National Institute of Environmental Health Sciences
Division of the National Toxicology Program

Variability in Reference Test Method Data and the Impact on NAM Evaluations

D Allen[1], J Rooney[1], K To[1], N Choksi[1], P Ceger[1], A Daniel[1], A Karmaus[1], J Strickland[1], J Truax[1], WM Casey[2], and N Kleinstreuer[2]

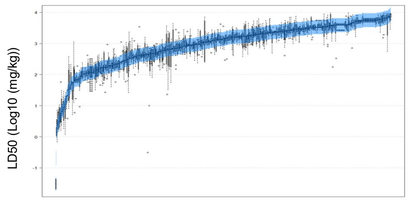
[1]Inotiv, RTP, NC, USA; [2]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

Highlights

- Historically, toxicity testing has been conducted using in vivo test methods.
- Confidence in data from these methods is such that regulatory hazard classification and labeling systems have been designed around their results.
- To establish confidence in new approach methodologies (NAMs), we must demonstrate that they are as good as or better than the existing in vivo test method.
- One approach to doing this is characterizing the inherent variability of the in vivo tests, which will directly affect the expectations for performance of NAMs that seek to replace them.
- In this study, we characterized the variability of in vivo reference test methods for multiple endpoints, including skin and eye irritation, skin sensitization, and acute systemic toxicity.

Acute Oral Toxicity Testing

- The graph below shows variability among 2441 replicate point estimate LD50 values quantified as mean absolute deviation (MAD) across replicate LD50 values per chemical.
- MADs were bootstrapped to compute margin of uncertainty (+/- 0.24 $\log_{10}$ mg/kg) for evaluation of NAMs.



LD50 points estimates per chemical with margin of uncertainty (median LD50 +/- 0.24 $\log_{10}$ mg/kg) highlighted, showing that the margin of uncertainty generally encompasses most replicate in vivo study LD50 values.

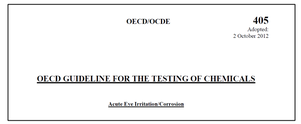
- The table below shows conditional probabilities calculated to predict the hazard classification outcome of multiple studies on the same chemical (United Nations Globally Harmonized System of Classification and Labeling of Chemicals: GHS).
- The shaded cells on the diagonal show that the probability of a subsequent study on the same chemical identifying the same GHS hazard category ranges from 48-75%.

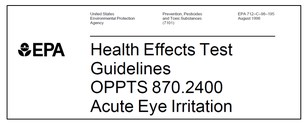
		Conditional Probability of Subsequent Study Categorization				
		1	2	3	4	5
Prior Type	1	53%	34%	2%	5%	5%
	2	8%	48%	33%	9%	1%
	3	2%	7%	62%	29%	2%
	4	0%	0%	11%	66%	2%
	5	0%	0%	1%	24%	75%

Karmaus et al. submitted

Eye Irritation Testing

Eye Irritation Classification	
GHS Category	**In Vivo Effect**
1	≥1 animal with CO=4 at any time or ≥2 animals with mean CO≥3 or IR≥1.5 or ≥1 animal at day 21 with CO or IR≥1 or CC or CR≥2.
2A	≥2 animals with mean CO or IR≥1 or CC or CR≥2 which reverses within 21 days.
2B	≥2 animals with mean CO or IR≥1 or CC or CR≥2 which reverses within 7 days.


OECD/OCDE 405 Adopted: 2 October 2012
OECD GUIDELINE FOR THE TESTING OF CHEMICALS
Acute Eye Irritation/Corrosion


United States Environmental Protection Agency | Prevention, Pesticides and Toxic Substances (7101) | EPA 712-C-98-195 August 1998
EPA Health Effects Test Guidelines
OPPTS 870.2400
Acute Eye Irritation

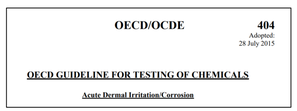
		Conditional Probability of Subsequent Study Categorization			
		1	2A	2B	NC
Prior Type	1	73%	16%	0.4%	10%
	2A	4%	33%	4%	59%
	2B	0.2%	4%	16%	80%
	NC	1%	4%	2%	94%

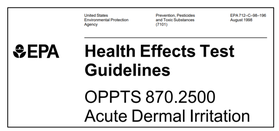
N = 491 substances with at least 2 rabbit eye tests (Luechtefeld et al. 2016 – DOI: 10.14573/altex.1510053)

Conditional probabilities: How likely is the same hazard category if the same chemical is tested multiple times? Variability is greatest when testing mild and moderate eye and skin irritants.

