Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Toxicology

Estimating uncertainty in the context of new approach methodologies for potential use in chemical safety evaluation

Ly Ly Pham[1,2], Thomas Y. Sheffield[1,2], Prachi Pradeep[1,2], Jason Brown[2,3], Derik E. Haggard[1,2], John Wambaugh[2], Richard S. Judson[2] and Katie Paul Friedman[2]

Abstract

Acceptance of new approach methodologies (NAMs) for use in characterizing chemical hazard and risk requires informed expectations regarding the minimum precision and maximum accuracy of their results. Uncertainty in NAMs derived from variability in the traditional reference data used to train or validate performance of the NAM, and uncertainty in the modeling procedures themselves, limits NAM performance. Herein, we review current approaches to characterizing uncertainty in NAMs. We discuss variability in *in vivo* data used as a reference for NAM development and validation; the quantitative uncertainty in concentration–response modeling for high-throughput *in vitro* bioactivity screening; the uncertainties associated with *in vitro* to *in vivo* extrapolation using toxicokinetic information; and the quantitative uncertainty in the experimental inputs and modeled outputs from quantitative structure activity relationship models for prediction of point of departure doses. Communication of the amount of uncertainty, both in the input and output for NAMs, often involves a confidence or prediction interval around a given potency estimate, derived from an understanding of the variability in the data modeled. Tuning expectations of NAM performance to an understanding of the reproducibility and variability, both of traditional approaches and NAM approaches, provides a path for the adoption of NAMs as alternatives in screening chemicals for risk.

Addresses

[1] ORISE Postdoctoral Research Participant, USA

[2] National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

[3] ORAU, Contractor to U.S. Environmental Protection Agency Through the National Student Services Contract, USA

Corresponding author: Judson, Richard (judson.richard@epa.gov)

Current Opinion in Toxicology 2019, **15**:40–47

Edited by **Anne Marie Vinggaard** and **Richard Judson**

This review comes from a themed issue on **Risk Assessment in Toxicology**

Available online 30 April 2019

For a complete overview see the Issue and the Editorial

https://doi.org/10.1016/j.cotox.2019.04.001

2468-2020/Published by Elsevier B.V.

Keywords

Uncertainty, New approach methodologies, *In vitro*, Chemical safety.

Introduction

In toxicology and chemical safety evaluation, the use of new approach methodologies (NAMs) is increasing because of recognition that traditional animal tests can never be used to assess all chemicals in commerce, a push to eliminate the use of animals wherever possible on animal welfare grounds, as well as concerns regarding how well animal models predict human toxicity [4,15,36,84]. An NAM is any one of the several methods that yield a prediction of the key outcomes currently used from animal studies: a point of departure (POD) or estimate of potency (a no observable adverse effect level [NOAEL] or lowest observable adverse effect level [LOAEL]), and the specific types of effects that may be caused by a chemical (e.g., liver hypertrophy, cholinesterase inhibition). NAMs to predict hazard, for example, *in vitro* assays, can be combined with *in vitro* to *in vivo* extrapolation (IVIVE), to predict *in vivo* PODs. Another main category of NAMs involves quantitative structure activity relationships (QSARs) or read-across methods that are purely computational and use only chemical structure and computed properties as model inputs to predict hazard. NAMs, being models, are always imperfect, and therefore, users should understand the uncertainties associated with NAM predictions. One can break uncertainty into two pieces: uncertainty in the input to the models, arising from the animal data used to develop and evaluate them, and uncertainty in the outputs of the model, arising from the approximations made in the modeling approaches themselves. The first section of this article focuses on our analysis of uncertainty in the animal data, and the later sections focus on uncertainty in *in vitro* data used in IVIVE models and uncertainty in QSAR models of toxicity end points. Finally, we conclude by assessing future directions for our research to characterize the sources and magnitude of uncertainty for NAMs.

Uncertainty in animal toxicity studies

Often the perceived gold standard for validation of NAMs in chemical hazard prediction is comparison to animal toxicity studies [37,38]. Beyond consideration of how well animal studies may or may not predict human toxicity, benchmarking to animal data is practically challenging because of study reproducibility limitations, which include (1) qualitative uncertainty in the observable effect and (2) quantitative uncertainty in the dose at which an effect is observed, that is, the variability in the POD. There is a need for a refined understanding of these uncertainties as the inherent variability of animal toxicity studies informs the upper limit on predictive performance for any NAM trained using these reference data. In this section, we review current perspectives on evaluating qualitative and quantitative uncertainty in *in vivo* data and demonstrate the practical use of this quantitative uncertainty for setting expectations of NAM performance.

