ALTEX, accepted manuscript
published June 27, 2023
doi:10.14573/altex.2210171

Research Article

The Impact of Biostatistics on Hazard Characterization Using *In Vitro* Developmental Neurotoxicity Assays

Hagen Eike Keßel[1], Stefan Masjosthusmann[1], Kristina Bartmann[1], Jonathan Blum[2], Arif Dönmez[1], Nils Förster[4], Jördis Klose[1], Axel Mosig[4], Melanie Pahl[1], Marcel Leist[2], Martin Scholze[5] and Ellen Fritsche[1,3*]*

[1]IUF - Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany; [2]*In vitro* Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Konstanz, Germany; [3]Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany; [4]Bioinformatics Group, Ruhr University Bochum, Bochum, Germany; [5]Brunell University, London, UK

Abstract

In chemical safety assessment, benchmark concentrations (BMC) and their associated uncertainty are needed for the toxicological evaluation of *in vitro* data sets. A BMC estimation is derived from concentration-response modelling and results from various statistical decisions, which depend on factors such as experimental design and assay endpoint features. In current data practice, the experimenter is often responsible for the data analysis and therefore relies on statistical software often without being aware of the software default settings and how they can impact the outputs of data analysis. To provide more insight into how statistical decision-making can influence the outcomes of data analysis and interpretation, we have developed an automatic platform that includes statistical methods for BMC estimation, a novel endpoint-specific hazard classification system, and routines that flag data sets that are outside the applicability domain for an automatic data evaluation. We used case studies on a large dataset produced by a developmental neurotoxicity (DNT) *in vitro* battery (DNT IVB). Here we focused on the BMC and its confidence interval (CI) estimation as well as on final hazard classification. We identified five crucial statistical decisions the experimenter must make during data analysis: choice of replicate averaging, response data normalization, regression modelling, BMC and CI estimation, and choice of benchmark response levels. The insights gained in are intended to raise more awareness among experimenters on the importance of statistical decisions and methods but also to demonstrate how important fit-for-purpose, internationally harmonized and accepted data evaluation and analysis procedures are for objective hazard classification.

1 Introduction

In 2007, the National Research Council (NRC) of the United States proposed a new strategy for toxicity testing in the 21st century centering around a shift from *in vivo* experiments in animals to mechanism-based *in vitro* testing (NRC, 2007). Since then, major advances in the field of *in vitro* toxicology have been made, including development and establishment of medium and high throughput screening (HTS) assays, as well as bioinformatics tools for data generation, management and analysis (Leist et al., 2014; Wheeler et al., 2015; Villeneuve et al., 2019). These efforts are contributing to next generation risk assessment, which aims at using new approach methods (NAMs) for exposure-based, hypothesis-driven risk assessment without the generation of new animal data (Li et al., 2021; Dent et al., 2021; Pallocca et al., 2022).

Typically, an *in vitro* HTS test system produces hazard data for a relatively large number of test concentrations and thus makes it most suitable for concentration-response regression modelling. This statistical approach allows the interpolative estimation of a concentration value at a given effect level (effect or inhibitory concentration), and of particular regulatory interest is hereby the benchmark concentration (BMC) and its associated uncertainty, expressed as lower limit of a one-sided 95% confidence interval (BLL). In analogy with the benchmark dose (BMD) approach for *in vivo* studies, a BMC is considered as lowest concentration of the test compound that produces a pre-defined small "relevant" change to the control reference's response level (Crump, 1995; Krebs et al., 2020), and as a consequence, the benchmark response (BMR) value should be as "close as possible" to the control response.

* contributed equally

Received October 17, 2022; Accepted May 31, 2023;
Epub June 27, 2023;

ALTEX 40(#), ###-###. doi:10.14573/altex.2210171

Correspondence: Hagen Eike Keßel, PhD
IUF - Leibniz Research Institute for Environmental Medicine
Leibniz-Institut für umweltmedizinische Forschung GmbH
Auf'm Hennekamp 50
40225 Düsseldorf, Germany

ALTEX, accepted manuscript
published June 27, 2023
doi:10.14573/altex.2210171

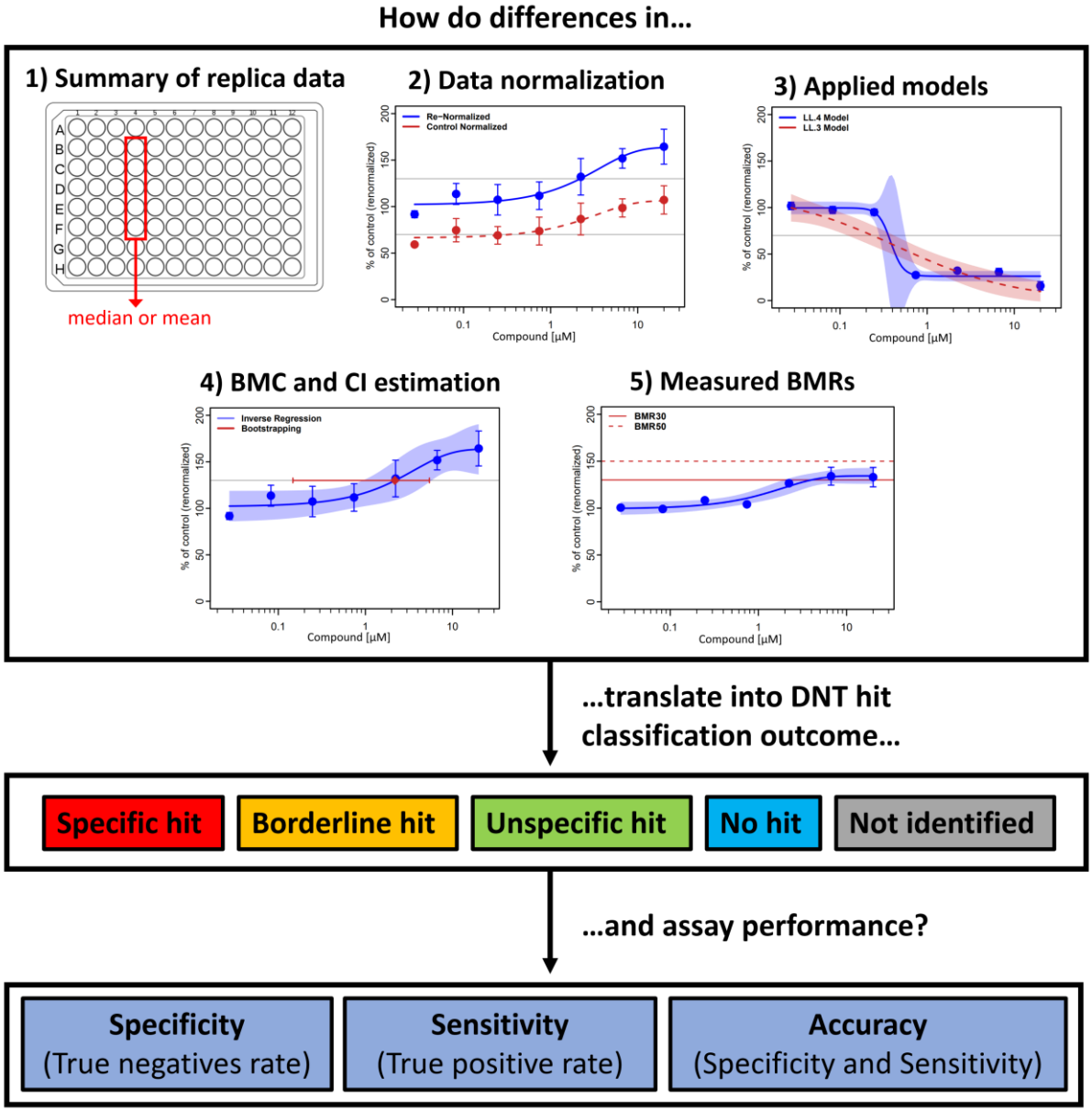

Fig. 1: Study overview
Several biostatistical data analysis and evaluation steps were analyzed for their impact on a BMC estimation and subsequent hazard characterization from developmental neurotoxicity (DNT) data: i) how to average replicate responses from an experiment, ii) how to normalize concentration-response data, iii) how to describe concentration-response data by regression modelling, iv) how to estimate a benchmark concentration (BMC) and its uncertainty, and v) which benchmark response (BMR) level to select. Changes between statistical methods were recorded for 148 compounds tested on up to 22 assay endpoints, and their impact translated into the compound's DNT hit classification and the predictivity performance of the overall assay battery.

In vitro test systems represent a huge variety of different types of assays, from cell-free, cell and tissue-based methods up to multi-response organoid systems, and as a consequence, concentration-response data between these systems vary enormously with respect to their test-specific experimental designs, data variability, dynamic ranges and concentration-response pattern. Unique to HTS systems is also that assay outputs are produced in microplate multi-well readers, with concentration-response data from the same concentration and experiment are considered to reflect technical (intra-replicate) variation and data from repeated experiments more indicative for "biological" (between-study) variation. These hierarchical data are usually simplified by using an average response value per test concentration and experiment (replicate average) as statistical unit for the concentration-response analysis, with the argument that the BMC and BLL estimation should reflect mainly biological and between-study variability.

The BMC estimation consists of various statistical decisions to be made in the concentration-response analysis, which depends largely on the experimental design, the concentration-response data and assay endpoint features, and which require statistical knowledge that is usually only warranted by experienced biostatisticians. In current data practice, the experimenter is often responsible for the data analysis and therefore relies on statistical software without being aware about the software default settings and how they can impact the outputs of data analysis (Jensen et al., 2020). Existing guidelines for concentration

response data analysis are often too general (OECD, 2006; EFSA, 2017), and no clear consensus on a common and standardized biostatistical method for *in vitro* toxicity data have been achieved (Wheeler et al., 2015; Sand et al., 2017).

