Moving Beyond Animal Data as the Gold Standard

Nicole Kleinstreuer
Deputy Director, NICEATM
SACATM
5-6 September, 2018
When Mice Mislead

Tackling a long-standing disconnect between animal and human studies, some charge the researchers need to take safer and better statistics to ensure their science is in line with reality. But the authors had simply left them out of the paper. From analysis of their stroke drug, however, revealed that if mice had had an important advantage to bear. The therapy had to be tested but could help.

"This isn't real," says Clough, who is often critical of mice. Others suggest that a mouse study for many reasons, he says, "is not controlled, accepted part of the culture. You look at your data, see no role. People make mistakes at times when they feel they can't do it."

That had been, he believes, one of several things that were normal studies. So research, science, pharmacological companies, drug regulations, the general public have heightened awareness that mice studies are much of anything it is in mice. Much attention has focused on whether mice will be tested for safety. The authors are not.

News Focus

The Reproducibility Crisis

Rigor Mortis

How sloppy science creates worthless cures, crushes hope, and wastes billions

Richard Harris

Sloppy reporting on animal studies proves hard to change

Scientists appear to ignore guidelines adopted 7 years ago

By Martin Enserink

Sloppy reporting on animal studies proves hard to change

Scientists appear to ignore guidelines adopted 7 years ago

"We just don't seem to make much progress," says Merel Ritsma-Hogeling of Radboud University Medical Center in Nijmegen, the Netherlands, who co-organized a 26 September roundtable in Edinburgh where scientists meet with journal editors and funders such as the United Kingdom Medical Research Council and the Wellcome Trust to discuss ways of speeding up implementation of the guidelines. One problem may be that ensuring compliance can take a bit of work, both for authors and journals.

The 38-member panel provided a "gold standard," says Malcolm Macdonald, a neuroscientist at the University of Edinburgh, who has studied the problems in animal experimentation. The list covers a wide range of issues, from a paper's title and study design to how the animals were bred and reared, and conflicts of interest. But a 2014 survey showed almost no improvement in reporting in journals of Nature Publishing Group (NPG) and PLOS during the first 3 years after the guidelines were introduced, even though both publishers had endorsed ARRIVE. That study's last author, Sandra Arron of U of University Medical Center in Amsterdam, says that an unrepublished analysis shows that things wasn't much better in the 2013-15 period.
Validation Workflow

Importance of Curated Reference Data

Kleinstreuer et al. 2018 RTX in press
Addressing Data Quality
Ex: Rat oral acute toxicity LD50 Database

- Identify transcription errors (e.g. 20005000 mg/kg, >10 mg/kg, confidence intervals as values)
- Manual curation of highly variable chemicals; identify source data

<table>
<thead>
<tr>
<th>Data source</th>
<th>Number of LD50 values</th>
<th>Number of unique chemicals</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECHA ChemProp</td>
<td>5,533</td>
<td>2,136</td>
</tr>
<tr>
<td>NLM HSDB</td>
<td>3,981</td>
<td>2,205</td>
</tr>
<tr>
<td>JRC AcutoxBase</td>
<td>637</td>
<td>138</td>
</tr>
<tr>
<td>NLM ChemIDplus</td>
<td>13,072</td>
<td>12,977</td>
</tr>
<tr>
<td>NICEATM PAI</td>
<td>364</td>
<td>293</td>
</tr>
<tr>
<td>OECD eChemPortal</td>
<td>10,119</td>
<td>2,290</td>
</tr>
</tbody>
</table>
Defining a Confidence Range

Bootstrapping of the standard deviations for repeat test chemicals identified a 95% confidence interval for LD50 values of $\pm 0.31 \log_{10}(\text{mg/kg})$.
## Variation in Classification

