# Quantitative and qualitative variability in repeat dose toxicity studies: Implications for benchmarking NAMs

Katie Paul Friedman

October 10, 2018

Presented to the APCRA3 Meeting in Ottawa, Ontario

Based on collaboration with A*STAR, ECHA, EFSA, EPA-OLEM, EPA-ORD, Health Canada, and the JRC

# US and European statutory language on use and acceptability of NAMs

- In US, Section 4(h) in amended TSCA says –
  - "…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
- In Europe, REACH says –
  - Article 13: "Information on intrinsic properties of substances may be generated by means other than tests, provided that the conditions set out in Annex XI are met (…) for human toxicity, information shall be generated whenever possible by means other than vertebrate animal tests, through the use of alternative methods…"
  - Annex XI: "Results obtained from suitable *in vitro* methods may indicate the presence of a certain dangerous property or may be important in relation to a mechanistic understanding, which may be important for the assessment…" BUT confirmation using standard *in vivo* tests are still required unless:
    - Results are derived from an in *vitro* method whose scientific validity has been established by a validation study, according to internationally agreed validation principles; AND
    - Results are adequate for the purpose of classification and labelling and/or risk assessment; AND
    - Adequate and reliable documentation of the applied method is provided.

# Types of variability in traditional animal toxicity tests

- When comparing NAMs and traditional animal data, the variability may limit the observed predictive performance
- What is the variability in traditional data?

## Qualitative

*Challenge to binarization:*
*What if effect is not 100% reproducible across replicate studies?*

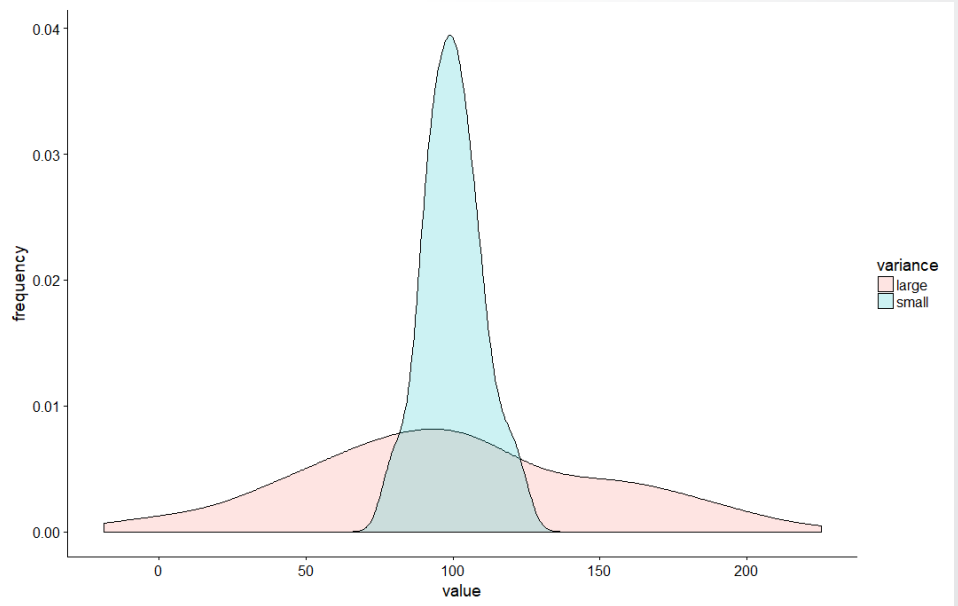
| | | "Truth" (traditional toxicology) | |
|---|---|---|---|
| | | Negative | Positive |
| Predicted (NAM) | Negative | True negative | False negative |
| | Positive | False positive | True positive |

## Quantitative

*Challenge to variance estimation:*
*What defines a study replicate?*

**Variance is a measure of how far values are spread from the average.**

We need to know what the "spread" or variability of traditional points-of-departure might be to know the range of acceptable or "good" values from a NAM.



