U.S. EPA Conference on the State of the Science on Development and Use of New Approach Methods (NAMs)

Conference Summary

October 12–13, 2022

U.S. Environmental Protection Agency (EPA)

Table of Contents

Conference Overview

Conference Purpose

The purpose of the 2022 EPA State of the Science on Development and Use of NAMs for Chemical Safety Testing Conference was to engage a cross-sectional mix of stakeholders and scientific experts in a dialogue about using new approach methods to evaluate chemicals for potential health effects. The conference consisted of presentations by scientific experts from inside and outside of the agency to inform attendees on advances in the NAMs field. EPA presenters also provided an update on the deliverables outlined in the EPA NAM Work Plan. Speakers presented on the state of the science in NAMs, related to the variability and relevance of traditional toxicity tests as well as the evolution of validation and scientific confidence frameworks focusing NAMS to better extrapolate effects on human health.

As outlined in the EPA NAM Work Plan, the Agency hosts regular conferences to exchange information and solicit feedback. This conference was the third EPA NAMs conference held since the concept of the Work Plan was discussed during the first conference in December 2019. This report summarizes the discussions and presentations which occurred during the 2022 EPA NAM Conference.

Conference Dates, Location, and Materials

October 12 and 13, 2022
U.S. EPA Conference Center and Online via Zoom
Link to NAMs conference website for agenda and presentations - https://www.epa.gov/chemical-research/epa-nams-conference

Participant Summary

There were 97 in-person participants and 711 virtual participants at the two-day conference. The participants were from various sectors including academia, government, and industry.

Day 1 Summary

Welcome: Current Status and Key Goals for the Meeting

Dr. Chris Frey opened the conference and discussed President Biden's commitment to scientific integrity and EPA's commitment to the development, translation, and use of the best available science to advance EPA's mission. Dr. Frey defined New Approach Methods (NAMs) as any technology and methodology approach that can provide information on chemical hazard and risk assessment to avoid the use of animal testing. NAMs have the potential to provide cost effective, relevant information on potential risks. The development, refinement, and adoption of scientific methodology requires careful collaborative work in establishing baselines, measurements, and reporting mechanisms to track progress in meeting goals, establishing scientific confidence, developing NAMs to address information gaps, and engaging with stakeholders.

The December 2021 updated NAMs Work Plan focuses on four key areas:

1) Expansion of species covered in the work plan to include all vertebrate animals to be consistent in TSCA.
2) Modified deliverable timelines that reflect the expansion of covered species and incorporate feedback received over the preceding years.
3) New case studies for building confidence and demonstrating NAMs applications.
4) A pilot study to develop NAMs training courses and materials.

The new additions to this Work Plan demonstrate the strong commitment of EPA to continuing the scientific work of the Agency to develop and apply NAMs. This is a highly active partnership between the Office of Research and

Development (ORD) and the Office of Chemical Safety and Pollution Prevention (OCSPP). Dr. Frey thanked the scientific leads and laid out the NAMs conference agenda before introducing Dr. Russell Thomas.

Dr. Thomas provided the status of EPA NAMs Work Plan deliverables and key goals for the meeting. The EPA research planning deliverable in the Work Plan focuses on EPA strategic research action plans (StRAPs), which include research projects to develop and apply NAMs. The released FY23-26 StRAPs outline the next four years of ORD research activities, with more than 100 research products directly related to research on NAM development and application. Dr. Thomas discussed EPA Science to Achieve Results (STAR) Grants supporting NAMs. Dr. Thomas listed the recipients of the NAMs focused STAR grants. STAR Grants on the Development of Innovative Approaches to Assess the toxicity of chemical mixtures are proposed to begin in 2023.

EPA partnered with five national and international organizations to develop a framework for establishing scientific confidence in NAMs. The EPA NAM website houses information about NAM efforts and progress related to the Work Plan deliverables. EPA NAM activities include a public NAMs training website, new EPA tools and databases, a May 2022 interactive training on ECOTOX knowledgebase, and a new NAMs Update email bulletin used to share progress and updates. He also provided metrics on the number of vertebrate animals used for toxicity tests.

Dr. Thomas discussed the goals for the meeting, including socializing with colleagues, stimulating discussions on the variability and relevance of current animal models, and identifying key elements in the development of a scientific confidence framework for NAMs. Dr. Thomas thanked the members and introduced Dr. Christoph Helma.

Variability and Relevance of Traditional Toxicity Tests

Christoph Helma (In Silico Toxicology Gmbh): Variability in Chronic Rodent Bioassays – Virtual

Dr. Christoph Helma presented on the variability in chronic rodent bioassays. The content of Dr. Helma's presentation included endpoints from rodent carcinogenicity studies and lowest observed adverse effect level (LOAEL) values from chronic rodent bioassays. Dr. Helma described the Nestle and FSVO databases, their LOAEL values, and unique chemical structures. Dr. Helma concluded that the carcinogenicity classifications seem to have poor reproducibility.

Summary

- Dr. Helma began with an overview of the two endpoints covered in his presentation: rodent carcinogenicity classifications (from Gottmann et al., 2001) and LOAEL values from chronic rodent bioassays (from Helma et al., 2018).
- Rodent Carcinogenicity:
    - The carcinogenicity data is from the 1990s and is based on the Carcinogenic Potency Database (CPDB, Gold et al., 1997) which consists of 1,289 unique compounds broken into two subsets (NTP and general literature). 121 compounds are common to both subsets and only 57% had the same classification, indicating poor reproducibility of sex, species, and organ-specific effects.
    - Dr. Helma presented a figure illustrating the correlation of carcinogenicity TD50 values from the NTP/NCI and the literature of the CPDB (r^2=0.63).
    - Carcinogenicity caveats included low sample size and no standardized protocols for literature data.
- Chronic LOAEL Values:
    - The LOAEL data is from chronic studies of rats after oral administration. It consists of two databases (Nestle and FSVO). The combined dataset had 155 unique chemical structures that occur in both databases, representing 375 LOAEL values.
    - Dr. Helma presented a figure illustration of LOAEL variability. All datasets have almost the same experimental variability (Nestle, FSVO, and combined) with standard deviations of 0.56 – 0.57 log mg/kg/d. He also showed a figure illustrating LOAEL correlation (r2=0.52; RMSE = 0.59 log10 mg/kg/d).
- Dr. Helma presented the conclusions of the two studies. Carcinogenicity classification seems to be poorly represented (~57% concordant for repeated experiments). In addition, experimental LOAEL values have a variability of approximately 1.5 log units. He discussed what might cause the variability in chronic, in vivo bioassays, including biological complexity, long term experimental conditions, evaluation complexity, and statistical limitations. Lastly, he concluded that good *in silico* models have the same accuracy as biological experiments for compounds in their applicability domain.

Thomas Steger-Hartmann (Bayer): Using Big Data to Evaluate the Concordance of Toxicity of Pharmaceuticals in Animals and Humans – Virtual

Thomas Steger-Hartmann reviewed the use of Big Data to evaluate the concordance of toxicity of pharmaceuticals between animals and humans. Based on a pharmaceutical perspective, despite the development of NAMs, Thomas Steger-Hartmann asserted animal studies will continue to deliver pivotal contributions to human safety assessment in the next decade. The methodology in evaluating the concordance of toxicity between animals and humans includes a Big Data approach using PharmaPendium. Thomas Steger-Hartmann next discussed ToxHub, a translational system for safety assessment from the European Innovative Medicines Initiative project eTRANSAFE. He described an example for matching for the term steatosis in preclinical and clinical data to show the translational analysis of safety data and went through the ToxHub website and a use case.

Summary

- Thomas Steger-Hartmann began explaining why we are interested in the concordance between animal studies and human outcomes. Despite the development of NAMs, animal studies will remain to deliver pivotal contributions to human safety assessment in the next decade.
- He shared some quotes and studies to demonstrate that while some authors have shown that results from animal studies do not necessarily translate to humans (Bailey at al. ATLA 41, 335-350, 2013), other studies have shown that animals do predict human outcomes (Olson et al. Regulatory Toxicology and Pharmacology 32, 56–67, 2000; Monticello et al. Toxicology and Applied Pharmacology 334, 100–109, 2017).
- Thomas Steger-Hartmann went over his big data approach to a systematic analysis, which was conducted using PharmaPendium.
    - Methodology: PharmaPendium contains over 1.5 million preclinical observations and adverse event reports for human, rat, dog, and other models. It has nearly 4,000 drugs and drug formulations, and the curations are conducted by the PharmaPendium curators.
    - Results: The highest rates of true positives (i.e., a preclinical species predicts the outcome for man) are found for rats and dogs. He showed a graph illustrating true positives per organ class and species, adjusted for the frequency of species used. He showed another graphic indicating which effect has the highest positive likelihood and therefore the best concordance for each toxicological endpoint.
    - Conclusions: The project concluded that certain safety-related findings in animals (e.g., cardiac disorders) are highly predictive for humans, though negative predictivity (i.e., absence of toxicity in animals predicts human safety) is generally low. In addition, predictivity of observations is highly species-specific but also influenced by frequency of animal use for specific endpoints. Statistical analyses are influenced by the size of data, data subset, and subjective terminology assignment.
- Thomas Steger-Hartmann then presented on ToxHub, a translation system for safety assessment developed within the eTRANSAFE project. Using this tool, one can search compounds or drug events and browse data, analyze visualizations, understand mechanisms, and predict models. He described ToxHub functionalities, data sources, and translational analysis of safety data.
    - He presented a use case for ToxHub investigating the translational value of animal data using skin toxicity caused by kinase inhibitors as an example. The main conclusion of the example shows that the higher translational value of the rat over other species regarding skin findings confirms previous findings.
- In summary, Thomas Steger-Hartmann stated that access to big data and application of advanced data science technologies will improve our understanding of the translational value of animal studies and may in the future contribute to a re-design of preclinical programs. Thomas Steger-Hartmann asserted that they will also be able to reverse translate and analyze what might have been missed, which will complement NAM efforts to reduce animal testing.

Alan Boobis (Imperial College): Conservation of Pharmacodynamic and Pharmacokinetic Modes-of-Action in Rodents and Humans – Virtual

Dr. Alan Boobis presented on the conservation of pharmacodynamic and pharmacokinetic modes-of-action (MOA) in rodents and humans. Modes-of-action are key events on the causal pathway from exposure to a chemical to the adverse outcome. These key events are linked by dose response relationships, and they reflect perturbations of

fundamental biological processes. Dr. Boobis discussed an example of the MOA for neurotoxicity of carbofuran. The mode-of-action of acetaminophen hepatotoxicity was also presented as an example, and it was stated that it is a highly conserved mode-of-action in both rodents and humans. Not all MOAs observed in rodent studies are relevant to humans. Dr. Boobis concluded that there is considerable conservation of biochemistry, cellular signaling pathways, anatomy and physiology between rodents and humans with many shared adverse outcome pathways (AOPs)/MOAs.