To provide more insight into how statistical decision making can influence the outcomes of data analysis and interpretation, we have used case studies on a large dataset produced by a developmental neurotoxicity (DNT) *in vitro* battery (DNT IVB; Masjosthusmann et al., 2020, Crofton and Mundy, 2021, Blum et al. 2022). In this DNT IVB, 148 compounds were tested across up to ten test methods representing the neurodevelopmental key events (KE) of neural progenitor cell (NPC) proliferation, migration of neural crest and radial glia cells, neurons and oligodendrocytes, neuronal differentiation, neurite outgrowth of peripheral and central nervous system neurons, as well as oligodendrocyte differentiation, and accomplished by various endpoints measuring cell viability and cytotoxicity (Masjosthusmann et al., 2020, Blum et al. 2022). Some of the DNT-specific endpoints are derived from primary and organotypic cultures which mimic a system of high physiological complexity and cell type heterogeneity, and thus are more prone to a data variability typically observed in animal studies. Here we focused on the BMC and its confidence interval (CI) estimation, as well as the final hazard classification. For this purpose, we identified five crucial statistical decisions the experimenter have to face during the data analysis (Figure 1):

(i) Replicate distribution and choice of location parameter for regression analysis: shall the median of all replicate responses of an experiment be used, which makes no assumption to the data and thus reduces the negative impact of potential data outlier, or the replicate mean, which is more efficient when the replicate responses follow a symmetric distribution, but if violated, can lead to a biased estimation of the replicate mean?

(ii) Response data normalization: shall the responses of an experiment always be normalized to the control's response even if the exposure responses provide clear evidence against the use of control data, or shall in that case the "control reference" be estimated directly from responses observed at low exposure concentrations ("re-normalization", Krebs et al., 2018)?

(iii) Regression model: shall the concentration-response data always be described by the same and supposedly flexible mathematical model, or is it better to use several models and either subsequently select the best model by means of goodness-of-fit criteria ("best-fit method", Scholze et al., 2001) or estimate an average of all model fits ("model averaging", Claeskens et al., 2008)?

(iv) Uncertainty of a BMC estimation: shall the confidence level of a BMC be calculated by a simple and commonly used statistical approximation technique ("Delta method", Cox, 1990) which can lead to a less accurate confidence interval (Moerbeek et al., 2004), or by alternative approaches such as bootstrapping or inverse regression (Jensen et al., 2019)?

(v) Benchmark response (BMR): shall a response level most close to the control reference be selected, which might not always be applicable for the statistical concentration-response analysis and thus might fail to provide a reliable BMC estimation, or a higher BMR, which guarantees a statistically more robust BMC estimation but might fail for compounds that has produced weak responses below the intended BMR?

We designed a standard data evaluation protocol ("standard protocol") which we used as reference to alternative statistical methods, so that their BMCs and confidence intervals (CIs) estimated to the same DNT IVB data could be compared. The statistical methods to be changed were chosen along the questions outlined in (i) to (v). This was supplemented by measuring their impact on hazard alerts derived from hit classifications, which separate cytotoxic concentration ranges from the respective BMC of the specific DNT endpoint, and by measuring their impact on the DNT IVB's capability on predicting DNT adversity in terms of specify, sensitivity and accuracy.

2 Methods

2.1 DNT data

All concentration response data used in this study are from a DNT *in vitro* battery of 8 assays with 22 endpoints, in which a total of 148 compounds were tested. 120 compounds were tested across all assays, while 28 compounds were tested in at least 2 assays. Fourteen assay endpoints represent major key neurodevelopmental processes, and 8 endpoints measure general cell viability and cytotoxicity (Table 1). This DNT *in vitro* battery was developed in collaboration with EFSA with the aim to advance the application of *in vitro* DNT testing for regulatory purposes. The term "BMC" was used equally for data from DNT-specific, cytotoxicity and viability endpoints.

Depending on the assay, fluorescent readouts using a multiplate reader or fluorescence and brightfield imaging with subsequent artificial intelligence-based image analysis (Schmuck et al., 2017; Förster et al., 2022) was performed as endpoint

Tab. 1: Number of endpoint-specific DNT hit classifications judged by experts
The numbers of hit classifications by expert judgement are presented as percentage of all classifications that were supervised by the hazard decision trees.

Method	NPC [%][a]	UKN [%][b]
Standard protocol	2.23	17.59
Replicate mean	2.15	16.67
Control-normalized	3.27	15.74
LL3rm	2.23	12.50
Delta method	8.09	18.52
Bootstrapping	3.12	17.13
Model averaging	2.90	15.74
BMR30+50	1.93	12.96

[a] NPC = Data outcomes from NPC assays; [b] UKN = Data outcomes from UKN assays

assessment. NPC Assays were conducted with three to five independent experiments and 5 replicates each, UKN Assays with three independent experiments and 6 replicates each (Tab. S1[1]). Each compound was tested in at least eight concentrations per experiment. An overview of the assays, the cell model and the respective endpoints is given in Table S1[1], and more detailed information about the assay-specific experimental testing procedures and test outcomes can be found elsewhere (Masjosthusmann et al., 2020, Crofton and Mundy, 2021, Blum et al., 2022).

The BMR for each endpoint was derived from the between experimental variability as the coefficient of variation of median plate medians (after normalization) measured at the lowest test concentration and across all independent experiments (Masjosthusmann et al., 2020). To achieve a better comparability across the endpoints, the BMRs were then rounded to the next higher value, resulting into three BMRs: a 10% change was selected for endpoints from the NPC2a and NPC1-5 cytotoxicity assay (BMR10), and a 30% change for endpoints from the NPC1, NPC2a, NPC2b, NPC3-5 and NPC1-5 viability assay (BMR30). For all UKN assays a 25% change was decided (BMR25), and for the viability of the UKN2 a BMR10 was chosen.

2.2 Data evaluation platform

For data processing and evaluation, the R package *drc*[2] (Ritz et al., 2019) was extended and optimized for the use of data from multi-well plate experiments. The biostatistical data evaluation software is freely available as open source under the name CRStats[3], an interactive R Markdown document is available and can freely be assessed for use. We defined a standard protocol for the evaluation of DNT IVB data with the following statistical methods: (i) average replicate per experiment estimated by median (2.2.3), (ii) control-normalization followed by re-normalization (2.2.4), (iii) application of several mathematical models to find the 'best fit' regression model for a BMC estimation (2.2.6), (iv) CI estimation of the BMC by inverse regression (2.2.7), and (v) selecting endpoint specific BMRs for the hazard classification as outlined in Table S1[1].

2.2.1 Minimal data requirements

Data were accepted for data analysis only if the following three minimal data requirements were fulfilled: (i) at least two replicas per concentration are available, otherwise all readouts from this concentration were excluded, (ii) at least five concentrations per experiment have provided readouts otherwise the whole experiment was excluded, and (iii) at least two control replicas are available otherwise the whole experiment was excluded.

2.2.2 Pre-processing

CRSTATS uses different assay-specific pre-processing steps in order to obtain a single response value for each well. For example, the neuronal differentiation in the NPC3 assay is calculated as the number of neurons divided by the total number of cells with a nucleus:

$$NPC3\ neuronal\ differentiation\ [120h] = \frac{NPC3\ number\ neurons\ [120h]}{NPC3\ number\ cells\ [120h]} \quad \text{(Eq 1)}$$

All assay specific pre-processing methods that are currently implemented in CRSTATS are listed in Table S2[1].

2.2.3 Replicate averaging

The average assay response for controls and treatments from the same experiment was either estimated by the arithmetic mean or by the median. The variability between replicates was calculated as standard deviation (SD; for the mean) or as median absolute deviation (MAD; for the median). Outlier detection procedures were not applied and data points from wells where technical problems were known or obvious (e.g., scanned images were blurred or empty, staining did not work properly on all cells) were excluded from the data analysis.

2.2.4 Effect data normalization

CRSTATS offers different normalization methods which allows the translation of pre-processed effect data into relative values. For this study, we used the following two methods:

(i) Control normalization: effect responses are normalized to the mean or median of the solvent controls as

$$\frac{\text{replicate response}}{\text{median or mean (solvent control responses)}} \quad \text{(Eq 2)}$$

(ii) Control re-normalization: normalized effect responses (Equation 2) are further normalized by a mean value that has been estimated by regression modelling at the lowest test concentration, i.e.

$$\frac{\text{normalised replicate response}}{\text{model estimate of normalised response at lowest test concentration}} \text{ (Krebs et al., 2018)} \quad \text{(Eq 3)}$$

2.2.5 Significance analysis

The presence of at least one exposure concentration that had produced an effect response which differs statistically significantly from the responses of all remaining exposures is a crucial factor in the hazard classification method (2.2.8). To account for that, significant differences between treatment means were identified by using the Tukey Honest Significant Differences test (alpha=5%, two-sided) (Tukey HSD; Yandell, 1997), with hypothesis testing conducted on normalized replicate averages from at least three independent experiments. On sample size (≥ 3 independent experiments) and family-wise error rate (≥ 5 concentrations) considerations we expected the statistical limit of detection to match the effect size of the endpoint-dependent BMRs. As an average control value was always set to 100% (2.2.4), controls were excluded from the significance analysis. Data provided no evidence against the Gaussian assumption.

[1] doi:10.14573/altex.2210171s
[2] https://www.R-project.org/
[3] github.com/ArifDoenmez/CRStats

2.2.6 Concentration-response regression analysis

The R packages *drc* (Ritz et al., 2015) and *bmd* (Jensen et al., 2020) were used for regression analysis and the estimation of a BMC and its associated uncertainty. The *drm* function fits a pre-defined regression model to the concentration-response data, with several options implemented to provide more flexibility for the estimation method. A large number of mathematical nonlinear regression functions was applied to the same data set (Table S3[1]), and the best fitting model then selected on basis of the Akaike´s Information Criterion (AIC) ("best fit method", Scholze et al., 2001; Portet, 2020). AIC is commonly used to compare the relative goodness-of-fit among different models and to then choose the model of best predictive power by balancing data support against model complexity. As all effect endpoints in this study are continuous, the estimation method of ordinary least-squares (OLS) was used. OLS relies on two assumptions, i.e. (i) effect data (here replicate average) follow a symmetrical distribution, and (ii) variance homogeneity across all treatment groups. Both assumptions were checked prior to data analysis on basis of pooled endpoint-specific data from all experiments (Breusch-Pagan test for heteroscedasticity, Breusch et al., 1979; Triples test for symmetry, Randles et al., 1980): data variability differed in average by maximally 20% between the treatment groups, with the highest variability often occurring at highest test concentration, and no overall clear evidence was detected that normalized replicate means did not follow a symmetric distribution. These findings were deemed as acceptable for using the unweighted OLS regression analysis. Count data from assay endpoints were considered as continuous as counts were well above zero.

