Performance Expectations for New Approach Methods in Predicting Effects from Repeat Dose Animal Studies

Katie Paul Friedman, PhD

paul-friedman.katie@epa.gov

Toxicologist, Center for Computational Toxicology and Exposure, Office of Research and Development, US EPA

September 23, 2022
OECD WNT Webinar

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

Overview: Why are traditional animal data being discussed in a session on NAM acceptance?

- Background on expectations of new approach methods (NAMs) for application, where NAMs include machine-learning or modeling approaches to prediction.
- Highlights of published and ongoing research into variability in traditional animal repeat dose toxicity data:
  - Part I: Benchmarks on quantitative reproducibility of systemic findings in repeat dose animal studies
  - Part II: Incorporating estimates of variance into quantitative structure-activity relationship models for repeat dose point of departure (POD) estimates
  - Part III: Benchmarks on reproducibility of organ-level findings in repeat dose animal studies
- Conclusions

What is needed to understand the acceptability of NAMs for risk assessment?

- In US, 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
- "Directive to Prioritize Efforts to Reduce Animal Testing" memorandum signed by Administrator Andrew Wheeler on September 10, 2019
  - "1. Validation to ensure that NAMs are equivalent to or better than the animal tests replaced."

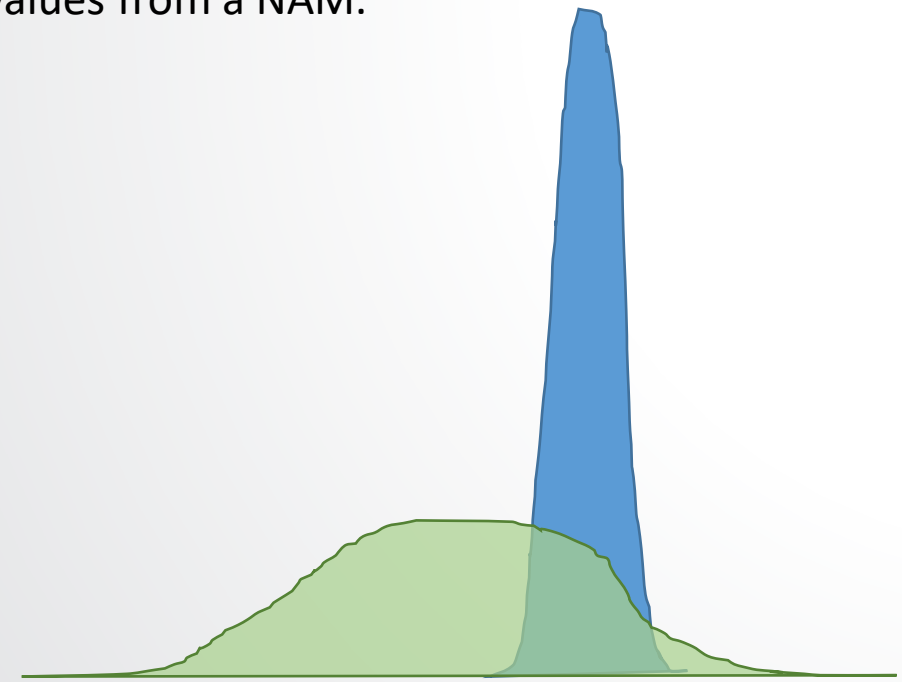
How do we define expectations of *in silico, in chemico,* and *in vitro* models for predicting repeat-dose toxicity?

***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 do we typically express variability in traditional animal toxicity tests used 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 (e.g., lowest effect levels, LELs, or lowest observable adverse effect levels, LOAELs) 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.

		"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.

ToxRefDB v2.0 is a source for data to address these questions of quantitative variability in repeat dose studies

