Variability in repeat dose toxicity studies: Implications for scientific confidence in NAMs

Katie Paul Friedman, PhD

November 11, 2021

2021 Nordic Training on New Approach Methodologies (virtual)

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

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

- In US, Section 4(h) in the Lautenberg amendment to TSCA:
  - "…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."

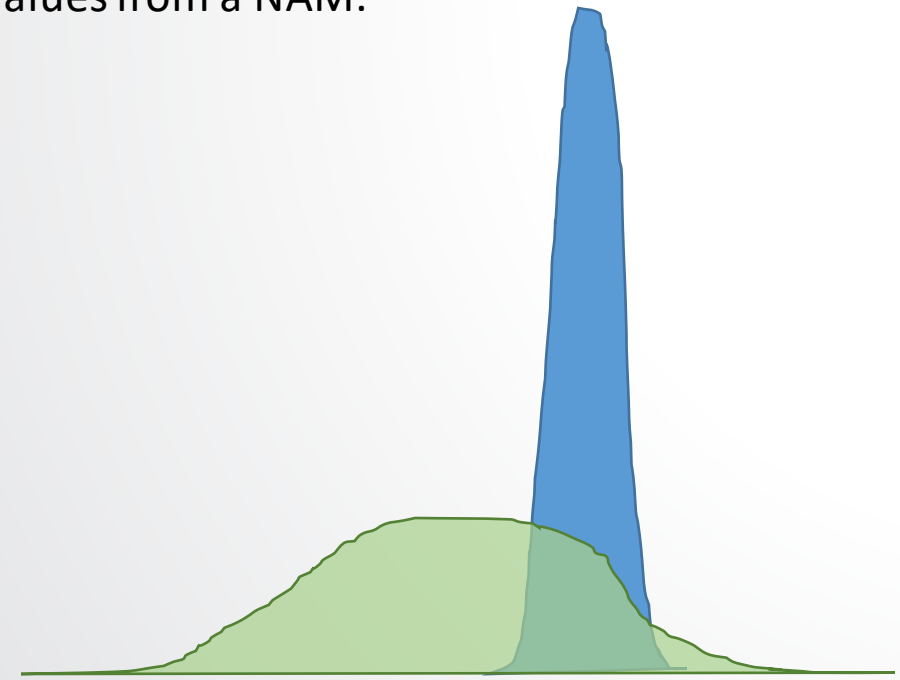
How do we define expectations of *in silico, in chemico,* and *in vitro* models for predicting repeat-dose toxicity?

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

How do we express variability in traditional animal toxicity tests?

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

Research questions for understanding this variability

3 main questions	What is the range of possible systemic effect values (mg/kg/day) in replicate studies?	What is the maximal accuracy of a model that attempts to predict a systemic effect values for an unknown chemical?	What is the probability that an effect in adult animals will be observed in replicate studies?
Statistical approach to the question	• 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.	• 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.	• Understand the reproducibility of treatment-related changes in specific endpoint targets (e.g., any effect on liver).

ToxRefDB v2.0 is a source for a dataset to address these questions of quantitative variability.

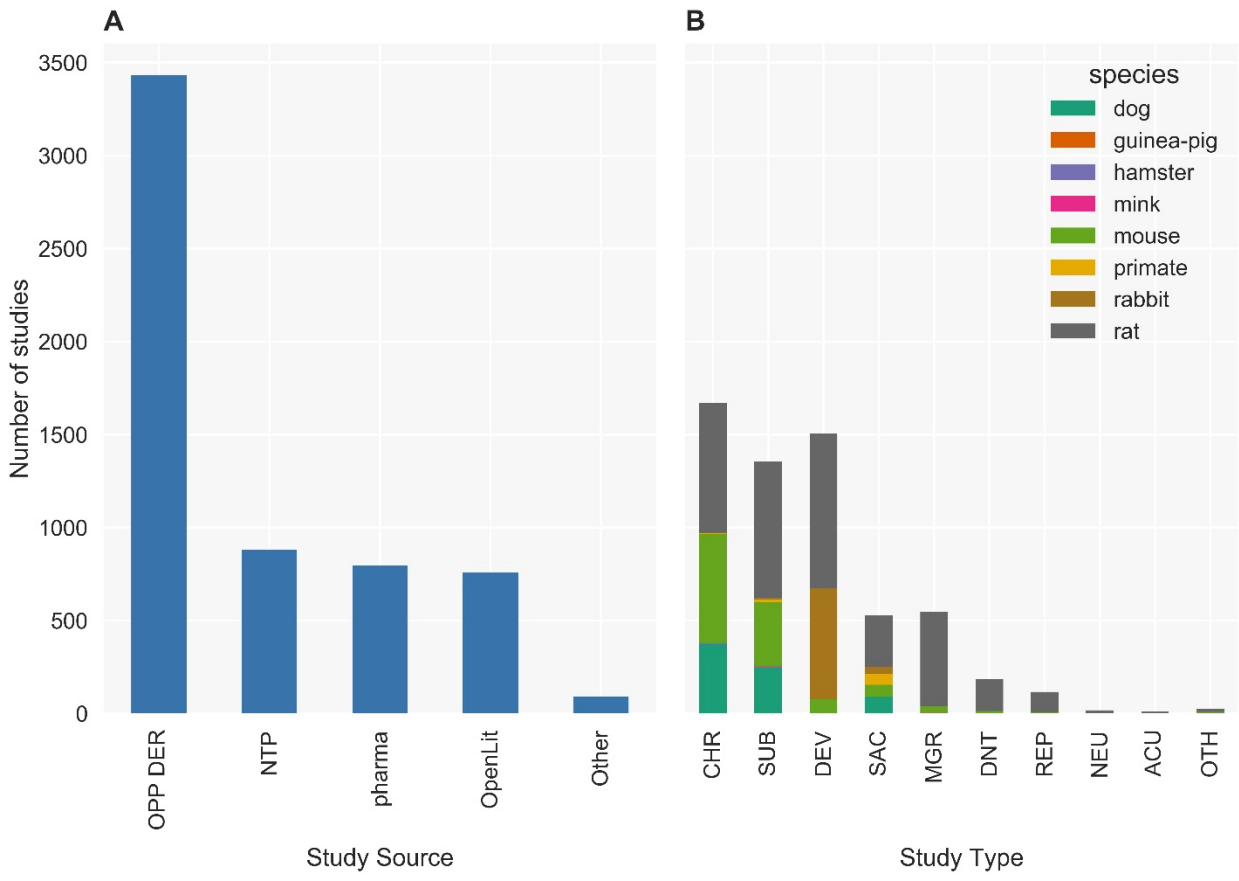


Figure 1. Number of studies by study type and species in ToxRefDB v2.0. *The study designs include chronic (CHR), sub-chronic (SUB), developmental (DEV), subacute (SAC), multigeneration reproductive (MGR), developmental neurotoxicity (DNT), reproductive (REP), neurotoxicity (NEU), acute (ACU), and other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