Qualitative assessments such as concordance analysis often binarize the observed effects for easy comparison across studies [6,27,37]. These concordance analyses typically focus on whether the specific effects observed are consistent between two studies with the same design or between two studies with the same design but using different species. A recent and well-studied example is the local lymph node assay (LLNA) for prediction of skin sensitization: even within species and using the same vehicle solvent, repeat LLNA studies to determine the positive or negative sensitization effects for 87 chemicals were concordant only 78% of the time, with a 35% chance that a 'negative' chemical would test 'positive' if the LLNA was repeated [14,27]. This highlights an important issue for validation of NAMs: they cannot predict the reference data with higher accuracy than the rate of concordance in the reference data set [7]. The LLNA example also provides a demonstration of the importance for interspecies extrapolation as the LLNA results in mice predicted human sensitization with approximately 68% balanced accuracy [38]. Indeed, the positive predictive values among species, depending on the end point measured, may range considerably. An evaluation of 37 National Toxicology Program repeat dose toxicity studies demonstrated 0–100% concordance, with a median of approximately 70%, in the noncarcinogenic effects observed between rats and mice, depending on the biological end point or tissue measured [73]. Concordance among rat and mouse models of carcinogenicity has been shown to range from 57% [22] to 76% [21,25], depending on the data set and analysis. Concordance in findings of anemia ranged from 40 to 90% depending on species and study type being compared [33]. These findings corroborate ongoing evaluation of reproducibility of effects within study data in ToxRefDB [40,46], a large database of detailed *in vivo* animal toxicity data. The concordance in observed effects between repeated studies within one species varies from 57 to 87%, depending on the species and the effect ([53]., in prep). Variability in the qualitative concordance among animal reference data may not be fully explained by parameters such as the species used, thereby limiting NAM predictive performance if evaluated using classification accuracy.

Quantitative evaluation of animal toxicity studies can also address two primary questions: (1) for a chemical of interest, what is the range of 'true' POD values that might be expected and (2) what is the maximal predictive accuracy of an NAM, that is, the amount of variance that the NAM can explain in the reference data set? Answering these questions requires a measure of variance, which is often unavailable as it requires availability of a database with replication of study information. Previous efforts have been made to quantify the variance in effect levels of interest from animal studies, for example, standard deviations around the doses that causes 50% lethality (LD50) in rodents (0.68–0.84 $\log_{10}$-mg/kg/day) [26] or derivation of a 95% confidence interval on the prediction of an NOAEL from chronic studies in rats ($\pm$0.64 $\log_{10}$-mg/kg/day) [47]. The variance in discrete NOAEL or LOAEL values is highly subject to the shape of the dose–response curve, the sample size, and the doses selected [43,65], and consideration of the variance among replicate studies might be reduced if modeled values of responses, for example, benchmark dose (BMD) values, could be used instead of these POD values. The second question, aimed at defining the maximal predictive accuracy of an NAM, given some reference data, may be approximated using estimates of variance in the reference data to define a reasonable prediction interval, for example, assuming a normal distribution, a reasonable prediction interval at the 95% confidence level for a given POD estimate might be approximately $\pm$1.96 times the variance estimate.

Ongoing work (Pham et al. in prep) aims to describe the variance in POD estimates for systemic toxicity effects in ToxRefDB, version 2.0 [45,46,75], and from that approximation of variance, to define the upper limit of predictive performance for an NAM to predict a systemic toxicity POD. Work by Pham et al. uses parameterization approaches on the basis of multilinear regression and cell means models approximate the total variance in systemic POD estimates, LOAEL or lowest effect level values, from replicate studies, and then to identify the amount of variance accounted for by study design descriptors such as the study type or the number of doses. In both models, the variance left unexplained by study design descriptors is estimated by the mean square error (MSE) or the average squared difference between the estimate and the reference value. In modeling applications, the MSE limits the maximal coefficient of determination, or R^2, which is the proportion of the variance in the prediction that can be

accounted for using the descriptors available. Utilization of NAMs as replacements for animal studies necessitates a reasonable expectation of NAM performance, that is, the amount of variance explained by the NAM cannot exceed the explained variance in the training or validation data set. The general relationship between R^2 and MSE is given by Equation (1).

$$R^2 = \frac{Total\ \ Variance - MSE}{Total\ \ Variance} \quad (1)$$

In preliminary findings from the study by Pham et al. [53], depending on the data subset and modeling method used to approximate the variance, only approximately 60–80% of the total variance could be explained by study descriptors in ToxRefDB. Consequently, the R^2 on an NAM to predict these systemic POD values would likely be bounded at approximately 60–80%. The root-mean-square error, a variance estimate in $\log_{10}$-mg/kg/day units just like the animal data used, ranged from approximately 0.4 to 0.6 $\log_{10}$-mg/kg/day, similar to previous findings on the quantitative variability in rat chronic LOAEL data [47]. Work to quantify variability in large *in vivo* POD data sets establishes what the impact of variance might mean for understanding good performance of NAMs to predict a POD and demonstrates practically what toxicologists have long understood: running a study more than once may result in different quantitative responses. Modeling uncertainty in *in vivo* studies has a long history [10,18,28,42,59,64], and the major contribution to that literature from the currently described work is the use of a large database that includes thousands of studies across hundreds of chemicals. In current chemical safety evaluation practice, an important tool for managing uncertainty is BMD modeling [10,18,28,59,64]. This type of analysis informs a POD and explicitly accounts for variability within a study derived from interindividual differences in treatment groups. The application of expertly selected uncertainty factors in current chemical safety evaluation may also account for study-to-study variability [20,62]. Both of these sources of variability, that is, intrastudy and interstudy variability, contribute to limitations on NAM-based predictions of POD.

