Bioinformatics, 37(14), 2021, 1990–1996
doi: 10.1093/bioinformatics/btab043
Advance Access Publication Date: 30 January 2021
Original Paper

Gene expression

Comparison of observation-based and model-based identification of alert concentrations from concentration–expression data

Franziska Kappenberg [1,*], Marianna Grinberg [1,†], Xiaoqi Jiang [2,†], Annette Kopp-Schneider [2], Jan G. Hengstler [3] and Jörg Rahnenführer [1]

[1]Department of Statistics, TU Dortmund University, Dortmund 44221, Germany, [2]Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany and [3]Department of Toxikologie/Systemtoxikologie, Leibniz Research Centre for Working Environment and Human Factors (IfADo), TU Dortmund University, Dortmund 44139, Germany

*To whom correspondence should be addressed.
[†]Present address: Department of R&D Global Biostatistics, Epidemiology and Medical Writing, Merck KGaA, Darmstadt 64293, Germany
[†]Present address: AI Solutions—Services & Core, BASF SE, Ludwigshafen am Rhein 67061, Germany

Associate Editor: Jonathan Wren

Received on August 22, 2020; revised on January 15, 2021; editorial decision on January 18, 2021; accepted on January 25, 2021

Abstract

Motivation: An important goal of concentration–response studies in toxicology is to determine an ‘alert’ concentration where a critical level of the response variable is exceeded. In a classical observation-based approach, only measured concentrations are considered as potential alert concentrations. Alternatively, a parametric curve is fitted to the data that describes the relationship between concentration and response. For a prespecified effect level, both an absolute estimate of the alert concentration and an estimate of the lowest concentration where the effect level is exceeded significantly are of interest.

Results: In a simulation study for gene expression data, we compared the observation-based and the model-based approach for both absolute and significant exceedance of the prespecified effect level. Results show that, compared to the observation-based approach, the model-based approach overestimates the true alert concentration less often and more frequently leads to a valid estimate, especially for genes with large variance.

Availability and implementation: The code used for the simulation studies is available via the GitHub repository: https://github.com/FKappenberg/Paper-IdentificationAlertConcentrations.

Contact: kappenberg@statistik.tu-dortmund.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Concentration–response studies are applied across a wide range of fields, including pharmacology, pharmacokinetics, toxicology and clinical research. In toxicology, such studies are conducted to investigate and quantify exposure-related effects. An important goal is to determine an 'alert' concentration where a critical level of the relevant response, referred to here as a ‘threshold’, is exceeded. In gene expression studies, typically only the nominal concentrations tested are considered as potential alert concentrations. However, by fitting a parametric model to the data, concentration–response curves with a monotonic relationship between concentration and response are better suited to more precisely estimate alert concentrations, since also non-tested concentrations are potential estimates. In the following, we use the term ‘measured concentrations’ for the concentration applied to cells, in contrast to ‘estimated alert concentrations’, derived from different approaches.

Different alert concentrations can be considered: The effective concentration EC_{50} refers to the concentration that induces 50% of the maximal effect. The estimate is defined in terms of lower and upper asymptote of the fitted curve and therefore heavily depends on these values. A similar alternative is the benchmark dose (BMD) methodology, which identifies the lowest concentration with a noticeable effect compared to the normal response (Jensen *et al.*, 2019). It has also been proposed to estimate the Absolute Lowest Effective Concentration (ALEC), the concentration where a fixed and prespecified critical effect level is attained (Jiang, 2013).

A common model class for modelling concentration–response data, or more specifically concentration–gene expression data, are log-logistic models with up to four parameters. Extensions are generalized log-logistic models or hormesis models. Further model classes include log-normal and Weibull models (Ritz *et al.*, 2019, pp. 178–188). One of the most common models is the four-parametric log-logistic (4pLL) model (e.g. Ritz, 2010), with four parameters corresponding to the upper and lower asymptote, inflection point and slope of the curve. Based on the 4pLL model, we propose a statistical test to determine the lowest concentration where the response significantly exceeds the response for the control, modelled by the left-sided asymptote of the 4pLL model, by a given threshold. We denote this concentration with LEC (Lowest Effective Concentration).

Model-based alert concentrations considered here are continuous alternatives to the discrete observation-based alert concentrations ALOEC (Absolute Lowest Observed Effective Concentration) and LOEC (Lowest Observed Effective Concentration). The ALOEC is the lowest measured concentration where the mean difference between all replicates of the concentration and of the control exceeds a given threshold, while the LOEC corresponds to the lowest measured concentration where the difference significantly exceeds the threshold (e.g. Delignette-Muller *et al.*, 2011). The significance can be assessed for example with a standard two-sample *t*-test.

In this article, we first we present the 4pLL model and calculation of the ALEC with a corresponding confidence interval. We then introduce the LEC as an extension of the ALEC, taking significance into account. We propose a 4pLL-based test for significant difference in the expression of two concentrations and present a real case study where derivation of the LEC is of interest.

All four alert concentrations, obtained from the two observation-based methods (ALOEC and LOEC) derived from the classical approach and from the two model-based methods (ALEC and LEC), were compared in a simulation study. In this study, three different true underlying concentration–gene expression profiles were considered, with different standard deviations of the replicates for the measured concentrations, respectively. The alert concentrations were calculated and interpreted with respect to the number of valid estimates, under- and overestimation of the true alert concentration, and the coverage probability of the confidence interval of the ALEC. Finally, the methods were applied to the real data from a study in which the expression values for 54675 probe sets for 7 different concentrations of the compound valproic acid plus a control were measured.

2 Material and methods

2.1 Statistical methods

Based on the assumption that the relationship between concentration and response can be described by a sigmoidal curve, a four-parameter log-logistic model (4pLL) can be fitted to the data (e.g. Ritz, 2010). For a concentration x, $x \geq 0$ and a parameter vector $\phi = (\phi^{(b)}, \phi^{(c)}, \phi^{(d)}, \phi^{(e)})^\top$ with $\phi^{(e)} > 0$, the model is defined as

$$f(x, \phi) = \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \exp\{\phi^{(b)}[\log(x) - \log(\phi^{(e)})]\}}. \quad (1)$$

A frequently used re-parameterisation, which provides more accurate estimates for small datasets, is given by $\phi^{(e)^*} = \log(\phi^{(e)})$ (Ritz, 2010).