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

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**

Chemical
Study
Observed Effect Level

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

MLR and RLR

- *Chemical*
- *Study Type*
- *Study Source*
- *Strain group*
- *Sex*
- *Admin Mthd*
- # Doses
- Dose Spacing
- Study Year
- % Sub Purity

ACM

- *Chemical*
- *Study Type*
- *Species*
- *Sex*
- *Admin Mthd*
- # Doses
- Dose Spacing
- Study Year
- % Sub Purity

- Unknowns
- Undocumented study parameters

Using two approaches:

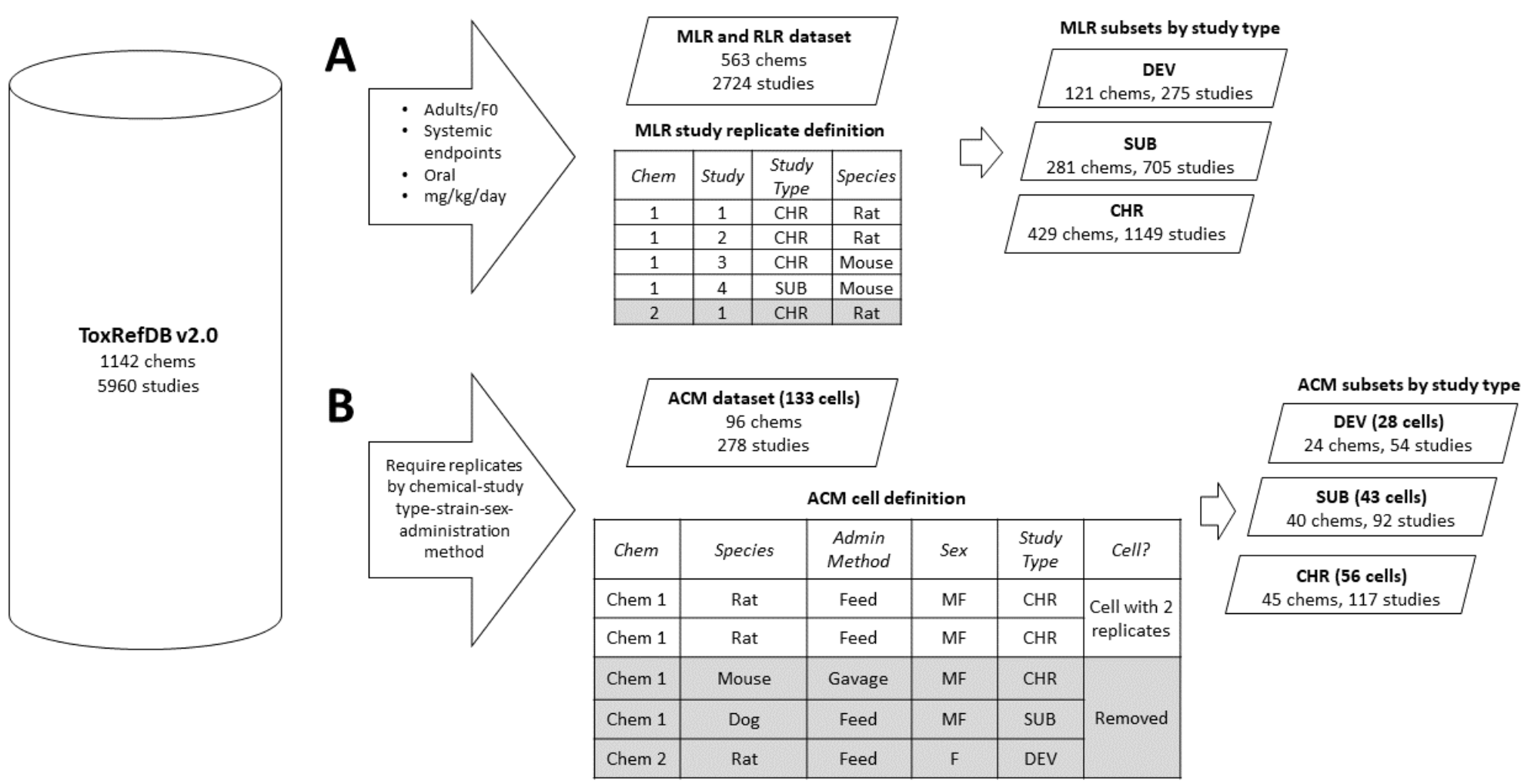
	Multilinear regression (MLR, RLR)	Augmented cell means (ACM)
Aggregation level	Chemical	Chemical-Study Type-Species-Sex-Admin Method combination
Replicate definition stringency	Not stringent	Stringent
N	Maximized; ↓ impact of outliers/database error rate	Small; may bias variance estimate
Study descriptors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



