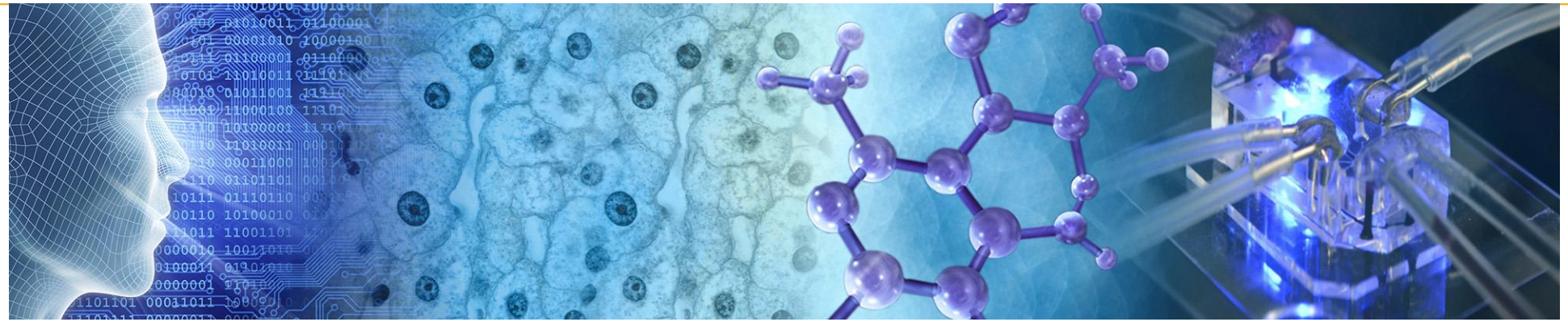
NTP
National Toxicology Program



Variability of animal studies for acute toxicity, skin sensitization, and mechanistic responses

Nicole C. Kleinstreuer, PhD
NICEATM

December 17, 2019

The Reproducibility Crisis

Papers often don't state whether animals were randomized to control and treatment groups.

REPRODUCIBILITY

Sloppy reporting on animal studies proves hard to change

Scientists appear to ignore guidelines adopted 7 years ago

By **Martin Enserink**

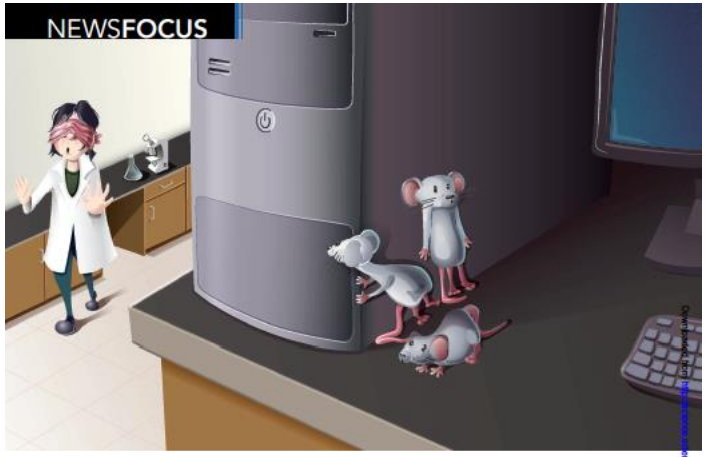
Closely read any paper on an animal experiment, and you're likely to have many questions. What strain of mice was used, and what were their sex and age? Were animals randomly assigned to control and treatment groups? Was the researcher who examined outcomes blinded to what group they were in? The absence of such details partly explains why between 51% and 89% of animal studies aren't reproducible. It may also help explain why so many treatments reported to work in animals have flopped in humans (*Science*, 22 November 2013, p. 922). Yet it's proving surprisingly hard to solve the problem.

In 2010, the U.K. National Centre for the Replacement, Refinement & Reduction of Animals in Research (NC3Rs) in London developed a checklist of items that any paper about in vivo research ought to include. More than 1000 scientific journals and two dozen funding agencies have endorsed the so-called ARRIVE guidelines—short for Animal Research: Reporting of In Vivo Experiments. (*Science* has not officially endorsed them, but encourages authors to comply.) But 7 years later, studies suggest that many scientists are either unaware of the guidelines or are ignoring them.

"We just don't seem to make much progress," says Merel Ritskes-Hoitinga of Radboud University Medical Center in Nijmegen, the Netherlands, who co-organized a 25 September roundtable in Edinburgh where scientists met with journal editors and funders such as the United Kingdom's 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 lot of work, both for authors and journals.

The 38 items in the checklist provide a "gold standard," says Malcolm Macleod, a neurologist 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 cared for, results, 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 2 years after the guidelines were introduced, even though both publishers had endorsed ARRIVE. That study's last author, Sandra Amor of VU University Medical Center in Amsterdam, says that an as-yet-unpublished analysis shows that things weren't much better in the 2012–15 period.

Macleod and colleagues have tested one



When Mice Mislead

Tackling a long-standing disconnect between animal and human studies, some charge that animal researchers need stricter safeguards and better statistics to ensure their science is solid

THREE MICE HAD VANISHED. AND ULRICH Dirnagl had a hunch about where they'd ended up: in the metaphorical dustbin housing animals—and there are lots of them—that line up at an experiment's starting line but are discarded before the finish. The paper that Dirnagl, director of the Center for Stroke Research at Charité University Medicine Berlin, was reviewing described how a new drug protected a rodent's brain after a stroke. The authors used 20 mice, half of which got the therapy. But mysteriously, only seven of the 10 treated animals appeared in a graph analyzing the results.

"I wrote to the editor and said, 'I cannot judge this paper, I need to know where the three mice went.'" Dirnagl recalls. For 6 months, radio silence. Then, the editor responded. He'd heard from the authors, he told Dirnagl. The three mice, suffering from massive strokes, had died, and the authors had simply left them out of the paper. Extra analysis of their stroke drug, however, revealed that those mice had an important message to bear: The therapy harmed the brain rather than helping it.

"This isn't fraud," says Dirnagl, who often works with mice. Dropping animals from a research study for any number of reasons, he explains, is an entrenched, accepted part of the culture. "You look at your data, there are no rules. … People exclude animals at their whim, they just do it and they don't report it." That bad habit, he believes, is one of several that plague animal studies.

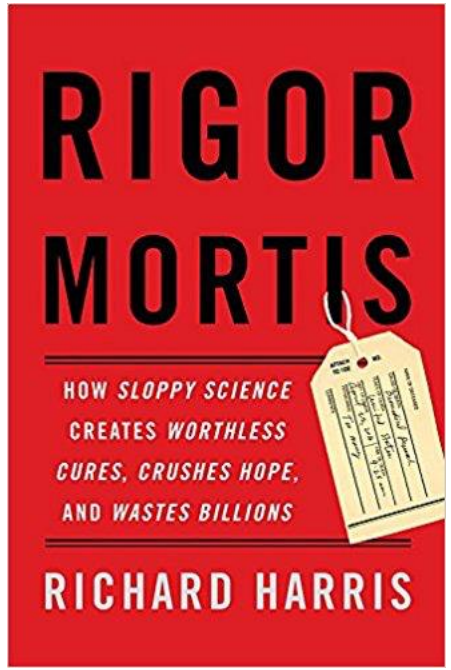
For years, researchers, pharmaceutical companies, drug regulators, and even the general public have lamented how rarely therapies that cure animals do much of anything for humans. Much attention has focused on whether mice with different diseases accurately reflect what happens in sick people. But Dirnagl and some others suggest there's another equally acute problem. Many animal studies are poorly done, they say, and if conducted with greater rigor they'd be a much more reliable predictor of human biology.

It's hard to generalize, of course: Animal studies cut across a massive swath of biology, tracking everything from the activity of single molecules in a healthy organ to side effects of a new drug poised for human testing. And many who stake their careers on animal studies conduct them with care, judiciously weighing how to structure their experiments and chasing the science wherever their furry subjects take it.

