

RESEARCH ARTICLE

Model selection characteristics when using MCP-Mod for dose–response gene expression data

Julia C. Duda  | Franziska Kappenberg  | Jörg Rahnenführer 

Department of Statistics, TU Dortmund University, Dortmund, Germany

Correspondence

Julia C. Duda, Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany.
Email: duda@statistik.tu-dortmund.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: LivSysTransfer (031L0119); Deutsche Forschungsgemeinschaft, Grant/Award Number: RTG 2624 (427806116)



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

We extend the scope of application for MCP-Mod (Multiple Comparison Procedure and Modeling) to in vitro gene expression data and assess its characteristics regarding model selection for concentration gene expression curves. Precisely, we apply MCP-Mod on single genes of a high-dimensional gene expression data set, where human embryonic stem cells were exposed to eight concentration levels of the compound valproic acid (VPA). As candidate models we consider the sigmoid E_{\max} (four-parameter log-logistic), linear, quadratic, E_{\max} , exponential, and beta model. Through simulations we investigate the impact of omitting one or more models from the candidate model set to uncover possibly superfluous models and to evaluate the precision and recall rates of selected models. Each model is selected according to Akaike information criterion (AIC) for a considerable number of genes. For less noisy cases the popular sigmoid E_{\max} model is frequently selected. For more noisy data, often simpler models like the linear model are selected, but mostly without relevant performance advantage compared to the second best model. Also, the commonly used standard E_{\max} model has an unexpected low performance.

KEYWORDS

dose–response curves, gene expression, MCP-mod, model selection, toxicology

1 | INTRODUCTION

In drug development, two major steps are of interest when a new compound is examined. First, changes in the dose or concentration of the compound are intended to cause changes in the response. Once this relation is established, the precise modeling of the dose–response curve is the next goal. It aims at finding the target dose for the confirmatory Phase III trials.

If multiple comparison procedures (MCPs) are used for signal detection, this can lead to less flexibility as target dose estimation is restricted to the tested dose levels. One major methodological advancement in this field is the Multiple Comparison Procedure and Modeling (MCP-Mod) approach by Bretz et al. (2005). It combines MCP and a modeling (Mod) step by proposing a multistage procedure. MCP-Mod received a positive qualification opinion and a “fit for purpose” designation by the EMA and FDA in 2014 and 2016, respectively, as statistical methodology to analyze Phase II dose-finding studies under model uncertainty (European Medicines Agency, 2015; Food and Drug Administration, 2016).

This work extends the usual scope of application of MCP-Mod from clinical Phase II to gene expression data. As a practical example, human embryonic stem cells are analyzed (O’Quigley et al., 2017, Chap. 12.3). Valproic acid (VPA) is

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

used for treating epilepsy but it is known to be embryo-toxic when taken in the first trimester of pregnancy (Genton et al., 2006). The MCP-Mod framework can help to gain insights on concentration–response relationships between the concentration of VPA and gene activity.

Specifically, we are interested in two aspects of MCP-Mod when applied on concentration–response data: the detection of genes where VPA has an effect on the dose–response curve (power) and model selection. We investigate these properties in analyses on real and on simulated data. Further, the model performance or goodness-of-fit of selected models is evaluated to identify which models are suitable for gene expression dose–response data.

Model selection and model performance differ substantially in the underlying theory. In model selection a statistical model from a set of candidate models has to be selected, given a data set. The aim is to select the model that represents the true, unknown model function best (Chap. 1 of Claeskens & Hjort, 2008; Schorning et al., 2016)). In addition to selecting the best model among the candidates, we also aim at identifying necessary or dispensable models. Therefore, we use the goodness-of-fit measure R_{adj}^2 . We combine the three aspects power, model selection, and goodness-of-fit in a newly proposed score that summarizes the suitability of a model set. This approach is applied on the VPA data set and on simulated data.

In the context of clinical Phase II trials, model uncertainty for dose–response modeling is considered to increase precision in target dose estimation—Ting (2006), Wheeler and Bailer (2009), Bornkamp et al. (2011) among many others. In Phase II trials, decisions on the model set can be based on expert knowledge and concentrate on a single compound and dose–response relationship. For gene expression data, model selection must be considered for thousands of genes simultaneously and it is not straightforward to find or use prior knowledge on the dose–response profile of each gene. House et al. (2017) and Filer et al. (2016) propose experimental pipelines that include concentration–response modeling and model selection for toxicological gene expression data. They consider a flat model, the sigmoid E_{max} model with all four parameters or with the lower asymptote fixed to zero, and a gain–loss model that is similar to the beta model considered here. However, detailed investigations on the necessity of model selection and on appropriateness of candidate model sets for gene expression concentration–response data are lacking, which motivates our work.

This paper is structured as follows. The VPA data set is introduced in Section 2. The statistical methodology including MCP-Mod and both established performance measures and a newly proposed one are presented in Section 3. Our analysis procedures and results that are based on the VPA data set are presented in Section 4. Different controlled simulation setups and corresponding results follow in Section 5. Final conclusions are summarized in Section 6. Source code to reproduce the results is available as [Supporting Information](http://onlinelibrary.wiley.com/doi/xxx/supinfo) on the journals web page (<http://onlinelibrary.wiley.com/doi/xxx/supinfo>).

2 | GENE EXPRESSION DATA SET

The data set was first presented in the study of Krug et al. (2013), where VPA is applied, among others, to human embryonic stem cells (hESC). VPA is widely used to treat different forms of epilepsy. However, it is linked to an increased incidence in congenital abnormalities (Cotariu & Zaidman, 1991). Krug et al. (2013) state that identifying changes in the transcriptome induced by toxic substances illustrates interesting mechanistic insights.

Gene expression levels of the hESCs are measured repeatedly for different concentrations, using the GeneChip R Human Genome U133 Plus 2.0. The data are preprocessed with the Robust Multi-Chip Average algorithm by Irizarry et al. (2003), such that the expression data are on the logarithmic scale with base 2.

The data set contains $G = 54,675$ probe sets, which will be referred to as genes in the following, for simplicity. For every gene, expression values corresponding to the concentrations $d_1 = 0$, $d_2 = 25$, $d_3 = 150$, $d_4 = 350$, $d_5 = 450$, $d_6 = 550$, $d_7 = 800$, and $d_8 = 1000 \mu\text{M}$ VPA are available. For the control level d_1 , $n_1 = 6$ replicates were measured. For all other concentrations there are $n_2 = \dots = n_8 = 3$ replicates. There are $N = 27$ measurements per gene. The replicates are biological replicates since different cells were used for each experiment. Due to functional relationships between genes, we cannot assume independence between the measurements from different genes. Further, six or three replicates per concentration is small for statistical modeling approaches. These problems are addressed in Section 4.

3 | MCP-MOD METHODOLOGY AND PERFORMANCE MEASURES

In this section, the methodology is presented. First, the MCP-Mod approach is outlined. Then, the performance measures precision and recall for evaluating the model selection in MCP-Mod are explained. Additionally, the newly proposed measure S_M is presented.

