365	<0.0001
Parathion and Disulfoton	16 (num), 18 (denom)	5.805	0.000597

Table 4b. SD estimates for LD_{50}s for species mixed and species matched datasets. The 95% CIs were approximated using the BCa method with 10,000 bootstrap replications as implemented in the "boot" package in R.

Description	**species mixed**	**species matched**
Number of LD_{50}s (n)	1420	404
Number of unique chemicals (N)	306	70
Standard Deviation (SD)		
25th percentile (95% CI)	0.17 (0.15-0.18)	0.17 (0.11-0.20)
Median (95% CI)	0.26 (0.23-0.28)	0.27 (0.22-0.29)
Mean (95% CI)	0.34 (0.32-0.36)	0.35 (0.30-0.41)
75th percentile (95% CI)	0.37 (0.31-0.44)	0.36 (0.31-0.45)

Table 5a. SD estimates for LOAELs and NOAELs. All available endpoints were used, i.e., distributions were based on full datasets. However, if an identical dose was recorded for multiple systems, it was included only once.

L(N)OAELs – full datasets						
	Acute		**Intermediate**		**Chronic**	
Description	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**
Number of endpoints (n)	740	848	1209	1529	368	547
Number of unique chemicals (N)	93	104	121	132	58	75
Standard deviation						
25th percentile (95% CI)	0.43(0.33-0.44)	0.35(0.32-0.44)	0.40(0.33-0.45)	0.47(0.43-0.51)	0.32(0.2-0.38)	0.49(0.35-0.55)
Median (95% CI)	0.53(0.46-0.6)	0.52(0.48-0.60)	0.55(0.50-0.62)	0.62(0.58-0.68)	0.49(0.37-0.58)	*0.63(0.58-0.68)*
Mean (95% CI)	0.73(0.65-0.84)	0.68(0.60-0.78)	0.66(0.61-0.72)	0.73(0.68-0.78)	0.63(0.55-0.71)	0.73(0.66-0.81)
75th percentile (95% CI)	0.77(0.65-0.89)	0.71(0.63-0.84)	0.72(0.68-0.84)	0.84(0.77-0.93)	0.77(0.60-0.91)	0.76(0.71-0.94)

Table 5b. SD estimates for LOAELs and NOAELs. For each chemical, endpoints were categorized by their test species and if >= 3 observations were available for any given species, SD corresponding to it was estimated and used in the analysis.

L(N)OAELs – species matched subsets						
	Acute[a]		**Intermediate**		**Chronic**[a]	
Description	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**
Number of endpoints (n)	587	682	1017	1365	265	422
Number of unique chemicals (N_1)	73	84	103	123	43	60
Number of samples (N_2)	104	123	148	197	61	93
Standard deviation						
25th percentile (95% CI)	0.36(0.32-0.41)	0.35(0.31-0.42)	0.34(0.3-0.38)	0.38(0.34-0.43)	0.28(0.21-0.34)	0.34(0.30-0.42)
Median (95% CI)	0.48(0.44-0.54)	0.49(0.44-0.54)	0.46(0.42-0.51)	0.53(0.50-0.58)	0.45(0.34-0.58)	0.58(0.50-0.63)
Mean (95% CI)	0.60(0.55-0.66)	0.62(0.56-0.70)	0.59(0.54-0.65)	0.68(0.63-0.74)	0.65(0.55-0.77)	0.67(0.61-0.76)
75th percentile (95% CI)	0.66(0.59-0.75)	0.67(0.59-0.78)	0.65(0.60-0.71)	0.77(0.68-0.83)	0.67(0.59-0.88)	0.77(0.67-0.84)

Table 5c. SD estimates for LOAELs and NOAELs for all three exposure durations. For each chemical, endpoints were categorized by their target systems and if >= 3 observations were available for any given system, SD corresponding to it was estimated and used in the analysis.

L(N)OAELs – target systems matched subsets

	Acute		Intermediate		Chronic	
Description	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**
Number of endpoints (n)	519	830	1155	2618	278	1275
Number of unique chemicals (N_1)	58	67	83	97	36	54
Number of samples (N_2)	103	189	219	542	65	322
Standard deviation						
25th percentile (95% CI)	0.32(0.26-0.36)	0.25(0.23-0.30)	0.32(0.28-0.35)	0.31(0.27-0.34)	0.25(0.17-0.31)	0.32(0.26-0.35)
Median (95% CI)	0.51(0.43-0.56)	0.41(0.36-0.47)	0.50(0.45-0.53)	0.49(0.46-0.54)	0.44(0.37-0.54)	0.49(0.45-0.52)
Mean (95% CI)	0.66(0.58-0.76)	0.56(0.52-0.63)	0.64(0.60-0.70)	0.64(0.61-0.70)	0.62(0.54-0.73)	0.61(0.57-0.65)
75th percentile (95% CI)	0.74(0.64-0.81)	0.64(0.59-0.70)	0.73(0.65-0.81)	0.67(0.62-0.70)	0.66(0.56-0.89)	0.69(0.65-0.77)

Table 6a. Outputs from the KS test for the comparison of distribution of variances between different species. 3 species with the highest number of observations were selected for comparisons. [a]Not enough sample size for other comparisons.

Dataset	KS test species pairs	n	D	p-value
LD_{50}s species mixed	Rat and Mouse	66; 19	0.193	0.5654
Acute LOAELs	Rat and Mouse	65; 29	0.183	0.51261
	Rat and Human	65; 7	0.448	0.11131
	Mouse and Human	29; 7	0.330	0.47801
Acute NOAELs	Rat and Mouse	75; 30	0.213	0.28362
	Rat and Rabbit	75; 8	0.217	0.88652
	Mouse and Rabbit	30; 8	0.358	0.31335
Intermediate LOAELs[a]	Rat and Mouse	96; 36	0.139	0.69341
Intermediate NOAELs	Rat and Mouse	108; 60	0.267	0.00829
	Rat and Dog	108; 9	0.306	0.41992
	Dog and Mouse	9; 60	0.394	0.17503
Chronic LOAELs	Rat and Mouse	39; 13	0.231	0.67676
	Rat and Dog	39; 5	0.446	0.24871
	Dog and Mouse	5; 13	0.523	0.21078
Chronic NOAELs	Rat and Mouse	91; 58	0.208	0.44958
	Rat and Dog	91; 26	0.185	0.93005
	Dog and Mouse	26; 58	0.258	0.65833

Table 6b. Outputs from the KS test for the comparison of distribution of variances between different target systems. 3 systems with the highest number of observations were selected for comparisons. [a]When two largest systems with identical sample size were present, both systems were included.

Dataset	KS test groups	n	D	p-value
Acute LOAELs [a]	Hepatic; Developmental	31; 11	0.139	0.90863
	Developmental; Neurological	11; 18	0.19	0.76274
	Hepatic; Neurological	31; 18	0.216	0.42814
	Hepatic; Body weight change	31; 11	0.237	0.17516
	Developmental; Body weight change	11; 11	0.26	0.32037
	Neurological; Body weight change	18; 11	0.29	0.18666
Acute NOAELs	Hepatic; Body weight change	29; 39	0.209	0.46381
	Hepatic; Neurological	29; 29	0.138	0.95144
	Body weight change; Neurological	39; 29	0.183	0.55879
Intermediate LOAELs[a]	Hepatic; Neurological	51; 21	0.216	0.42814
	Neurological; Body weight change	21; 33	0.29	0.18666
	Hepatic; Body weight change	51; 33	0.237	0.17516
	Renal; Hepatic	21; 51	0.249	0.26281
	Renal; Neurological	21; 21	0.333	0.19631
	Renal; Body weight change	21; 33	0.152	0.88031
Intermediate NOAELs	Hepatic; Body weight change	60; 70	0.233	0.05932
	Hepatic; Renal	60; 57	0.14	0.61235
	Body weight change; Renal	70; 57	0.158	0.41575
Chronic LOAELs[a]	Hepatic; Body weight change	19; 14	0.718	0.00016
	Body weight change; Neurological	14; 7	0.214	0.98119
	Body weight change; Renal	14; 7	0.643	0.03313
	Hepatic; Neurological	19; 7	0.504	0.10479
	Hepatic; Renal	19; 7	0.293	0.65557
	Neurological; Renal	7;7	0.571	0.21212
Chronic NOAELs[a]	Hepatic; Body weight change	33; 35	0.229	0.33295
	Body weight change; Renal	35; 29	0.139	0.91955
	Hepatic; Renal	33; 29	0.119	0.98083
	Hepatic; Hematological	33; 29	0.167	0.78132
	Body weight change; Hematological	35; 29	0.335	0.04282
	Renal; Hematological	29; 29	0.241	0.36685

Table 7a. Outputs from the KS test for the comparison of distribution of variances of endpoints based on the availability of their corresponding MRLs.

