Bench Marks!

Determination of Benchmark Concentrations and Their Statistical Uncertainty for Cytotoxicity Test Data and Functional *In Vitro* Assays

Alice Krebs[1,2,#], Johanna Nyffeler[1,3#], Christiaan Karreman[1], Béla Z. Schmidt[4], Franziska Kappenberg[5], Jan Mellert[6], Giorgia Pallocca[7], Manuel Pastor[8], Jörg Rahnenführer[5] and Marcel Leist[1,7]

[1]In vitro Toxicology and Biomedicine, Department inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Konstanz, Germany; [2]Konstanz Research School Chemical Biology (KoRS CB), University of Konstanz, Konstanz, Germany; [3]present address: Center for Computational Toxicology & Exposure, US EPA, Research Triangle Park, NC, USA; [4]Switch Laboratory, VIB-KU Leuven Center for Brain & Disease Research, Department of Cellular and Molecular Medicine, Katholieke Universiteit Leuven, Leuven, Belgium; [5]Department of Statistics, Technical University of Dortmund, Dortmund, Germany; [6]Faculty of Business and Economics, Macroeconomics Dortmund University, Technical University of Dortmund, Dortmund, Germany; [7]CAAT-Europe, University of Konstanz, Konstanz, Germany; [8]Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Many toxicological test methods, including assays of cell viability and function, require an evaluation of concentration-response data. This often involves curve fitting, and the resulting mathematical functions are then used to determine the concentration at which a certain deviation from the control value occurs (e.g., a decrease of cell viability by 15%). Such a threshold is called the benchmark response (BMR). For a toxicological test, it is often of interest to determine the concentration of test compound at which a pre-defined BMR of, e.g., 10, 25 or 50% is reached. The concentration at which the modelled curve crosses the BMR is called the benchmark concentration (BMC). We present a user-friendly, web-based tool (BMCeasy), designed for operators without programming skills and profound statistical background, to determine BMCs and their confidence intervals. BMCeasy allows simultaneous analysis of viability plus a functional test endpoint, and it yields absolute BMCs with confidence intervals for any BMR. Besides an explanation of the algorithm underlying BMCeasy, this article also gives multiple examples of data outputs. BMCeasy was used within the EU-ToxRisk project for preparing data packages that were submitted to regulatory authorities, demonstrating the real-life applicability of the tool.

1 Introduction

Investigating the onset of toxicity of chemicals and determining their non-cytotoxic concentration ranges is a pivotal task of toxicology. The no-observed-adverse-effect-level (NOAEL) and the lowest-observed-adverse-effect-level (LOAEL) have been used as points of departure (PoD) for hazard estimation and risk assessment in classical toxicological studies. The NOAEL is the highest dose/concentration tested that shows no adverse effect in a particular experiment, while the LOAEL is the lowest dose/concentration at which an adverse effect was observed. The determination of the NOAEL or LOAEL is straightforward; however, uncertainty of these threshold values is poorly defined, and the data is strongly affected by the study design (e.g., sample size, dose

# contributed equally

Received December 2, 2019;

ALTEX 37(1), 155-163. doi:10.14573/altex.1912021

Correspondence: Marcel Leist, PhD
In vitro Toxicology and Biomedicine
Dept inaugurated by the Doerenkamp-Zbinden Foundation
University of Konstanz, Box 657, 78457 Konstanz, Germany
(Marcel.Leist@uni-konstanz.de)

selection and dose spacing) (Crump, 1984; Haber et al., 2018; Kimmel and Gaylor, 1988; Barnes and Dourson, 1988; Dourson et al., 1985).

To overcome this limitation, the benchmark dose concept was developed. It is broadly used and recommended for animal-based studies (EFSA, 2017; Davis et al., 2011; U.S. EPA, 1995, 2012), and it has also been adapted to analyze data sets from *in vitro* tests or so-called new approach methods (NAM). As concentrations are the most common dose metric for cell-based tests (Kisitu et al., 2019), benchmark concentrations (BMC), rather than benchmark doses, are relevant for new approach methods.

The BMC is the concentration at which a specific, pre-defined change in an assay endpoint occurs, and the pre-defined change is called the benchmark response (BMR) (Crump, 1984). The application of this concept is most straightforward for situations where the concentration-response behavior is monotonic and a mathematical model can be fitted to the experimental data. The BMC is determined as the concentration at which the modeled curve intersects the pre-defined BMR (Fig. 1). For example, in the case of viability as test endpoint, a BMR of 12% would define a reduction of viability from 100% to 88%. The uncertainty of the BMC can be calculated and expressed as the confidence interval (CI). The lower limit of the CI (bench mark concentration lower bound; BMCL) can be considered the highest concentration of a compound that does not affect the assay endpoint with a pre-defined confidence level. In this case, "not affecting the assay endpoint" would be interpreted as being "non-cytotoxic", and the CI would relate to the given BMR. In the example, the BMCL indicates the highest concentration at which one can be 97.5% certain that the viability does not deviate more than 12% (BMR) from that of untreated or negative control cells.