The function f describes the response (here the logarithmic expression values) as a function of the concentration x. The parameters $\phi^{(c)}$ and $\phi^{(d)}$ specify the lower and upper limit of f, respectively. The parameter $\phi^{(b)}$ is proportional to the slope of the curve at $\phi^{(e)}$, the half-maximal effective concentration. This concentration is typically called EC_{50}, the concentration that induces 50% of the maximal effect. Different parameterizations of the EC_{50} exist, we use the logarithmized estimator $\phi^{(e)^*}$.

An alternative to EC_{50} is the ALEC, which is also a measure of the toxicity of a test compound (Jiang, 2013). For a parametric regression model function $y = f(x, \phi)$ the ALEC estimates the lowest effective concentration for a pre-specified critical effect level λ and is defined as the inverse function of f applied to λ:

$$f(\text{ALEC}, \phi) = \lambda \Rightarrow \text{ALEC} = f^{-1}(\lambda, \phi).$$

From formula (1), the ALEC can be calculated as a function $h(\lambda, \phi)$:

$$\text{ALEC} = h(\lambda, \phi) = \phi^{(e)} \left(\frac{\phi^{(d)} - \lambda}{\lambda - \phi^{(c)}} \right)^{1/\phi^{(b)}}. \quad (2)$$

The ALEC can only be estimated for an effect level λ which lies within the range of the lower and upper limit $\phi^{(c)}$ and $\phi^{(d)}$ of the concentration–response curve.

Due to the non-linearity of the function in (1), f is approximated with the least squares method using the Gauss-Newton algorithm. Since an iterative method is used for the estimation of the parameter vector ϕ, there is no guarantee to reach the global minimum. The estimation of the four parameters depends on the choice of the start values (Ritz *et al.*, 2015).

Uncertainties of the ALEC can be quantified by confidence intervals. Jiang (2013) showed that using the delta method (van der Vaart, 1998, p. 25) for approximating the variance of $h(\phi)$ from term (2) results in the following $(1 - \alpha)$ confidence interval for the ALEC:

$$\exp\left(\log(\widehat{\text{ALEC}}) \pm t_{(1-\alpha/2),\nu} \sqrt{\hat{\text{var}}[\log(\widehat{\text{ALEC}})]}\right)$$

where $t_{\nu,(1-\alpha/2)}$ is the $(1 - \alpha/2)$ quantile of a t-distribution with $\nu = n - 4$ degrees of freedom for n observations, see Jiang (2013) for an exact derivation of the confidence interval.

In addition to the ALEC, an alert concentration is of interest where the effect significantly exceeds a prespecified critical effect level λ. The effect of interest is the fold change (FC) i.e. the mean difference in logarithmic gene expression values. This concentration is called the LEC and is calculated based on a newly derived test statistic.

We first present a general form of this test, in which the difference between the response values of two concentrations is tested. Then we consider the case of interest, in which the statistical significance of the difference between the response value for a specific concentration and for the left asymptote, increased by the effect level λ, is assessed.

In order to test whether the modelled expression values for two concentrations $x_1 > 0$ and $x_2 > 0$ differ significantly, we derived a test statistic from the 4pLL model function in (1) for testing the hypothesis $H_0 : f(x_1, \phi) = f(x_2, \phi)$. The test statistic is given by

$$t_{4\text{pLL}} := t_{4\text{pLL}}(x_1, x_2, \hat{\phi}) = \frac{\hat{f}(x_1, \hat{\phi}) - \hat{f}(x_2, \hat{\phi})}{\sqrt{\hat{\text{Var}}[\hat{f}(x_1, \hat{\phi}) - \hat{f}(x_2, \hat{\phi})]}}.$$

The estimated variance of the difference $\hat{f}(x_1, \hat{\phi}) - \hat{f}(x_2, \hat{\phi})$ is derived using the delta method. Since $\hat{f}(x_1, \hat{\phi})$ and $\hat{f}(x_2, \hat{\phi})$ are highly correlated, the covariance term in $\hat{\text{Var}}[\hat{f}(x_1, \hat{\phi}) - \hat{f}(x_2, \hat{\phi})]$ remains in the calculation. We can estimate $\text{Var}[\hat{f}(x_1, \hat{\phi}) - \hat{f}(x_2, \hat{\phi})]$ using the approximation

$$\begin{aligned}
&\text{Var}[f(x_1, \phi) - f(x_2, \phi)] \\
&= \text{Var}[f(x_1, \phi)] + \text{Var}[f(x_2, \phi)] - 2\text{Cov}[f(x_1, \phi), f(x_2, \phi)] \\
&\approx \nabla f(x_1, \phi)^\text{T} \Sigma \nabla f(x_1, \phi) + \nabla f(x_2, \phi)^\text{T} \Sigma \nabla f(x_2, \phi) \\
&\quad - 2\nabla f(x_1, \phi)^\text{T} \Sigma \nabla f(x_2, \phi).
\end{aligned}$$

Σ corresponds to the covariance matrix of the four parameters and $\nabla f(x, \phi)$ to the gradient of f with respect to the parameter vector ϕ,

$$\nabla f(x, \phi) = \left(\frac{\partial f(x,\phi)}{\partial \phi^{(b)}}, \frac{\partial f(x,\phi)}{\partial \phi^{(c)}}, \frac{\partial f(x,\phi)}{\partial \phi^{(d)}}, \frac{\partial f(x,\phi)}{\partial \phi^{(e)}} \right)^T. \quad (3)$$

Under the null hypothesis, asymptotically $t_{4\text{pLL}} \sim \mathcal{N}(0, 1)$. The null hypothesis is therefore rejected at level α, if the observed value of $t_{4\text{pLL}}$ exceeds $z_{1-\alpha/2}$ or is smaller than $z_{\alpha/2}$, with z_q denoting the q% quantile of the standard normal distribution.