Ex: ECHA Ocular Data

<table>
<thead>
<tr>
<th>CASRN</th>
<th>ECHA Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-41-4</td>
<td>Not Irritating</td>
</tr>
<tr>
<td>100-41-4</td>
<td>Category 2A</td>
</tr>
<tr>
<td><strong>100-74-3</strong></td>
<td><strong>Category 1</strong></td>
</tr>
<tr>
<td><strong>100-74-3</strong></td>
<td><strong>Category 2A</strong></td>
</tr>
<tr>
<td>102-06-7</td>
<td>Category 1</td>
</tr>
<tr>
<td>102-06-7</td>
<td>Category 2</td>
</tr>
<tr>
<td><strong>10213-78-2</strong></td>
<td><strong>Category 1</strong></td>
</tr>
<tr>
<td><strong>10213-78-2</strong></td>
<td><strong>Category 2A</strong></td>
</tr>
<tr>
<td>103-50-4</td>
<td>Not Irritating</td>
</tr>
<tr>
<td>103-50-4</td>
<td>Category 2B</td>
</tr>
<tr>
<td><strong>10361-93-0</strong></td>
<td><strong>Category 1</strong></td>
</tr>
<tr>
<td><strong>10361-93-0</strong></td>
<td><strong>Category 2A</strong></td>
</tr>
</tbody>
</table>
Reproducibility of Animal Data

Binary Hazard Classification

- Uterotrophic: ~74%
- Hershberger: ~72%
- Skin Sensitization: ~78%
- Acute Systemic: ~81%
- Skin Irritation: ~76%
- Eye Irritation: ~84%

Kleinstreuer et al. 2016; Browne et al. 2018; Kleinstreuer et al. 2018a; Dumont et al. 2016; Hoffmann et al. 2018; Kleinstreuer et al. 2018b; Karmaus et al. in prep; Leuchtefeld et al. 2018
Reproducibility of Animal Data

Ocular Potency Categorization

Conditional probability of Draize evaluations given a previous test result

491 substances with at least two Draize studies and extractable eye irritation category in REACH registrations 2008-2014

<table>
<thead>
<tr>
<th>Prior Type</th>
<th>1</th>
<th>2A</th>
<th>2B</th>
<th>Non</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73%</td>
<td>16.1%</td>
<td>0.4%</td>
<td>10.4%</td>
<td>46</td>
</tr>
<tr>
<td>2A</td>
<td>4.2%</td>
<td>32.9%</td>
<td>3.5%</td>
<td>59.4%</td>
<td>138</td>
</tr>
<tr>
<td>2B</td>
<td>0.2%</td>
<td>4%</td>
<td>15.5%</td>
<td>80.2%</td>
<td>86</td>
</tr>
<tr>
<td>Non</td>
<td>1.1%</td>
<td>3.5%</td>
<td>1.5%</td>
<td>93.9%</td>
<td>400</td>
</tr>
</tbody>
</table>

Leuchtfeld et al. 2017
How to Benchmark Alternative Models
How to Benchmark Alternative Models

Animal data reproducibility as threshold for performance

Defined Approaches (AOP WoE and KE 1/3 STS) accepted by EPA based on comparison to LLNA (mouse) data

Diagram:

- Test Chemical
- KE a
- KE b
- Concordant?
- YES
  - Classify based on concordance
- NO
  - KE c
  - Classify based on 2/3 concordance
Development of Predictive Models for Acute Oral Toxicity

- Use large database of rat oral LD50 values to train (and test) QSAR models to predict acute oral systemic toxicity
- 32 groups from the US, Europe, and Asia responded with 135 models for LD50, EPA and GHS categories, and binary nontoxic vs all others and very toxic vs all others.
- Models were qualitatively and quantitatively assessed and combined into consensus models.
- Consensus model performance compared with animal test reproducibility for binary, categorical, and quantitative models

https://ntp.niehs.nih.gov/go/tox-models
### Predictive Models for Acute Toxicity: Performance vs Animal Data

**Rat Oral LD50: Reproducibility**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>VT</td>
<td>63%</td>
<td>99%</td>
<td>81%</td>
</tr>
<tr>
<td>NT</td>
<td>96%</td>
<td>82%</td>
<td>89%</td>
</tr>
<tr>
<td>EPA</td>
<td>74%</td>
<td>91%</td>
<td>82%</td>
</tr>
<tr>
<td>GHS</td>
<td>66%</td>
<td>92%</td>
<td>79%</td>
</tr>
</tbody>
</table>

**Consensus Model Performance (Tr/Ts Avg)**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>BA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>77%</td>
<td>95%</td>
<td>86%</td>
</tr>
<tr>
<td></td>
<td>82%</td>
<td>92%</td>
<td>87%</td>
</tr>
<tr>
<td></td>
<td>62%</td>
<td>94%</td>
<td>78%</td>
</tr>
<tr>
<td></td>
<td>54%</td>
<td>92%</td>
<td>73%</td>
</tr>
</tbody>
</table>