TABLE 1 Dose–response models $f(d, \theta)$, their standardized versions $f^0(d, \theta^0)$, and the guesstimates for θ^0 for the analysis. For the beta model B is defined as $B(\delta_1, \delta_2) = (\delta_1 + \delta_2)^{\delta_1 + \delta_2} / (\delta_1^{\delta_1} \delta_2^{\delta_2})$ and $D = 1200$

| Model | $f(d, \theta)$ | $f^0(d, \theta^0)$ | θ^0 |
|--------------------|--|--|---------------------------------------|
| E_{\max} | $E_0 + E_{\max}d/(ED_{50} + d)$ | $d/(ED_{50} + d)$ | $ED_{50} \in \{100\}$ |
| Sigmoid E_{\max} | $E_0 + E_{\max}d^h/(ED_{50}^h + d^h)$ | $d^h/(ED_{50}^h + d^h)$ | $ED_{50} = 450, h = 5.117$ |
| Exponential | $E_0 + E_1\{\exp(d/\delta) - 1\}$ | $\exp(d/\delta) - 1$ | $\delta \in \{144.455\}$ |
| Linear | $E_0 + \delta d$ | d | \emptyset |
| Quadratic | $E_0 + \beta_1 d + \beta_2 d^2$ | $d + (\beta_2/ \beta_1)d^2$ | $\delta = \beta_2/ \beta_1 = -0.001$ |
| Beta | $E_0 + E_{\max}B(\delta_1, \delta_2)(d/D)^{\delta_1} \cdot (1 - d/D)^{\delta_2}$ | $(d/D)^{\delta_1}(1 - d/D)^{\delta_2}$ | $\delta_1 = 2, \delta_2 = 1$ |

3.1 | MCP-Mod

The MCP-Mod approach was originally developed by Bretz et al. (2005) to model dose–response relationships in Phase II clinical trials under model uncertainty. For details see also Xun and Bretz (2017) and Bornkamp et al. (2009).

The MCP-Mod methodology comprises two analysis steps. First, in the MCP-step, a statistically significant signal in a gene is determined by an optimal-contrast test for a prespecified set of candidate models. If such a signal is found for at least one model, a significant result of the multiple comparison procedure (signifMCP) is present for the gene. This means that an effect of VPA on the gene activity is established. The second step, Mod, refers to the modeling. From the set of models, for which a signifMCP has been established, one model fit is chosen and used as final fit for the data. Alternatively, model averaging can be performed.

Denote d_1 as placebo concentration and $d_2 < \dots < d_k$ as active concentrations with n_i replicates. For concentration $i = 1, \dots, k$ and $j = 1, \dots, n_i$, $N = n_1 + \dots + n_k$, the (preprocessed) expression values are modeled as

$$y_{ij} = \mu(d_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad (1)$$

with homogeneous variance $\sigma^2 > 0$. The mean response $E(y_{ij}) = \mu_i = f(d_i, \theta)$ at concentration d_i is assumed to follow a concentration–response model with parameter vector θ and ε_{ij} as independent errors.

For the MCP step, a set \mathcal{M} of M candidate models needs to be prespecified. Models commonly used for dose–response relationships are summarized in Table 1.

All models summarized in the first column of Table 1 can be reformulated as

$$f(d, \theta) = \theta_0 + \theta_1 f^0(d, \theta^0), \quad (2)$$

(see second column of Table 1), where $f^0(d, \theta^0)$ is the standardized version of a model. Introduction of the standardized model shape concept is crucial for choosing optimal contrasts in the MCP step, as their choice is scale invariant.

It remains to determine initial guesses for the parameter θ^0 . In practice, for a Phase II study, careful considerations and prior knowledge on expected percentages of maximal effects at certain doses are translated into guesstimates for θ^0 . Here, the large number of genes makes individual, gene-dependent decisions on θ^0 difficult. We therefore use the same guesstimates for all genes. Figure 1 displays the (rescaled) model shapes $f^0(d, \theta^0)$ used for the analysis. The guesstimates are listed in Table 1.

To the best of our knowledge there is little preliminary work on dose–response model selection in the context of gene expression data (Filer et al., 2016; House et al., 2017). In toxicology, often monotone dose–response relationships are assumed. Especially the E_{\max} model, a special case of the sigmoid E_{\max} model with $h = 1$, was found to be appropriate for the majority of dose–response relationships in a large meta-analysis of clinical dose–response studies (Thomas et al., 2014). The inclusion of these two monotonic models in the candidate model set is therefore obligatory. The linear model is added as a reference or baseline model. For genes where the true underlying model might be a sigmoid E_{\max} model, but at the maximal considered dose, the turning point has not yet been reached, the exponential model might be more suitable. The quadratic and the beta model are included as nonmonotone shapes. They are similar to the gain–loss model used by Filer et al. (2016). There might be a nonmonotone relationship between concentration and gene activity, for example, for metabolic genes. Such genes might be activated at lower VPA concentrations but successively deactivated at increasing, highly toxic concentrations.

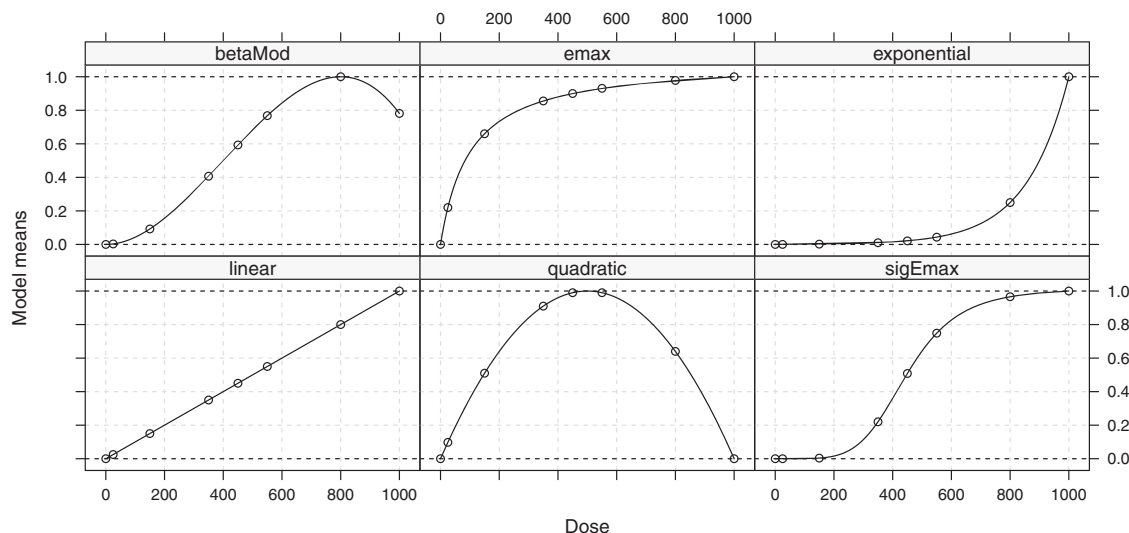


FIGURE 1 Candidate model shapes used for the six concentration–response models

The specific guesstimates for θ^0 for each model are chosen such that a wide range of dose–response shapes is covered. Further, we consider that during the experimental design stage, the concentrations were chosen with the expectation that the dose with the half maximal effect (ED_{50}) is close to 450 and a plateau is reached at concentration 1000, which translates into the second guess that 95% of the maximal effect is reached at concentration 800. With these two assumptions ($ED_{50} \approx 450$ and $ED_{95} \approx 800$), the guesstimate θ^0 for the sigmoid E_{\max} model can be calculated analytically. For the E_{\max} model, a guess of an ED_{50} of 300 is used. And for the exponential model, an ED_{50} of 700 is assumed.

