Qualitative and Quantitative Variability of Repeat Dose Animal Toxicity Studies

Katie Paul Friedman, PhD

paul-friedman.katie@epa.gov

Toxicologist, Center for Computational Toxicology and Exposure,

Office of Research and Development, US EPA

March 21, 2023

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA

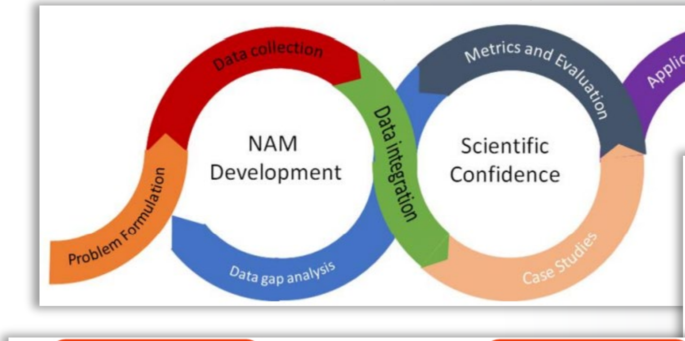
Conflict of interest

- This presentation has been reviewed by the US EPA Office of Research and Development, Center for Computational Toxicology and Exposure.
- The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.
- No conflicts of interest.

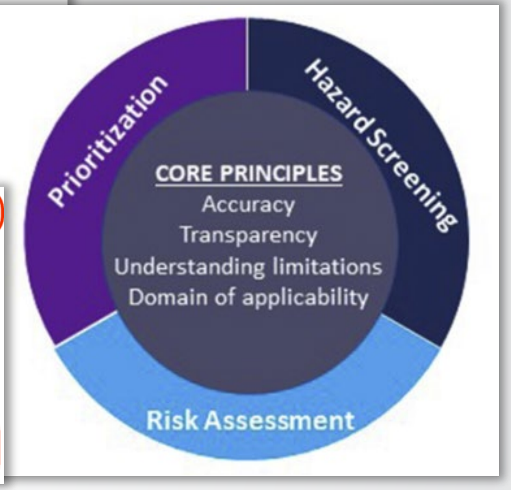
Variability of *in vivo* repeat dose data informs NAM performance expectations and a part of scientific confidence

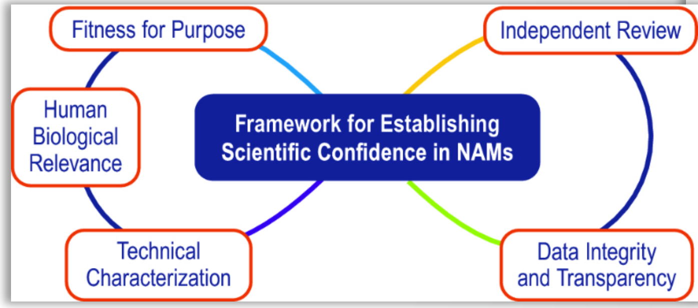
- In Section 4(h) in the Lautenberg amendment to Toxic Substances Control Act:
  - *"…Administrator shall reduce and replace, to the extent practicable and scientifically justified…the use of vertebrate animals in the testing of chemical substances or mixtures…"*
  - New approach methods (NAMs) need to provide *"information of equivalent or better scientific quality and relevance…"* than the traditional animal models
- Multiple frameworks suggest scientific confidence may depend in part on fitness for purpose, biological relevance, and characterization of NAM performance, which in some cases relates to traditional animal study performance or reference data.

US EPA NAMs WorkPlan (2020-2021)



Parish et al. (2020). 10.1016/j.yrtph.2020.104592





Fitness for Purpose
Independent Review
Human Biological Relevance
Framework for Establishing Scientific Confidence in NAMs
Technical Characterization
Data Integrity and Transparency

van der Zalm et al. (2022). 10.1007/s00204-022-03365-4

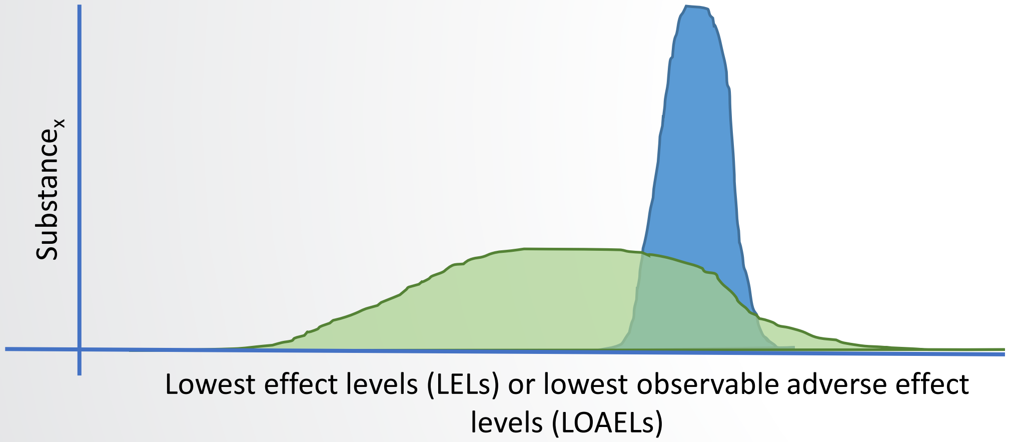
How do we define expectations of *in silico, in chemico,* and *in vitro* models for predicting repeat-dose toxicity?

***In silico, in chemico,* and *in vitro* models cannot predict *in vivo* systemic effect values from animal studies with greater accuracy than those animal models reproduce themselves.**

How can variability in traditional animal studies be expressed for use as reference or training data?

Quantitative: variance is a measure of how far values are spread from the average.

We need to know what the “spread” or variability of traditional effect levels might be to know the range of acceptable or “good” values from a NAM.



Qualitative: We need to know if a specific effect is always observed or not.

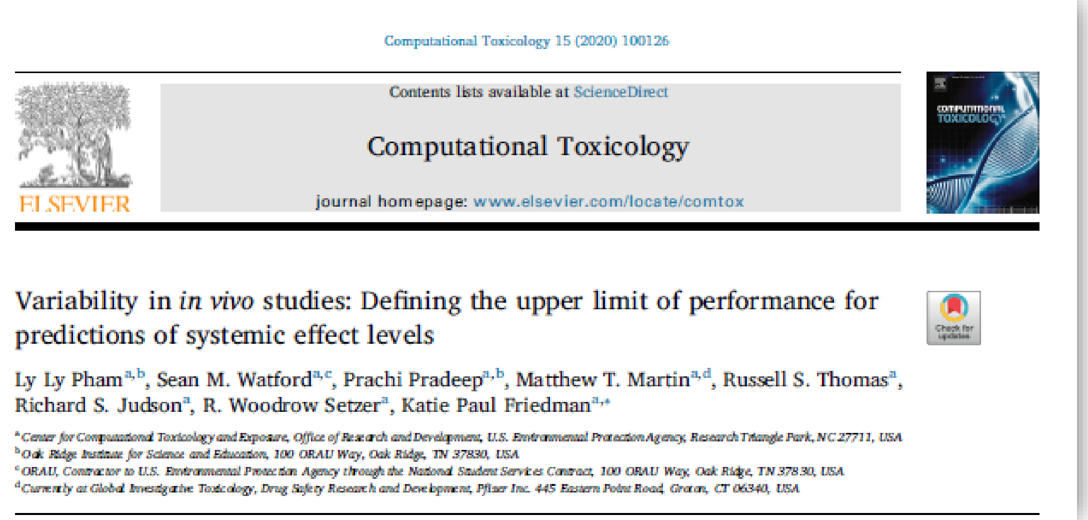
We need to know something about classification performance or about reference data for a phenotype.

		“Truth” (traditional toxicology)	
		Negative	Positive
Predicted (NAM)	Negative	True negative	False negative
	Positive	False positive	True positive

If we are going to learn from variable and uncertain data, we will propagate this variability and uncertainty to any NAMs developed.

If we are going to evaluate NAM performance based on comparison to *in vivo* data, we should account for variability and uncertainty in these reference data.