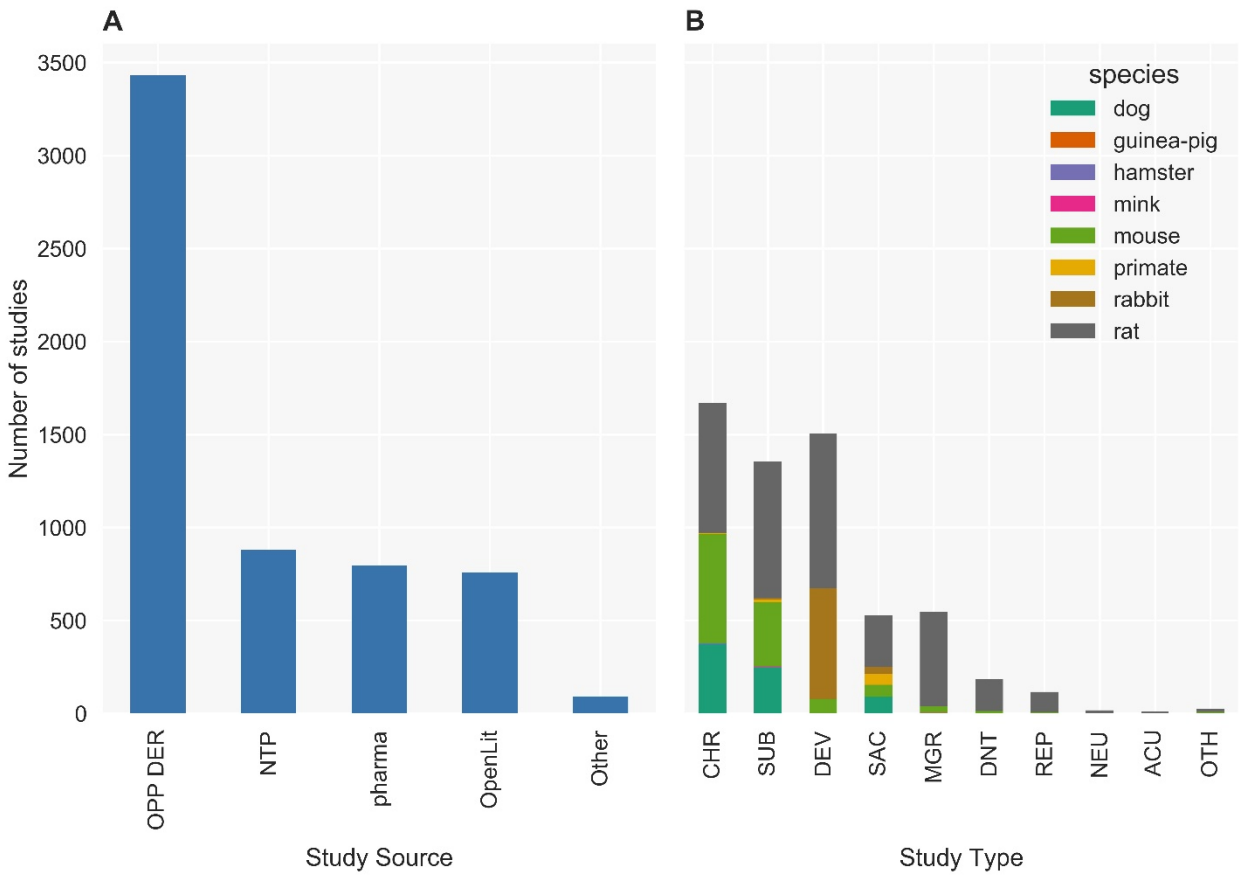
# Characterizing the variability in traditional animal toxicity tests

| | *Quantitative variability in traditional animal toxicity tests* | *Quantitative limitations for traditional vs. NAM concordance* | *Qualitative variability in traditional animal toxicity tests* |
|---|---|---|---|
| **3 main questions** | **What is the range of possible "true" systemic effect values (mg/kg/day) given a particular chemical?** | **What is the maximal precision of a model that attempts to predict a systemic POD for an unknown chemical?** | **What is the probability that an effect in adults will be observed for a given chemical?** |
| **Statistical approach to the question** | • Need an estimate of variance.<br>• Residual root mean square error (RMSE) is an estimate of variance in the same units as the systemic effect values.<br>• The RMSE tells us the potential range of the "true" effect level for a given observed effect level. | • Need to understand how much of the total variance can be explained by study descriptors.<br>• The mean square error (MSE) is used to approximate the unexplained variance.<br>• This unexplained variance limits the R-squared on a new model.<br>• The RMSE can also be used to define a reasonable prediction interval, or estimate range, for a model. | • Initially a qualitative exercise to understand the reproducibility of treatment-related changes in specific endpoint targets (e.g., any effect on liver). |

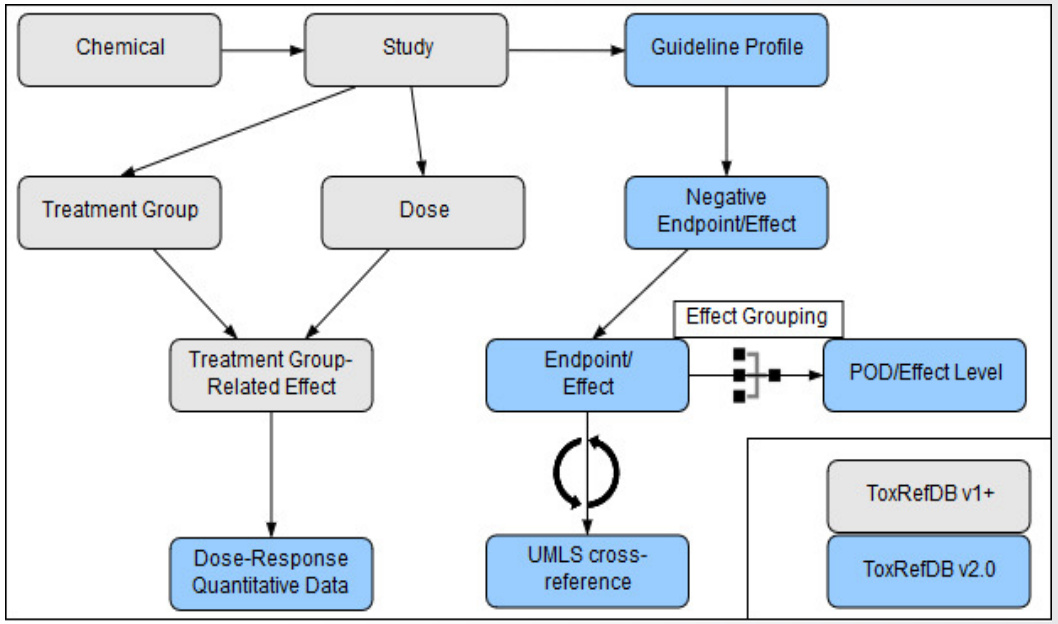
**EPA's Toxicity Reference Database (ToxRefDB) provides a comprehensive public resource to address these questions for regulatory toxicology.***