Each candidate shape, $m = 1, \dots, M$, defines a respective mean response vector $\mu_m = (\mu_{m1}, \dots, \mu_{mk})$. For the MCP-step, a contrast t -test as first described by Abelson et al. (1963) is calculated. The test is constructed based on the linear contrast $\mathbf{c}_m^\top \mu_m$ where $\mathbf{c}_m = (c_{m1}, \dots, c_{mk})^\top$ is chosen to maximize the power of the test for the assumed mean response μ_m (Bornkamp et al., 2009). This yields the hypotheses $H_0^m : \mathbf{c}_m^\top \mu_m = 0$ and $H_1^m : \mathbf{c}_m^\top \mu_m \neq 0$.

The test statistics for the contrasts are given by

$$T_m = \frac{\sum_{i=1}^k c_{mi} \bar{y}_i}{S \sqrt{\sum_{i=1}^k c_{mi}^2 / n_i}}, \quad m = 1, \dots, M, \quad (3)$$

where \bar{y}_i is the observed mean at dose d_i and $S^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (N - k)$ is the mean squared error. Under H_0 and (1), $(T_1, \dots, T_m)^\top$ follows a central, multivariate t -distribution.

A dose–response signal is established if $T_{\max} = \max(T_1, \dots, T_m) > q_{1-\alpha}$, where q_α is the equicoordinate α -quantile of the null distribution. This approach leads to multiple testing adjustment for $\{H_0^m, H_1^m\}$ with a strict control of the family wise error rate at level α . The models with $T_m > q_{1-\alpha}$ form the set $\mathcal{M}^* = \{M_1, \dots, M_L\}$ of L significant models with established signifMCP. The modeling step is only executed if $\mathcal{M}^* \neq \emptyset$, that is, a dose–response signal is established for at least one model.

During the Mod-step, either one fitted model of those that passed the MCP-step can be chosen for a final fit or all fitted models that passed the MCP-step can be averaged. If a single model is selected, criteria as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) as well as the largest test statistic ($\max T$) can be used to pick a model. For model averaging, standardized weights based on the AIC or BIC can be calculated for the models in \mathcal{M} , and the final model is the resulting weighted average of each of the fitted models.

Calculations are done with the DoseFinding R package, version 0.9-17, and the statistical software R, version 4.0.2 (R Core Team, 2020). For the numerical estimation of the nonlinear parameters, we use the default boundary setting of the DoseFinding package. As the maximum concentration is 1000, this leads to boundaries for the ED_{50} parameter of $[1, 1500]$ and $[1/2, 10]$ for the h parameter of the sigmoid E_{\max} model.

3.2 | Measures

In this section, we briefly present for our context the definitions of the evaluation measures precision, recall, and R_{adj}^2 . Further, a new measure $S_{\mathcal{M}}$ is proposed specifically for the context of using MCP-Mod with a fixed candidate set \mathcal{M} on many dose–response data sets.

For a set of genes and a specific model $M \in \mathcal{M}^* = \{M_1, \dots, M_L\}$, the precision is defined as the conditional probability that a model is correct, given that it has been selected. Accordingly, the recall is defined as the conditional probability that a model is selected, given that it is correct (Buckland & Gey, 1994). Formally, we denote

$$\begin{aligned} \text{precision} &= \hat{P}(\text{Model is correct} \mid \text{Model is selected}) = \frac{tp}{tp + fp}, \\ \text{recall} &= \hat{P}(\text{Model is selected} \mid \text{Model is correct}) = \frac{tp}{tp + fn}, \end{aligned}$$

where tp , fp , and fn are the number of true positives, false positives, and false negatives. Precision and recall values are in the interval $[0,1]$, and a larger value corresponds to a better performance. They can only be evaluated in simulations where the correct model is known.

For a model fit $f(\cdot, \hat{\theta})$ and data y_{ij} of a specific gene, $i = 1, \dots, k$, $j = 1, \dots, n_i$, we use R_{adj}^2 defined as

$$R_{\text{adj}}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p}, \quad R^2 = 1 - \frac{SSE}{SST}, \quad (4)$$

where $SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - f(d_i, \hat{\theta}))^2$ and $SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ is the sum of squared errors and total sum of squares, respectively. The number of parameters is p and the total number of measurements is $N = \sum_{i=1}^k n_i$.

We further propose a new measure, the suitability of model set score $S_{\mathcal{M}}$. It can be used in a descriptive manner when MCP-Mod is applied to dose–response or concentration–response data of many genes. The score balances two desired properties. First, the number of detected signals (signifMCPs) is desired to be large. Additionally, the detected signals are also desired to be clear, that is, to have a large R_{adj}^2 value. The score balances the number of detected signals and the model performance, that is, power and goodness-of-fit. It is defined as

$$S_{\mathcal{M}} = \frac{1}{G} \sum_{g=1}^G \mathbb{1}\{\text{Gene } g \text{ has significant MCP after adjustment}\} \cdot R_{\text{adj}}^2, \quad (5)$$

where G denotes the total number of genes and \mathcal{M} the considered set of candidate models. For a given set \mathcal{M} , the score summarizes the proportion of genes with significant MCP after adjustment, weighted by the goodness-of-fit of the respective genes. Adjustment means that the false discovery rate (FDR) is controlled using the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995).

In the context of MCP-Mod, this means that for each gene, the smallest p -value from the MCP tests of all candidate models is chosen. Consequently, each gene is represented by a single p -value. These p -values are adjusted with the BH procedure. If a BH-adjusted p -value is below 0.05 then this results in a multiplicity-adjusted significant concentration–response signal. As performance measure, the value of R_{adj}^2 of the winner model w.r.t. AIC for the corresponding gene is used. In general, when comparing two values $S_{\mathcal{M}_1}$ and $S_{\mathcal{M}_2}$, the larger value indicates a favorable choice of the candidate set, since both the number of detected signals and the model performance are taken into account. For improved clarity, in addition both the proportion of genes with detected signal and the mean R_{adj}^2 of the fit of the winner models corresponding to those genes will be reported.