Uncertainty in high-throughput bioactivity data

In vitro data can be used to predict both the potency of a chemical, that is, the dose where an effect may occur and to predict the types of effects that may be observed. The basic approach to potency prediction is to measure the active concentration *in vitro* for a chemical using a range of cell types and assays, and to select the appropriate minimum concentration causing any biological effect. This concentration value is then put into a dose context using a toxicokinetic (TK) model that produces the 'administered equivalent dose' with respect to the *in vitro* potency [31,57,71,78–80], a process known as *in vitro* to *in vivo* extrapolation (IVIVE). This section focuses on uncertainty in quantitative potency estimates, and we only mention that qualitative effect prediction from *in vitro* assays is also an active area of research, often using the Adverse Outcome Pathway framework that links molecular or cellular key events which are measurable *in vitro*, to apical adverse outcomes seen at the whole organism level [1,39,52,67,68]. The uncertainty in separating positive responses from assay interference or screening artifacts, for example, autofluorescence or cytotoxicity, is a separate process from quantitative uncertainty in curve-fitting [17,30] and typically requires knowledge of the given assay as well as other information that could be used to filter concentration–response data. Here, quantitative uncertainty in curve-fitting *in vitro* bioactivity data and the potential for IVIVE uncertainty from the high-throughput TK information and modeling are considered.

The US EPA ToxCast program generates high-throughput screening (HTS) bioactivity data for use in various predictive toxicology applications [32,35]. The ToxCast data pipeline (tcpl) [17] has been used to normalize and curve-fit data for nearly 1400 assay end points, thus enabling first-tier data processing of heterogeneous bioactivity data from HTS. For each chemical sample:assay end point pair with at least 4 concentrations available, the tcpl curve-fitting procedure attempts to fit a Hill, a gain-loss, and a constant model, with the model selection based on a maximum likelihood estimate [17]. The tcpl analysis generates concentration–response parameters on the basis of the winning model, including the 50% activity concentration (AC50), the potency estimate. There are multiple sources of potential variability in these AC50s, resulting from biological variance, experimental error, or curve-fitting procedures. One method for quantifying uncertainty in ToxCast curve-fitting is called toxboot, an R package extension of tcpl that implements smooth nonparametric bootstrapping (a statistical method that uses resampling and added noise to determine uncertainty in a series) [76,77]. This addition of random, normally distributed noise to the series allows one to be more confident in the winning model if similar models are produced in each iteration. By resampling and adding normally distributed noise over many iterations, a general picture of the confidence in a curve fit can be ascertained. Two examples of the quantitative measures of uncertainty produced by toxboot are hit percent and an AC50 confidence interval. Hit percent is the number of active hit-calls of the total number of resamples, currently with 1000 resamples per curve. This is potentially informative because the binary hit-call for a borderline response, positive or negative, is particularly susceptible to minor fluctuations in the data, especially

for weak or borderline responses. With a more continuous statistic (ranging from 0 to 1), such as hit-call percent, the borderline hit-calls may be more obvious; currently, roughly 61% of positive hit-calls (hitcall equals 1) correspond to a hit percent of 100 in the latest ToxCast database release [16]. Additionally, tcpl can be executed upon each resampled set and generate an AC50 (or AC10, AC20, area under the curve, etc.) and further generate a median value and a 95% confidence interval, on the basis of the results. These approaches could be used as a method to determine the confidence in potency estimates derived from concentration series. A summary of curve-fitting uncertainty information is available in level 7 of the latest ToxCast database release, invitrodb version 3.1 [16].

The issue of quantitative uncertainty in fitting *in vitro* activity data is not unique to ToxCast data, and different approaches have been developed to handle this uncertainty. Previously, the US EPA's Benchmark Dose software, used commonly in modeling dose–response information from *in vivo* toxicity studies to define PODs, has been adapted for use in modeling gene expression data sets as BMDExpress [3,82]. BMDExpress enables automated analysis of continuous transcriptomic data via identification of genes that demonstrate significant dose–response behavior, followed by curve-fitting with statistical models that can be used to define a confidence interval around a potency estimate for significant changes in expression for a given gene. The upper and lower bounds on the confidence interval for a BMD thus give a sense of the uncertainty in fitting the available transcriptomic data. A commonality between the BMDExpress and toxboot approaches, both suited to HTS data of different types, is the desire to communicate a confidence interval around the potency estimate. BMDExpress is flexible in that the upper and lower bounds (BMDU and BMDL, respectively) of the confidence interval around a calculated potency value can be modified, allowing for differing levels of uncertainty in the analysis. Using default parameters, for a differentially expressed gene, BMDExpress will calculate a 95% confidence interval around a BMD for a 10% increase or decrease in response compared to the background of the control samples.