Taking the viability example above, one may conclude that all data above 88% viability mean that there is no cytotoxicity, while all data below the BMR (< 88% viability) indicate cytotoxicity. One could then conclude that all test compound concentrations that are below the BMC are non-cytotoxic (have a viability above the BMR). However, this interpretation does not account for the statistical uncertainty of the BMC, and one cannot be sure that toxicant concentrations at the BMC level are indeed non-cytotoxic. If one wants to have a certain confidence that a concentration is non-cytotoxic, it should be lower than the BMCL. Thus, the BMCL provides a better estimation of non-cytotoxic concentrations than the BMC alone. Knowing this threshold is important for choosing test concentrations for functional assays or for gene expression studies (Krug et al., 2013; Rovida et al., 2014; Nyffeler et al., 2017a,b; Waldmann et al., 2014; Rempel et al., 2015; Shinde et al., 2017; House et al., 2017). As the BMC method uses the information of the whole dataset, including the data uncertainties (i.e., variability produced by experimental errors and fluctuations of experimental parameters), it is less influenced by concentration spacing than a NOAEL, and it provides an estimate of statistical uncertainty of a point of departure (PoD). It is therefore a preferred method for modern toxicological approaches (Hartung and Leist, 2008; Leist et al., 2010, 2012, 2014), and the concept has been used for a number of algorithms to define non-cytotoxic concentrations (Stadnicka-Michalak et al., 2018; Hsieh et al., 2019; Calderazzo et al., 2019; Behl et al., 2015). Several software tools have been developed to calculate the BMC (Filer et al., 2017; Slob and Setzer, 2014; Ritz and Streibig, 2005)[1,2], but applying the BMC concept still requires considerably more mathematical/statistical and programming skills than the NOAEL/LOAEL approach and can be challenging for experimental biologists.

The ideal tool for an experimental *in vitro* toxicologist without programming skills and working on low to medium throughput test methods would have the following characteristics: (1) requires neither programming skills nor extensive mathematical and statistical knowledge; (2) handles data from cytotoxicity endpoints as well as from functional endpoints (note that the lower

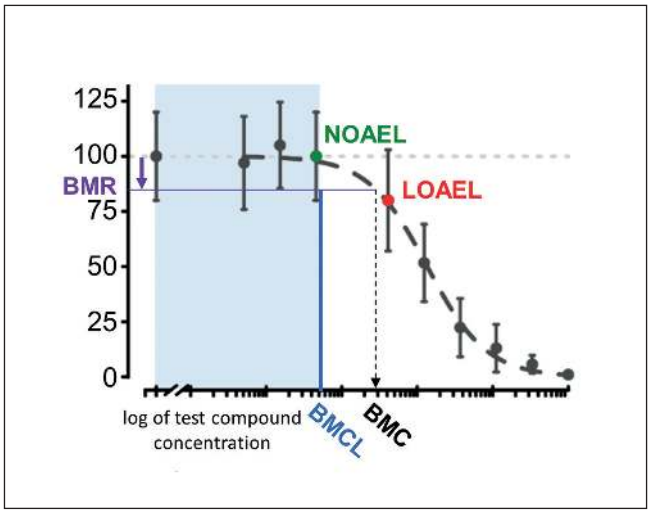

Fig. 1: The benchmark concentration concept for *in vitro* toxicology data

An example dataset of concentration-response data of a toxicant is shown. Older curve evaluation strategies use the NOAEL (maximal concentration that has no observable adverse effect) and the LOAEL (lowest concentration that evokes an adverse effect) to define the onset of toxicity. The benchmark concept requires definition of a benchmark response level (BMR). The compound concentration that causes the test endpoint (e.g., viability) to drop to the BMR is called the benchmark concentration (BMC). A two-sided confidence interval (CI) can be calculated for the BMC. The *lower* limit of the CI is defined as BMCL, the *upper* limit as BMCU. Concentrations lower than the BMCL are likely to produce effects lower than the BMR with a pre-defined level of confidence (usually 97.5%).

[1] https://www.rivm.nl/en/proast

[2] https://cran.r-project.org/web/packages/drc/drc.pdf

Abbreviations

BMC, benchmark concentration; BMCL, benchmark concentration lower limit; BMCU, benchmark concentration upper limit; BMR, benchmark response; CI, confidence interval; LOAEL, lowest observed adverse effect level; NOAEL, no observed adverse effect level

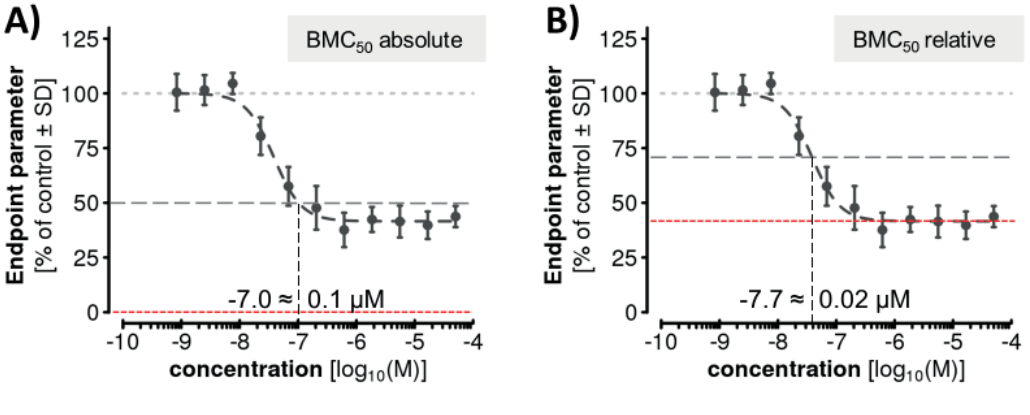

Fig. 2: Graphical representation of the absolute versus the relative BMC_{50}

A) The absolute BMC_{50} is defined independently of the data and their curve fit, the curve's inflection point or its lower asymptote. The absolute BMC_{50} is thus always defined as a 50% reduction of the respective endpoint, independent of the lower asymptote.