4 | DATA-BASED ANALYSIS

In this section, setups and results of the data-based analyses are presented. In the following, they are referred to as Analysis I and Analysis II. In Analysis I, MCP-Mod is applied on the real VPA data set and an overview on model selection results

TABLE 2 Overview of the analyses scenarios and their respective data generation details (Section 4) as well as for the simulation study of Section 5

| Part | | Data (generation) |
|-------------|------|---|
| Analysis I | main | Original VPA data |
| Analysis II | LOMO | Original VPA data |
| Simulation | | $n_1 = \dots = n_g \in \{3, 5, 10\}$ |
| | | $\sigma = q(0.5) \cdot \text{range} (\sigma = q_{\text{null}}(0.5) \text{ for null-model})$ |

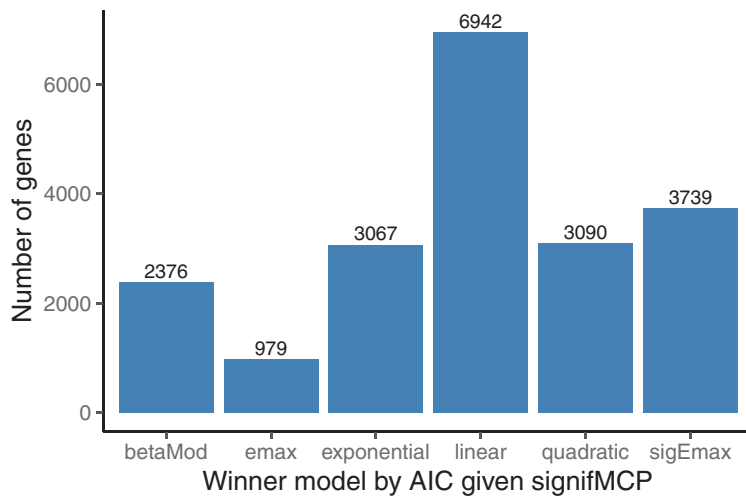


FIGURE 2 Distribution of winner counts per model

is provided. An additional goal is to check if any model can be omitted from the candidate model set because it can be easily substituted by another model. Analysis I is extended by Analysis II through leave-one-model-out (LOMO) analyses. These include that the entire analysis of Analysis I is repeated several times, and each time one of the candidate models of \mathcal{M} is omitted. For an overview of the different analyses and the simulation, see Table 2.

4.1 | Setup for Analysis I

In Analysis I, MCP-Mod is applied independently on each gene of the VPA data set. As we cannot assume only increasing or decreasing effects, each gene is tested with two-sided contrast tests with significance level $\alpha = 0.05$.

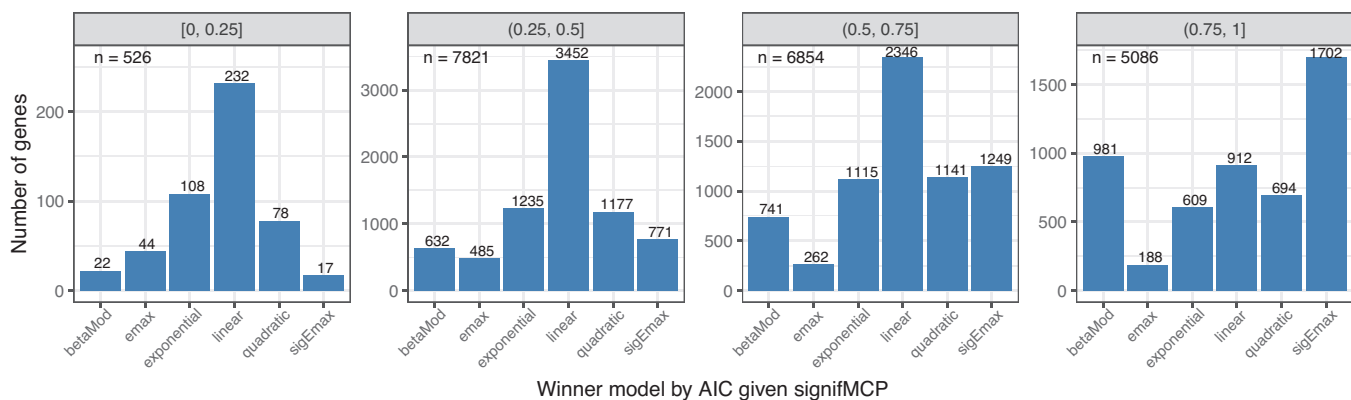
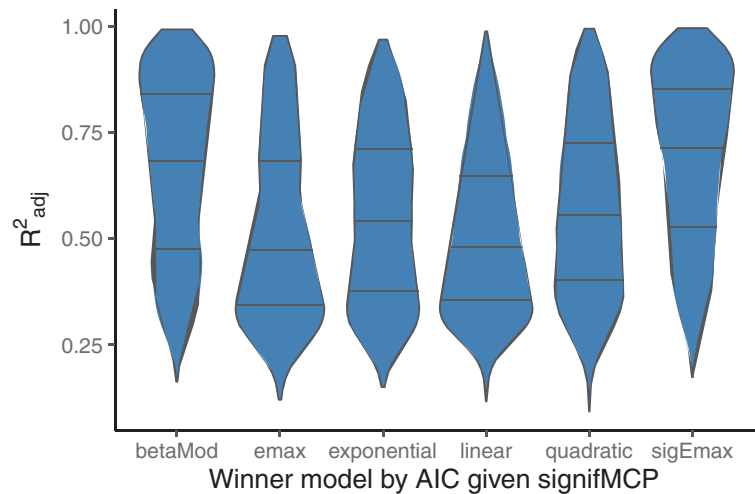
We apply multiplicity adjustment between genes by controlling the FDR using the BH procedure as described in Section 3.2. For each gene, if a dose–response signal is detected and hence at least one model passes the MCP-step, the AIC is used as model selection criterion. For small sample sizes, Schorning et al. (2016) show that the AIC outperforms the BIC, especially if the true underlying model is a complex one among the considered models. There are $N = 27$ observations per gene in the VPA data set. Thus we use the AIC to avoid too low selection counts of possibly correct, more complex models. In our analysis we will see that even with the AIC the simple linear model is often selected.

4.2 | Results for Analysis I

Of the 54,675 genes, when controlling the FDR, 20,193 (36.9%) genes pass the MCP-step, that is, their concentration–response profile significantly differs from a flat profile. VPA has a significant effect on the activation (deactivation) of these genes. For each gene one winner model is selected by AIC as a final fit (Figure 2). The linear model is selected most often (34.4%), because the AIC penalizes more complex models. However, Figure 3 clearly shows that the linear model performs comparatively poorly with respect to the R_{adj}^2 .

The popular E_{max} model (cf. Thomas et al., 2014, among many others) wins the fewest times and when it does win, its fit has low R_{adj}^2 values (Figures 2, 3). Figure 4 shows the distribution of model winner counts w.r.t. AIC stratified by R_{adj}^2 . Less noisy genes are represented by the rightmost plot. Assuming (1), we refer to more (less) noisy genes as genes whose

FIGURE 3 Performance of winner models

FIGURE 4 Distribution of winner counts per model stratified by R^2_{adj}

underlying model has larger (smaller) standard deviations σ in relation to the response range, which leads to smaller (larger) R^2_{adj} values. While the sigmoid E_{max} and the beta model win most often for the least noisy stratum, the E_{max} model is chosen rarely, regardless of the strata. The nonmonotone beta and quadratic model are chosen considerably often. For more noisy genes the linear model is preferred. For these genes, none of the models explains a lot of variance, which favors the linear model in terms of AIC. Hence, if the linear model is selected by AIC, one should hesitate to assume a true linear concentration–response relationship. Some example fits are visualized in Figure 5.