Beyond the uncertainty in curve-fitting high-throughput bioactivity data, there are also uncertainties associated with the TK parameters and models needed for IVIVE. As with *in vivo* toxicity data, there is a paucity of *in vivo* TK data sufficient for performing IVIVE. Instead, NAMs developed by the pharmaceutical industry have been adapted for screening more diverse chemicals. High-throughput toxicokinetics (HTTK) is an NAM that relies on chemical-specific data from *in vitro* experiments along with generalized TK mathematical models to develop tentative predictions of quantities such as the steady-state plasma concentration (C_{ss}) for a 1 mg/kg body weight/day dose [79,81]. Because HTTK requires the development of chemical-specific analytical chemistry methods, the uncertainty associated with the *in vitro* measurements varies significantly from chemical to chemical [72]. In addition, uncertainty is increased by many factors that are not always covered in high-throughput IVIVE approaches, including extrahepatic metabolism and active transport. Predictions from HTTK are reasonably concordant with *in vivo* data for quantities such as peak plasma concentration (C_{max}) and time-integrated plasma concentration (area under the curve) but have greater uncertainty for predicting C_{ss} [63,69,70,74,83]. For pharmaceutical chemicals, IVIVE methods predict plasma concentrations within a factor of three of *in vivo* observations [74]. For more diverse chemistries, however, only roughly half of predictions are this accurate, although the overall accuracy (or lack thereof) can itself be predicted [70]. TK NAMs allow for the estimation of human population variability [29], and the most sensitive 95th percentile C_{ss} with respect to variability is often used for IVIVE [79,81]. Recent work propagates both chemical-specific measurement uncertainty and population variability into the C_{ss} 95th percentile which in turn can be propagated into IVIVE [72].

A major assumption common to IVIVE applications is that chemical partitioning between medium and cells *in vitro* is equivalent to chemical partitioning between blood and tissue *in vivo* [57,79]. Significant differences in how a chemical may partition between cells, medium, and plastic *in vitro* may make this assumption less accurate, thus introducing additional uncertainty to IVIVE [23,41,48,49]. However, the extent of this problem is unknown, that is, the extent of differential partitioning between *in vitro* and *in vivo* situations and the number of chemicals that differentially partition are currently uncertainties in IVIVE. Understanding the relationship between chemical structure and medium to cell partitioning may provide for more accurate IVIVE; indeed, existing models suggest that the chemical-specific octanol:water partitioning coefficient and the concentration of serum (including lipids and proteins) in *in vitro* media may be strong determinants of *in vitro* chemical disposition [2,19]. Empirical determination of *in vitro* chemical disposition may help augment these models and establish the extent of this problem, thereby reducing uncertainty in IVIVE [9].

Uncertainty in QSAR modeling

As part of the NAM paradigm, computational methods (e.g. QSAR and read-across) are commonly used to predict hazard in the absence of experimental data [13,50,51,56]. However, computational methods are prone to uncertainty because of various sources which introduce limitations in the development, evaluation, validation, reliability, and regulatory acceptance of

computational models. Uncertainty in QSAR models can be attributed to two primary sources: (1) experimental data (as described previously) and (2) computational modeling approach. Computational model-driven uncertainty can arise because of (a) the size, distribution, and balance of the training data set [24]; (b) choice of molecular descriptors [34,55]; and (c) the choice of modeling algorithm and hyperparameter selection. This last issue is usually dealt with by tuning the hyperparameters of the model using multiple cross-validation [8] and by development of ensemble models [11,44,56]. For read-across, uncertainties usually arise from similarity justification and the completeness of the method used, necessitating a need for structured reporting [60] and uncertainty characterization [5] of read-across predictions. Another key aspect to understanding and quantifying model uncertainty is to define the applicability domain [12] and quantify the 'statistical' applicability domain [58]. Characterization of various sources of data variability, adequate incorporation of variability in computational model development, and quantification of data-driven uncertainty in model predictivity are critically needed to improve the reliability and acceptance of computational models.

As a first example of uncertainty analysis in QSAR, we describe a model for predicting *in vivo* PODs to illustrate methods to account for the variability in underlying training data. The *in vivo* data are taken from the EPA's ToxValDB, a compilation of information on about 3000 unique chemicals from a variety of public data sources, available through https://comptox.epa.gov. The standard distribution of POD values, per chemical, for different species and study type combination for the data set used in this study ranged from 0.40 to 0.60 $\log_{10}$-mg/kg/day. As described previously, it has been shown that the typical study-to-study variability (even within study type) may be about 0.5 $\log_{10}$-mg/kg/day (Pham et al. in prep). Based on these observations, it was assumed that PODs for multiple studies for a chemical follow a log-normal distribution. Thus, a POD distribution was constructed for each chemical (regardless of whether multiple studies were available) where the mean of the distribution was set as the median experimental POD value for that chemical and the standard deviation of the distribution was set as 0.5 log-units. This resultant distribution was used to develop N bootstrap QSAR models where for each model the training POD value for each chemical was drawn randomly from the distribution. The predictions from the model were used to estimate the uncertainty bounds in output model predictions. This is in contrast to typical QSAR models, which provide only point estimates.