The application of this test to determine the LEC as alert concentration leads to the formulation of the null hypotheses

$$\mathrm{H}_0 : f(x,\phi) - f(0,\phi) \leq \lambda \qquad \text{for an increasing curve,} \tag{4}$$

$$\mathrm{H}_0 : f(x,\phi) - f(0,\phi) \geq -\lambda \qquad \text{for a decreasing curve,} \tag{5}$$

where λ is the prespecified effect level of interest.

In cases where the direction of the curve is known beforehand, e.g. from the biological background, only the corresponding null hypothesis needs to be tested. The test statistic for an increasing curve is given by

$$t_{4\text{pLL; inc}} := t_{4\text{pLL; inc}}(x,\hat{\phi},\lambda) = \frac{\hat{f}(x,\hat{\phi}) - \left(\hat{f}(0,\hat{\phi}) + \lambda\right)}{\sqrt{\hat{\mathrm{Var}}[\hat{f}(x,\hat{\phi}) - \hat{f}(0,\hat{\phi})]}}.$$

The corresponding P-value is calculated as $1 - \Phi(t_{4\text{pLL; inc}})$, where Φ denotes the distribution function of the standard normal distribution. Analogously, the test statistic for a decreasing curve is given by

$$t_{4\text{pLL; dec}} := t_{4\text{pLL; dec}}(x,\hat{\phi},\lambda) = \frac{\hat{f}(x,\hat{\phi}) - \left(\hat{f}(0,\hat{\phi}) - \lambda\right)}{\sqrt{\hat{\mathrm{Var}}[\hat{f}(x,\hat{\phi}) - \hat{f}(0,\hat{\phi})]}}$$

and the corresponding P-value is calculated as $\Phi(t_{\ 4\text{pLL; dec}})$.

In general, it is not known in advance whether a curve is increasing or decreasing. In this case, a two-sided P-value is calculated as

$$2 \cdot \min\left(1 - \Phi(t_{4\text{pLL; inc}}), \Phi(t_{4\text{pLL; dec}})\right). \tag{6}$$

For estimating the LEC, a search within the tested concentration range is performed (in our data example 0–1000 µM). First, it is tested whether the response for the highest concentration significantly exceeds the prespecified effect level λ in comparison to the control [i.e. a P-value as in (6) is calculated for the highest concentration and compared to the prespecified significance level α]. If not, no LEC can be determined. Otherwise, the LEC is determined via a bisection method: The starting limits of the first interval are the lowest and highest concentration considered. A P-value as in (6) is calculated for the mean concentration of the interval. If the P-value is smaller than α, the parameter space is restricted to the lower half of the considered interval, and to the upper half of the interval if it is larger. The algorithm stops if the length of the remaining interval is smaller than a small prespecified threshold value ε.

This approach is an alternative to the classical standard approach, where each measured concentration is tested separately to determine whether the critical level λ is exceeded significantly. In this case, the LOEC, determined with hypothesis testing, refers to the lowest concentration where the difference between treatment and control (the effect) significantly exceeds a given fold change. Similarly, the ALOEC is defined as the concentration where the average value of the fold change exceeds the critical effect level, without significance testing. Note that the letter 'O' in the names ALOEC and LOEC indicates that for these methods, only observed concentration values are potential candidates for alert levels, whereas ALEC and LEC allow arbitrary positive values as alert level.

A popular method in the field of toxicology for comparing several treatments with a control is the Dunnett-test (Dunnett, 1955). This test is a multiple-comparison procedure that tests for significant differences between responses to several treatments, e.g. increasing concentrations of a compound, and to a control. Multiplicity adjustment is executed by using a multivariate test statistic that incorporates the correlations between the treatment situations. Although this procedure would technically be preferred to the t-test, in this work we mainly report the results from widely used standard two-sample t-tests and we refer to Supplementary for results obtained with the Dunnett-test.

Table 1. Methods for estimating alert concentrations from concentration–gene expression data

	Observation-based t-test	Model-based 4pLL
FC	ALOEC	ALEC
FC & P-value	LOEC	LEC

Note: Rows indicate the cut-off criteria and columns the methods for estimating fold changes. An alert means that either a given fold change value is exceeded (FC) or that additionally the corresponding P-value is below a cut-point (FC & P-value). The P-value results either from the t-test or from the 4pLL modelling approach.

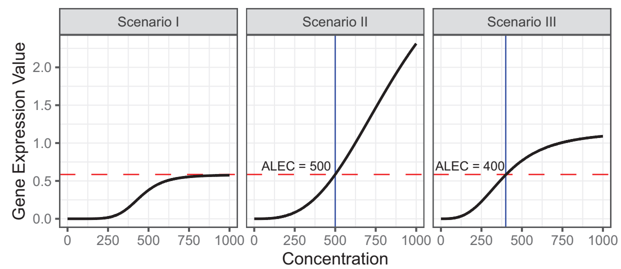

Fig. 1. Visualization of the three scenarios of concentration–response profiles considered in the simulation study. The y-axis shows the logarithmic expression values. The prespecified effect level $\lambda = \log_2(1.5)$ is visualized by a red line. For Scenarios II and III the true ALEC value can be calculated and is indicated by a blue vertical line. For Scenario I, the upper asymptote of the curve attains the value $\log_2(1.5)$ and therefore no ALEC value can be calculated

Table 1 summarizes the four estimators that were compared in this article, both in an exemplary simulation study and on real data.

2.2 Exemplary case study

A case study was conducted to investigate the development of human embryonic stem cells (hESC) to neuroectoderm (Krug *et al.*, 2013). Cells were treated *in vitro* with valproic acid (VPA) at seven different concentrations (25, 150, 350, 450, 550, 800 and 1000 µM). Each concentration was assessed in three replicate experiments. The cells were exposed to the compound over the entire differentiation process. In addition, six replicates for untreated measurements were available. The study was carried out within the framework of the European Commission-funded research consortium (ESNATS) which targeted the prediction of toxicity of drug candidates for the use of embryonic stem cell-based novel alternative tests.