That said, even animal research that has a big effect on human drug studies—like the work Dirnagl reviewed—is governed by far fewer standards than clinical trials in people. There, volunteers are randomly assigned by

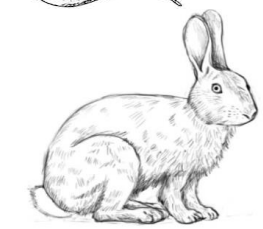
Downloaded from www.sciencemag.org on August 22, 2018

Published by AAAS



Reproducibility of Animal Data: Hazard

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: Potency

Potency Categorization

Conditional probability given a previous test result: Eye Irritation

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

Prior Type	1	2A	2B	Non	Total
1	73%	16.1%	0.4%	10.4%	46
2A	4.2%	32.9%	3.5%	59.4%	138
2B	0.2%	4%	15.5%	80.2%	86
Non	1.1%	3.5%	1.5%	93.9%	400

Leuchtefeld et al. 2017

Historical Insight

TOXICOLOGY AND APPLIED PHARMACOLOGY 19, 276–360 (1971)

Study of Intra- ar
Results of Rab

CARROL

Mellon

Esso
Medical Re

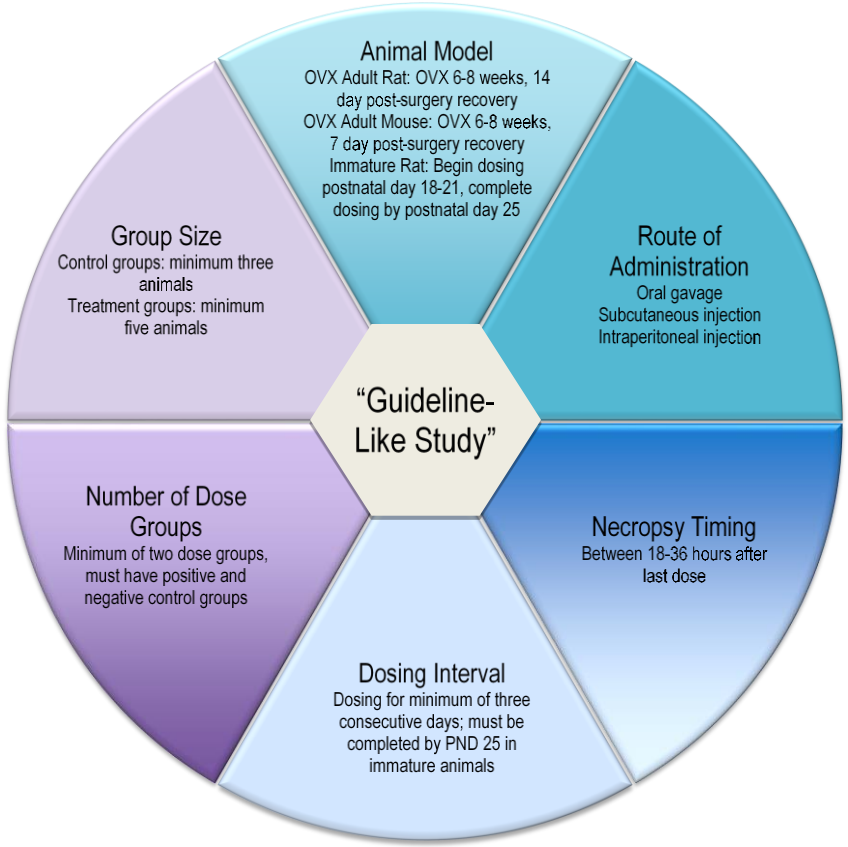
TABLE 12

MINIMUM-MAXIMUM SCORE FOR EYES OF INDIVIDUAL RABBITS—24 HR REFERENCE PROCEDURE

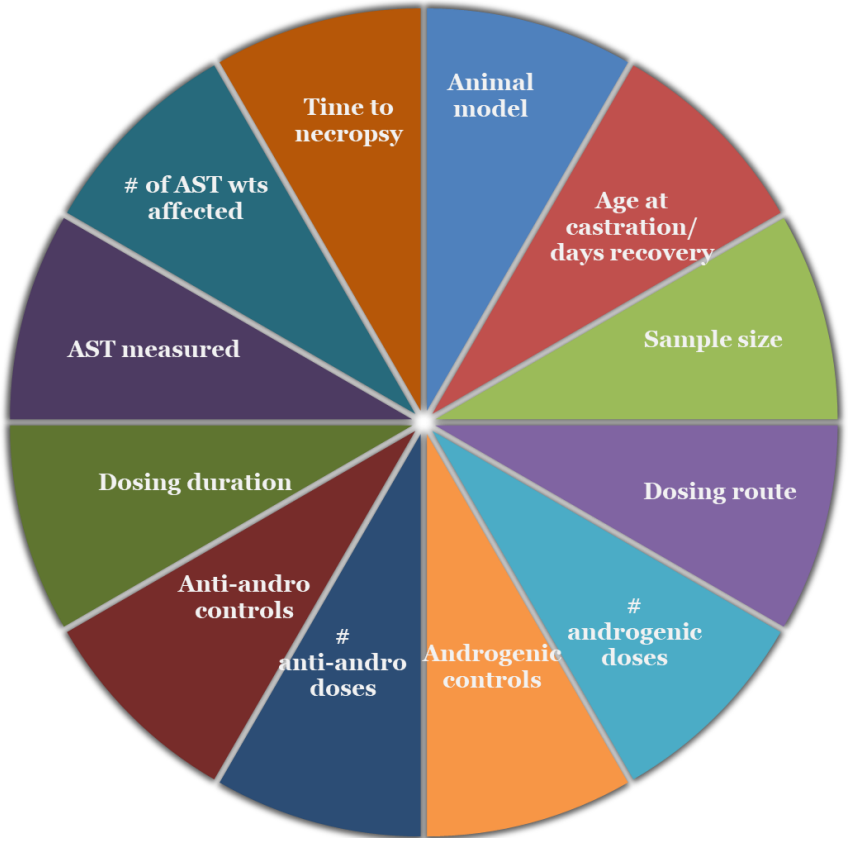
Laboratory No.[a]	Material									Any material
	A	F	G	J	K	L	M	N	P	
14	38–58	58–60	26–60	56–60	18–62	40–68	52–60	21–60	26–52	18–68
29	25–65	58–70	37–90	28–71	14–43	32–53	30–37	23–38	6–36	6–90
4	30–62	46–94	0–25	9–57	59–99	24–65	37–83	37–63	26–63	0–99
12	17–44	30–59	11–90	39–59	14–63	39–63	34–46	37–46	39–46	11–90
1	23–59	35–57	55–86	23–48	10–37	31–59	23–45	6–18	18–64	6–86
10	9–33	40–57	39–59	26–55	8–42	40–57	6–59	2–53	9–35	2–59
31	25–55	26–35	13–53	19–33	8–52	24–53	9–72	9–57	9–37	8–72
22	13–37	30–41	4–79	6–37	10–35	29–39	0–39	0–35	10–35	0–79
2	4–100	25–54	0–23	4–57	4–49	44–63	4–92	8–61	0–65	0–100
9	20–37	26–34	12–61	0–43	10–37	26–37	30–37	18–34	11–32	0–61
11	11–28	11–37	37–71	2–37	6–39	14–39	32–37	23–34	2–34	2–71
23	15–35	24–35	0–61	35–41	8–41	30–39	26–39	9–34	18–37	0–61
25	35–39	23–35	68–92	10–36	6–39	11–37	33–37	4–35	10–37	4–92
21	9–43	74–80	22–56	49–83	10–12	32–76	0–6	6–15	0–8	0–83
5	8–27	6–31	17–29	20–31	18–26	20–24	17–25	20–27	17–31	6–31
19	0–12	15–19	33–70	9–43	17–55	22–53	0–45	20–49	6–13	0–70
16	15–25	8–18	8–110	17–23	18–30	19–25	8–22	8–23	4–16	4–110
24	13–21	8–37	35–63	8–37	6–23	19–34	6–23	15–21	8–17	6–63
7	2–36	42–50	40–72	11–26	11–30	0–11	0–4	0–2	0–2	0–72
13	0–37	2–39	0–57	0–30	0–14	6–39	6–16	6–16	12–32	0–57
18	0–28	16–21	17–62	2–13	6–45	7–16	16–26	2–12	10–18	0–62
8	0–29	8–18	55–86	2–47	2–11	2–14	6–15	4–25	6–35	0–86
27	2–23	14–24	2–66	0–26	all 4	7–26	2–29	2–31	4–21	0–66
30	2–6	12–19	2–30	8–16	13–22	13–25	2–42	13–15	0–4	0–42
Any laboratory	0–100	2–80	0–110	0–83	0–99	0–76	0–83	0–63	0–65	—

Controlling for Study Quality

Uterotrophic



Hershberger



- Systematic literature review to identify "guideline-like" studies
- Identify *in vivo* reference chemicals
  - Active chemicals verified in ≥2 independent studies
  - Inactive chemicals verified in ≥2 independent studies (with no positive results in any study)