Dataset	KS test groups	n	D	p-value
Acute LOAEls	MRL present vs. MRL not present	45; 48	0.213	0.24505
Acute NOAELs	MRL present vs. MRL not present	49; 55	0.246	0.08685
Intermediate LOAELs	MRL present vs. MRL not present	71; 50	0.123	0.76316
Intermediate NOAELs	MRL present vs. MRL not present	78; 54	0.127	0.68419
Chronic LOAELs	MRL present vs. MRL not present	35; 23	0.125	0.98107
Chronic NOAELs	MRL present vs. MRL not present	43; 32	0.152	0.72054

Table 7b. SD estimates for LOAELs and NOAELs. Only endpoints from chemicals with corresponding MRLs were included.

L(N)OAELs – MRLs present subsets						
	Acute		**Intermediate**		**Chronic**	
Description	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**
Number of endpoints (n)	404	477	804	1045	249	352
Number of unique chemicals (N)	45	49	71	78	35	43
Standard deviation						
25th percentile (95% CI)	0.41(0.28-0.43)	0.34(0.27-0.45)	0.38(0.32-0.43)	0.49(0.41-0.52)	0.29(0.17-0.39)	0.51(0.37-0.60)
Median (95% CI)	0.49(0.43-0.58)	0.51(0.44-0.57)	0.54(0.47-0.62)	0.62(0.55-0.69)	0.50(0.33-0.65)	0.67(0.59-0.72)
Mean (95% CI)	0.60(0.52-0.68)	0.57(0.51-0.65)	0.61(0.56-0.67)	0.7(0.65-0.76)	0.64(0.53-0.75)	0.75(0.66-0.84)
75th percentile (95% CI)	0.71(0.58-0.78)	0.62(0.53-0.71)	0.70(0.64-0.82)	0.83(0.72-0.89)	0.82(0.6-0.95)	0.87(0.71-1.03)

Table 7c. SD estimates for LOAELs and NOAELs. Only endpoints from chemicals without corresponding MRLs were included.

L(N)OAELs – MRLs not present subsets						
	Acute		**Intermediate**		**Chronic**	
Description	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**	**LOAELs**	**NOAELs**
Number of endpoints (n)	336	371	405	484	119	195
Number of unique chemicals (N)	48	55	50	54	23	32
Standard deviation						
25th percentile (95% CI)	0.44(0.33-0.51)	0.42(0.25-0.47)	0.43(0.29-0.5)	0.47(0.37-0.52)	0.36(0.17-0.47)	0.46(0.29-0.56)
Median (95% CI)	0.59(0.49-0.68)	0.60(0.47-0.67)	0.58(0.5-0.67)	0.63(0.54-0.72)	0.49(0.37-0.55)	0.60(0.50-0.68)
Mean (95% CI)	0.83(0.70-1.00)	0.76(0.65-0.93)	0.72(0.62-0.85)	0.76(0.68-0.86)	0.61(0.52-0.75)	0.70(0.61-0.87)
75th percentile (95% CI)	0.89(0.67-1.20)	0.83(0.68-0.86)	0.77(0.64-0.86)	0.89(0.71-1.05)	0.66(0.52-0.97)	0.74(0.64-0.95)

Table 8. Scaling factors based on estimated median SDs to derive 95% lower bound on a geometric mean of endpoints or a single endpoint. For example, to obtain a lower bound on an acute LOAEL, a division factor of 7.40 could be applied.

Endpoint	Duration	Species mixed (full)	Species matched	Target systems matched
LD$_{50}$s		2.67 (2.38 - 2.88)	2.77 (2.38 - 2.99)	na
LOAELs	Acute	7.40 (5.68 - 9.64)	6.13 (5.27 - 7.68)	6.86 (5.07 - 8.29)
	intermediate	7.98 (6.61 - 10.39)	5.68 (4.88 - 6.86)	6.61 (5.47 - 7.40)
	Chronic	6.36 (4.04 - 8.94)	5.47 (3.61 - 8.94)	5.27 (4.04 - 7.68)
NOAELs	Acute	7.13 (6.13 - 9.64)	6.36 (5.27 - 7.68)	4.70 (3.89 - 5.90)
	Intermediate	10.39 (8.94 - 13.04)	7.40 (6.61 - 8.94)	6.36 (5.68 - 7.68)
	Chronic	10.79 (8.94 - 13.04)	8.94 (6.61 - 10.79)	6.36 (5.47 - 7.13)

Table 9. LD$_{50}$s for Acrylamide and the proposed lower bound based on the scaling factor of **2.7**

LD$_{50}$	1	2	3	4	5	6	7	8	9	10	11
species	**rat**	**rat**	**rat**	**rat**	**rat**	**rat**	**rat**	**mouse**	**mouse**	**gn. pig**	**rabbit**
Dose (mg/kg-day)	124	175	180	203	294	316	413	107	195	180	150
GM	**196.79**										
LD$_{50_L}$	**72.89**										

Appendix B: Figures

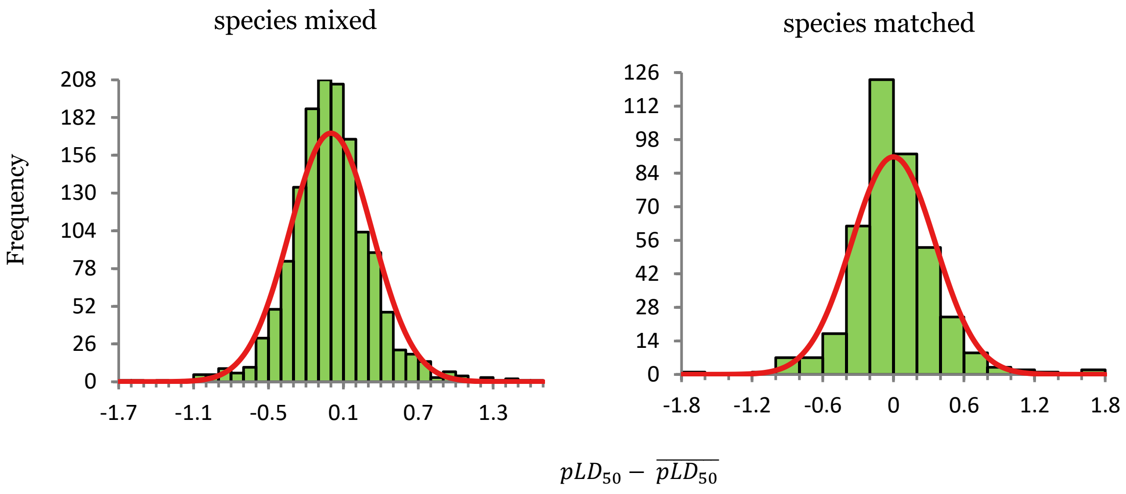


Figure 1a. Histograms representing the distributions of LD_{50}s (without removing outliers). Each point represents a mean-centered (by chemical specific mean) log-molar LD_{50}. Red curves represent the normal fit.