### Performance vs Animal Data

<table>
<thead>
<tr>
<th></th>
<th>R2</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD50</td>
<td>0.8</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Communicating Variability to Stakeholders

Review

A Curated Database of Rodent Uterotrophic Bioactivity
Nicole C. Kirsteins*, Patricia C. Craps*, David G. Allen, Judy Straitfield, Xiaoning Jiang, T. J. Amsden, and Wendy M. Corev

Integrated Laboratory Systems, Inc., support of the National Toxicology Program Interagency Center for the Validation of Alternative Methods (NTP-ICVAM), National Institute of Environmental Health Sciences, National Institutes of Health, North Carolina, USA

ABSTRACT: Novel in vivo methods are being developed to identify chemicals that may interfere with estrogen receptor signaling, but the need for in vivo to in vitro biological context remains a challenge. Uterotrophic assays are robust tools for identifying potential endocrine disruptors due to the predictive value of uterine weight increase and because of the lack of high-quality, in vitro reference data. The Organisation for Economic Co-operation and Development (OECD) and the ICVAM Recommendations for the Validation of Alternative Methods (ICVAM-SCRAM) outline uterine weight increase assays that are considered the "gold standard" for identifying potential ER agents.

OBJECTIVES: We performed a comprehensive review to identify data sources from in vitro experiments and to evaluate uterine variability.

Materials: We collected 235 articles on in vitro uterine weight increase experiments using 205 unique chemicals. Study designs, such as quasireplicates, nonreplicates, dosing regimens, and dose levels, were not uniform. Many of the articles are not peer-reviewed, and the results are not always reproducible. The studies included a range of uterine weight increase models (64% active vs. 36% inactive) due to the variability of in vitro cultures. The studies’ findings varied, all of which are considered guideline level 2 (C2) and were subsequently analyzed.

Results: The uterine weight control was not used for 76% of the GI studies. Active outcomes were more prevalent across more results (74%) than across mean results (60%) active. Of the 235 chemical tests used at least five GI studies, 184 (78%) did not demonstrate outliers and were classified as both active and inactive. More reliable results were attributable to differences in study design variability, which included the use of tissue culture media, cell lines, plate types, and uterine species.

Conclusions: The uterine weight increase assay provides a valuable resource for understanding in vitro uterine variability and evaluating the performance of in vivo assays that measure systemic estrogenic activity.

Keywords: Uterotrophic, estrogen receptor, in vitro, reference data.

Table of contents

Introduction

Selection of defined approaches

Qualitative evaluation criteria

Dose-response

Results

Qualitative evaluation

References

Table

Figure

Acknowledgments

References

Predictive Models for Acute Oral Systemic Toxicity: A Workshop to Bridge the Gap from Research to Regulation

Authors: Kirsteins NC, Kermanshah, M, Mounsi K, Allen D, Fitzgerald P, Parlowitz G

Acknowledgments: The authors thank the ICVM-SCRAM members and the Predictive Models for Acute Oral Toxicity Workshop Organizing Committee: O. Axtin, S. Bell, L. Buu-Ba, D. Cronk, J. Gearhart, J. Gordon, S. Aimey, M. Milchak, E. Dendenker, P. Gospodski, L. Scarno, and J. Shideler

Highlights:

- Ongoing implementation of the ICVM Strategic Roadmap, a global modeling project organized to build predictive in silico models for acute oral systemic toxicity.
- An international workshop was held in April 2014 at the NIH to discuss the results of the modeling project, with a diverse group of experimentalists and regulatory agencies working together.
- Relative strengths and weaknesses of the models for different regulatory purposes were discussed, and recommendations and next steps were presented.

Table of contents

Section 1: Overview

Section 2: Methodology

Section 3: Results

Section 4: Conclusions

Highlights

- Towards implementation of the ICVM Strategic Roadmap, a global modeling project organized to build predictive in silico models for acute oral systemic toxicity.
- An international workshop was held in April 2014 at the NIH to discuss the results of the modeling project, with a diverse group of experimentalists and regulatory agencies working together.
- Relative strengths and weaknesses of the models for different regulatory purposes were discussed, and recommendations and next steps were presented.

