www.epa.gov

Variability in *in vivo* Toxicity Studies: Defining the upper limit of predictivity for models of systemic effect levels

Ly Ly Pham[1,2], Richard Judson[1], R. Woodrow Setzer[1], **Katie Paul Friedman[1*]**

[1]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 27711, USA
[2]Oak Ridge Institute for Science and Education, 100 ORAU Way, Oak Ridge, TN 37830.

ORCID: 0000-0002-2710-1691
Katie Paul Friedman | paul-friedman.katie@epa.gov

L.L.P. was supported by appointment to the Research Participation Program of the U.S. Environmental Protection Agency, Office of Research and Development, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. EPA. *This views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.*

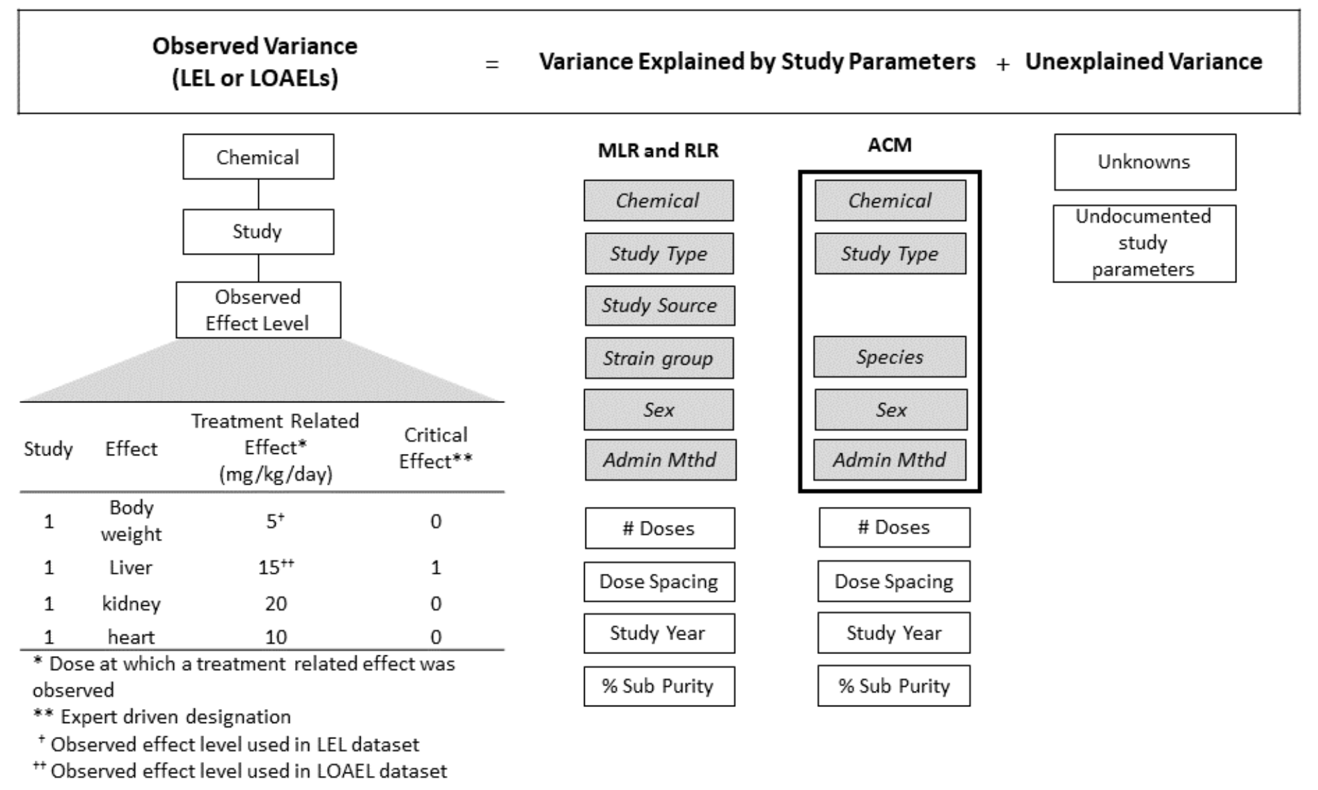
Abstract

New approach methodologies (NAMs) for hazard are often evaluated via comparison to animal studies; however, variability in animal study data limits NAM accuracy. The US EPA Toxicity Reference Database (ToxRefDB) enables consideration of variability in effect levels, including the lowest effect level (LEL) for a treatment-related effect and the lowest observable adverse effect level (LOAEL) defined by expert review, from subacute, subchronic, chronic, multi-generation reproductive, and developmental toxicity studies. The objectives of this work were to quantify the variance within systemic LEL and LOAEL values, defined as potency values for effects in adult or parental animals only, and to estimate the upper limit of NAM prediction accuracy. Multiple linear regression (MLR) and augmented cell means (ACM) models were used to quantify the total variance, and the fraction of variance in systemic LEL and LOAEL values explained by available study descriptors (e.g., administration route, study type, species). The MLR approach considered each study descriptor as an independent