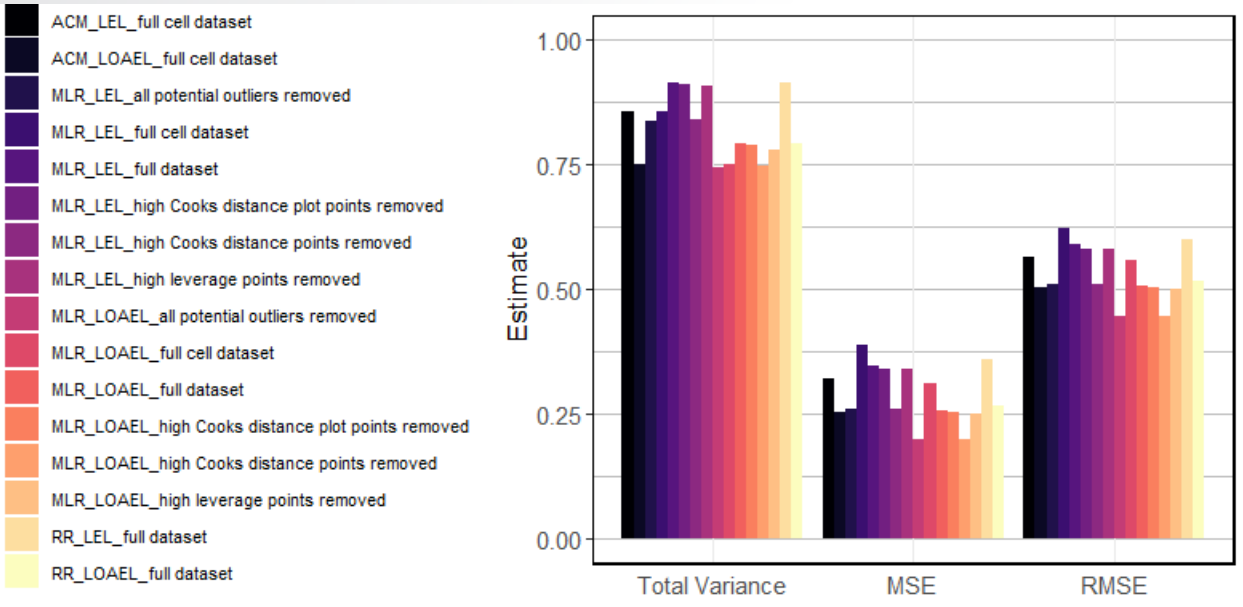
Chem	*Study*	*Study Type*	*Species*
1	1	CHR	Rat
1	2	CHR	Rat
1	3	CHR	Mouse
1	4	SUB	Mouse
2	1	CHR	Rat

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	

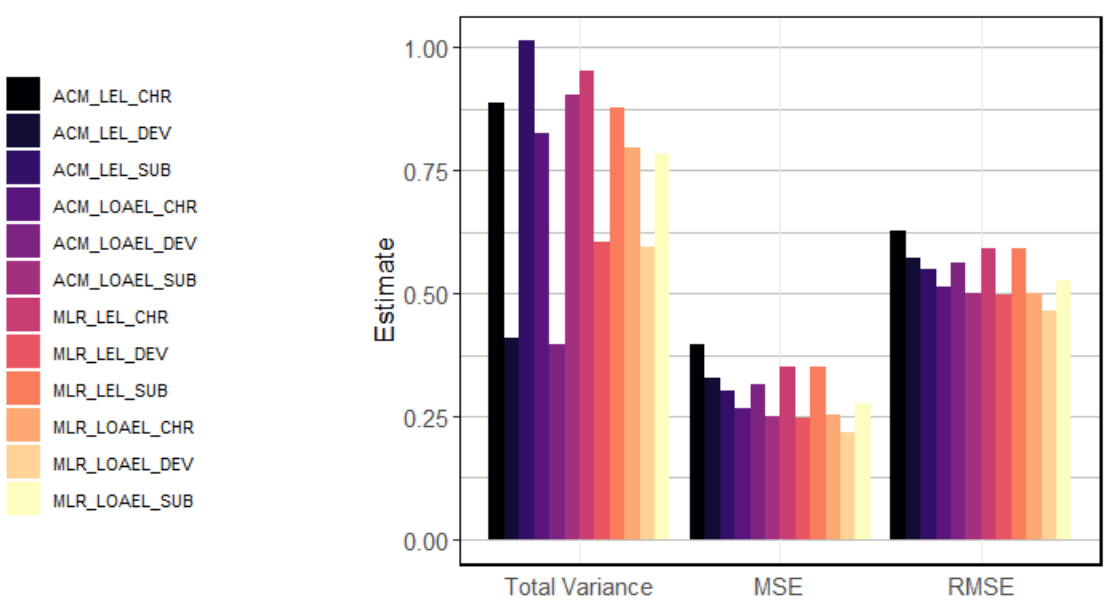
Figure 1. Variance estimation workflow.
CHR = chronic; DEV = developmental (adults only); SUB = subchronic; cells are defined by the factor of all categorical variables; MF = males and females; F = females; MLR = multilinear regression; RLR = robust linear regression; ACM = augmented cell means.

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



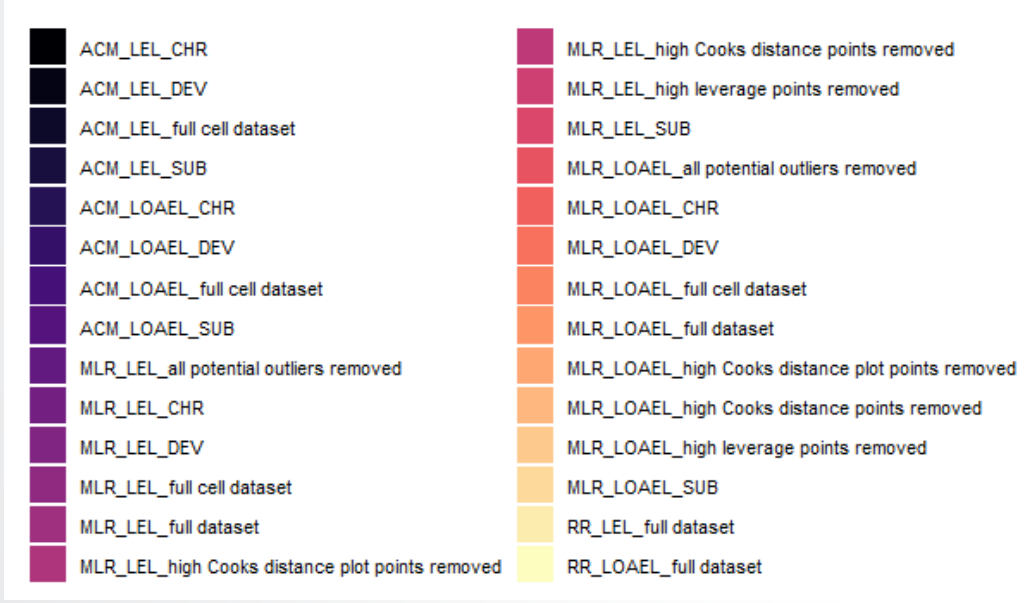
- Total variance in systemic toxicity effect values likely approaches 0.75-1 (units of $(\log_{10}\text{-mg/kg/day})^2$)
- MSE (unexplained variance) is 0.2 – 0.4 (units of $(\log_{10}\text{-mg/kg/day})^2$)
- RMSE is 0.45-0.60 $\log_{10}$-mg/kg/day
- RMSE is used to define a 95% minimum prediction interval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