Validating NAMs for Endocrine Disruptor Screening

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

Review

Development of a curated Hershberger database

P. Browne[a,*], N.C. Kleinstreuer[b], P. Ceger[c], C. Deisenroth[d], N. Baker[e], K. Markey[f], R.S. Thomas[d], R.J. Judson[d], W. Casey[b]

Reproductive Toxicology 81 (2018) 272–280

Contents lists available at ScienceDirect

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

Evaluation of androgen assay results using a curated Hershberger database☆,☆☆

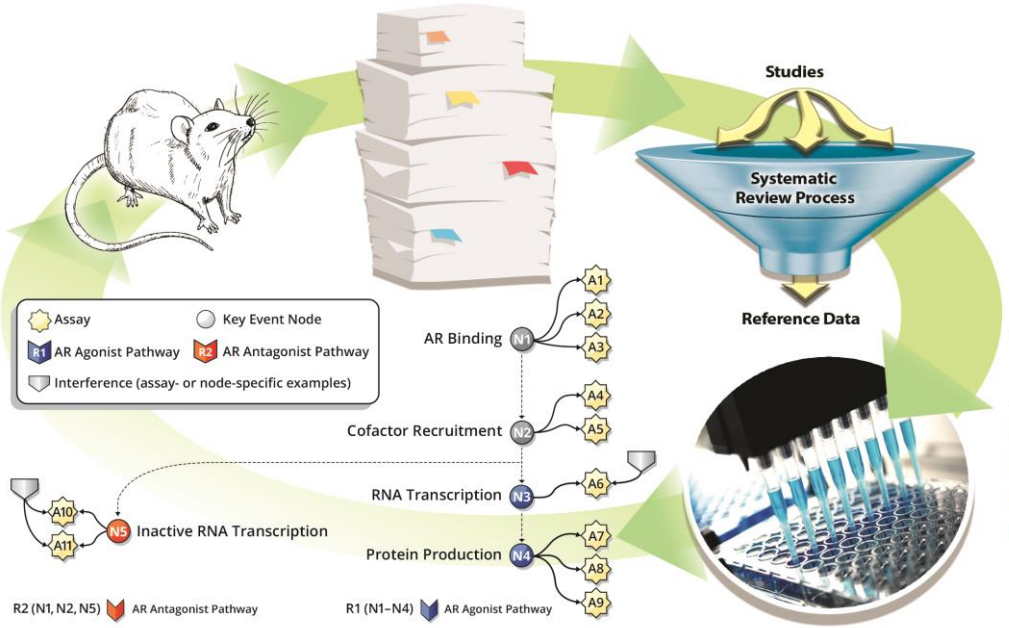
N.C. Kleinstreuer[a,*], P. Browne[b], X. Chang[c], R. Judson[d], W. Casey[a], P. Ceger[c], C. Deiser N. Baker[e], K. Markey[f], R.S. Thomas[d]

ENVIRONMENTAL HEALTH PERSPECTIVES

A Curated Database of Rodent Uterotrophic Bioactivity

Nicole C. Kleinstreuer, Patricia C. Ceger, David G. Allen, Judy Strickland, Xiaoqing Chang, Jonathan T. Hamm, and Warren M. Casey

http://dx.doi.org/10.1289/ehp.1510183 Received: 7 May 2015
Accepted: 30 September 2015
Advance Publication: 2 October 2015



OXFORD SOT Society of Toxicology www.toxsci.oxfordjournals.org

TOXICOLOGICAL SCIENCES, 148(1), 2015, 137–154
doi: 10.1093/toxsci/kfv168
Advance Access Publication Date: August 13, 2015
Research Article

Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 *In Vitro* High-Throughput Screening Assays for the Estrogen Receptor

Richard S. Judson,*[,1] Felicia Maria Magpantay,[†] Vijay Chickarmane,[‡]

ENVIRONMENTAL Science & Technology

Article

pubs.acs.org/est

Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model

Patience Browne,*[,†] Richard S. Judson,[‡] Warren M. Casey,[§] Nicole C. Kleinstreuer,[∥] and Russell S. Thomas[‡]

Chemical Research in Toxicology®

Subscriber access provided by NATIONAL INST OF HEALTH

Article

Development and Validation of a Computational Model for Androgen Receptor Activity

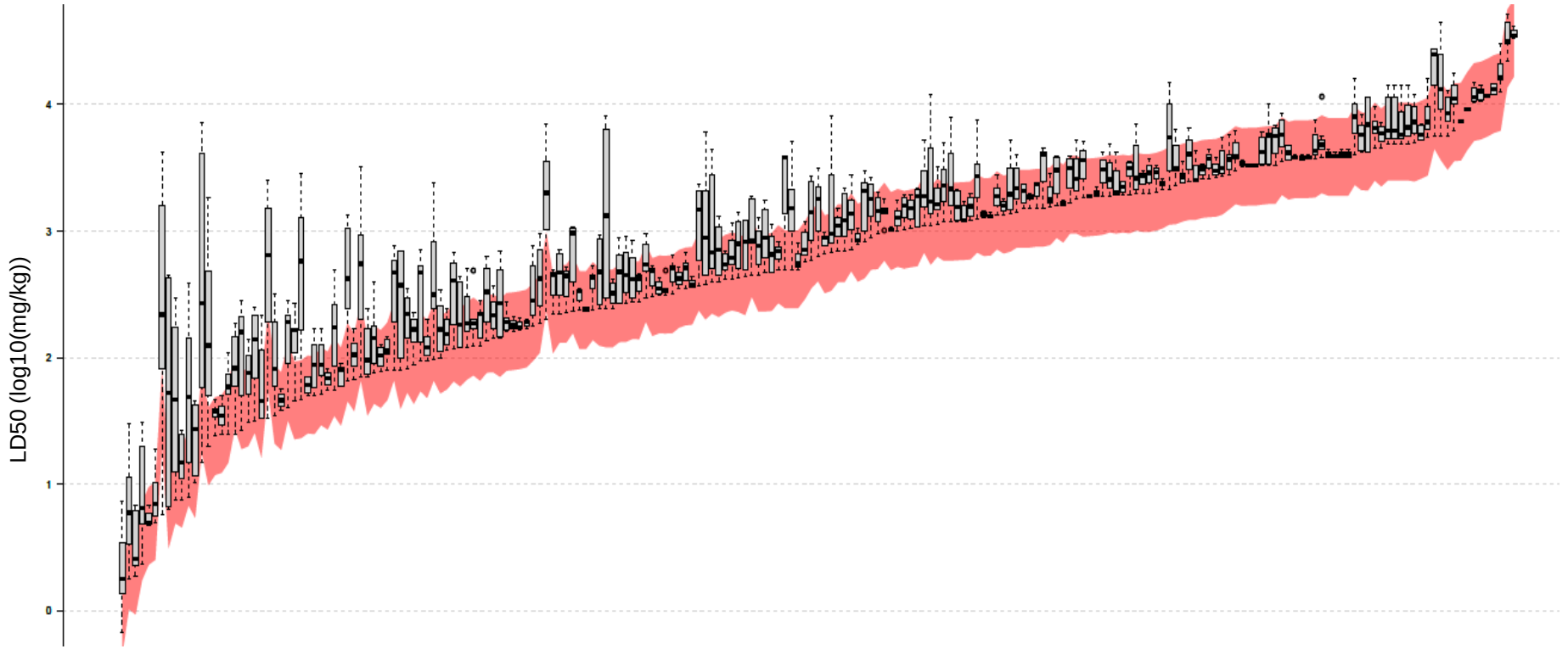
Nicole C. Kleinstreuer, Patricia Ceger, Eric D. Watt, Matthew Thomas Martin, Keith A. Houck, Patience Browne, Russell S. Thomas, Warren Casey, David Dix, David Allen, Srilatha Sakamuru, Menghang Xia, Ruili Huang, and Richard S. Judson

Chem. Res. Toxicol. **Just Accepted Manuscript** • DOI: 10.1021/acs.chemrestox.6b00347 • Publication Date (Web): 18 Nov 2016

