National Institute of Environmental Health Sciences
*Division of Translational Toxicology*

**Abstract 3022**
**Poster A123**

# Quantifying the Reproducibility and Variability of In Vivo Guideline Toxicology Studies

**A. Kreutz[1], A.L. Karmaus[1], O.B. Oyetade[1], K. Paul Friedman[2], D.G. Allen[3], E.N. Reinke[1], M. Paparella[4], N.C. Kleinstreuer[5]**

*[1]Inotiv, RTP, NC; [2]US EPA/ORD/CCTE/BCTD, RTP, NC; [3]ICCS, Raleigh, NC;*
*[4]Medical University Innsbruck, Innsbruck, Austria; [5]NIH/NIEHS/DTT/NICEATM, RTP, NC*

## Introduction

- Data from guideline in vivo toxicology studies are used by regulatory agencies to make decisions about chemical classification and labeling for human safety and to inform hazard assessment.
- Guideline toxicology studies remain the "gold standard" against which new approach methodologies (NAMs) are compared for regulatory consideration.
- Retrospective analyses of guideline studies have revealed variance in quantitative and qualitative reproducibility attributed to inherent experimental and biological variability.
- A review of in vivo toxicology guideline study variability and reproducibility can help set expectations for performance evaluation of NAMs.
- Here we present a review of variability and reproducibility analyses of in vivo regulatory toxicology studies.

**Several key questions/points addressed:**

- Which study types have had retrospective analyses of variability?
- How often is the same outcome reproduced in repeat studies?
  - Categorically (e.g., hazard classification)
  - Quantitatively (e.g., points of departure)
- How does variability affect confidence and context for interpreting results?
- What are potential impacts and takeaways for consideration when evaluating NAMs?

## Computing Reproducibility

- Several approaches to characterizing variability, uncertainty, and reproducibility can be informative for retrospectively evaluating in vivo toxicity testing outcomes.
- With results from guideline studies often used for hazard classification and labeling, one approach is to review variability of these categorical endpoints. This has been done using conditional probabilities where reproducibility is determined by calculating how often the same category is identified across replicate studies on the same substances. An example from four studies is presented in Table 1.

**Table 1: Calculating Conditional Probabilities**

| 1st Study Outcome \ Subsequent Study Outcome | Prior Type | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| | 1 | 0% | 66% | 33% | - |
| | 2 | 33% | 33% | 33% | - |
| | 3 | 33% | 66% | 0% | - |
| | 4 | - | - | - | - |

**Chemical X outcomes**

- Study 1: category 3
- Study 2: category 2
- Study 3: category 2
- Study 4: category 1

- Since some of these studies also produce continuous values such as the lethal dose 50 (LD50), another approach is to analyze quantitative variability by computing statistical metrics, including:
  - Standard deviation (SD) - amount of variation around the mean
  - Confidence interval (CI) - interval expected to contain estimated parameter
  - Mean squared error (MSE) - average of squares of difference between true & predicted values
  - Root MSE (RMSE) - square root of the MSE; commonly used for assessing model fit

## Study Types Evaluated

- Published retrospective evaluations of in vivo toxicological guideline study variability and/or reproducibility were retrieved and summarized. While not all study types have data amenable to such analyses, a broad coverage of toxicity endpoints were reviewed.