Table: Table of contents

Section 1: Overview

Section 2: Methodology

Section 3: Results

Section 4: Conclusions

Highlights

- Towards implementation of the ICVM Strategic Roadmap, a global modeling project organized to build predictive in silico models for acute oral systemic toxicity.
- An international workshop was held in April 2014 at the NIH to discuss the results of the modeling project, with a diverse group of experimentalists and regulatory agencies working together.
- Relative strengths and weaknesses of the models for different regulatory purposes were discussed, and recommendations and next steps were presented.

Table of contents

Introduction

Selection of defined approaches

Qualitative evaluation criteria

Dose-response

Results

Qualitative evaluation

References

Table

Figure

Acknowledgments

References

Predictive Models for Acute Oral Systemic Toxicity: A Workshop to Bridge the Gap from Research to Regulation

Authors: Kirsteins NC, Kermanshah, M, Mounsi K, Allen D, Fitzgerald P, Parlowitz G

Acknowledgments: The authors thank the ICVM-SCRAM members and the Predictive Models for Acute Oral Toxicity Workshop Organizing Committee: O. Axtin, S. Bell, L. Buu-Ba, D. Cronk, J. Gearhart, J. Gordon, S. Aimey, M. Milchak, E. Dendenker, P. Gospodski, L. Scarno, and J. Shideler

Highlights:

- Ongoing implementation of the ICVM Strategic Roadmap, a global modeling project organized to build predictive in silico models for acute oral systemic toxicity.
- An international workshop was held in April 2014 at the NIH to discuss the results of the modeling project, with a diverse group of experimentalists and regulatory agencies working together.
- Relative strengths and weaknesses of the models for different regulatory purposes were discussed, and recommendations and next steps were presented.

Table: Table of contents

Section 1: Overview

Section 2: Methodology

Section 3: Results

Section 4: Conclusions

Highlights

- Towards implementation of the ICVM Strategic Roadmap, a global modeling project organized to build predictive in silico models for acute oral systemic toxicity.
- An international workshop was held in April 2014 at the NIH to discuss the results of the modeling project, with a diverse group of experimentalists and regulatory agencies working together.
- Relative strengths and weaknesses of the models for different regulatory purposes were discussed, and recommendations and next steps were presented.

Table of contents

Introduction

Selection of defined approaches

Qualitative evaluation criteria

Dose-response

Results

Qualitative evaluation

References

Table

Figure

Acknowledgments

References
Recent Workshop:
Modelers + Regulators

Predictive Models for Acute Oral Systemic Toxicity

William H. Natcher Conference Center
National Institutes of Health, Bethesda, Maryland
April 11 – 12, 2018

Attendees in-person: 89; webcast: 215
Model Accessibility and Transparency

Mansouri et al. OPERA models
https://github.com/kmansouri/OPERA
How to Benchmark Alternative Models

Human data and human biology as the gold standard

C. Chandrasekera/CCAAM
How to Benchmark Alternative Models
Example: Skin Sensitization

Defined Approaches (DAs) combine *in vitro* and *in silico* data using simple decision trees or machine learning algorithms to predict skin sensitization.
How to Benchmark Alternative Models

Example: Skin Sensitization

All non-animal defined approaches evaluated perform as well or better than the mouse at predicting human skin sensitization:

- Hazard: 74% (mouse) vs. 75-85% (DAs)
- 3-class Potency: 59% (mouse) vs. 55-69% (DAs)
Summary of major recommendations

A true shift in paradigm will require greater emphasis to be placed on human relevance, from top-down funding decisions to data generation, to building of databases and/or knowledge management tools.

International and interagency collaboration is critical: formal collaboration between major organizational and funding bodies should be established.

Funding should be prioritized for researching human-based biology (versus ‘improved’ animal models) and promoting open access data.

Human data should be collected in collaborative, open-access high-quality databases.

Common reporting formats and common ontologies should be established for collecting and collating human biology information, from different ‘omics technologies to human clinical data.

There is a need to establish formal processes for cross-sector communication.