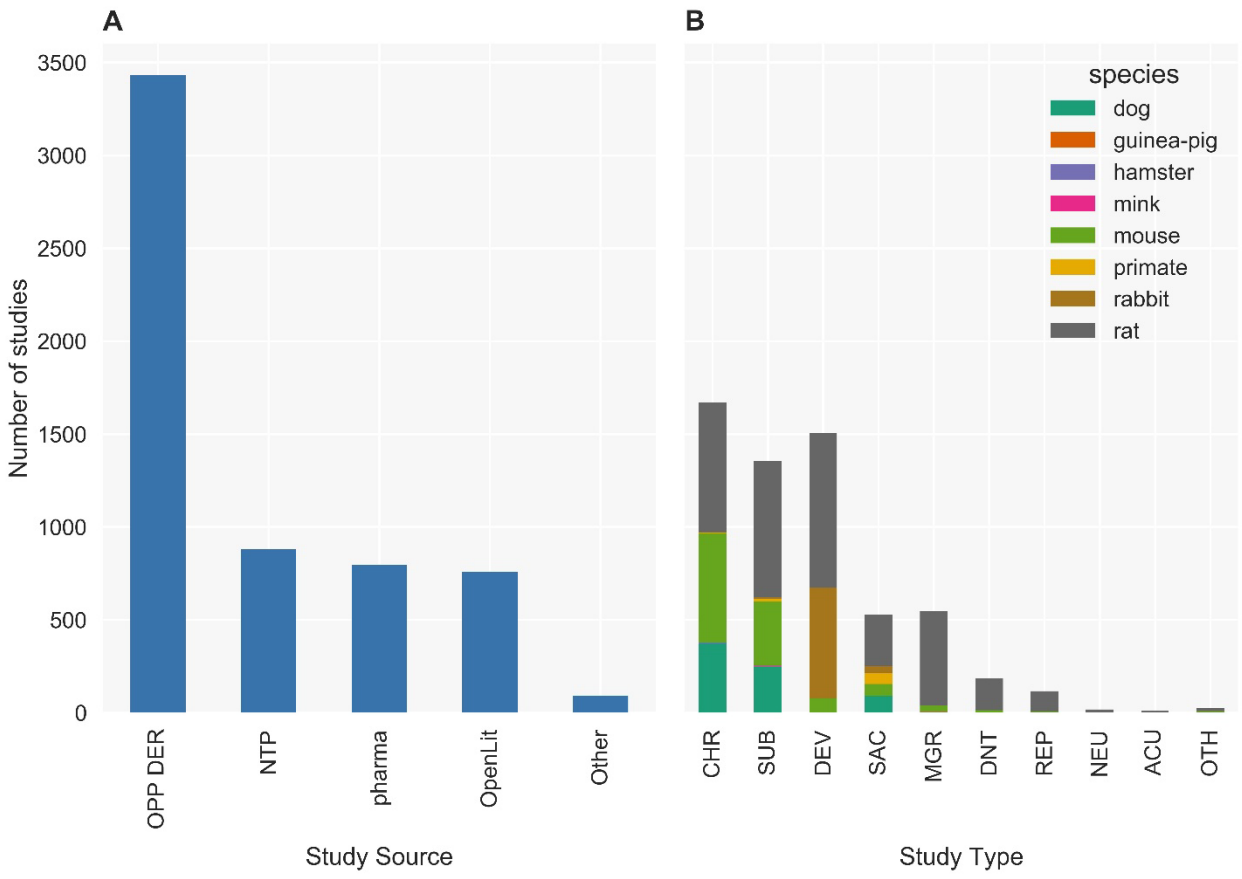


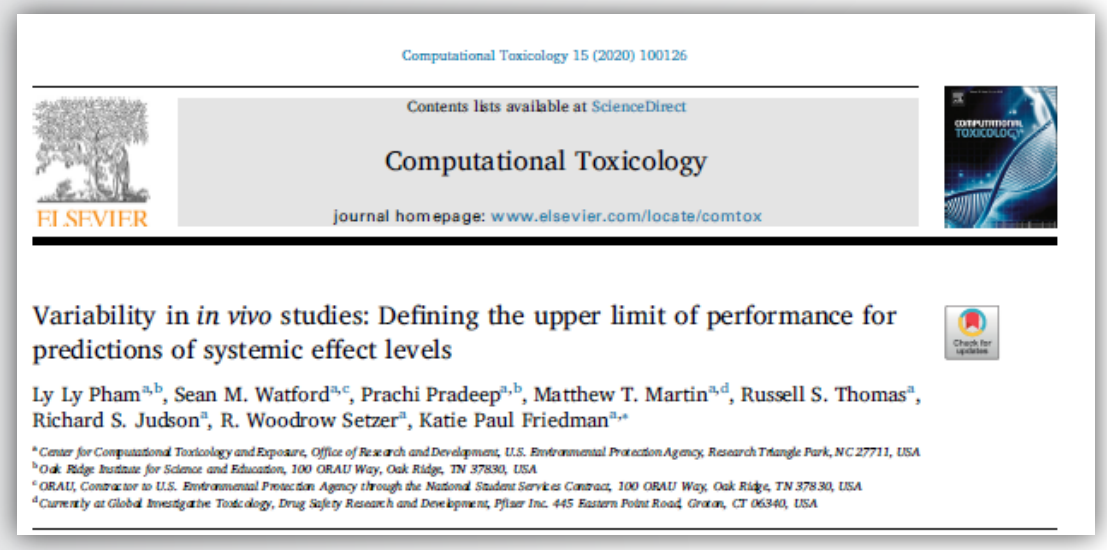
Figure 1. Number of studies by study type and species in ToxRefDB v2.0. *The study designs include chronic (CHR), sub-chronic (SUB), developmental (DEV), subacute (SAC), multigeneration reproductive (MGR), developmental neurotoxicity (DNT), reproductive (REP), neurotoxicity (NEU), acute (ACU), and other (OTH) for numerous species, but mostly for rat, mouse, rabbit, and dog.*

ToxRefDB v2.0 contains relevant study data to evaluate variability in traditional data for >1000 chemicals and >5000 studies.

https://doi.org/10.23645/epacomptox.6062545.v3

Figure from Watford S, Pham LL, Wignall J, Shin R, Martin MT, Paul Friedman K. 2019. "ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses." Reproductive Toxicology; 89: 145-158. https://doi.org/10.1016/j.reprotox.2019.07.012

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

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

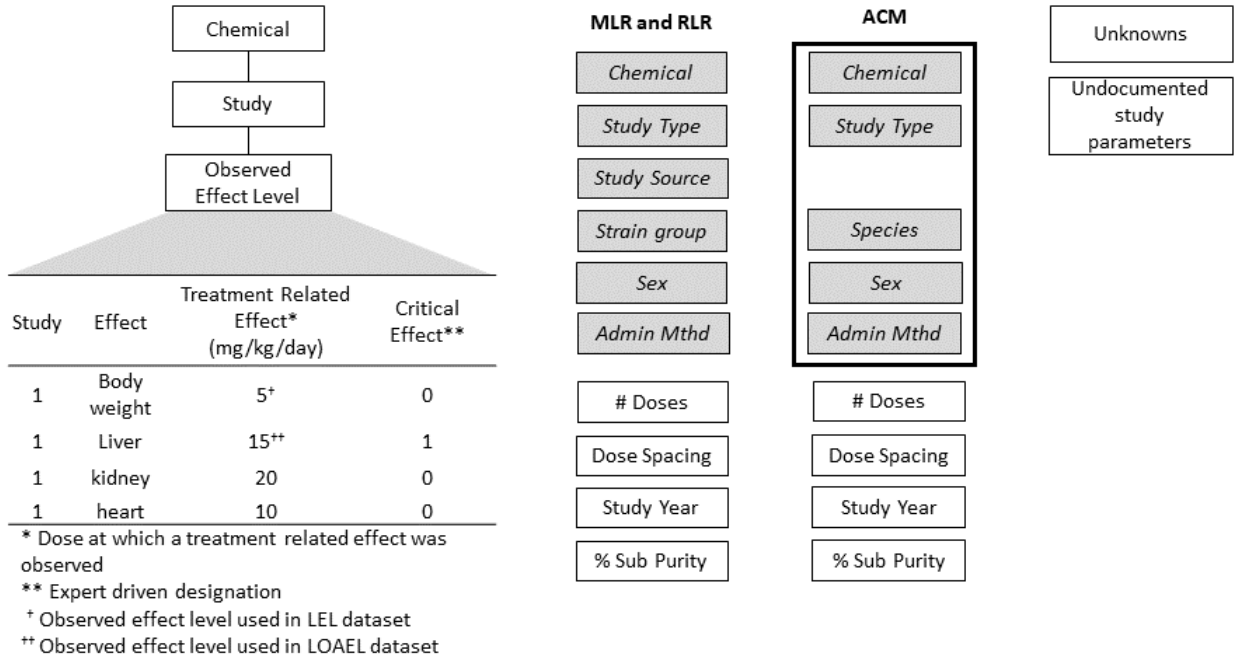
Primary Research Question	Statistical approaches
What is the range of possible effect values (mg/kg/day) in replicate studies for a given chemical?	• Residual root mean square error (RMSE) is an estimate of variance in the same units as the systemic effect values. • The RMSE can also be used to define 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). • This unexplained variance limits the 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:

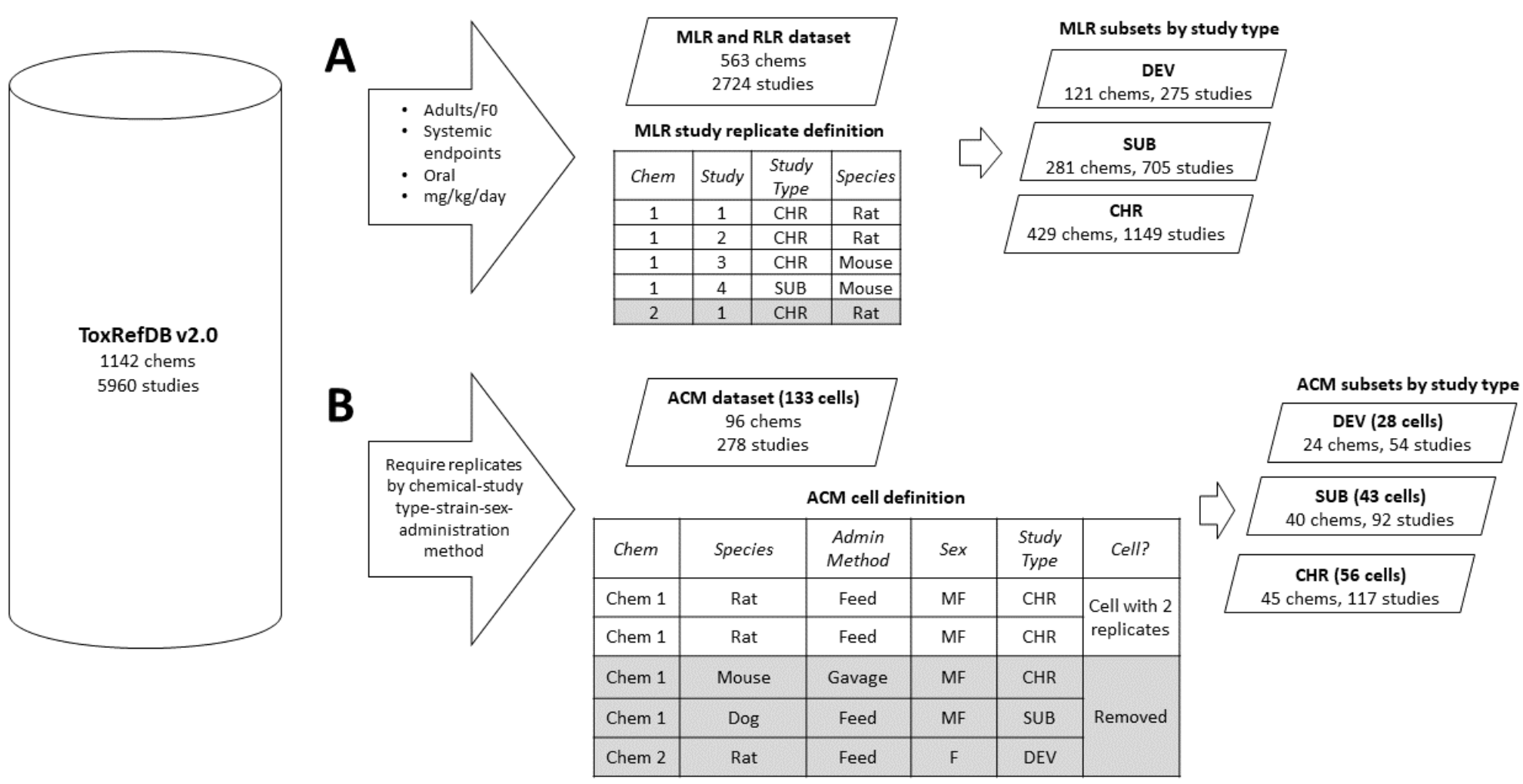
	Multilinear regression (MLR, RLR)	Augmented cell means (ACM)
Aggregation level	Chemical	Chemical-Study Type-Species-Sex-Admin Method combination
Replicate definition stringency	Not stringent	Stringent
N	Maximized; ↓ impact of outliers/database error rate	Small; may bias variance estimate
Study descriptors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

.