B) The relative BMC_{50} is dependent on the data, in particular on the lower asymptote, because it is defined by the midpoint of the difference between the upper and the lower asymptotes. For a 4-parameter sigmoid fit, this is identical to the curve's inflection point. The lower asymptote is shown in red, the upper asymptote in grey. The dashed vertical line (black) indicates the BMC_{50}. BMCeasy uses the absolute BMC_{50} definition as in A.

asymptote of functional endpoint data is often > 0); (3) calculates *absolute* BMC and BMCL values in a standardized way to allow comparison in inter-laboratory projects or between different assays in a test battery (Fig. 2), (4) ensures transparency and reproducibility by creating publication-quality graphical output and providing background on the fitting procedure as well as documenting all parameters of the curve fitted to the concentration-response data. To the best of our knowledge, no such program was easily available, therefore we created BMCeasy[3], a web-based application fulfilling all these requirements.

BMCeasy was programmed in R[4] (R Core Team, 2019) using the well-established drc package for curve fitting and BMC estimation[5] (Ritz and Streibig, 2005; Ritz, 2010) and the R Shiny package to produce the web interface[6]. The tool is suitable for data obtained from a wide variety of *in vitro* test methods. It has been used and optimized in practice during collaborative work in the EU-ToxRisk project[7] and several other international collaborative research activities involving laboratories from both academia and industry.

In the present work, we aim to give an overview of the workflow, together with some example applications of the program. A detailed user manual of the BMCeasy graphical user interface[3] and the code[8] are available.

2 Curve fitting and calculation of BMR data

Data for up to two different endpoints can be uploaded at a time, and any BMR can be specified. For both endpoints, three concentration-response models with slightly different constraints are available to fit the data. The best model is selected automatically by BMCeasy. The outputs are the BMCs, their confidence intervals with lower and upper boundaries (the BMCL and BMCU), the parameters used for curve fitting, and a high-quality graph.

2.1 Pre-processing of the data and data input

The program was designed on purpose to require a manual check and pre-processing of data: Initial normalization of raw data, detection of outliers and possible re-normalization of data sets (Krebs et al., 2018) is left to the user and needs to be done prior to data upload in standard laboratory programs such as GraphPad or Excel. It is of utmost importance that the user is aware of

[3] http://invitrotox.uni-konstanz.de/BMCeasy/
[4] https://www.r-project.org/
[5] http://www.bioassay.dk/
[6] https://shiny.rstudio.com/
[7] https://www.eu-toxrisk.eu/
[8] https://github.com/JohannaNyff/BMCeasy