Defining a Confidence Range

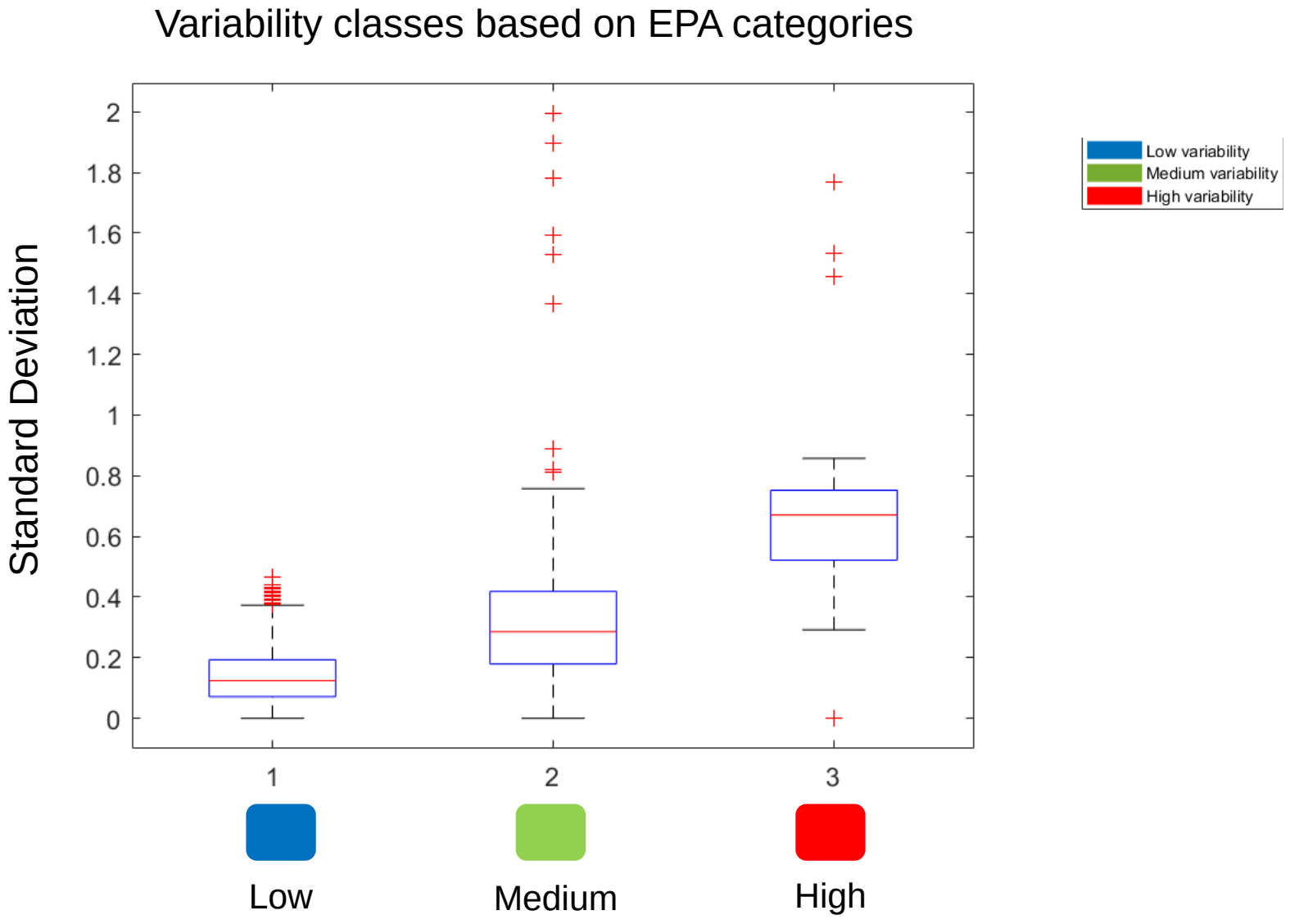
Ex: Acute Oral Toxicity

Bootstrapping of the standard deviations for 1120 repeat test chemicals identified a 95% confidence interval for LD50 values of ± 0.31 $\log_{10}$(mg/kg)