A similar model was built for ecological points of departure in fish [61] using data largely from the EPA ECOTOX database [66]. This data set incorporates information from tens of thousands of studies covering thousands of unique chemicals and hundreds of fish species. To maximize the amount of useable data, different types of experiments were aggregated by using experimental covariates, such as end point type or exposure route, as binary features. Taxonomic groupings were also entered as binary features corresponding to the species, genus, and so on of the fish used in each experiment. These features were pared down to a manageable number by eliminating groupings that were uncorrelated with the end point, multicollinear, or present in too few or too many experiments. Because validation performance can vary because of the size and choice of test set, the standard deviation of the error was offered in addition to the usual root-mean-square error. These quantities are approximately equal for a large, centered error distribution, as in this model. Well-known formulas for the confidence intervals of the standard deviations were then used to bound the error estimates for a given model and test set. With thousands of chemicals available, externally cross-validated error could be restricted to a total range of 0.06 $\log_{10}$(mg/L), allowing stronger conclusions to be drawn about relative model performance.

Future directions

The work to define uncertainty in traditional and NAM approaches for chemical risk assessment is ongoing and focused on several areas. The applicability of current evaluations of the variance and reproducibility of *in vivo* data sets to a general understanding of the 'true' variability in animal studies would be enhanced by expansion of the curated *in vivo* data available, including data from nonmammalian species. Further, modeling of the uncertainty in the IVIVE approach for high-throughput bioactivity data is needed, especially around the development and refinement of models for *in vitro* disposition of chemicals in high-throughput assays (i.e., moving beyond the nominal micromolar concentration to considerations of the behavior of chemicals *in vitro*) and uncertainties in NAMs used to indicate the key physiological processes for chemical disposition. The expanded use and quality of NAMs can also serve as a feedback loop to identify possible experimental errors (i.e. during data curation or misidentification by the researcher). Even when the quantified uncertainty for an NAM may appear large, application of this information to different prioritization applications is important for understanding the precision possible when using the NAM. Of course, as investigations of the variability in traditional methods continue, practical benchmarks for the predictivity and reproducibility of NAMs will be elucidated. Further characterization of the specific sources of uncertainty at each step of applying NAMs will better qualify their use and help focus additional research on information that will reduce uncertainty.

Conflict of interest

The authors declare no conflicts of interest.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