**Table 2: Studies Evaluated in Retrospective Variability/Reproducibility Analyses**

| Study type | OECD TG | Studies |
|---|---|---|
| Ocular irritation: Draize rabbit eye irritation test | 405 | Weil & Scala 1971, PMID: 5570948; Earl et al. 1997, PMID: 20654315; Cormier et al. 1996, PMID: 8661334; Blein et al. 1991, PMID: 20732076; Luechtefeld et al. 2016, PMID: 26863293 |
| Dermal sensitization: local lymph node assay (LLNA) | 429 | Roberts et al. 2016, PMID: 27470439; Hoffmann et al. 2015, PMID: 26168096; Dumont et al. 2016, PMID: 27085510 |
| Dermal irritation/corrosion | 404 | Rooney et al. 2021, PMID: 33757807 |
| Acute lethality: oral LD50 | 420 | Hoffman et al. 2010, PMID: 20709128; Karmaus et al. 2022, 35426934 |
| Acute lethality: inhalation LC50 | 433 | Hull et al. in prep |
| Subchronic/chronic lethality: repeat dose study | | Paul Friedman et al. 2023, PMID: 37990691 |
| Carcinogenicity: chronic study | 451 | Gottmann et al. 2001, PMID: 11401763 |
| Developmental neurotoxicity | 426 | Paparella et al. 2020, PMID: 32970822 |
| Hershberger assay for androgenic activity | 441 | Browne et al. 2019, PMID: 26066997 |
| Uterotrophic assay for estrogenic activity | 440 | Kleinstreuer et al. 2016, PMID: 26431337 |

OECD: Organisation for Economic Cooperation and Development; TG: test guideline

- Chemicals are typically grouped into categories based on study results, with each of the categories defined based on study outcome.
  - The United Nations Globally Harmonized System of Classification and Labeling of Chemicals (GHS) is the most common classification scheme for chemical hazard categorization.
  - The U.S. Environmental Protection Agency (EPA) also has a separate categorization scheme for several toxicity endpoints, including dermal irritation and acute lethality.

## Categorical Reproducibility of Acute Toxicity Hazards

**Table 3: Ocular Irritation/Corrosion (Draize Rabbit Eye Test)**

| First Study Category \ Second Study Category | 1 | 2A | 2B | NC | Total Studies |
|---|---|---|---|---|---|
| 1 | 73.0% | 16.1% | 0.4% | 10.4% | 46 |
| 2A | 4.2% | 32.9% | 3.5% | 59.4% | 138 |
| 2B | 0.2% | 4.0% | 15.5% | 80.2% | 86 |
| NC | 1.1% | 3.5% | 1.5% | 93.9% | 400 |

NC: Not categorized — Luechtefeld et al. 2016

**Additional Studies:**

| % Reproducible | Number of Test Articles | Number of Studies | Reference |
|---|---|---|---|
| GHS Cat 1: 62.5%<br>GHS Cat 2A/2B: 71.4%<br>GHS NC: 90% | 42 substances | 89 | Barroso et al. 2016, PMID: 26997338 |
| GHS Cat 1: 95%<br>GHS Cat 2A/2B: 88%<br>GHS NC: 100% | 1826 substances | 1860 | Adriaens et al. 2014, PMID: 24374802 |

**Table 4: Dermal Sensitization (LLNA)**

| First Study Category \ Second Study Category | 1A | 1B | Neg | Total Studies |
|---|---|---|---|---|
| 1A | 79% | 23% | 8% | 36 |
| 1B | 18% | 68% | 14% | 65 |
| Neg | 11% | 23% | 66% | 35 |

Dumont et al. 2016

| % Reproducible | Number of Test Articles | Number of Studies | Reference |
|---|---|---|---|
| EC3 NS: 80%<br>EC3 Weak: 68%<br>EC3 Moderate: 63%<br>EC3 Strong: 58%<br>EC3 Extreme: 92% | 38 substances | 333 | Hoffmann et al. 2015 |

EC3: dose that would give stimulation index of 3; NS: nonsensitizer

**Table 5: Dermal Irritation/Corrosion (Draize Rabbit Skin Test)**

| First Study Category \ Second Study Category | I (Corrosive) | II | III | IV | Total Studies |
|---|---|---|---|---|---|
| I (Corrosive) | 86.3% | 4.2% | 7.1% | 2.5% | 207 |
| II | 14.1% | 44.9% | 20.5% | 20.5% | 35 |
| III | 6.9% | 5.2% | 53.6% | 34.3% | 133 |
| IV | 0.9% | 2.0% | 9.1% | 88.0% | 690 |