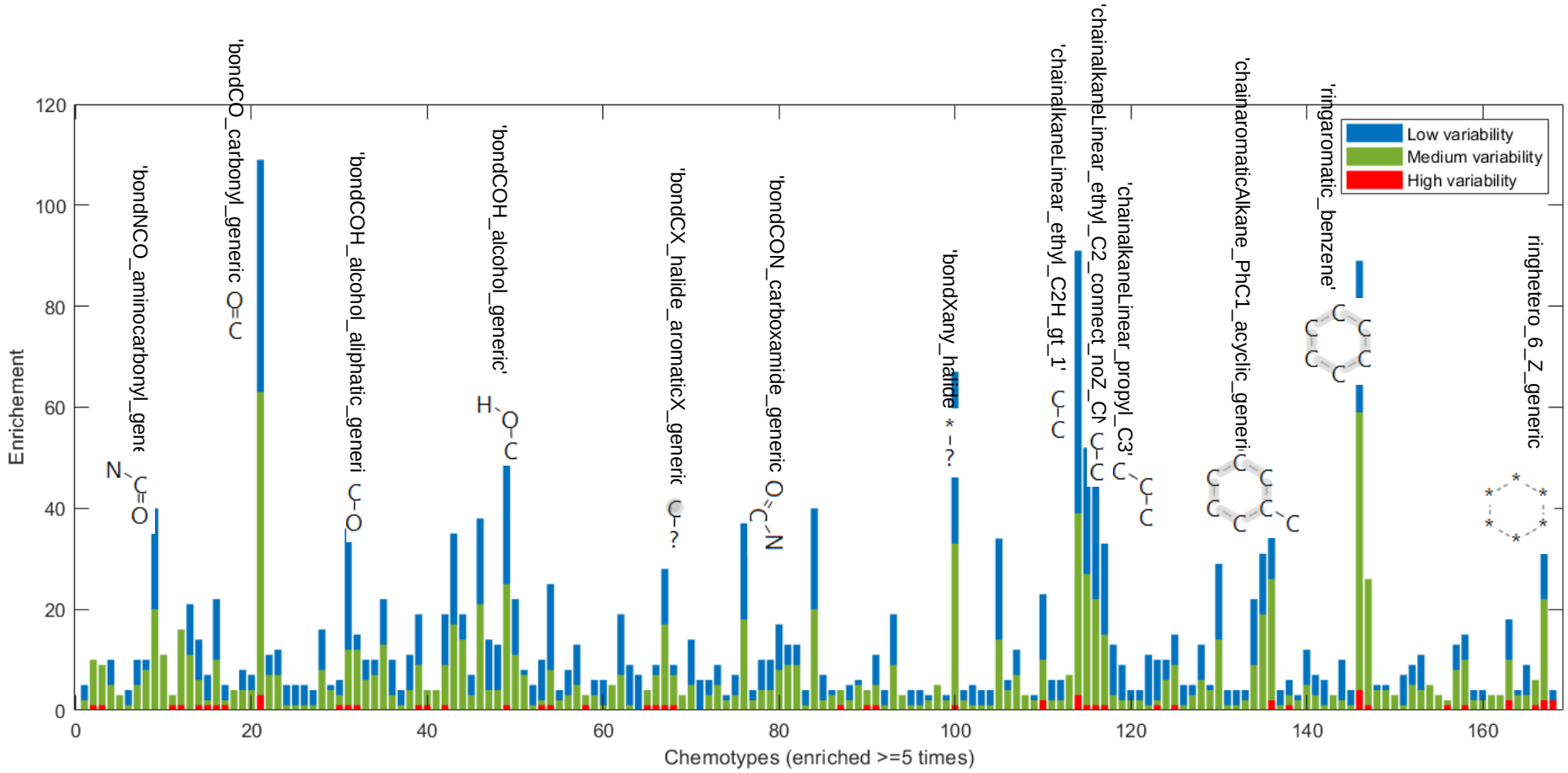
Investigating Sources of Variability

Acute Toxicity Dataset: Chemicals Spanning EPA Hazard Categories



Investigating Sources of Variability

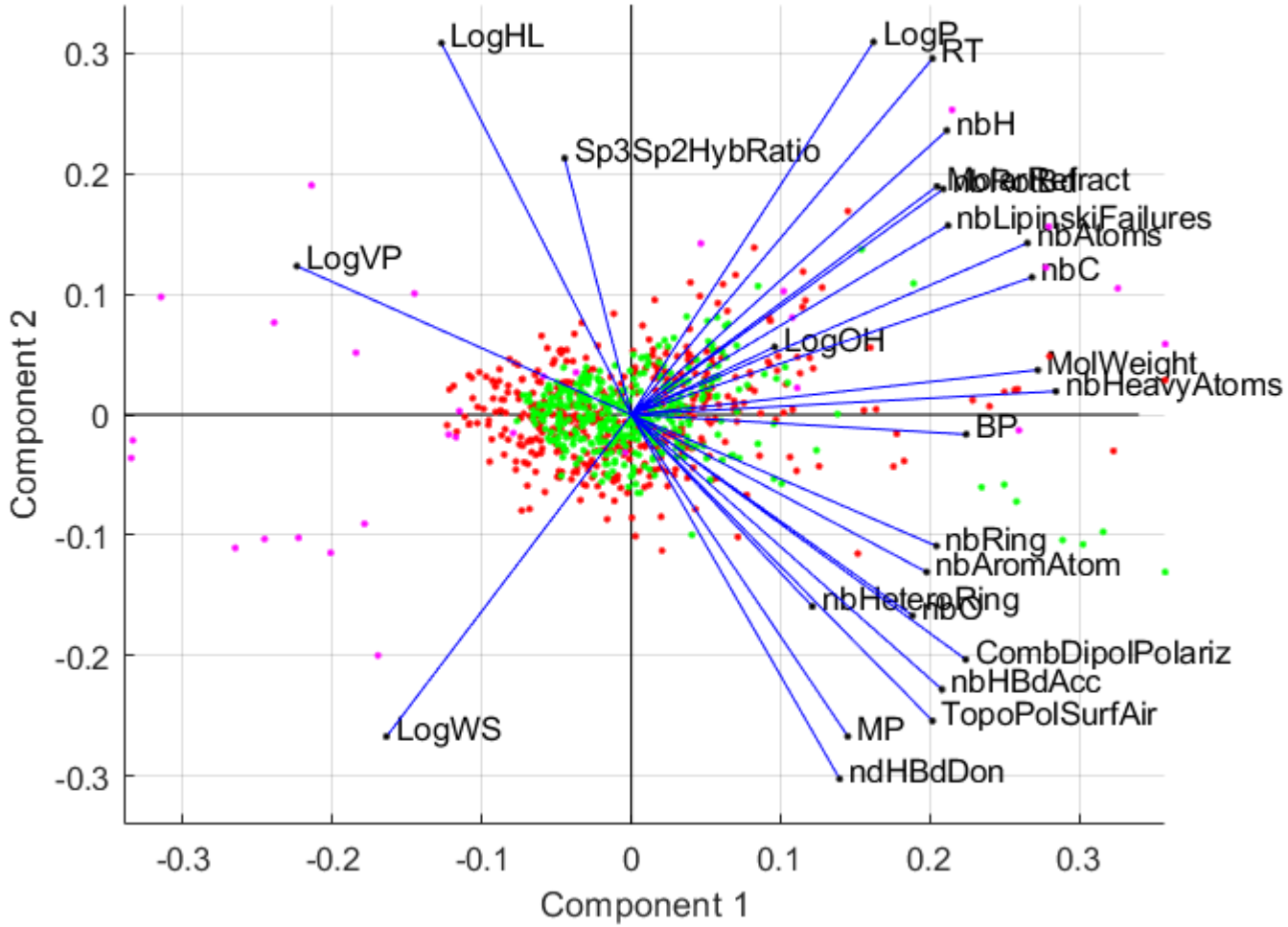
Acute Toxicity Dataset: Chemicals Spanning EPA Hazard Categories



No significant differences found in ToxPrint Chemotype enrichment

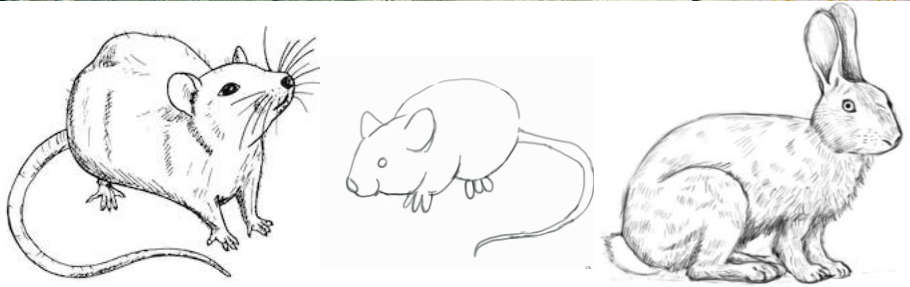
Investigating Sources of Variability

Acute Toxicity Dataset: Chemicals Spanning EPA Hazard Categories



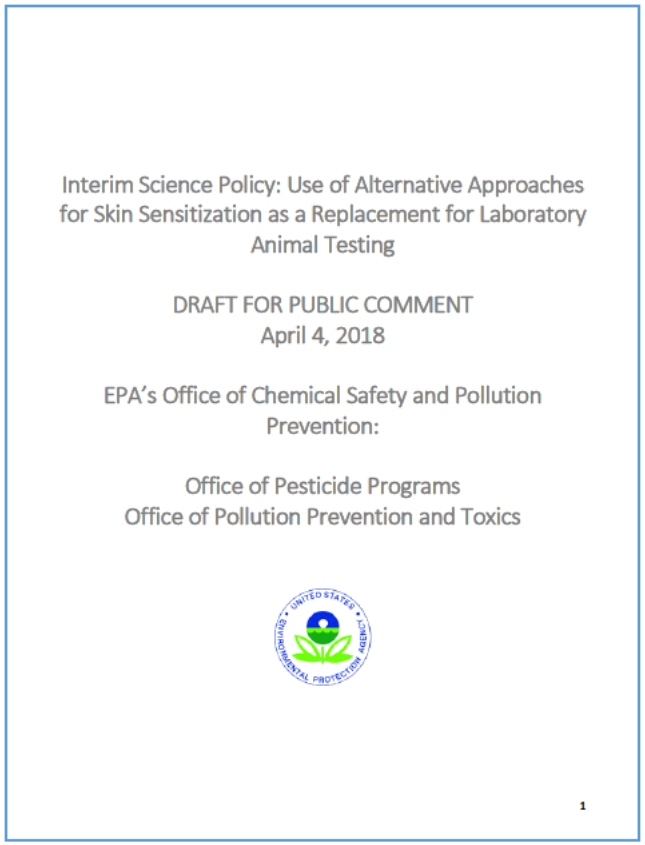
No significant differences driven by physicochemical properties

Benchmarking Alternative Models



Benchmarking Alternative Models

Animal data reproducibility as threshold for performance



Interim Science Policy: Use of Alternative Approaches for Skin Sensitization as a Replacement for Laboratory Animal Testing

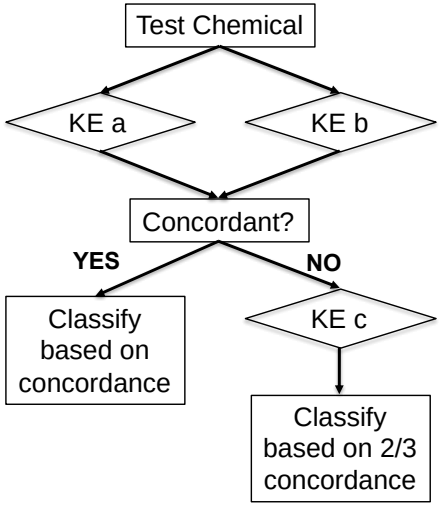
DRAFT FOR PUBLIC COMMENT
April 4, 2018

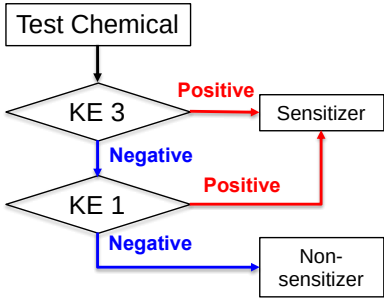
EPA's Office of Chemical Safety and Pollution Prevention:

Office of Pesticide Programs
Office of Pollution Prevention and Toxics

1

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





Skin Sensitization: Lab Animal vs Human Data (n≈150)

LLNA

Hazard	Potency
72%-82%	54% - 60%

GPMT / Buehler

Hazard	Potency
~72%	~60%

Reproducibility of Multiple Tests (~100 chems)