2.3 Simulation study

The simulation study was performed with three different scenarios for the true concentration–gene expression relationship, seeFigure 1. The critical effect level λ was chosen to represent a FC of 1.5. The value of 1.5 was chosen from a biological motivation, taking the typical range of gene expression values for the given type of data into account. Since the data in the simulation study was intended to resemble the exemplary case study with $\log_2$-transformed values, a FC of 1.5 corresponds to the critical effect level $\lambda = \log_2(1.5) \approx 0.585$. The scenarios were chosen as follows, covering a broad range of concentration–response curves observed in real-data situations:

- In Scenario I, the true parameters of the curve were set to $\phi^{(b)} = -6$, $\phi^{(c)} = 0$, $\phi^{(d)} = 0.58$ and $\phi^{(e)} = 450$ such that the curve never exceeds the threshold $\lambda = 0.585$. In this case the true ALEC cannot be calculated. This Scenario corresponds to the null situation.

- In Scenario II, the true parameters were $\phi^{(b)} = -3$, $\phi^{(c)} = 0$, $\phi^{(d)} = 4$ and $\phi^{(e)} = 900$, and the curve clearly exceeds the given threshold, with true ALEC= 500. The curve is however not saturated in the range of considered concentrations, which is a challenge for modelling.
- Scenario III represents a situation with an almost saturated sigmoidal curve, with parameters $\phi^{(b)} = -3$, $\phi^{(c)} = 0$, $\phi^{(d)} = 1.16$ and $\phi^{(e)} = 400$, and with true ALEC= 400. This Scenario represents the best situation for modelling, as the effect attains values corresponding to the upper asymptote within the range of considered concentrations and the curve clearly exceeds the threshold.

The setup of the simulation study was inspired by the exemplary case study, with three replicates per concentration (in total $n = 24$) and the same concentration values as in the VPA study (0, 25, 150, 350, 450, 550, 800, 1000 µM), where a concentration of 0 refers to the control. The true parameters were used to calculate the true ALEC values and to generate simulated data. For each concentration, gene expression data from a normal distribution with mean $f(x, \phi)$ were generated, where f corresponds to the true 4pLL model function and x to the respective concentration. The analysis of the real data example revealed a correlation between the range and the standard deviation (w.r.t. concentration-wise replicates) of the gene expression values. A linear regression (with intercept) was fitted to this relationship. For each gene, the range was calculated as the difference between the mean of the response values for the highest concentration and for the control. The three scenarios considered correspond to ranges of 0.58 (Scenario I), 2.31 (Scenario II) and 1.16 (Scenario III). Corresponding estimated standard deviations (SD) were 0.189, 0.261 and 0.231. In addition to these 'medium' values of SD, using the factors 0.5 and 2, 'small' values (0.095, 0.131, 0.107) and 'large' values (0.379, 0.522, 0.427) were considered. These values were still observed remarkably often when considering the relationship between SD and range in real data. In the following, small, medium and large SD are abbreviated with 'small SD', 'medium SD' and 'large SD'.

For each scenario, the simulation procedure was repeated 1000 times to obtain simulated courses of 1000 genes. We estimated ALEC and ALOEC, as well as LEC and LOEC values. The LEC estimates were calculated using our proposed iterative algorithm, and the LOEC estimators resulted from the t-test approach.

2.4 Statistical analyses

The following analyses were performed using the statistical programming language R, version 4.0.0 (R Core Team, 2020). For the normalization of the entire set of 27 Affymetrix gene expression arrays, the extrapolation strategy (RMA+) (Harbon *et al.*, 2007) algorithm was used. RMA+ applies the steps background correction, $\log_2$ transformation, quantile normalization and a linear model fit to the normalized data in order to obtain a value for each probe set on each array. As reference, the normalization parameters obtained in earlier analyses were used (Krug *et al.*, 2013). After normalization, at each concentration the difference between averaged gene expression and averaged control values was calculated. The significance of this difference was assessed with a two-sample-t-test. The 4pLL model was fitted using the R package drc (Ritz *et al.*, 2015).

3 Results

The observation-based and the model-based estimates were compared. A main interest focused on the hypothesis-driven procedures yielding the LEC and the LOEC. The simulation study compared both estimators (FC and FC & P-value) with respect to their accuracy for the two methods (observation-based and model-based). The same analysis was performed on the data of the exemplary case study. Here, the analysis was restricted to those genes that showed a significant change in gene expression for at least one of the measured concentrations. We applied an analysis of variance to exclude probe sets with no effect at all and only kept probe sets with an unadjusted P-value smaller than 0.001, resulting in 9460 out of the initial 54675 probe sets.

3.1 Simulation results

Only results for the medium SD and large SD have been reported here, results for small SD are summarized in Supplementary Figure S1. Results for the LOEC obtained using a two-sample t-test have been summarized. Analogous figures for the LOEC obtained using the Dunnett-procedure are shown in Supplementary Figures S2–S4.

Some simulated genes had to be excluded from the analysis, when the numerical estimation of the covariance matrix of the parameters, Σ, resulted in implausible negative diagonal entries. As the diagonal entries correspond to the respective variances of the parameters, a negative result is an indicator of numerical difficulties when estimating Σ, and these results impair the calculation of the 4pLL test. The numbers of excluded genes for each situation were 14, 4 and 2 for medium SD, and 112, 8 and 16 for large SD (Scenarios I, II and III, respectively). In the observation-based approach, only expression profiles with unambiguous direction were considered, profiles with values above the upper threshold ($\log_2(1.5)$) at one concentration and below the lower threshold ($-\log_2(1.5)$) at another concentration were excluded, for the calculation of both LOEC and ALOEC.

The main results of the simulation study are summarized in Figures 2 (medium SD) and 3 (large SD). Key figures for modelling-based alert concentrations are shown in Table 2.

The total number of alerts differed from the total number of considered genes (1000) for several reasons. Firstly, the algorithm for model fitting may not have converged; the upper asymptote may not have exceeded the threshold λ, and, lastly, the respective (A)LEC estimate may have been larger than the highest measured concentration. For the (A)LOEC, no alert could be determined when the mean difference to the mean of the control value (significantly) exceeded the pre-defined threshold λ for none of the concentrations.