Summary

- Dr. Boobis explained the similarity between the structure and function of rodent and human anatomy. He used carbofuran as an example of conservation – in this case, the mode of action for neurotoxicity is completely conserved.
- Absorption, distribution, metabolism, and excretion (ADME) processes determine exposure. Passive processes are reasonably well conserved. Dr. Boobis discussed xenobiotic disposition — there is a degree of conservation of function despite variability.
- Dr. Boobis presented a figure from Sarver et al. (1997) that compares plasma half-lives across species. While qualitatively the chemical reactions might be similar, the quantitative rate can vary significantly. This is important for quantitative risk assessment. This data can be extrapolated to predict what would happen in an exposed human population.
- Dr. Boobis pointed out that not all MOAs observed in rodent studies are relevant to humans and gave some examples of this (e.g., bladder tumors by sodium saccharin; mammary tumors in female rats by atrazine).
- Conclusions:
    - There is considerable conservational overlap of biochemistry, signaling pathways, anatomy, and physiology between rodents and humans. While there are many shared AOPs/MOAs, there are also some quantitative differences in dose-response and response-response relationships.
    - Some AOPs/MOAs are rodent specific. Many of these were identified early as the focus was on disproving human relevance and overall are relatively well understood.
    - There are qualitative similarities in toxicokinetics, but many important quantitative differences, often conservation when TK plays a key role in MOA (e.g., metabolic activation, active uptake).

Variability and Relevance of Current Animal Tests and Expectations for NAMs

Nicole Kleinstreuer (NICEATM): Variability and Relevance of Animal Studies for Acute Toxicity, Skin Sensitization, and Mechanistic Responses – In-person

Dr. Nicole Kleinstreuer (NICEATM) presented on acute topical and acute systemic toxicities and how variability has been used to assess new approach methodologies. The presentation also covered the role variability plays in looking at human relevance data and setting new standards for evaluating NAMs. This work contributes to the EPA NAM plan by establishing scientific confidence in NAMs and by improving the quality of NAMs with human mechanistic data.

Summary

- Dr. Kleinstreuer discussed the importance of variability, highlighting that data from traditional mammalian guideline toxicity studies are used by agencies to make decisions about chemical classification and labeling, guideline studies are also the reference upon which alternative methods are assessed, and that better characterizing variability allows setting appropriate expectations for NAMs.
- Dr. Kleinstreuer discussed evaluating reproducibility for categorical endpoints, noting that it cannot be assessed quantitatively, followed by an example.
- She provided an example related to rabbit eye test scoring using health effects test guidelines that examine the cornea, iris, and conjunctiva and are subjectively evaluated and scored.
- These scoring systems are incorporated into eye irritation hazard classifications, which look at the severity of these effects.
- Dr. Kleinstreuer discussed reproducibility of the Global Harmonized System (GHS) hazard classifications for eye irritation based on rabbit Draize eye test. The results showed that chemicals showing severe irritation and no irritation were generally qualitatively reproducible, but moderate to mild irritants showed low reproducibility.

- The OECD guidelines for *in vitro/ex vivo* eye irritation testing were applied based on comparison to the rabbit test, only predicting a top-down or bottom-up approach. This may not be an appropriate approach to evaluating NAMs.
- A similar analysis was performed evaluating the reproducibility of *in vivo* skin irritation using the rabbit model. The results showed that chemicals showing severe irritation and no irritation were generally qualitatively reproducible, but moderate to mild irritants showed low reproducibility.
- She then described variability of the acute oral toxicity test, comparing both EPA and GHS categories. With respect to EPA categories, there was a comprehensive compilation of data from multiple global resources. Data is heavily curated manually and includes limited tests and point estimate data. The reproducibility of the categories ranged from 50 – 80%. The GHS categories were similar to the EPA categories, but there was better reproducibility than with the EPA categories.
- The margin of uncertainty bootstrapped access MADs derived from replicate LD50 values per chemical and shows defined ranges of 0.24 log10 (mg/kg) encompassing most experimental LD50 values
- Dr. Kleinstreuer discussed the value of collaborative crowd-sourcing in the development and validation of predictive computational models. Using the acute toxicity dataset, consensus QSAR models predicted LD50 values with performance equal to variability in the *in vivo* studies.
- Dr. Kleinstreuer highlighted the value of incorporating mechanistic information and human relevance in evaluating NAMs. She pointed to Clippinger et al., 2021 Cut Ocu Tox as a proof of concept to use human biological and mechanistic relevance as the basis for evaluating methods for eye irritation. This represents an exemplar for other endpoints in establishing confidence.
- Nicole Kleinstreuer explained the defined approaches of the skin sensitization guideline which was adopted by the OECD as the first guideline.
- She summarized four key points: (1) *in vivo* data have been used to derive thresholds for hazard categorization, precautionary labeling, and quantitative risk assessments execution, (2) establishing confidence in NAMs should include considerations of variability for *in vivo* test methods, (3), in vivo variability should also be considered to determine if concordance with NAMs is an appropriate comparison, and (4) mechanistic relevance to humans should also be carefully considered to adequately determine confidence.

Katie Paul Friedman (U.S. EPA): Qualitative and Quantitative Variability of Repeat Dose Animal Toxicity Studies – In-person

Dr. Katie Paul Friedman presented work to characterize the variability of *in vivo* study data as a key element in establishing benchmark performance expectations when evaluating NAMs and establishing scientific confidence. This presentation covered both quantitative and qualitative variability in traditional animal study data generated for regulatory toxicology purposes by examining repeat dose study data in the U.S. EPA Toxicity Reference Database (ToxRefDB). This work is critical to defining expectations for NAMs as multiple frameworks suggest that fitness for purpose, biological relevance, and performance required of NAMs should be equivalent or better traditional animal study performance.

Summary

- Dr. Paul Friedman noted the importance of characterizing the variability of in vivo repeat dose data which informs NAM performance expectations and a part of scientific confidence.
- Variability can be expressed quantitatively where variance is a measure of how far values are spread from the average, as well as qualitatively in knowing if a specific effect is always observed or not.
- In Part 1, Dr. Paul Friedman described the quantitative reproducibility of systemic findings in repeat dose animal studies based on her previously reported work in Pham et al. 2020.
- ToxRefDB 2.0 was used to develop statistical models of the variance in quantitative systemic effect level values [i.e., Low Effect Levels (LEL), Low Observed Adverse Effect Levels (LOAEL)] from repeated dose toxicity studies.
- The two statistical modeling approaches used to estimate variance in these data were multi linear regression models and augmented cell means modeling.
- The variance results suggest that repeat dose studies for regulatory toxicology, as conducted and subsequently curated, may have an inherent irreducible amount of unexplained variance.
- The amount of explained variance (explained by study descriptors) approaches 55 – 73%.
- In Part 2, Dr. Paul Friedman described the reproducibility organ-level effects in replicate repeat dose toxicity studies, where organ-level effects inform biological reproducibility.

- Depending on how the data were grouped, the sample size, organ, and species have a big effect on qualitative concordance of effects, which ranged from 33-88%. Within species concordance tended to be greater than within study concordance. Organs associated with more negative chemicals had higher rates of concordance than organs with high frequencies of findings (liver and kidney).
- Previous literature reports suggest that the observed qualitative concordance in organ-level effects is similar to inter-species concordance of carcinogenic findings.
- Estimates of variance in curated LELs and/or LOAELs typically approach 0.5 log10-mg/kg/day at the study-level and the organ-level.
- Analysis using odds ratios suggested that it is unlikely to observe a qualitative positive at the organ-level in a chronic study if the chemical-matched sub-chronic study was negative for that organ. Additionally, differences between sub-chronic and chronic organ-level effect LELs were similar in size to estimates of replicate study variance. Together, these two findings suggest that sub-chronic and chronic LEL values could potentially be combined in training datasets for prediction of repeat dose point of departure.
- In summary, the main conclusions drawn were:
    - Variability in *in vivo* toxicity study data used in training or evaluation of NAM performance limits the resultant predictive accuracy of NAMs.
    - Understanding that a prediction of an animal systemic effect level within ± 1 $\log_{10}$ mg/kg/day fold demonstrates a very good NAM is important for acceptance of NAMs for chemical safety assessments.
    - For qualitative and quantitative reproducibility of organ-level effect observations in repeat dose studies of adult animals, the highest concordance was seen within species and there were similar estimates of quantitative variance to previous work at the study level.
    - The construction of NAM-based effect level estimates that offer an equivalent level of public health protection as effect levels produced by methods using animals may provide a bridge to major reduction in the use of animals as well as identification of cases in which animals may provide scientific value.

Additional points made:

- A conference participant asked, given complex biological systems in the sensitivity to initial conditions, how much variability should be expected in complex in vivo studies. Dr. Paul-Friedman answered that the true amount of variability will never be known, due to limitations on curation of the data, reporting of the data, and guideline structures. Based on current annotations and curation of legacy toxicity studies, approximately 55–73% of the total variability can be explained.
- A conference participant next asked how much variability should be accepted in a complex biological system. Dr. Paul Friedman noted that it depends on where in the AOP the endpoints measured fall (e.g., may observe less biological variability in effects on molecular initiating events than effects on apical outcomes).

Chad Blystone (Division of Translational Toxicology): Inter-Species Concordance of Toxicological Endpoints – Virtual

Dr. Chad Blystone presented on inter-species endpoint comparisons using NTP bioassays. Dr. Blystone gave background on NTP carcinogenicity assays with two species and two models and discussed how level of evidence calls are made for each sex and species. The presentation contributes to the EPA NAM work plan by showing how concordance across species can strengthen interpretation of studies.

Summary

- Dr. Blystone provided background on the ~600 NTP carcinogenicity bioassays, with two species and two sexes, primarily rat or mouse, with a large variety of chemicals and routes of exposures evaluated.
- There are level of evidence calls that are made for each sex and species.
- Dr. Blystone noted that there is a higher level of concordance between species and sexes than endpoint concordance.
- He defined the species endpoint concordance, noting that genetics play a key role in response within a species.
- Dr. Blystone added that study design and conduct are similar and that the evaluations of outcomes are not necessarily interpreted independently (i.e., a strong response in one species may influence interpretation in another species).

- Dr. Blystone explained the neoplastic response within tissues, including the GI tract, urinary bladder, thyroid gland follicular cell, mammary gland, and lung, with high overall species concordance across all tissues.
    - For the GI tract, 20 chemicals with positive tumor calls with 18 tested in both mice and rats. Sex concordance was 70% for positive calls, while species concordance was 0%.
    - In urinary bladder, 21 chemicals with positive tumor calls. Sex concordance was 52%, while species concordance was 10%.
    - For thyroid gland follicular cells, 32 chemicals with positive calls, 31 tested in both mice and rats. Sex concordance was 41%, while species concordance was 26%.
    - For mammary gland, 44 chemicals with positive calls, 38 tested in both mice and rats. Sex concordance was 14%, while species concordance was 29%.
    - For lung, 71 chemicals with positive calls, 61 tested in both mice and rats. Sex concordance was 58%, while species concordance was 20%.
- He summarized that species neoplastic concordance of positive fundings varies across tissues and that sex concordance is greater than species concordance, as expected.
- Dr. Blystone concluded that concordance across species will strengthen interpretation and covers wider genomic background.

Additional points made:

- A conference participant asked where the group is looking at interspecies connection to humans. Dr. Blystone noted that there are previous publications showing that but nothing recent.
- A conference participant asked if rat LOAEL is lower than mouse LOAEL, showing that rat is protective. Dr. Blystone noted that it depends on the tissue, and from a quantitative standpoint there is no specific answer.

Tom Monticello (Amgen): Concordance of the Toxicity of Pharmaceuticals in Animals and Humans: Lessons from the DruSafe Consortium – In-person

Dr. Tom Monticello from Amgen presented on toxicity concordance of pharmaceuticals between animals and humans. This work is done by the IQ Consortium, and this data is presented on behalf of the IQ Working Group, a consortium of about 40 pharmaceutical companies with the mission of advancing science and tech to develop translational work and transformational solutions. Dr. Monticello's presentation discussed the limited data that addresses the correlation between observed toxicities in animal models and those observed in clinical trials.