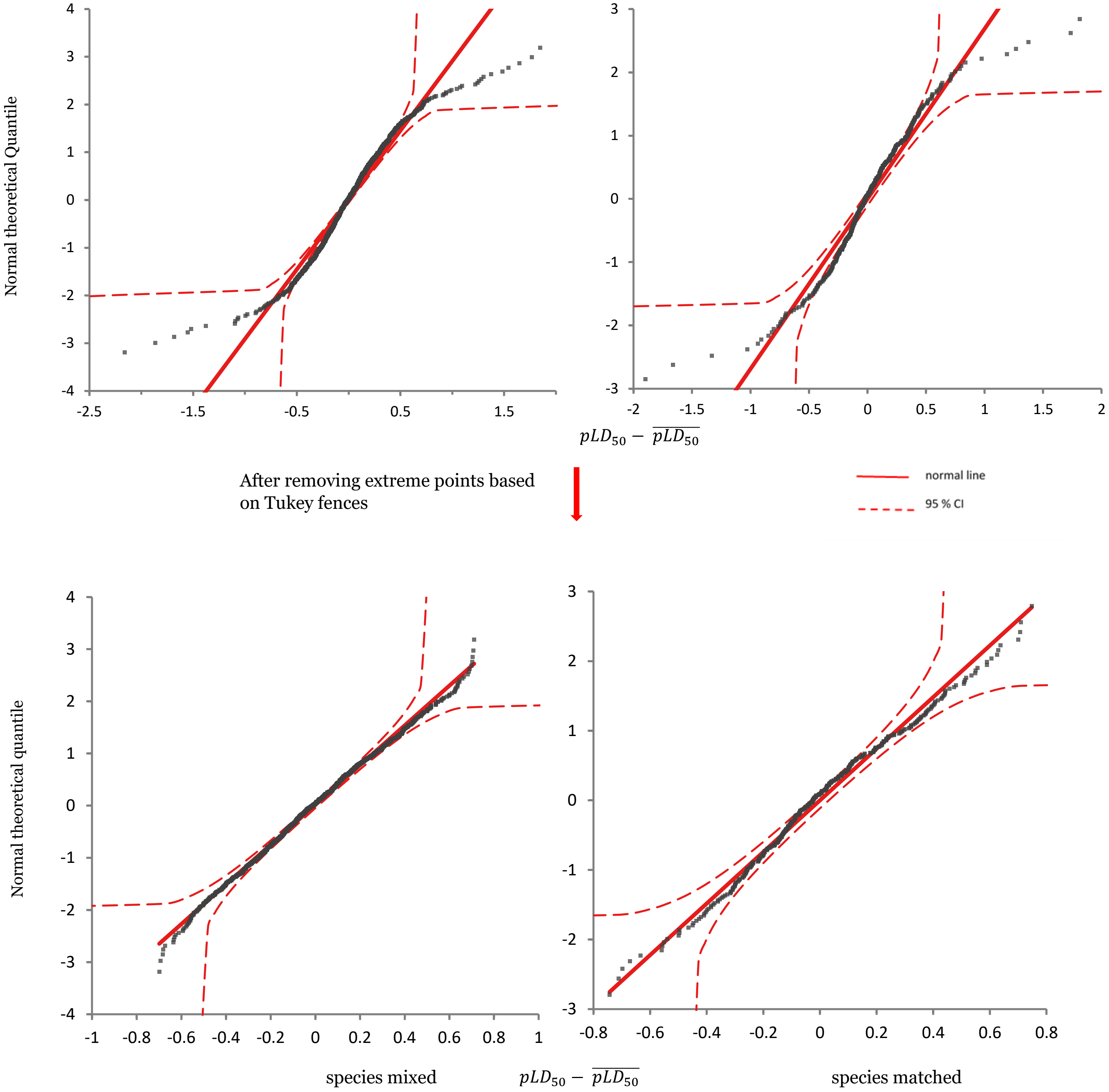


Figure 2b. Normal Q-Q plots for the distributions of LD_{50}s. Each point represents a mean-centered (by chemical specific mean) log-molar LD_{50}. Plots before and after removing outliers are provided.

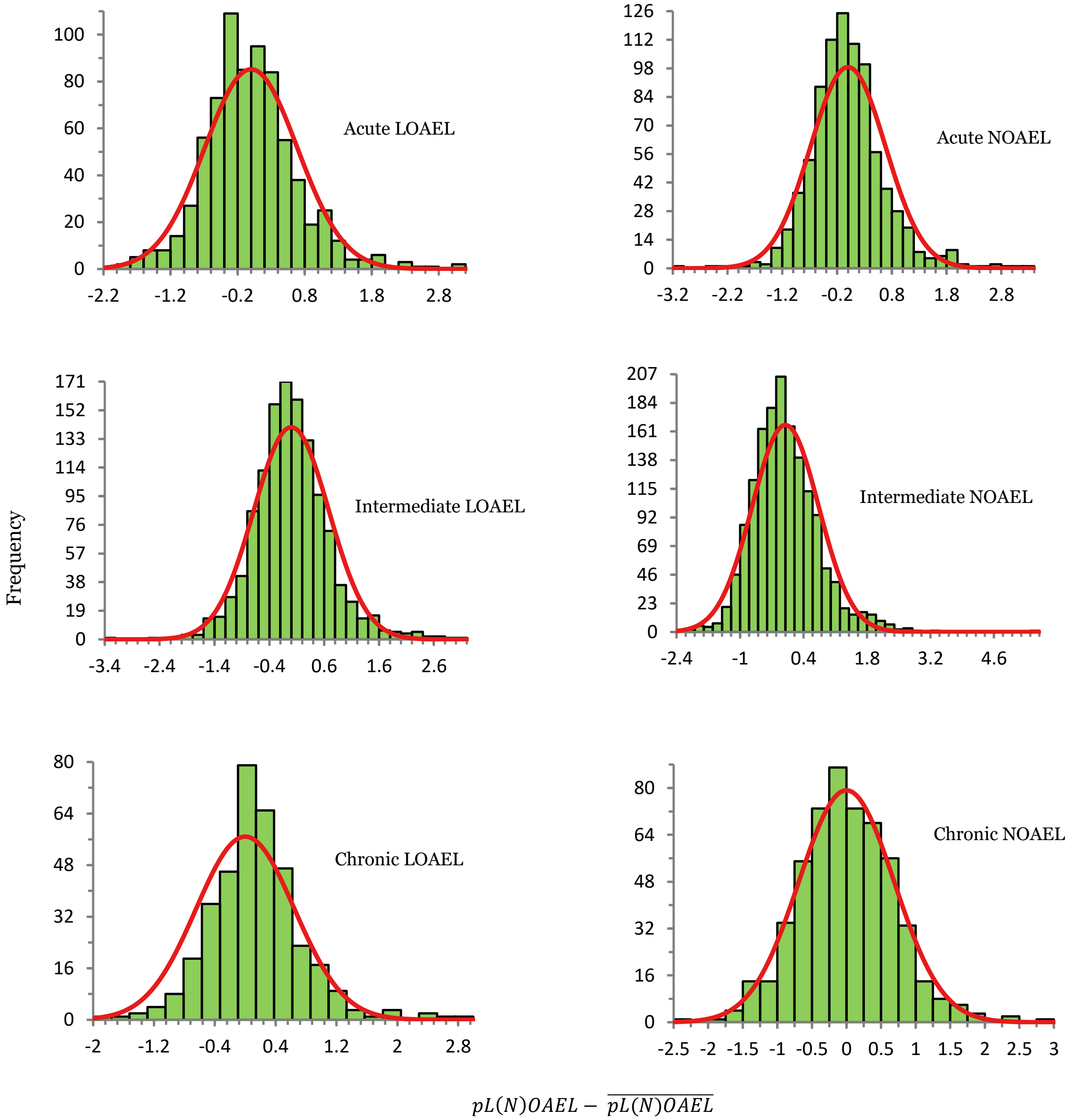


Figure 1c. Histograms representing the distributions of L(N)OAELs. Each point represents a mean-centered (by chemical specific mean) log-molar endpoint. The red curves represent the normal fit.