**Now starting a collaboration with ECHA to expand our access to relevant data to answer this question*

# ToxRefDB v2.0 contains relevant study data to evaluate uncertainty in traditional data for >1000 chemicals




**Number of studies by study type and species in ToxRefDB v2**




**Generalized schema of ToxRefDB v2**

Figures from Watford, Paul-Friedman, in prep, "ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses."

# Two approaches for estimating variance

Variance(Observed LEL or LOAEL) = Variance(Explained by Reported Study Descriptors) + Unexplained Variance

**Total variance**

**Fraction of total variance explained by information in the database**

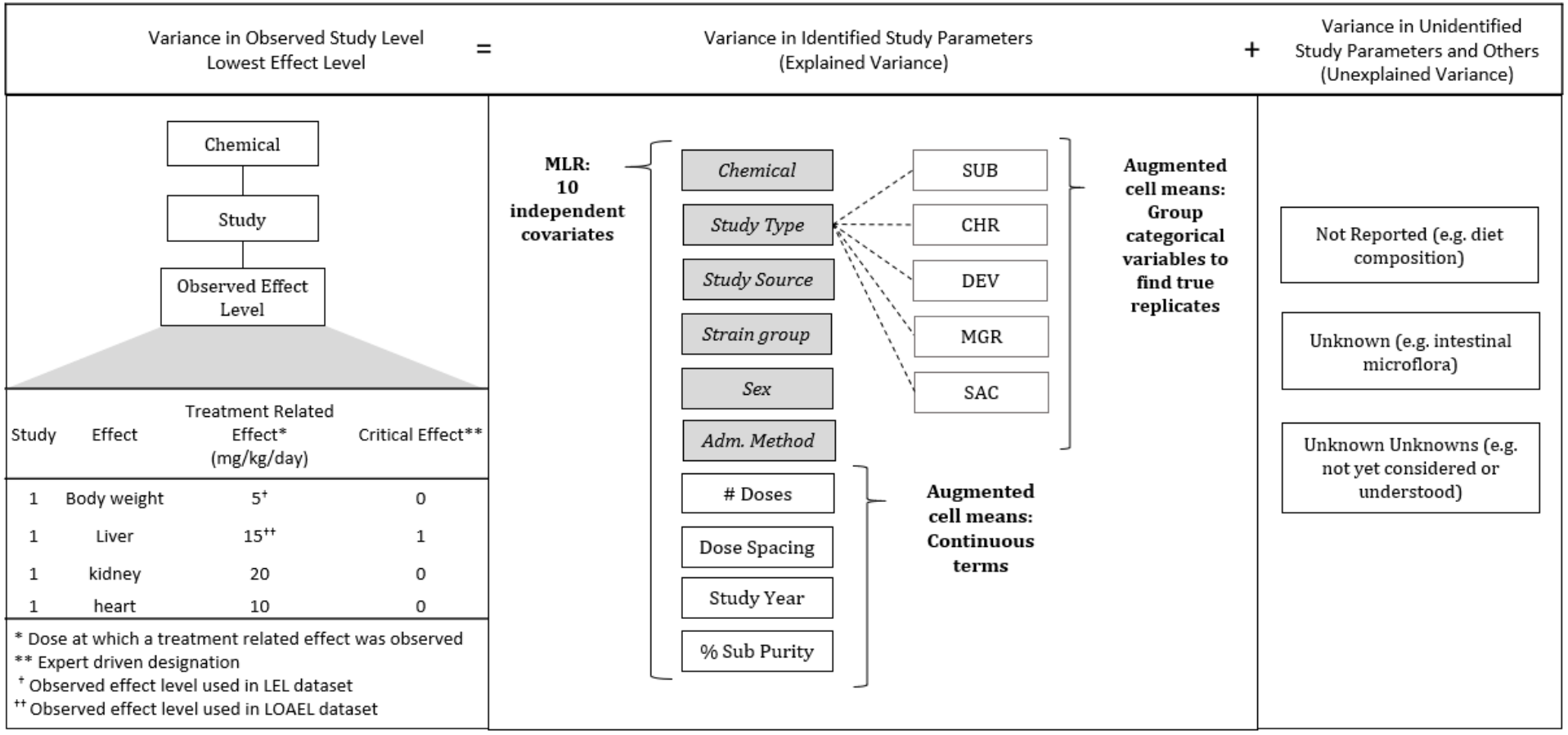
**Approximated by the mean square error (MSE)**