To ensure that the low number of E_{max} winners is not only due to too strict parameter constraints in the numerical optimization, we visualize the ED_{50} parameter estimates for genes where the E_{max} model won (Appendix, Figure A1). There is no evidence that the poor performance of the E_{max} model is due to optimization constraints, but instead due to the often low ED_{50} estimates. For an E_{max} model, a low ED_{50} translates to an early plateau, which can lead to an SSE close to the SST and therefore to a small R^2_{adj} .

It is also of interest if any of the models in the candidate set is redundant such that it can be substituted by another model. Removing such a model from the candidate set would increase power as the number of hypotheses would be decreased in the MCP-step of each gene. If for many genes the R^2_{adj} for the winner model and the second best model differ substantially, the winning model should be considered for future analyses. If the quadratic model is the winning model with a good fit, many genes cannot be modeled well by the second best model (Figure 6).

The sigmoid E_{max} and the beta model performances also differ by a considerable amount to the second best model's performance across the whole range of explained variance. The E_{max} and the exponential model can mostly be replaced by other models without substantial loss in R^2_{adj} . This especially applies to genes with larger explained variance. The linear model can always be replaced with minimal loss in explained variance, as it is a special case of the E_{max} model and the quadratic model.

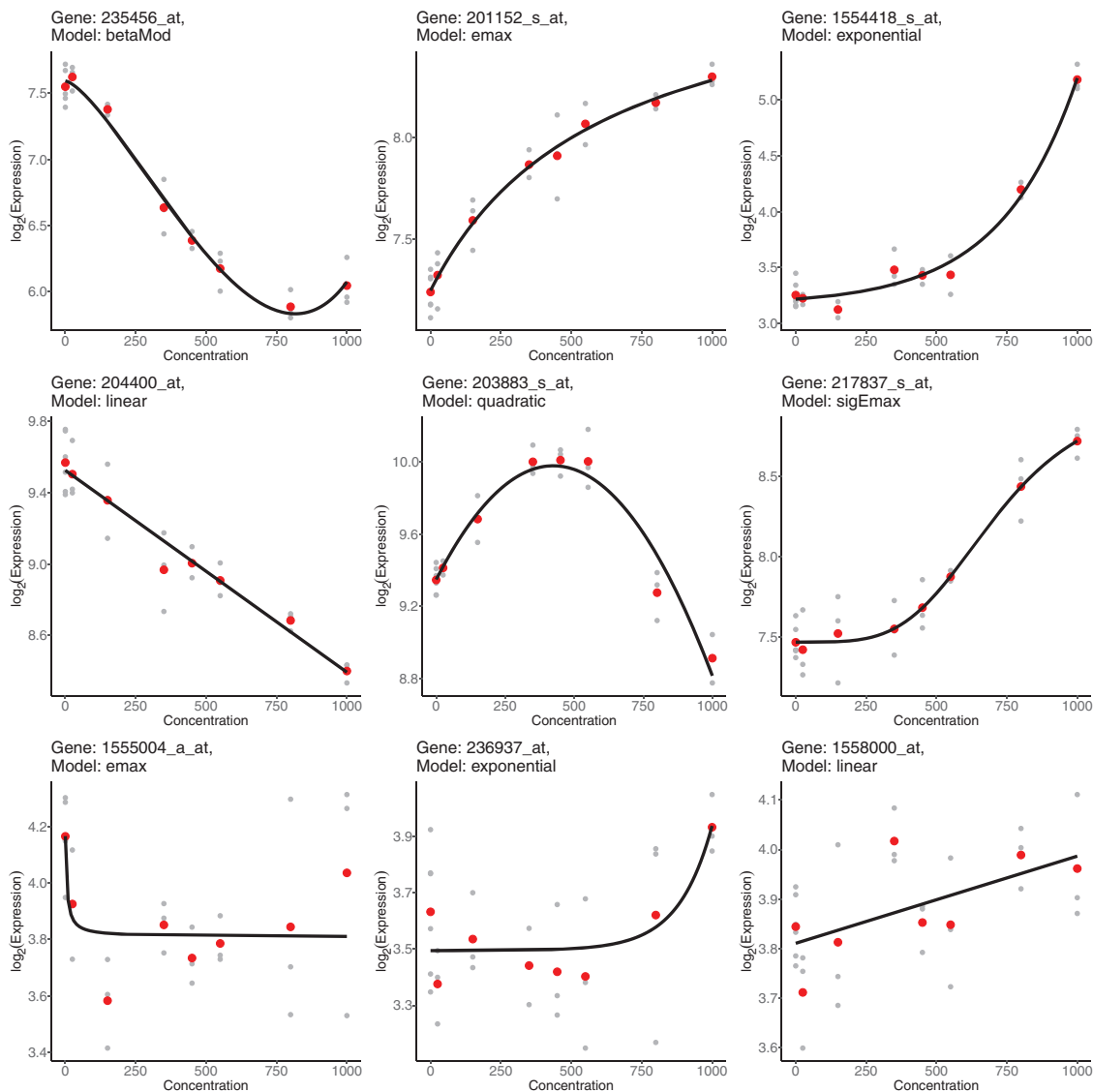


FIGURE 5 Example selection of nonnoisy (rows 1 and 2) and noisy (row 3) genes of the VPA data set with significant concentration–response model, with added fit of the model selected as winner w.r.t. AIC. Gray dots represent single response values, and the red dots indicate the mean responses per concentration. Each model has an R^2_{adj} of at least 0.75 (rows 1 and 2) or below 0.25 (row 3)

4.3 | Setup for Analysis II

Analysis II offers further insights into possibly expendable models in the candidate set. The analysis setup is similar to the one of Analysis I. Analysis I is redone several times, but each time one model from the candidate model set is omitted. We refer to these as LOMO analyses.