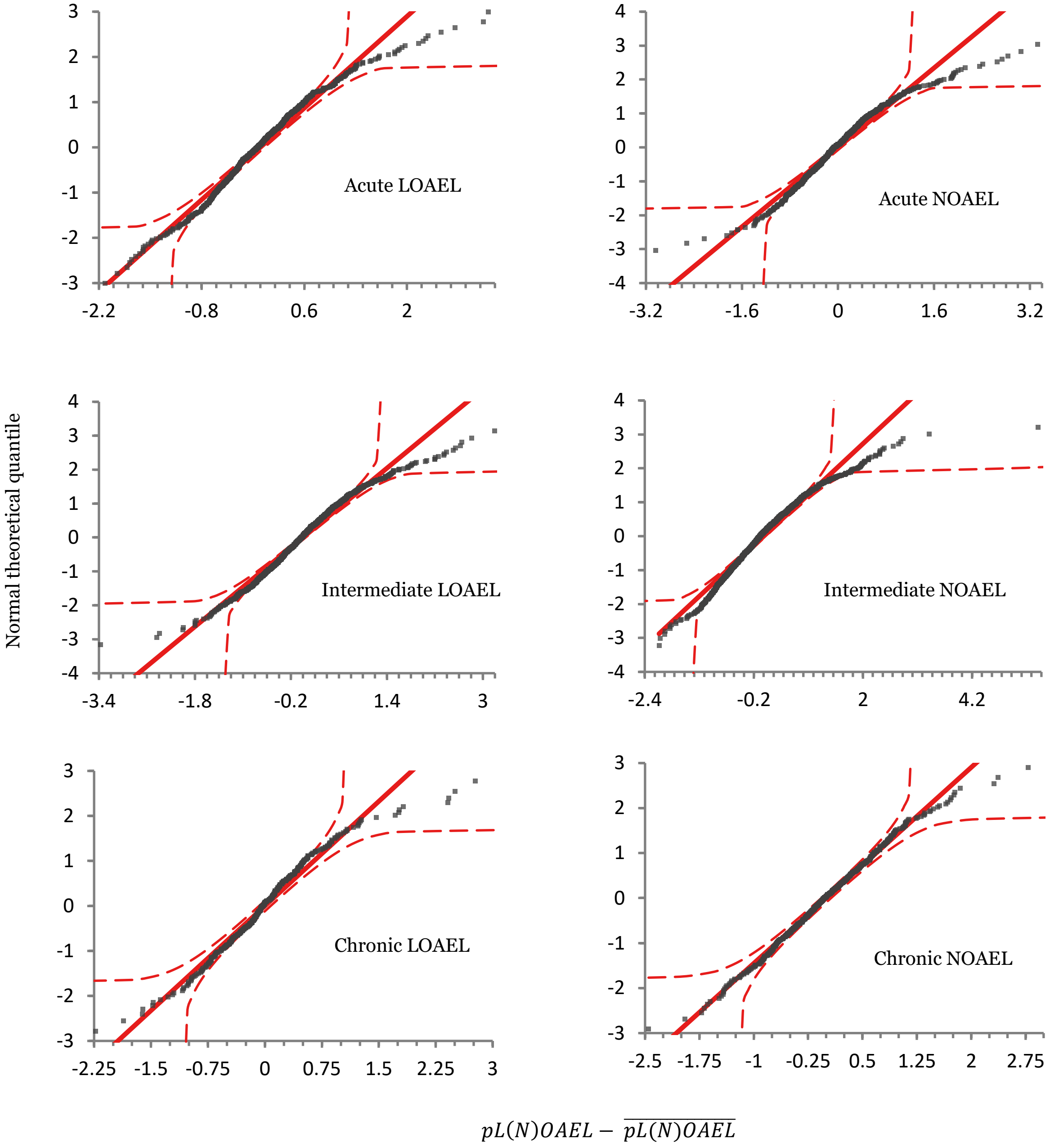


Figure 1d. Normal Q-Q plots for the distributions of L(N)OAELs. Each point represents a mean-centered (by chemical specific mean) log-molar endpoint.

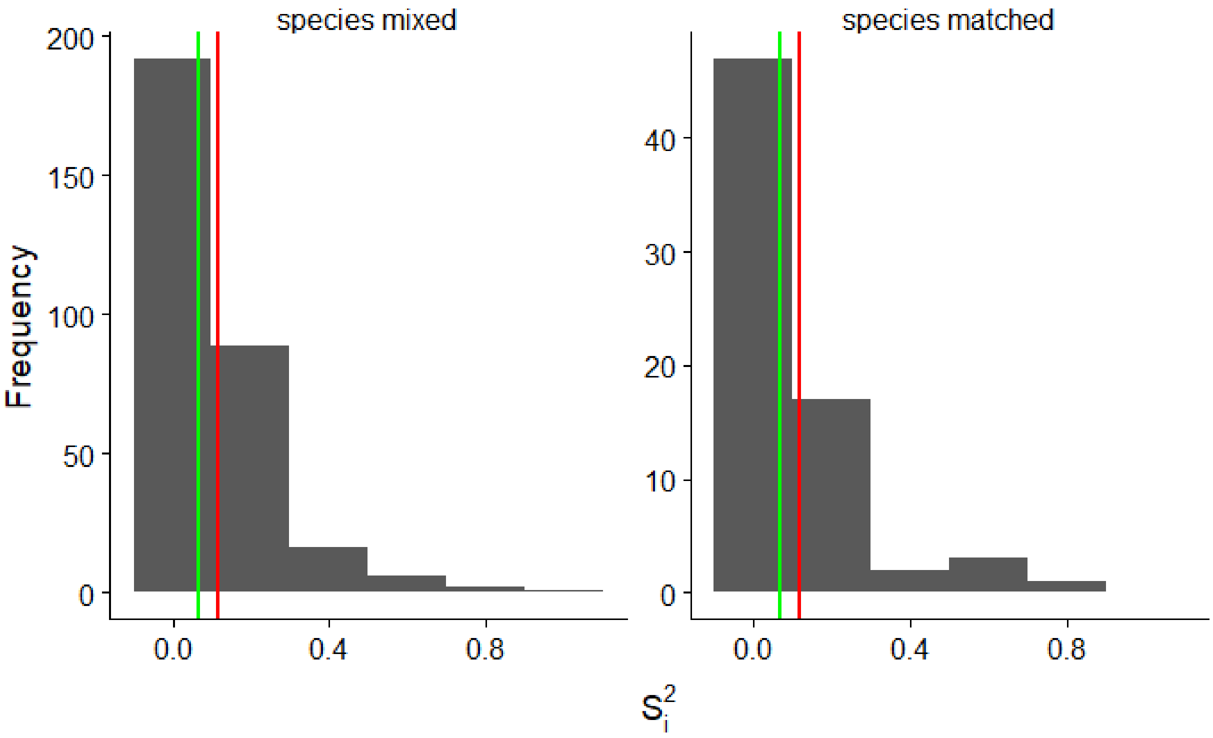


Figure 2a. Distributions of variances (S^2) of log-molar LD_{50}s. Green and red lines represent the median and mean of distributions, respectively.