Part I: Benchmarks on quantitative reproducibility of systemic findings in repeat dose animal studies



Computational Toxicology 15 (2020) 100126

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

ELSEVIER

Variability in *in vivo* studies: Defining the upper limit of performance for predictions of systemic effect levels

Ly Ly Pham[a,b], Sean M. Watford[a,c], Prachi Pradeep[a,b], Matthew T. Martin[a,d], Russell S. Thomas[a], Richard S. Judson[a], R. Woodrow Setzer[a], Katie Paul Friedman[a,*]

[a] Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[b] Oak Ridge Institute for Science and Education, 100 ORAU Way, Oak Ridge, TN 37830, USA
[c] ORAU, Contractor to U.S. Environmental Protection Agency through the National Student Services Contract, 100 ORAU Way, Oak Ridge, TN 37830, USA
[d] Currently at Global Investigative Toxicology, Drug Safety Research and Development, Pfizer Inc. 445 Eastern Point Road, Groton, CT 06340, USA

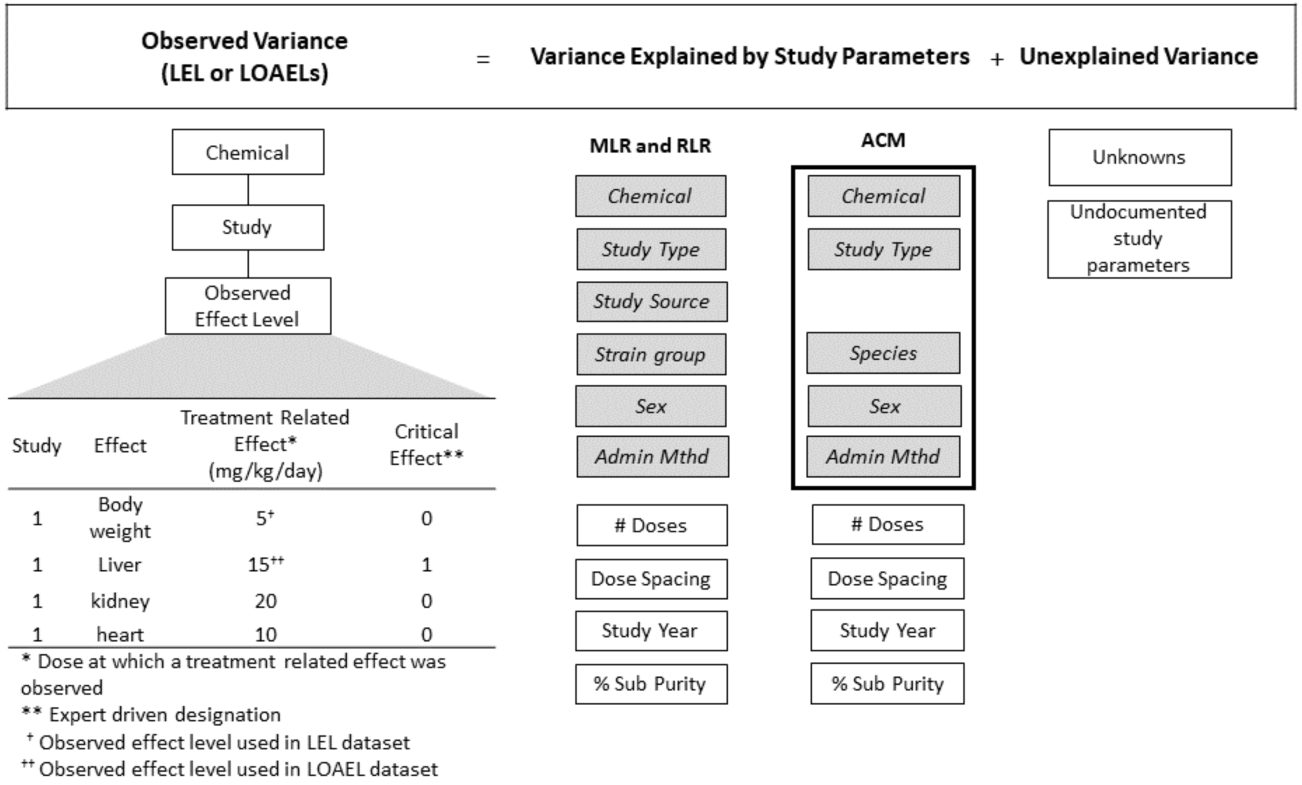
Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

Primary Research Question	Statistical approaches
What is the range of possible effect values (mg/kg/day) in replicate studies for a given chemical?	• Residual root mean square error (RMSE) is an estimate of variance in the same units as the systemic effect values. • The RMSE can be used to define a minimum prediction interval, or estimate range, for a model.
What is the maximal accuracy of a new model that attempts to predict effect values for a chemical?	• The mean square error (MSE) is used to approximate the unexplained variance (not explained by study descriptors) for comparison to total variance. • This % unexplained variance limits the maximal R-squared on a new model.

Based on the study descriptors in ToxRefDB v2.0, we developed statistical models of the variance in quantitative systemic effect level values.

Total variance

Approximated by mean square error



Observed Variance (LEL or LOAELs) = **Variance Explained by Study Parameters** + **Unexplained Variance**

Study	Effect	Treatment Related Effect* (mg/kg/day)	Critical Effect**
1	Body weight	5^{+}	0
1	Liver	15^{++}	1
1	kidney	20	0
1	heart	10	0

* Dose at which a treatment related effect was observed

** Expert driven designation

[+] Observed effect level used in LEL dataset

[++] Observed effect level used in LOAEL dataset

Using two approaches:

	Multilinear regression (MLR, RLR)	Augmented cell means (ACM)
Aggregation level	Chemical	Chemical-Study Type-Species-Sex-Admin Method combination
Replicate definition stringency	Not stringent	Stringent
N	Maximized; ↓ impact of outliers/database error rate	Small; may bias variance estimate
Study descriptors	Contribute independently to variance	Accounts for possible interactions among descriptors

Figure 2. Statistical model of the variance. *LEL = lowest effect level; LOAEL = lowest observable adverse effect level. The LEL is the lowest treatment-related effect observed for a given chemical in a study, and the LOAEL is defined by expert review as coinciding with the critical effect dose level from a given study. Multiple studies for a given chemical yield multiple LELs and LOAELs for computation of variance. MLR = multilinear regression; RLR = robust linear regression; ACM = augmented cell means; Adm. Method = administration method; % Sub Purity = % substance purity used in the study. The gray shaded study descriptor boxes are categorical variables, and the white study descriptor boxes are continuous variables. The box around five categorical study descriptors for the ACM indicates these were concatenated to a factor to define study replicates.*

Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

.

Repeat dose studies for regulatory toxicology, as conducted and curated, may have inherent irreducible amount of unexplained variance.

- 28 different statistical models were constructed.
- RMSE is used to define a 95% minimum prediction interval (i.e., based on the standard deviation or spread of the residuals).
- The % explained variance (amount explained by study descriptors) likely approaches 55-73%.
- This means that the R^2 on some new, predictive model would approach 0.55 to 0.73 as an upper bound on accuracy.

	Total Variance (log$_{10}$-mg/kg/day)^2	Unexplained Variance (MSE) (log$_{10}$-mg/kg/day)^2	RMSE (log$_{10}$-mg/kg/day)	% explained variance	Minimum prediction interval (log$_{10}$-mg/kg/day)
Range	0.744 - 1.013	0.2 - 0.395	0.448 - 0.629	54.9 - 73.3	± 0.878 - ± 1.23
Median (MAD)	0.825 (0.065)	0.301 (0.068)	0.549 0.061	66.1 4.89	± 1.07 (0.12)
Mean (SD)	0.838 (0.070)	0.300 (0.055)	0.545 (0.050)	65.3 (4.86)	± 1.07 (0.098)