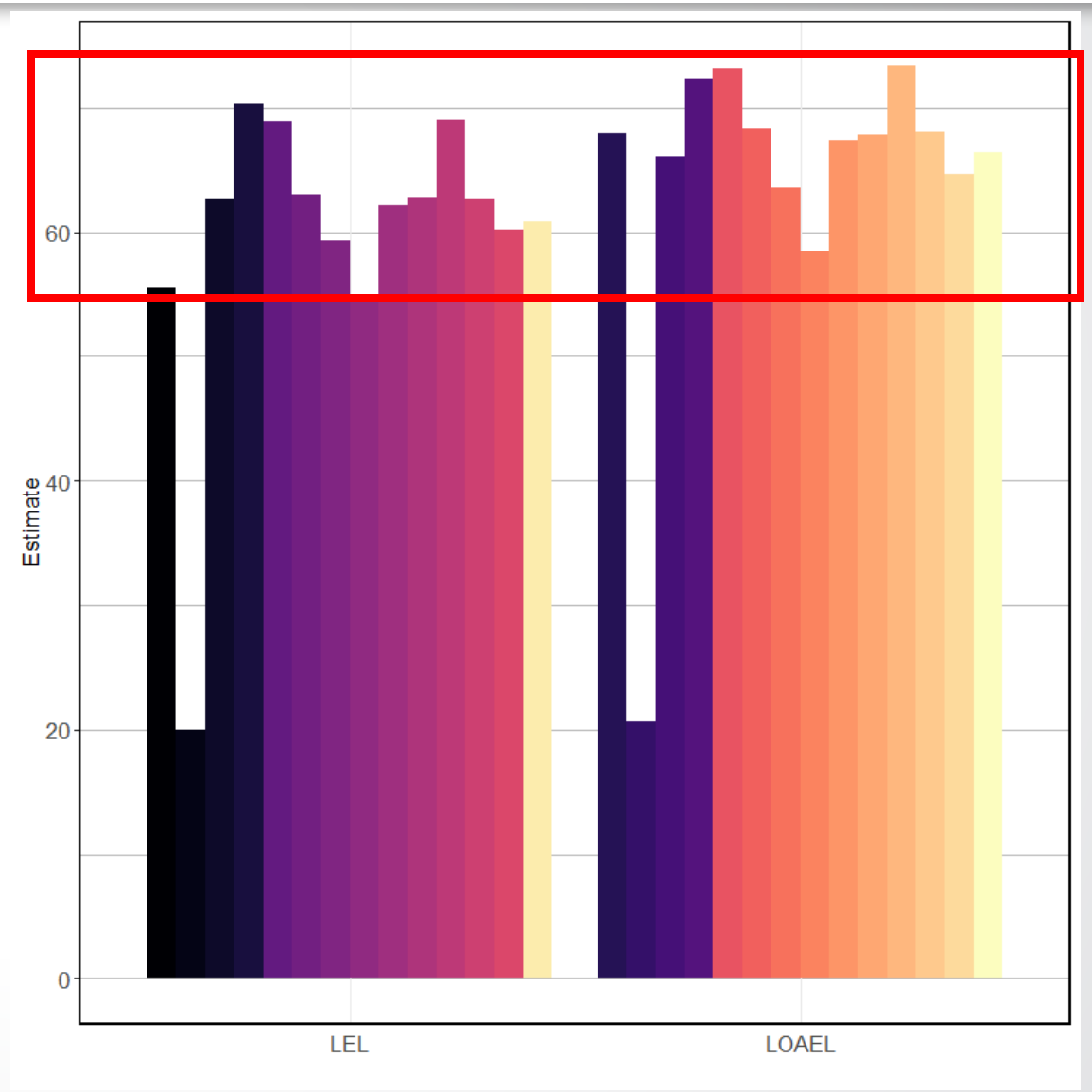
Based on tables from Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

.

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.