Summary

- Dr. Monticello described that IQ is an international consortium for innovation and quality in pharmaceutical development.
- The problem statement shows that current nonclinical testing paradigm based on tradition and ICH guidance assumes that the animal models and study are predictive of human hazards, while limited data exists that addresses the correlations between observed toxicities in animal models to those in the clinic.
- Initial studies were primarily retrospective with only clinically observed adverse events being analyzed; therefore, prevalence was 100% and potential false positives could not be evaluated.
- The DruSafe translational database enabled a prospective, blinded study of 182 molecules with data obtained from regulatory dossiers. This enabled a full evaluation of both sides of the 2X2 concordance table.
- The concordance table summarizes the agreement between a nonclinical animal finding and the respective clinical finding. The table is populated for each target organ for each molecule so then concordance statistics can be determined.
- The first-in-human (FIH)-enabling toxicology study placed emphasis on positive predictive value (PPV) and negative predictive value (NPV) as they are more aligned with nonclinical to clinical translation.
- Prevalence impacts positive predictive value. So when prevalence is low, PPV is low even when sensitivity and specificity are high.
- The DruSafe database has been expanded since the original publication (Monticello et al., 2017).
- The updated database includes a higher number of monoclonal antibodies and primate studies compared to the original database, but otherwise their characteristics are similar.
- Overall conclusions are that the PPV of preclinical studies remains low, while NPV remains high. A similar trend was observed in rodents, dogs, and monkeys.
- Translational observations from longer term testing were disproportionately greater for the small molecule modality and the oncology therapeutic indication.

Additional points made:

- A conference participant asked how the dose-response is considered in the clinical vs preclinical. Dr. Monticello responded that the dose-response is not built in because exposure is important for its effect on the animal vs human, rather than the specific dose.
- A conference participant asked in the chat if "fallen angel" compounds were included in the analysis or were they excluded due to legal implications. Dr. Monticello responded that that data is not shown here because those compounds were never nominated to move to the clinic.
- A conference participant asked in the chat if the NPV at lower doses could predict a safe dose. Dr. Monticello responded that yes it could be predicted because the pharmaceutical scheme is set up for false positives, to promote and identify target organ toxicity even at unreasonable doses.

Breakout Group Discussions with Discussion Leads

The conference participants were split into multiple breakout groups and were asked to address the following charge questions:

1. Given the presentations you heard today, are there any generalizable conclusions from the studies evaluating the qualitative and quantitative variability and inter-species concordance of laboratory mammalian toxicity studies and what are the implications when using them to establish the performance of NAMs?

2. The scientific community has a long history of using traditional animal toxicity studies to assess the risks of chemicals to human health. How has the conservation of mode-of-action between the animal toxicity testing models and humans been incorporated into those risk assessment decisions and were the differences in mode-of-action primarily qualitative or quantitative in nature? What are opportunities for NAMs to better inform decision making related to mode-of-action?

3. Evidence in pharmaceutical safety testing suggests that that traditional animal toxicity studies are better at identifying the absence of an effect (i.e., negative predictive value) than accurately identifying specific adversities. How could this evidence influence the expectations on the use of NAMs in toxicity testing?

Breakout Group Room 1

EPA Lead: Anna Lowit

Participants: Bill Eckel, Carrie Brown, Shadia Catalano, Nicole Kleinstreuer, Annette Guiseppi-Elie, Raja Settivari, Thomas Hartung, Joe Manuppello, Marco Corvaro, Martin Phillips, Andrea Hindman, Natalia Vinas, Patience Browne, Kellie Fay, Rashmi Joglekar, and Athena Keene.

Summary

- The group had a lengthy discussion of negative *in vitro* findings. Some key areas of discussion on this issue involved:
    - EPA's experience with CATMOS was consistent with the concept that low toxicity (e.g., EPA Category III or IV) is more confidently predicted.
    - The group discussed the extent to which it is feasible to compile negative animal data for comparing with NAMs.
    - Some questioned whether there is confidence in negative findings given the potential for co-exposure to multiple chemicals and cumulative exposure.
    - The group explored how we can prove a negative and that negative findings in a single assays does not equal lack of bioactivity. Thus, there needs to be an expansion of defining negatives beyond there being no activity.
- The group also discussed dose/concentration selection *in vitro* testing.
    - The group considered if we should consider testing at more environmental or human relevant concentrations.
    - The fact that *in vitro* studies are often tested up to high concentrations which result in cytotoxicity or hit the limit solubility was discussed.

- The point that negatives are not equal to lack of bioactivity was made.
    - It was stated that we must be cautious around positives and negatives versus the need for dose response.
- The importance of problem formulation was also discussed.
    - For problem formulation, we must start with the exposure and the context of use.
    - There is a need to use the concept of tiered testing to avoid early testing and/or duplicative testing.
    - Building problem formulation should be based on tiered testing.
    - The group examined the analysis of critical effects that drive PODs.

The breakout group began by discussing Question 3. A participant shared that a research group found that if the predicted LD50 was greater than 2000 mg/kg then it agreed with the animal testing data (usually an unbounded value of >2000 or >5000 mg/kg), adding that predicting what is not toxic is just as useful as predicting what is toxic. The group discussed pharmaceutical research and how toxic compounds are sometimes not addressed because they do not reach clinical trial. These compounds are not intended to kill mammals, thus contributing to the expectation that there will be low toxicity. A participant added that the utility of animal data adverse specific effects should be considered if it is agreed that the NPV is most relevant. The group discussed the utility of animal data over 50 years old and noted that it could be used in benchmarking. A participant added that by knowing that something is toxic with a margin of exposure, it is easier to omit such data. The participants discussed whether there could still be confidence in negative findings. Other points made in the discussion of negative findings included: in a panel of animal assays how NAMs would or would not be able to replace the entire panel; that negatives do not translate to bioactivity (e.g., in human primary cells with fruit and vegetable extracts, there was a clear distinction between the phytochemicals and pesticides that were tested in the same systems). The participants discussed what would be seen in terms of bioactivity and its range and how comparators in the groups of compounds would be created in this example, concluding that it is a spectrum and needs to be considered with mixture effects. While “clearly toxic” and “no effect” categories are easy to predict, the NAMs categories are more difficult.

The participants continued with discussing negative results in NAMs. Having a panel of negative results may not assist in the assessment, and participants expressed concerns regarding how confidence in the negative data would be acknowledged without assuming that these chemicals are safe. A participant noted that moving to human systems using stem cells could address this. Another participant continued that a NAM in a ten-fold range would have exposure effects that were either too low or too high; a participant responded that if physical chemical properties have no effect on saturation, then that is as high as one must up to. The group discussed the use of NAMs to eliminate tests of various lengths, for example, NAMs could eliminate acute and 14-day assays and could predict whether to start at the 20-day assay. A participant added that the accumulation of PFAS is an example that could be missed with acute dosing and asked how to incorporate that into the NAM prediction and negative prediction. Another participant noted that by narrowing down to what the exposure is and who is being exposed, a scenario is created, and proving that it is negative is important to having confidence.

The group discussed the industrial chemical sector, where the modes of action are unknown with only a small number of studies on pesticide testing. A participant pondered how confidence would be built with no testing in the toxic space. Data is needed in the industrial chemical space for this, and one participant noted that there is a lot of data with over 10,000 tests done multiple times, but that the negatives are important in repeated-testing scenarios. A participant noted that there is abundant negative data that has not been published. The group concluded that for repeat dose toxicity studies, there are many gaps in NAMs that must be understood. A participant raised the point that ‘6 packs’ should only be looked at for animal points and focus should be brought to repeat dose and mode of action of humans – implementation of this shift is a work-in-progress and not enough effort has been put into the repeated dose to see the effects on some organs. A participant noted that prioritization would be made easier if there was access to data, risk assessments, and insight on the critical effects on the outcome. The group discussed the potential for developing a battery of NAMs in the negative space, which Unilever has done to predict toxicity for realistic use cases that cover the spectrum of expected use. Starting with toxicokinetic and a predictive based would help in determining how to bin those models. A participant brought up that the endocrine disruptor screening program – strong positives could be a way to identify responses, rather than just looking at the negatives. The breakout group concluded with discussion of the benefit of coming up with a battery of assays and identifying controls, which may be informative in understanding predictive toxicology.

Breakout Group Room 2

EPA Lead: Monique Perron

Participants: Don Ward, Anna van der Zalm, Elizabeth Webb, Menghang Xia, Gina Hilton, Amy Clippinger, Maurice Whelan, Katie Paul Friedman, Sue Leary, Krystle Yozzo

Summary

- Risk management has a role to play in decision making and NAMs should inform risk management (NAMs should inform applied decision-making).
- There is a wider discussion to be had about how people perceive risk and appropriate/acceptable levels of risk.
- The context of the use for NAMs is key to their value.
- Key points are:
    - Data validation is difficult: animal testing is itself not well validated, but it is still the standard. With this in mind, how do we validate NAMs?
    - Contexts and scenarios for NAMs (the fit-for-purpose approach to NAM development) are highly meaningful.
    - Risk management should be part of the NAM framework development process.

The breakout group lead posed Question 1, asking whether there were any generalizable conclusions and how those might affect researchers' ability to establish parameters for NAM performance. The group discussed the expectations surrounding NAMs research; one participant noted that there is an expectation of having lots of reference chemicals and statistical concordance, but that there are fewer reference chemicals as endpoints get more complex. Another stated that there is some amount of technical variability that researchers must live with when using NAMs, but that there is a double standard for NAMs because of resistance to NAMs adoption. The group discussed if there is uncertainty around whether human biological relevance is reflected in animal testing as well as in NAMs testing, concluding that there is a difference between uncertainty and variability, and that the scientific framework can account for both. They considered whether animal testing would be as relevant as a NAM for predicting human toxic effects. One participant specified some differences between language used in the European Union (EU) and the United States – he used "reliability" for technical aspects and "relevance" for mechanistic aspects and stated that people who prefer conventional animal testing might consider animal-based studies less uncertain and closer to human relevance than NAMs. The group then discussed whether the first charge question is EPA-specific and how the discussion varies a lot based on the regulatory body and the government involved. While the conference is not focused only on the EPA, but that research must be fit-for-purpose and requires an understanding of who work is being submitted to, and that fitness can improve the usefulness of the NAM. One participant added that the discussion is also relevant for the EU, where the conversation around animal testing and NAMs for regulatory decision-making is complex. The group discussed that while NAMs should reduce animal testing where possible, it cannot fully replace animal testing in the current environment. As new chemicals are developed and released, confidence in QSAR and predictive model methodologies can drop outside of known domains.

The group moved on to discussing Question 2. They talked about MOAs, and how knowing these can reduce uncertainty in risk assessment. A participant commented that MOAs tend to be targeted for a specific endpoint or pathway, and that NAMs can use that mechanistic data. Another participant stated that with *in vitro* assays, isolating cell types can result in similar toxicities in testing when there would be different results because of pharmacokinetics in the body. The group discussed how mechanistic data can also help with interpretation of effects and reducing uncertainty; one participant pointed to the example of endocrine disruptor regulations in the UK, which require whole animal effects and MOA information based on World Health Organization (WHO) guidelines. Those requirements call for higher-tier and longer-duration animal studies and criteria must be developed to assess the chemicals in a particular class. The group then discussed the role of regulators, and that from a regulatory perspective there does not need to be an exact understanding of all events in a pathway (only key pieces of information required for protective standards). In addition, risk managers should be an important part of the conversation, since many public health problems are chronic illnesses like cancer, and the goal of risk management should be to reduce the probability that a chemical on the consumer market would contribute to its incidence. A participant emphasized that NAMs should be used for decision-making, and especially for changing the paradigm around carcinogenicity testing and product development. She stated that data-rich agencies could focus on using that data to build confidence in NAMs for MOA testing. Another participant added that NAMs could be especially

valuable for early screening of chemicals that are bad actors and discarding chemicals and chemical relatives with known adverse effects.