4.4 | Results for Analysis II

The number of FDR adjusted significant concentration–response relationships is similar to the main analysis where no model is left out (Table 3). This finding is consistent with the results of Pinheiro et al. (2006). If the quadratic model is omitted from the candidate model set, the total number of signifMCPs increases at the cost of reduced mean R^2_{adj} for the remaining genes. This is due to the different, rarely appropriate shape of the quadratic model compared to all other models. Measured by the S_M score, it is proposed to drop the E_{max} model from the candidate model set (indicated in bold). The

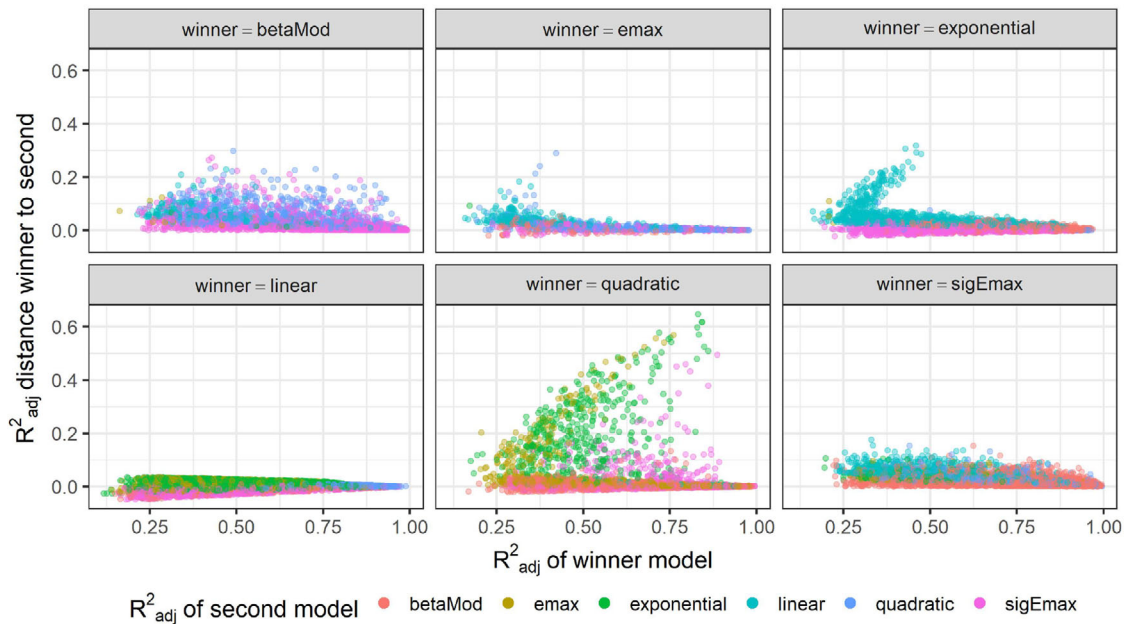


FIGURE 6 Scatter plots stratified by winner model (by AIC) showing the R^2_{adj} value of the winner on the x-axis and the difference to the R^2_{adj} value of the second best model (by AIC) on the y-axis

TABLE 3 Total number of FDR adjusted significant genes in each LOMO analysis, number of gained and lost genes compared to the case when no model is left out (Analysis I), $S_{\mathcal{M}}$ score, rate of FDR adjusted significant genes, and mean R^2_{adj} among significant genes. Largest $S_{\mathcal{M}}$ score indicated in bold

| Model | Total | Gained | Lost | $S_{\mathcal{M}}$ | signifMCP rate | mean R^2_{adj} |
|-------------------|--------|--------|------|-------------------|----------------|------------------|
| Sigmoid E_{max} | 20,122 | 61 | 132 | 0.2114 | 0.3680 | 0.5743 |
| Quadratic | 20,459 | 697 | 431 | 0.2119 | 0.3742 | 0.5664 |
| Beta | 20,221 | 150 | 122 | 0.2120 | 0.3698 | 0.5732 |
| Exponential | 20,164 | 529 | 558 | 0.2120 | 0.3688 | 0.5748 |
| Linear | 20,178 | 91 | 106 | 0.2127 | 0.3691 | 0.5763 |
| None | 20,193 | 0 | 0 | 0.2134 | 0.3693 | 0.5778 |
| E_{max} | 20,349 | 330 | 174 | 0.2138 | 0.3722 | 0.5745 |

sigmoid E_{max} model, which contains the E_{max} model as a special case, decreases the score the most when removed from the candidate set.

We are further interested by which model an originally selected model after its omission is typically substituted in the modeling step (Figure 7). The beta model is selected more often, if the sigmoid E_{max} model is removed and vice versa. If the often selected linear model is omitted, the exponential model is most often replacing it.

Two additional evaluations regarding the validity of the $S_{\mathcal{M}}$ score were conducted. First, a copy of the VPA data set was generated and all 3067 genes where the exponential model won by AIC were removed and the LOMO analyses were repeated. As expected, in this artificial scenario the $S_{\mathcal{M}}$ score suggests to drop the exponential model (Table 4, second column from the left).

Second, Analysis I was repeated but with a single model as candidate model (Table 4). When a single candidate model is used, the $S_{\mathcal{M}}$ score is always smaller than when only one or no model is omitted from the candidate model set and the original VPA data are used. The lowest score of 0.0825 is obtained if the quadratic model is the only candidate model. This is because the quadratic model shape passes the MCP-step for only 12.27% of the genes. When using only one candidate model, the sigmoid E_{max} model has the largest score of 0.2036.

The absolute differences in $S_{\mathcal{M}}$ scores might appear small but must not be misinterpreted as irrelevant. For the artificial scenario where genes with the exponential model as winner model are removed, omitting the exponential model from the candidate model set is considered reasonable by construction. Therefore, the corresponding difference in the $S_{\mathcal{M}}$ score of

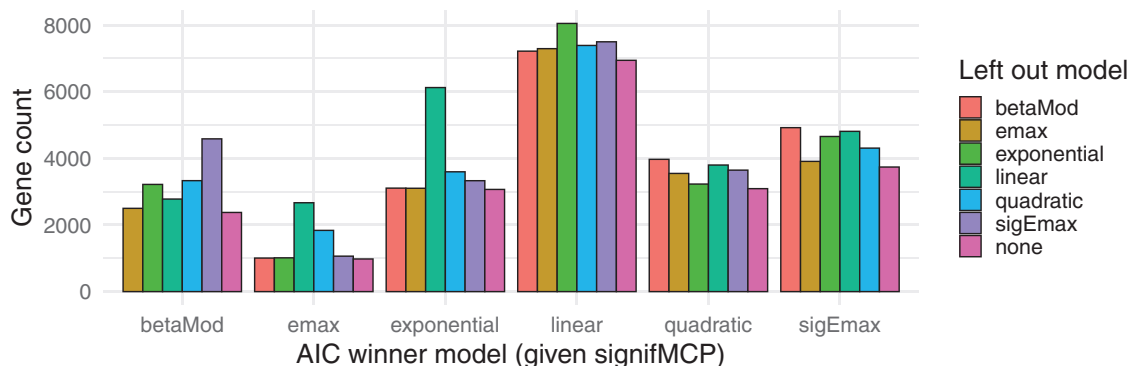


FIGURE 7 Absolute count of selections by AIC in the modeling step of each model that had a signifMCP in the MCP-step for each LOMO scenario