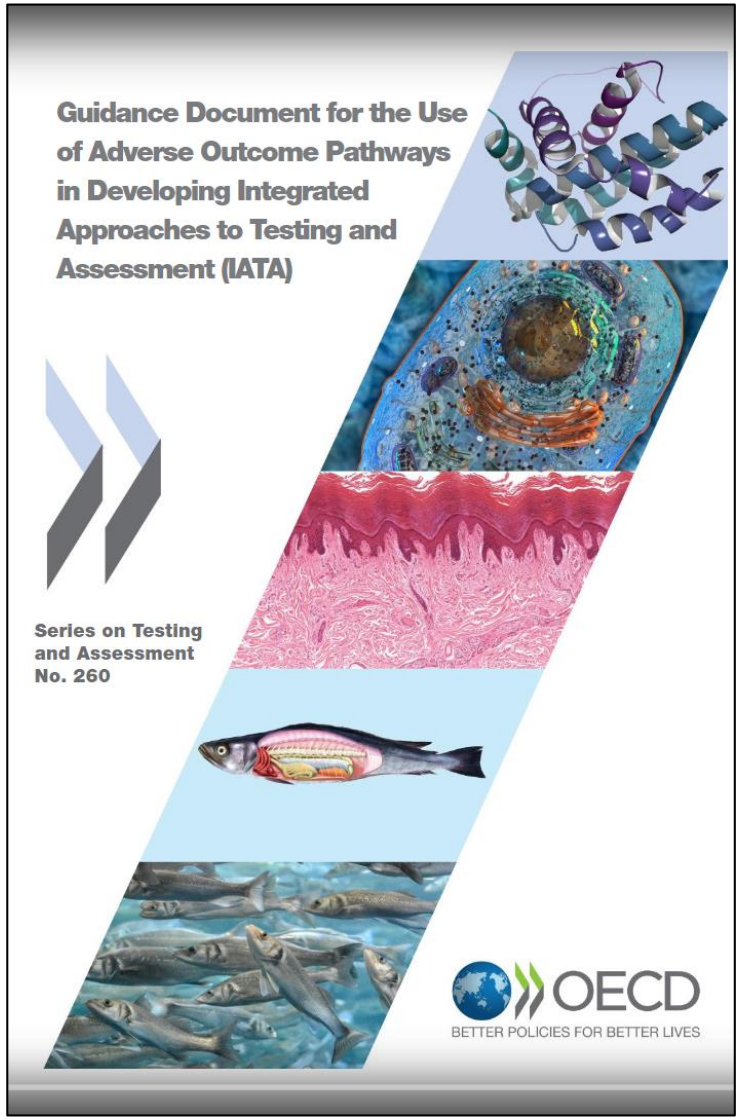
Hazard	Potency
~78%	~62%

ICCVAM. 1999. NIH Publication No. 99-4494
ICCVAM. 2010. NIH Publication No. 11-7709
Urbisch et al. 2015. Reg Tox Pharm 71:337-351.
Dumont et al. 2016. Tox In Vitro 34: 220-228
Kleinstreuer et al. 2018 Crit Rev Toxicol 48(5);359-374

Benchmarking Alternative Models

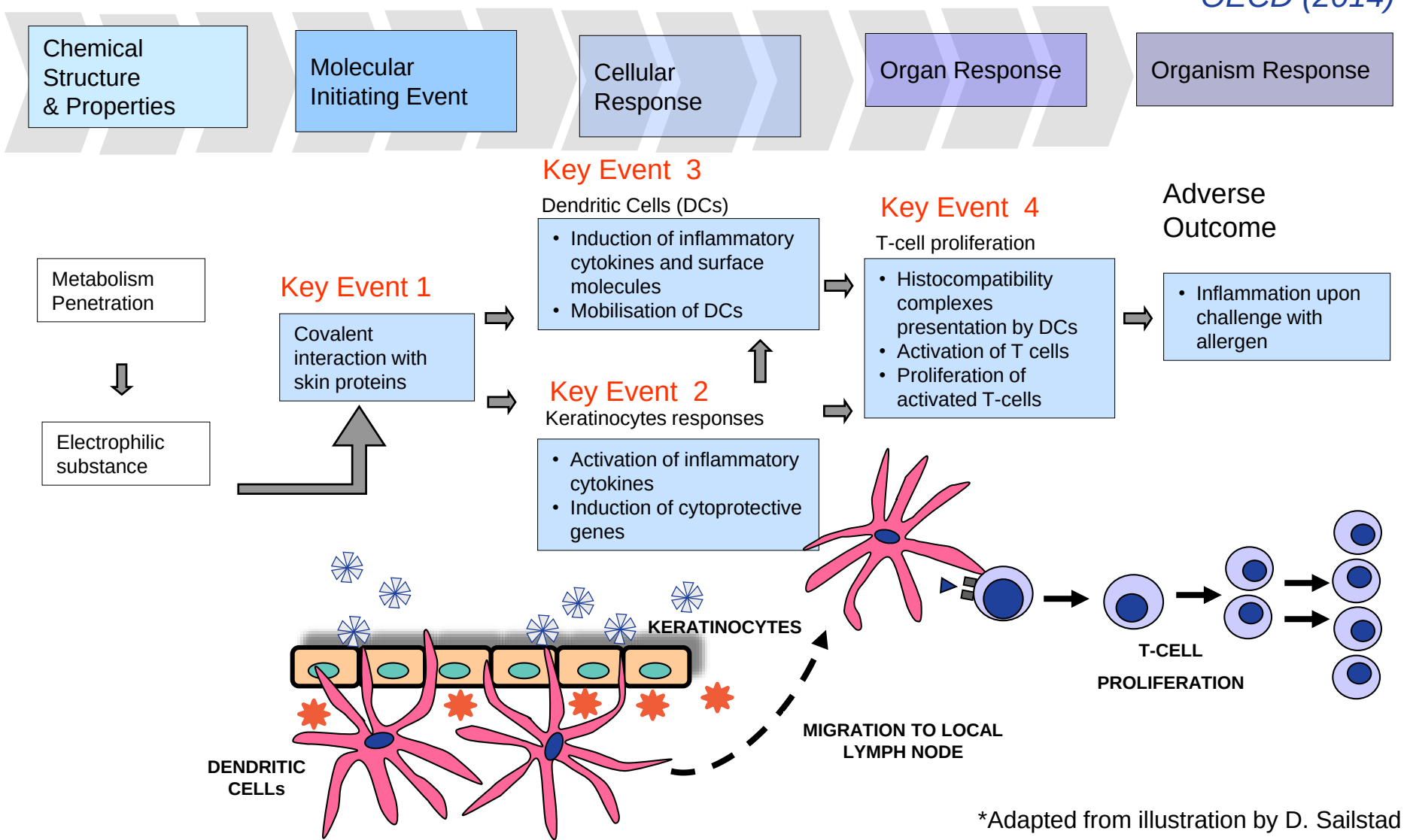
Human data and human biology as the gold standard

Using the AOP framework to develop testing strategies



Adverse Outcome Pathway (AOP) for Skin Sensitization

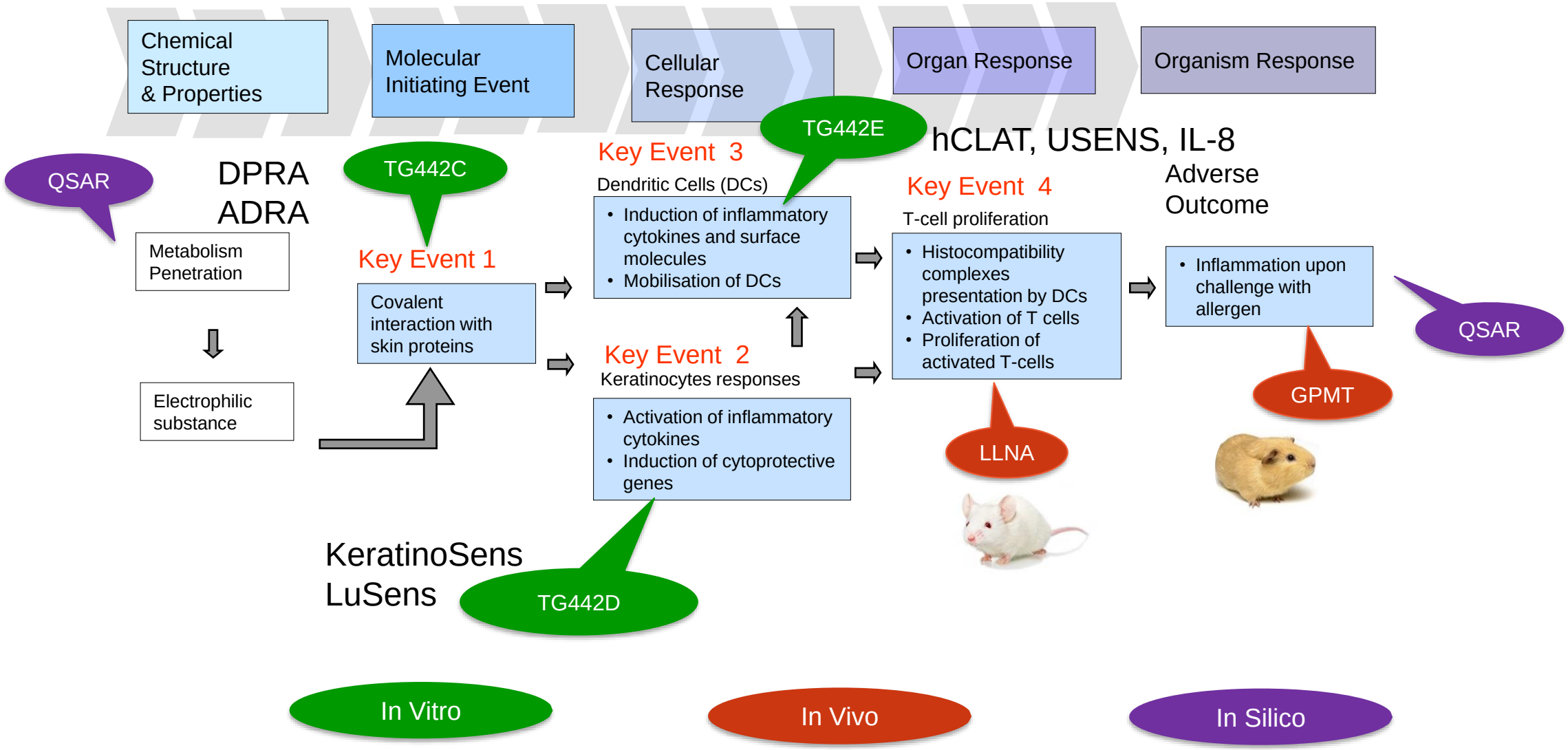
OECD (2014)