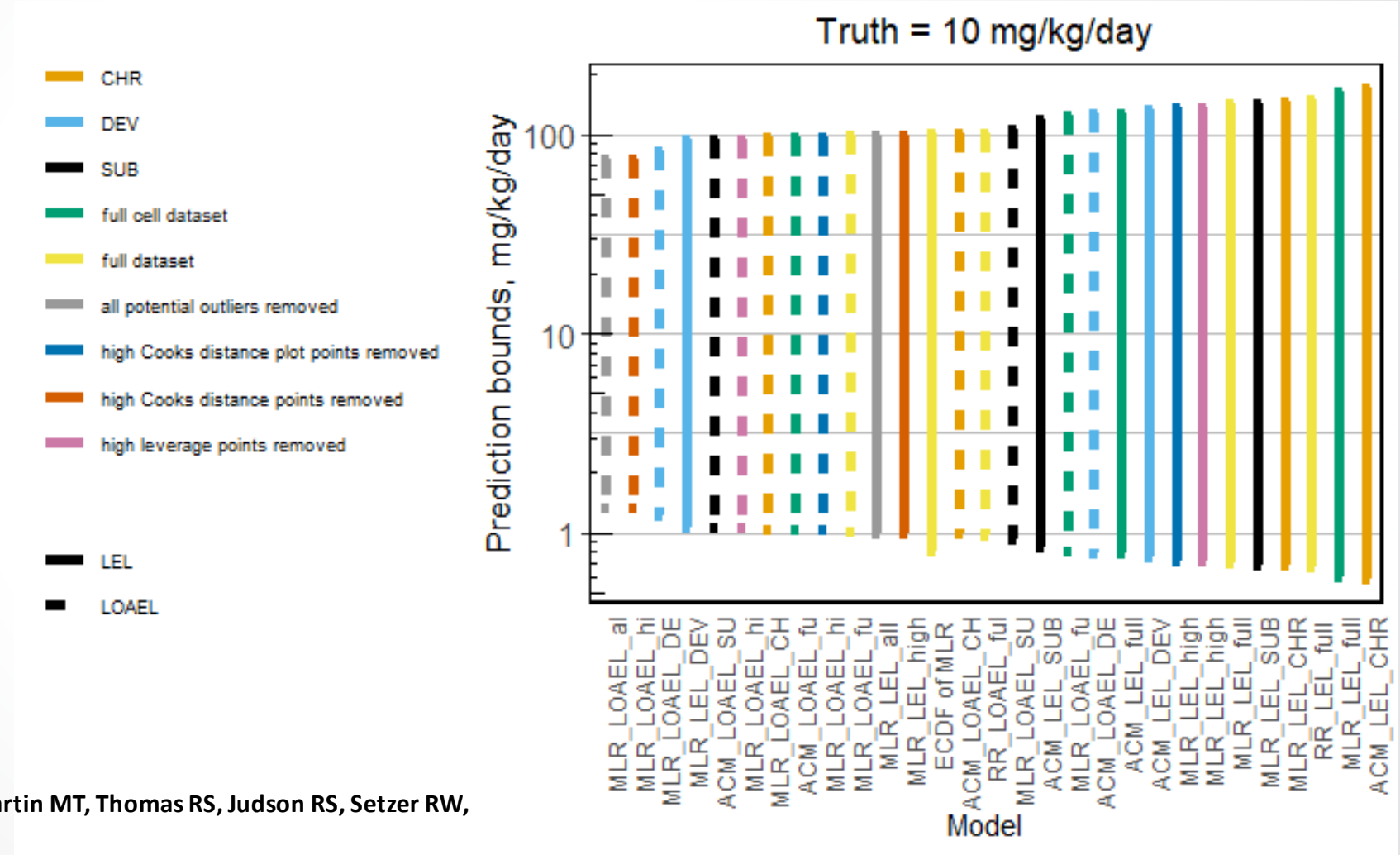




Based on tables from Pham LL, Watford S, Pradeep P, Martin MT, Thomas RS, Judson RS, Setzer RW, Paul Friedman K. 2020. 10.1016/j.comtox.2020.100126

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

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).

Few examples of quantitative variability in this domain to cite, 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.

Primary conclusions of our work

- Variability in *in vivo* toxicity studies limits predictive accuracy of NAMs.
- Total variance in systemic effect levels and the fraction explained were quantified.
- 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.
- The estimate of variance (RMSE) in curated LELs and/or LOAELs approaches a 0.5 log10-mg/kg/day.
- **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.**

Data variability informs model uncertainty

Model Uncertainty

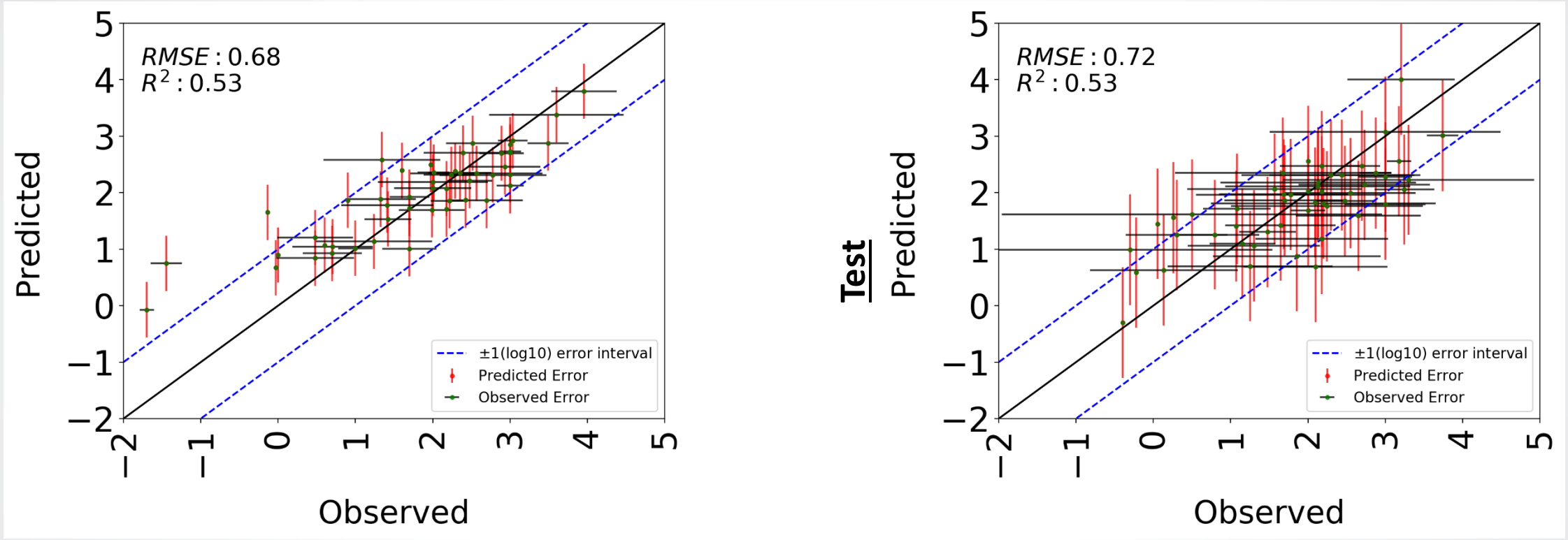
- A model gives a result (a POD), but this is an estimate of the "true" POD. The true POD is mostly unknown.
- Uncertainty in the evaluation data will lead to uncertainty in the model and our estimate of its quality

Point-estimate with confidence interval models

- A POD distribution was constructed for each chemical (μ = Median experimental POD value from all studies, σ = 0.5 log_{10}-units)
- *100* bootstrap models were built with random sampling of POD values for each chemical from the pre-generated POD distribution.
- Predicted POD_{QSAR} = mean of 100 bootstrap predictions
- Confidence interval of POD_{QSAR} = ±1 standard deviation of 100 bootstrap predictions

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

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



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

Observed versus predicted plot for 50 (random) chemicals with the observed and predicted confidence intervals

- The predicted 95% confidence interval (error bar) for each chemical is calculated as two standard deviations of the predictions from the models.
- The observed 95% confidence interval (error bar) is calculated as two standard deviations of the experimental data for each chemical.