We employ two methods rooted in classical statistics to provide a range of reasonable estimates of variance in lowest effect levels (LELs) or lowest observable effect levels (LOAELs):

| | Aggregation level | Replicate definition | Dataset size | How study descriptors are treated |
|---|---|---|---|---|
| **Multilinear Regression Model** | Chemical | Not stringent | Maximized, reduce impact of possible outliers or database errors | Assumes study descriptors contribute to variance independently |
| **Augmented Cell Means Model** | Chemical-Study Type-Species-Sex-Admin Method combination | Stringent | Small, may bias variance estimate | Account for possible interactions among study descriptors |

*Consider that if we were asked to compute a hazard:exposure ratio using existing legacy data, we might combine all study data and take a minimum value*

# Models to estimate variance rely on the legacy data curation in ToxRefDB




**The math is simple and looks like this:**

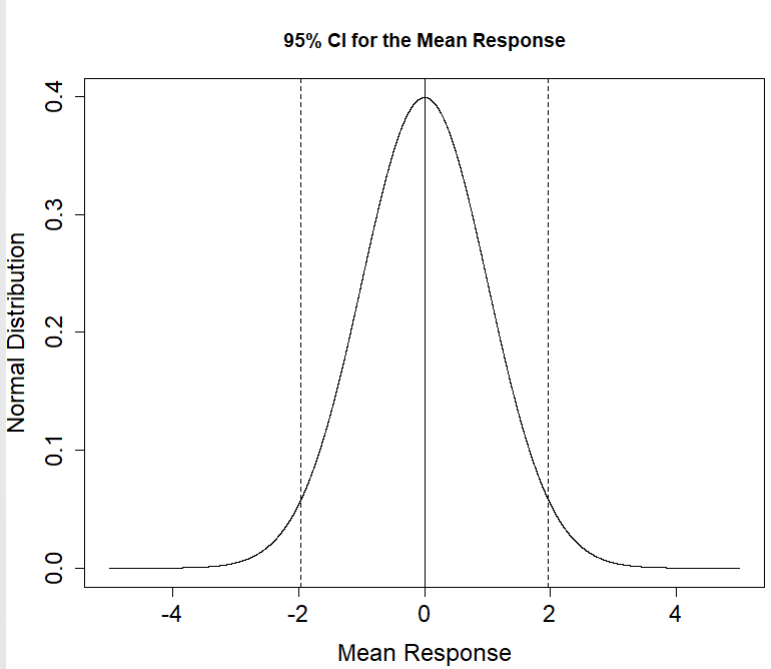
Response = Fit + Residual (MSE)

$$Y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$

**Conceptual view of the variance models.**
Pham, Paul Friedman, in prep. "Variability in *in vivo* Toxicity Studies: Defining the Upper Limit of Predictivity for Models of Systemic Effect Levels."

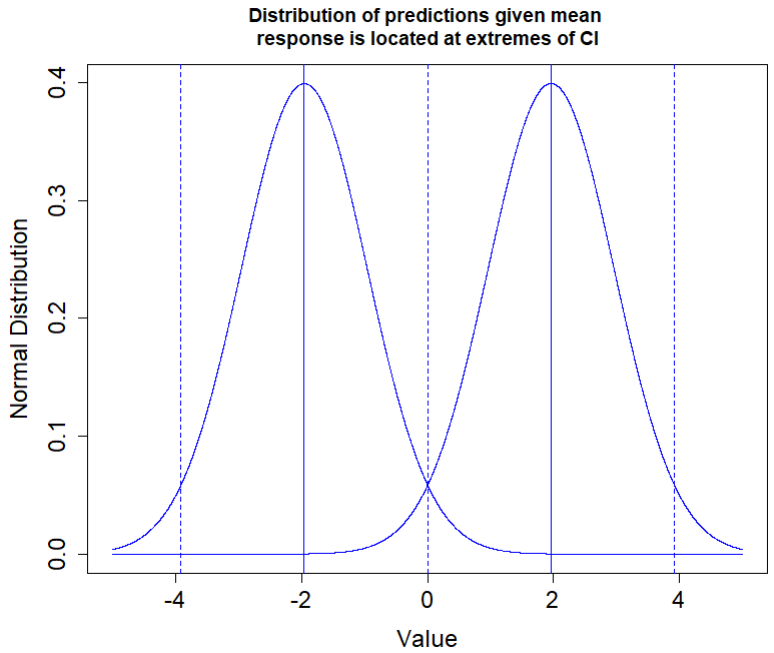
# From the variance estimate (RMSE) we can estimate a prediction interval.