Based on tables from Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

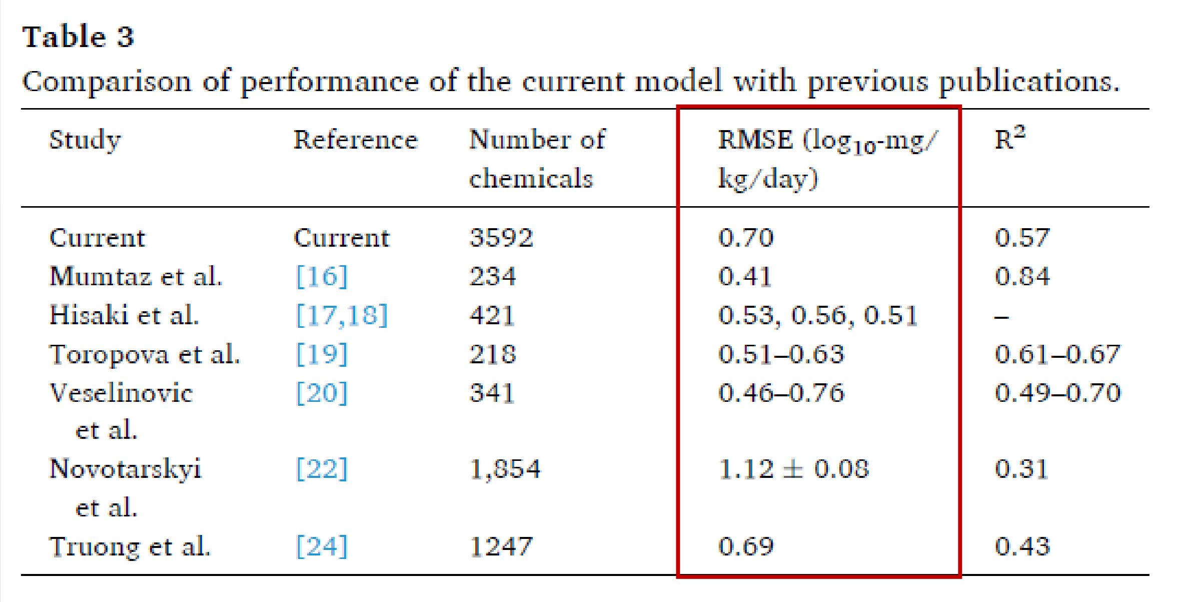


Table 3

Comparison of performance of the current model with previous publications.

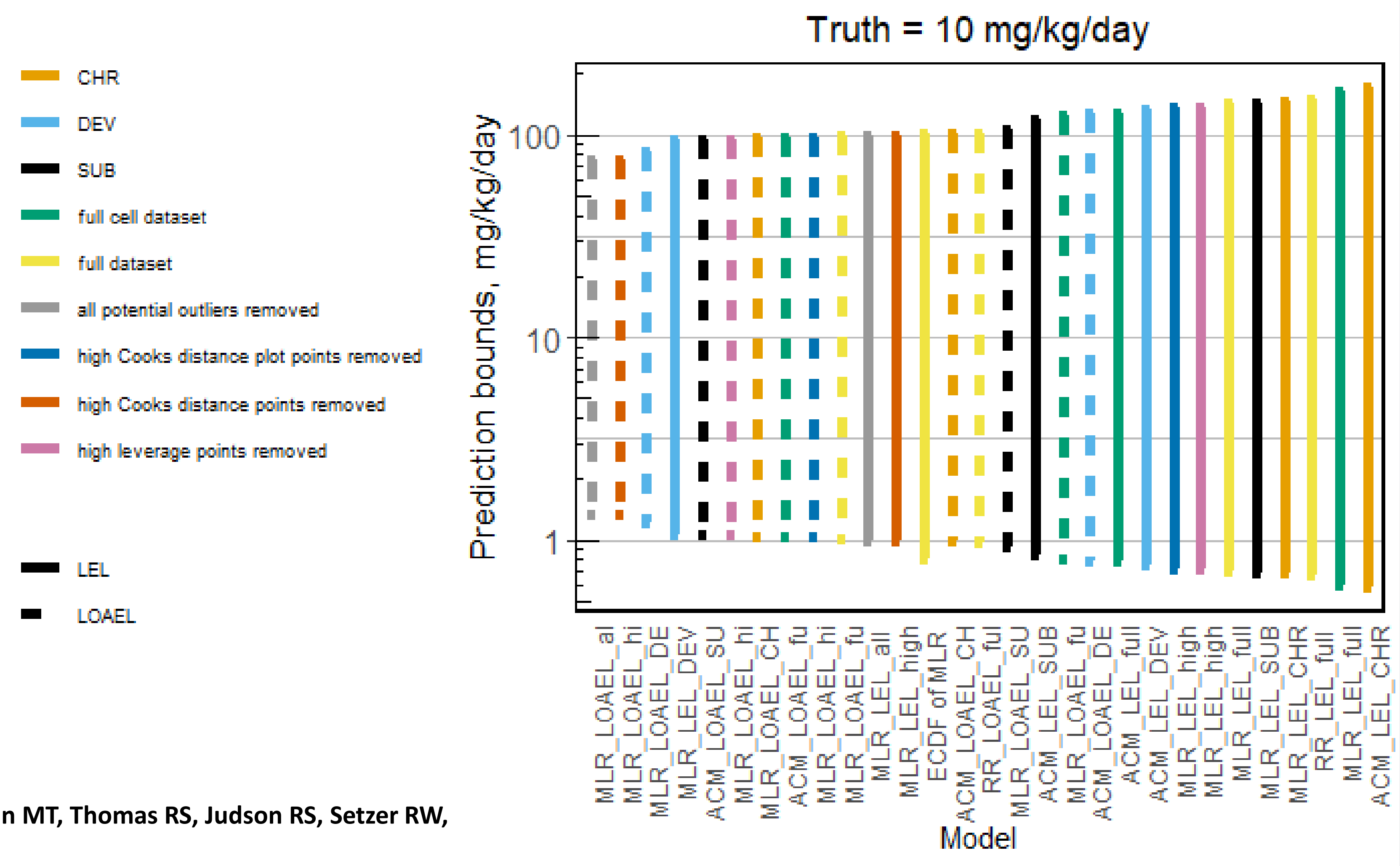
Study	Reference	Number of chemicals	RMSE (log$_{10}$-mg/kg/day)	R^2
Current	Current	3592	0.70	0.57
Mumtaz et al.	[16]	234	0.41	0.84
Hisaki et al.	[17,18]	421	0.53, 0.56, 0.51	–
Toropova et al.	[19]	218	0.51–0.63	0.61–0.67
Veselinovic et al.	[20]	341	0.46–0.76	0.49–0.70
Novotarskyi et al.	[22]	1,854	1.12 ± 0.08	0.31
Truong et al.	[24]	1247	0.69	0.43

- A multi-linear regression QSAR model of chronic oral rat LOAEL values for approximately 400 chemicals, demonstrated a RMSE of 0.73 log$_{10}$(mg/kg-day) which was similar to the size of the variability in the training data, ±0.64 log$_{10}$(mg/kg-day), suggested that the error in the model approached the error in the reference data from different laboratories (Mazzatorta et al. 2008; Helma et al. 2018).

Pradeep P, Paul Friedman K, Judson RS. (2020). 10.1016/j.comtox.2020.100139

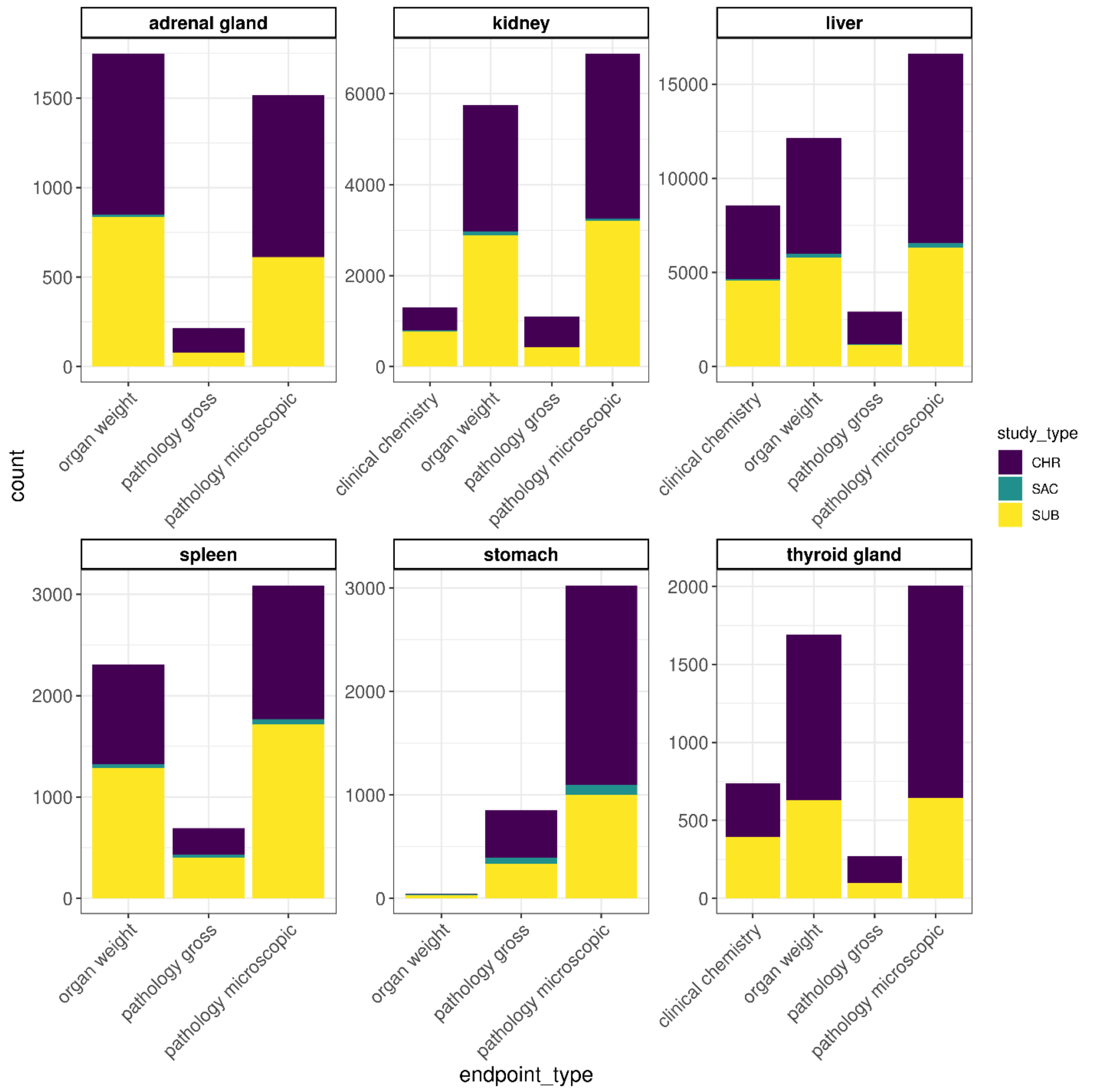
Range of 95% minimum prediction intervals across the modeling approaches, effect levels, and study types is 58-284-fold