*Adapted from illustration by D. Sailstad

Test Methods Mapped to AOP

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

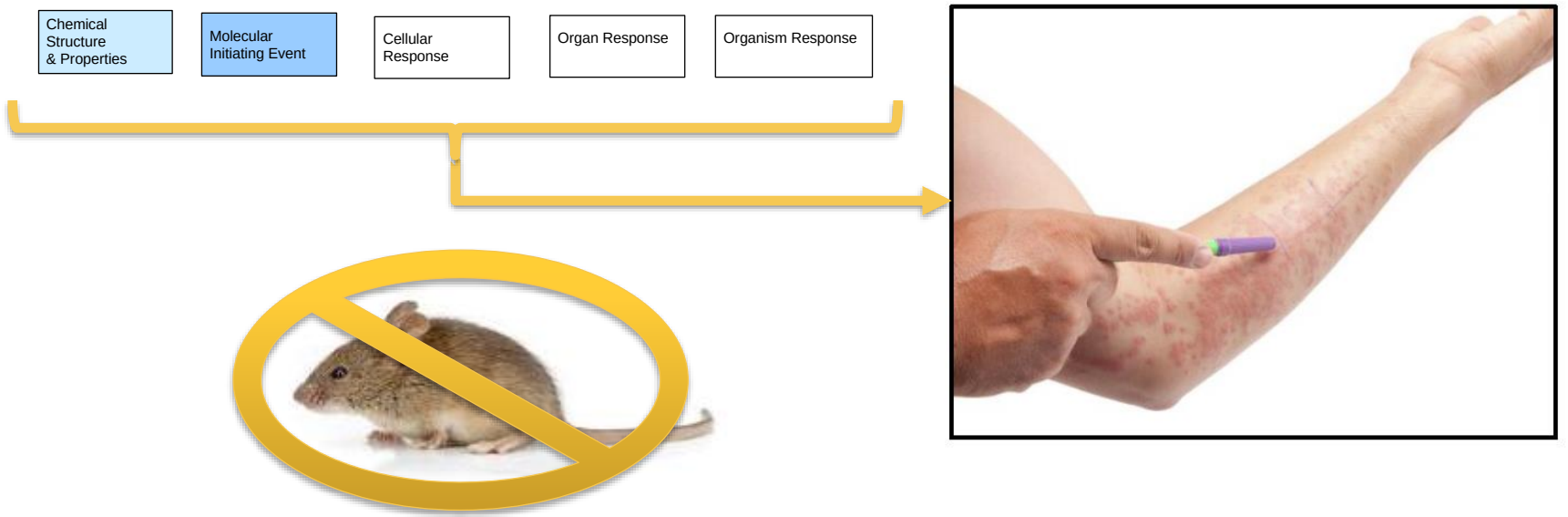


Skin Sensitization DA Performance

All non-animal AOP-based DAs evaluated perform as well or **better** than the animal test at predicting human skin sensitization:

Hazard: 74% (mouse) vs. 75-85% (DAs)

3-class Potency: 59% (mouse) vs. 55-69% (DAs)



Kleinstreuer et al. 2018 Crit Rev Toxicol

Eye Irritation: Reconstructed Human Tissue Models

OECD/OCDE **492**

Adopted:
25 June 2018

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.

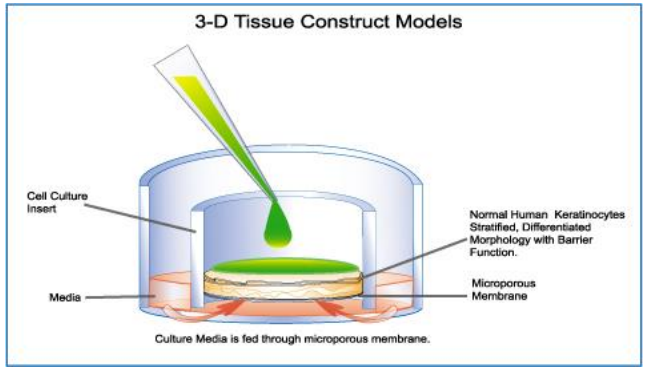
2. The assessment of serious eye damage/eye irritation has typically involved the use of laboratory animals (OECD Test Guideline (TG) 405; adopted in 1981 and revised in 1987, 2002, 2012 and 2017) (2). The choice of the most appropriate test method and the use of this Test Guideline should be seen in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

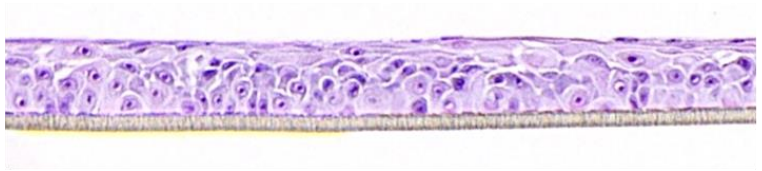
© OECD, (2018)

You are free to use this material subject to the terms and conditions available at http://www.oecd.org/termsandconditions/.

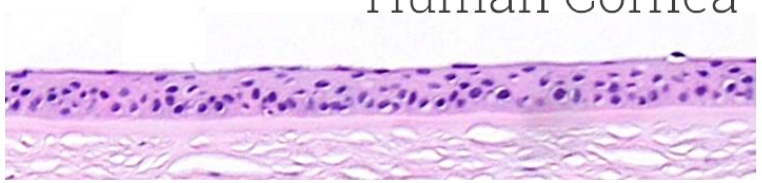
In accordance with the Decision of the Council on a Delegation of Authority to amend Annex I of the Decision of the Council on the Mutual Acceptance of Data in the Assessment of Chemicals [C(2018)49], this Guideline was amended by the OECD's Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology by written procedure on 25 June 2018. guideline was adopted by the OECD Council by written procedure on 25 June 2018.



