Reproducibility and Variance of Liver Effects in Subchronic and Chronic Repeat Dose Toxicity Studies

Pham, Ly L.[1]; Watford, Sean[2,3]; Woodrow, Setzer[4]; Paul Friedman, Katie[4]

[1]ORISE Fellow, RTP/EPA *
[2]University of North Carolina Chapel Hill
[3]ORAU National Student Services Contractor, USEPA National Center for Computational Toxicology
[4]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC
The views expressed are those of the authors and do not necessary reflect the view or policy of the US Environmental Protection Agency

www.epa.gov

Background

In vivo studies provide reference data to evaluate alternative methods for predicting toxicity. However, the reproducibility and variance of effects observed across multiple *in vivo* studies is not well understood. The US EPA's Toxicity Reference Database (ToxRefDB) stores data from EPA guideline studies including subchronic (SUB) and chronic (CHR) studies. The current work focused on the reproducibility of liver effects in SUB and CHR studies, as evaluation of liver weight, gross and micropathology are required by these guidelines. The objectives of this work include determination of: (1) the variance in observed liver effects in SUB and CHR studies; (2) the probability that liver effects were observed in replicate SUB or CHR studies and, (3) the potential for prediction of CHR liver effects from SUB studies.

Methods

Data Preparation

Data source: US EPA's Toxicity Reference Database
(ToxRefDB v 2; Abstract Number 2532; Poster number P894)
- >5,000 *in vivo* toxicity studies for >1,000 chemicals.
- Guideline or guideline-like studies from various sources.

Data was filtered to include:
- Adult animals in the F0 generation
- Systemic toxicity endpoints from CHR and SUB studies
- Administration Route: Oral
- Species: mouse, rat, and dog
- Non-control group data
- Organ effect was observed if any treatment related effect was observed in the organ weight, pathology gross, and/or pathology microscopic.
- Evaluated organs: liver, kidney, spleen, testes, adrenal gland, heart, and thyroid gland.
- Organs with the most number of positive observation at the study level were chosen for evaluation.

Study Conditions observed:
- Chemical _id (factor)
- Strain type (factor)
- Study type (factor)
- Dose spacing
- Number of doses
- Substance purity
- Study source (factor)
- Sex (factor)

Analysis

Variance Estimation
- Multilinear Regression (MLR) was used to partition the total variance in the observed lowest effect level (LEL) per organ into an unexplained component and a component attributable to different study design factors, and ANOVA was used to compare the significance of individual components.
  - LEL_{organ} ~ MLR(chemical , study conditions)

Probability Estimation
- Logistic Regression (LR) was used to calculate the probability of observing an effect in the organ given the study conditions.
  - Organ (binary) ~ LR(chemical, study conditions)
- Probability >0.5 is a positive prediction ; <0.5 is a negative prediction
- Compare prediction back to original data to see how well the prediction matched itself/source data.

Concordance Analysis
- Concordance per chemical was assessed for organs by comparing if a treatment related effect was observed across all SUB and CHR studies separately. 1 indicates all studies showed a treatment related effect, 0 indicates all studies showed no treatment related effect, and M (mixed) indicates that the studies did not agree if there was a treatment related effect for that chemical (Figure 1).
- Concordance rates within the SUB and CHR studies were then compared with each other to create a probability matrix with nine possible patterns (Figure 2).
  - E.g.,
    - p1 = 100% concordance that no effect was observed at the SUB studies and the CHR studies for a particular chemical.
    - p2 = 100% concordance that no effect was observed at the SUB studies but CHR studies showed mixed results a particular chemical, etc.

Figure 1: Concordance within Study Type

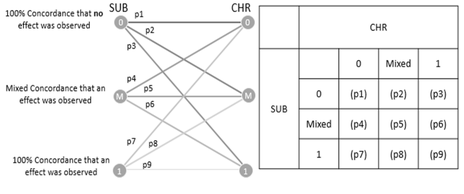
Chemical x	Liver	Kidney
SUB Study 1	1	1
SUB Study 2	1	0

→ 1 (Liver), M (Kidney)

Chemical x	Liver	Kidney
CHR Study 1	1	0
CHR Study 2	1	0

→ 1 (Liver), 0 (Kidney)

Figure 2: Concordance Matrix for SUB and CHR



SUB \ CHR	0	Mixed	1
0	(p1)	(p2)	(p3)
Mixed	(p4)	(p5)	(p6)
1	(p7)	(p8)	(p9)

Results

Ly Ly Pham | pham.lyly@epa.gov | 0000-0001-8467-2645

Variance Estimation

Table 1: Variance Estimation for the organ LEL For each organ, the total variance, mean squared error (MSE), and root mean squared error (RMSE) were calculated for the indicated number of studies and chemicals. The RMSE is in (log10(mg/kg/day)) units.