Lastly, the group discussed Question 3. One participant responded that he thought predictive values were very interesting for thinking about the psychology of decision-making. The group talked about risk, and the idea that it is impossible to eliminate risk, but that researchers and regulators have to manage it in relation to chemicals that are useful but have some risk associated. Participants agreed that researchers should reconsider what animal testing is really being used for; NAMs are inherently conservative to some degree, since small changes at the cellular/tissue level from a NAM do not necessarily reflect an adverse outcome downstream. Sometimes a health protective estimate is all that is needed from the regulatory standpoint. The discussion concluded with a participant noting that it is not helpful to compare the speed of adoption of NAMs to animal testing when establishing the performance of NAMs. Finally, a participant posed open questions to the group: How can MOA-based decision-making be a viable method? Is this work with NAMs translatable to other fields?

Breakout Group Room 3

EPA Lead: Lindsay O'Dell

Participants: Jim Stevens, Alison Harrill, Iris Camacho, Kristie Sullivan, Katherine Groff, Eryn Slankster-Schmierer, Zhongyu (June) Yan, Jeff Frithsen, Kathie Dionisio, and Heidi Bethel.

Summary

- Sensitivity and specificity should be the metrics used to evaluate NAMs, rather than positive or negative predictive value.
- Understanding that a predictability of 70% for a NAM is an acceptable value given *in vivo* predictability is also in this range was a key point made.
- Uncertainty may be more tolerable when NAMs are combined in a test battery, so that the sum is greater than its parts.
- Certain *in vivo* tests are not worth doing at all because their sensitivity is so poor.
- There is a lot of work to be done to understand the complex biological mechanisms and pathways involved in toxicologic responses. The participants recognized that there is an opportunity to link NAMs to key events in the AOP.

For Question 1, breakout group participants brought up key considerations such as understanding the chemical domains tested and the reproducibility of animal studies. They discussed the standards by which the predictivity of new methods is judged, noting that flexibility is important (i.e., 70% predictability is an A+). Participants discussed whether researchers could expect to see a different standard for hazard NAMs in particular. One participant noted that it is helpful to think of uncertainty in the context of what a NAM is providing, e.g., ability to predict a true negative; another commented that their conclusion was that there is not a lot of available data. A participant added that researchers should celebrate that negative predictive abilities are stronger than positive predictive abilities, and that biological variability is an issue with NAMs where one-third of the variance cannot be explained. The group then discussed sensitivity and specificity: one participant commented that these should be considered over positive and negative predictive value because sensitivity and specificity estimates are more prevalent in datasets. Another questioned whether the "acceptable" limits of sensitivity and specificity would change when discussing NAMs as a component of a battery versus a single, standalone assay, and whether researchers would accept a lower level of either measure for an individual NAM, if multiple NAMs were being used together. According to another participant, in Dr. Tom Monticello's paper, if either measure was positive, the positive predictive value went up noticeably. Another participant added an example of Dr. Richard Judson's work on endocrine disrupting chemicals, stating that too much data can be equally problematic as not enough data; test robustness is more important than having many tests. The group discussed how current animal testing contains many endpoints that are not human health relevant. Participants noted that focus on human health relevance is important, and researchers should be thinking about which NAM can fill which data gap.

Participants then discussed Question 2. One participant noted that in the pharmaceutical industry, arguments made for a lack of species relevance of toxicity findings are used in both qualitative and quantitative contexts. Another stated that uncertainty factors play a role in risk assessments from a health protective standpoint, but that

some NAMs offer an opportunity to be more precise in determining the human point-of-departure because of increased human relevance owing to using human cells. The group discussed the potential for knowledge gaps toward applying standard uncertainty factors to evaluate human risks, depending on the biological context of the NAM - especially if a pathway used in a NAM exists only in humans and not in other animals.

The discussion of Question 3 began with one participant noting that the question was worded poorly, and that specificity and sensitivity are more relevant values to consider than positive and negative predictive value due to methodological considerations in deriving the values. The group then discussed how the limited availability of historical toxicology data for benchmarking NAMs can make evaluation difficult. There was general agreement that confidence in sensitivity, the fit for purpose of a test, and understanding of the biological system being tested were all key elements of acquiring useful information from a given NAM. Two conference participants both mentioned the importance and difficulty of testing large numbers of chemicals, and the need for a process to prioritize which ones to start testing now to provide robust datasets for NAM benchmarking. A participant mentioned that approaches to regulatory decision-making differ between pharmaceutical and chemicals sectors – EPA will make different choices about what to prioritize in NAMs development than the pharmaceutical sector might due to its responsibilities. The group discussed whether exposure information could help bring researchers closer to a tiered testing structure to decide which chemicals in the environment to prioritize for further testing in NAMs, highlighting the bioactivity-exposure relationship work being done to inform testing prioritization strategies. One participant commented that qualitative decisions rely less on exposure information, and another emphasized that uncertainty is important to consider, but that regulators and researchers might view the same uncertainty values differently. The group ended the discussion by agreeing that NAMs need to be context-specific.

Virtual Breakout Group Room

EPA Lead: Monica Linnenbrink

Summary

- There is a need to harmonize a quantitative method for scoring variability among studies.
- The larger the size of the effect, more robust and repeatable the studies tend to be.
- Subjective variability among staff is comparable to variability in NAMS.
- There is a need to reduce the number of animals and refine assays before pushing to replace animal assays.

For Question 1, Participants discussed a need for consistency; animal testing is full of false positives, and that the issue moving forward with NAMs is that it cannot be a standard without outside validation. A participant stated that some projects are much closer to the goal of practical application, while others are much more theoretical. Data harmonization and a consistent method for scoring variability are important – a participant noted that there are sources of subjective variability like staff training, staff attitude and well-being that can contribute to variability. Another participant stated that size of an effect can make a study more robust and repeatable, which has to be accounted for when setting standards for NAMs. The group discussed categories like ‘dangerous’ and ‘non-dangerous’ and their utility for human health protection and the consideration of inter-lab reproducibility.

The participants then discussed Question 2. Participants discussed what an appropriate method of digitalizing, testing, and interpreting would be to get more standardized data. A participant noted that some NAMs are globally harmonized, but that the current classifications scheme is not reproducible, which brought up whether that opens the possibility of creating new classifications with a more robust and human relevant model. Test guidelines focus heavily on the lab portion of a study but not on the interpretation portion where statistical testing happens – participants emphasized that interpretation should be a focus to help with reproducibility. A participant added that NAMs have an advantage based on acceptance criteria; human relevance also requires researchers to specify a population (it requires the context of either a large population or a smaller cohort).

In response to Question 3, a participant responded that it will be very difficult to replace rodent testing with models because of their 90% similarity to humans. A participant agreed that models will not be able to completely displace rodent testing, but that chemicals of lower concern could be assessed computationally while only more harmful chemicals would be assessed with animals. Another participant asked whether the 90% genetic certainty is relevant to endpoints for testing. A responder noted that systemic toxicity has too many endpoints to hypothesize fully, and that is where the conflict emerges between global testing and targeted testing. NAMs applicability does

depend on a toxicity endpoint, and researchers need to know what they are testing for and the limitations of their methods. The group discussed how the failure rate of test efficacy has not changed in many years, and researchers will have to use complex methods for complex exposures in the future. A participant commented that the success rate could improve by defining applicable limitations to in vitro assays and by devising a better chemical reference list. Animal models are not predictive for cancer endpoints in humans, and researchers must be objective about flaws in different models, including animals. The discussion concluded with a participant noting that refining assays and reducing the number of animals used must come before totally replacing animal models with NAMs.

Day 2 Summary

Welcome and Opening

Dr. Tala Henry introduced Rick Keigwin, who is the Deputy Assistant Administrator for Management for the EPA's Office of Chemical Safety and Pollution Prevention. He provides advice and support to political leadership, and he has "grown up" professionally in EPA's programs. He started in Toxics and, prior to this position, spent many years within the Office of Pesticide Programs. He was very supportive of bringing NAMs into the science space and is now doing that as a primary focus.

Rick Keigwin welcomed the second-day participants and emphasized how important the work is. He stated that it is valuable to have experience, insights, and input on this topic. As a non-scientist, he enjoys having scientists' input. Even starting in the 1980s, he was fascinated by predictive technology like QSARs and has enjoyed following the evolution of the science over decades. The incorporation of new techniques is important from a risk-management standpoint as it gives EPA the ability to do things more quickly and brings a more human-relevant or species-relevant angle. He acknowledges that skepticism around the topic exists and asks participants to consider how to bring skeptics into the conversation, so that the confidence by the public is built and maintained.

Report Out from Discussion Groups

Dr. Tala Henry (PhD, EPA's Deputy Director of the Office of Chemical Safety and Pollution Prevention) gave a summary of yesterday's breakout room conversations. She organized the collective feedback around the three questions that were posed to the groups, and she acknowledged that different groups focused more time and attention on different questions. Dr. Henry summarized takeaways for each question, including some highlights: animal studies may not be the gold standard, even though benchmarking to animal studies is common (Question 1); tiered testing strategies could eliminate the need for *in vivo* testing in certain contexts (Question 2); NAMs could be applied to "unfuzzy the middle", to better understand moderate or weak effect chemicals (Question 3). Finally, Dr. Henry summarized observations across groups, which converged on recommendations that data standardization and data harmonization play an important role in NAMs development and implementation.

Summary (report out from discussion groups)

- Question 1:
    - Animal studies are not necessarily the "gold standard," and PK and PD can contribute to better understanding potential chemical effects. We should strive to create NAMs that address refinements.
    - Quantitative variability in study and organ level effects was generally ±10 fold, and qualitative variability in organ and whole animal responses varied significantly by endpoint but generally did not exceed 70%.
    - In discretized responses for classification and labeling, the high and low response bins were more concordant while middle bins were consistently discrepant.
    - Inter-species concordance of negative responses was generally good, but inter-species concordance of positive responses was generally less than 30%.
    - It is important not to conflate technical variability associated with a testing system with uncertainty related to inter-species concordance.
- Question 2:

  - Many important biological systems are conserved across species, suggesting that many (but not all) MOAs will be qualitatively similar, but have different dose responses, at least in mammals.
  - Consider the application of NAMs as a part of a tiered testing paradigm (NAMs are used to identify potential human relevant MOAs in the initial tiers).
  - From data rich chemicals with known MOAs, review of existing animal in vivo results should evaluate several different factors to frame NAMs development.
- Question 3:
  - Experiences with pesticides is consistent with the concept that low toxicity is more confidently predicted.
  - The evidence from pharmaceutical safety testing of animal studies being better at identifying the absence of effects is consistent with current environmental risk assessment practices. For NAMs, a similar protective approach could be used to identify the dose showing an absence of significant changes in biological processes and pathways.
  - Development of an *in vitro*/NAM battery that is designed to detect chemicals that alter the organ systems that are most sensitive to adverse effects and drive risk assessment should be taken on.
  - Benchmarking the battery can be accomplished using negative animal test data from data calls and registration dossiers, the strong positive chemicals identified from human clinical data and animal studies, and the moderate and weak chemicals for the endpoint of interest, which will take time and experience to develop.
- Overall observations:
  - Most groups noted the importance for standardizing protocols.
  - Most groups identified a need to move toward combining methodologies to fill needs, not just developing individual options.
  - Most groups identified a need to replicate more complex biological mechanisms.
  - Groups suggested considering testing more relevant concentrations in assays.
  - Most groups noted difficulties in predictability between *in vivo* models and *in vitro* NAMs.
- Additional points made:
  - A conference participant commented that if one tests multiple species (of mammals) and the results disagree, the results cannot be trusted as much.
  - A conference participant noted that it is more about evaluating the variability and uncertainty in existing methods, in scenarios where two species disagree.
  - To the question from the chat, Dr. Henry responded that we should not be advocating that NAM assays need to perform superior to the existing animal studies.