TABLE 4 S_M score, rate of significant genes, and mean R^2_{adj} among significant genes for two added analyses: The LOMO analyses on the modified VPA data set where genes with exponential winner model are removed and Analysis I repeated but with only one model in the candidate model set. Largest S_M score indicated in bold

| Model | LOMO analyses on VPA data set without exponential winner genes | | | Only model in candidate model set on original VPA data set | | |
|----------------|--|-------------------|------------------|--|-------------------|------------------|
| | S_M | signif. genes (%) | mean R^2_{adj} | S_M | signif. genes (%) | mean R^2_{adj} |
| sig. E_{max} | 0.1789 | 0.3062 | 0.5844 | 0.2036 | 0.3771 | 0.5399 |
| Beta | 0.1795 | 0.3077 | 0.5834 | 0.2013 | 0.3647 | 0.5520 |
| Quadratic | 0.1795 | 0.3117 | 0.5758 | 0.0825 | 0.1227 | 0.6722 |
| Linear | 0.1803 | 0.3071 | 0.5869 | 0.1855 | 0.3768 | 0.4922 |
| None | 0.1810 | 0.3078 | 0.5882 | - | - | - |
| E_{max} | 0.1813 | 0.3097 | 0.5854 | 0.1690 | 0.3235 | 0.5225 |
| Exponential | 0.1834 | 0.3173 | 0.5778 | 0.1669 | 0.3011 | 0.5544 |

0.0024 can be interpreted as relevant. Only using the sigmoid E_{max} model compared to using the full candidate model set differs by 0.0098, which can hence be viewed as a relevant difference such that it would not suffice to use a single model. The interpretation of the S_M score is not straightforward, which is discussed in Section 6.

5 | SIMULATION-BASED ANALYSIS

The simulation gives insights on the effect of the number of replications per concentration level while standard deviation of the noise is fixed (Table 2). Opposed to the data-based analysis, the correct model is known such that precision, recall, and goodness-of-fit can be evaluated.

5.1 | Setup

Concentration–response data sets are generated for 10,000 genes for each of the six considered models and for the null case, as well as for three different numbers of replicates n_i and a fixed standard deviation to range ratio (see Table 2). Details on how the range and standard deviation are chosen are explained in the following. The null case means that a constant model is used to generate the data. In order to have a realistic data generation process, it is based on the real VPA data set. For each considered n_i , a data set structurally similar to the VPA data set but with 70,000 genes, 10,000 for each of the six nonflat models, and 10,000 for the flat null model, is generated as follows.

Consider a model $f = f(d, \theta) \in \mathcal{F}$ where $|\mathcal{F}| = 7$ and for the null case, $f = f(d, \theta) = f(d, c) = c > 0$. The assumed ratio of standard deviation to range denoted by $q(0.5)$ is explained below. Define $\mathcal{G}^*(f)$ as the set of all genes g for which

TABLE 5 Summary of the simulation results, stratified by correct model and chosen model, respectively. Signif. genes (%) is the rate of FDR adjusted, detected dose–response signals among the 10,000 generated dose–response data for each model, with n_i replicates at each dose level. For the recall rate, the model is the correct model. For the precision rate, the model is the selected model

| Measure | n_i | Beta | E_{\max} | Exponential | Flat | Linear | Quadratic | sig. E_{\max} |
|-------------------|-------|-------|------------|-------------|-------|--------|-----------|-----------------|
| Signif. genes (%) | 3 | 0.989 | 0.982 | 0.979 | 0.037 | 0.988 | 0.988 | 0.997 |
| | 5 | 1.000 | 1.000 | 1.000 | 0.042 | 1.000 | 1.000 | 1.000 |
| | 10 | 1.000 | 1.000 | 1.000 | 0.045 | 1.000 | 1.000 | 1.000 |
| Recall | 3 | 0.447 | 0.496 | 0.561 | 0.963 | 0.705 | 0.543 | 0.471 |
| | 5 | 0.567 | 0.552 | 0.657 | 0.958 | 0.750 | 0.638 | 0.567 |
| | 10 | 0.712 | 0.616 | 0.754 | 0.955 | 0.770 | 0.723 | 0.717 |
| Precision | 3 | 0.582 | 0.658 | 0.682 | 0.925 | 0.416 | 0.547 | 0.510 |
| | 5 | 0.627 | 0.746 | 0.745 | 0.999 | 0.529 | 0.565 | 0.586 |
| | 10 | 0.676 | 0.790 | 0.822 | 1.000 | 0.713 | 0.614 | 0.689 |

model f was the selected winner model by AIC in the VPA data set in Analysis I. Further, $\underline{\mathcal{G}}(f)$ is a sample of 10,000 genes drawn with replacement from $\mathcal{G}^*(f)$. For a gene $g \in \underline{\mathcal{G}}(f)$, the true underlying concentration–response relationship is assumed to be the model fit $f^{(g)}(d, \hat{\theta})$ of model f on the VPA data of gene g . For the null model, the mean response $\bar{y}^{(g)}$ is used as true value for c .

Given this true concentration–response relationship of gene g , noise is added to generate a data set according to the model equation (1). For concentration levels $d \in \{d_1, \dots, d_8\}$ used in the original experiment (see Section 2), generate $y_{ij}^{(g)} = f^{(g)}(d_i, \hat{\theta}) + e_{ij}$, $j = 1, \dots, n_i$. The added noise values e_{ij} are independently drawn from $\varepsilon \sim \mathcal{N}(0, (\sigma(f^{(g)}, s))^2)$. If $f^{(g)}$ is not the null case, then $\sigma(f^{(g)}, s) := \text{range}(f^{(g)}) \cdot q(s)$, where the range for a gene with model f is calculated as $\text{range}(f^{(g)}) := \max_{d \in \{d_1, \dots, d_8\}} (f^{(g)}(d, \hat{\theta})) - \min_{d \in \{d_1, \dots, d_8\}} (f^{(g)}(d, \hat{\theta}))$. The term $q(s)$ is the empirical s -quantile of the ratio $S^{(g)}/\text{range}(f^{(g)})$ across all genes g with a detected signal and their respective fits in Analysis I. Here, $(S^{(g)})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij}^{(g)} - \bar{y}_i^{(g)})^2 / (N - k)$ is the estimated variance for each gene. Hence, for $s = 0.5$, we obtain $q(0.5) = 0.3222$, which is used for all nonflat models and all genes to calculate $\sigma(f^{(g)}, 0.5)$ (Appendix, Figure A2). If $f^{(g)}$ is the null case, then $\text{range}(f^{(g)}) = 0$. In this case, we use $\sigma(f^{(g)}, s) = q_{\text{null}}(s)$, which is the empirical s -quantile of S calculated for nonsignificant genes g of Analysis I. We obtain $q_{\text{null}}(0.5) = 0.1909$ (Appendix, Figure A3).

Using an adapted standard deviation per gene and model is preferred over using a fixed standard deviation, because it allows for comparability between different models and ranges with respect to goodness-of-fit (Kappenberg et al., 2021). Finally, the generated data set is analyzed as the original VPA data set in Analysis I.