For Scenario I, the effect level of 0.585 is not reached for the true curve. Thus, every identified alert concentration was a false positive. For Scenarios II and III, an estimated alert concentration was called false positive if it was below the true ALEC value (500 and 400, respectively). The total numbers of false positive alerts for the different methods are summarized in Table 3.

The main results from the simulation study for each scenario are summarized separately below.

Scenario I: In Scenario I, with the true curve below the threshold no alerts were expected; however, both methods resulted in some false positives. The t-test approach with estimate $\widehat{ALOEC}$ resulted in over 600 false positives for both medium SD and large SD, and the 4pLL method with estimates $\widehat{ALEC}$ in 100 fewer false positives. Median values for the false positive $\widehat{ALEC}$ values were 652 (medium SD) and 554 (large SD), while most $\widehat{ALOEC}$ values were 550 and 800. For the more stringent criteria with significance testing, the methods resulted in only 33–44 false positive alerts in all cases.

Scenario II: For medium SD, in almost all cases a valid alert concentration (below 1000) was obtained. For $\widehat{ALEC}$, the median value of the 996 estimations was very close to the true value of 500, while for $\widehat{LEC}$, the median was a little higher. Both from the histograms in Figure 2 and from the standard deviations summarized in Table 2, a narrow distribution of the $\widehat{(A)LEC}$ values around their median values was observed. The $\widehat{ALOEC}$ was mostly between 350 and 800 and in more than 500 cases 550. The $\widehat{LOEC}$ values were larger, with a clear peak at 800, sometimes 550 or 1000, and rarely 450. For both $\widehat{ALOEC}$ and $\widehat{L(O)EC}$, the t-test approach yielded fewer false positive alerts than the 4pLL approach. These observations are summarized in the distribution functions in Figure 2, with similar starting points and initial slopes for $\widehat{ALOEC}$ and $\widehat{ALEC}$, although $\widehat{ALOEC}$ reached its maximal value at smaller alert concentrations.

For large SD, the model-based approaches resulted in more than 980 valid estimations (Fig. 3, second column, middle row), while the observation-based approaches resulted in only 832 (t-test) or 783 (4pLL). The median value of the $\widehat{ALEC}$ was again only slightly

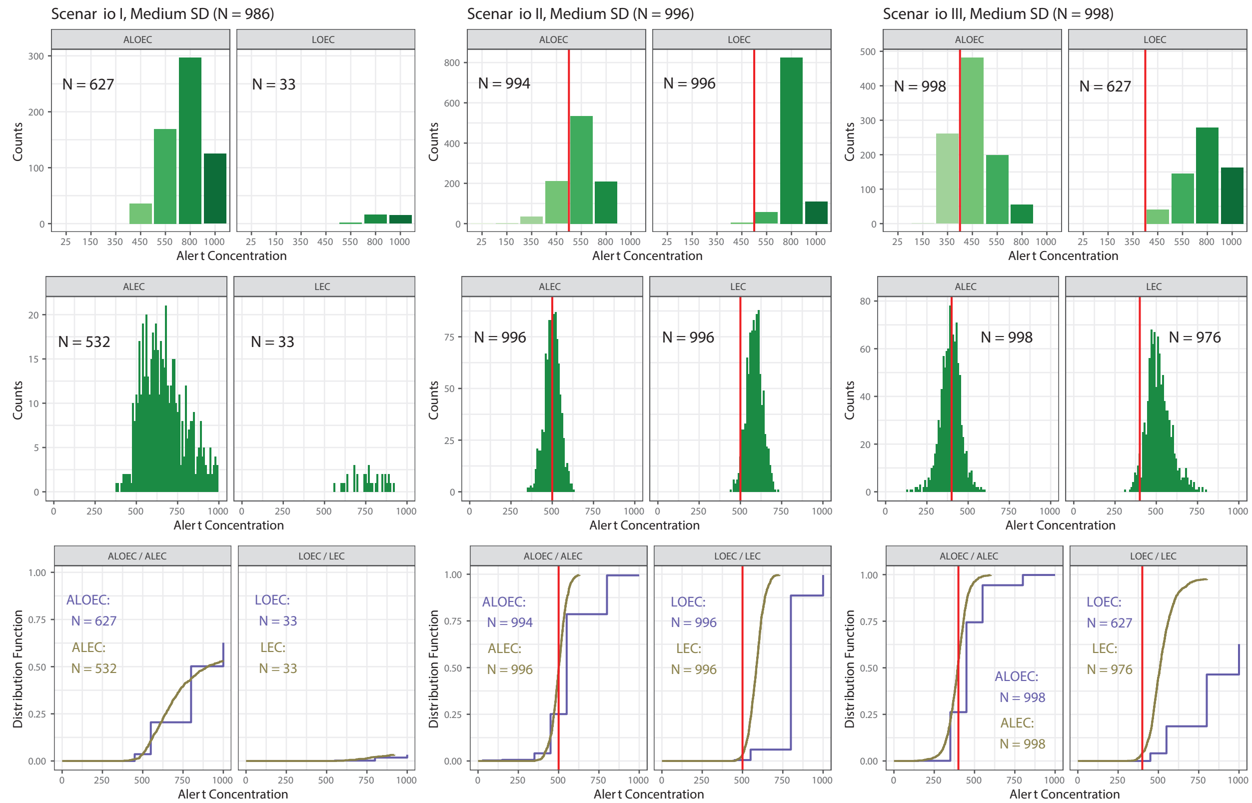

Fig. 2. Results of the simulation study for medium SD. The three columns represent from left to right Scenarios I, II and III. Each column is further subdivided into two columns. In the first of these columns, the AL(O)EC is displayed, i.e. the concentration obtained only by the FC criterion, and in the second column, the L(O)EC is displayed, i.e. the concentration for FC & *P*-value criterion. The top row corresponds to the observation-based methods (ALOEC and LOEC), the middle row to the model-based methods (ALEC and LEC) and the bottom row shows empirical distribution functions for both methods. For Scenarios II and III, the value of the true underlying ALEC is indicated by a red line. The number in each of the cells indicates the number of valid estimates in the range of concentrations considered, while the number in the respective columns' title corresponds to the total number of genes considered

Table 2. Summary statistics for the distributions of the ALEC and the LEC for Scenario I–III for the situations with medium SD and with large SD

		n		Med		SD	
		Medium	Large	Medium	Large	Medium	Large
ALEC	Scenario I	532	575	651.6	553.5	131.0	207.8
	Scenario II	996	989	500.7	510.9	45.9	95.7
	Scenario III	998	970	396.0	373.0	61.1	136.1
LEC	Scenario I	33	44	768.3	687.3	98.0	194.2
	Scenario II	996	988	585.5	674.1	46.8	92.8
	Scenario III	976	607	507.4	543.6	73.3	147.0

Note: The following parameters are presented: The total number of alerts (n), the median (Med) and the standard deviation (SD). The top three rows correspond to the ALEC values and the bottom three rows to the LEC values.