Evolution of Validation and Scientific Confidence Frameworks to Incorporate 21st Century Science

Warren Casey (NIEHS): ICCVAM Strategic Roadmap for Validating New Methods – Virtual

Dr. Warren Casey (NIEHS) discussed the evolution of ways in which validation is referred to in current scientific literature and alternative methods to approaching validation, such as the goal of only referring to it as "establishing confidence." He presented the history and the existing outdated guidance regarding establishing confidence, explaining how ICCVAM has been working to develop new guidance to assist end-users in developing NAMs with confidence.

Summary

- The first lesson is to refer to it as "establishing confidence" rather than "validation." This is due to the multiple meanings that have been ascribed to the word.
- Updated guidance should now refer to it as establishing confidence to help standardize the future use of NAMs in practice.
- The 2018 strategic roadmap NTP Interagency Center for the Evaluation of Alternative Toxicological Methods outlined three strategic goals to expedite the adoption of alternative methods:
  - Help end-users guide the development of new methods.
  - Use efficient and flexible approaches to establish confidence in new methods.
  - Encourage the adoption of new methods by federal agencies and regulated industries.

- Original efforts focused on "validation" lacked the complexity to be used in a variety of studies and NAMs, aiming for a one-size fits all form of guidance.
- Now, there is additional focus of the importance of NAMs to be fit for their intended purposes and how they will be used, rather than to only comply with validation/establishing confidence.
- Regulatory agencies, centers, and offices have overlapped and often worked together, but each follow independent acceptance criteria regarding establishing confidence. In order to validate/establish confidence for regulatory acceptance, NAMs must be specific during their development about which agency a method will be used and must acknowledge it could be used for multiple purposes.
- The "3Cs" were consistent in the exercises held to better understand validation/establishing confidence:
    - Communication
    - Collaboration
    - Commitment
- ICCVAM used results of these exercises to narrow down the needs of agencies, industries, and other stakeholders in NAMs with validation/establishing confidence.
- ICCVAM is introducing new guidance on validation/establishing confidence:
    - The existing guidance is outdated, last updated in March of 1997.
    - The new guidance draft document is expected to be ready in 2023.
- Guidance is currently in a transitional period, where new changes are being made to fit the needs of today.
    - Previous guidance: centralized, one-size fits all, binary status, and stand-alone NAMs.
    - Future guidance: decentralized, fit-for-purpose, evolving confidence, and integrative NAMs.
- ICCVAM's new guidance will retain OECD-34 underlying principles but introduce the "context of use" terminology.
    - New guidance principles will focus on biological relevance, data integrity, technical characterization, and information transparency.
    - Additional topics included in new guidance include the importance of quality reference data, roles of Legacy Animal Data, and a discussion surrounding "Good or Better Standard" in performance evaluation.
    - ICCVAM will ensure that processes are independent and that they can advise, share between agencies, and facilitate collaboration.
- Regarding context-of-use in a regulatory perspective, NAMs should consider which of the two questions are being answered, and for what purpose. Those questions are as follows:
    - Will this "predict" specific potential adverse health effects in humans?
    - Will this identify "no biological effect" levels for human exposures?
- Overall, there is an emphasis in the need for global harmonization, modified language regarding validation/establishing confidence, and assurance that new methods are developed to fit their intended purpose.

John Gordon (CPSC): CPSC NAM Guidance – Virtual

Dr. John Gordon presented on the development of the Consumer Product Safety Commission's guidance document for alternative methods for consumer product testing. Dr. Gordon discussed validity, using methods that were developed to be fit-for-purpose, and the goal of standardizing staff evaluation of alternative toxicological methods. He added that the final guidance document is now available but will continue to evolve.

Summary

- Dr. Gordon presented the background of the CPSC guidance.
- The Federal Hazardous Substances Act requires appropriate labeling on the hazards that may be present but does not require manufacturers to perform any specific toxicological tests to assess those hazards.
- Previously, CPSC had not issued any guidance describing what factors they would consider in evaluating a manufacturer's alternative test method and resulting data that would be submitted in support of a product's labeling.
- He went over who will use the new guidance document, including CPSC staff, manufacturers, test method developers, contract laboratories, ICCVAM, and other stakeholders.
- He also reviewed the purpose of the guidance document, which is to standardize the staff evaluation of alternative tox methods by providing factors staff should consider during technical review.
- He went over the guiding principles for evaluation method and data as well as the technical factors (see slide).

- Dr. Gordon presented an overview of the guidance. It is not mandatory, evaluation of the test methods will remain on a case-by-case basis, it is not a blueprint.
- The guidance document is now available. He added that future plans include updating the web page with guidance documents and any new methods reviewed and approved by the commission. The document will be updated as needed.

Suzanne Fitzpatrick (FDA): Predictive Toxicology Roadmap at FDA – In-person

Dr. Suzanne Fitzpatrick discussed the Food and Drug Administration's goals and timeline for contributing to the development of NAMs, starting with the 2017 release of the Predictive Toxicology Roadmap. Her presentation outlined current and future efforts on the FDA's part to find, develop, and integrate alternative methodologies.

Summary

- FDA Predictive Toxicology Roadmap was announced in December 2017. FDA decided to introduce this after seeing sister agencies publicly sharing guidance on alternative methods. The roadmap focused on six parts to evaluate new methodologies and technologies for their potential to expand FDA's toxicology predictive capabilities and to potentially reduce the use of animal testing.
- Part 1: The FDA formed a senior-level toxicology working group to foster enhanced communication and leverage FDA resources to advance integrating emerging methods and technologies into regulatory safety and risk assessment.
- Part 2: Updated training and continuing education of FDA regulators and researchers on NAMs are important.
- Part 3: Continued communication with stakeholders as part of the regulatory submission process and encourage sponsors to submit valid methods early in the regulatory process.
- Part 4: Collaboration with stakeholders both nationally and internationally are pivotal to identify needs, maintain momentum, and establish a community to support new predictive toxicology methods.
- Part 5: Research to identify data gaps and to identify the most promising technologies.
- Part 6: Oversight by the Office of the Commissioner and reports to the Chief Scientist annually. The goal is to ensure transparency, collaboration, and communication for NAMs development.
- Important to start with context of use regulatory question:
    - What question needs to be answered and for what purpose?
    - How much validation or qualification is needed for a particular assay will depend on the context of use?
    - Context of use helps define the applicability domain and limitations of a NAM, and additional context of use can be added later.
- Alternative Methods Working Group (AMWG) formed in the Office of Chief Scientist, Office of Commissioner to strengthen FDA's commitment to reducing animal testing.
- There is now a website (Advancing Alternative Methods at FDA) to share AMWG objectives, and information on the FDA webinar series on NAMs.
- The webinar series is an opportunity for developers to present new methods and methodologies to FDA. Webinars are held monthly and advertised to FDA scientists exclusively. FDA is not endorsing nor qualifying methods presented in the webinar series.
- FDA created a Tool Development Programs website, which includes the Drug Development Tool Qualification program and the ISTAND Pilot Process.
- There is a need for rapid-return NAMs for consumer protection and satisfaction.
- FDA has proposed a NAMs Program for new funding. This funding will be focused towards regulatory decisions depending on confidence in methods. FDA has stated it is committed to the development and qualification of NAMs.

Maurice Whelan (JRC): Evolution of Validation and Scientific Confidence in Europe – In-person

Dr. Maurice Whelan gave a presentation on the evolution of validation and scientific confidence in Europe, based on his experience with the EU's Chemical Strategy for Sustainability. He spoke on the EU's focus areas – international guidelines, technical standards, and academic studies – and on the use of various international guidelines for peer review and validation techniques. He emphasized the need for better reporting practices and data standardization, and for improved trust and transparency between international partners, scientists, and stakeholders.

Summary

- Dr. Whelan reviewed the EU's Chemical Strategy for Sustainability. He highlighted four aspects that are most relevant to this conference.
    - Promote innovative testing and assessment methods
    - Better assessment of critical effects for more chemicals
    - Internationally recognized standards and tools
    - Make better use of 'academic' data in regulatory processes
- He reviewed JRC's Survey on NAMs, which aimed to explore options to expand REACH information requirements. The survey found that many initiatives have different perspectives, that there are many methods but few solutions, that there is more demonstration than validation, and that there is a lot of variety but little standardization.
- Dr. Whelan then discussed the three focus areas for the EU – international guidelines, technical standards, and academic studies.
- He moved on to discussing the Guideline on Defined Approaches for Skin Sensitization, the first OECD guideline to combine multiple alternative methods in a testing strategy. It is the first time including computational methods in a guideline and includes DAs for both hazard identification and potency-based classification.
- He also touched on the validation of 'omics and machine learning – an independent scientific peer review by ESAC on two SenzaGen GARD methods for skin sensitization testing. The ESAC opinion sets a precedent.
- Dr. Whelan presented a slide on IATA for DNT. (See slide)
- In a recent paper, JRC reviewed the common principles/criteria of different validation frameworks employed by the toxicology community (Patterson et al., Comp Tox, 17, 100144, 2021). Many of these frameworks had common elements such as relevance, reproducibility, transparency, performance, etc.
- Validation and scientific credibility: scientific credibility is the willingness of others to use predictions to inform their decisions.
- He shared two reports to emphasize the importance of trust and transparency. The lack of these is a barrier to validation. He also shared OECD's performance standards for test methods.

Additional points made:

- A conference participant asked how to change people's minds if they do not like the results, Dr. Whelan responded that this is a human process where people want to be convinced and understand before they may accept something new. Understanding concerns and addressing them as scientific questions but also straight forward responses will help with this.

Patience Browne (OECD): OECD Perspectives on the Future of NAMS, Mutual Acceptance of Data, and Test Guidelines – In-person

Dr. Patience Browne (OECD) presented an overview of some of the OECD's activities of NAMs, particularly in the OECD Testing Guidelines and Hazard Assessment Programmes, along with the work in the IATA Case Studies Project to exchange experiences, identify aspects that can be standardized, propose internationally applicable solutions, and increase implantation of NAMs in a regulatory context.

Summary

- OECD Council Act on the Mutual Acceptance of Data (MAD) requires that results from OECD Test Guidelines conducted in a lab following GLP practices is accepted by all OECD member-countries and MAD-adherent countries that have a regulatory requirement for such type of data.
- OECD MAD plays a large role in reducing animal use and preventing duplicative testing.
- Many OECD Test Guidelines, which include NAMs, are currently in use and are revised and updated frequently.
- There are a variety of national and international initiatives to increase the use of NAMs for chemical assessment, and the main drivers for the increased uptake of NAMs are increased chemical throughput, reduced costs, increased relevance to target species (e.g., humans), and changing regulations that reduce or prohibit animal testing.
- Chemical regulations vary with respect to specific data requirements, flexibility, and national/organizational mandates. This creates potential divergence among MAD countries and regulatory authorities because of the variety of NAMs roadmaps, a lack of harmonization between NAMs acceptance, and it poses a potential threat to MAD.
- OECD strives to find aspects of NAM implementation that can be internationally harmonized.