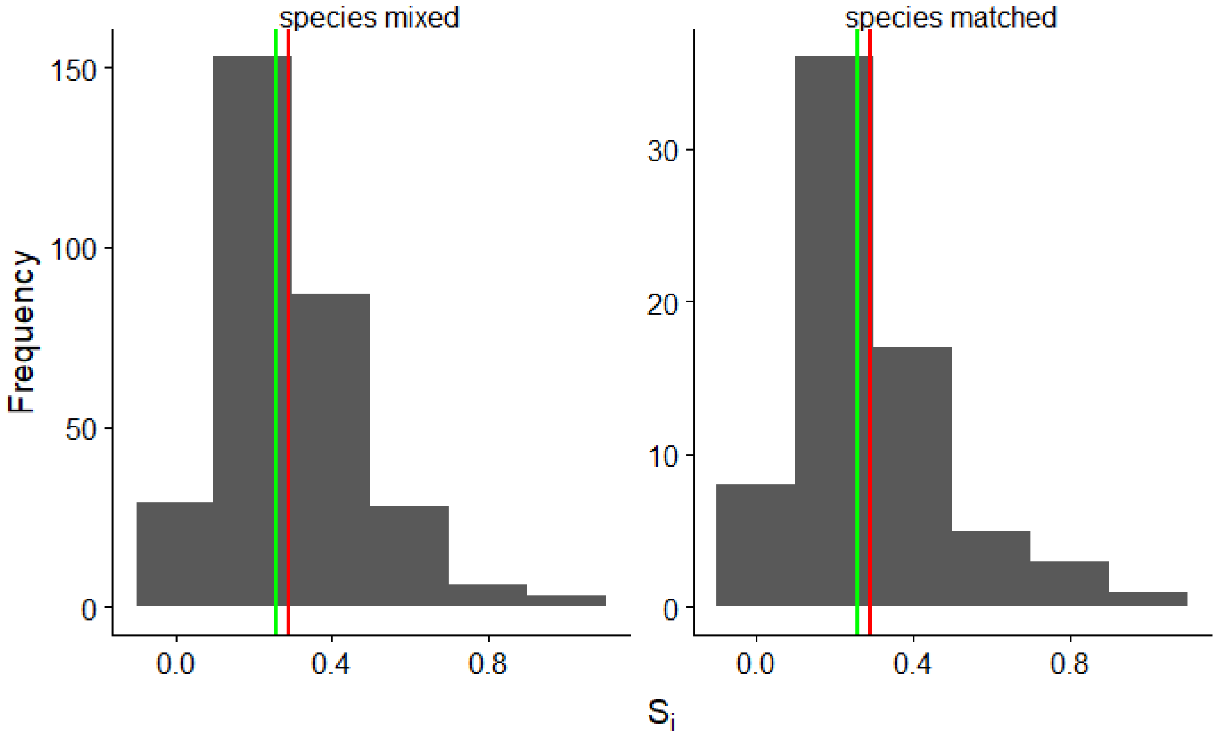


Figure 2b. Distributions of SDs (S) of log-molar LD_{50}s. Green and red lines represent the median and mean of distributions, respectively.

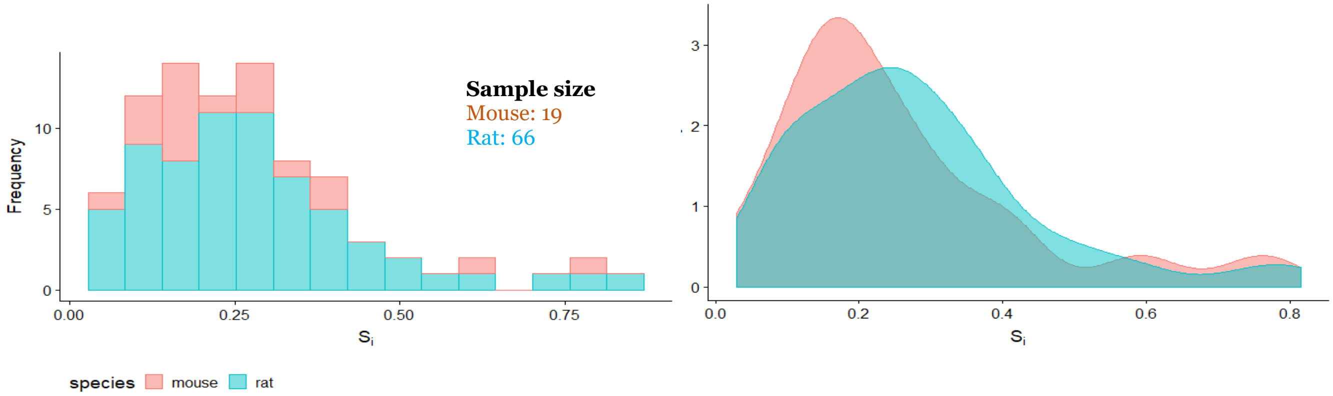


Figure 2c. Histograms and density plots of SDs for the comparison of variability between rat and mouse. The density plots were constructed using the default bandwidths in the R density function.

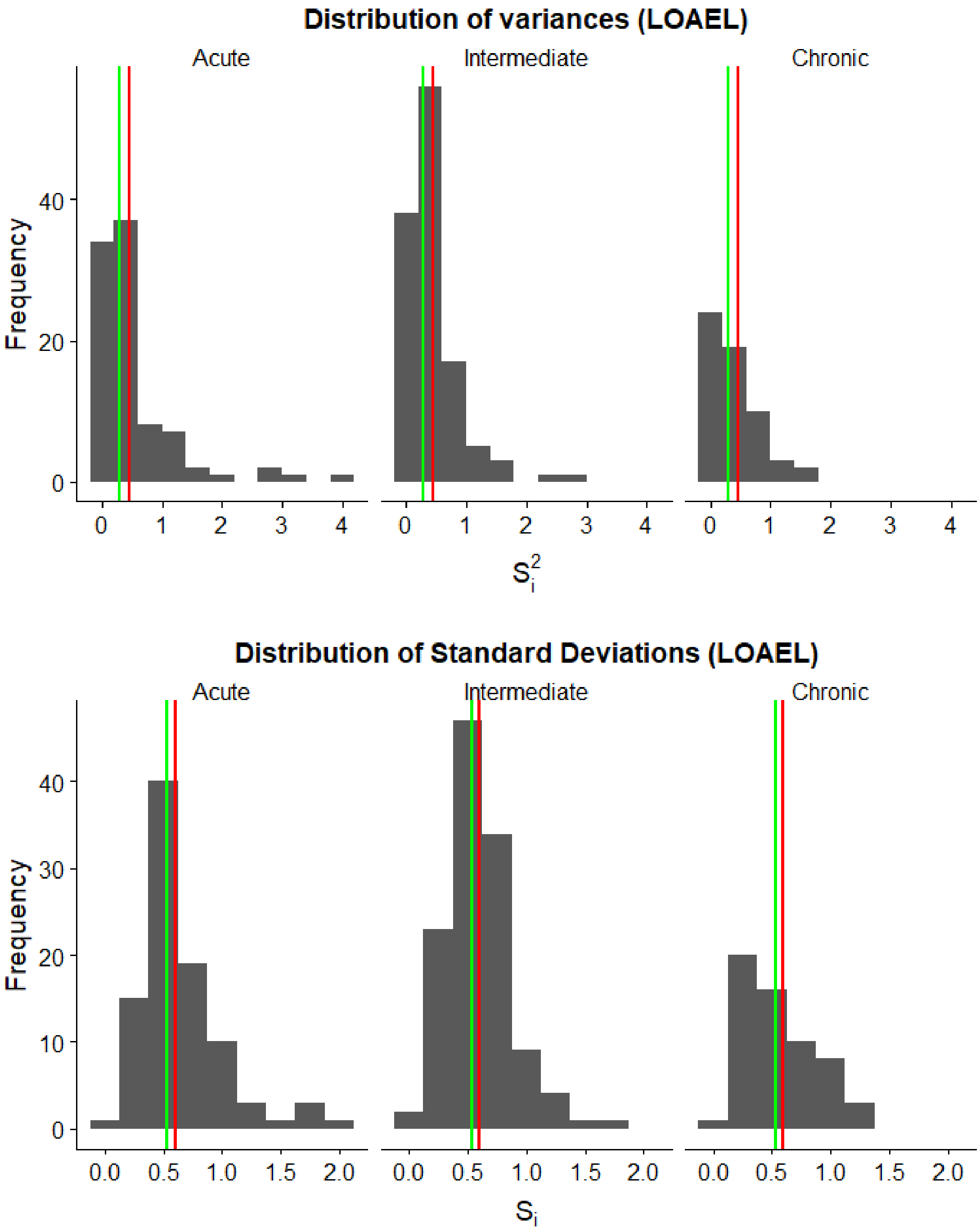


Figure 3a. Distributions of variances (S^2) and SDs (S) of log-molar LOAELs. Green and red lines represent the median and mean of distributions, respectively.