2.2.7 BMC and its uncertainty

In the standard protocol the BMC was estimated directly from the best fit model. We also considered model averaging as an alternative option where, similar to the previous best fitting method, a number of suitable concentration–response models were fitted to the same data, but in this case, all resulting model fits were combined to provide a weighted average of BMC estimates (Ritz et al., 2013). Uncertainty was always expressed as α=5%, i.e. the lower limit (BLL) corresponds to the 2.5% limit and upper limit (BUL) to the 97.5% limit. BLL and BUL were derived by three different methods, i.e. inverse regression, the delta method and bootstrapping. The estimation of the BMC and its 95% CI by model averaging was always performed in combination with bootstrapping. Inverse regression was used after having fitted the regression model to the data (Buckley et al., 2009; Fang et al., 2015). Here we used a simple, pragmatic approach by anchoring both BLL and BUL directly to the 95% CI of the regression curve, i.e., the intersection of the horizontal BMR line with the lower and upper 95% CIs of the regression fit determined the BLL and BUL. This approach puts high emphasizes on a successful regression fit in terms of robustness and reliability, rests on a Wald-based confidence interval which is only asymptotically valid, and assumes that the model parameters are close to being unbiased and normally distributed. The delta method is an asymptotic approach which combines information of the estimated model parameters to derive a Wald-type interval (Jensen et al., 2020). Bootstrapping uses computer-intensive simulation techniques that resamples the original dataset to create a huge number of so-called bootstrap samples, with each sample mirroring the original data set with an identical experimental design but newly simulated effect responses. On each bootstrap sample the same statistical data analysis was performed, resulting in a distribution of resampled BMC values around the original BMC estimation. If the median of this distribution equals the original BMC (unbiased resampling), then the 2.5% and 97.5% quantiles are expected to mirror the BUL and BLL of the original BMC, respectively. For each bootstrap sample, always the same regression model was used as part of the best-fit method, or one model-averaged BMC if model averaging was performed. To simplify the model averaging method, only three regression models were considered (four-parameter loglogistic, four-parameter Weibull and three-parameter exponential model). Bootstrapping was always conducted on 1000 resampled datasets, and due to the small sample sizes, we used always the parametric version (Efron and Tibshirani, 1994). All resampling was performed by the function *bmdMA* of the R package *bmd* (Jensen et al., 2020). Bootstrapping can simulate a bootstrap sample which do not allow a BMC estimation or which leads to an unreliable BMC estimation that is well outside the tested concentration range. Therefore, a resampled BMC was excluded from the resampling distribution if it was 1.5-times above the highest test concentration or below the lowest tested concentration.

To allow a better comparison between BMCs from different data scenarios, the BMC was transformed to a relative BMC on a log10 scale by relating the 100-fold BMC estimation to the highest test concentration of the data set:

$$relative\ BMC = log10(\frac{100 * BMC}{highest\ test\ concentration}) \quad \text{(Eq 4)}$$

A relative BMC of 1 corresponds to a BMC that is tenfold below the highest test concentration of the data set, and a relative BMC above 2 corresponds to a BMC that has been extrapolated beyond the highest test concentration. The lower the relative BMC value, the more likely the estimation is supported by effect data from more concentrations.

2.2.8 Hazard classification

CRSTATS uses a hazard classification approach which judges if data evidence is sufficient to define a compound as active for the specific DNT endpoint and if this can be distinguished from an activity observed in cell health related endpoints (viability and cytotoxicity). Accordingly, the endpoint-specific hazard of a compound is classified into five categories:

- *No hit:* no observed effect on the DNT-specific endpoint or on general cell health.
- *Unspecific hit:* the effect on the DNT-specific endpoint cannot be separated from an effect on the cell health related endpoint.
- *Borderline hit:* the separation between the effects on the DNT-specific endpoint and the effect on cell health related endpoint is statistically not clear (Leontaridou et al., 2017).
- *Specific hit:* the effect on the DNT-specific endpoint is clearly separated from an effect on the cell health related endpoint.
- *Not identified:* data are incomplete und do not allow any classification.

If the automatic classification failed due to a high uncertainty of the BMC or a missing BMC for the cell health related endpoint, the classification was recorded as **expert judgement** and classification into one of these five categories was done by manual inspection according to a guidance document (see Section 1.4 and 1.5 in the supplementary file[1] for more details). Each classification was done independently by two experts with a high interrater reliability (Kappa Statistic > 0.9). An overview

ALTEX, accepted manuscript
published June 27, 2023
doi:10.14573/altex.2210171

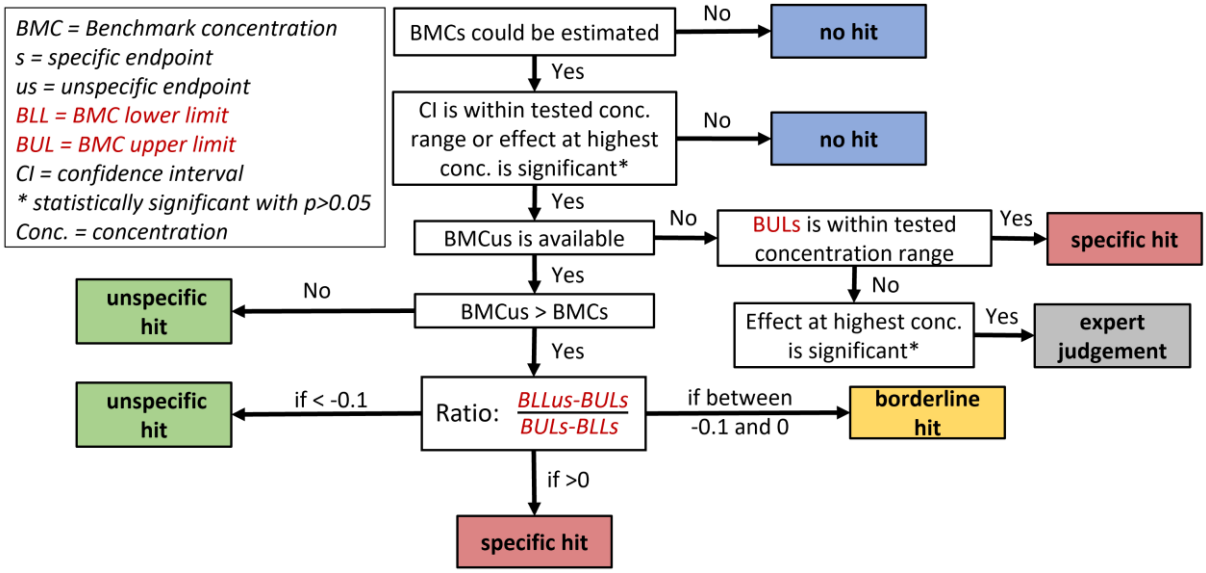

Fig. 2: Decision tree for the NPC hazard classification of inhibitory effects
The decision tree shows for NPC1-5 data with decreasing concentration-response pattern how BMC estimations and their uncertainty (expressed as 95% confidence intervals, CI) for data from both specific and unspecific endpoints are used to classify the compound into one of the DNT hit categories (colored boxes). Hits with the category "expert judgement" (grey box) will be classified into one of the DNT hit categories by manual inspection on the basis of all data evidence.

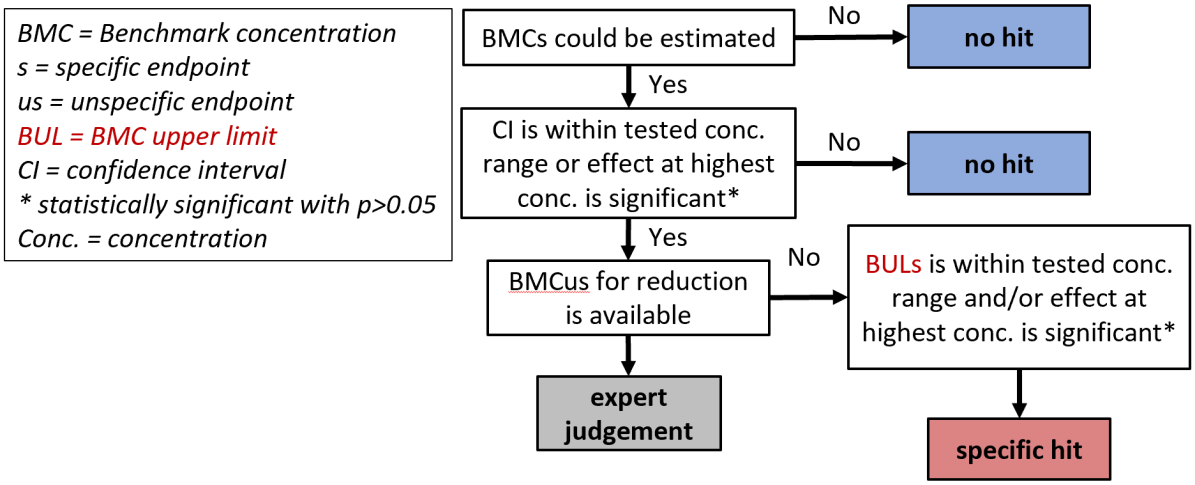

Fig. 3: Decision tree for NPC hazard classification of increasing effects
The decision tree shows for NPC1-5 data with increasing concentration-response pattern ("induction") how BMC estimations and their uncertainty (expressed as 95% confidence intervals, CI) for data from both specific and unspecific endpoints are used to classify the compound into a specific or no hit (colored boxes). The presence of a cytotoxic responses can lead to an artefact in the DNT-specific endpoint and is therefore initially categorized as "expert judgement". These hits will be classified into one of the DNT hit categories by manual inspection on the basis of all data evidence.

of all flagging alerts leading to expert judgement is given in Table S5[1]. The guidance used for expert judgement is given in Fig. S1[1].