Our workflow for evaluating variance in repeat dose toxicity information



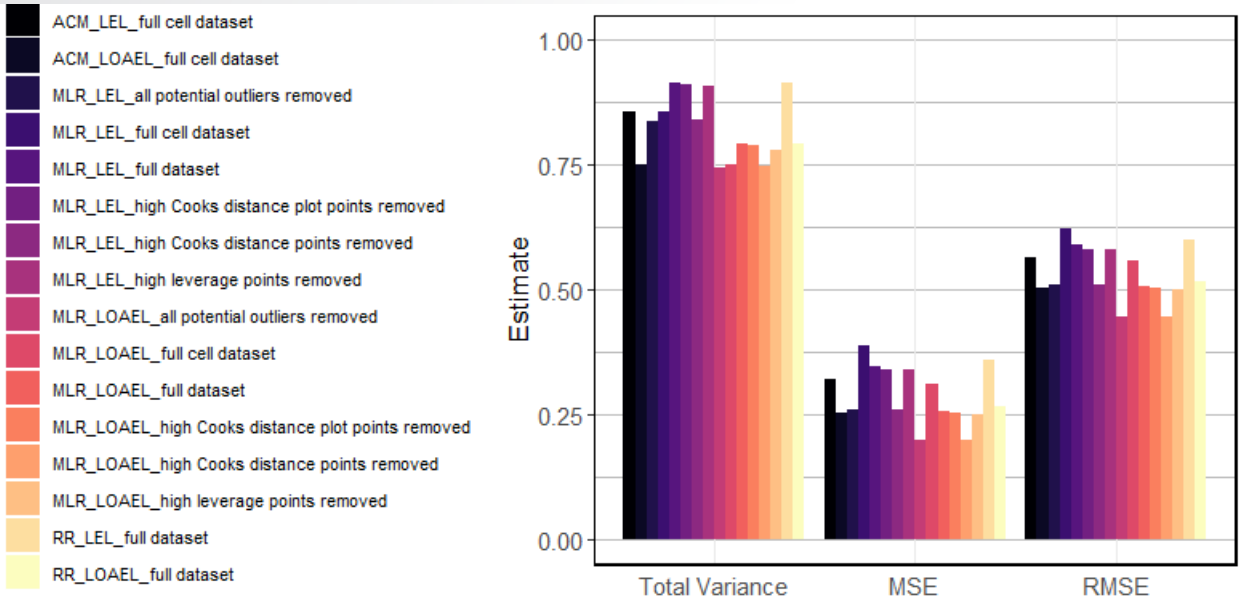
Chem	*Study*	*Study Type*	*Species*
1	1	CHR	Rat
1	2	CHR	Rat
1	3	CHR	Mouse
1	4	SUB	Mouse
2	1	CHR	Rat

Chem	*Species*	*Admin Method*	*Sex*	*Study Type*	*Cell?*
Chem 1	Rat	Feed	MF	CHR	Cell with 2 replicates
Chem 1	Rat	Feed	MF	CHR	
Chem 1	Mouse	Gavage	MF	CHR	Removed
Chem 1	Dog	Feed	MF	SUB	
Chem 2	Rat	Feed	F	DEV	

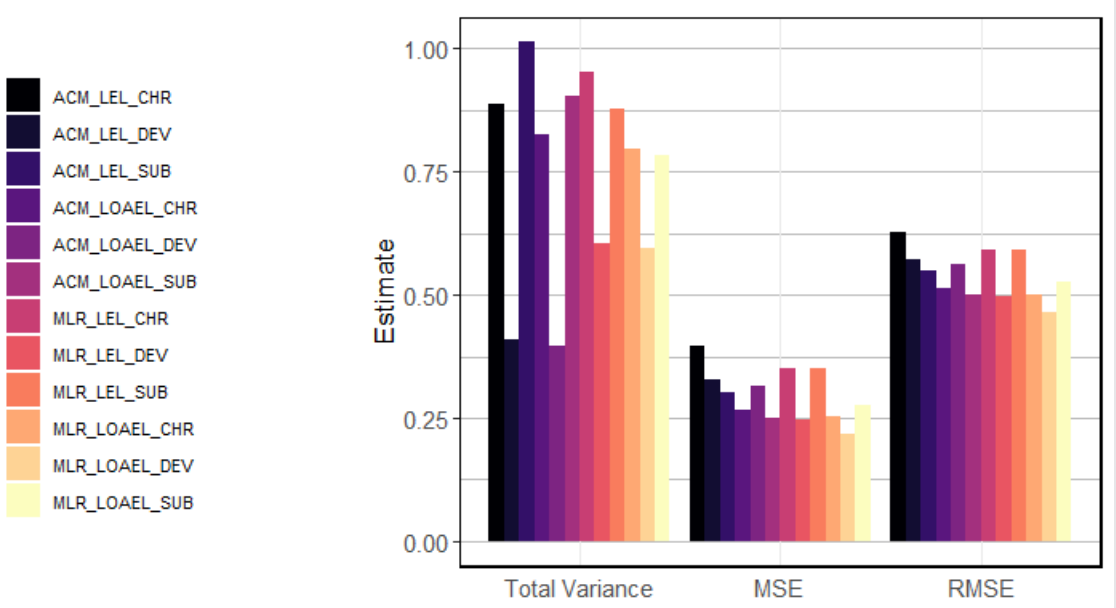
Figure 1. Variance estimation workflow.
CHR = chronic; DEV = developmental (adults only); SUB = subchronic; cells are defined by the factor of all categorical variables; MF = males and females; F = females; MLR = multilinear regression; RLR = robust linear regression; ACM = augmented cell means.

28 models to approximate total variance, unexplained variance (MSE), and then the spread of the residuals from the statistical models (RMSE)