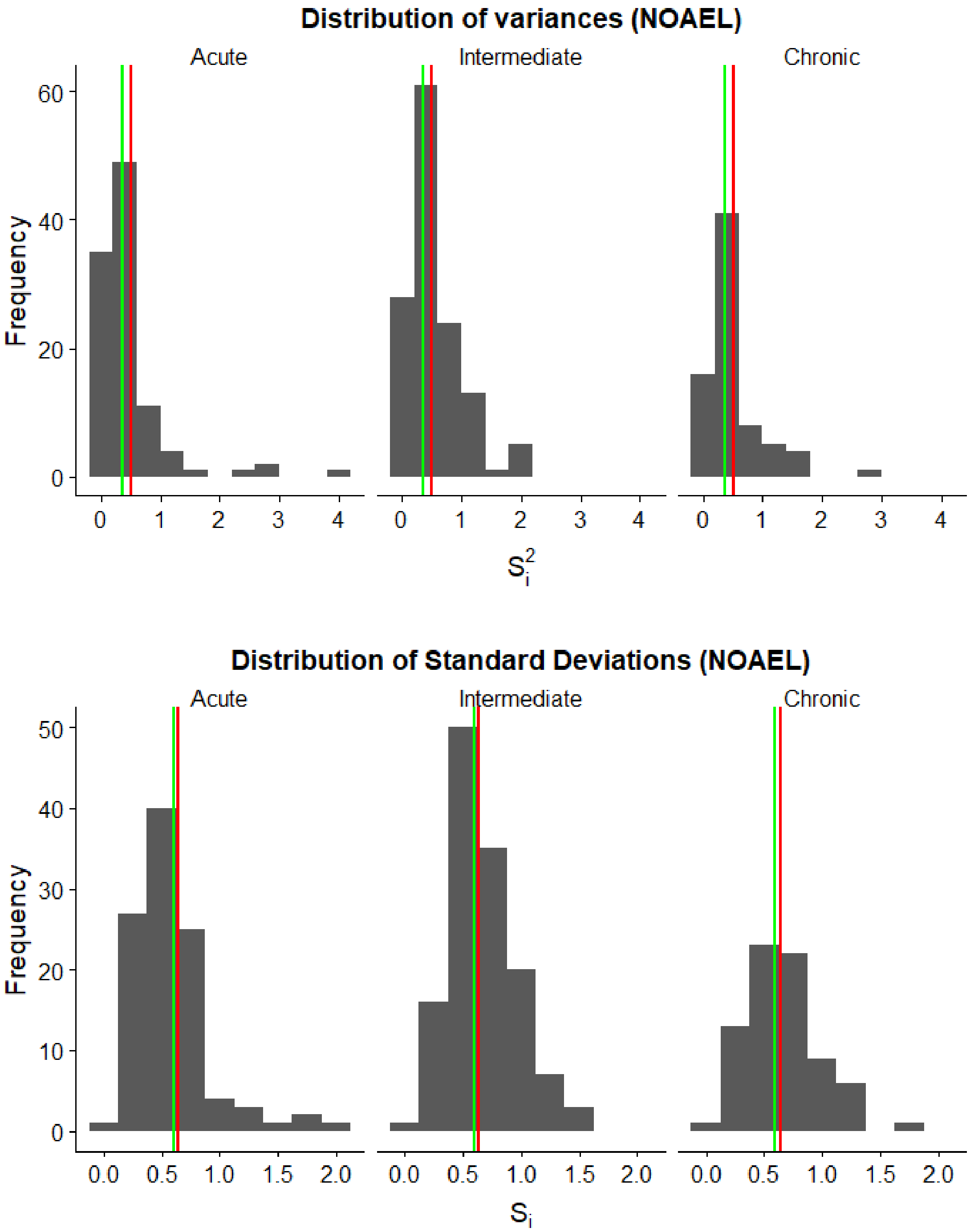


Figure 3b. Distributions of variances (S^2) and SDs (S) of log-molar NOAELs. Green and red lines represent the median and mean of distributions, respectively.