Rooney et al. 2021

**Table 6: Rat Acute Oral Lethality (GHS Classification)**

| First Study Category \ Second Study Category | 1 | 2 | 3 | 4 | 5 | Total Studies |
|---|---|---|---|---|---|---|
| 1 | 53.3% | 34.9% | 1.5% | 5.1% | 5.1% | 104 |
| 2 | 7.7% | 48.9% | 33.2% | 8.9% | 1.3% | 342 |
| 3 | 0.2% | 7.1% | 61.9% | 28.9% | 1.9% | 1166 |
| 4 | 0.1% | 1% | 11% | 66.1% | 21.8% | 3095 |
| 5 | 0% | 0.2% | 1% | 23.8% | 75% | 2867 |

Karmaus et al. 2022

| % Reproducible | Number of Test Articles | Number of Studies | Reference |
|---|---|---|---|
| 54% of compounds reproduce GHS category | 97 substances | 1060 | Hoffmann et al. 2010 |

**GHS Categories**

Packing Group: I (≤ 5 mg/kg); II (>5 ≤ 50 mg/kg); III (>50 ≤ 300 mg/kg); IV (>300 ≤ 2000 mg/kg); NC (> 2000 mg/kg)
Hazard: OSHA

**Table 7: Rat Acute Oral Lethality (EPA Classification)**

| First Study Category \ Second Study Category | I | II | III | IV | Total Studies |
|---|---|---|---|---|---|
| I | 57.9% | 34.5% | 6.2% | 1.3% | 446 |
| II | 5.7% | 66.5% | 27.5% | 0.4% | 1694 |
| III | 0.5% | 11% | 79.8% | 8.7% | 4646 |
| IV | 0.1% | 0.6% | 44.7% | 54.6% | 788 |

Karmaus et al. 2022

**EPA Categories**

Hazard: I (≤ 50 mg/kg); II (>50 ≤ 500 mg/kg); III (>500 ≤ 5000 mg/kg); IV (>5000 mg/kg)

**Table 8: Rat Acute Inhalation Lethality (EPA OPP Classification)**

| First Category \ Second Category | I | II | III | IV |
|---|---|---|---|---|
| I | 70.3% | 24.6% | 0.4% | 4.7% |
| II | 10.7% | 68.0% | 13.8% | 7.5% |
| III | 0.3% | 25.5% | 46.8% | 27.4% |
| IV | 0.6% | 3.9% | 9.9% | 85.7% |

OPP: Office of Pesticide Programs — Hull et al. in prep

**Table 9: Other Categorical Reproducibility Studies**

| Study Type | % Reproducible | Number of Test Articles | Number of Studies | Reference |
|---|---|---|---|---|
| Carcinogenicity (Chronic testing) | Carc/Non-Carc: 65% between rat sexes and 36% between species (rat and mouse) | 313 substances | 379 (349 in rat, 339 mice) | Haseman et al. 1993 |
| | Carc/Non-Carc: 86% between sexes 74% between species (rat/mouse) | | 379 | Huff et al. 1991 |
| | <50% for tumors in same GHS Cat | 121 substances | | Gottmann et al. 2001 |
| Hershberger | Pos/Neg: 72% | 25 substances | 2 or more studies per chemical | Browne et al. 2019 |
| Uterotrophic | Pos/Neg: 74% | 118 substances | 458 studies | Kleinstreuer et al. 2016 |
| Developmental Neurotoxicity | Pos/Neg for each of six endpoints*: 50-100% | 7 substances | 8 labs | Catalano et al. 1997, PMID: 9457734; Paparella et al. |

*For computing categorical reproducibility statistics, endpoints were grouped into six categories: convulsive, autonomic, neuromuscular, sensorimotor, excitability, activity.