The hazard classification approach was operationalized by hazard decision trees which reflect specific assay features and the directionality of the observed concentration response pattern (i.e., either reduction or inhibition). Common to all decisions trees is that they compare the BMC of the DNT-specific endpoints to the respective BMC of the unspecific endpoint (i.e., cytotoxicity or cell viability). For the NPC and UKN assays slightly different versions were developed, with all NPC assay endpoints accounting directly for the statistical uncertainties of both BMC estimations by using their corresponding CIs, and all UKN assay endpoints using pe-defined acceptance ranges instead. The principles of the hazard decision tree for data sets with decreasing concentration-response pattern (reduction) measured in NPC assays (NPC1, NPC2, NPC3 and NPC5, Table S1[1]) are shown in Figure 2, and for increasing concentration-response pattern (induction) in Figure 3. Inductions are handled separately, because the specific and unspecific endpoints do not have the same relationship during an induction, compared to a reduction in the endpoint. A loss in general cell viability for example will likely result in an effect in cell proliferation, while an induction in cell viability does not necessarily increase cell proliferation. If migration (NPC2a) is affected, only cytotoxicity is used as a reference for all specific endpoints of NPC2-5. A reduction in migration also reduces cell viability due to the lower number of cells in the migration area and not necessarily due to cell death. If so, it cannot be used as valid reference to discriminate between a specific and unspecific effect. The same applies to effects in cell viability. In these cases, only

cytotoxicity is used as general cell health reference for according specific NPC endpoints. More details can be found in the supplementary material (Section 1.3[1]) and in Table S4[1], and details about the classification tree applied to data from the UKN assays can be found in Masjosthusmann et al. (2020). The lower and upper confidence interval of the BMC always refers to the central two-sided 95% confidence interval (BLL=2.5% percentile, BUL=97.5% percentile).

2.3 Assay performance

From the 148 compounds tested in the DNT IVB, a set of 45 reference compounds (17 negative compounds that are known not to cause DNT; 28 positive compounds with proven DNT adversity in humans or mammals) was used for an evaluation of the DNY IVB predictivity. Hit decisions were derived from the hazard decision trees developed in 2.2.8, and the following performance parameters were used:

$$Specificity = \frac{\#\ true\ negative\ hits}{\#\ negative\ compounds} \quad \text{(Eq 5)}$$

$$Sensitivity = \frac{\#\ true\ positive\ hits}{\#\ positive\ compounds} \quad \text{(Eq 6)}$$

$$Accuracy = \frac{\#\ true\ negative\ hits + \#\ true\ positive\ hits}{\#\ negative\ compounds + \#\ positive\ compounds} \quad \text{(Eq 7)}$$

A negative compound was considered as true negative if it was not classified as specific hit or borderline in any of the assays. A positive compound was considered as true positive, if it was classified as specific hit or borderline in at least one assay.

3 Results

The impact of different statistical methods was quantified by comparing outcomes of the standard protocol with those from the following alternatives: 1) average replicate per experiment estimated by the arithmetic mean, 2) control normalization without re-normalization, 3) using a three-parameter log-logistic regression model (LL3rm) for the BMC estimation, 4) using model-averaging for the BMC estimation, 5) CI estimation of the BMC by the delta method, 6) CI estimation of the BMC by bootstrapping, 7) CI estimation of the BMC by model averaging, and 8) increasing the endpoint specific BMR by 20% (BMR30 and 50). Differences in the BMC estimation and its lower 95% CI , the endpoint-specific hazard classification of the compound and the final assay performance were quantified and compared across the various specific assay endpoints.

In total, 148 compounds were tested on up to 14 DNT-specific and 8 cytotoxicity and viability endpoints, leading to a total of 2426 datasets of which 2385 (98.31%) fulfilled the minimal data requirements of the data evaluation pipeline. According to the standard protocol, it was possible to perform a regression analysis for 2385 data sets (1953 NPC and 432 UKN) and a hazard hit categorization for 1563 data sets from DNT-specific endpoints (1347 NPC and 216 UKN). Three-parametric models were suggested as best-fit choice for 55.85% of these data sets, the exponential function with two model parameters for 31.74%, four-parameter models for 10.48%, and the most complex model (5-parameter general log-logistic) for 1.93%.

3.1 Impact of different data evaluation methods on the BMC estimation

The relative BMCs from the standard and alternative statistical protocols are shown in Figures 4 A-E for five statistical parameters that were changed, with the BMC of the alternative protocol always referring to the x-axis and the BMC of the standard protocol to the y axis. If a regression analysis could be performed but a BMC not established due to missing data support for the BMR, the BMC was flagged as “BMRnr” (BMR not reached) and included in the plot at the end of the BMR axes, i.e., a BMRnr value on the right side of the plot indicate a BMC estimation which was only possible for the standard protocol, and similarly, a BMC value on the top of the plot area indicate a BMC estimation that could only be established for the alternative protocol. Data sets for which none of the protocols were able to produce a BMC were excluded. Color-coded symbols refer to the 22 bioassay endpoints, and a data point on (or close to) the solid line of identity indicates a perfect agreement between the BMCs from both protocols. Three-fold BMC differences are highlighted by a belt around the line of identity (i.e., values outside of the belt have above three-fold change), and the percentage number of successful regression fits for the alternative protocol are included on top of each plot, with reference to the 1953 data sets for which a successful regression modelling was conducted according to the standard protocol. To identify general deviation patterns, we performed trend regression analyses between the relative BMCs, and the corresponding value of the goodness-of-fit criterion (R^2) is provided in the plot: the higher the coefficient, the more consistent the results between the two protocols. For the trend analysis, we set a relative BMC = 2.47 for a BMRnr, i.e. a 3-fold difference between the highest concentration and a fictional BMC was assumed. Not shown are BMC differences for the bootstrapping and delta method, as both refer to the same BMC and thus would have resulted always into identical BMCs in the plot.

We found the most profound BMC differences between the data re-normalization and control normalization (Fig. 4B), with an R^2 of 0.3. The main reason for the huge number of BMC disagreements is due to huge number of BMRnr’s, i.e., regression fits that could establish a reliable BMC for the endpoint-specific BMR in only one of the protocols. Using the mean as replicate average instead of the median (Fig. 4A), using a predefined regression model (LL3rm) instead of the best fit method (Fig. 4C), and using a higher BMR resulted in moderate BMC changes, with R^2’s between 0.59-0.61. The best agreement between relative BMC values was observed for the comparison between the outcomes from model averaging against the best-fit method (Fig. 4D) with an R^2 of 0.85. The number of datasets for which a regression model could be fitted for the alternative protocol was related to the number of fits for the standard protocol and expressed as relative “fit success rate”. All changes of statistical methods lead to similar success rates, with the exception of the sole application of the three-parameter log-logistic model which led to a noteworthy loss of successful regression fits (68.55% success rate).

To further explore differences between BMC estimates, the number of BMRnr cases that only occurred in the alternative protocol (i.e. the standard protocol did result in a BMC while the alternative protocol did not; Fig. 4F, blue shaded area of bar), the number of BMRnr cases that only turned out in the standard protocol (Fig. 4F, green shaded area of bar) and