Skin Irritation Testing

Skin Irritation Classification			
EPA Category	**PDII**	**Signal Word**	**Effect**
I	Corrosive	DANGER	Corrosive (tissue destruction into the dermis and/or scarring)
II	>5.0	WARNING	Severe irritation (severe erythema or edema)
III	2.1-5.0	CAUTION	Moderate irritation
IV	0-2.0	CAUTION	Mild or slight irritation


OECD/OCDE 404 Adopted: 28 July 2015
OECD GUIDELINE FOR TESTING OF CHEMICALS
Acute Dermal Irritation/Corrosion

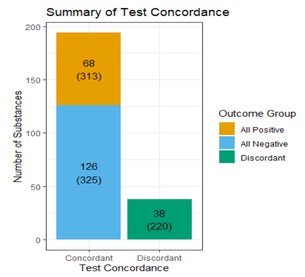

United States Environmental Protection Agency | Prevention, Pesticides and Toxic Substances (7101) | EPA 712-C-98-196 August 1998
Health Effects Test Guidelines
OPPTS 870.2500
Acute Dermal Irritation

		Conditional Probability of Subsequent Study Categorization			
		COR	II	III	IV
Prior Type	COR	86%	4%	7%	2%
	II	14%	45%	20%	20%
	III	7%	5%	54%	34%
	IV	1%	2%	9%	88%

N = 425 substances with at least 2 rabbit skin tests (Rooney et al. 2021 – DOI: 10.1016/j.yrtph.2021.104920)

Skin Sensitization Testing – Human Patch Test

- Variability and uncertainty of binary test outcomes (sensitizer or non-sensitizer) and DSA1+ (dose per skin area with exactly one positive outcome) were evaluated in a human predictive patch test (HPPT) database.
- Binary outcome variability was evaluated in 232 substances that had at least 2 binary test outcomes. Substances were categorized by overall test concordance. 38 substances had discordant test outcomes (graph above right).
- DSA1+ variability was evaluated in 91 substances that had at least 2 numeric DSA1+ variability MAD(DSA1+) was calculated for each substance.
- The table (below right) indicates the range of responses in this dataset based on chemicals with the minimum, median, and maximum median DSA1+.
- Differences in experimental parameters and physicochemical characteristics (below) were evaluated for association with test concordance and MAD(DSA1+). There were no variables that were significantly associated with higher variability of test outcomes.
- **More information: visit Strickland et al., Abstract 3387 / Poster P579.**



CASRN	Median DSA1+	MAD(DSA1+)
97-00-7	1.51	1.29
144-74-1	1293.75	583.77
34131-99-2	15517.24	0

Experimental Parameters	Physicochemical Characteristics	
• Dose per skin area	• Molecular Weight	• Octanal-Air Partition Coefficient
• Concentration	• Boiling Point	• Octanal-Water Distribution Coefficient
• Sample size	• Henry's Law Constant	• Octanal-Water Partition Coefficient
• Skin patch area	• Melting Point	• Vapor Pressure
• Vehicle	• Negative Log of Acid Dissociation Constant	• Water Solubility
• Test Type		

Summary

- These results indicate that in many cases data from in vivo test methods are highly variable.
- Establishing confidence in NAMs includes considerations of test method variability. It is unrealistic to expect a NAM to achieve a level of concordance with an in vivo test higher than the intrinsic level of concordance exhibited by that test.
- Variability is just one aspect of determining if a NAM is as good or better than the existing in vivo test method.
- Ongoing work involves incorporating human biological relevance into NAM assessment (e.g. Clippinger et al. 2021, doi: 10.1080/15569527.2021.1910291; IVAM SS Best Paper).

More Information

This project was funded with federal funds from the National Institute of Environmental Health Sciences, National Institutes of Health, under Contract No. HHSN273201500010C.

To get announcements of NICEATM activities, visit the NIH mailing list page for NICEATM News at https://list.nih.gov/cgi-bin/wa.exe?SUBED1=niceatm-l&A=1 and click "Subscribe."