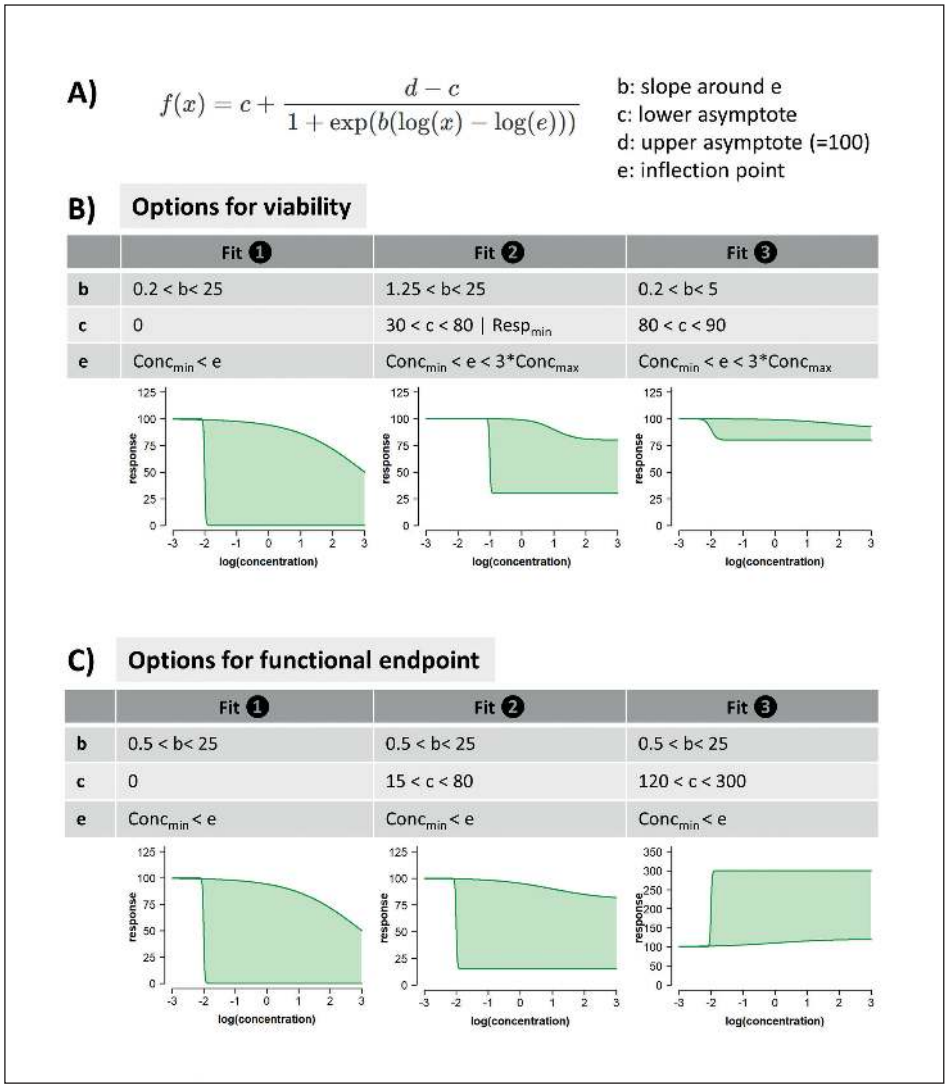

Fig. 3: Overview of the different models to fit concentration-response data
A) The log-logistic 4-parameter model on which all fits are based. B) The three different fits (based on different parameter constraints) available to fit viability endpoint data. C) The three different fits (based on different parameter constraints) available to fit functional endpoint data. The green areas are based on the respective upper and lower parameter limits. They demonstrate the full range of curves that can be generated by the model. Conc, concentration

the quality of data necessary to yield reasonable outputs. In some cases, repetitions of an experiment may be necessary to provide data sets of sufficient quality. Sometimes data sets may need to be excluded from analysis (e.g., when a compound used in a viability test makes cells proliferate (reaching values > 100%). As the software performs poorly with low-quality data and requires manual pre-processing, BMCeasy is not suitable for high-throughput data processing.

BMCeasy input datasets are simple lists containing the name of the chemical tested and, in each line, a test concentration, the concentration unit, and the response (normalized to 100% relative to a vehicle control or to another reference point). In case of several replicates, they may be entered as individual values or as mean, standard deviation, and number of replicates. A sample MS Excel file is provided[3].

BMC calculation requires datasets with a downward direction (towards 0%) for increasing concentrations of test chemical. For example, a decrease in cell viability or inhibition of cell migration are typical assay responses that can be evaluated. Upward responses (e.g., enhanced cell proliferation, or activation of a reporter gene) cannot be evaluated directly; such datasets need prior transformation (e.g., inversion or mirroring on a line parallel to the x-axis) (Weimer et al., 2012).

2.2 Curve modeling

Concentration-response input data is fitted using the log-logistic function of the drc package. The curve fits differ in their constraints and possible ranges of curve fit variables (Fig. 3A)[9]. The parameters were not derived from and validated by stringent statistical procedures. They rather reflect the experience gained