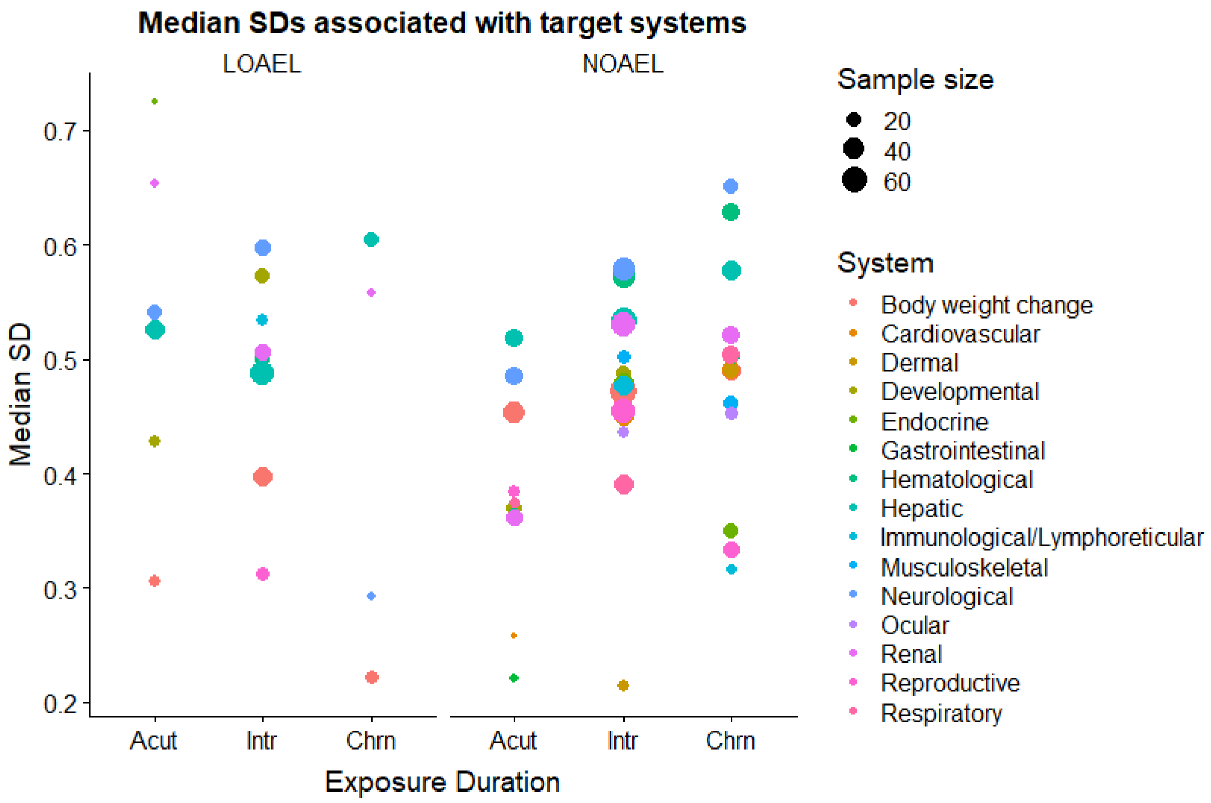
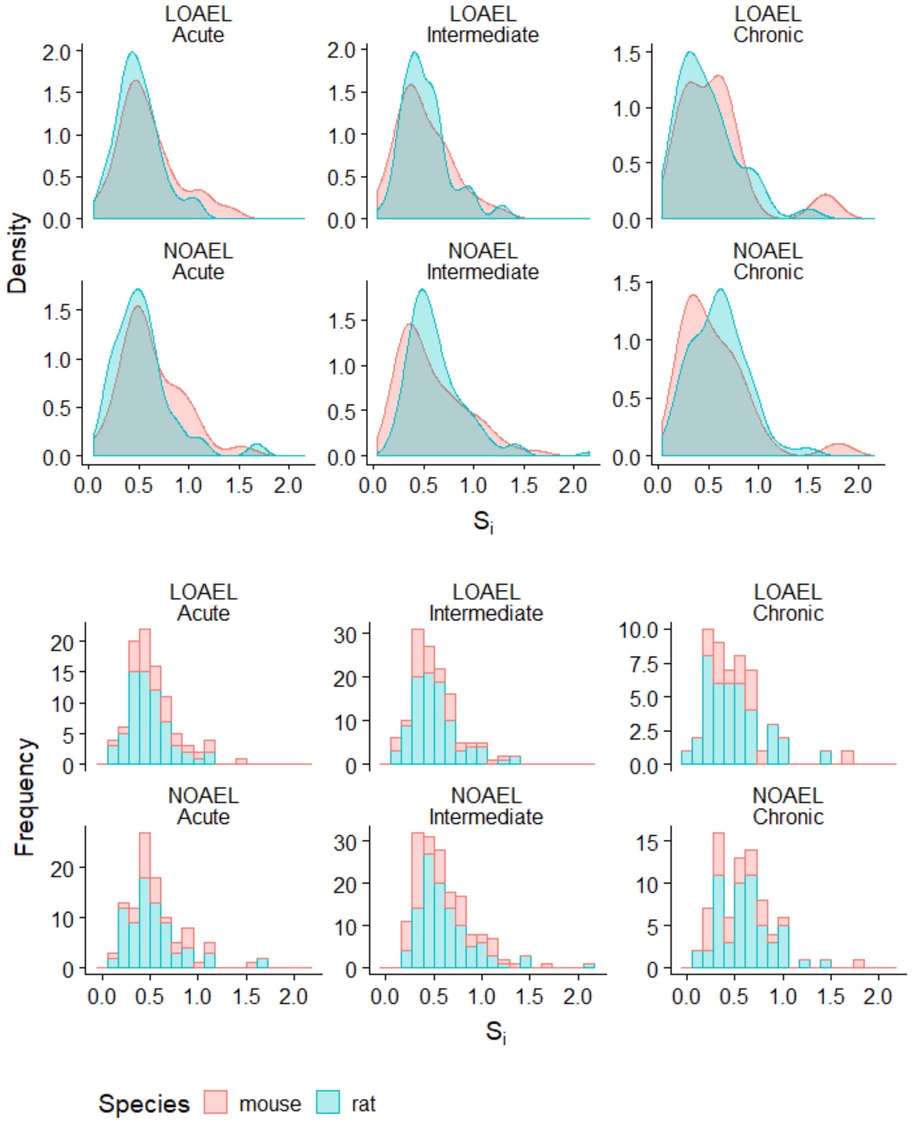
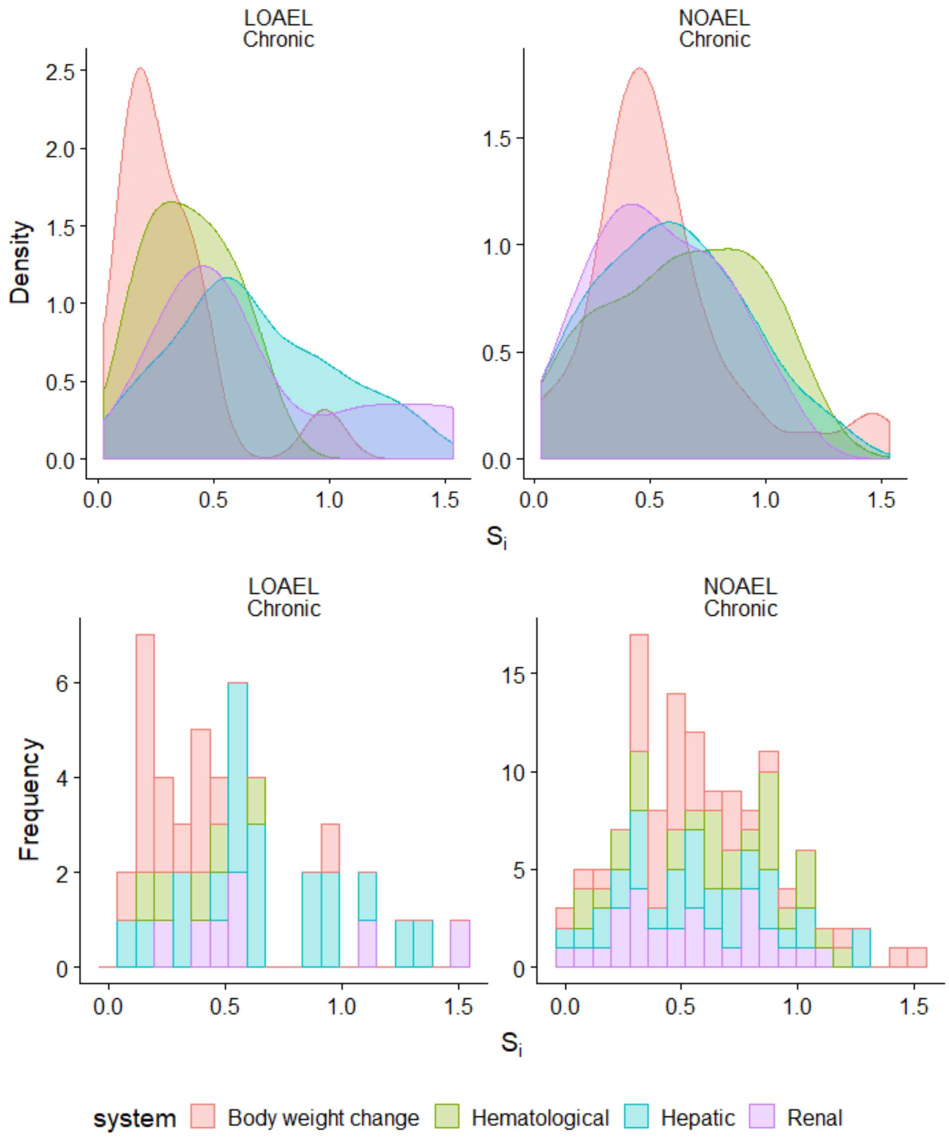


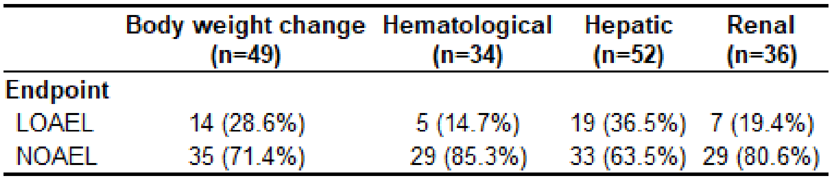
Figure 4a. Median SDs corresponding to endpoints based on different target systems. For better visualization, only systems >5 observations per endpoint per exposure duration were included.



	LOAEL		NOAEL	
	mouse (n=78)	rat (n=200)	mouse (n=115)	rat (n=237)
Duration				
Acute	29 (37.2%)	65 (32.5%)	30 (26.1%)	75 (31.6%)
Intermediate	36 (46.2%)	96 (48.0%)	60 (52.2%)	108 (45.6%)
Chronic	13 (16.7%)	39 (19.5%)	25 (21.7%)	54 (22.8%)