If attempting to use a NAM-based predictive model for prediction of a reference systemic effect level value of 10 mg/kg/day, it is likely that given the variability in reference data of this kind, that a model prediction of somewhere between 1 and 100 mg/kg/day would be the greatest amount of accuracy achievable (100-fold wide).



Based on tables from Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

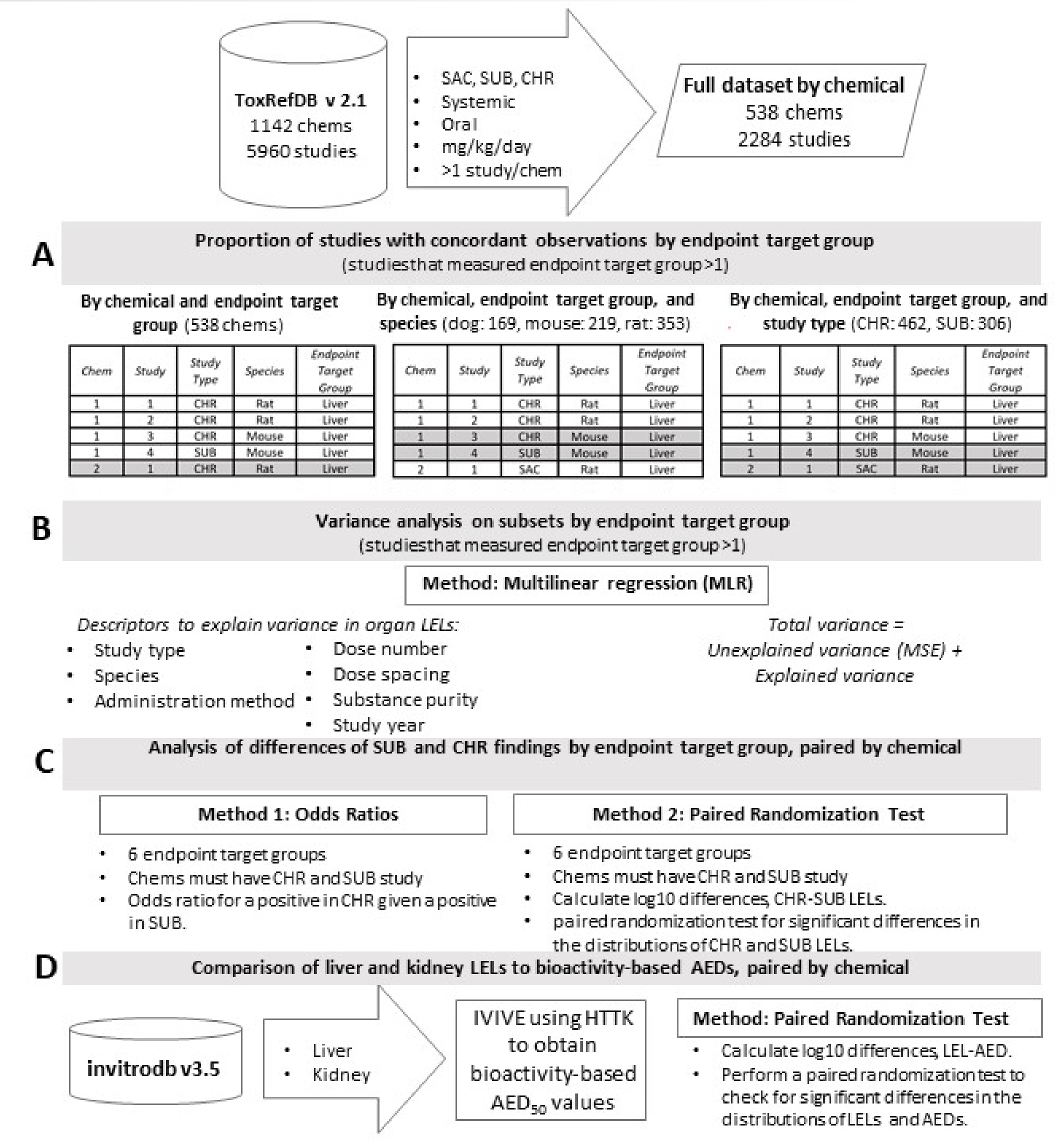
Part II: What if we only considered reproducibility in the context of specific organ-level effects?



Supp Fig 2, Paul Friedman et al. (in prep).

- Hypothesis: focusing on organ-level effects will result in reduced variance because the target site is conserved
- 6 tissues with the most positive reporting
- Exclude non-specific systemic effects (BW, food consumption)
- How reproducible are these types of effects in replicate studies?

How reproducible are organ level effects in replicate studies and studies of different duration?



ToxRefDB v 2.1
1142 chems
5960 studies

- SAC, SUB, CHR
- Systemic
- Oral
- mg/kg/day
- >1 study/chem

Full dataset by chemical
538 chems
2284 studies

A

Proportion of studies with concordant observations by endpoint target group
(studiesthat measured endpoint target group>1)

By chemical and endpoint target group (538 chems)

Chem	*Study*	*Study Type*	*Species*	*Endpoint Target Group*
1	1	CHR	Rat	Liver
1	2	CHR	Rat	Liver
1	3	CHR	Mouse	Liver
1	4	SUB	Mouse	Liver
2	1	CHR	Rat	Liver

By chemical, endpoint target group, and species (dog: 169, mouse: 219, rat: 353)

Chem	*Study*	*Study Type*	*Species*	*Endpoint Target Group*
1	1	CHR	Rat	Liver
1	2	CHR	Rat	Liver
1	3	CHR	Mouse	Liver
1	4	SUB	Mouse	Liver
2	1	SAC	Rat	Liver

By chemical, endpoint target group, and study type (CHR: 462, SUB: 306)

Chem	*Study*	*Study Type*	*Species*	*Endpoint Target Group*
1	1	CHR	Rat	Liver
1	2	CHR	Rat	Liver
1	3	CHR	Mouse	Liver
1	4	SUB	Mouse	Liver
2	1	SAC	Rat	Liver

B

Variance analysis on subsets by endpoint target group
(studiesthat measured endpoint target group>1)

Method: Multilinear regression (MLR)

Descriptors to explain variance in organ LELs:

- Study type
- Species
- Administration method
- Dose number
- Dose spacing
- Substance purity
- Study year

Total variance =
Unexplained variance (MSE) +
Explained variance

C

Analysis of differences of SUB and CHR findings by endpoint target group, paired by chemical

Method 1: Odds Ratios

- 6 endpoint target groups
- Chems must have CHR and SUB study
- Odds ratio for a positive in CHR given a positive in SUB.

Method 2: Paired Randomization Test

- 6 endpoint target groups
- Chems must have CHR and SUB study
- Calculate log10 differences, CHR-SUB LELs.
- paired randomization test for significant differences in the distributions of CHR and SUB LELs.