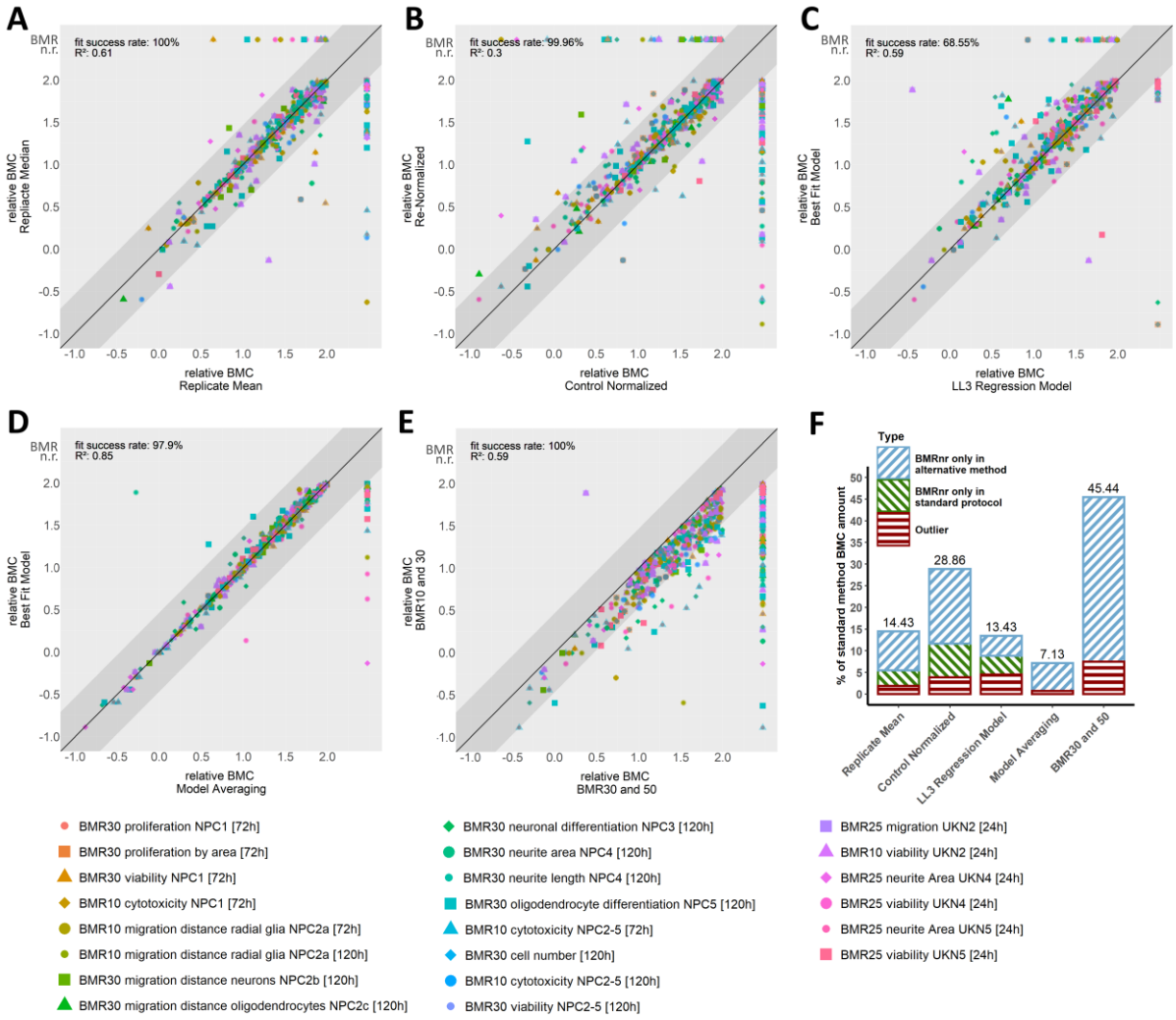

Fig. 4: Impact of methodological changes in the data evaluation on the BMC estimation
BMCs for 148 compounds tested on up to 22 endpoints from 8 assays were estimated using the standard protocol and opposing alternative methods. A-E): A relative BMC was expressed as the log10-transformed ratio between the 100-fold BMC and the maximum test concentration, and relative BMCs from all data sets and endpoints but different statistical methods were plotted against each other. The solid black line of identity indicates no differences between the relative BMCs, the grey interval around the line of identity indicates values within a three-fold range. Values outside this interval are considered as relevantly different between the opposing methods. If a relative BMC could be calculated for only one method, the missing value of the opposing method is plotted as BMRnr area on the right or upper side of the graph. Relative BMCs are colored according to their bioassay endpoint. To indicate the strength of agreement between both data evaluation protocol, the goodness-of-fit coefficient from a trend regression analysis between both relative BMCs is included (R^2; top left), and the percentage of successfully applied regression models of the alternative protocol in relation to the standard protocol is shown top left ("fit success rate"). A) Experimental median replicates versus mean replicates (n = 568). B) Re-normalized data versus control-normalized data (n = 630). C) Best fit approach versus a predefined three parameter log-logistic regression model (n = 520). D) Inverse regression versus model averaging (n = 604). E) BMR10+30 (BMR10+25 for UKN) versus BMR30+50 (n = 604). F) Percentage of all data sets for which the protocol change lead to a BMC change in terms of BMRnr (i.e. a BMC could not be determined from the regression fit) or an above three-fold BMC change.

large differences outside the belt ("outliers", Fig. 4F red shaded area of bar) were compared to the total number of BMCs that were estimated by the standard protocol. Most protocols that lead to less successful BMCs were caused by the inability of the data to support the regression modelling for the intended BMR level. All alternative protocols together led to less BMCs but more BMRnr cases, with protocol changes to control normalization and higher BMRs resulting into the highest increase towards BMnr cases (i.e. less BMCs), with an increase of 17.25% and 37.98% of BMRnr cases, respectively. Taking only the cases with huge BMC differences into account ("outliers"), the number of BMCs that were either lost or gained due to the protocol change was further quantified: model averaging led to the smallest number of relevant changes (7.13%), followed by replicate averaging by mean, fixed regression model (LL3rm) and control-normalization with moderate changes (13.43%-28.86%), up to >40% changes were reached if a higher BMR was used. More details on how changes to the standard evaluation protocol influenced the overall uncertainty of the BMC estimation are provided in the supplementary results (Fig. S2[1]).

The impact of each methodological change on the BMCL (lower 5% confidence interval of the BMC) was assessed as ratio between the BMCLs from the changed to the standard protocol, with the ratio distributions summarized in Figure 5.

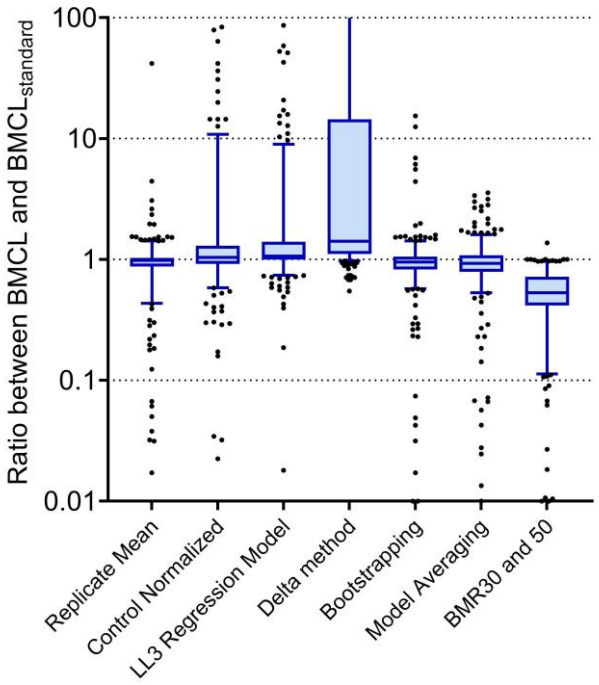

Fig. 5: Impact of methodological changes on the BMCL for DNT-specific assay endpoints
BMCLs for 148 compounds tested on up to 14 DNT-specific endpoints from 8 assays were estimated using the standard protocol and opposing alternative methods. BMCL differences are expressed as ratio (black dots), with variation shown by box and whisker plot (Box = median ± interquartile range, Whisker = 5-95 percentiles)

With the exception of the delta approximation and an increased BMR, the methodological changes provided in average no systematic deviation towards larger or smaller BMCLs, with the occurrence of 10 times larger or smaller BMCLs well below 1%. The delta approximation led in more than 90% of all data sets (N=409) to a larger BMCL (Median= 1.42, P95= 271.4), and increasing the BMR by 20% (BMR30 and 50) lead for 99.5% of all data sets to a smaller BMCL. The latter indicates a higher statistical certainty of the BMC estimate with increasing BMR, however, at cost of less data sets for which a sufficient data support was given to allow a BMC estimation (Figure 4E).

3.2 Method impact on hazard classification

An important application of the BMC estimation is the endpoint-specific hazard classification of the test compound into one of five hit categories, i.e. if the compound produced sufficient data evidence to be judged as a DNT-specific hit, borderline hit, unspecific hit, no hit, or as not identifiable (due to missing data support). Although all decision trees were setup as automatic systems, some data scenario provided insufficient data and were flagged for an expert judgement. The number of data scenarios for which the hazard classification was performed by "expert judgement" are listed in Table 1 for the standard protocol and seven methodological changes, divided according to the main decision trees developed for data from NPC or UKN assays. In total, 1563 classifications were conducted (NPC: 1347, UKN: 216), of which 68 (NPC: 30, UKN: 38) were flagged for an expert judgement according to the standard protocol. All protocol changes led to similar numbers, with the exception of the delta method applied to data outcomes from NPC assay endpoints which required expert input for three-times more classifications. A marked difference was observed between the decision trees for NPC and UKN assay endpoints, with up to 5 times more classifications flagged for expert judgement for UKN outcomes depending on the statistical method chosen. The main reason for this discrepancy is how the two classification trees dealt with data sets where the highest effect responses from an unspecific endpoint were below the BMR: these were always marked for an expert judgement in the UKN classification tree, but more thoroughly checked by automatized algorithms in the NPC classification model before judged for expert judgement (Fig. 2 and 3).

Due to the poor performance of the delta method on the BMC uncertainty (Figure 5) and the consequence of a more likely expert intervention in the automatic hazard classification (Table 1), we judged this method as too unreliable and thus excluded it from all remaining analyses.

Exemplary data sets are shown for three different classification scenarios: (i) a specific DNT hit decision for a significantly inhibited oligodendrocyte differentiation at exposure concentrations above 0.25 µM, but only a marginally reduced cell viability (marker for cytotoxicity) at 20 fold higher concentrations (Figure 6A), (ii) an unspecific hit decision for a significantly inhibited oligodendrocyte differentiation and cytotoxicity observed at same concentration ranges (0.24 to 2.2 µM) (Figure 6B), and (iii) a data scenario which was flagged for an expert judgement because for the specific endpoint a weak but statistically significant effect reduction at highest test concentration (20 µM) was observed which was not supported by a reliable BMC10 estimation. On closer inspection of the experimental data (Figure 6C, with each color-coded symbol representing the replicate median from an independent experiment) it was decided that responses from both the specific and unspecific endpoint were not distinguishable, and thus the weak response reduction of the specific endpoint was classified as unspecific.