contributor to variance, whereas the ACM approach combined all categorical descriptors into cells to define replicates. Using these approaches, total variance in systemic LEL and LOAEL values (in log_{10}-mg/kg/day units) ranged from 0.74 to 0.92, and the unexplained variance, approximated by the residual mean square error (MSE), ranged from 0.20-0.39. Considering subchronic, chronic, or developmental study designs separately resulted in similar values. Based on the relationship between MSE and R-squared for goodness-of-fit, the maximal R-squared for a systemic effect level model using these data may approach 55 to 73%. The root mean square error (RMSE) ranged from 0.47 to 0.63 log_{10}-mg/kg/day, depending on dataset and regression approach, suggesting that a two-sided minimum prediction interval for systemic effect levels may have a width of 58 to 284-fold. These findings may have important implications for evaluation criteria used for NAM predictions of systemic toxicity.

- ***Predictive models cannot predict animal effect values with greater accuracy than those animal models reproduce themselves.***

- ***Defining the quantitative variability, or variance, in traditional systemic toxicity data informs the upper limit of predictivity for new approach methods and assists with acceptance of new approach methods with similar or better performance.***

Approach to estimating variance in systemic toxicity information from ToxRefDB



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

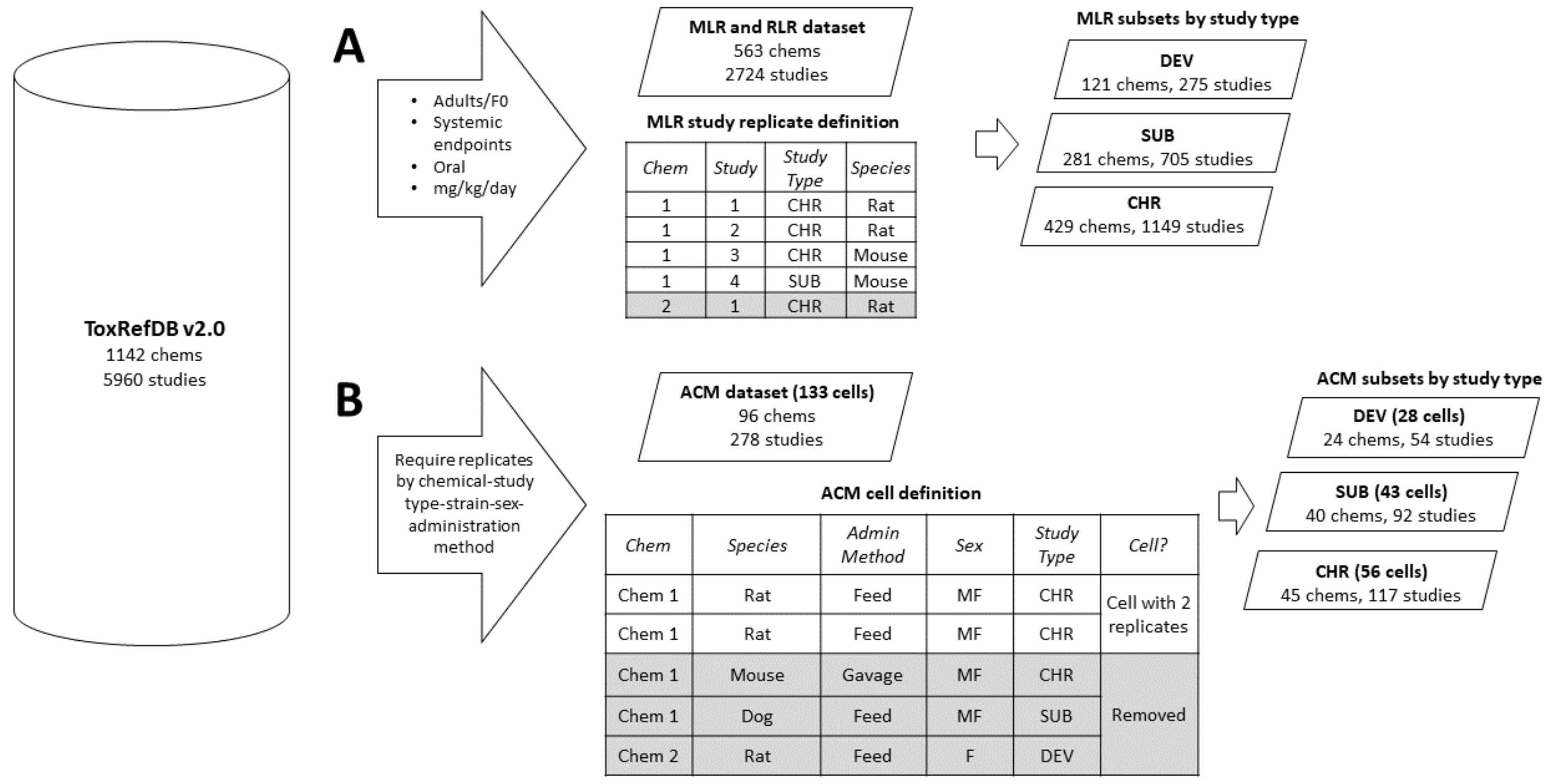
Figure 1. Variance models.
MLR = multilinear regression; RLR = robust linear regression; ACM = augmented cell means; Adm. Method = administration method; % Sub Purity = % substance purity used in the study. Gray boxes indicate categorical study descriptors whereas white boxes indicate quantitative study descriptors.