9 https://cran.r-project.org/web/packages/drc/drc.pdf

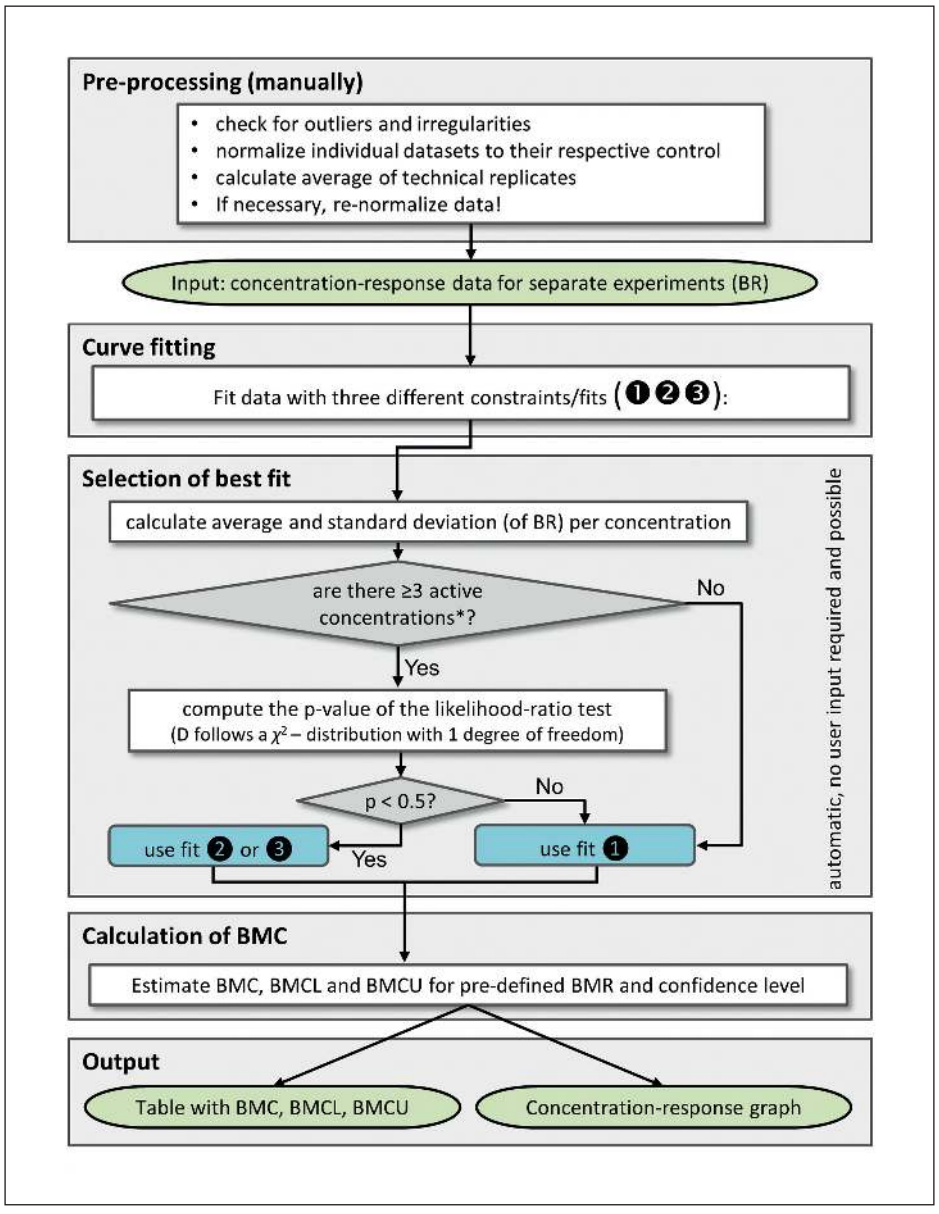

Fig. 4: Overview of data processing with the BMCeasy tool to determine the BMC and its confidence interval
The data are pre-processed manually before they are uploaded. Ideally, datasets from several independent experiments (BR, biological replicates) are used for input. Three different fits are available (❶❷❸; see also Fig. 3) to fit the concentration-response data. The program automatically identifies the most suitable concentration-response model and uses it to calculate the BMCs and the respective BMCL/BMCU.
* at least three concentrations with a mean response exceeding the noise level. The noise level was defined as 5% for viability and 10% for the functional endpoint. logL, log-likelihood

during several years of experimental work with the fitting of real-life data sets (Nyffeler et al., 2017a,b). Best results are obtained for low BMRs (e.g., 10-30%), as BMCeasy was mainly developed to determine the onset of toxicity. Accordingly, the program performance may be poor for BMRs > 50%.

For viability data (Fig. 3B), Fit 1 works best for datasets with a steep concentration-response and a lower asymptote near 0%. The parameter setting takes into account the biological background that most viability data sets should theoretically approach 0% at high test concentrations. It also accounts for frequent experimental findings that there is sometimes an imperfect baseline correction or blank subtraction. Moreover, the parameter setting takes advantage of our experience that imperfect curve fits in the lower part of the curve affect BMCs with a low BMR (often BMR10 is selected for viability data) only to a small extent. The rationale was to rather tolerate a reduced fit quality at high concentrations than to accept the biologically rather implausible assumption that exactly 8% (number chosen only as example of an exact lower asymptote) of the cells will survive even the highest test concentrations.

A second case for which Fit 1 often works well is a weak, non-sigmoidal response, like a linear decrease, e.g., to 80% viability, at high test concentrations. However, users need to be careful with data that have a very gentle slope, as they are error-prone and uncertainties can be high. This is a typical case where users need to decide whether BMC determination is meaningful. Fit 2 and Fit 3 are suitable for sigmoidal responses with a lower asymptote that is clearly higher than 0%. An example case may be rel-

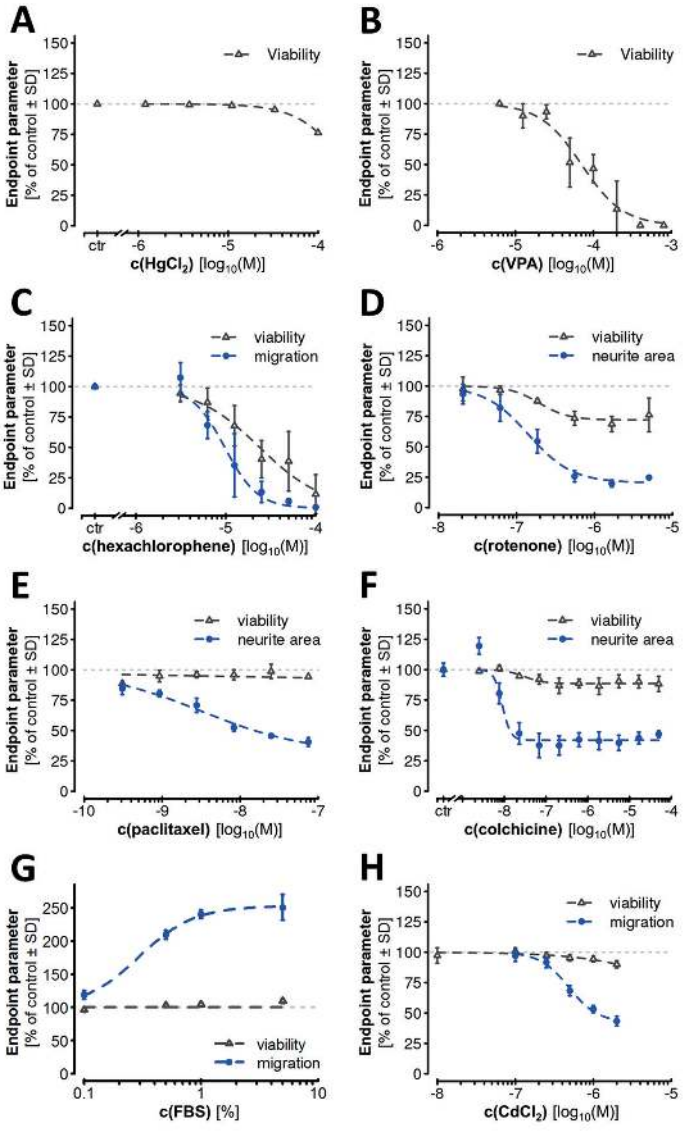

Fig. 5: Examples of real-life outputs from BMCeasy demonstrate how curves are fitted to various experimental datasets