Figure 6D provides an overview about the total number of hit classifications that changed in response to changes of the standard protocol. Expressed as percentages and for each methodological change, the changes of hit classifications are further divided in "gains", i.e. the percentual increase of hazard hits in relation to the standard protocol, and "losses", i.e. the percentual decrease of hazard hits in relation to the standard protocol. Here a change toward replicate averaging by mean, control normalization and bootstrapping caused the lowest number of classification changes (<5%), followed by

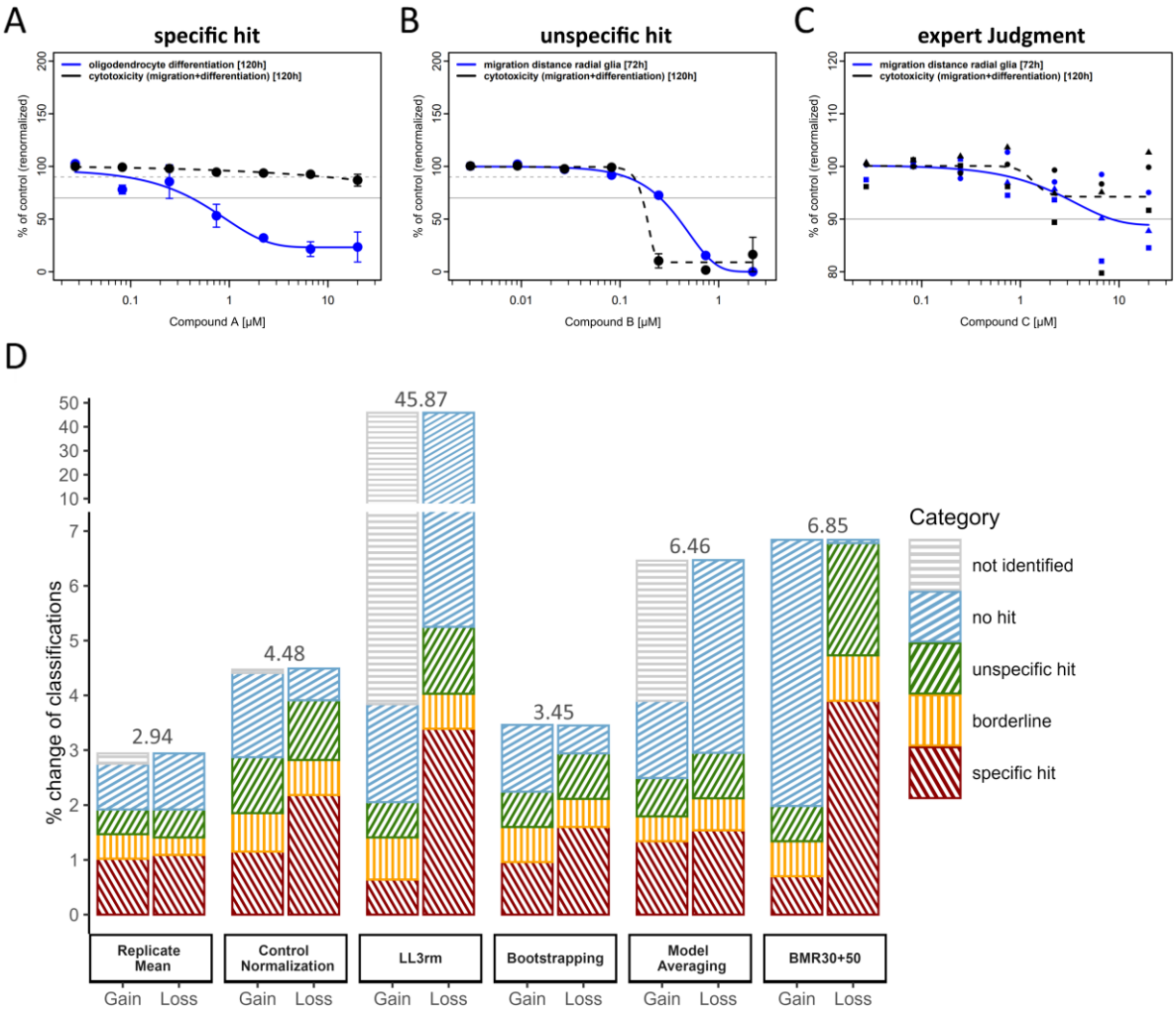

Fig. 6: Number of endpoint-specific DNT hit classification changes in response to changes in the standard data evaluation protocol
A-C) Exemplary data sets for three different classification scenarios: concentration-response data from the specific (blue) and unspecific (black) endpoints are from 5 independent experiments, with effect responses re-normalized to the regression estimate at lowest test concentration and summarized as mean±SEM. Horizontal lines indicate the BMR levels for the BMC estimation, where straight lines indicate the specific endpoint BMR and dotted lines the unspecific endpoint BMR (if they differ). Data were always analyzed according to the standard data evaluation protocol A) Specific hit: the specific endpoint (oligodendrocyte differentiation) is impacted at non-toxic concentrations. B) Unspecific hit: inhibition of oligodendrocyte differentiation and cell viability are observed at similar concentration ranges. C) Hit classification by expert judgement: an automatic hit classification was prevented by ambiguous data, but judged as "unspecific" by experts. D) For each methodological change to the standard protocol, the number of hit changes is expressed as percentage of the total number of hit classifications, divided into in "gains" (i.e. the percentual increase of hazard hits in relation to the standard protocol) and "losses" (i.e. the percentual decrease of hazard hits in relation to the standard protocol). Different bar segments represent the different classification categories.

methodological changes towards model averaging or higher BMR levels which led to almost 7% different hit classifications. Here, model averaging increased the number of "not identified" classifications by 2.56%, mostly at the cost of "no hit" classifications, and the higher BMR levels led to 4.86% more "no hit" classifications. The by far most severe changes of hit classifications were observed if only the LL3rm regression model was used to describe the experimental concentration response data (45.87% total difference), which led to 42.03% more "not identified" classifications. The latter is most likely the consequence of unsuccessful regression modelling (and corresponding BMC estimation) due to lack of sufficient data support for this model (see 3.1 and 3.2).

3.3 Assay performance

To assess how changes in the data evaluation protocol might impact the evaluation of the DNT IVB's predictivity, 28 reference chemicals of known DNT and 17 negative control chemicals were selected (Masjosthusmann et al., 2020, Blum et al., 2022), with all 45 substances tested in the DNT IVB, and the overall performance of the DNT IVB was quantified by its specificity, sensitivity and accuracy. Outcomes are shown for the standard protocol as well as all relevant changes in Figure 7: (i) Specificity (Fig. 7A): standard protocol and changes of it led always to a specificity between 87.5% and 100%, i.e. a truly DNT negative substances were almost always also judged as negative by the DNT IVB, and the standard protocol seems to be robust against methodological changes in judging false-negatives. (ii) Sensitivity (Fig. 7B): 23 of the 28 DNT substances (82.1 %) were

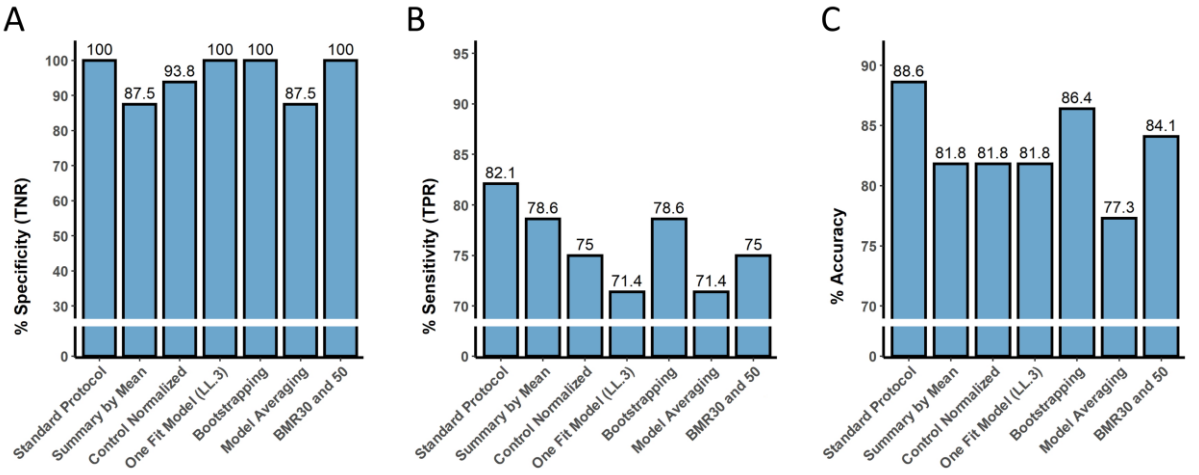

Fig. 7: Evaluation of the predictive performance of the DNT IVB based on the standard data evaluation protocol and changes
Bar graphs show the results of the predictive capability of the DNT IVB for 28 substances of known DNT and 17 negative control substances in terms of specificity, sensitivity and accuracy.

successfully identified by the DNT IVB if the standard protocol was used, but changes to it led always to a lower sensitivity. (iii) Accuracy (Fig. 7C): The best performance was achieved for the standard protocol (88.6%), followed by a methodological change to bootstrapping (86.4%), higher BMR levels (84.1%), mean replicates, control normalization, pre-defined regression model (all 81.8%) and model averaging (77.3%). The latter performed 11.3% below the accuracy value of the standard protocol. A detailed overview over the hit definition of all control compounds is given in supplementary segment 2.1 (Tab. S6-S8[1]).

4 Discussion

The basis for this biostatistical study is a compound screening project performed on behalf of an EFSA procurement during the years 2017-2020 (OC/EFSA/PRAS/2017/01). Twelve DNT *in vitro* test methods with accompanying cytotoxicity and viability assays belonging to an OECD DNT IVB (Crofton and Mundy, 2021) were challenged with 148 compounds from different compound classes including expected negative control compounds (Masjosthusmann et al., 2020, Blum et al., 2022). DNT-specific assays endpoints are capable of assessing biologically more complex systems like changes in key processes in brain development over time, and as consequence their test outcomes are more diverse in terms of data variability and concentration-response pattern than observed typically for data from simple reporter gene assays. Here we tested the hypothesis if the selection of common biostatistical concentration-response methods can affect the performance of the DNT IVB, with the focus on the BMC estimation, the DNT hit classification and its overall predictive performance for DNT adversity. All study outcomes are discussed at given experimental design and data, that is a data set is considered as pooled data from at least three independent experiments.