Statistical models for LELs and LOAELs for the full dataset



Statistical models for LELs and LOAELs for datasets subset by study type



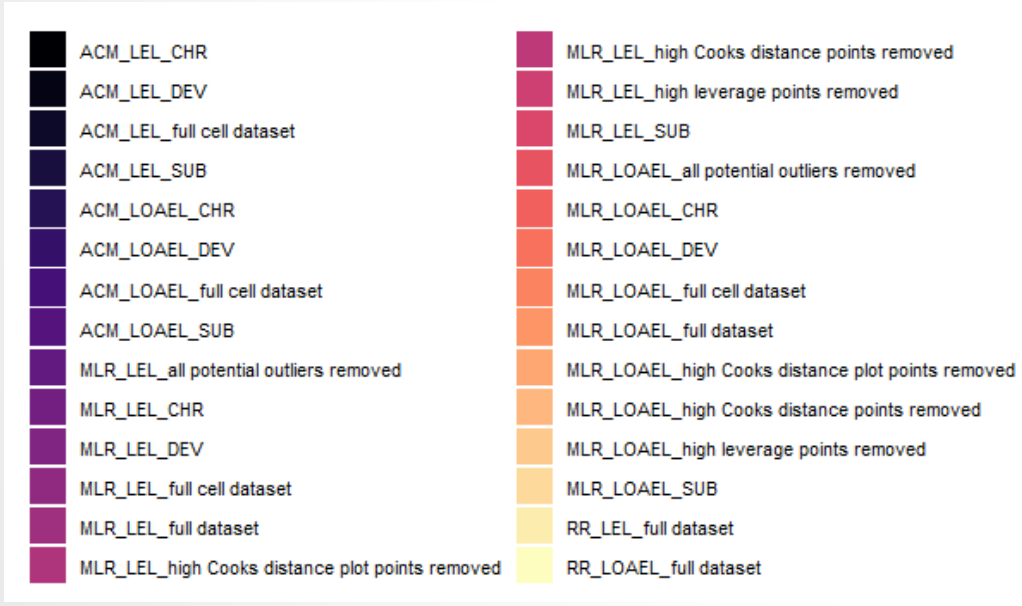
- Total variance in systemic toxicity effect values likely approaches 0.75-1 (units of $(\log_{10}\text{-mg/kg/day})^2$)
- MSE (unexplained variance) is 0.2 – 0.4 (units of $(\log_{10}\text{-mg/kg/day})^2$)
- RMSE is 0.45-0.60 $\log_{10}$-mg/kg/day
- RMSE is used to define a 95% minimum prediction interval (i.e., based on the standard deviation or spread of the residuals)

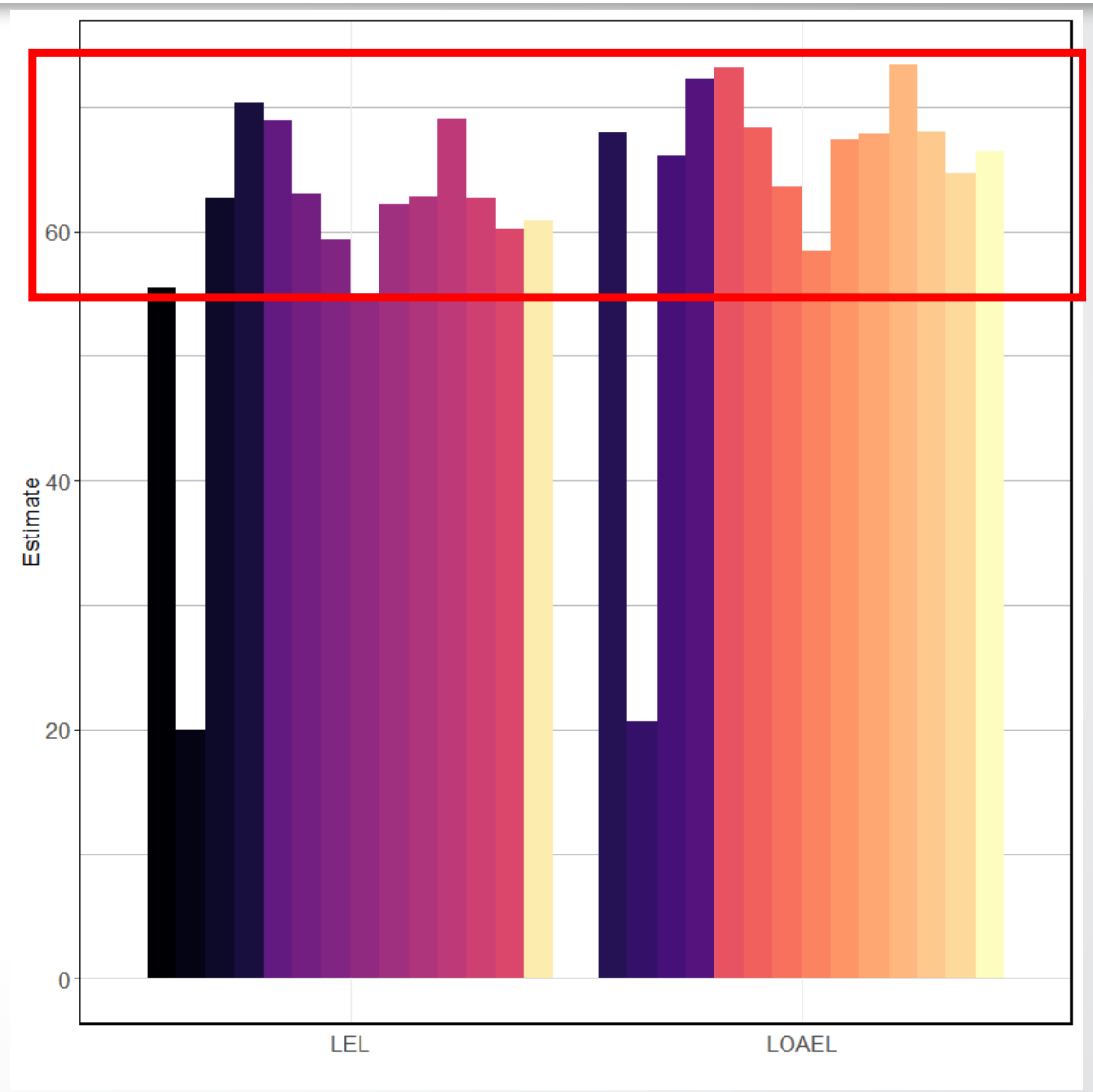
Based on tables from Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

.

Percent explained variance is also stable across statistical models.

- 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.