A) Concentration-response data of mercury chloride from the PBEC-ALI assay (air-exposed primary bronchial epithelial cells). B) Percent of viable zebrafish embryos after 96 h exposure to valproic acid (VPA). C) Effect of hexachlorophene on migrating neural crest cells (cMINC assay). D) Concentration-response data of rotenone in peripheral neurons (PeriTox assay). E) Paclitaxel effects on the neurite area of peripheral neurons (PeriTox assay). F) Data of the NeuriTox assay using developing dopaminergic neurons (LUHMES cells). G) Fetal bovine serum (FBS) enhances cell migration of neural crest cells (cMINC). H) Cadmium chloride ($CdCl_2$) inhibits neural crest cell migration (cMINC). G) and H) were adopted from Nyffeler et al. (2017b). Data provided by the Leiden University Medical Center (LUMC), The Netherlands for panel A, by University of Heidelberg, Germany, for panel B, and by the University of Konstanz, Germany for panels C-H.

ative cell count measurements for a cytostatic (but not cytotoxic) compound (Fig. 3B). Fit 3 is intended to allow modelling of datasets with a lower asymptote > 80%.

For functional endpoints (Fig. 3C), Fit 1 and 2 are similar to those for cytotoxicity data, while Fit 3 was included to account for datasets with upward concentration-response curves (Fig. 3C). Notably, Fit 3 is not meant to generally capture all responses > 100% (e.g., motility, proliferation, reporter assays, etc.). It was included in the software to account for unexpected counter-regulations (e.g., cytostatic drugs that increase cell numbers; potential enzyme inhibitors that activate an enzyme, etc.).

2.3 Best model selection

Following principles of the Akaike information criterion[10], concentration-response models were compared for their goodness of fit and for the risk of overfitting. Fit 2 and 3 have three degrees of freedom, while Fit 1 has only two degrees of freedom. Models with more degrees of freedom typically fit the data better, but they may also overfit them. Therefore, their use must be justified. The selection procedure was set in a way that Fit 1 is preferred over Fit 2 or 3 unless either of these models the data considerably better. Moreover, Fit 2 and Fit 3 can be selected only if there are at least three data points on the concentration-response

[10] https://www.statisticshowto.datasciencecentral.com/akaikes-information-criterion/

Tab. 1: Advantages and limitations of the BMCeasy tool

Advantages/strengths	Limitations
Forces operator to visually check data and output	Handles only one compound at a time
Customized for continuous data from *in vitro* test methods	Only handles effects that decrease relative to the control response
Deals with two different readouts/endpoints in parallel	Sensitivity to misspelling
Calculates BMCs of different BMRs in parallel	Only ≤ 2 endpoints simultaneously
Free choice of BMR	Manual re-normalization required (see Krebs et al. 2018)
Confidence level can be chosen	Normalization to control (= 100%) has to be done manually before data entry
No need for coding and scripting	Data need to fit standard requirements
No program installation needed	Needs internet connection
High quality graphical output	No further graphic modification

curve with a mean response exceeding the noise level. The noise levels were defined empirically as 5% for viability and 10% for the functional endpoint because these automatic settings worked well in practice. We set a higher noise level (10% instead of 5%) for functional endpoints because we observed that these types of endpoints often have higher variance than viability measurements. We consider fixed thresholds important to compare different experiments within a given assay and across different assays. A different strategy, i.e., a flexible noise level calculated from negative control data, has been chosen by others (Behl et al., 2019). To decide between Fit 2 and Fit 3, their log-likelihood is determined (Fig. 4).

2.4 BMC calculation

The best fits for viability or the functional endpoint are then used to derive the BMC and to calculate the respective BMCU and BMCL at the desired BMR.

To ensure that the estimated confidence interval is of best possible data quality, it is preferable that data from as many replicates as possible are included in the analysis; the number of concentrations tested as well as their spacing will also affect the output. For calculation of the CI, the ED function within the drc package (tlfs function pre-selected) is used (Ritz and Streibig, 2005). The resulting CI is usually asymmetrical on a linear axis.

It is important to note that BMCeasy defines BMC as decline/reduction of an *absolute* effect. For example, BMC_{50} is a 50% reduction of the respective endpoint. For a curve with, e.g., an asymptote at 40%, this point does not coincide with the inflexion point of the curve (Fig. 2). This feature differs from other modelling solutions (e.g., tcpl) (Filer et al., 2017). According to the definition of the BMC underlying BMCeasy, an increase in response cannot be defined. This means that no BMC can be calculated if Fit 3 is selected for the functional endpoint.

2.5 Output

The calculated BMCs, BMCUs and BMCLs are returned in a table and can be downloaded as an Excel file. A graph with the concentration-response data of both endpoints and the respective best model fitting the data is plotted in publication quality. There are multiple options for layout and design of the graph (appearance of curves, legends, axes and data points). However, for users who prefer their own graphical representation, the parameters of the curve fit are returned. This allows reproduction in other programs.

BMCeasy has been used for processing the results of a broad variety of methods (Fig. 5). Within the EU-ToxRisk project, its use spanned from air-exposed bronchial epithelial cells, through zebrafish embryos, neural crest cells and neurons to hepatocytes and kidney cell models. The example plots show that the tool can handle various challenges, such as curves with almost no decline (Fig. 5A), with large data uncertainty (Fig. 5B, C), data with a lower asymptote > 0% (Fig. 5D, F) or curves with a very gentle decrease (Fig. 5E) as well as data with increasing response (Fig. 5G).