D

Comparison of liver and kidney LELs to bioactivity-based AEDs, paired by chemical

invitrodb v3.5

- Liver
- Kidney

IVIVE using HTTK to obtain bioactivity-based AED_{50} values

Method: Paired Randomization Test

- Calculate log10 differences, LEL-AED.
- Perform a paired randomization test to check for significant differences in the distributions of LELs and AEDs.

Figure 1, Paul Friedman et al. (in prep).

A. What is the qualitative reproducibility of organ-level findings in repeat dose animal studies?

B. Are variance estimates reduced for organ-level effects in repeat dose animal studies when compared to systemic effects, using LELs, BMDs, etc.?

C. Understanding NAM alternatives are not necessarily 1:1 replacements, would estimates of subchronic and chronic effect levels be necessary?

D. Are NAM-based PODs within estimates of variability in replicate repeat dose studies?

A: How qualitatively reproducible are organ level findings in repeat dose studies?

Primary Research Question	Statistical approaches
How concordant are organ-level effects for multiple repeat dose study observations?	Calculate concordance of findings between replicate studies when grouped by chemical and organ; chemical, organ, and species; and chemical, organ, and study type

$$\%\,Concordance = \frac{chemical\ with\ positive\ finding\ in\ all\ studies + chemicals\ with\ negative\ finding\ in\ all\ studies}{total\ chemicals\ tested}$$

- Qualitative reproducibility of organ-level effect observations in repeat dose studies of adult animals was 33-88%, depending on grouping.
- Organs associated with more negative chemicals (stomach, thyroid, adrenal) had higher rates of concordance.
- Within-species concordance tended to be greater than within-study concordance.

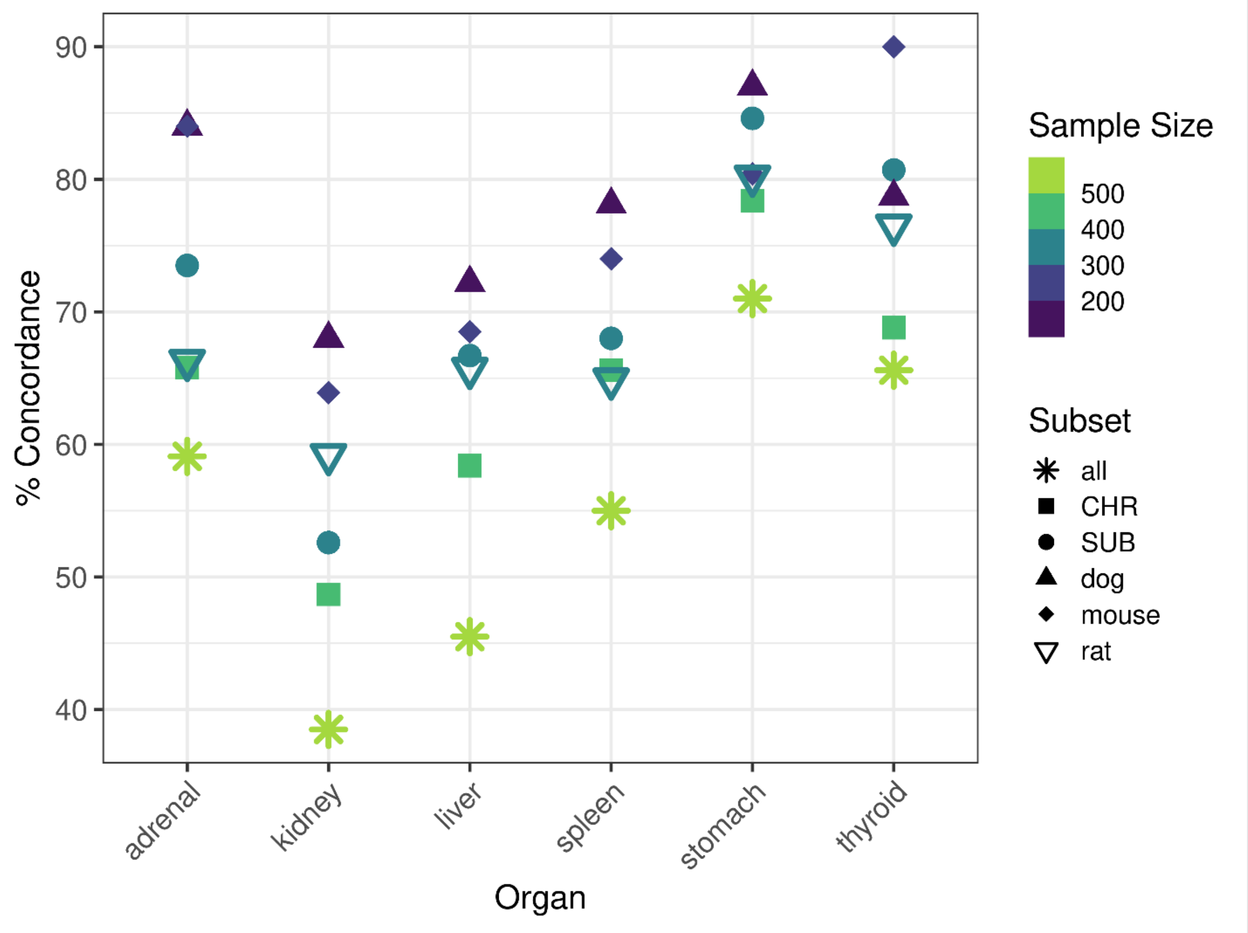


Figure 2, Paul Friedman et al. (in prep).

Previous estimates of inter-species concordance are within the range we observed

Comparison type	Effect type	Species	Description of N	% Concordance	Reference
Intraspecies (species-sex) concordance	Site-specific carcinogenesis	Male/Female site-specific carcinogenesis, average of within-mouse and within-rat	146 chemicals for rat; 159 chemicals for mouse	65-66	Haseman and Lockhart, 1993 https://doi.org/10.1289/ehp.9310150
Interspecies concordance	Site-specific carcinogenesis, average for all sites	Rat/Mouse	173 site-specific cancer positives in rat divided by positives in mouse, by chemical	35	Haseman and Lockhart, 1993 https://doi.org/10.1289/ehp.9310150
Interspecies concordance	Site-specific carcinogenesis, average for all sites	Mouse/Rat	167 site-specific cancer positives in mouse divided by positives in rat, by chemical	37	Haseman and Lockhart, 1993 https://doi.org/10.1289/ehp.9310150
Intraspecies concordance	Carcinogen/non-carcinogen	Mouse	NCI/NTP studies vs. CPDB literature component; 70 chemicals	49	Gottmann *et al.*, 2001 10.1289/ehp.01109509
Intraspecies concordance	Carcinogen/non-carcinogen	Rat	NCI/NTP studies vs. CPDB literature component; 71 chemicals	62	Gottmann *et al.*, 2001 10.1289/ehp.01109509
Interspecies	Carcinogen/non-carcinogen	Rat vs. Mouse	NTP studies, 313 chemicals	74.4	Huff *et al.*, 1991 10.1289/ehp.9193247

Table 4, Paul Friedman et al. (in prep).

Examining organ effect levels specifically failed to reduce estimates of variance (RMSE)