There is an immediate need for the creation of case studies to demonstrate applications and benefits of predictive, mechanism-based approaches in the context of translation and human disease biology, and for the identification of new therapeutics.
Example: Eye Irritation

OECD/OCDE

OECD GUIDELINE FOR THE TESTING OF CHEMICALS

Reconstructed human Cornea-like Epithelium (RhCE) test method for identifying chemicals not requiring classification and labelling for eye irritation or serious eye damage

INTRODUCTION

1. Serious eye damage refers to the production of tissue damage in the eye, or serious physical decay of vision, following application of a test chemical to the anterior surface of the eye, which is not fully reversible within 21 days of application, as defined by the United Nations Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS) (1). Also according to UN GHS, eye irritation refers to the production of changes in the eye following the application of a test chemical to the anterior surface of the eye, which are fully reversible within 21 days of application. Test chemicals inducing serious eye damage are classified as UN GHS Category 1, while those inducing eye irritation are classified as UN GHS Category 2. Test chemicals not classified for eye irritation or serious eye damage are defined as those that do not meet the requirements for classification as UN GHS Category 1 or 2 (2A or 2B) i.e., they are referred to as UN GHS No Category.

2. The assessment of serious eye damage/eye irritation has typically involved the use of laboratory animals (OECD Test Guidelines (TG) 405, adopted in 1981, and revised in 1987, 2002, 2012 and 2017 (2)). The choice of the most appropriate test method and the use of this Test Guideline should be seen in the context of the OECD Guidance Document on an Integrated Approaches to Testing and Assessment (IATA) for Serious Eye Damage and Eye Irritation (3).

3. This Test Guideline describes an in vitro procedure allowing the identification of chemicals (substances and mixtures) not requiring classification and labelling for eye irritation or serious eye damage in accordance with UN GHS. It makes use of reconstructed human cornea-like epithelium (RhCE) which closely mimics the histological, morphological, biochemical, and physiological properties of the human corneal epithelium. Four other in vitro test methods have been validated, considered scientifically valid and

© OECD, 2018

You are free to use this material subject to the terms and conditions available at http://www.oecd.org/ourwork/conditions
## Eye Irritation Classification:
OECD TG 492 Proficiency Chemicals vs. Sigma SDS

<table>
<thead>
<tr>
<th>Chemical Name</th>
<th>CASRN</th>
<th>OECD TG 492 (in vivo data)</th>
<th>SDS (in vivo data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methythioglycolate</td>
<td>2365-48-2</td>
<td>Category 1</td>
<td>Category 2A</td>
</tr>
<tr>
<td>2,5-Dimethyl-2,5-hexanediol</td>
<td>110-03-2</td>
<td>Category 1</td>
<td>Not classified</td>
</tr>
<tr>
<td>1-Ethyl-3-methylimidazolium ethylsulphate</td>
<td>342573-75-5</td>
<td>Not Classified</td>
<td>Category 1</td>
</tr>
<tr>
<td>Diethyl toluamide</td>
<td>134-62-3</td>
<td>Category 2B</td>
<td>Category 2A</td>
</tr>
<tr>
<td>Camphene</td>
<td>79-92-5</td>
<td>Category 2B</td>
<td>Category 2A</td>
</tr>
</tbody>
</table>
Mechanistic Mapping of HTS Assays

Example: Developmental Toxicity

Human Teratogenic Mechanisms

- Endocrine disruption
- Oxidative stress
- Vascular disruption
- Folate antagonism
- Neural crest cell disruption
- Specific receptor- or enzyme-mediated

Van Gelder et al. 2010; Knudsen and Kleinstreuer 2011
Mechanistic Mapping of HTS Assays

Example: Carcinogenicity

Hallmarks of Cancer & Characteristics of Carcinogens

- Inflammation
- Oxidative stress
- Genotoxicity/instability
- Angiogenesis
- Immortalization/proliferation
- Immunosuppression
- Invasion/metastasis
- Specific receptor- or enzyme-mediated

Hanahan & Weingberg 2011; Smith et al. 2016; Guyton et al. 2018; Chiu et al. 2018
Addressing Risk Probabilistically

- Human-Relevant Mechanistic Information (from HTS assays, 3D organotypic systems, QSAR models, targeted animal studies, etc.)
- Exposure Data and Population Genetics (from biomonitoring studies, high-throughput transcriptomics, GWAS studies, etc.)

Risk

Prior

Posterior

% Population
Challenges

• Scientific
  – Considering population/genetic variability
  – Incorporating metabolic competence
  – Developing complex systems models
  – Reporting and collection of reference data

• Non-scientific
  – Increasing awareness, education, and training
  – Cross-sector communication
  – Funding for human-centric research and education