3 Discussion and conclusion

We created a web-based tool to help experimental *in vitro* toxicologists without programming skills or extensive mathematical and statistical knowledge to calculate absolute BMC and BMCL values in a standardized way. As BMCeasy does not contain a quality check module for the input data, it requires the user to judge which types of data sets can be handled by the program and where problems may occur (Tab. 1). It is important to note that the program will to some extent also handle "non-suitable" data, and in such cases the delivered BMC data may not be meaningful. Like any research tool, the program has an applicability domain and a working environment that must be known and respected. For instance, non-monotonic responses or no-effect datasets cannot yield a reasonable BMC. Moreover, curve shapes that are very different from those obtained in typical cell-based assays potentially will not result in good quality estimates. The tool is not suitable for high-throughput efforts, as batch processing of multiple test compounds has not been implemented. A detailed manual[3] explains how to set up the analysis.

Currently, only the log-logistic curve is modelled. It has been suggested that a broad range of functions should be available for

curve modelling (U.S. EPA, 2012; Ritz, 2010). Even solutions for non-monotonic (e.g., hormesis) effects have been developed[11,12]. A broad range of curve models is particularly important for high throughput data that are not manually controlled. The drc package used for BMCeasy has many curve models implemented, but we found that a limitation to the log-logistic function yields sufficiently good and more comparable results. Exceptions to this rule are possible and need to be excluded by visual control.

The most severe limitation of BMCeasy is that it requires that datasets are normalized to 100% (for negative controls). A typical example of data that do not work is determining the percentage of lactate dehydrogenase (LDH) released from cells. In a typical assay, negative controls may show 15% release (spontaneous cell death) and positive controls 90% release. Such datasets may require complex transformation and renormalization (i.e., expressing data in terms of lactate dehydrogenase not released: 85% for negative controls, 10% for positive controls, then normalizing the negative control to 100% by dividing all data by 0.85).

Many test methods yield two endpoints, a viability/cytotoxicity endpoint and a functional endpoint (e.g., cell migration, electrical network activity, substrate transport or secretion of a biomolecule). Often the relation of the two endpoints is used to determine acceptance criteria or to identify hits. Therefore, it is often desired to represent them in one graph and to use the same tool for BMC calculations. BMCeasy is ideal for such cases. This allows experimenters, for example, to rapidly decide if the concentration spacing was chosen appropriately or if they should repeat their experiment with different compound dilutions.

References

Barnes, D. G. and Dourson, M. (1988). Reference dose (RfD): Description and use in health risk assessments. *Regul Toxicol Pharmacol 8*, 471-486. doi:10.1016/0273-2300(88)90047-5

Behl, M., Hsieh, J. H., Shafer, T. J. et al. (2015). Use of alternative assays to identify and prioritize organophosphorus flame retardants for potential developmental and neurotoxicity. *Neurotoxicol Teratol 52*, 181-193. doi:10.1016/j.ntt.2015.09.003

Behl, M., Ryan, K., Hsieh, J. H. et al. (2019). Screening for developmental neurotoxicity at the national toxicology program: The future is here. *Toxicol Sci 167*, 6-14. doi:10.1093/toxsci/kfy278

Calderazzo, S., Tavel, D., Zurich, M. G. et al. (2019). Model-based estimation of lowest observed effect concentration from replicate experiments to identify potential biomarkers of in vitro neurotoxicity. *Arch Toxicol 93*, 2635-2644. doi:10.1007/s00204-019-02520-8

Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundam Appl Toxicol 4*, 854-871. doi:10.1016/0272-0590(84)90107-6

Davis, J. A., Gift, J. S. and Zhao, Q. J. (2011). Introduction to benchmark dose methods and US EPA's benchmark dose software (BMDS) version 2.1.1. *Toxicol Appl Pharmacol 254*, 181-191. doi:10.1016/j.taap.2010.10.016

Dourson, M. L., Hertzberg, R. C., Hartung, R. et al. (1985). Novel methods for the estimation of acceptable daily intake. *Toxicol Ind Health 1*, 23-33. doi:10.1177/074823378500100404

EFSA (2017). Update: Guidance on the use of the benchmark dose approach in risk assessment. *EFSA J 15*, 4658.

Filer, D. L., Kothiya, P., Setzer, R. W. et al. (2017). tcpl: The ToxCast pipeline for high-throughput screening data. *Bioinformatics 33*, 618-620. doi:10.1093/bioinformatics/btw680

Haber, L. T., Dourson, M. L., Allen, B. C. et al. (2018). Benchmark dose (BMD) modeling: Current practice, issues, and challenges. *Crit Rev Toxicol 48*, 387-415. doi:10.1080/10408444.2018.1430121

Hartung, T. and Leist, M. (2008). Food for thought ... On the evolution of toxicology and the phasing out of animal testing. *ALTEX 25*, 91-102. doi:10.14573/altex.2008.2.91

House, J. S., Grimm, F. A., Jima, D. D. et al. (2017). A pipeline for high-throughput concentration response modeling of gene expression for toxicogenomics. *Front Genet 8*, 168. doi:10.3389/fgene.2017.00168