Organ	Chemical (n)	Study (n)	Total Variance	MSE	RMSE	% Variance Explained
liver	271	908	0.72	0.32	0.56	27.56
kidney	181	513	0.72	0.31	0.56	28.94
spleen	94	247	0.65	0.34	0.58	12.69
testes	63	160	0.67	0.22	0.46	34.16
adrenal gland	61	153	0.75	0.39	0.62	23.00
heart	59	137	0.76	0.26	0.51	41.79
thyroid gland	48	128	0.73	0.28	0.53	34.64

A variance analysis was performed for seven organs (Table 1). The organs were chosen based on the amount of data available. The total variance ranged from ~0.65 - 0.76, by organ. The mean square error (MSE) or variance that can not be explained by the study conditions varied more (~ 0.22-0.39). The root mean square error (RMSE), which is calculated as the square root of the MSE, can provide a prediction interval around an estimated LEL_{organ}.

The amount of variance that can be explained ranged from ~12.69-41.79% of the total variance. The large range can indicate that the available reported study conditions account for more of the variation in observed lowest effect levels (LEL) related to the heart than the spleen.

Probability Estimation

One of the new updates to ToxRefDBv2 is the ability to indicate true positives and true negatives. In this analysis, for all chemicals with at least two studies (SUB and/or CHR), we calculated the probability of observing a treatment related effect (0 or 1) in a particular organ (Table 2). The modeled probability was compared back to the original data to assess performance. The self prediction is to provide a benchmark level of performance. The balanced accuracy for all evaluated organs ranged from ~0.73-0.83. All of the organs had better specificity values than sensitivity values, except for liver.

Table 2: Probability Prediction for 7 organ systems using true positive and true negative. The probability model was compared to the observed data to get the confusion matrix of true positives/negatives. The performance of the prediction was assessed using accuracy, false discovery rate, sensitive, specificity, and balance accuracy. TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative

organ	TP	TN	FP	FN	accuracy	FDR	sensitivity	specificity	Balance accuracy
liver	893	349	114	86	0.861	0.113	0.912	0.754	0.833
kidney	476	693	125	148	0.811	0.208	0.763	0.847	0.805
spleen	217	1024	73	128	0.861	0.252	0.629	0.933	0.781
testes	159	1111	57	115	0.881	0.264	0.58	0.951	0.765
adrenal gland	143	1133	47	119	0.885	0.247	0.546	0.96	0.753
heart	111	1168	52	111	0.887	0.319	0.5	0.957	0.728
thyroid gland	132	1173	41	96	0.905	0.237	0.579	0.966	0.772

Concordance within and across study types

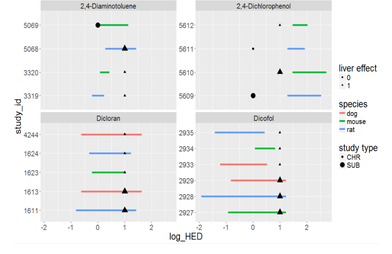
Liver	CHR			
SUB		0	M	1
	0	8 (p1)	8 (p2)	3 (p3)
	M	4 (p4)	26 (p5)	12 (p6)
	1	6 (7)	22 (p8)	58 (p9)

In this test case, we evaluated liver concordance within a chemical-study type pair and then across study type to evaluate if SUB studies can predict the CHR studies. The confusion matrix for liver (Figure 1) shows that strong positive concordance in the SUB studies tends to correspond to strong positive concordance in the CHR studies. For the M category, we wanted to evaluate if the disagreement was due to species differences or lack of overlapping dose ranges. Figure 2 shows an example of four chemicals. In preliminary analysis, it seems that, as expected, the discordance is complex and not simply consistently the result of species or dose-range.

Figure 3: A confusion matrix for the liver concordance analysis of the SUB and CHR studies.

For 2,4-diaminotoluene, studies 5069 (SUB) and 3320 (CHR) had overlapping dose ranges and were performed in mouse; however, the SUB study failed to note a liver effect. Conversely, studies 5068 (SUB) and 3319 (CHR) were performed in rat and had non-overlapping dose ranges, but liver effects were observed for both.

For Dicloran, all studies (3 species) had overlapping dose ranges and all demonstrated observed liver effects.



For 2,4-Dichlorophenol, there were clear differences by species; the mouse (green) studies both observed liver effects, whereas the SUB and CHR studies for rat (blue) failed to demonstrate liver effects.

For Dicofol, all three species agreed across study types.

Figure 4: An example of 4 chemicals used in the liver test case. The dose range, study, type, species tested, and absence or presence of liver effects for each study of a chemical

Conclusions and Future Directions

Conclusion:
- The variance and prediction interval (based on the RMSE) for a LEL at the organ level differed by organ.
- Certain study features significantly accounted for a percentage of the observed variance.

Future Directions:
- Quantify and standardize the evaluation of species and dose spacing as factors in concordance.
- Evaluate if some chemotypes are better predictors of organ-level endpoints via enrichment analysis of the possible patterns of response represented by the confusion matrix.
- Try to predict CHR study effects with SUB study effects, using this probability analysis and machine learning approaches.

*This project was supported in part by an appointment to the Internship/Research Participation Program at the National Center for Computational Toxicology, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.