- LEL = lowest treatment-related effect observed for a given chemical in a study; LOAEL = 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.
- Study descriptors can be used to construct statistical models of variance using: multilinear regression (MLR), robust linear regression (RLR), and augmented cell means (ACM) regression.
- ACM creates a factor of the categorical descriptors to more stringently define "replicate" studies, whereas MLR/RLR approaches allow for larger datasets. ACM better accounts for interactions between descriptors, whereas MLR/RLR assume the study descriptors contribute independently to variance.

Workflow: Construct multiple statistical models of systemic toxicity data to estimate variance.

Figure 2. Variance estimation workflow.
CHR = chronic; DEV = developmental (adults only); SUB = subchronic; cells are defined by the factor of all categorical variables.

- (A) outlines the workflow for more permissively defined study replicates to enable a larger dataset for consideration of variance coupled with MLR and RLR;
- (B) outlines the workflow for more stringently defined study replicates using the ACM modeling approach.
- Both MLR and ACM datasets were subset by study type and statistically modelled to estimate variance.



MLR study replicate definition

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

ACM cell definition

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	

Results

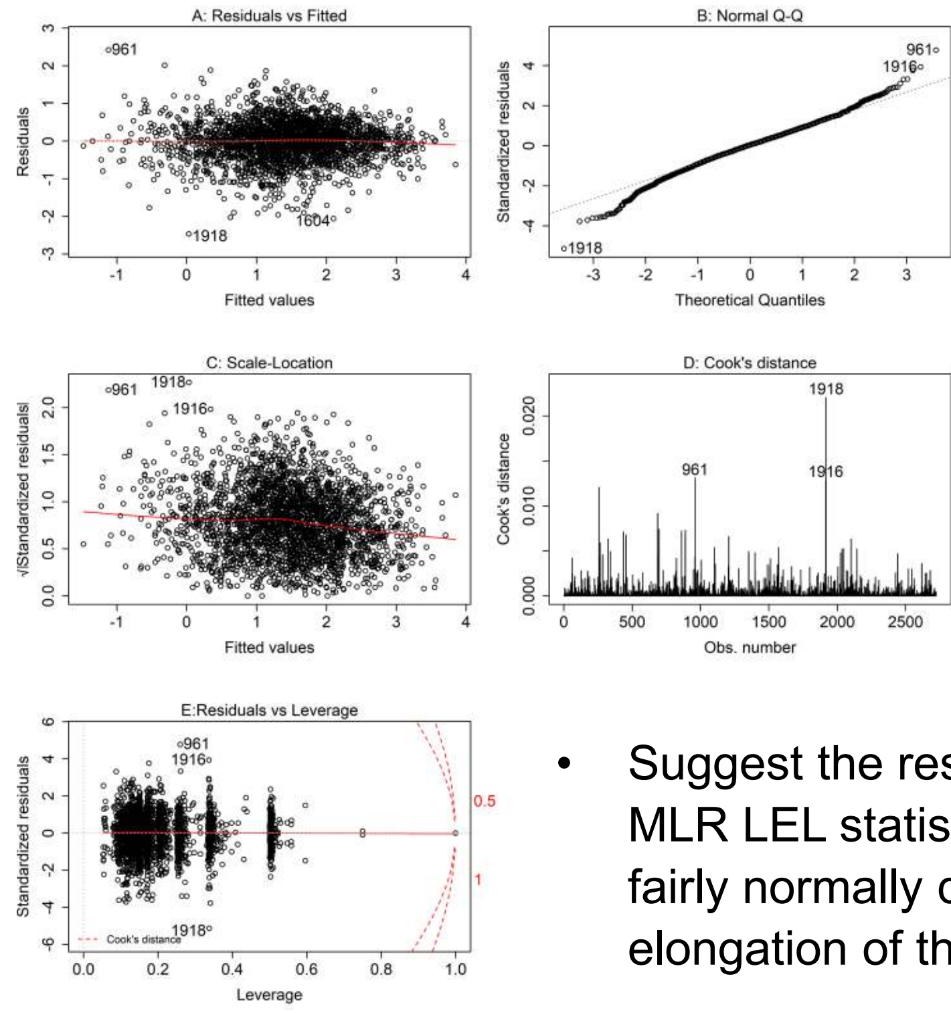
Table 1. Statistical model results from full datasets.

Regression Type	Data	LEL				LOAEL				N
		Total Variance	MSE	RMSE	% exp.	Total Variance	MSE	RMSE	% exp.	
RLR	full dataset	0.92	0.36	0.60	61	0.79	0.27	0.52	66	2724
MLR	full dataset	0.92	0.35	0.59	62	0.79	0.26	0.51	67	2724
MLR	high leverage points removed	0.91	0.34	0.58	63	0.78	0.25	0.50	68	2709
MLR	high Cooks distance plot points removed	0.91	0.34	0.58	63	0.79	0.25	0.50	68	2721
MLR	high Cooks distance points removed	0.84	0.26	0.51	69	0.75	0.20	0.45	73	2614
MLR	all potential outliers removed	0.84	0.26	0.51	69	0.74	0.20	0.45	73	2603
ACM	full cell dataset	0.86	0.32	0.57	63	0.75	0.25	0.50	66	278
MLR	full cell dataset	0.86	0.39	0.62	55	0.75	0.31	0.56	58	278

Table 2. Statistical model results from datasets subset by study type.

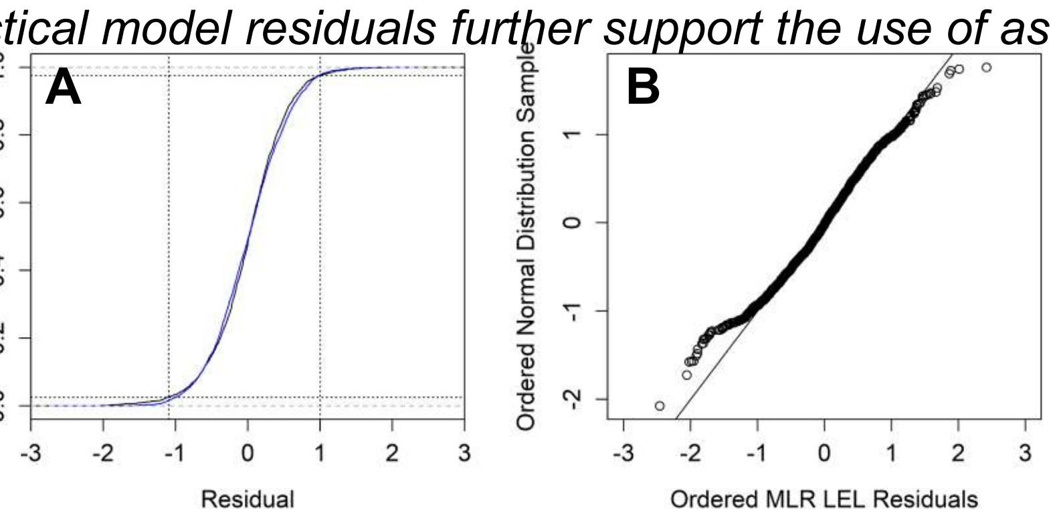
Regression Type	Data	LEL				LOAEL				N
		Total Variance	MSE	RMSE	% exp.	Total Variance	MSE	RMSE	% exp.	
MLR	SUB	0.88	0.35	0.59	60	0.78	0.28	0.53	65	705
ACM	SUB	1.0	0.30	0.55	70	0.90	0.25	0.50	72	92
MLR	CHR	0.95	0.35	0.59	63	0.80	0.25	0.50	68	1149
ACM	CHR	0.89	0.40	0.63	55	0.83	0.27	0.52	68	117
MLR	DEV	0.60	0.25	0.50	59	0.59	0.22	0.47	64	275
ACM	DEV	0.41	0.33	0.57	20	0.40	0.32	0.56	21	54