4.1 Experimental mean or median replicate

Altogether, our study outcomes strengthen the argument for using the median as average response per test concentration and experiment (replicate average) as statistical unit in the concentration-response regression analysis. The alternative use of the arithmetic mean led in average only to minor changes in BMC estimations and hazard classification outcomes, which most likely refer to those data sets where either no outliers were present or outliers occurred at concentrations with only little influence on the regression analysis. Nevertheless, for a few data sets a decision towards the median or mean had a strong influence on the best-fit regression analysis such that, at worst case, the subsequent BMC estimation was prevented ("BMRnr") and performance parameters (specificity, sensitivity, accuracy) for identifying DNT adversity were lowered. If effect data follow a symmetric distribution (or can be transformed accordingly) and are not affected by an outlier, the sample mean is known to be more efficient than the median, e.g., for normal distributed data the relative standard error of the median is ca. 25% greater than the standard error of the mean (Gerrard et al., 2010). However, to protect the mean against an outlier would not only require outlier detection methods, which are *per se* problematic for assay endpoints of relatively high within-experimental variability (such as DNT IVB) and small sample sizes (typically N=3), but also a decision on how to handle these values in further data analyses (e.g., removing, winsorization, or trimming). The median provides a simple way to circumvent these complex statistical decisions and can thus be considered as an ideal choice to enhance the robustness of an automatic data evaluation pipeline.

4.2 Data normalization to control and re-normalization

The normalization of effect data to the outcomes of test concentrations can be a viable option to safeguard against an ill-defined negative control reference and therefore avoid a biased BMC estimation and incorrect hazard alerts. The re-normalization of response data should be done whenever sufficient data evidence is provided for non-exposure related effect responses at lowest concentrations and which have been confirmed by independent experiments. If not justified, an existing exposure effect can be judged wrongly as technical or biological artifact and misused as zero effect response in the statistical concentration-response

analysis. Although a random explanation is theoretically possible, e.g., the control readouts were not representative and only rare "unlucky" outcomes, the confirmation by independent experiments points to a non-random, compound-specific cause. However, reasons for this phenomenon are unclear, with biological effects and technical issues discussed (Krebs et al., 2018), and unknown whether this a phenomenon specific to the DNT field or functional endpoints.

The experimental design for the assays from the DNT IVB was chosen such that the lowest test concentrations were expected to produce no treatment-related effect responses, and for more than 90% of all data sets the three lowest concentrations provided non-distinguishable effect responses. This and the frequent occurrence of misleading negative control responses for some of the assay endpoints was deemed as sufficient for using the control re-normalization on all data sets (as part of the standard protocol). We cannot fully rule out that it was wrongly used for some data sets, and more robust decision rules are required for assuring a successful data re-normalization. The criteria we recommend are (i) a certain minimum magnitude between the original and re-defined control references based on statistical reasoning (e.g., outside the detection limit), (ii) a minimum number of test low concentrations for which the effect responses provide no indications for a positive or negative trend (e.g., N=3), and (iii) a minimum concentration range at which no effect can be judged (e.g., > factor 10). However, in case a control normalization can neither be judged by an automatic ruling system nor by a human expert we suggest as last solution always a repetition of the experiment at lower test concentrations.

4.3 Concentration-response model and BMC estimation

The biological complexity of DNT assay endpoints rules out a unique mechanistic model that would allow a 100% accurate quantitative representation of every possible shape of a concentration-response pattern. Therefore, various mathematical candidate functions were used as empirical models to describe the data in the best possible way, with their number of model parameters mirroring different degrees of data complexity, and the model with the lowest number of parameters favored over a more complex model as long as it can describe the data almost as accurate as the more complex model (parsimony). The number of different concentration-response functions that were selected as best fit model suggests that not an individual regression function is capable of describing every possible data scenario, which was demonstrated at the example of the 3-parameter log-logistic function (LL3, Table S2[1]), a reparametrized form of the Hill function (but without the log10 transformation of the concentration term) which is often used as standard regression model in pharmacology and toxicology. In line with theoretical expectations and previously reported simulation studies (Zhu et al., 2007; West et al., 2012; Piegorsch et al., 2013), the best-fit model approach responded more flexible to data sets and therefore resulted often to BMC estimations that differed significantly from those derived by this pre-fixed single model.

For 1 out of 3 data sets, the two-parameter exponential function was chosen as best fit model, indicating a relatively low model complexity necessary to describe the observed data pattern, and which corresponded typically to a DNT data set where only the two highest concentrations had produced significant effects. Therefore, the simplest mathematical functions (e.g., exponential and linear) should always be included in the pool of candidate models in order to ensure a BMC estimation also for data sets with only a limited data support for the regression modelling. It should be noted that the best-fit approach is purely data-driven and therefore different regressing models can be chosen for the same assay endpoint, but as a data set is defined as pooled from independently repeated experiments, it avoids that different models are used for the same test compound and bioassay endpoint.

Model averaging is historically motivated by the typically small number of doses in animal studies that can provide meaningful data for the regression modeling, and the subsequent problem that different regression models can describe the observed dose-response data equally well but interpolation into a dose region with little or no data may result into very different response (and BMD) estimates (EFSA, 2017). A statistical argument in favor of model averaging is that uncertainty of the model selection process of the best fit method is not incorporated in the BMD and associated BMDL estimation (West et al., 2012). Our study shows no big differences between both methods, and we attribute the higher number of failed BMC estimates for model averaging (Figure 4D) due to the fact that the simple exponential function was excluded from the pool of candidate models for model averaging, but proved to be superior for data sets where maximally two (or less) concentrations responded with significant but often weak assay responses. Due to statistical reasoning, model averaging should be the preferred approach, however, the corresponding inference can only be expressed either by a Wald type of confidence interval, which can produce a negative value for the BMDL, or by the use of computer-intensive parametric bootstrap (Aerts et al., 2020), which in our study too often failed for "poor" data scenarios.

We used various statistical methods which are implemented in the drc and bmd R package to derive a confidence belt around the BMC, which is required for the BMCL (5% percentile of the lower one-sided CI) or, in case of the hazard classification, the BLL and BUL (2.5% and 97.5% percentile of the two-sided CI). It should be noted that these methods (delta approximation, inverse regression, resampling methods) do not change the BMC estimation but try to calculate the uncertainty of the BMC estimation from the estimated regression model and experimental data. All methods have their pros and cons with different requirements to the data and regression models, and none of them can *a priori* be ruled out as inappropriate for the BMC estimation of a DNT IVB data set. Only further computer-intensive simulation studies could reveal the potential bias and coverage of the estimators at given data and model complexity, but this was not the aim of our comparative study. Nevertheless, our results show that simple common statistical methods such as the delta method do not necessarily guarantee a reliable estimation about the BMC confidence, and more sophisticated methods such as resampling require data support which is often not given by the experiments. For instance, the bootstrap method puts a strong emphasis on the representativeness of the original data set from which repeatedly samples are drawn (virtual data sets), and if violated, can be prone to a biased interval estimation (i.e., mode of the resampled BMC distribution differs from the original BMC estimation), or, in worst-case, led to an interval that hardly mirrors the observed data variability. Typically, DNT IVB endpoints are characterized by a relatively high between-study variability (illustrated by the BMRs, Table S1[1]), small sample sizes (3-5 experimental medians) and a small number of test concentrations at which significant responses were observed. If all these data characteristics come together, regression resampling had a high chance for failure, independently whether non-parametric, residual or parametric sampling were used. Until generally applicable decision rules about the minimal data requirements for bootstrapping can be implemented in an

automatic data evaluation platform, it is only difficult for the non-expert to make decisions about the usefulness of resampling for a particular data scenario, especially as the bootstrapping can be performed in various ways.

4.4 BMR selection

The BMR level should be chosen as close as possible to the control level without compromising the statistical concentration-response analysis. Setting it too high (e.g., 50% for an IC50) involves the danger of overlooking hazard responses which can lead to erroneous hazard hit classifications. We used the 1.5 sigma rule for the selection of the BMR, with sigma estimated as standard deviation from the between-experimental variation from a large set of historical data sets (Masjosthusmann et al., 2020). For a sample size of 3-5 independent experiments, we expected for the majority of data sets the estimation of a BMC if a true BMC was present in the data, but an unexpected high data variability might have contradicted the 1.5 sigma rule such that the BMR was too low for a BMC estimation. If the scatter between the experimental replicate medians always followed the Gaussian distribution, we expect this to be the case in less than 1% of all cases.

4.5 Hazard identification

An endpoint-driven hazard classification method is essential for a reliable identification of hazard alerts, and DNT-specific endpoints should always take general cell health into account. However, neither a clear biological rationale on how to differentiate DNT cell functionality from general cell heath exists nor a universally agreed quantitative approach on how to distinguish cytotoxic from DNT relevant concentrations. Our proposed hazard classification method puts a strong emphasize on cell health before declaring a test compound as a specific DNT hit, by allowing only a small fraction of the central 95% CI of the DNT BMC to be overlapped by the central 95% CI of the BMC for cell health (Figure 2). Here a BUL corresponds to a 2.5% percentile of the central 95% CI, which for cell health data sets with a relatively high data variability could mean more than a factor of 10 between the BMC and its BUL. This rigorous way of accounting for statistical uncertainty in the BMC estimates has the consequence that test compounds with a weak DNT activity (which are often close to cytotoxic concentration ranges) are more likely to be classified as "borderline hit". This should not be confused with an unspecific hit, especially as a mildly affected cell viability does not necessarily have a measurable impact on cell functionality. Nevertheless, defining the desired degree of statistical uncertainty in hazard classification methods requires a general acceptance in the DNT field, and other approaches might be in the same way viable.

4.6 Conclusion

Our study on 148 compounds which were tested in a large number of DNT *in vitro* assays demonstrates that statistical decisions which seem to be of minor importance can become decisive if it comes to the hazard classification of a test substance. Although this study was conducted on concentration response data from only the DNT IVB, we think many of the conclusions can be generalized to data from other specific toxicological endpoints, especially in the rising field of organotypic/stem cell-based cultures. To our experience, the proposed standard protocol is for an automated data evaluation pipeline the best compromise between the various statistical methods without "overcomplicating" the regression analysis and the corresponding BMC estimation, but we also acknowledge that the selected methods are not necessarily optimal for every data set. The drawback of an automated analysis is always the danger of not being prepared to deal with an unexpected data set, a scenario that can only be avoided by a case-by-case expert analysis. The strength of our data evaluation platform is the integration of endpoint-specific hazard classifications, including flagging systems for uncertain cases, which to our knowledge is novel. We consider it crucial for the hazard assessment to differentiate between general cell toxicity and specific DNT hits.