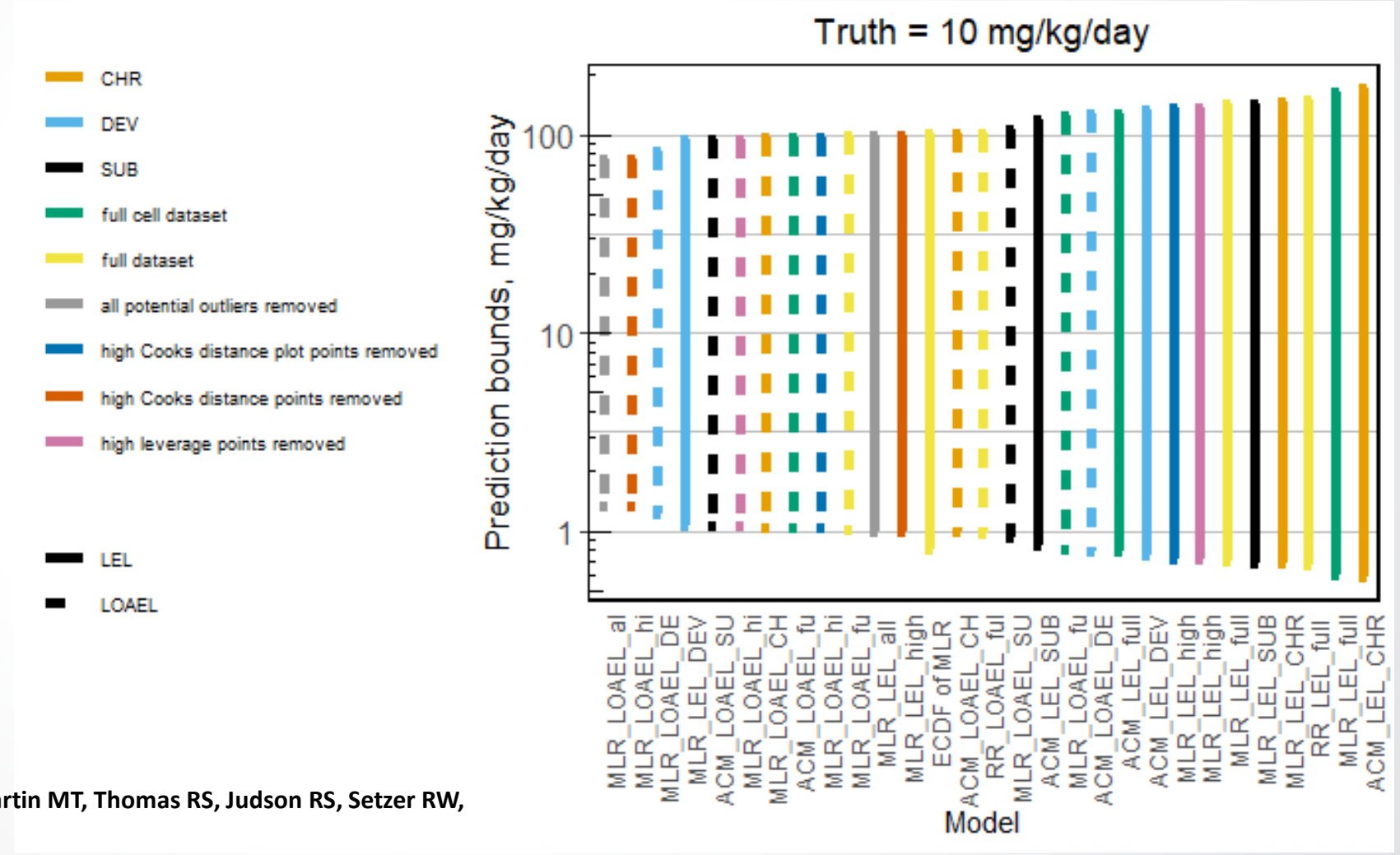




Based on tables from Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

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.



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: Incorporating estimates of variance into QSARs for repeat dose

New Journal and we have not received input yet 16 (2020) 100139

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology

ELSEVIER

Structure-based QSAR models to predict repeat dose toxicity points of departure

Prachi Pradeep [a,b,*], Katie Paul Friedman [b], Richard Judson [b]

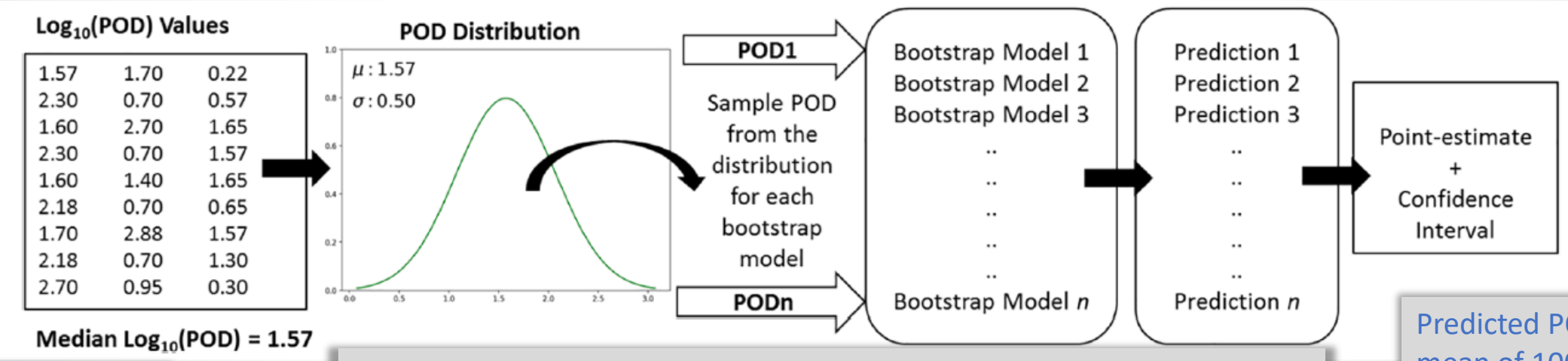
[a] Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA
[b] Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

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

Primary Research Question	Statistical approaches
Does expanding the training data set for repeat dose point-of-departure (POD) values improve performance?	• Point-estimate QSARs for POD values were developed and compared using performance metrics (RMSE and R^2).
Can POD_{QSAR} values be represented as a confidence interval based on estimates of variance in training data?	• **QSAR models were constructed using a constructed POD distribution and bootstrap resampling to represent the confidence interval for POD_{QSAR} values.**

Data variability should inform model uncertainty

- A model gives a result (a POD), but this is an estimate of the "true" POD. The true POD is unknown.
- Variability and uncertainty in the reference/training data will lead to variability and uncertainty in the model and our estimate of its quality.