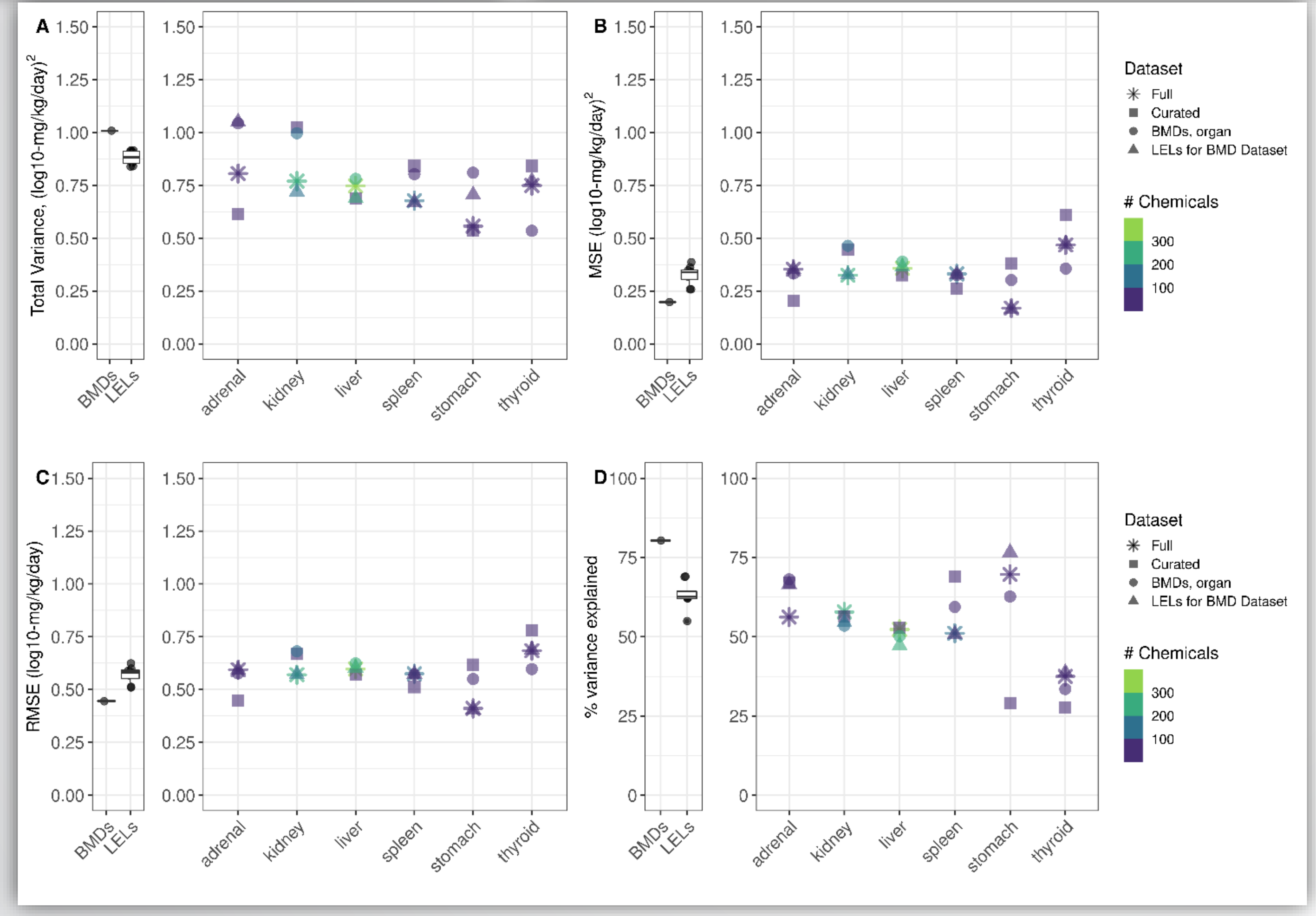


Figure 3, Paul Friedman et al. (in prep).

Primary Research Question	Statistical approaches
Can the estimate of variance for chemicals with replicate studies be reduced by estimating variance in specific organs?	Use multi-linear regression to approximate total variance, unexplained variance (MSE), RMSE, and % variance explained.

Predictions of an organ-level finding within ±1 log10-mg/kg/day may be an upper limit expectation on NAM performance.

Qualitative reproducibility of organ-level findings between SUB and CHR studies may inform NAM strategy

- *In silico* NAMs for repeat dose toxicity could potentially be improved by combining SUB and CHR data for greater chemical coverage in training/testing.
    - Is it reasonable to expect similar organs will be affected by different study durations?

- Would a strategy focused on identification of a protective repeat dose point of departure using shorter-term studies or NAMs, without a chronic exposure study, miss organ-level effects?
    - NAM strategy could include cheminformatics and toxicoinformatics to identify substances with longer serum half-life.
    - Exclude consideration of adversity of the findings in the organ.

Odds ratios for a positive in a tissue in a CHR given a negative in SUB are all less than 1, indicating this is an unlikely scenario.

Primary Research Question	Statistical approaches
What are the odds a chemical will produce any organ-level effect in a chronic (1-2 yr) study if the subchronic study was negative?	Calculate odds ratios for chemicals with subchronic and chronic study information

Possible indication: a repeat dose POD for a target organ at 90 days, particularly for liver and kidney where we have the largest datasets, is likely protective for a chronic finding.

(without accounting for level of adversity)

A negative in the SUB indicates a greater likelihood of negative in the CHR.

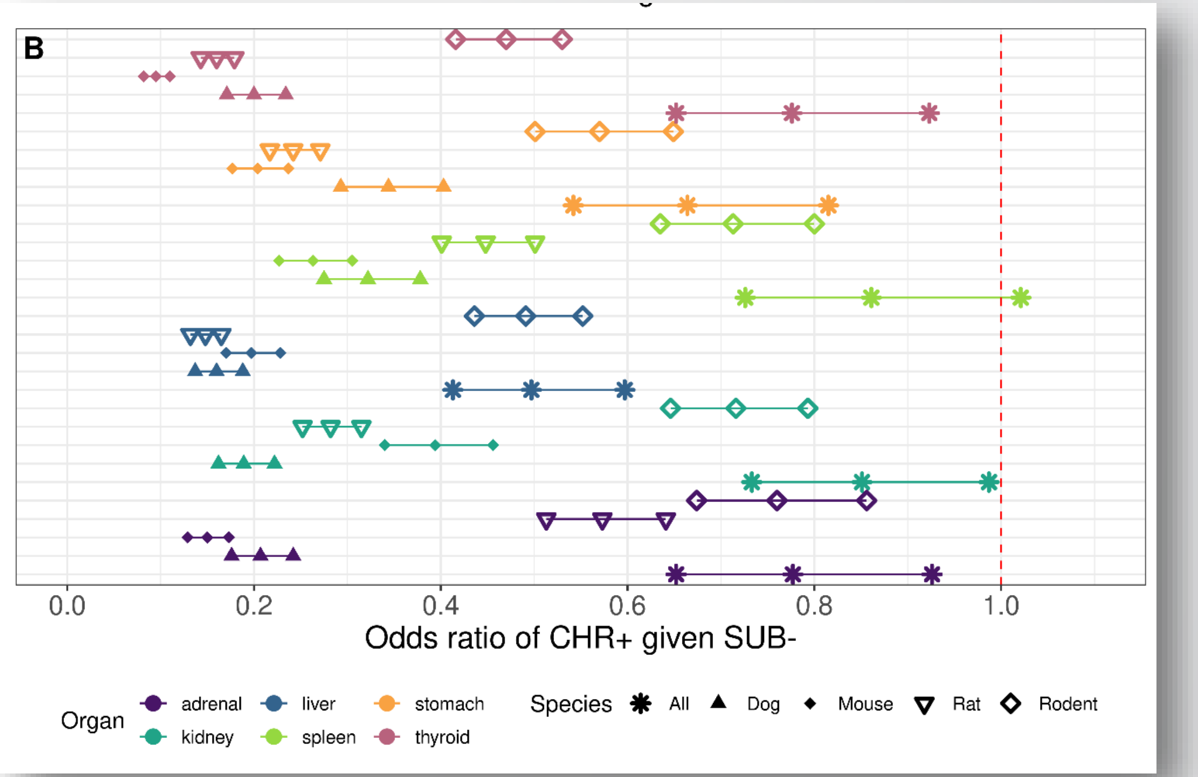
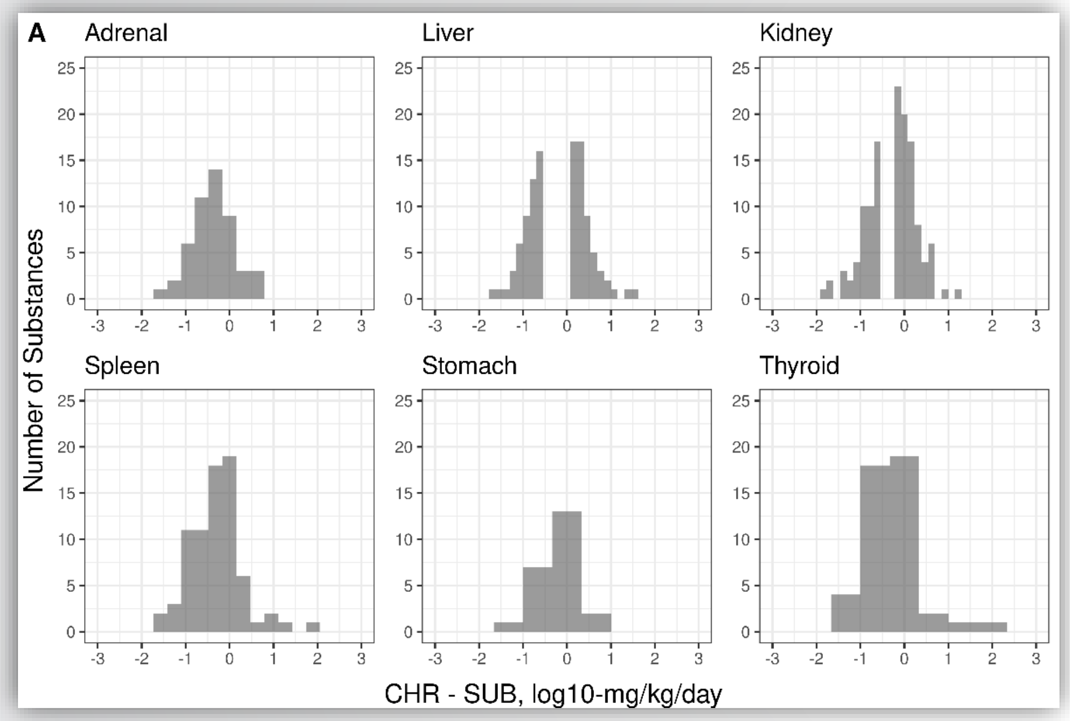