References

1. Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, *et al.*: **Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment**. *Environ Toxicol Chem* 2010, **29**:730–741.
2. Armitage JM, Wania F, Arnot JA: **Application of mass balance models and the chemical activity concept to facilitate the use of in vitro toxicity data for risk assessment**. *Environ Sci Technol* 2014, **48**:9770–9779.
3. Auerbach S: **The NTP proposed approach to genomic dose-response modeling - scott auerbach, ph.D., NIH/NIEHS/DNTP**. In *Peer review of draft NTP approach to genomic dose-response modeling webinar series*; 2017.
4. Bailey J, Thew M, Balls M: **Predicting human drug toxicity and safety via animal tests: can any one species predict drug toxicity in any other, and do monkeys help?** *Altern Lab Anim* 2015, **43**:393–403.
5. Blackburn K, Stuard SB: **A framework to facilitate consistent characterization of read across uncertainty**. *Regul Toxicol Pharmacol* 2014, **68**:353–362.
6. Browne P, Judson RS, Casey WM, Kleinstreuer NC, Thomas RS: **Screening chemicals for estrogen receptor bioactivity using a computational model**. *Environ Sci Technol* 2015, **49**: 8804–8814.
7. Casati S: **Integrated approaches to testing and assessment**. *Basic Clin Pharmacol Toxicol* 2018. https://doi.org/10.1111/bcpt.13018.
8. Consonni V, Ballabio D, Todeschini R: **Evaluation of model predictive ability by external validation techniques**. *J Chemom* 2010, **24**:194–201.
9. Crizer DM, Harrill JA, Sipes N, Wetmore BA, Waidyanatha S, Richard A, Williams A, Ferguson SS, Honda GS, Wambaugh J, *et al.*: **Where is the chemical?: the *in vitro* disposition of Tox21 chemicals**. In *Society of toxicology 2019 annual meeting (Vol. Abstract 3136, Baltimore, MD)*; 2019.
10. Davis JA, Gift JS, Zhao QJ: **Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1**. *Toxicol Appl Pharmacol* 2011, **254**:181–191.
11. Dietterich TG. *Ensemble methods in machine learning*, **1857**; 2000:1–15.
12. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan O: **A stepwise approach for defining the applicability domain of SAR and QSAR models**. *J Chem Inf Model* 2005, **45**:839–849.
13. Dudek A, Arodz T, Galvez J: **Computational methods in developing quantitative structure-activity relationships (QSAR): a review**. *Comb Chem High Throughput Screen* 2006, **9**:213–228.
14. Dumont C, Barroso J, Matys I, Worth A, Casati S: **Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches**. *Toxicol In Vitro* 2016, **34**:220–228.
15. ECHA: **New approach methodologies in regulatory science**. In *Proceedings of a scientific workshop, April 19-20, 2016*; 2016. Helsinki, Finland. Available at: https://echa.europa.eu/documents/10162/22816069/scientific_ws_proceedings_en.pdf/a2087434-0407-4705-9057-95d9c2c2cc57. Accessed 3 September 2018.
16. EPA, U: **ToxCast database (invitroDB)**. In *T. U. S. E. P. A. s. N. C. f. C. Toxicology*; 2019. https://doi.org/10.23645/epacomptox.6062623.v33.1.
17. Filer DL, Kothiya P, Setzer RW, Judson RS, Martin MT: **tcpl: the ToxCast pipeline for high-throughput screening data**. *Bioinformatics* 2017, **33**:618–620.
18. Filipsson AF, Sand S, Nilsson J, Victorin K: **The benchmark dose method–review of available models, and recommendations for application in health risk assessment**. *Crit Rev Toxicol* 2003, **33**:505–542.
19. Fischer FC, Henneberger L, Konig M, Bittermann K, Linden L, Goss KU, Escher BI: **Modeling exposure in the Tox21 in vitro bioassays**. *Chem Res Toxicol* 2017, **30**:1197–1208.
20. Gaylor DW, Kodell RL: **A procedure for developing risk-based reference doses**. *Regul Toxicol Pharmacol* 2002, **35**(2 Pt 1): 137–141.
21. Gold LS, Bernstein L, Magaw R, Slone TH: **Interspecies extrapolation in carcinogenesis: prediction between rats and mice**. *Environ Health Perspect* 1989, **81**:211–219.
22. Gottmann E, Kramer S, Pfahringer B, Helma C: **Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments**. *Environ Health Perspect* 2001, **109**: 509–514.
23. Groothuis FA, Heringa MB, Nicol B, Hermens JL, Blaauboer BJ, Kramer NI: **Dose metric considerations in in vitro assays to improve quantitative in vitro-in vivo dose extrapolations**. *Toxicology* 2015, **332**:30–40.
24. Haibo H, Garcia EA: **Learning from imbalanced data**. *IEEE Trans Knowl Data Eng* 2009, **21**:1263–1284.
25. Haseman JK: **Using the NTP database to assess the value of rodent carcinogenicity studies for determining human cancer risk**. *Drug Metab Rev* 2000, **32**:169–186.
26. Hoffmann S, Kinsner-Ovaskainen A, Prieto P, Mangelsdorf I, Bieler C, Cole T: **Acute oral toxicity: variability, reliability, relevance and interspecies comparison of rodent LD50 data from literature surveyed for the ACuteTox project**. *Regul Toxicol Pharmacol* 2010, **58**:395–407.
27. Hoffmann S, Kleinstreuer N, Alepee N, Allen D, Api AM, Ashikaga T, Clouet E, Cluzel M, Desprez B, Gellatly N, *et al.*: **Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database**. *Crit Rev Toxicol* 2018, **48**: 344–358.
28. Izadi H, Grundy JE, Bose R: **Evaluation of the benchmark dose for point of departure determination for a variety of chemical classes in applied regulatory settings**. *Risk Anal* 2012, **32**: 830–835.
29. Jamei M, Marciniak S, Feng K, Barnett A, Tucker G, Rostami-Hodjegan A: **The Simcyp® population-based ADME simulator**. *Expert Opin Drug Metabol Toxicol* 2009, **5**:211–223.
30. Judson R, Houck K, Martin M, Richard AM, Knudsen TB, Shah I, Little S, Wambaugh J, Woodrow Setzer R, Kothya P, *et al.*: **Analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space**. *Toxicol Sci* 2016, **152**:323–339.
31. Judson RS, Kavlock RJ, Setzer RW, Cohen Hubal EA, Martin MT, Knudsen TB, Houck KA, Thomas RS, Wetmore BA, Dix DJ: **Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment**. *Chem Res Toxicol* 2011, **24**:451–462.
32. Judson RS, Martin MT, Egeghy P, Gangwal S, Reif DM, Kothiya P, Wolf M, Cathey T, Transue T, Smith D, *et al.*: **Aggregating data for computational toxicology applications: the U.S. Environmental protection agency (EPA) aggregated computational toxicology resource (ACToR) system**. *Int J Mol Sci* 2012, **13**:1805–1831.
33. Judson RS, Martin MT, Patlewicz G, Wood CE: **Retrospective mining of toxicology data to discover multispecies and chemical class effects: anemia as a case study**. *Regul Toxicol Pharmacol* 2017, **86**:74–92.

34. Katritzky AR, Gordeeva EV: **Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research**. *J Chem Inf Model* 1993, **33**:835–857.

35. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D, *et al.*: **Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management**. *Chem Res Toxicol* 2012, **25**:1287–1302.

36. Kavlock RJ, Bahadori T, Barton-Maclaren TS, Gwinn MR, Rasenberg M, Thomas RS: **Accelerating the pace of chemical risk assessment**. *Chem Res Toxicol* 2018, **31**:287–290.

37. Kleinstreuer NC, Ceger PC, Allen DG, Strickland J, Chang X, Hamm JT, Casey WM: **A curated database of rodent uterotrophic bioactivity**. *Environ Health Perspect* 2015, **124**.

38. Kleinstreuer NC, Hoffmann S, Alepee N, Allen D, Ashikaga T, Casey W, Clouet E, Cluzel M, Desprez B, Gellatly N, *et al.*: **Non-animal methods to predict skin sensitization (II): an assessment of defined approaches (*)**. *Crit Rev Toxicol* 2018, **48**:359–374.

39. Knapen D, Vergauwen L, Villeneuve DL, Ankley GT: **The potential of AOP networks for reproductive and developmental toxicity assay development**. *Reprod Toxicol* 2015, **56**:52–55.