EpiOcular



Human Cornea

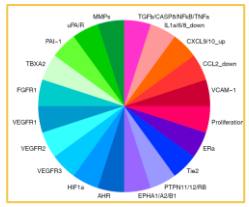


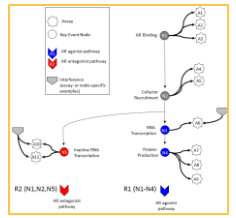
Mechanistic Mapping of HTS Assays

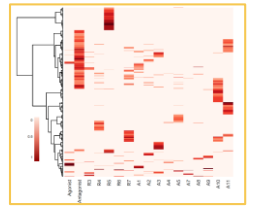
Example: Developmental Toxicity

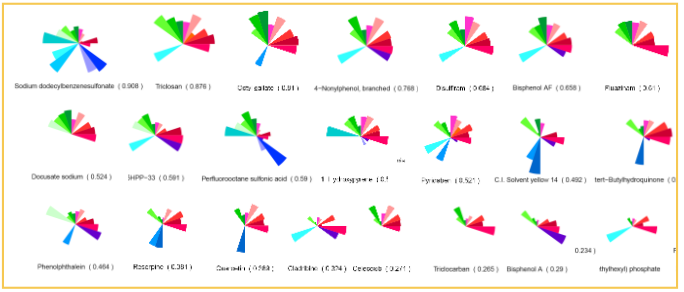
Human Teratogenic Mechanisms

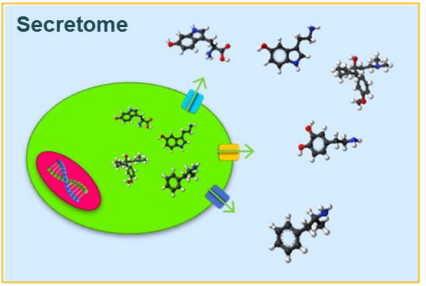
- Endocrine disruption
- Oxidative stress
- Vascular disruption
- Folate antagonism
- Neural crest cell disruption
- Specific receptor- or enzyme-mediated











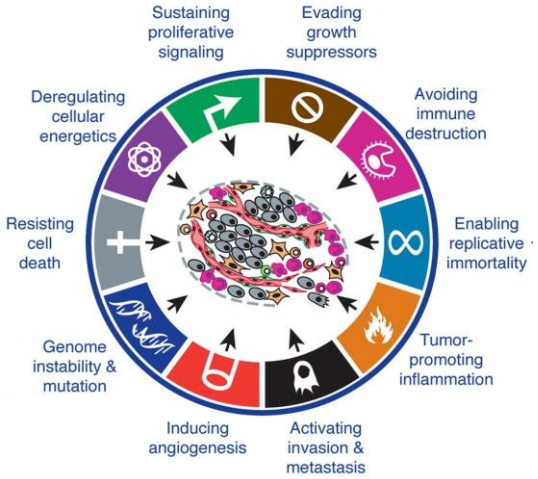
Van Gelder et al. 2010; Knudsen and Kleinstreuer 2011; Saili et al. 2019

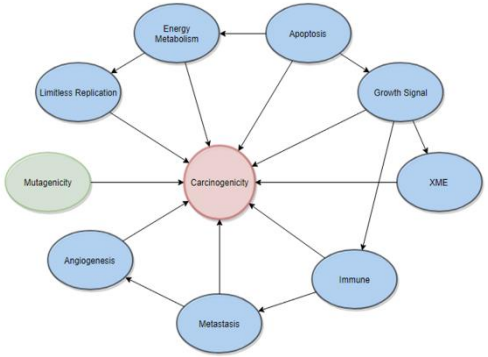
Mechanistic Mapping of HTS Assays

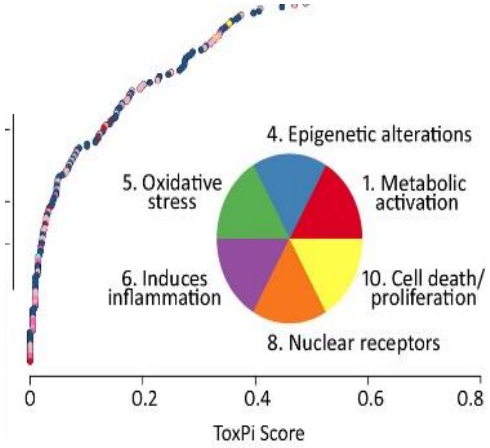
Example: Carcinogenicity

Hallmarks of Cancer & Characteristics of Carcinogens

- Inflammation
- Oxidative stress
- Genotoxicity/instablitiy
- 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

Acknowledgments

- ILS/NICEATM group
- Sciome collaborators
- ICCVAM agencies
- ICATM partners
- Patience Browne (OECD)
- Anna Lowit (EPA)
- Richard Judson (EPA)
- Tom Knudsen (EPA)
- Tony Williams (EPA)

Questions?

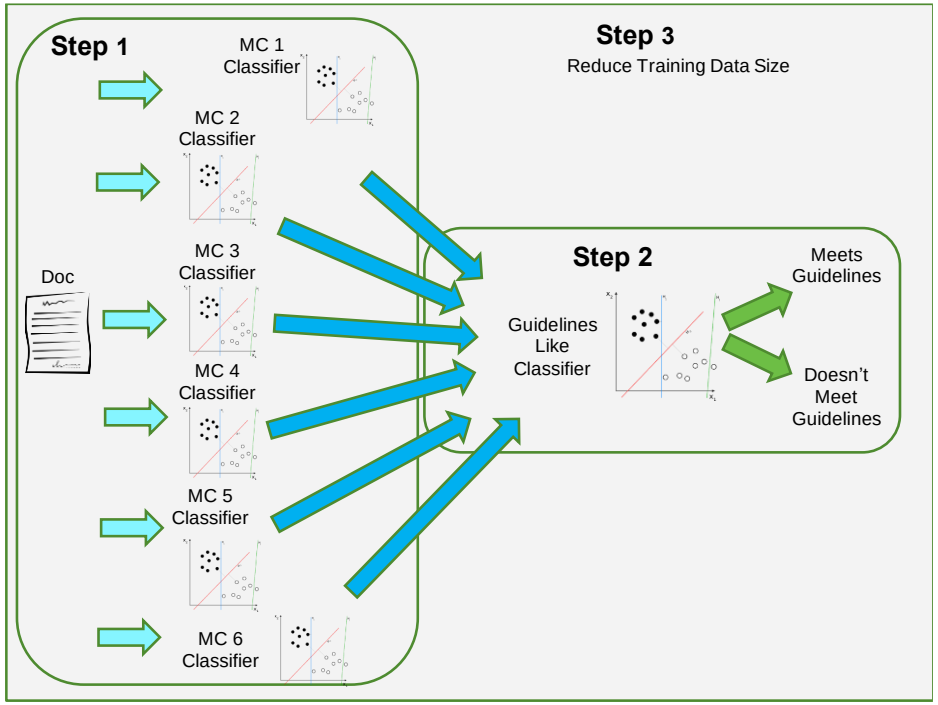
NIH National Institute of Environmental Health Sciences
Your Environment. Your Health.

Extra Slides

Challenges

- Scientific
  - Considering human population/genetic variability
  - Incorporating metabolic competence
  - Developing complex systems models
  - Reporting, collection, and curation of reference data
- Non-scientific
  - Increasing awareness, education, and training
  - Cross-sector communication
  - Funding for human-centric research and education

Automating Reference Data Identification

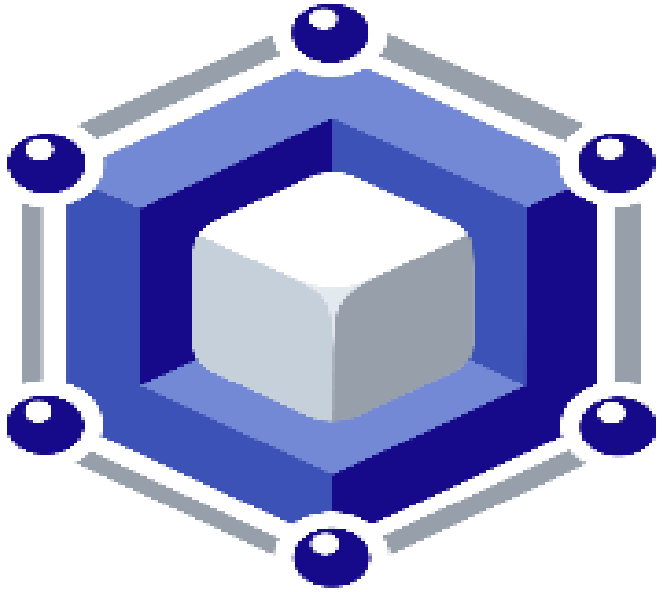


- Project with Oak Ridge National Labs (ORNL) and FDA CFSAN to apply text-mining (NLP) approaches & ML to identify high-quality data
- Semi-automated retrieval and evaluation of published literature (trained on uterotrophic database)
- Apply to developmental toxicity studies (with ICCVAM DARTWG)
  - Define literature search keywords, identify corpus
  - Extract/characterize study protocol details from regulatory guidelines: minimum criteria
  - Apply ML algorithms to identify high-quality studies, expert check

Integrated Chemical Environment (ICE)

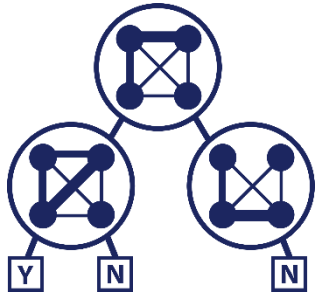
Validation Studies

Published Data



Databases

Computational Models



https://ice.ntp.niehs.nih.gov/