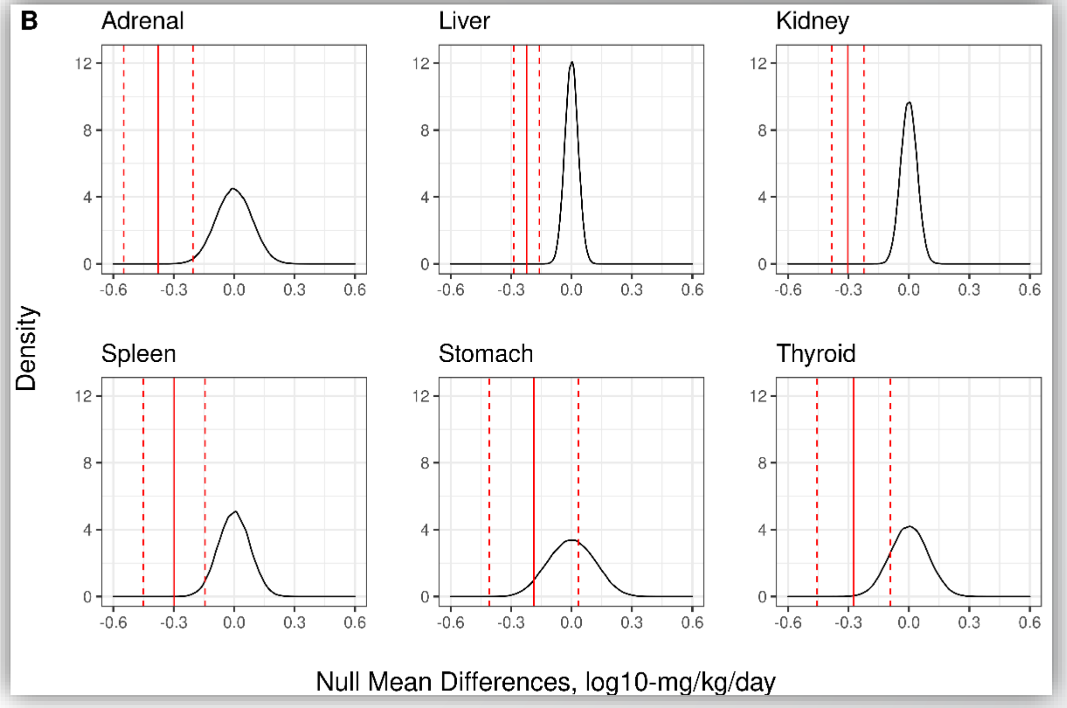
Figure 4B, Paul Friedman et al. (in prep).

Generally, the chronic effect level values are 0.3 log10-mg/kg/day less than subchronic effect level values

Raw differences in CHR –SUB LELs



Sample mean differences ± CI compared to distribution of null mean differences



—

sample mean difference from the original data (log10(CHR/SUB))

- - -

2-sided 95% confidence interval (p <0.05); *if the interval includes 0 then we cannot say that the true mean difference is different from 0*