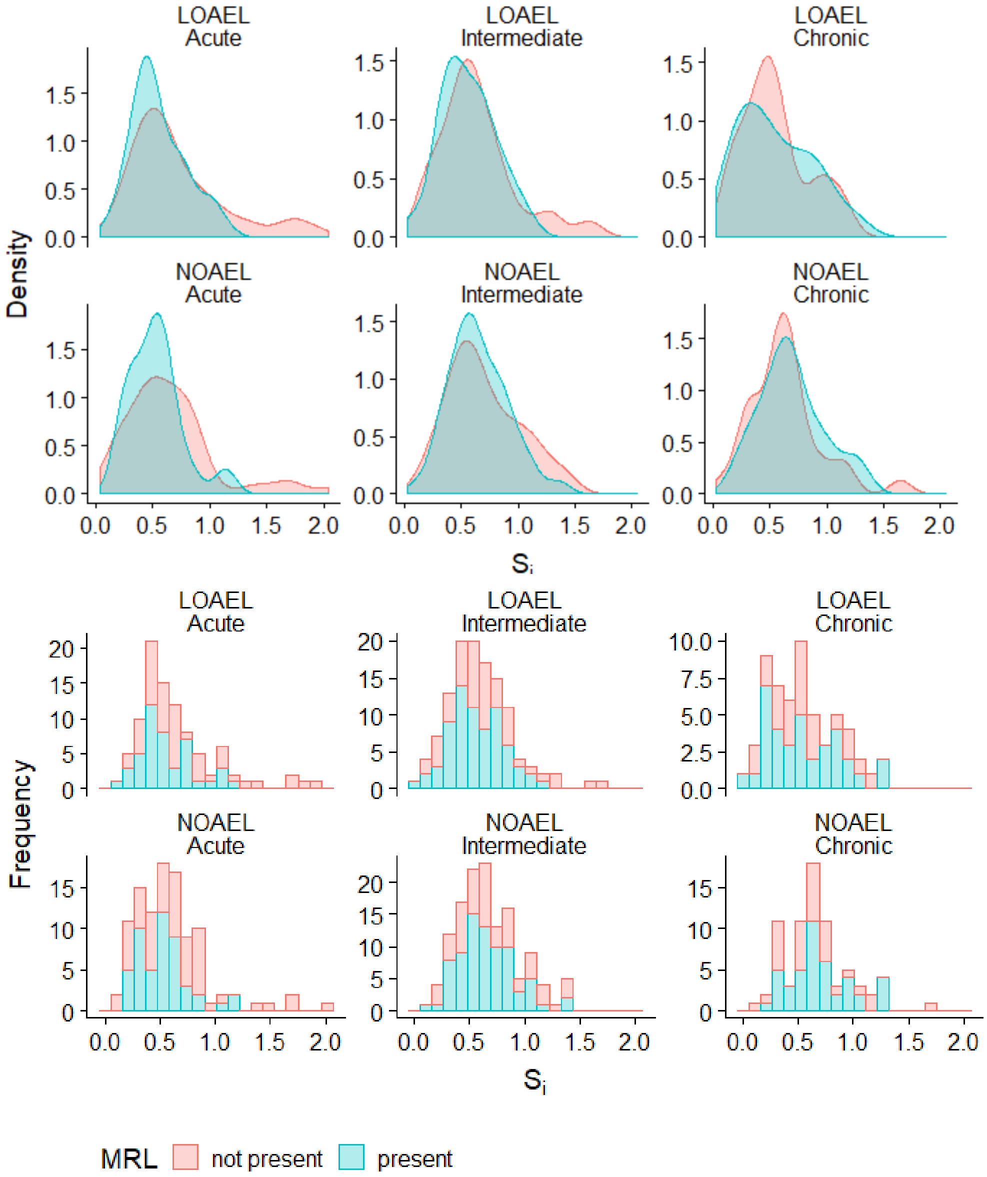
Figure 4b. Histograms and density plots of SDs for the comparison of variability between rat and mouse. The density plots were constructed using default bandwidths in the R density function. The table includes the sample size used to construct respective plots.





	Body weight change (n=49)	Hematological (n=34)	Hepatic (n=52)	Renal (n=36)
Endpoint				
LOAEL	14 (28.6%)	5 (14.7%)	19 (36.5%)	7 (19.4%)
NOAEL	35 (71.4%)	29 (85.3%)	33 (63.5%)	29 (80.6%)

Figure 4c. Histogram and density plots of SDs for the comparison of variability among Body weight change, Hematological, Hepatic, and Renal target system in chronic exposure duration (significant differences between the distributions of variances were observed in these datasets). The density plots were constructed using default bandwidths in the R density function. The table includes the sample size used to construct respective plots.



	Acute		Chronic		Intermediate	
	LOAEL (n=93)	NOAEL (n=104)	LOAEL (n=58)	NOAEL (n=75)	LOAEL (n=121)	NOAEL (n=132)
MRL						
not present	48 (51.6%)	55 (52.9%)	23 (39.7%)	32 (42.7%)	50 (41.3%)	54 (40.9%)
present	45 (48.4%)	49 (47.1%)	35 (60.3%)	43 (57.3%)	71 (58.7%)	78 (59.1%)

Figure 4d. Histograms and density plots of SDs for the comparison of variability between endpoints based on availability of corresponding MRLs. The density plots were constructed using default bandwidths in the R density function. The table includes the sample size used to construct respective plots.

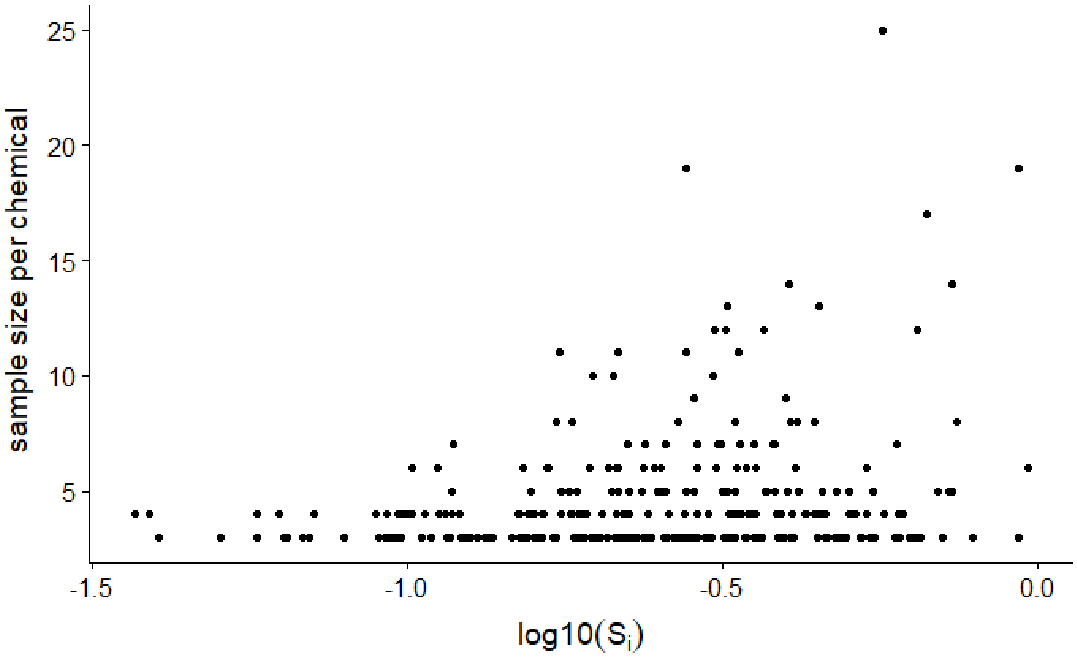


Figure 5a. Scatterplot of log10(S_i) versus sample size (n) of a chemical (using LD_{50}s full dataset). The variability of LD_{50}s per chemical did not necessarily decrease relative to the increase in number of studies available for that chemical.

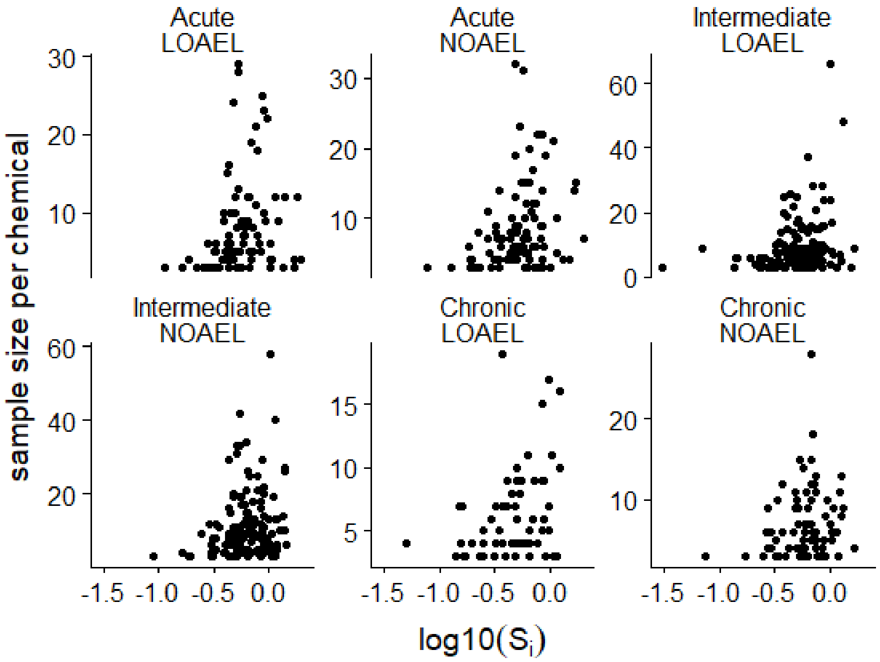


Figure 5b. Scatterplot of log10(S_i) versus sample size (n) of a chemical (using L(N)OAELs full datasets). Although the trend is less clear as compared to LD_{50}s in Figure 5a, the variability of endpoints per chemical did not necessarily decrease relative to the increase in number of studies available for that chemical.