Table 3. Total numbers of false positive alerts, i.e. estimates below the true ALEC value, in Scenario I all identified alerts. Rows indicate the cut-off criteria

		Scenario I		Scenario II		Scenario III	
		t-test	4pLL	*t*-test	4pLL	*t*-test	4pLL
AL(O)EC	Medium SD	627	532	251	491	262	536
	Large SD	679	575	419	444	430	587
L(O)EC	Medium SD	33	33	5	35	0	42
	Large SD	38	44	8	36	12	50

Note: An alert was identified when the given fold change value of 1.5 was reached exactly (ALÊC, row 1) or exceeded by the average value (ALÔEC, row 2), or exceeded significantly ($p \leq 0.05$) (rows 3 and 4).

higher (511) than the true value (500), whereas the median of the LÊC was much higher (674). As indicated by the larger standard deviations for ALÊC and LÊC in comparison to the 'medium' situation, the histograms for ALÊC and LÊC were wider. For the ALÔEC, approximately 150 simulation runs yielded an estimate of 25 or 150, the rest of the observations were divided between the concentrations 350, 450, 550 and 800, with a peak at 550. By contrast, for LÔEC, almost all estimated values were 800 or 1000. The number of false positive alerts for AL(Ô)EC was similar using *t*-test and 4pLL, while for L(Ô)EC, only very few false positive alerts were obtained.

The distribution functions for the AL(Ô)EC intersect: Due to several very low alerts for the ALÔEC, the corresponding distribution function started to increase for lower concentrations, but with a smaller slope. Distribution functions for ALÊC and LÊC were comparable in terms of their slope and differed mostly with respect to the starting point of increase. End points of the observation-based distribution functions were lower than when the model-based approach was used due to the smaller number of valid estimates of the respective alert concentration.

Scenario III: In comparison with Scenario II, the true underlying curve of Scenario III exhibits a smaller range and a lower inflection point and therefore is already almost saturated for the highest concentration.

For medium SD and the FC as alert criterion, almost no invalid estimates were obtained. Using the more stringent criterion based on FC and *P*-value, 976 valid estimates were obtained using the 4pLL

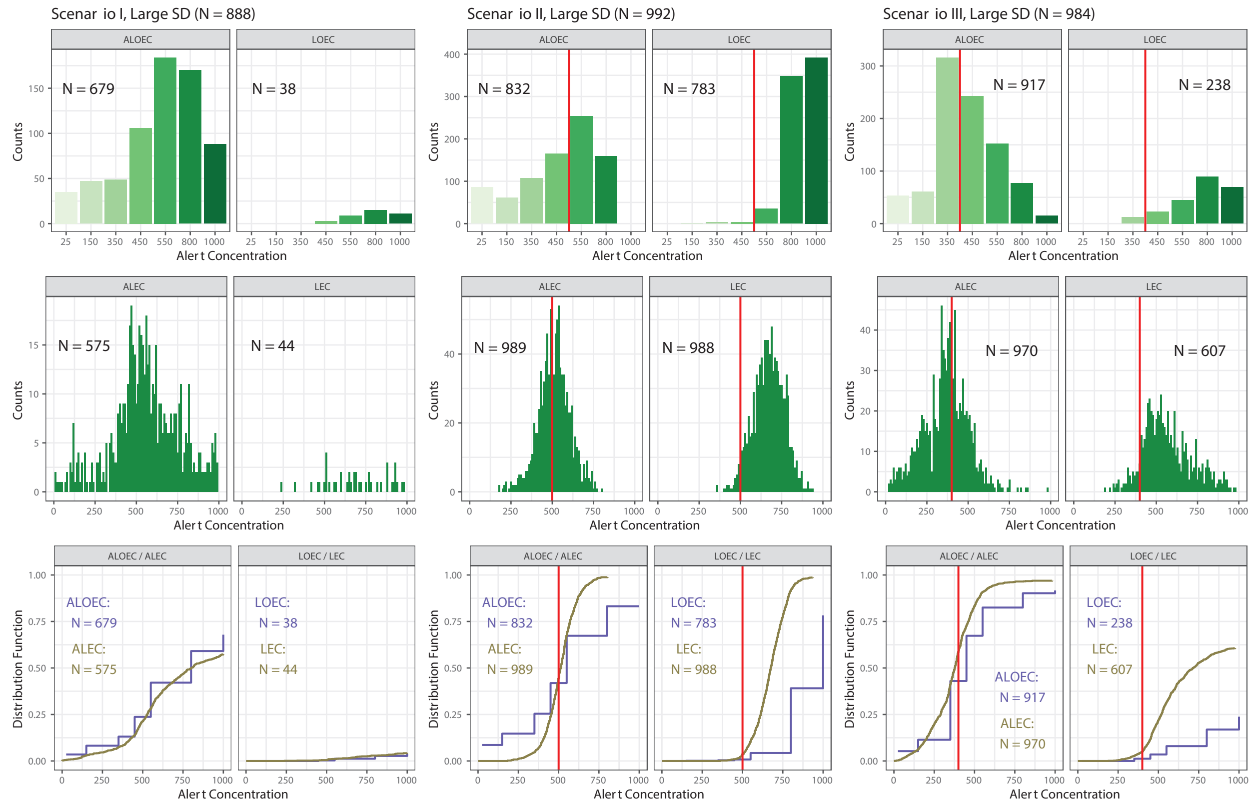

Fig. 3. Results of the simulation study for large SD, with the same structure as in Figure 2 for medium SD