## Confidence interval (CI) describes how well we have estimated the mean response



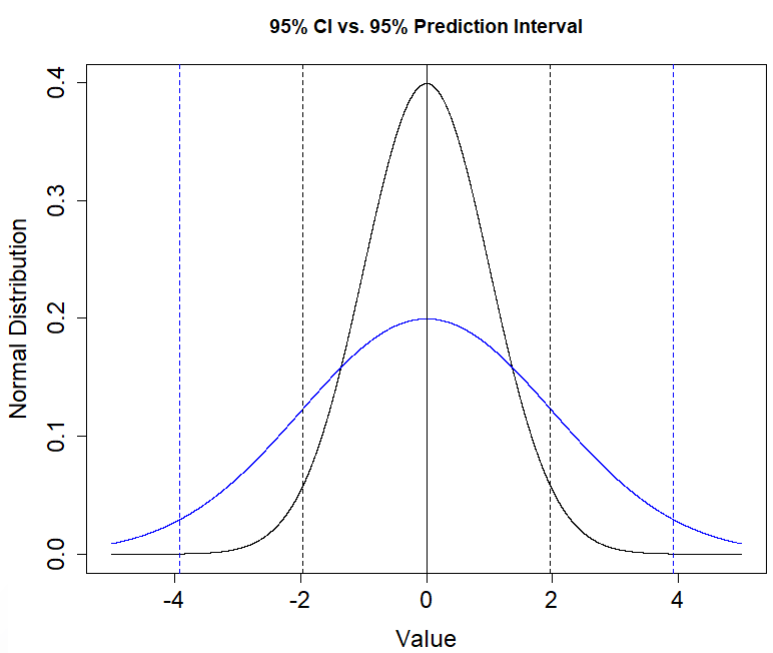

Given a sample size N, we estimate the mean response: the confidence interval describes where we would find the mean response if we had a different N.

## Where might the next data point be sampled from within that CI?




Next data point sampled (or next prediction) could be at the upper and lower bound of the CI for the mean response. The distribution of the prediction could have a center or median anywhere in the original CI.

## Prediction interval (blue) describes <u>where the next data point might be</u>




The prediction interval (blue) is much wider than the CI (black) because it has a tail to encompass the error around any sample/prediction within the CI