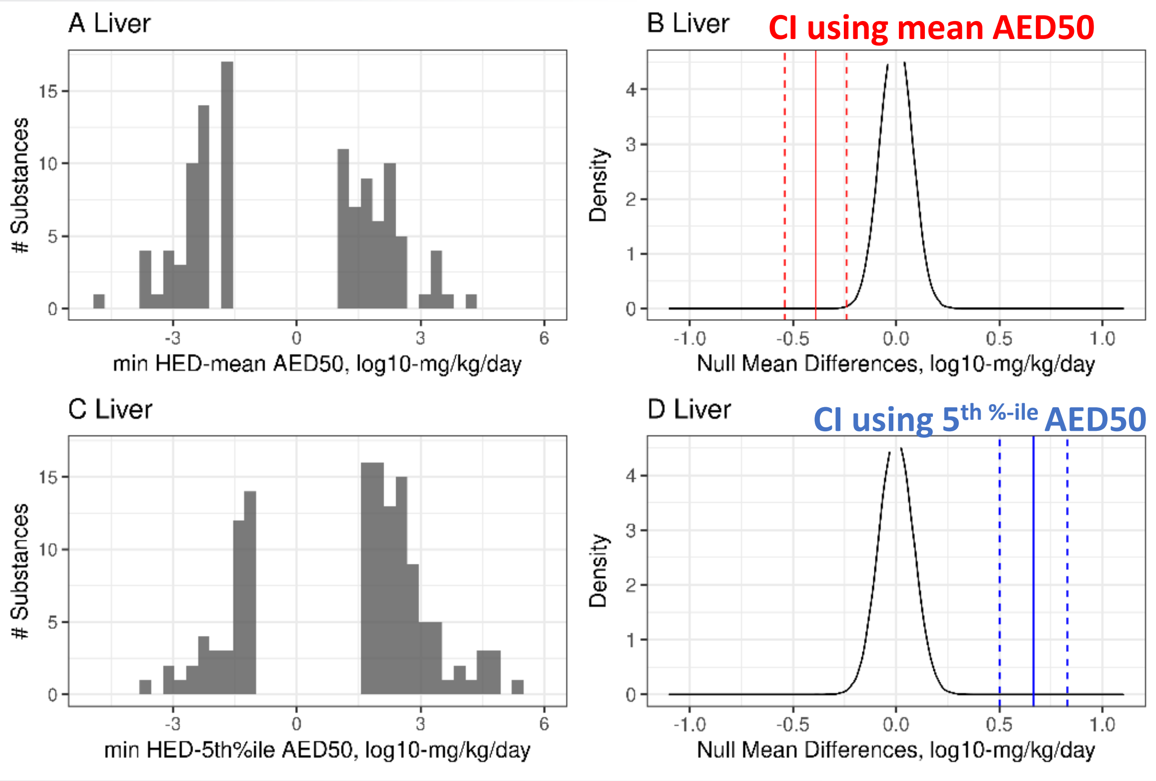
—

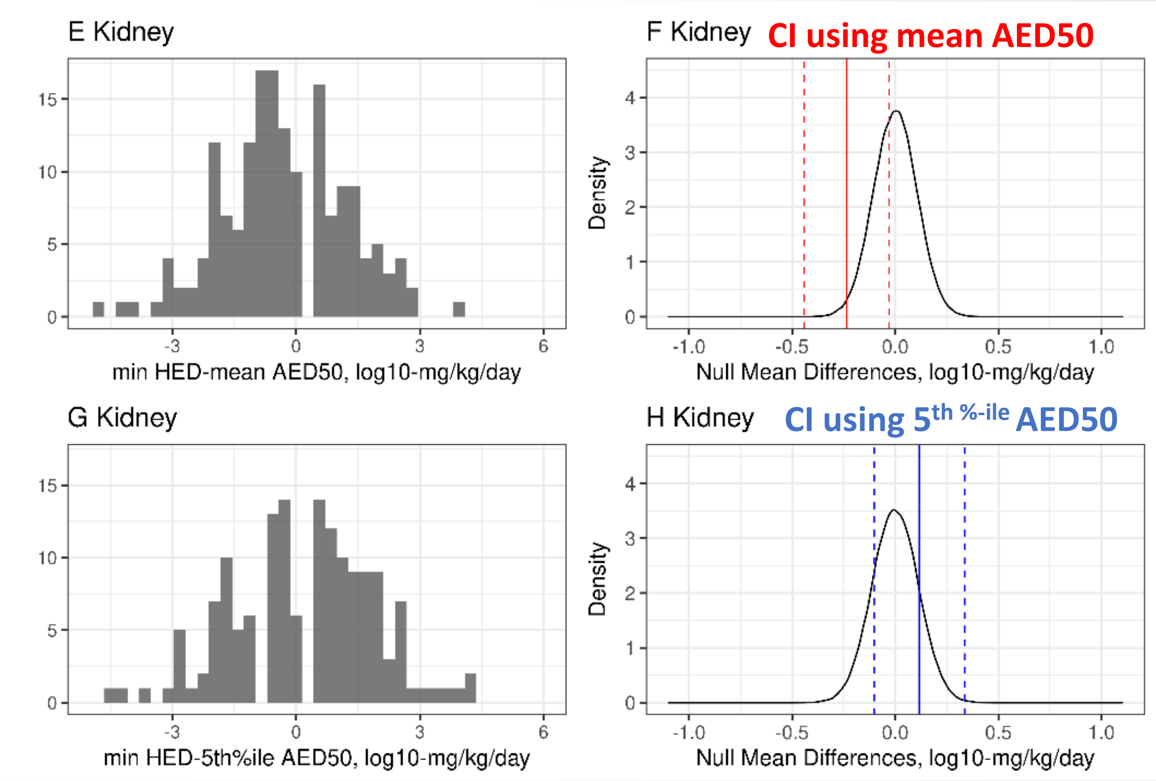
Distribution of log10-transformed LEL differences following 100,000 randomization tests

- The mean differences in CHR – SUB min LEL values by organ approach estimates of variance in replicate repeat dose studies.
- *In silico* and *in vitro* NAMs for repeat dose point-of-departure estimation could combine SUB and CHR data in training.
- Current uncertainty or adjustment factors for SUB to CHR are protective.

The distribution of LEL-AED$_{50}$ differences demonstrated very long tails, signaling the differences in LELs or HEDs and AEDs can be extreme

- Distributions of raw differences suggest the mean difference approaches 0, but these distributions demonstrated much longer tails than the differences in CHR-SUB organ level LELs, with minimum LEL to AED$_{50}$ comparisons at times suggesting differences in excess of 3 orders of magnitude in either direction at the tails
- The *mean* differences (HED or LEL – summary AED50 metrics) are all within 1 log10-mg/kg/day





The distribution of LEL-AED$_{50}$ differences demonstrated very long tails, signaling that for smaller numbers of chemicals, the differences in LELs and AEDs can be extreme

Organ	# Chemicals	In vivo POD (log$_{10}$-mg/kg/day)	AED type (log$_{10}$-mg/kg/day)	Mean difference, in vivo POD - AED (log10-mg/kg/day)	p-value	Lower CI bound	Upper CI bound
Liver	365	min LEL	mean AED	0.3203	<0.0001	0.1736	0.4670
Liver	365	min LEL	5th %-ile AED	1.3755	<0.0001	1.172	1.579
Kidney	194	min LEL	mean AED	0.5060	<0.0001	0.290	0.7223
Kidney	194	min LEL	5th %-ile AED	0.8586	<0.0001	0.608	1.110
Liver	365	min HED	mean AED	-0.3900	<0.0001	-0.5394	-0.2405
Liver	365	min HED	5th %-ile AED	0.6652	<0.0001	0.5013	0.8291
Kidney	194	min HED	mean AED	-0.2357	0.0245	-0.4418	-0.0295
Kidney	194	min HED	5th %-ile AED	0.1169	0.2953	-0.1027	0.3366

Table 3, Paul Friedman et al. (in prep).

It is possible that existing NAMs that indicate organ-level effects, on average, may predict liver- or kidney-related HEDs within estimates of variability in replicate *in vivo* studies, *but caution should be employed in viewing this result due to the tails on the distribution of raw differences*

Conclusions: Primary takeaways from this work

- Part I: Variability in *in vivo* toxicity studies used in training or evaluation limits predictive accuracy of NAMs.
  - Maximal R-squared for a NAM-based predictive model of systemic effect levels may be 55 to 73%; i.e., as much as 1/3 of the variance in these data may not be explainable using study descriptors *at the study and the organ level.*
  - The estimate of variance (RMSE) in curated LELs and/or LOAELs approaches a 0.5 log10-mg/kg/day *at the study and the organ level.*
  - **Understanding that a prediction of an animal systemic effect level within ± 1 log10-mg/kg/day fold demonstrates a *very good* NAM is important for acceptance of NAMs for chemical safety assessment.**
- Part II: Qualitative and quantitative reproducibility of organ-level effect observations in repeat dose studies of adult animals
  - Qualitative concordance of organ-level effects was 33-88%, with highest concordance within species.
  - Quantitative variability in organ-level effects are similar to estimates of variance at the study-level.
  - Subchronic and chronic *in vivo* observations can likely be combined for modeling to increase N.
  - It is unlikely that there are effects in organs like liver or kidney in a chronic study if these organs were unaffected in a subchronic study.
  - A repeat dose point of departure could be predicted by a NAM (e.g., QSAR) and adjusted to create a chronic-protective prediction.

Further application of the learnings herein

- The LEL-AED$_{50}$ and HED-AED$_{50}$ comparison points to the need for a multifaceted approach to quantitative POD prediction when moving beyond the existing paradigm based on long-term animal studies and protective estimates of uncertainty factors, including strategies such as QSAR, read across, bioactivity, and short-term animal studies.
- Construction of NAM-based effect level estimates that offer an equivalent level of public health protection as effect levels produced by methods using animals may provide a bridge to major reduction in the use of animals as well as identification of cases in which animals may provide scientific value.

Thank you for listening

Select References

Gold, L. S., et al. (1989). "Interspecies extrapolation in carcinogenesis: prediction between rats and mice." Environ Health Perspect **81: 211-219.**

Gottmann, E., et al., 2001. Data quality in predictive toxicology: Reproducibility of rodent carcinogenicity experiments. Environmental Health Perspectives. 109**,** 509-514.

Haseman, J. K. (2000). "Using the NTP database to assess the value of rodent carcinogenicity studies for determining human cancer risk." Drug Metab Rev **32(2): 169-186.**

Mazzatorta, P., et al., 2008. Modeling Oral Rat Chronic Toxicity. Journal of Chemical Information and Modeling. 48**,** 1949-1954.

Monticello, T. M., et al. (2017). "Current nonclinical testing paradigm enables safe entry to First-In-Human clinical trials: The IQ consortium nonclinical to clinical translational database." Toxicol Appl Pharmacol **334: 100-109.**

Pham L, Watford S, Pradeep P, Martin M, Judson R, Thomas RT, Setzer RW, Paul Friedman K. (2020). Variability in *in vivo* studies: Defining the upper limit of performance for predictions of systemic effect levels. Computational Toxicology. DOI: https://doi.org/10.1016/j.comtox.2020.100126

Pradeep P, Paul Friedman K, Judson RS. (2020). Structure-based QSAR models to predict repeat dose toxicity points of departure. Computational Toxicology. DOI: https://doi.org/10.1016/j.comtox.2020.100139

Toropov, A. A., et al., 2015. CORAL: model for no observed adverse effect level (NOAEL). Molecular diversity. 19**,** 563-75.

Toropova, A. P., et al., 2017. The application of new HARD-descriptor available from the CORAL software to building up NOAEL models. Food and Chemical Toxicology.

Toropova, A. P., et al., 2015. QSAR as a random event: a case of NOAEL. Environ Sci Pollut Res Int. 22**,** 8264-71.

Veselinović, J. B., et al., 2016. The Monte Carlo technique as a tool to predict LOAEL. European Journal of Medicinal Chemistry. 116**,** 71-75.

Wang, B. and G. Gray (2015). "Concordance of Noncarcinogenic Endpoints in Rodent Chemical Bioassays." Risk Anal **35(6): 1154-1166.**

Watford, S., et al., 2019. ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses. Reprod Toxicol. 89**,** 145-158.

Thanks especially to Richard Judson, Woody Setzer, Ly Ly Pham, Prachi Pradeep, MJ Foster, Sean Watford, and Rusty Thomas

Office of Research and Development
Center for Computational Toxicology & Exposure (CCTE)
Bioinformatic and Computational Toxicology Division (BCTD)
Computational Toxicology and Bioinformatics Branch (CTBB)