method and only 637 using the *t*-test. For the $\widehat{\mathrm{ALEC}}$, the median was again very close to the true value of 400, whereas the median of all $\widehat{\mathrm{LEC}}$ values was again clearly larger. For the ALOEC, estimates were divided between the values 350, 450, 550 and 800, with a peak at 450, the best possible estimate for the observation-based method in this scenario. Estimates of the LOEC were between 450 and 1000, with a peak for 800. The number of false positive alerts for the $\widehat{\mathrm{AL(O)EC}}$ was twice as large using the 4pLL method compared to the *t*-test method. For $\widehat{\mathrm{L(O)EC}}$, no false positive alerts occurred using the *t*-test method and only 42 using the 4pLL method. The distribution functions of $\widehat{\mathrm{ALOEC}}$ and $\widehat{\mathrm{ALEC}}$ were very similar, while the distribution function of $\widehat{\mathrm{LEC}}$ started to increase earlier and exhibited a larger slope and a higher endpoint than the distribution function of $\widehat{\mathrm{LOEC}}$.

For large SD, for the FC criterion almost all estimates of the alert concentrations yielded valid results. This was in contrast to the $\widehat{\mathrm{LOEC}}$ and the $\widehat{\mathrm{LEC}}$, with only 238 and 608 valid estimates, respectively. The $\widehat{\mathrm{ALEC}}$ slightly underestimated the true ALEC of 400, while the $\widehat{\mathrm{LEC}}$ yielded a larger median value. The $\widehat{\mathrm{LOEC}}$ would not be considered to be a suitable estimator in this case as the few valid estimates all had a peak at 800 and thus clearly overestimate the true ALEC. On the other hand, for the $\widehat{\mathrm{ALOEC}}$, as in Scenario II, also small estimates of 25 and 150 were obtained (in over 100 cases) as well as many other values, with a peak at 350, corresponding to underestimation. The number of false positive alerts was very high for both $\widehat{\mathrm{ALOEC}}$ and $\widehat{\mathrm{ALEC}}$. The most striking difference in the distribution functions is the different endpoint for $\widehat{\mathrm{L(O)EC}}$, while the distribution functions for $\widehat{\mathrm{AL(O)EC}}$ are very similar.

The 95% confidence intervals (CI) for the ALEC estimators were calculated. Based on these, coverage probabilities (CPs) were estimated as percentages of cases with true ALEC value inside the CI. Since no true ALEC value is available for Scenario I, CPs were calculated for Scenarios II and III only. Only CIs with a length smaller than 1000 were considered. For medium SD, in Scenario II the CP was 0.83 (996 considered CIs) and in Scenario III 0.84 (998 considered CIs). For large SD, in Scenario II the CP was 0.86 (979 considered CIs) and in Scenario III 0.79 (920 considered CIs). Hence CPs are generally relatively low.

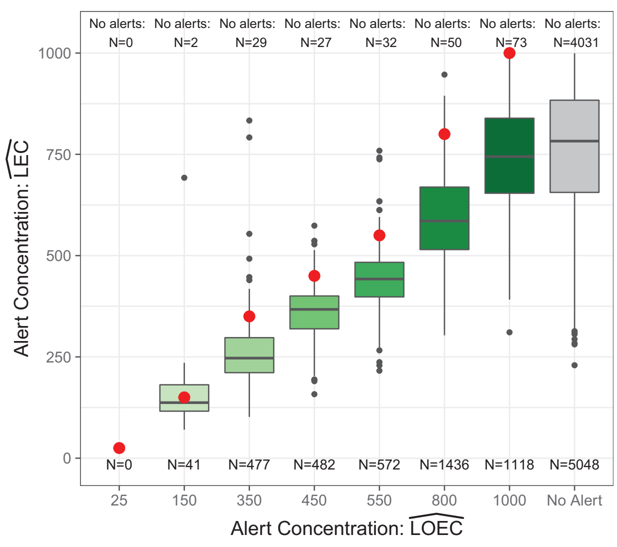

Fig. 4. Results of the analysis of the VPA dataset. Boxplots summarize values of the $\widehat{\mathrm{LEC}}$ values, stratified by corresponding $\widehat{\mathrm{LOEC}}$ values, which are also visualized by red dots. Numbers in the bottom row indicate total numbers of LOEC alerts and numbers in the top row cases with LOEC alert outside the permitted range. Hence each boxplot represents 'bottom number—top number' genes, e.g. for 550 µM $572 - 32 = 540$ genes

3.2 Exemplary case study

For the VPA study, only the results of the more stringent criteria of FC >1.5 and *P*-value $\leq$ 0.05 are shown. For simplicity, probe sets are referred to as genes. Again, some genes had to be excluded because of negative diagonal entries in the estimated covariance matrix Σ of the parameters in the 4pLL model. In this case 286 genes were excluded, keeping 9174 of the 9460 genes preselected with analysis of variance. Only 4128 and 4928 genes provided an estimate for the LOEC and LEC below 1000 μM respectively.

The boxplots in Figure 4 show values of the $\hat{\text{LEC}}$ alert concentration, stratified by the corresponding $\hat{\text{LOEC}}$ values that are also indicated by red dots. Additionally, Supplementary Figures S5 and S6 show boxplots for the same set of genes, but divided into the genes with an increasing and with a decreasing gene expression pattern. The direction is determined based on the fitted 4pLL model.

For almost all observed concentrations, the respective boxes are below the indicated points, i.e. in more than 75% of the cases the model-based approach yielded lower concentrations than the observation-based method. Note that the 4pLL model can be misspecified, and if the parametric assumptions do not hold, its estimates may be biased. In this sense, the results should be interpreted with caution. According to the *t*-test method, most alerts were identified at the concentrations 800 and 1000 μM, while our new 4pLL approach identified mostly smaller concentrations with quartiles 433 μM and 746 μM.