40. Knudsen TB, Martin MT, Kavlock RJ, Judson RJ, Dix DJ, Singh AV: **Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB**. *Reprod Toxicol* 2009, **28**:209–219.

41. Kramer NI, Di Consiglio E, Blaauboer BJ, Testai E: **Biokinetics in repeated-dosing in vitro drug toxicity studies**. *Toxicol In Vitro* 2015, **30**(1 Pt A):217–224.

42. Kuljus K, von Rosen D, Sand S, Victorin K: **Comparing experimental designs for benchmark dose calculations for continuous endpoints**. *Risk Anal* 2006, **26**:1031–1043.

43. Leisenring W, Ryan L: **Statistical properties of the NOAEL**. *Regul Toxicol Pharmacol* 1992, **15**:161–171.

44. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, *et al.*: **CERAPP: collaborative estrogen receptor activity prediction project**. *Environ Health Perspect* 2016, **124**.

45. Martin MT, Judson RS, Reif DM, Kavlock RJ, Dix DJ: **Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database**. *Environ Health Perspect* 2009a, **117**:392–399.

46. Martin MT, Mendez E, Corum DG, Judson RS, Kavlock RJ, Rotroff DM, Dix DJ: **Profiling the reproductive toxicity of chemicals from multigeneration studies in the toxicity reference database**. *Toxicol Sci* 2009b, **110**:181–190.

47. Mazzatorta P, Estevez MD, Coulet M, Schilter B: **Modeling oral rat chronic toxicity**. *J Chem Inf Model* 2008, **48**:1949–1954.

48. Meacham CA, Freudenrich TM, Anderson WL, Sui L, Lyons-Darden T, Barone Jr S, Gilbert ME, Mundy WR, Shafer TJ: **Accumulation of methylmercury or polychlorinated biphenyls in in vitro models of rat neuronal tissue**. *Toxicol Appl Pharmacol* 2005, **205**:177–187.

49. Mundy WR, Freudenrich TM, Crofton KM, DeVito MJ: **Accumulation of PBDE-47 in primary cultures of rat neocortical cells**. *Toxicol Sci* 2004, **82**:164–169.

50. National Academies of Sciences, E., and Medicine: *Application of systematic review methods in an overall strategy for evaluating low-dose toxicity from endocrine active chemicals*. Washington, D.C: National Academies Press (US); 2017. https://doi.org/10.17226/24758.

51. Patlewicz G, Ball N, Booth ED, Hulzebos E, Zvinavashe E, Hennes C: **Use of category approaches, read-across and (Q)SAR: general considerations**. *Regul Toxicol Pharmacol* 2013, **67**:1–12.

52. Perkins EJ, Antczak P, Burgoon L, Falciani F, Garcia-Reyero N, Gutsell S, Hodges G, Kienzler A, Knapen D, McBride M, *et al.*: **Adverse outcome pathways for regulatory applications: examination of four case studies with different degrees of completeness and scientific confidence**. *Toxicol Sci* 2015, **148**:14–25.

53. Pham, L. L., Watford, S., Pradeep, P., Martin, M. T., Judson, R., Setzer, R. W., and Paul Friedman, K. (in prep). Variability in in vivo Toxicity Studies: defining the upper limit of predictivity for models of systemic effect levels. In (.

55. Pradeep P, Mansouri K, Patlewicz G, Judson R: **A systematic evaluation of analogs and automated read-across prediction of estrogenicity: a case study using hindered phenols**. *Comput Toxicol* 2017, **4**:22–30.

56. Pradeep P, Povinelli RJ, White S, Merrill SJ: **An ensemble model of QSAR tools for regulatory risk assessment**. *J Cheminf* 2016, **8**.

57. Rotroff DM, Wetmore BA, Dix DJ, Ferguson SS, Clewell HJ, Houck KA, Lecluyse EL, Andersen ME, Judson RS, Smith CM, *et al.*: **Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening**. *Toxicol Sci* 2010, **117**:348–358.

58. Roy K, Ambure P, Aher RB: **How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models?** *Chemometr Intell Lab Syst* 2017, **162**:44–54.

59. Sand S, Victorin K, Filipsson AF: **The current state of knowledge on the use of the benchmark dose concept in risk assessment**. *J Appl Toxicol* 2008, **28**:405–421.

60. Schultz TW, Amcoff P, Berggren E, Gautier F, Klaric M, Knight DJ, Mahony C, Schwarz M, White A, Cronin MTD: **A strategy for structuring and reporting a read-across prediction of toxicity**. *Regul Toxicol Pharmacol* 2015, **72**:586–601.

61. Sheffield, T. Y., and Judson, R. S. (In prep). Ensemble QSAR modeling to predict multispecies fish toxicity points of departure. TBD..

62. Simon TW, Zhu Y, Dourson ML, Beck NB: **Bayesian methods for uncertainty factor application for derivation of reference values**. *Regul Toxicol Pharmacol* 2016, **80**:9–24.

63. Sipes NS, Wambaugh JF, Pearce R, Auerbach SS, Wetmore BA, Hsieh J-H, Shapiro AJ, Svoboda D, DeVito MJ, Ferguson SS: *An intuitive approach for predicting potential human health risk with the Tox21 10k library*. 2017.