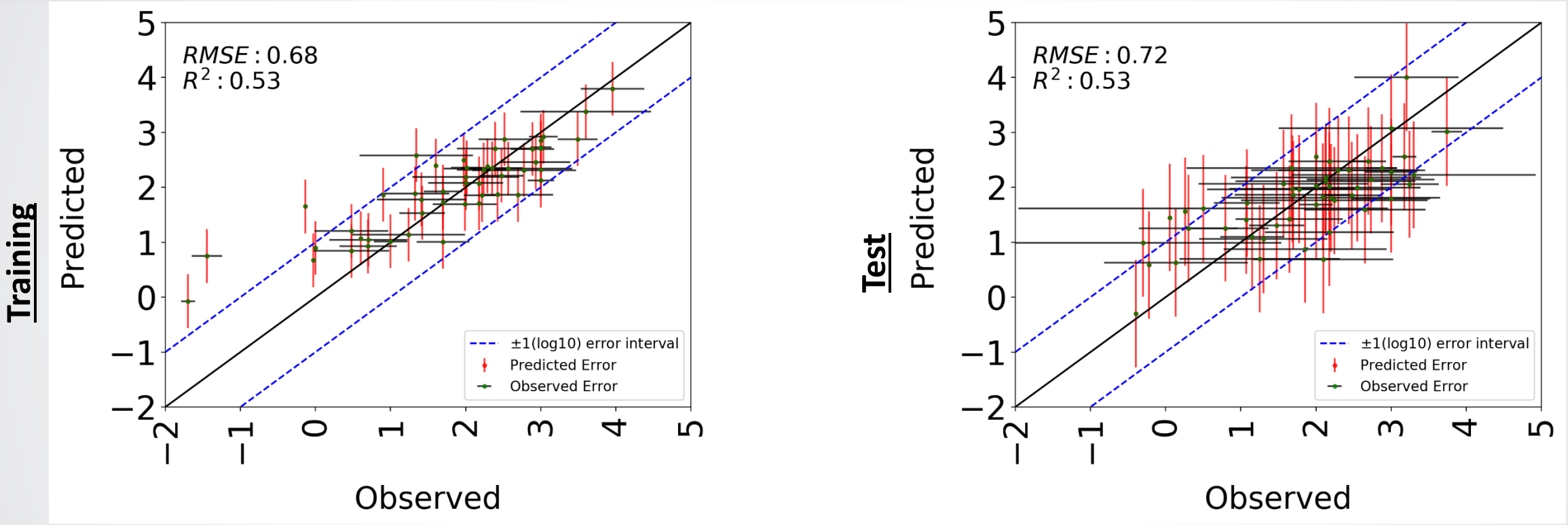


100 bootstrap models were built with random sampling of POD values for each chemical from the pre-generated POD distribution.

A POD distribution was constructed for each chemical (μ = Median experimental POD value from all studies, σ = 0.5 log_{10}-units)

Predicted POD_{QSAR} = mean of 100 bootstrap predictions
Confidence interval of POD_{QSAR} = ±2 standard deviation of 100 bootstrap predictions

A systemic toxicity prediction informed by variability: POD_{QSAR}



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

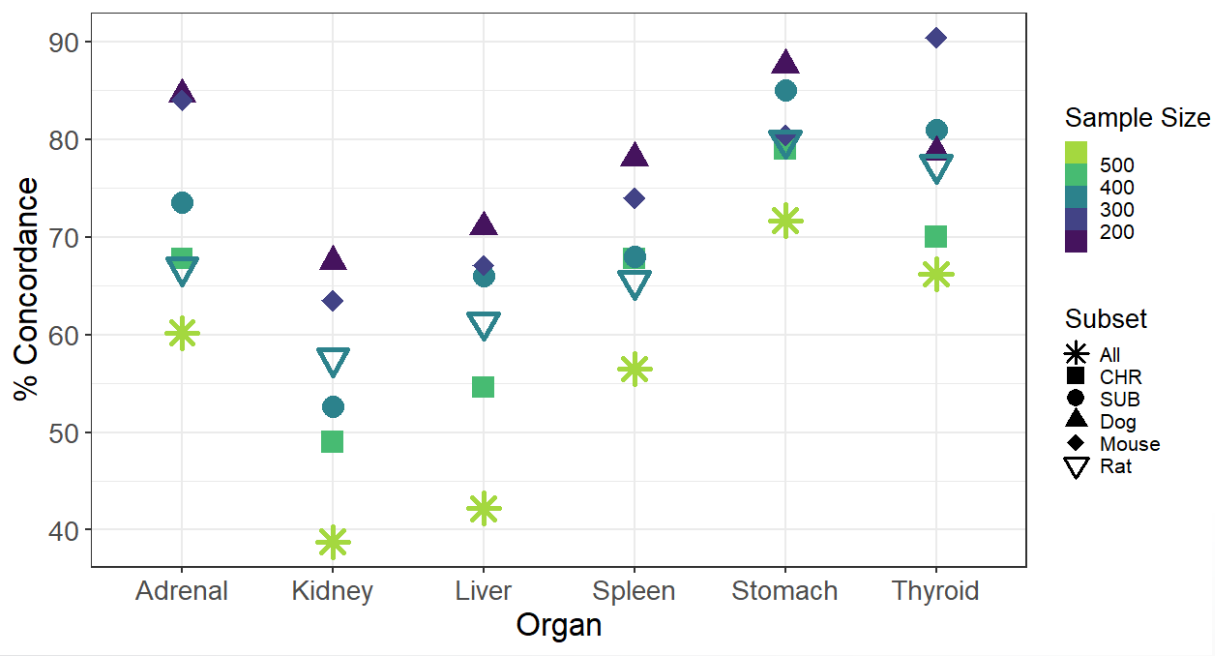
Enrichment analysis to evaluate the accuracy of POD_{QSAR} showed that 80% of the 5% most potent chemicals were found in the top 20% of the most potent chemical predictions, suggesting that the repeat dose POD_{QSAR} models presented here may help inform screening level human health risk assessments in the absence of other data.