4 Discussion and conclusion

A frequent goal in concentration–response studies is to determine an 'alert' concentration where a critical effect level is exceeded. It is important to distinguish between a pure estimate of this 'alert' concentration and a concentration where the effect level is even statistically significantly exceeded. The standard approach is to analyse each measured concentration separately (*t*-test method). An alternative is to fit a sigmoidal curve, considering statistical variation in the corresponding parameter estimates (4pLL method).

We compared both methods in terms of the accuracy of the estimates. We performed a simulation study with three independent scenarios, covering different situations, and also evaluated the results on a real data example. In all cases, our proposed model-based approach (ALEC and LEC) performed better than the classical *t*-test approaches (ALOEC and LOEC). In Scenario I, where the critical fold change is not reached, fewer false positive signals were identified. In Scenarios II and III, where the expression pattern followed a pronounced sigmoidal shape clearly crossing the threshold, the estimates of the model-based approach were closer to the true alert concentrations than those of the observation-based approaches. The same trend was observed in the real data example. Compared to the observation-based approaches, the model-based approaches yield alerts at lower concentrations. The model-based approach benefits from its independence of measured concentration levels by allowing arbitrary positive values as alert concentrations.

An advantage of the model-based approach is that the ALEC and LEC values can also be estimated reliably in the case of incomplete concentration–response data. The right-sided asymptote can still be extrapolated by fitting a curve, but slightly biased estimations of this asymptote may occur. Those have less impact on ALEC and LEC, as the given effect level is predefined, than they have on alternatives like the EC_{50} that heavily depends on values of the asymptote. Another property of the model-based approach is that the entire information about the concentration–response relationship is incorporated in the estimation for a specific concentration. This can be a problem in unsaturated scenarios, where the variance of the parameter estimates becomes very large.

While all analyses were conducted under the assumption of homogeneity, the application of the methods is also possible in the case of heteroscedasticity or non-normal residuals (Calderazzo *et al.*, 2019). Heteroscedasticity and skewed distributions can also be addressed using robust standard errors as in the case of robust linear regression (Venables and Ripley, 2002).

In general, the use of the model-based approach is only recommended if the parametric assumptions of the model hold. In this case, the structure in the data can be captured by the parametric model and the highest possible efficiency can be obtained. In order to obtain reasonable estimates, it is recommended to test explicitly for deviations from sigmoidal curve progressions (Schoyer, 1984). Our method can be extended and applied to other parametric models, such as the log-normal or Weibull model.

Alternatively, non-parametric methods can also be used including Kernel regression (Müller and Schmitt, 1988; Staniswalis and Cooper, 1988) or local linear regression (Kelly and Rice, 1990; Zhang et al., 2013), as well as mixture models, with weighted averages of parametric and non-parametric fits. Such approaches have already been taken by Yuan and Yin (2011), Nottingham and Birch (2000), Olkin and Spiegelman (1987), Mays *et al.* (2000) and Pickle *et al.* (2008), among others.

Funding

This work was supported by the Bundesministerium für Bildung und Forschung (LivSysTransfer 031L0119).

Conflict of Interest: none declared.

References

Calderazzo,S. *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.

Delignette-Muller,M.-L. *et al.* (2011) A new perspective on the Dunnett procedure: filling the gap between NOEC/LOEC and ECx concepts. *Environ. Toxicol. Chem.*, **30**, 2888–2891.

Dunnett,C.W. (1955) A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.*, **50**, 1096–1121.

Harbon,C. *et al.* (2007) RefPlus: an R package extending the RMA Algorithm. *Bioinformatics*, **23**, 2493–2494.

Jensen,S.M. *et al.* (2019) A review of recent advances in benchmark dose methodology. *Risk Anal.*, **39**, 2295–2315.

Jiang,X. (2013) Estimation of effective concentrations from in vitro dose-response data using the log-logistic model, PHD Thesis, Medical Faculty of Ruprecht-Karls-University in Heidelberg.

Kelly,C. and Rice,J. (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, **46**, 1071–1085.

Krug,A.K. *et al.* (2013) Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Arch. Toxicol.*, **87**, 123–143.

Mays,J.E. *et al.* (2000) An overview of model-robust regression. *J. Stat. Comput. Simul.*, **66**, 79–100.

Müller,H.G. and Schmitt,T. (1988) Kernel and Probit Estimates in Quantal Bioassay. *J. Am. Stat. Assoc.*, **83**, 750–759.

Nottingham,Q.J. and Birch,J.B. (2000) A semiparametric approach to analysing dose-response data. *Stat. Med.*, **19**, 389–404.

Olkin,I. and Spiegelman,C.H. (1987) A semiparametric approach to density estimation. *J. Am. Stat. Assoc.*, **82**, 858–865.

Pickle,S.M. *et al.* (2008) A semi-parametric approach to robust parameter design. *J. Stat. Plan. Inference*, **138**, 114–131.

R Core Team. (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna

Ritz,C. (2010) Toward a unified approach to dose-response modeling in ecotoxicology. *Environ. Toxicol. Chem.*, **29**, 220–229.

Ritz,C. *et al.* (2015) Dose–response analysis using R. *PLoS One*, **10**, e0146021.

Ritz,C. *et al.* (2019) *Dose–Response Analysis Using R*. CRC Press, Baco Raton, FL.

Schoyer,R.L. (1984) Sigmoidally constrained maximum likelihood estimation in quantal bioassay. *J. Am. Stat. Assoc.*, **79**, 448–453.

Staniswalis,J.G. and Cooper,V. (1988) Kernel estimates of dose response. *Biometrics*, **44**, 1103–1119.

Vaart,A.W.vd. (1998) *Asymptotic Statistics*. Cambridge University Press. Cambridge.

Venables,V.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S*. Springer, Berlin.

Yuan,Y. and Yin,G. (2011) Dose–response curve estimation: a semiparametric mixture approach. *Biometrics*, **67**, 1543–1554.

Zhang,H. *et al.* (2013) A strategy to model nonmonotonic dose–response curve and estimate IC50. *PLoS One*, **8**, e69301.