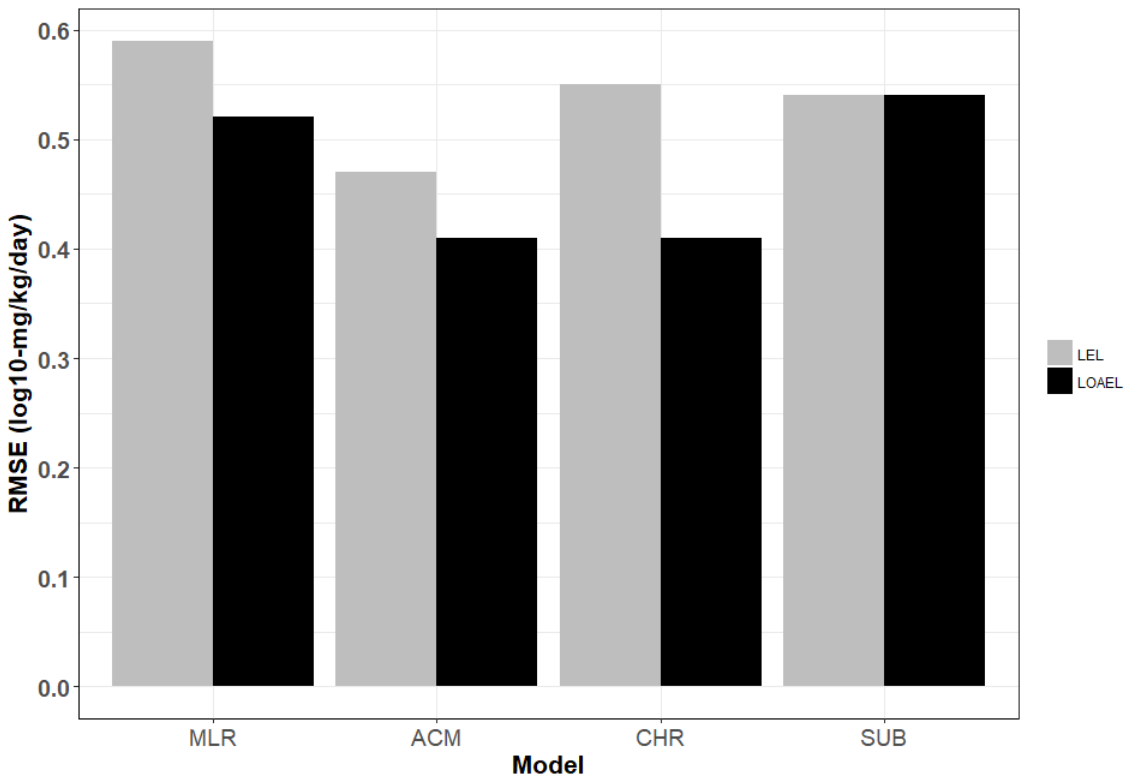


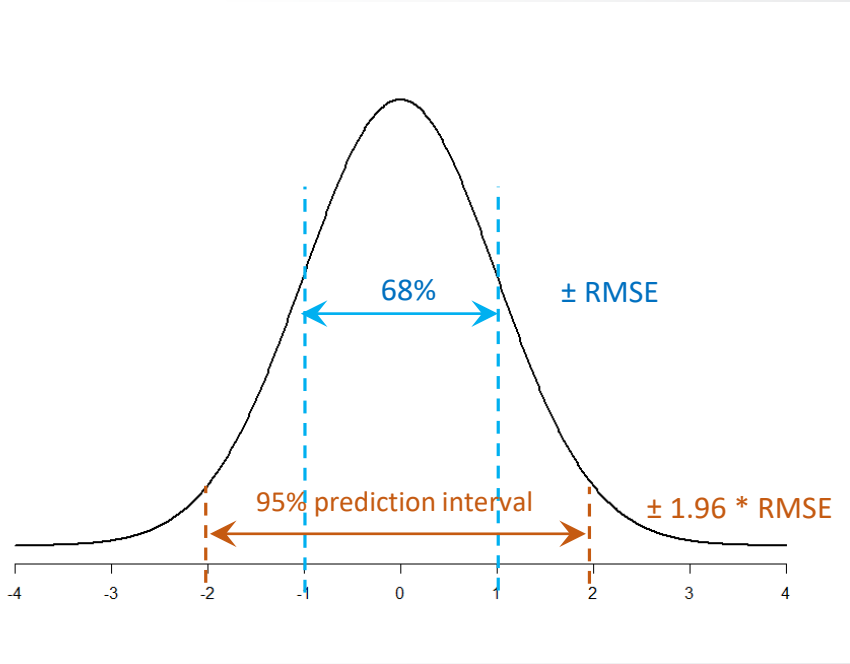
See <u>Applied Linear Statistical Models</u>, ISBN-13: 978-0073108742 and GraphPad Software (www.graphpad.com).

# The RMSE (estimate of variance or spread) of systemic LELs and/or LOAELs can be approximated as ~ 0.5 log10-mg/kg/day.

## Precisely, RMSE ranged from approximately 0.41 to 0.59 log10-mg/kg/day, depending on model and dataset

*RMSE estimates across different models and data sets*




Total size of the prediction interval is ~100-fold

Using RMSE=0.5, a low effect level (LEL) of:
1 mg/kg/day would be predicted as 0.1 – 9.5 mg/kg/day

**Summary of draft findings from Pham, Paul Friedman, in prep. "Variability in *in vivo* Toxicity Studies: Defining the Upper Limit of Predictivity for Models of Systemic Effect Levels."**

# Upper limit on the R-squared for a predictive model is related to MSE, or unexplained variance in the model.

*The MSE approximates the amount of variance not explained by any study descriptors*

**Total variance = Explained variance + Unexplained variance**

*R-squared is limited by the unexplained variance*

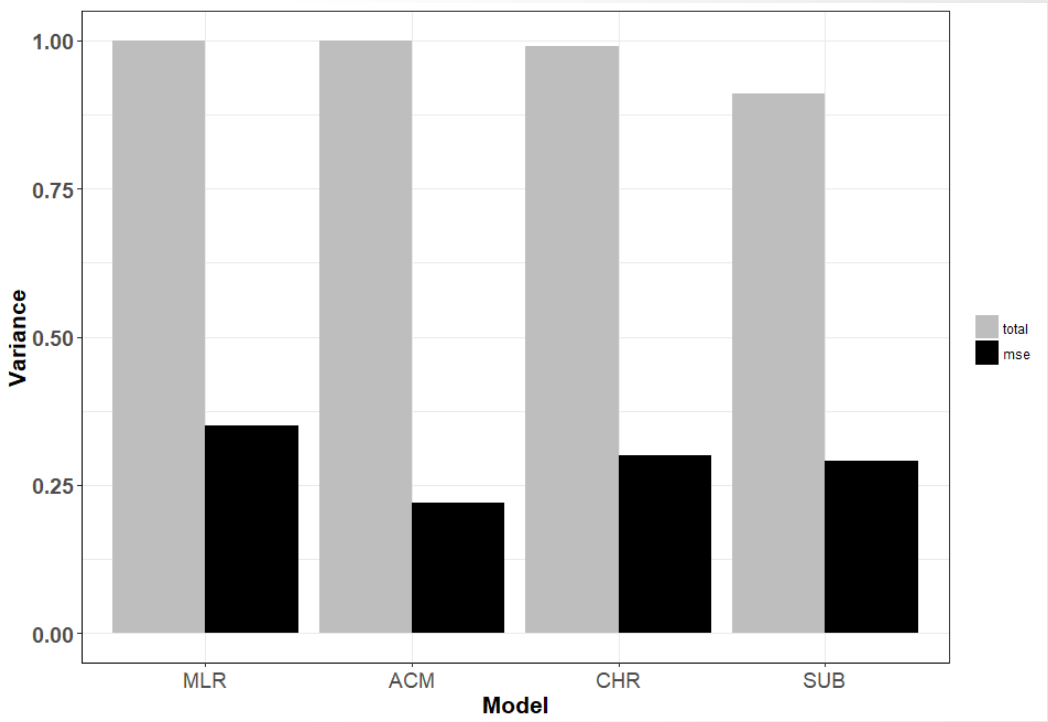
$$R^2 = \frac{Total\ Variance - MSE}{Total\ Variance}$$

$$R^2 = \frac{1-0.22}{1} = 0.78$$

$$R^2 = \frac{1-0.35}{1} = 0.65$$

**Depending on the dataset used in training, the upper limit on the R-squared for NAM for systemic effect POD is somewhere around 70%.**

*Ex. Total variance and MSE for LEL data sets*




For large combined sets of repeat dose studies, CHR studies and SUB studies alone, MSE ranged 0.22 to 0.35.