- While the definition of NAMs varies among agencies and organizations, the OECD's working definition of NAMs is intentionally broad and includes:
    - In chemico, in silico, in vitro and in vivo methods designed to fill data gaps;
    - Methods that contribute to faster time to safety decisions and use fewer resources;
    - Methods that are not exclusively "non-animal methods" but are aligned with the 3Rs (Replaced, Reduced, and Refined);
    - Defining methods that are "as good or better" than existing methods, meaning they must be reproducible and relevant to a regulatory application, sensitive to chemical changes, and may provide data for the >80% of chemicals for which chemical safety assessment information may be lacking (i.e., NAMs must be fit-for-purpose).
- In the Hazard Assessment Program, experiences and best approaches and practices for integrating information to come to a regulatory decision are discussed.
- The IATA Case Studies Project has been key to the Hazard Assessment Program's efforts on the implementation of NAMs. More than 35 case studies have been published as of September 2022. These case studies all begin with a specific problem formulation and undergo a stepwise process to determine suitability of the approach. Some IATAs may eventually be proposed as OECD Test Guidelines (and covered by MAD) while others may not be suitable as Test Guidelines, but based on the documentation, reporting of uncertainties of the approach, and independent peer-review may be optionally accepted by regulatory authorities.
- OECD focuses on internationally applicable solutions that can fit different regulatory contexts. Such solutions are likely to be a continuum of acceptable levels of uncertainty depending on the application.
- The first NAMs used in a regulatory context were often pathway-defined NAMs based on AOPs, and while this was required as proof-of-concept, these are not the only option. There may be a variety of pathway-undefined NAMs based on biology used in future chemical assessments.
- IATA is being used to develop templates and standards to harmonize reporting and review, as well as to discuss the criteria of determining states of readiness, acceptable levels of uncertainty based on context of use. The IATA Case Studies Projects involves stakeholders in this process, and continuous engagement with regulators is also necessary.
- OECD focuses on internationally applicable solutions that can fit different regulatory contexts.
- There is a need for data, continued engagement, clusters of case studies, and engagement of regulators and data submitters to get to "the future."

Additional points made:

- There will be a workshop held to identify the evolution of Test Guideline Program in December 2022. One discussion will be revolving around advancing performance standards.

Alison Harrill (EPA): Draft Outline for the EPA Scientific Confidence Framework – In-person

Dr. Alison Harrill presented on the EPA Scientific Confidence Framework, which is currently a draft outline. Dr. Harrill described the timeline for the EPA Scientific Confidence Framework final deliverable and discussed the development of key characteristics for a generalizable scientific framework that will prioritize the use of NAMs.

Summary

- Dr. Harrill presented the goal of the scientific confidence framework. She went over how NAMs are defined – in context, this includes not just hazard, but also dose-response, toxicokinetics, and exposure.
- She discussed the process towards the 2024 deliverable, beginning with the draft outline, then incorporating feedback and reports, creating the draft framework, and then feedback, revisions, and the finalized framework.
- She talked about the initial framing of the confidence framework. The hope for the framework is to be a bit more comprehensive than resources that already exist.
- Dr. Harrill went over the essential elements of the framework:
    - Fit-for-purpose: the NAM should be fit-for-purpose for a specific decision context and the context of use for the NAM should clearly be defined.
    - Transparent: the technology, method, and/or analysis procedure associated with the NAM should be transparently described and sufficiently detailed to enable independent review and evaluation.
    - Reliable: the reliability of the NAM should be characterized, clearly described, and considered within the context of intended use.

  - Relevance: the relevance of the NAM for the intended use should be described to the extent possible.
  - Uncertainty: uncertainties relating to the NAM should be well-described.

Sue Marty (Dow Chemical), Kristie Sullivan (PCRM), Rashmi Joglekar (Earth Justice): Panel Discussion on Validation and Scientific Confidence Frameworks – Virtual

Dr. Annette Guiseppi-Elie and Dr. Krystle Yozzo introduced the panel. The panelists introduced themselves and were given a charge question to respond to. The panelists offered their thoughts on different approaches to NAMs development, the purpose and context of NAMs in current toxicology work, and how science-based criteria can be applied to determine the limitations of NAMs. Overall, the panelists agreed that NAMs should be fit-for-purpose, follow standardized protocols, and contribute to risk prevention for human and environmental health.

Summary

- Dr. Doug Wolf stated that the goal of toxicology and chemical regulation is to prevent risk, not predict adversity. He cautioned against relying entirely on pharmacological approaches for developing environmental frameworks.
- Dr. Wolf stated that while animal studies will still be used for some time, NAMs can help us determine applicability of other tests.
- He stated that NAMs being used only for human safety and risk will not change the paradigm; he believes that ecological risk assessment is equally important.
- Regarding test validity and reliability, Dr. Wolf stated that NAMs do not need to be perfect to be useful.
- Dr. Sue Marty gave a brief presentation on the goals for NAMs use and anchoring NAMs in scientific data.
- NAMs ideally can be used to provide additional information when:
  - The standard test cannot be benchmarked or there is no existing standard test;
  - Animal toxicity results are not benchmarked to humans;
  - There is an absent or insensitive animal model.
- Dr. Marty stated that the ability to cover sufficient biological space to the point that one can feel confident in the decision is important. For example, the DNT battery being used to screen for the potential hazard is currently a challenge, as is anchoring the battery given that the animal model is considered insensitive (especially if the NAMs are based on human cells/tissues).
- Dr. Marty stated that biological relevance is a key element of fitness for purpose.
  - Positive and negative control chemical evaluation is very important, especially if we can define relevant MOAs.
- Kristie Sullivan described the key components of a confidence framework. These included:
  - Flexible standards, classifications, and cut-off values to account for uncertainty.
  - More effective technical evaluations that avoid redundant work and include peer review mechanisms.
  - Weight of evidence approaches.
  - Communication, engagement, and training to ensure that NAMs are implemented appropriately.
- Kristie Sullivan stated that more engagement with biomedical and environmental health communities is needed.
- Dr. Rashmi Joglekar gave a brief presentation on NAMs from the perspective of environmental justice.
- Dr. Joglekar discussed the Toxic Exposure & Health Program and using the new TSCA to protect workers and communities from toxic exposure.
  - Fence-line communities face and/or experience disproportionate harm.
  - Existing animal studies provide sufficient evidence for some endpoints in humans.
  - NAMs validation should be a transparent process.
  - NAMs should generate actionable evidence that can be used to make regulatory decisions.
    - There is a need for a regulatory framework that can use this evidence.
  - NAMs should be used to assess mixture toxicity, human variability, and susceptibility factors.
- The panel discussed various aspects of NAMs implementation and development.
- The panel concluded that:
  - End-user input is important for tool development and uptake.
  - NAMs must add value to regulatory decisions.
  - Confidence in NAMs must be as high as achievable, but it is not always possible to wait for long-term studies when harm is present.

- The three C's (communication, commitment, collaboration) should be kept in mind for groups involved in NAMs work.
- Systemic toxicity is still a new field and requires a paradigm shift to fully grasp.
- Risk managers, regulators, and other participants in the legal framework need to be a part of the conversations surrounding new approach methods.

Day 2 Closing Remarks

Maureen Gwinn (EPA) – In-person

Dr. Maureen Gwinn (Principal Deputy Assistant Administrator for Research and Development) thanked participants for their engagement and noted that great progress has been made since the first State of the Science conference three years ago. She expressed excitement about the work being done within the Office of Research and Development and across the EPA to achieve the goals laid out in the NAMs Work Plan.

Dr. Gwinn recapped some of the topics covered during the conference, including the progress within the agency to better understand the qualitative and quantitative variability in repeat dose animal toxicity data (presented by Dr. Katie Paul Friedman) and the scientific confidence framework (presented by Dr. Alison Harrill). She noted that a great aspect of the conference was bringing in people from around the world and hearing presentations that tied together recent and older research to inform regulatory decision making.

She thanked the panelists, noting that the panel tapped into future directions for addressing cumulative impacts and risk needs. She urged participants to apply the energy brought to this conference to future challenges that will inevitably need to be overcome. Lastly, Dr. Gwinn thanked the colleagues in ORD and OCSVP for coordinating the workshop and added that ORD remains very committed to applying NAMs to decision making and reducing animal use.

Appendix A: Questions and Answers

Conference attendees submitted questions for the speakers at the end of each presentation

Day 1

Variability and Relevance of Traditional Toxicity Tests

Dr. Christoph Helma (In silico Toxicology Gmbh): Variability in Chronic Rodent Bioassays

Questions following the presentation included:

- **Birgit Neumann:** The Variability of *in vivo* could also be caused by dose spacing and not only be the variability of the animal model. How is that captured?
    - **Dr. Helma (In silico Toxicology Gmbh)**: We did not have access to the raw data, so we always use condensed data for our comparisons. Our main objective was not to investigate the bioassays, but to provide a base level for our predictions. For this, we would have to go back to our original data to investigate, so I am sorry, but I cannot answer this question.
- **Xianglu Han:** Does it make sense to study variability in chronic assays based on benchmark doses (BMD)? Can this handle the dose spacing confounding factor?
    - **Dr. Helma (In silico Toxicology Gmbh)**: I am not familiar with the protocols, so I cannot comment on that. This question should be directed at the developers rather than the statisticians.
- **Saeed Alqahtani:** Is the software free and open to use?
    - **Dr. Helma (In silico Toxicology Gmbh)**: Yes, it is free and open to use.
- **Rosalie Elespuru (FDA):** Were the species and strains identical in the comparisons?
    - **Dr. Helma (In silico Toxicology Gmbh)**: Species were identical in both studies (rats and mice for carcinogenicity/ rats for LOAELs). The reproducibility of carcinogenicity studies did not improve when taking strains into account. For LOAELs strains were not reported in our database, but the authors asserted that LOAELs were similar across the three rat strains (although bodyweight and food intake was different).
- **Sue Anne:** Were LOAELs used to model represent different toxicological endpoints?
    - **Dr. Helma (In silico Toxicology Gmbh)**: No, only an overall LOAEL was reported in our database and we did not have access to further detail.

Dr. Thomas Steger-Hartmann (Bayer): Using Big Data to Evaluate the Concordance of Toxicity of Pharmaceuticals in Animals and Humans

Questions following the presentation included:

- **Nicole Kleinstreuer:** Do your plans include incorporating chemical identifiers to be able to link some of the i*n vitro* data that has been generated by screening programs to this large database of preclinical and clinical effects?
    - **Dr. Steger-Hartmann (Bayer)**: Yes, this is implemented. My example of kinase-inhibitors was used because it was easier to show. All the chemical searches are possible and reachable.
- **Dr. Alison Harrill:** I was wondering about interindividual variability and the responses you may see in the clinical population. If we can accept that there is a subset of the clinical population that is going to experience an adverse effect, to what extent should we be considering the effect of 'n' numbers in the animal studies in the ability to observe a clinically relevant effect?
    - **Dr. Steger-Hartmann (Bayer)**: This is a given. We are aware of the limitations of the statistics of the power of our studies, and this is one of the reasons that the negative prediction is short, because the numbers are simply too small. By beefing up the control groups and increasing the control number of animals, we get a higher sensitivity.
- **Martin Phillips:** Presumably, chemicals with strong adverse effects in the pre-clinical trials do not move on to clinical trials. Have you thought about how that might affect the interpretation of the analysis?