Part III: Benchmarks on qualitative reproducibility of organ-level findings

Paul Friedman et al. (unpublished). Reproducibility of organ-level effects in repeat dose animal 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

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



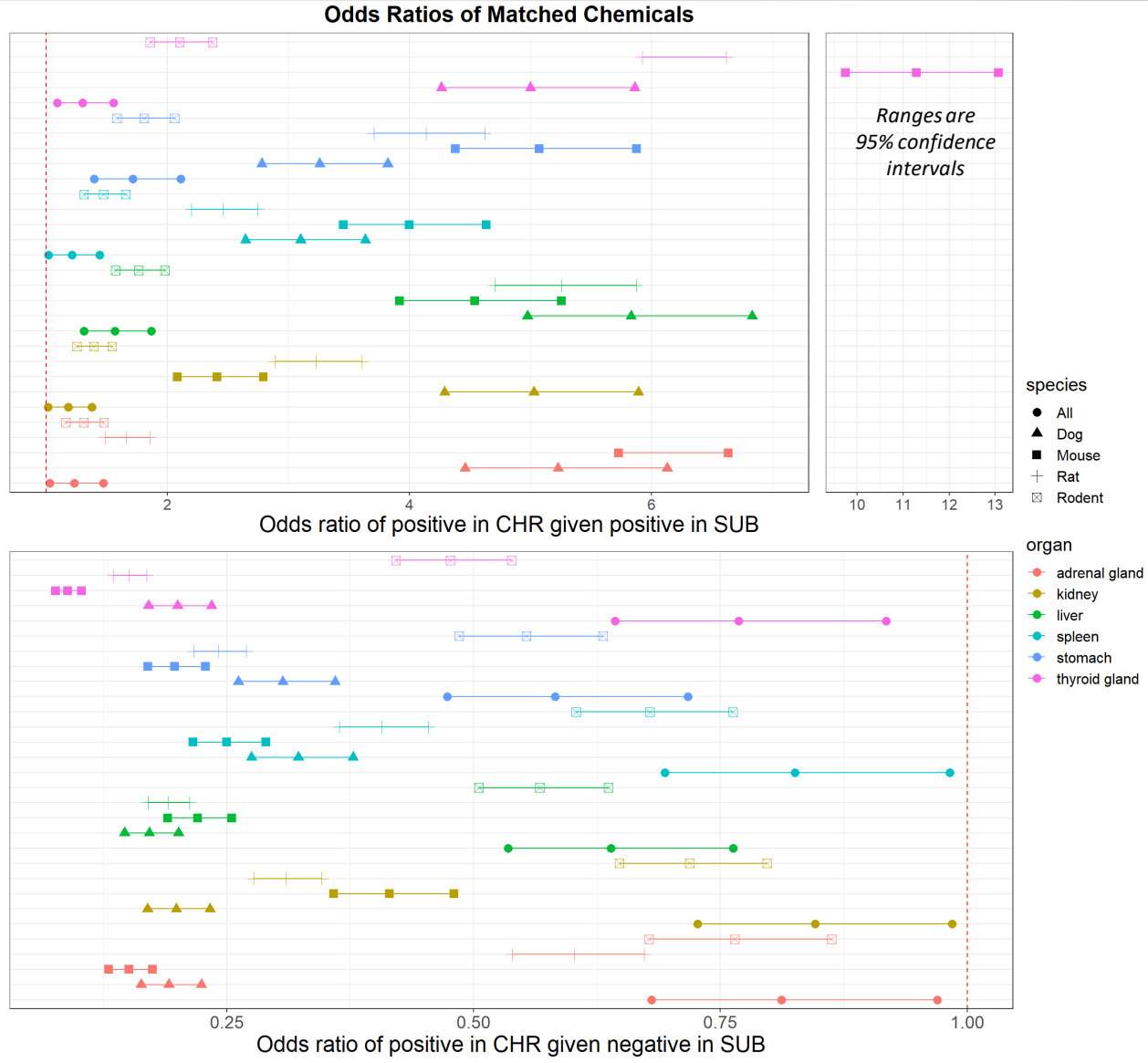
- Qualitative reproducibility of organ-level effect observations in repeat dose studies of adult animals was 33-88%, depending on grouping
- Organs associated with more negative chemicals (stomach, thyroid, adrenal) had higher rates of concordance
- Within-species concordance tended to be greater than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

If a substance failed to produce effects in a target organ at 90 days, what are the odds there would be a positive at 2 years?

Paul Friedman et al. (unpublished). Reproducibility of organ-level effects in repeat dose animal studies.

Primary Research Question	Statistical approaches
Can target organ effects be identified using subchronic (90 day) studies? 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

- A positive in SUB tends to indicate a greater likelihood of a positive in CHR at that tissue, with some variability by species and tissue.
- The odds ratio for a CHR positive given a SUB negative for each of these target organs was less than 1 in all cases, indicating that a negative in the SUB indicates a greater likelihood of negative in the CHR.
- *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 any chronic finding.*



Conclusions: How does this compare to previous work in this area?

- Previous QSAR models of subchronic oral rat NOAEL values: R^2 approaches 0.46-0.71, i.e. 46-71% of residual variance could be explained for the reference set (Veselinovic et al. 2016; Toropov et al. 2015; Toropova et al. 2017).
- 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).

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

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

Few examples of quantitative variability in repeat dose PODs but suggest that similar thresholds of 50-70% explained variance and RMSE of 0.5-0.7 may exist in other larger reference data sets for systemic toxicity in subchronic and chronic animal studies.

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: Variability in *in vivo* training set data should be incorporated in QSARs for predicting points-of-departure.
- Part III: Qualitative reproducibility of organ-level effect observations in repeat dose studies of adult animals was 33-88%, with highest concordance within species
  - Subchronic and chronic *in vivo* observations can be combined for modeling to increase N, and 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.
- Finally, 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

References

Congress, U. S., FRANK R. LAUTENBERG CHEMICAL SAFETY FOR THE 21ST CENTURY ACT. In: Congress, (Ed.), H.R.2576, Vol. Public Law 114-182, 2016.

Dumont, C., et al. (2016). "Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches." Toxicol In Vitro **34: 220-228.**

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.**

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.

Wheeler, A. R., Memorandum: Directive to Prioritize Efforts to Reduce Animal Testing. US Environmental Protection Agency, Washington, D.C., 2019.

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)