Research questions for understanding this variability

3 main questions	What is the range of possible systemic effect values (mg/kg/day) in replicate studies?	What is the maximal accuracy of a model that attempts to predict a systemic effect values for an unknown chemical?	What is the probability that an effect in adult animals will be observed in replicate studies?
Statistical approach to the question	• 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.	• 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.	• Understand the reproducibility of treatment-related changes in specific endpoint targets (e.g., any effect on liver).

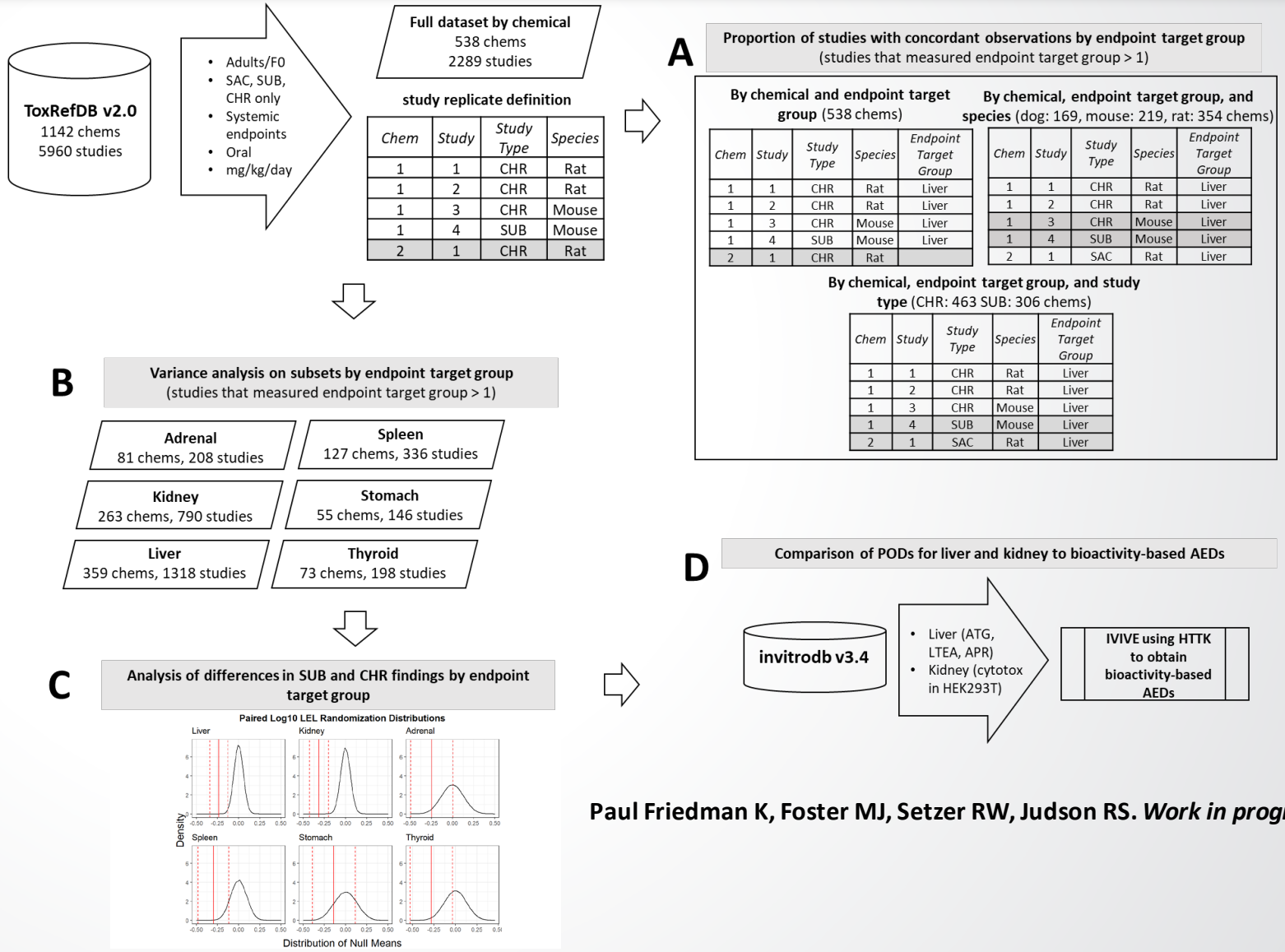
Preliminary outline of this work

(A) What is the qualitative reproducibility of organ-level findings across replicate repeat dose studies in animals?

(B) What is the quantitative variability of organ-level findings across replicate repeat dose studies in animals?

(C) If a NAM can predict an organ-level POD, is it necessary to adjust this POD to create separate predictions of subchronic and chronic organ-level effects?

(D) Can targeted NAMs predict liver or kidney level point-of-departure (POD) within the reference POD ± X*RMSE?



Paul Friedman K, Foster MJ, Setzer RW, Judson RS. *Work in progress.*

Table. Repeated concordance of organ-level findings.