Hsieh, J. H., Ryan, K., Sedykh, A. et al. (2019). Application of benchmark concentration (BMC) analysis on zebrafish data: A new perspective for quantifying toxicity in alternative animal models. *Toxicol Sci 167*, 92-104. doi:10.1093/toxsci/kfy258

Kimmel, C. A. and Gaylor, D. W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Anal 8*, 15-20. doi:10.1111/j.1539-6924.1988.tb01149.x

Kisitu, J., Hougaard Bennekou, S. and Leist, M. (2019). Chemical concentrations in cell culture compartments (C5) – Concentration definitions. *ALTEX 36*, 154-160. doi:10.14573/altex.1901031

Krug, A. K., Kolde, R., Gaspar, J. A. et al. (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: A transcriptomics approach. *Arch Toxicol 87*, 123-143. doi:10.1007/s00204-012-0967-3

Krebs, A., Nyffeler, J., Rahnenfuhrer, J. et al. (2018). Normalization of data for viability and relative cell function curves. *ALTEX 35*, 268-271. doi:10.14573/1803231

Leist, M., Efremova, L. and Karreman, C. (2010). Food for thought ... Considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317. doi:10.14573/altex.2010.4.309

Leist, M., Hasiwa, N., Daneshian, M. et al. (2012). Validation and quality control of replacement alternatives – Current status and future challenges. *Toxicol Res 1*, 8-22. doi:10.1039/c2tx20011b

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

Nyffeler, J., Dolde, X., Krebs, A. et al. (2017a). Combination of multiple neural crest migration assays to identify environmental toxicants from a proof-of-concept chemical library. *Arch Toxicol 91*, 3613-3632. doi:10.1007/s00204-017-1977-y

Nyffeler, J., Karreman, C., Leisner, H. et al. (2017b). Design of a high-throughput human neural crest cell migration assay to

[11] https://www.rdocumentation.org/packages/drc/versions/2.5-12/topics/cedergreen

[12] https://www.rdocumentation.org/packages/drc/versions/2.5-12/topics/braincousens

indicate potential developmental toxicants. *ALTEX 34*, 75-94. doi:10.14573/altex.1605031

R Core Team (2019). R: A language and environment for statistical computing. https://www.R-project.org/

Rempel, E., Hoelting, L., Waldmann, T. et al. (2015). A transcriptome-based classifier to identify developmental toxicants by stem cell testing: Design, validation and optimization for histone deacetylase inhibitors. *Arch Toxicol 89*, 1599-1618. doi:10.1007/s00204-015-1573-y

Ritz, C. and Streibig, J. C. (2005). Bioassay analysis using R. *Journal of Statistical Software 12*, 1-22. doi:10.18637/jss.v012.i05

Ritz, C. (2010). Toward a unified approach to dose-response modeling in ecotoxicology. *Environ Toxicol Chem 29*, 220-229. doi:10.1002/etc.7

Rovida, C., Vivier, M., Garthoff, B. et al. (2014). ESNATS conference – The use of human embryonic stem cells for novel toxicity testing approaches. *Altern Lab Anim 42*, 97-113. doi:10.1177/026119291404200203

Shinde, V., Hoelting, L., Srinivasan, S. P. et al. (2017). Definition of transcriptome-based indices for quantitative characterization of chemically disturbed stem cell development: Introduction of the STOP-Tox_{ukn} and STOP-Tox_{ukk} tests. *Arch Toxicol 91*, 839-864. doi:10.1007/s00204-016-1741-8

Slob, W. and Setzer, R. W. (2014). Shape and steepness of toxicological dose-response relationships of continuous endpoints. *Crit Rev Toxicol 44*, 270-297. doi:10.3109/10408444.2013.853726

Stadnicka-Michalak, J., Knobel, M., Zupanic, A. et al. (2018). A validated algorithm for selecting non-toxic chemical concentrations. *ALTEX 35*, 37-50. doi:10.14573/altex.1701231

U.S. EPA (1995). Use of the benchmark dose approach in health risk assessment. U. S. Environmental Protection Agency, Risk Assessment Forum. Washington, DC. EPA/630/R-94/007

U.S. EPA (2012). Benchmark Dose Technical Guidance. EPA/100/R-12/001

Waldmann, T., Rempel, E., Balmer, N. V. et al. (2014). Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. *Chem Res Toxicol 27*, 408-420. doi:10.1021/tx400402j

Weimer, M., Jiang, X., Ponta, O. et al. (2012). The impact of data transformations on concentration-response modeling. *Toxicol Lett 213*, 292-298. doi:10.1016/j.toxlet.2012.07.012

Conflict of interest

The authors declare no conflict of interest.

Authors' contributions

J.N. and M.L. conceived the presented idea. J.N. developed the application and C.K. made the application publicly accessible. A.K. wrote the manual and collected input and data from the consortium. B.Z.S., F.K., J.M., G.P., M.P., J.R. contributed to the figures and proofread the manuscript. A.K., J.N. and M.L. designed the figures and wrote the manuscript.

Acknowledgements

This work was supported by the BMBF, EFSA, the DK-EPA, and the DFG (Konstanz Research School of Chemical Biology; KoRS-CB). It has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 681002 (EU-ToxRisk) and No. 825759 (ENDpoiNTs).

We are grateful to collaborators R. von Hellfeld and T. Braunbeck (University of Heidelberg), and H. Vrieling and J. Boei (Leiden University Medical Center) of the EU-ToxRisk consortium for providing the experimental data. We are indebted to S. Förster for triggering this work and to other colleagues for insightful discussions.