References

Aerts, M., Wheeler, M. W., Abrahantes, J. C. (2020). An extended and unified modeling framework for benchmark dose estimation for both continuous and binary data. *Environmetrics 31*, e2630. doi:10.1002/env.2630

Blum, J., Masjosthusmann, S., Bartmann, K. et al. (2022). Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals. *Chemosphere 311*, 137035. doi:10.1016/j.chemosphere.2022.137035

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica 47*, 1287–1294. doi:10.2307/1911963

Buckley, B. E., Piegorsch, W. W., West, R. W. (2009). Confidence limits on one-stage model parameters in benchmark risk assessment. *Environ Ecol Stat 16*, 53–62. doi:10.1007/s10651-007-0076-2

Claeskens, G. and Hjort, N. (2008). *Model selection and model averaging*. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511790485

Cox, C. (1990). Fieller's theorem, the Likelihood and the delta method. *Biometrics 46*, 709–18. doi:10.2307/2532090

Crofton, K. M., Mundy, W. R. (2021). External Scientific Report on the Interpretation of Data from the Developmental Neurotoxicity *In Vitro* Testing Assays for Use in Integrated Approaches for Testing and Assessment. *EFSA supporting publication*; 18(10):EN-6924. 42 pp. doi:10.2903/sp.efsa.2021.EN-6924

Crump, K. S. (1995). Calculation of benchmark doses from continuous data. *Risk Anal 15*, 79-89, doi:10.1111/j.1539-6924.1995.tb00095.x

Dent, M. P., Vaillancourt, E., Thomas, R. S. et al. (2021). Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients. *Regul Toxicol Pharmacol 125*, 105026. doi:10.1016/j.yrtph.2021.105026

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press. doi:10.1201/9780429246593

EFSA Scientific Committee, Hardy, A., Benford, D. et al. (2017). Update: Guidance on the use of the benchmark dose approach in risk assessment. *EFSA J 15*, 4658. doi:10.2903/j.efsa.2017.4658

Fang, Q., Piegorsch, W. W., and Barnes, K. Y. (2015). Bayesian benchmark dose analysis. *Environmetrics* 26, 373– 382. doi:10.1002/env.2339

Förster, N., Butke, J., Keßel, H. E. et al. (2022). Reliable identification and quantification of neural cells in microscopic images of neurospheres. *Cytometry 101*, 411-422. doi:10.1002/cyto.a.24514

Gerrard, P., Maindonald, J. and Braun, W. J. (2010). *Data Analysis and Graphics Using R: An Example Based Approach*. Cambridge Series in Statistical and Probabilistic Mathematics. Third Edition, *Psychometrika 78*, 856-857. doi:10.1007/s11336-013-9349-x

Jensen, S. M., Kluxen, F. M., Ritz, C. (2019). A review of recent advances in benchmark dose methodology. Risk Anal 39, 2295-2315. doi:10.1111/risa.13324

Jensen, S. M., Kluxen, F. M., Streibig, J. C. et al. (2020). *bmd*: an R package for benchmark dose estimation. *PeerJ 8*, e10557. doi:10.7717/peerj.10557

Krebs, A., Nyffeler, J., Karreman, C. et al. (2020). Determination of benchmark concentrations and their statistical uncertainty for cytotoxicity test data and functional *in vitro* assays, *ALTEX 37*, 155–163. doi:10.14573/altex.1912021

Krebs, A., Nyffeler, J., Rahnenführer, J., & Leist, M. (2018). Normalization of data for viability and relative cell function curves. *ALTEX 35*, 268–271. doi:10.14573/1803231

Leist, M., Hasiwa, N., Rovida, C. et al. (2014) Consensus report on the future of animal-free systemic toxicity testing. *ALTEX 31*, 341–356. doi:10.14573/altex.1406091

Leontaridou, M., Urbisch, D., Kolle, S. N. et al. (2017). The borderline range of toxicological methods: Quantification and implications for evaluating precision. *ALTEX 34*, 525–538. doi:10.14573/altex.1606271

Li, H., Yuan, H., Middleton, A. et al. (2021). Next generation risk assessment (NGRA): Bridging *in vitro* points-of-departure to human safety assessment using physiologically-based kinetic (PBK) modelling – A case study of doxorubicin with dose metrics considerations. *Toxicol in Vitro 74*, 105171. doi:10.1016/j.tiv.2021.105171

Masjosthusmann, S., Blum, J., Bartmann, K. et al. (2020). Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. *EFSA supporting publication 17*, EN-1938. 152 pp. doi:10.2903/sp.efsa.2020.EN-1938

Moerbeek, M., Piersma, A. H., Slob, W. (2004). A comparison of three methods for calculating confidence intervals for the benchmark dose. *Risk Anal 24*, 31-40. doi:10.1111/j.0272-4332.2004.00409.x

OECD (2006). *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A guidance to application (annexes to this publication exist as a separate document)*. OECD Series on Testing and Assessment, No. 54, OECD Publishing, Paris. doi:10.1787/9789264085275-en.

Pallocca, G., Moné, M. J., Kamp, H. et al. (2022). Next-generation risk assessment of chemicals – Rolling out a human-centric testing strategy to drive 3R implementation: The RISK-HUNT3R project perspective, *ALTEX 39*, 419–426. doi:10.14573/altex.2204051

Piegorsch, W. W., An, L., Wickens, A. A. et al. (2013), Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics 24*, 143-157. doi:10.1002/env.2201

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. *Infect Dis Model 5*, 111-128. doi:10.1016/j.idm.2019.12.010

Randles, R. H., Fligner, M. A., Policello, G. E. II, Wolfe, D. A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *J Am Stat Assoc 75*, 168-172, doi:10.1080/01621459.1980.10477448

Ritz, C., Baty, F., Streibig, J. C., Gerhard, D. (2015) Dose-response analysis using R. *PLoS ONE 10*, e0146021. doi:10.1371/journal.pone.0146021

Ritz. C., Gerhard, D., Hothorn, L. A. (2013). A unified framework for benchmark dose estimation applied to mixed models and model averaging. *Statistics in Biopharmaceutical Research 5*, 79–90; doi:10.1080/19466315.2012.757559

Ritz, C., Jensen, S. M., Gerhard, D., Streibig, J. C. (2019). *Dose–response analysis using R*. Boca Raton: Chapman & Hall, Pages 226, doi:10.1201/b21966

Sand, S., Parham, F., Portier, C. J. et al. (2017). Comparison of points of departure for health risk assessment based on high-throughput screening data. *Environ Health Perspect 125*, 623–633; doi:10.1289/EHP408

Schmuck, M. R., Temme, T., Dach, K. et al. (2017). Omnisphero: a high-content image analysis (HCA) approach for phenotypic developmental neurotoxicity (DNT) screenings of organoid neurosphere cultures *in vitro*. *Arch Toxicol 91*, 2017-2028. doi:10.1007/s00204-016-1852-2

Scholze, M., Boedeker, W., Faust, M. et al. (2001), A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. *Environ Toxicol Chem 20*, 448-457. doi:10.1002/etc.5620200228

Villeneuve, D. L., Coady, K., Escher, B. I. et al. (2019). 38, 12-26. doi:10.1002/etc.4315

West, R. W., Piegorsch, W. W., Peña, E. A. et al. (2012), The impact of model uncertainty on benchmark dose estimation. *Environmetrics 23*, 706-716. doi:10.1002/env.2180

Wheeler, M. W., Park, R. M., Bailer A. J., Whittaker, C., (2015). Historical context and recent advances in exposure-response estimation for deriving occupational exposure limits. *J Occup Environ Hyg 12* Supp1, S7-S17. doi:10.1080/15459624.2015.1076934

Yandell, B. S. (1997). *Practical data analysis for designed experiments* (1st ed.). Routledge. doi:10.1201/9780203742563

Zhu, Y., Wang, T. and Jelsovsky, J. Z. (2007). Bootstrap estimation of benchmark doses and confidence limits with clustered quantal data. *Risk Analysis 27*, 447-465. doi:10.1111/j.1539-6924.2007.00897.x

Conflict of interest

Kristina Bartmann, Arif Dönmez, Ellen Fritsche and Axel Mosig are co-founders of the start-up company DNTOX.

Data availability

The raw data of this study was submitted to the US-EPA and published via the ToxCast database assessable via this link: https://clowder.edap-cluster.com/spaces/62bb560ee4b07abf29f88fef

Author contribution

All authors read, commented, and approved the manuscript. **Hagen Eike Keßel**: study conception, data analysis, software development, supervision, figure design, writing of article. **Stefan Masjosthusmann**: study conception, figure design, supervision. **Kristina Bartmann**: investigation. **Jonathan Blum**: investigation. **Arif Dönmez**: software development, data analysis. **Nils Förster**: software development, data analysis. **Jördis Klose**: investigation. **Axel Mosig**: software development, supervision. **Melanie Pahl**: investigation. **Marcel Leist**: supervision. **Martin Scholze**: supervision, software development, data analysis, writing of article. **Ellen Fritsche**: study conception, supervision, funding acquisition, project administration, writing of article.

Acknowledgements

The authors are grateful to Katie Paul Friedmann (US EPA) for data integration and transfer to the ToxCast data base. We would like to thank Signe Marie Jensen (University of Copenhagen) for her help with proper application of the *drc* and *bmd* R packages for data analysis. This work was supported by the European Food Safety Authority (EFSA- Q - 2018 – 00308), the Danish Environmental Protection Agency (EPA) under the grant number MST-667-00205 and the project CERST (Center for Alternatives to Animal Testing) of the Ministry for culture and science of the state North-Rhine Westphalia, Germany (file number 233- 1.08.03.03- 121972/131 – 1.08.03.03 – 121972). Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 417677437/GRK2578.