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

% Concord = percent concordant chemicals; Chem = total # chemicals tested at the endpoint target group; +Pos = # chemicals with positive observations in all available studies; -Neg = # chemicals with negative observations in all available studies; Mixed = chemicals with at least 1 study that was not positive

Endpoint target group	% Concord	Chem	+Pos	-Neg	Mixed
adrenal	**60.2**	538	8	316	214
kidney	**38.8**	538	54	155	329
Liver	**42.4**	538	149	79	310
spleen	**56.5**	538	17	287	234
stomach	**71.7**	538	14	372	152
thyroid	**66.2**	538	11	345	182

Endpoint target group	Study Type	% Concord	Chem	+Pos	-Neg	Mixed
adrenal	CHR	**67.8**	463	8	306	149
kidney		**49**	463	58	169	236
liver		**54.6**	463	160	93	210
spleen		**67.8**	463	16	298	149
stomach		**79**	463	22	344	97
thyroid		**70**	463	10	314	139
adrenal	SUB	**73.5**	306	10	215	81
kidney		**52.6**	306	65	96	145
liver		**66**	306	143	59	104
spleen		**68**	306	24	184	98
stomach		**85**	306	10	250	46
thyroid		**81**	306	11	237	58

Endpoint target group	Species	% Concord	Chem	+Pos	-Neg	Mixed
adrenal	dog	**84.6**	169	8	135	26
	mouse	**84**	219	6	178	35
	rat	**66.9**	354	17	220	117
kidney	dog	**67.5**	169	20	94	55
	mouse	**63.5**	219	43	96	80
	rat	**57.6**	354	106	98	150
liver	dog	**71**	169	86	34	49
	mouse	**67.1**	219	96	51	72
	rat	**61.3**	354	157	60	137
spleen	dog	**78.1**	169	9	123	37
	mouse	**74**	219	16	146	57
	rat	**65.5**	354	31	201	122
stomach	dog	**87.6**	169	2	146	21
	mouse	**80.4**	219	7	169	43
	rat	**79.9**	354	11	272	71
thyroid	dog	**78.7**	169	8	125	36
	mouse	**90.4**	219	3	195	21
	rat	**77.4**	354	28	246	80

Table. Results of MLR to estimate unexplained and explained variance in organ LELs.

$$organLEL \sim b_0 + chemical * b_1 + species * b_2 + study\ type * b_3 + administration\ method * b_4 + dose\ spacing * b_5 + number\ of\ dose\ levels * b_6 + study\ year * b_7 + \%\ substance\ purity * b_8$$

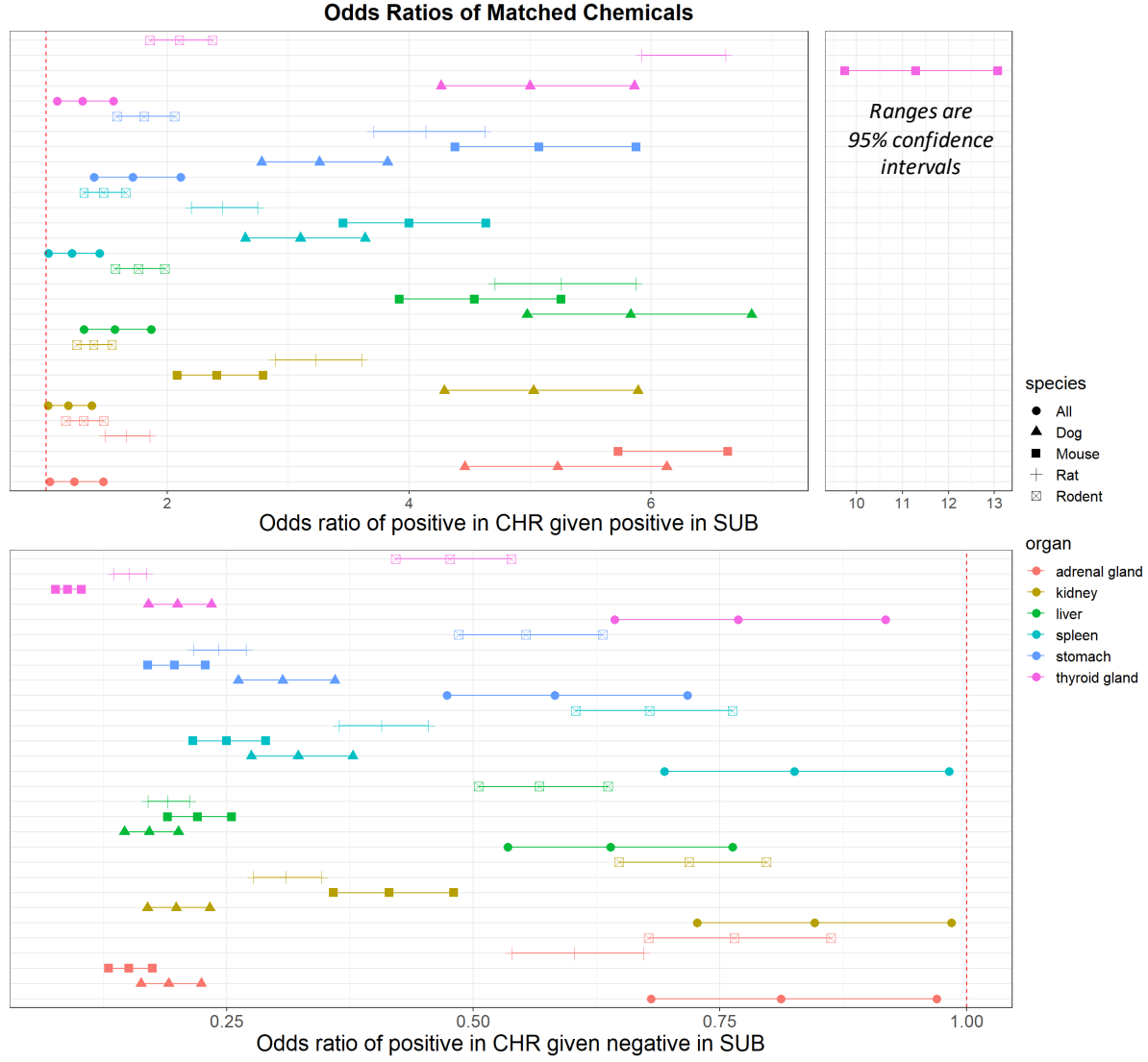
Chems = # chemicals; N = number of studies; Var = total variance; MSE = mean square error on the model; RMSE = root residual mean square error; % var explained = % of total variance explained.by study descriptors

Endpoint Target Group	Chem	N	Var	MSE	**RMSE**	% var explained
adrenal	81	208	0.756	0.349	**0.591**	53.8
kidney	263	790	0.765	0.316	**0.562**	58.7
liver	359	1318	0.745	0.355	**0.596**	52.3
spleen	127	336	0.671	0.318	**0.564**	52.6
stomach	55	146	0.553	0.173	**0.416**	68.7
thyroid	73	198	0.721	0.378	**0.615**	47.6

Total variance at the organ level is generally less than or equal to total variance at the study-level. The RMSE at the organ level is similar to the study level RMSE in Pham *et al.* The % variance explained is similar to the lower estimate of % variance explained at the study level in Pham *et al.*

Paul Friedman K, Foster MJ, Setzer RW, Judson RS. *Work in progress.*

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?