**Summary of draft findings from Pham, Paul Friedman, in prep. "Variability in *in vivo* Toxicity Studies: Defining the Upper Limit of Predictivity for Models of Systemic Effect Levels."**

# Preliminary conclusions for our quantitative variance work

- The MLR and ACM approaches yielded a similar range of variance values.
- The estimate of variance (RMSE) in curated LELs and/or LOAELs approaches a 0.5 log10-mg/kg/day.
- The unexplained variance (MSE) across different models and datasets suggests that a NAM built to predict a systemic POD would have an maximum $R^2$ around 70%, i.e. as much as 1/3 of the variance in these data may not be explainable by curated study descriptors.
- Using the RMSE to estimate a reasonable prediction interval for systemic POD suggests the interval would 1.5 to 2 log10-mg/kg/day wide (again, depending on the model and data subset).
- Definition of a "study replicate" for the ACM approach is complex because the legacy data is often curated to the summary or report level, rather than the study level. *Currently we are manually reviewing each cell in the ACM approach to confirm what was done programmatically.*

# How does this compare to previous work in this area?

- The Monte Carlo approach available in CORAL software(http://www.insilico.eu/coral) has also been used in a number of cases to model subchronic oral rat NOAEL values, producing a spectrum of R-squared values from 0.46-0.71, suggesting that the inputs to these models could only account for 50-70% of the variance in the reference data (Veselinovic et al. 2016; Toropov et al. 2015; Toropova et al. 2017).

*An $R^2$ approaching 70% for these types of data may indicate a fairly reasonable model based on our estimates*

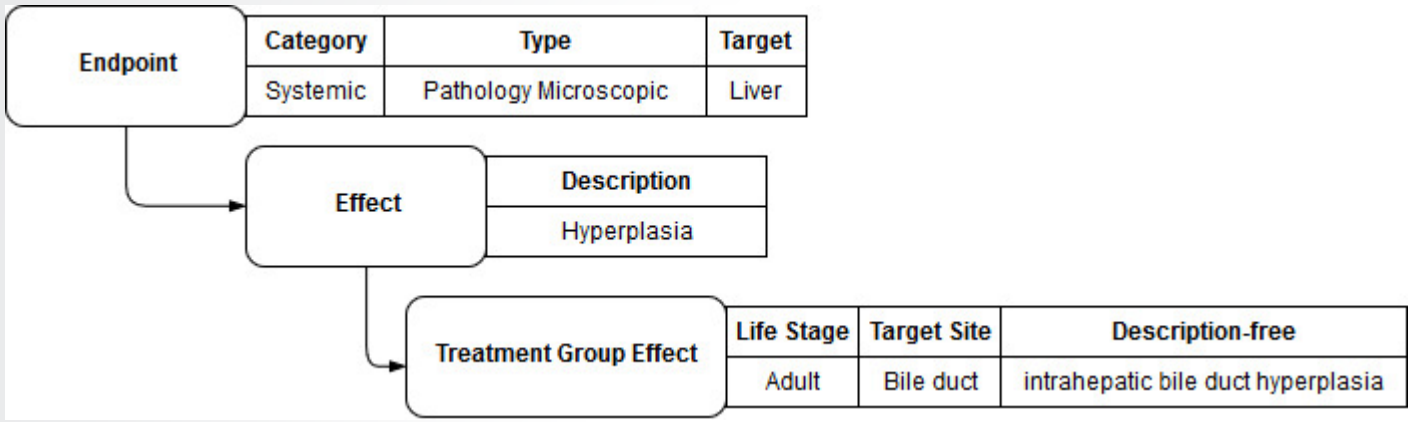
- A multi-linear regression QSAR model of chronic oral rat lowest observable adverse effect level (LOAEL) values for approximately 400 chemicals, demonstrated a root mean square error (RMSE) of 0.73 $\log_{10}$(mg/kg-day), which was similar to the size of the variability in the training data, ±0.64 $\log_{10}$(mg/kg-day), approximated as two times the mean standard deviation; these findings suggested that the error in the model approached the error in the reference data from different laboratories (Mazzatorta et al. 2008).

*A dataset (of rat chronic LOAEL) seems to suggest similar amount of variability (though methods are different from ours)*

- Importantly, the variance in discrete NOAEL or LOAEL values is highly subject to the shape of the dose-response curve, the sample size, and the doses selected . As such, we might hypothesize that modeled (BMD) values may decrease the variance.

*We could hypothesize that BMD/BMDL/other curve fit values would decrease variance estimates*

# Qualitative uncertainty analysis is enabled by controlled vocabulary in ToxRefDBv2




**Example of the controlled effect vocabulary in ToxRefDB v2. (Watford, Paul Friedman, et al., in prep)**

For examining reproducibility, we work at the endpoint level.