64. Slob W, Moerbeek M, Rauniomaa E, Piersma AH: **A statistical evaluation of toxicity study designs for the estimation of the benchmark dose in continuous endpoints**. *Toxicol Sci* 2005, **84**:167–185.

65. Slob W, Setzer RW: **Shape and steepness of toxicological dose-response relationships of continuous endpoints**. *Crit Rev Toxicol* 2014, **44**:270–297.

66. US EPA: *ECOTOX knowledgebase*. 2018. Available at: https://cfpub.epa.gov/ecotox/. Accessed 13 August 2018.

67. Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, *et al.*: **Adverse outcome pathway (AOP) development I: strategies and principles**. *Toxicol Sci* 2014a, **142**:312–320.

68. Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, *et al.*: **Adverse outcome pathway development II: best practices**. *Toxicol Sci* 2014b, **142**:321–330.

69. Wambaugh JF, Hughes MF, Ring CL, MacMillan DK, Ford J, Fennell TR, Black SR, Snyder RW, Sipes NS, Wetmore BA: **Evaluating in vitro-in vivo extrapolation of toxicokinetics**. *Toxicol Sci* 2018, **163**:152–169.

70. Wambaugh JF, Wetmore BA, Pearce R, Strope C, Goldsmith R, Sluka JP, Sedykh A, Tropsha A, Bosgra S, Shah I, *et al.*: **Toxicokinetic triage for environmental chemicals**. *Toxicol Sci* 2015a, **147**:55–67.

71. Wambaugh JF, Wetmore BA, Pearce R, Strope C, Goldsmith R, Sluka JP, Sedykh A, Tropsha A, Bosgra S, Shah I, *et al.*: **Toxicokinetic triage for environmental chemicals**. *Toxicol Sci* 2015. https://doi.org/10.1093/toxsci/kfv118.

72. Wambaugh JF, Wetmore BA, Ring CL, Nicholas CI, Pearce R, Honda GS, Dinallo R, Angus D, Gilbert J, Sierra T, *et al.*: *Assessing the impact of uncertainty and variability on high-throughput predictions of chemical toxicokinetics*. 2019.

73. Wang B, Gray G: **Concordance of noncarcinogenic endpoints in rodent chemical bioassays**. *Risk Anal* 2015, **35**:1154–1166.

74. Wang Y-H: **Confidence assessment of the simcyp time-based approach and a static mathematical model in predicting clinical drug-drug interactions for mechanism-based CYP3A inhibitors**. *Drug Metab Dispos* 2010, **38**:1094–1104.

75. Watford, S., Pham, L. L., Wignall, J., Shin, R., Martin, M. T., and Paul Friedman, K. submitted for publication. ToxRefDB version 2.0: improved utility for predictive and retrospective toxicology analyses. *In Review*. Special Issue, "Translational Use of In Vitro- and In Silico-Based Developmental Toxicity Systems for Human Health Assessment." Reproductive Toxicology.

76. Watt ED: *Toxboot: bootstrap methods for ToxCast high throughput screening data*. 2019.

77. Watt ED, Judson RS: **Uncertainty quantification in ToxCast high throughput screening**. *PLoS One* 2018, **13**, e0196963.

78. Wetmore BA, Allen B, Clewell 3rd HJ, Parker T, Wambaugh JF, Almond LM, Sochaski MA, Thomas RS: **Incorporating population variability and susceptible subpopulations into dosimetry for high-throughput toxicity testing**. *Toxicol Sci* 2014, **142**: 210–224.

79. Wetmore BA, Wambaugh JF, Allen B, Ferguson SS, Sochaski MA, Setzer RW, Houck KA, Strope CL, Cantwell K, Judson RS, *et al.*: **Incorporating high-throughput exposure predictions with dosimetry-adjusted in vitro bioactivity to inform chemical toxicity testing**. *Toxicol Sci* 2015a, **148**: 121–136.

80. Wetmore BA, Wambaugh JF, Ferguson SS, Li L, Clewell 3rd HJ, Judson RS, Freeman K, Bao W, Sochaski MA, Chu TM, *et al.*: **Relative impact of incorporating pharmacokinetics on predicting in vivo hazard and mode of action from high-throughput in vitro toxicity assays**. *Toxicol Sci* 2013, **132**: 327–346.

81. Wetmore BA, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, Freeman K, Clewell 3rd HJ, Dix DJ, Andersen ME, Houck KA, *et al.*: **Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment**. *Toxicol Sci* 2012, **125**:157–174.

82. Yang L, Allen BC, Thomas RS: **BMDExpress: a software tool for the benchmark dose analyses of genomic data**. *BMC Genom* 2007, **8**:387.

83. Yoon M, Efremenko A, Blaauboer BJ, Clewell HJ: **Evaluation of simple in vitro to in vivo extrapolation approaches for environmental compounds**. *Toxicol In Vitro* 2014, **28**:164–170.

84. Zbinden G: **Predictive value of animal studies in toxicology**. *Regul Toxicol Pharmacol* 1991, **14**:167–177.