Figure 3. The distribution of the MLR LEL model residuals evaluated using standard diagnostic plots.



- Used to identify potential outliers and influential values for trimming the full dataset to build additional statistical models.
- Suggest the residuals from the MLR LEL statistical model are fairly normally distributed (minor elongation of the tails).

Figure 4. Empirical cumulative distribution of MLR LEL statistical model residuals. .
*(A) **C**omparison to normal distribution with same N, mean & standard deviation and (B) ordered normal distribution sample vs. ordered MLR LEL statistical model residuals further support the use of assumptions of normality.*



Statistics reference

Term	Concept
Accuracy	The degree to which a value matches the "true" value; in context, NAM accuracy to predict reference in vivo data cannot exceed the reference in vivo data accuracy for predicting itself.
Explained variance	Amount of the total variance that can be explained by the regression model built using study descriptors, where the unexplained variance is approximated by MSE.
Minimum prediction interval	A prediction interval is the possible range for a new value given some dataset and model with their own contributions to variance. The minimum prediction interval defined in this work is the possible range of a new value given the variance in the in vivo data available for training. Thus, only a perfect model could have a "minimum prediction interval" because all other models will contribute additional variance and width to the prediction interval.
MSE, also known as the residual mean square error	MSE for the regression model is the residual sum of squares divided by the degrees of freedom for the regression model, where the residual sum of squares is equal to the sum of the squared difference between each empirical observation Y_i and the predicted value for observation i ($f(x_i)$), and the degrees of freedom are equal to the number of observations, n, and the number of covariates (in this case, study descriptors).
% total variance explained	% total variance explained by study descriptors; this is the variance
Predictive model	A model that is constructed for forward prediction of unavailable values, typically trained on reference data.
Regression model	A statistical model of the existing data; seeks to explain variance in the current dataset rather than creating a forward prediction.
RMSE, also known as residual root mean square error	RMSE is the square root of the MSE and gives a measure of the residual spread or standard deviation for the regression model, in the same units as the LEL and LOAEL values (whereas the total variance and MSE are unitless). For normally distributed residual values, 95% of residuals should fall between ± 1.96*RMSE. In this work, RMSE is used to approximate what a minimum prediction interval might be for a prediction model using these data as a reference.
R-squared or R^2	The proportion of variance in a dependent variable that be explained by a regression model or independent variable. The maximum R^2 for a model representing some data is limited by the percent of the total variance that is explained by the available regression model parameters (in this work, study descriptors).
Total variance	Explained + Unexplained variance; the sum of the squared deviations of every observation from the sample mean divided by the degrees of freedom for the sample
Uncertainty	When applied to reference in vivo data, uncertainty might be quantified as a confidence interval for a mean value or perhaps the minimum prediction interval for a new predicted LEL or LOAEL value.
Unexplained variance	The portion of the variance that is not explained by the regression model built using study descriptors. This is estimated as the MSE.
Upper bound of predictivity	In reference to a predictive model; the limit on how precise a predictive model could be given the reference data used in training. In this work, the upper bound of predictivity includes the upper bound on an R^2 for a model of these data and the maximum accuracy of a prediction model for systemic toxicity values (i.e., the minimum prediction interval).
Variability	The spread or dispersion of some data.