    - **Dr. Steger-Hartmann (Bayer)**: In the original numbers we were only looking at those who progressed. In the second try we attempted to leverage the clinical data from various projects which failed. The compilation of the sharing of the data is not trivial for legal reasons, so we are not as far as we wanted to be. It is a valid point and a lot of work to be done.
- **George Hinkal:** Can this data be translated from drug development into predictive industrial chemical safety? With mixtures?
    - **Dr. Steger-Hartmann (Bayer)**: No. One would need to think about defining the LOAELS and see whether you can derive any conclusion when you add the LOAELS. For the moment, I cannot imagine how that data would show this.
- **Darshan Mehta:** Apparently, porcine skin data is known to more accurately reflect the findings in humans. I am wondering if skin data for pigs was evaluated in this analysis?
    - **Dr. Steger-Hartmann (Bayer)**: Pig skin data was not looked at. The coverage would be too small.
- **Manjeet Singh:** Big data can be helpful and problematic-Curse of dimensionality. Data quality is also equally important. Do you think Findable, Accessible, Interoperable and Reusable (FAIR) data standards like Elixir-EU can provide better insights in AI, ML and traditional drug development and chemicals risk assessment.
    - **Dr. Steger-Hartmann (Bayer)**: FAIRification of preclinical data is a prerequisite for translational analysis. Its implementation will indeed contribute to a better risk assessment.

Dr. Alan Boobis (Imperial College): Conservation of Pharmacodynamic and Pharmacokinetic Modes-of-Action in Rodents and Humans

Questions following the presentation included:

- **Shadia Catalano:** How do you see the application of those now to the in vitro space where we are?
    - **Dr. Boobis (Imperial College)**: I think our knowledge on AOPs and MOAs is being used to design intelligent NAMs. These need to be anchored. And we need to ask ourselves how much we need to preserve a key event to propagate the AOPs. That is where legacy data needs to be harvested, to help inform us about the biological significance of the changes we are seeing. We can compare the sensitivity of human targets to animal targets and skew the responses likely in humans without having to measure it apically in humans.
- **Saeed Alqahtani:** I believe this presentation highlights the importance of *in vivo* evaluation until we can develop a better prediction model. Is this correct?
    - **Dr. Boobis (Imperial College)**: Yes and no. I think there are some very bright people that are developing NAMs and are thinking more carefully about what a prediction model needs to be to use these NAMs meaningfully. The roadmaps presented show that there needs to be an effective way to translate that into quantitative human relevance. I don't think we should be replacing animals for the sake of replacing animals, we need to put in our efforts to do this properly.
- **Manjeet Singh:** In addition to differences between phase 1&2 enzymes, what are the effects of other factors like food effects, chronobiology, liver, and renal impairments in actual human exposure and changes in TK-TD profiles of chemicals?
    - **Dr. Boobis (Imperial College)**: Some of this thinking underlies the PK predictive platform, where what they have tried to do is build knowledge of the variability and factors that vary the expression activity of the enzymes and transporters involved in chemical disposition. We understand how liver disease affects the expression of the enzymes of drug metabolisms, and we can make predictions about the impact of liver disease of the disposition of our chemicals. The scientific community also looks at how the food matrix affects absorption and disposition of compounds.
- **Sayak:** How well do these conserved modalities generalize for different chemicals? You made the case for acetaminophen, but can that claim be generalized to other chemicals?
    - **Dr. Boobis (Imperial College)**: It will depend upon the MOA itself. If we think about something that causes neurotoxicity by affecting cholinesterase, then any chemical that can inhibit cholinesterase will cause neurotoxicity at a high enough dose in both animals and humans. For some of the chemicals it will be case by case, but as we understand more about the underlying biology, we will be able to make more generic predictions.
- **Sue Marty (Dow Chemical):** If interested in how much perturbation in a KE drives an adverse effect, does this mean that KEs closer to the adverse outcome will be better for predictions vs. MIEs?

- Additional research would be needed to fully answer this question. There are several opinions.

Dr. Katie Paul Friedman (U.S. EPA): Qualitative and Quantitative Variability of Repeat Dose Animal Toxicity Studies

Questions following the presentation included:

- **James Stevens:** Given complex bio systems and sensitivity, how much variability should be expected in complex studies? I would assume it is a lot in models, and a lot in humans. What is the objective?
  - **Dr. Paul Friedman (U.S. EPA)**: We will never know the true variability in animals, though there are experiments that could be conducted to better characterize interindividual variability in animals. We are limited by available data and by the structures of data/experimental design, so we are not looking at populational variability and microbiome, for instance. Based on our work, you cannot capture more than 55-73% using the curated meta-data, or study descriptors, that we have available based on how these studies are conducted and reported.
  - **James Stevens**: How much variability should we accept if that is the reality of working in a complex bio system?
  - **Dr. Paul Friedman (U.S. EPA)**: Depends on where you are in the adverse outcome pathway framework in terms of biological complexity. You may be able to make measurements in a particular biological model system with much less variability than the data we have, which is largely of apical endpoints such as body weight changes or organ histopathology, but it depends on the endpoint being measured.
- **Frank Barile:** "False positive" and "false negative" used to be known as Specificity and Sensitivity. Are these terms still applicable and how do the concepts work into the variance analysis?
  - The terms false positive and false negative are useful when one can classify responses for a reference data set. In the work reported in this presentation, we focused largely on quantitative variability, which is a description of the spread of replicate values rather than a classification exercise. For qualitative reproducibility, we looked at concordance and did not require any *a priori* knowledge of whether a chemical was a “true positive” or “true negative.” Curation steps would need to be taken to try to identify these “true positives” and “true negatives,” and then these results would need to be evaluated. In terms of sensitivity and specificity, these terms quantify how often the classification exercise is successful (i.e., true positive rate and true negative rate), which was not the focus of the work presented.
- **Julija Filipovska:** You can explain about 50% of variability by differences in descriptors of the study. Is the rest just inherent stochastic variability of measurement or further analysis may contribute to mechanistic understanding?
  - Unfortunately, it is currently unknown how much of the unexplained variance is due to systematic measurement error, e.g., measurement error on endpoints such as body weight and organ histopathology, or inherent biological variability of the animals in the study designs typically conducted for the 870 Series Health Effect Guidelines. The statistical models herein attempted to approximate the total variance, the unexplained variance, and then the spread of LEL/LOAEL/BMD values observed for replicate studies.
- **Manjeet Singh:** Is animal tox data in your study from GLP studies? Did you compare between GLP VS Non-GLP data? What if you used BMD as POD instead of NOAEL. Will that make any difference? Any gender or strain or specific effects especially in rats VS dogs.
  - Many but not all of the studies in ToxRefDB v2.0 are GLP-compliant. We do not currently have enough replicate data of GLP-compliant and non-GLP to perform an analysis to compare these studies. However, GLP is not the only determinant of study quality; for instance, some of the curated “non-GLP” studies were conducted at the National Toxicology Program. In our current unpublished work not included in this presentation, we have tried using BMDs instead of LEL or LOAEL values, and this failed to reduce the estimate of variance (RMSE) significantly, such that an estimate of 0.5 log10-mg/kg/day remains a reasonable estimate of variance for BMDs or LELs or LOAELs at the study-level (unpublished, Paul-Friedman et al., in preparation). For more details on sex and

species/strain differences in study level POD variance, I would direct the commenter to Pham et al. 2020. Species/strain do account for some of the explained variance in the statistical model. For dog studies, there are an insufficient number of strains to perform an analysis to compare differences in strain, as the available dog studies are typically annotated to a single strain (Beagle).

Dr. Chad Blystone (Division of Translational Toxicology): Inter-Species Concordance of Toxicological Endpoints

Questions following the presentation included:

- **Gina Hilton (PETA):** It looks like there is a good handle on variability within bioassays. Where is the group looking at interspecies connection to humans? Wondering if there is thinking about environmental exposures and epidemiology data in this framework for human protection.
    - **Dr. Blystone (Division of Translational Toxicology):** There have been previous publications but not evaluated recently. Trying to determine if there is a human carcinogenic response is very difficult compared to animal models. Short answer, it has not been evaluated recently.
- **Zhongyu Yan (Agriscience):** If the rat LOAEL is lower than the mouse LOAEL, does that show that the rat is protective?
    - **Dr. Blystone (Division of Translational Toxicology):** I do not think we can say that. It depends on the tissue. If we did not care as much about mouse liver implications, we would use that a lot, but from a quantitative standpoint I guess there is no answer for that.

Dr. Tom Monticello (Amgen): Concordance of the Toxicity of Pharmaceuticals in Animals and Humans: Lessons from the DruSafe Consortium

Questions following the presentation included:

- **Martin Phillips (U.S. EPA)**: When you are looking at these matrices with different rates, you are discarding the dose response (DR) data from clinical side. How do you take the DR into account? Do you compare preclinical and clinical dose effect?
    - **Dr. Monticello (Amgen):** The way it is entered into the informational brochure is based on target organ effects at dose. We did not really build into this a dose response because all we care about is what exposure related to the adverse effect in animals, and how does that relate to adverse effect in humans.
- **William Irwin:** Were "fallen angel" compounds included in the analysis which had considerable human toxicity and dropped out of the pipeline, or were those compounds excluded due to legal implications?
    - **Dr. Monticello (Amgen):** No. We might have multi-series of compounds being run through assays and exploratory animal tox studies. If it is causing liver damage that might happen in clinic, the chemical does not move on to the clinic because it is excluded/not nominated to go to clinic. Info brochure is only chemicals that have been nominated and are “best bets.” GLP toxicology studies, an accidental kill would never make it to the database
- **William Mattes:** If you look only at lower doses could the NPV at those levels predict a safe dose?
    - **Dr. Monticello (Amgen):** Yes, probably. A lot of times the higher dose pharmacology scheme is set up for false positives. We want to promote target organ toxicity at extreme doses to ensure high sensitivity of the method for detecting target toxicity.
- **Scott Auerbach:** If the data set were reduced to just oncology drugs where the doses in the tox studies more closely approximate humans’ doses would the PPV increase?
    - Further research is needed to answer this question.
- **Unknown:** The three percent additional toxicities identified in long-term studies, were some of them serious or life-threatening?
    - **Dr. Monticello (Amgen):** That type of inquiry of the database is still ongoing.

Day 2

Dr. Tala Henry (U.S. EPA): Report out from Discussion Groups

Questions following the presentation included:

- **James Stevens:** One theme from yesterday: if you use two different species and they disagree, that may mean we should trust the data less. So, by analogy, if you look at genetic toxicology tests, they will not always agree, but in combination, they catch different mechanisms of genetic aberration potential. If you consider the two different species to be two separate assays, I would argue that combining them adds power, we will always need a battery of assays. The fact that the assays disagree, it does not necessarily mean the assays are bad, it just means they are detecting different things. I am not arguing about replacing the animals but suggesting that because they disagree is not necessarily an adequate justification.
    - **Dr. Paul Friedman (U.S. EPA)**: At least from my perspective, looking at whether a species agrees with another is aimed at understanding a benchmark of what NAMs would need to predict. How do I benchmark the NAM that I am trying to use to replace those two studies with if those two studies disagree (i.e., which study should be predicted with some amount of accuracy)? It is more about evaluating uncertainty in existing methods to see what the NAMs would need to tell me and with what level of certainty.
- **Xianglu Han (LANXESS Corporation):** Can someone explain why we are studying in vivo study variability in the process of evaluating *in vitro* assays? We will need to consider human population variability for human health risk assessment purposes, so unless variability from animal studies is strongly related to human variability for chemicals, I do not see a clear need for assessing animal study variability. Admittedly, for some other purposes, we do want to study variability among animal studies.
    - **Dr. Henry (U.S. EPA):** I think the point around assessing the variability around animal studies is to try and say whether NAMs are comparable and comparable in what way. NAMs are not necessarily superior to *in vivo* animal studies, but it has not been clearly put out there.
- **Dr. Rusty Thomas (U.S. EPA):** It is required by the language in TSCA that if we are replacing an *in vivo* study with an alternative, that alternative has to be equivalent or better than the existing tests. Biological relevance, variability associated with existing models — we have to demonstrate NAMs are as good or better. Part of this evaluation is to ask how well NAMs must perform.
    - **Dr. Henry (U.S. EPA):** He is referring to TSCA 2016 amendments which mentioned the standard to replace a scientific test is another test that is of equal or better quality. The words in the regulatory context set a bar. A part of these studies [of *in vivo* variability] is establishing how well animal studies reproduce themselves as a benchmark on the accuracy that can be expected from NAM approaches built to replace these animal studies.