A draft analysis conducted in 3 parts:

(1) Simple concordance analysis for endpoint level observations

*(2) Probability of endpoint level observations using a logistic regression model*

*(3) Average concordance within the cells defined in the quantitative variance work*

| Organ | # chemicals | | | % Concordance |
|---|---|---|---|---|
| | 0 | M | 1 | |
| Liver | 46 | 118 | 126 | 59.31 |
| kidney | 72 | 170 | 48 | 41.38 |
| spleen | 150 | 123 | 17 | 57.59 |
| testes | 158 | 120 | 12 | 58.62 |
| adrenal gland | 164 | 111 | 15 | 61.72 |
| heart | 179 | 107 | 4 | 63.10 |

| Study Type | # chemical x study | | | % Concordance |
|---|---|---|---|---|
| | 0 | M | 1 | |
| SUB | 64 | 45 | 156 | 83.02 |
| CHR | 73 | 59 | 149 | 79 |
| SUB | 108 | 70 | 87 | 73.58 |
| CHR | 115 | 83 | 83 | 70.46 |
| SUB | 167 | 57 | 41 | 78.49 |
| CHR | 187 | 60 | 34 | 78.65 |
| SUB | 180 | 48 | 37 | 81.89 |
| CHR | 197 | 52 | 32 | 81.49 |
| SUB | 190 | 44 | 31 | 83.39 |
| CHR | 203 | 48 | 30 | 82.92 |
| SUB | 191 | 49 | 25 | 81.51 |
| CHR | 216 | 45 | 20 | 83.98 |

| Species | # chemical x species | | | % Concordance |
|---|---|---|---|---|
| | 0 | M | 1 | |
| dog | 20 | 26 | 46 | 71.74 |
| mouse | 29 | 39 | 69 | 71.53 |
| rat | 41 | 70 | 130 | 70.95 |
| dog | 49 | 33 | 10 | 64.13 |
| mouse | 61 | 51 | 25 | 62.77 |
| rat | 59 | 103 | 79 | 57.26 |
| dog | 64 | 21 | 7 | 77.17 |
| mouse | 91 | 31 | 15 | 77.70 |
| rat | 130 | 83 | 29 | 65.70 |
| dog | 65 | 20 | 7 | 78.26 |
| mouse | 109 | 19 | 9 | 86.13 |
| rat | 133 | 85 | 23 | 64.73 |
| dog | 76 | 12 | 4 | 86.96 |
| mouse | 107 | 23 | 7 | 83.21 |
| rat | 139 | 83 | 19 | 65.56 |
| dog | 72 | 19 | 1 | 79.35 |
| mouse | 112 | 20 | 5 | 85.40 |
| rat | 155 | 69 | 17 | 71.36 |

# How does this compare to previous work?

- Local lymph node assay (LLNA): with same species & vehicle solvent, repeat LLNA were concordant only 78% of the time, with a 35% chance that a “negative” chemical would test “positive” if the LLNA was repeated (Hoffmann et al., 2018; Dumont et al., 2016).
- Kleinstreuer and colleagues showed that even in high quality studies for the rodent uterotrophic bioactivity assay, concordance was only achieved 74% of the time for replicate uterotrophic assays.
- An evaluation of 37 National Toxicology Program repeat dose toxicity studies demonstrated 0-100% concordance, with a median of approximately 70%, in the non-carcinogenic effects observed between rats and mice, depending on the biological endpoint or tissue measured (Wang & Gray, 2015).
- Concordance among rat and mouse models of carcinogenicity has been shown to range from 57% to 76% (Gottman et al. 2001; Gold et al., 1989; Haseman 2000).

# How can we relate this work on variance to ongoing collaboration via APCRA?

- **NCCT and ECHA will share data resources from EPA's Toxicity Reference Database (ToxRefDB) and ECHA's IUCLID to provide a comprehensive public resource to estimate anticipated "spread" of repeat dose toxicity POD values.**
- Estimate quantitative uncertainty in available subchronic POD values (for as many chemicals as possible) supports:
  - reasonable expectations of NAM replacement for subchronic data; and,
  - what to do in the Prospective approach when only a subchronic study for a chemical is available (i.e., how much uncertainty should we account for in distinguishing chemicals with small BER)?

Ongoing work, Rusty Thomas and Katie Paul-Friedman (NCCT)
Panos Karamertzanis & Tomasz Sobanski (ECHA)

# Acknowledgements within NCCT

- Ly Ly Pham
- Sean Watford
- Richard Judson
- Woody Setzer

EPA's National Center for Computational Toxicology

**Pham, L.L.,** S. Watford, P. Pradeep, R. Judson, C. Grulke, W. Setzer, M. Martin, K. Paul Friedman. *"Variability in in vivo Toxicity Studies: Defining the upper limit of predictivity for models of systemic effect levels"* (in preparation, for submission in late 2018).

**Pham, L.L,** T. Sheffield, P. Pradeep, J. Brown, D. Haggard, J. Wambaugh, R. Judson, K. Paul Friedman. *"Estimating Uncertainty in the Context of New Approach Methods Used in Risk Assessment"* (invited mini-review, submission imminent).