- Positive = any gross or histopathological change, or associated hormones (in the case of thyroid gland) or clinical chemistry (in the case of kidney)
- 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 positive 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 POD in a target organ at 90 days, particularly for liver and kidney where we have the largest datasets, is likely protective for any chronic finding.*

Paul Friedman K, Foster MJ, Setzer RW, Judson RS. *Work in progress.*

A randomization test of the ratio of CHR/SUB LEL values from ToxRefDB suggests that liver and kidney PODs are smaller for CHR studies

Interpretation: We are 95% confident that the log10 difference in CHR – SUB is on average between -0.1261 and -0.3416 for the liver data available in ToxRefDB.

We can also exponentiate (10^) this difference and turn this back into LELs, and this becomes a ratio. The LEL ratio of CHR/SUB for liver would between 0.4554 and 0.7479.

Organ	Observed Mean of log10(CHR-SUB)	Upper Bound	Lower Bound	P value
Liver	-0.2339	-0.1261	-0.3416	P<0.0001
Kidney	-0.3142	-0.201	-0.4274	P<0.0001
Adrenal	-0.2445	0.0057	-0.4948	0.054
Spleen	-0.2979	-0.1147	-0.481	0.0011
Stomach	-0.1383	0.1144	-0.3911	0.2991
Thyroid	-0.2817	-0.0357	-0.5276	0.0229

Paul Friedman K, Foster MJ, Setzer RW, Judson RS. *Work in progress.*

How much should administered equivalent doses (AEDs) be adjusted when predicting *in vivo* LELs?

- AEDs are the mg/kg/day external dose predicted to correspond to in in vitro bioactive concentrations, based on a reverse dosimetry approach that relates the in vitro bioactive concentration to the human plasma concentration.
- The goal of this organ-specific AED to LEL comparison is to understand the adjustment factor that might be needed when doing NAM-based assessments of repeat dose toxicity observed in target tissues.
- Here we only have enough data in liver and kidney for the union between tissue-specific assay endpoints in invitrodb and organ-level LELs in ToxRefDB.

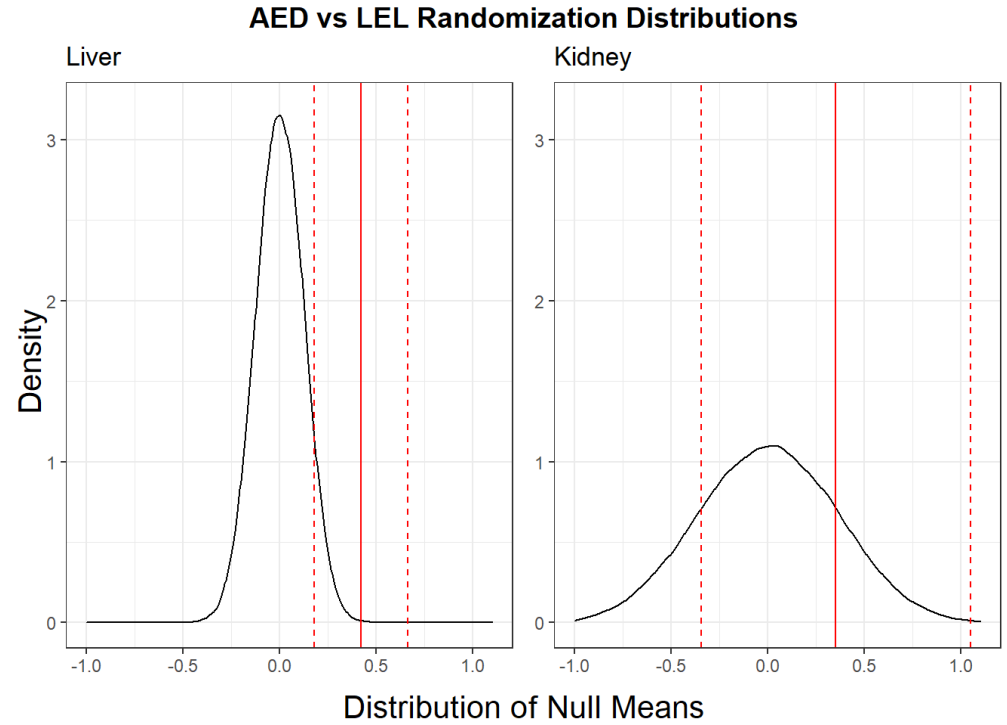
Organ	Number of unique substances
Liver	137
Kidney	25

LELs from ToxRefDBv2.0, calculated by organ

AEDs calculated using library(httk); 3 compartment steady state model and only assays that are in liver or kidney associated cell lines or primary cells

Paul Friedman K, Foster MJ, Setzer RW, Judson RS. *Work in progress.*

Preliminary work suggests that depending on the IVIVE approach, the AEDs by tissue may be within the estimate of variance in organ level LELs



Differences calculated as LEL-AED (log10-mg/kg/day)

Organ	Observed Mean Difference (log10LEL-log10AED)	Lower Bound	Upper Bound	P value
Liver	0.4193	0.1761	0.6626	0.0007
Kidney	0.3512	-0.3455	1.0478	0.3329

For liver, there is a statistically significant difference, but it is between 0.17 and 0.66 log10-mg/kg/day (with the AED being essentially 0.5 log10-mg/kg/day more conservative).

There are only 25 chemicals in the kidney dataset, which is relatively small for making inferences. However, the mean observed difference for kidney and liver is within the estimate of variance for replicate repeat dose studies.

Paul Friedman K, Foster MJ, Setzer RW, Judson RS. *Work in progress.*

Primary conclusions of our work

- Variability in *in vivo* toxicity studies limits predictive accuracy of NAMs.
- Total variance in systemic effect levels and the fraction explained were quantified.
- 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.**
- 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.
  - Existing QSAR for repeat dose POD may be informative for rapid workflows.
  - Work is in progress to support best practices for estimating *in vivo* PODs at the organ level.

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.

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