Conclusions

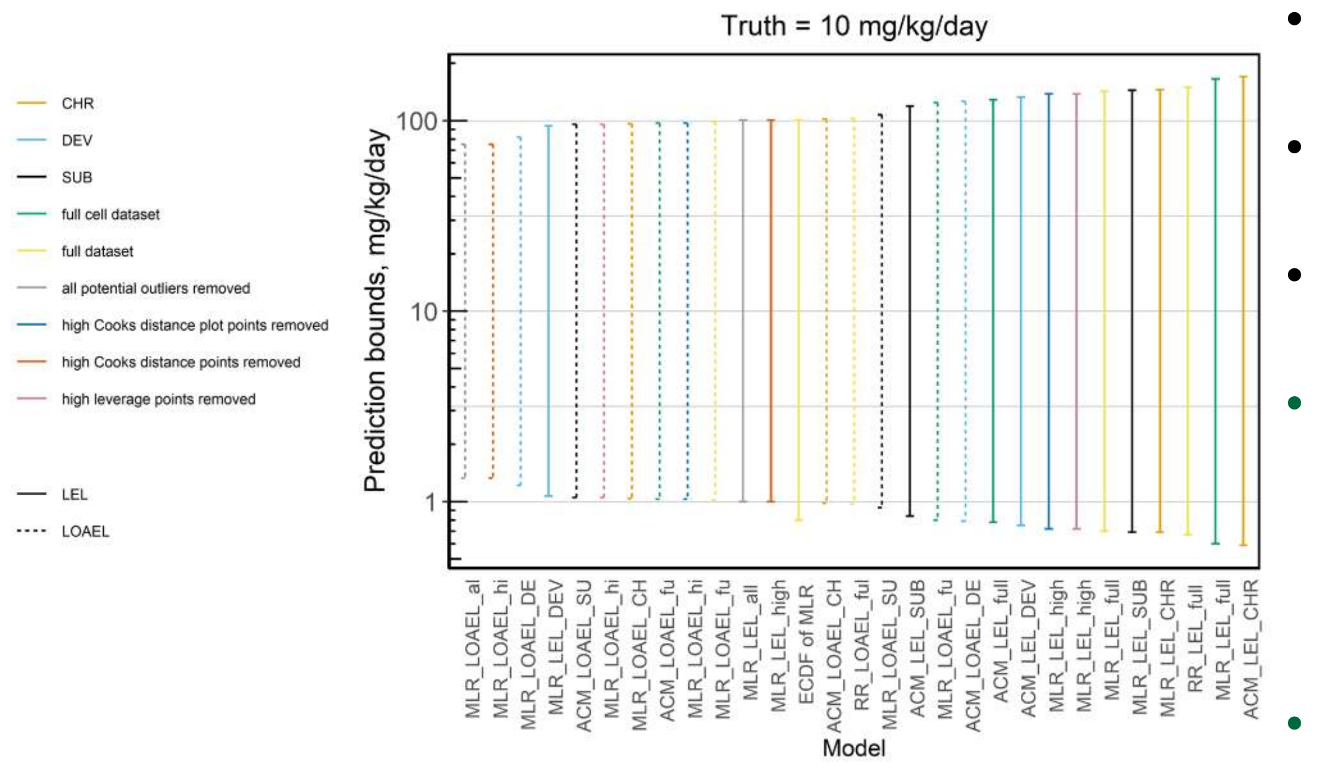


Figure 5. Visualization of minimum prediction intervals based on variance in systemic toxicity data.
Based on the RMSE of the statistical models of the systemic toxicity data, the two-sided minimum prediction interval tends to be approximately 2 orders of magnitude on a log10-mg/kg/day scale. For a "truth" of 10 mg/kg/day, a reasonable prediction from a predictive model might be between ~0.06 to 17 mg/kg/day.

- Variability in *in vivo* toxicity studies limits predictive accuracy of NAMs.
- The US EPA Toxicity Reference Database includes repeat dose toxicity studies.
- Total variance in systemic effect levels and the fraction explained were quantified.
- ***Maximal R-squared for a predictive model of systemic effect levels may be 55 to 73% based on the unexplained variance approximated as residual mean square error.***
- ***Estimated minimum prediction intervals for systemic effect levels were 58 to 284-fold. This is based on the amount of explained variance (RMSE) for different statistical models of these data.***

U.S. Environmental Protection Agency
Office of Research and Development