## Quantitative Uncertainty and Confidence Intervals

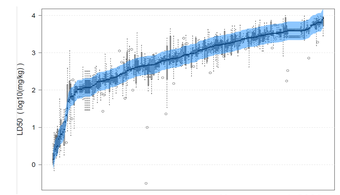
**Table 10: Quantitative Uncertainty**

| | Measure compared | Variability* | Number of Test Articles | Number of Studies Evaluated ** | Reference |
|---|---|---|---|---|---|
| **Ocular Irritation** (Draize rabbit eye irritation test) | | | | | |
| | MAS | Intralaboratory CV 42-59% | 9 | 24 labs | Weil & Scala 1971<br>Earl et al. 1997 |
| | MAS | Intralaboratory CV 38% | 1 | 4 labs, 13 tests | Cormier et al. 1996 |
| | MAS | Intralaboratory CV 3-65% | 4 | 2 labs, 5 occasions | Blein et al. 1991 |
| **Dermal Sensitization** (LLNA) | | | | | |
| | EC3 | SD 0.147 logEC3 | 12 | 94 assays | Roberts et al. 2016 |
| **Acute Lethality** (Oral LD50) | | | | | |
| | LD50 | SD <0.42 log(mg/kg) | 62 | 504 | Hoffman et al. 2010 |
| | LD50 | 95% CI 0.24 log (mg/kg) | 1885 | 5826 | Karmaus et al. 2022 |
| **Subchronic/Chronic Repeated Dose** (Repeat dose study LEL) | | | | | |
| **Study-level** | LEL | Full dataset LEL model: RMSE 0.589 log10-mg/kg/day | 563 | 2724 | Pham et al. 2020, PMID: 33426408 |
| **Organ-level** | LEL | RMSE 0.41-0.68 log10-(mg/kg/day)<br>mean RMSE across organ-level LEL models = 0.59 ± 0.09 log10-mg/kg/day | 58-364, depending on target organ | 151-1353 | Paul Friedman et al. 2023 |
| **Carcinogenicity** (Chronic testing) | | | | | |
| | TD50 | $R^2$ 0.63 mg/kg/d | 121 | 70 | Gottmann et al. 2001 |

MAS: maximum average score; LEL: lowest effect level; TD50: dose resulting in tumors in half of test animals.
*replicate of same chemical, **numbers indicate number of studies unless otherwise specified.

### Defining a Margin of Uncertainty for Acute Oral LD50

- Curated point-estimate LD50 values were used to compute a margin of uncertainty.
- Bootstrapping across mean absolute deviations derived from replicate LD50 values per chemical was applied.
- Blue shading shows defined range of 0.24 $\log_{10}$(mg/kg), which encompassed most experimental LD50 values.




Karmaus et al. 2022

## Summary

- We provide an analysis of variability and reproducibility for numerous in vivo mammalian guideline toxicology studies.
  - Replicate studies are available for dozens to hundreds of chemicals, per study type, allowing for robust retrospective analyses.
  - Study types include acute lethality, (sub-)chronic lethality, dermal irritation/corrosion, ocular irritation/corrosion, carcinogenesis, developmental neurotoxicity, and endocrine activity.
- Comparing the results across retrospective analyses reveals some consistent findings:
  - Hazard classification reproducibility, regardless of study type, is generally lowest for categories describing mild to moderate effects.
  - The most/least potent classification categories tend to have higher reproducibility.
- Categorical reproducibility ranges from 15-94% with the greatest variation being observed in data from the Draize eye test.
- There is a general trend for greater reproducibility in classification schemes with fewer classification categories.
- Quantifying variability and/or reproducibility measures as benchmarks for assay performance can build confidence in NAM robustness. Based on this analysis, we cannot expect NAMs to perfectly replicate in vivo rodent outcomes, which are themselves unreproducible.

## Acknowledgments

*This abstract does not reflect EPA or NIEHS policy, nor does mention of trade names or products constitute endorsement or recommendation for use.* This project was funded in whole or in part with federal funds from the NIEHS, NIH under Contract No. HHSN273201500010C. M. Paparella was financed by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, Department V/5 – Chemicals Policy and Biocides.

To subscribe to the NICEATM News email list scan the QR code or visit https://list.nih.gov/cgi-bin/wa.exe?SUBED1=niceatm-l&A=1.

A. Karmaus' current affiliation is Syngenta Crop Protection, LLC, Greensboro, USA.