Evolution of Validation and Scientific Confidence Frameworks to Incorporate 21st Century Science

Dr. Warren Casey (NIEHS): ICCVAM Strategic Roadmap for Validating New Methods

Questions following the presentation included:

- **Dr. Rusty Thomas (U.S. EPA):** Can you explain on the scope of the framework? How wide are you casting the net?
    - **Dr. Casey (NIEHS):** FDA has some guidance on how to evaluate analytical methods. It is intended to be very broad. There are a few deep dives like that, but it is intended to be broadly applicable.
- **Dr. Maureen Gwinn (U.S. EPA):** The way it was worded on the slide (vs slide) the way it was framed seems like only high dose? Or would it be both?
    - **Dr. Casey (NIEHS):** I was commenting on the requirement that we have to push up doses until we find a sensitive organ. Is it better to find a safe dose regardless of what would happen if you went to unrealistic exposure levels?
- **Emma Grange (Cruelty Free International):** With respect to the "gold standard" set by animal test-based regulatory information requirements; can we agree that this standard is, and always has been, arbitrary (in fact the term "gold standard" likely refers to the use of gold to benchmark the value of currency, not any intrinsic value) – it really is important to scrutinize just what the animal tests offer.

    - Comment from participant.
- **William Mattes:** Couldn't a high dose study guide what NAM is used to assess dose-response?
    - **Dr. Casey (NIEHS):** I do not know, this is one of the things we need to start looking out for.
- **Jeffrey Morgan:** What are the top 5-10 NAMs and what are the attributes that have made them successful?
    - The top 5-10 attributes could be dependent on the intended purpose or use of a NAM.

Dr. John Gordon (CPSC): CPSC NAM Guidance

Questions following the presentation included:

- **Katie Groff:** I just had a question about implementation of the document. Are there plans to track increased NAMS use or decreased animal use based on the implementation of this guidance document?
    - **Dr. Gordon (CPSC):** Yes, we track the use of NAMs as part of our CPSC metrics. We will track numbers and which assays are used for FHSA labeling. We cannot track animal use because we do not do the animal testing or require that testing, so we have no way to get that from testing bodies.
- **Charles Kovatch:** Since being published, what has CPSC seen to be working well in the document's implementation?
    - **Dr. Gordon (CPSC):** The guidance has a nomination form at the end of it; it helps organize the information. It helps with the evaluation process. That is the best thing we have seen from the document so far, is getting nominations from people about which methods they want to see CPSC use.
- **Dr. Rashmi Joglekar:** Is making this process transparent and accessible to the public something you plan to do?
    - **Dr. Gordon (CPSC):** Yes, it should all be online. All the reviews and briefing package/packet will be on the website for the public to look at.
- **Dr. Annette Guiseppi-Elie:** Do you have anything that you would like to share about validation frameworks ahead of the panel?
    - **Dr. Gordon (CPSC):** We know that validation and confidence are paramount. We have been working on the skin sensitization. We have found a nice way to test statistical uncertainty, with a statistical model within that plate. We were able to use custom positive calls. That also helps us to have more confidence in that system. There are a lot of components to assessing confidence.

Dr. Suzanne Fitzpatrick (FDA): Predictive Toxicology Roadmap at FDA

Questions following the presentation included:

- **Dr. Alison Harrill (U.S. EPA):** Since you have qualification programs, have you collected data on the length of time it takes to get a NAM qualified?
    - **Dr. Fitzpatrick (FDA):** As we start working on food, that would be a good idea. The more you put out guidance, those will help us too. We are thinking about more guidance.
- **Jim Stevens:** You touched on the use in non-regulatory space. What can FDA do to help provide incentive for industry to share their experience? Is there something FDA can do for inceptive?

Dr. Maurice Whelan (JRC): Evolution of Validation and Scientific Confidence in Europe

Questions following the presentation included:

- **Dr. Rusty Thomas (U.S. EPA):** You said that you learned that when somebody does not necessarily like some parts of data, they will dismiss it. So, what did you learn about how to overcome that?
    - **Dr. Whelan (JRC):** I heard that from my discussion group yesterday. We did not go into the process as naïve, but my sense of what I was hearing, was “such member country wants this analysis done” and people arguing over applicability domains and having to go back and do redundant work… So what I learned is that we must be real about it. It is an actual human process. People take time and want to be convinced and understand. We must be smart in understanding that sometimes people’s concerns are being expressed as a scientific question, but it is not their real concern, there is

something else they are not articulating. It is a social-technical process. That would be a tricky guidance document to write. What is often missing from our guidance is how do you create the socialization of that process, and there is no easy answer there. Being aware is half the problem.

- **Athena Keene**: How exactly are you going about international acceptance of data?
    - **Dr. Whelan (JRC):** The next speaker is going to talk about mutual acceptance of data. Patience is going to answer your question.

Dr. Patience Browne (OECD): OECD Perspectives on the Future of NAMS, Mutual Acceptance of Data, and Test Guidelines

Questions following the presentation included:

- **Yad Bhuller:** Is the OECD considering the Joint Review process for agrochemicals as a mechanism to build science, regulatory, and public confidence for NAMs? This is opposed to industry sending independent national submissions.
    - **Dr. Browne (OECD):** Yes, if supported by member countries. There is currently interest among some countries in Joint Reviews of Minor Use pesticides and some biocides. The hope is that these can be used as case studies to evaluate additional opportunities for shared chemical assessments.

Dr. Alison Harrill (EPA): Draft Outline for the EPA Scientific Confidence Framework

Questions following the presentation included:

- **Emma Grange (Cruelty Free International):** does the EPA agree with the OECD definition of NAM, i.e., that it could include *in vivo* tests?
    - The current definition in the NAMs Work Plan does not include in vivo tests. However, EPA's Office of Research and Development views novel *in vivo* systems in alternate species as a key component to data collection. Exploration of a NAMs definition change to include in vivo testing would have to be done under consultation with EPA's regulatory offices.
- **Piper Hunt (FDA):** Flies, worms, and zebrafish are animals. Does the EPA consider alternative small animal model assays to be NAMs?
    - These are not currently covered under the definition of a NAM, but alternative species are an important component of toxicology research efforts both within the Agency and beyond.
- **Craig Rowlands:** Currently, there is no standardized approach to validating *in silico* models for predictions. However, these *in silico* approaches are used for risk decisions e.g., TSCA PMNs and Risk Evaluations of existing chemicals. The validation of these *in silico* models varies significantly, usually a peer review is mentioned, and sometimes there is no record of a validation, but these models are still used for regulatory decisions. Will a standardized approach be developed by EPA and other agencies to validate *in silico* models?
    - This is certainly of interest. Note that some types of in silico approaches do have guidance around them, an example is OECD's "Guidance Document on the Validation of (Q)SAR Models".
- **Darshan Mehta:** Within this framework, which core principle does "variability" belong to?
    - This is a difficult question to answer, because "variability" can have different meanings and considerations depending on the context of the question. This response focuses on a couple of common considerations around variability. Within the framework, biological variability should be considered in at least two of the core principles. Under the core principle of "Fit for Purpose," the decision context must be clearly defined – this could involve considerations for a specific life stage (e.g., developmental or aging populations) or molecular sequence variation (e.g., sequence homology for a receptor binding pocket). Under the core principle of "Relevance," mechanistic interpretability of the NAM is described and could include considerations of biological variability.
    - With regard to quantitative or qualitative variability in the data derived from a NAM, considerations would be included in the core principles of "Uncertainty" and "Reliability." These two domains involve considerations around describing a lack of data, applicability domain, or an incomplete understanding of NAM components (Uncertainty) or measurements of the extent to which results of

using a test method for a defined purpose can be reproduced within and between laboratories over time when performed using the same protocol (Reliability).

- **Shaun McCullough:** How would in vitro NAM reproducibility be considered/evaluated? Would it be at the qualitative level (e.g., two or more groups both get a "positive" result from a positive control) or is there a quantitative threshold for a statistical measure of variation?
    - Reliability and reproducibility of the NAM could have both qualitative and quantitative considerations and should consider both the chemical- and/or endpoint-specific domains of applicability.
- **Roper Clive:** Should we be balancing the uncertainty of the test we are replacing as well in order to draw out the advantages of the new as well as the uncertainties of the new?
    - There would likely be advantages to understanding and quantitating uncertainties around the test to be replaced as well as the NAM to put uncertainties around the NAM's output into context.
- **Andrew White:** in defining the uncertainty is that always bounded by what was said about as good as or better than current regulatory tests?
    - In some cases, this would be helpful, but not all NAMs are designed to replace an existing test and are instead created to meet an unmet need. We are interested in pursuing discussions around contextualizing uncertainty in such cases.

Panel Discussion

Questions following the presentation included:

- Is the EPA focusing on non-vertebrate NAM tests (as stated in TSCA) or non-animal tests?
    - The main focus is non-vertebrate as outlined in the EPA NAM Work Plan.
- **Gina Lento:**
    - As Michael Whelan stated, a lot of innovation is happening in small companies. How can the agencies support this kind of work and what is the process by which a small company can approach the agencies with a NAM?
        - We encourage them to visit EPA's Small Business Innovation Program website to learn about opportunities for working with EPA - https://www.epa.gov/sbir
    - Suzie Fitzpatrick mentioned their open invitation to present to a webinar series at the FDA, but what about the EPA and OECD?
        - EPA and OECD have numerous webinar series. Learn more about existing webinar series here: https://www.epa.gov/chemical-research/computational-toxicology-communities-practice https://www.oecd.org/education/career-readiness/conferences-webinars
    - It appears that there are two schools of thought around the suitability or relevance of a NAM. One is that a NAM is only useful if one can elucidate knowledge of a pathway or mode of action from its readouts. The other is that a NAM is useful if it can simply give sound indication of likely hazard potential. What views do the panelists have on whether these perspectives are mutually exclusive? Please contact the panelists if interested in their perspectives beyond what was presented at the conference.
- **Jeffrey Morgan:** Much of the past 2 days have been concerned with high level thinking about defining criteria of NAMs which is very valuable. Something that would be very valuable to NAM developers would be to drill down on specific NAMs. The kind of info that would be valuable would be what are the top 5-10 NAMs, what are the attributes that have made them successful, how do they stake up to the criteria that are being proposed? What are their strengths and weaknesses, etc.
- **Wendy Wang:** How will NAMs change/affect regulatory data requirement on agrochemicals? Thanks!
- **Craig Rowlands (SACC):** How will NAMs or what NAMs will be able to model not just dose, but exposure duration, e.g., sub-acute, sub-chronic, chronic etc., and especially DART that at least some have proposed there are at least 18 etiologies documented for chemical induced DART adverse effects so far.