5.2 | Results

Table 5 summarizes the results of the simulation w.r.t. signal detection (power), recall, and precision. For $n_i = 3$, which mimics the conditions in the real VPA data set, a signal is almost always detected if it is present. However, the recall and precision rates for nonlinear and nonflat models for this scenario are below 0.69. If $n_i = 10$, the rates of these model selection errors are still large, even though the sample size $N = 8 \cdot n_i = 80$ is rather large in the context of toxicology. For example, if the sigmoid E_{\max} model is correct, for 31.1% of the generated dose–response data another model is incorrectly selected. Due to the penalty term of the AIC used in model selection, complex correct models as the sigmoid E_{\max} or the beta model have a comparatively large increase in recall rates when n_i is increased. For comparatively noisy scenarios, these models are rarely selected. The opposite holds for the least complex model, the linear model. It has a comparatively very low precision rate (41.6%) and very high recall rate (70.5%) for $n_i = 3$, but not for $n_i = 10$. Precision values naturally have more practical value, as they give insight on how confident one can be with the model selection. The precision rate increases from 92.5% to 99.9% for the flat model if n_i increases from 3 to 5. For nonflat models, the precision rate does not exceed 82.2% at any n_i .

In practice, often one is not mainly interested in selecting the true underlying model but to have a sufficiently good fit. Figure 8 summarizes the relative loss in model fit by considering the log-ratio in root-mean-square error (RMSE) between the winner and the true model, that is, between the actually selected model and the fitted model if the correct

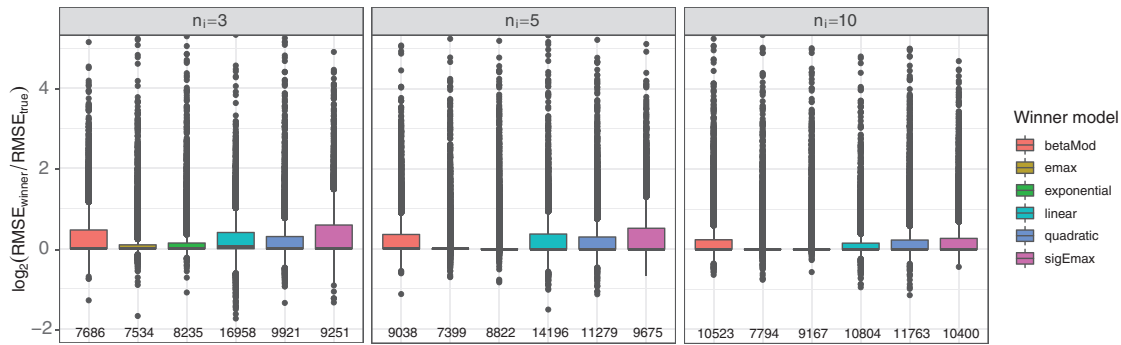


FIGURE 8 Distributions of $\log_2(RMSE_{winner}/RMSE_{true})$ in the simulation. $RMSE_{true}$ is the root mean squared error (RMSE) if the correct model is fitted and $RMSE_{winner}$ the RMSE of the selected winner model. For 108 genes, the log-ratio is greater than 5 and not displayed

model is selected. For both, the RMSE is calculated with respect to the true dose–response curve that is known in the simulation. If the correct model is actually selected, this ratio is 0, because the true and selected model are the same. If the ratio is greater than one, then the selected model differs from the correct model and has a worse RMSE. The ratio of the RMSE values is independent of n_i and of the range of the respective gene. It only captures the effect of the model selection.

In general, the relative loss in RMSE decreases with increasing n_i but is still present for $n_i = 10$, although $N = 8 \cdot n_i = 80$ is already a large sample size in toxicology. For $n_i = 3$ the median of the log-ratio is 0.0000 for all winner models, but 0.0543 for the linear model. This demonstrates the low precision of the linear model for small n_i . For small $n_i = 3$, the penalty of the AIC is comparatively large, yielding too many selections of the simpler linear model in cases where a more complex model might be required. If a more complex model such as the beta or sigmoid E_{max} model is selected, the ratio's upper quartile are largest with $2^{0.464} = 1.480$ and $2^{0.594} = 1.509$, respectively. Hence, for 25% of the generated genes where the sigmoid E_{max} model is selected, the selection is not correct and the RMSE is at least 50.9% larger than the RMSE of the correct model. For the E_{max} and for the exponential model, the upper quartiles of the ratio are closest to zero for each n_i . With increasing n_i , the penalty term of the AIC becomes comparatively weak. For $n_i = 10$, this heavily affects the linear model. It is selected less often and has log-ratios closely concentrated around zero. For the beta, quadratic, and sigmoid E_{max} model, the upper quartile of the log-ratios remain comparatively far from 0. If the beta model is selected, for 25% of the generated genes the selection is incorrect and the RMSE is at least $2^{0.235} = 1.177$ times the RMSE if the correct model was selected. Thus, not selecting the correct model results in a noteworthy relative loss in goodness-of-fit, even when larger sample sizes are used in toxicology.

6 | CONCLUSION

In this work, MCP-Mod was used as model selection approach for gene expression concentration–response data. For the data set at hand, human embryonic stem cells were exposed to varying concentrations of VPA. For 54,675 probe sets or genes the expression is measured. The data set indicated that modeling gene expression concentration–response data requires the consideration of several models, that is, a candidate model set. Only considering the popular E_{max} or sigmoid E_{max} model might not be sufficient. Especially nonmonotone models like the quadratic model should also be taken into account. When using MCP-Mod, frequent selections of a linear model should not be misinterpreted as evidence for a true, linear concentration–response relationship. A large noise-to-signal ratio, or, more precisely, a large standard deviation to true response range ratio, favors the selection of the linear model. Also, there is typically no notable loss in goodness-of-fit, when instead of the linear model the second best model is used.

Using a newly proposed score, S_M , it was observed that the E_{max} model can be omitted from the candidate set despite its popularity, as long as the more general sigmoid E_{max} model is included in the candidate set. Further, the score discourages to omit the linear model, even though it can be easily substituted with respect to goodness-of-fit. Including the linear model in the candidate set aims to detect unclear concentration–response signals rather than modeling detected signals well. If the linear model is omitted, one might fail to identify potentially interesting genes. Simulation studies based on the data set indicate that the confidence in the correctness of the selected model, measured by the precision, is not very high.

Even when increasing the sample size per concentration from 3 to 10, which is very large for this type of toxicological experiments, the precision of nonflat models does not exceed 0.83. Thus, increasing the number of experiments does not increase the precision in model selection proportionally. The relative loss in goodness-of-fit due to model selection mistakes decreases with increasing sample size, but remains notable even for 10 replicates per concentration.

The newly proposed S_M score served as a help to summarize analysis and simulation results. For a given candidate set, it considers the power, that is, number of detected genes, and the goodness-of-fit of genes with a detected signal simultaneously. Despite its simple form, its interpretability is not straightforward, which allows for improvements. If one does not want to consider both power and goodness-of-fit at the same time but, for example, focuses on optimizing power, the score is not an adequate tool.

The data basis of this work is a single data set, which, despite its size and quality, is an obvious limitation. Similar analyses on other gene expression concentration–response data would be valuable to confirm our results. Another promising approach, which is not considered in this work, is model averaging. It would be interesting to analyze the influence of the different approaches and parameters on target dose estimation.

ACKNOWLEDGMENTS

This work has received funding from LivSysTransfer (BMBF: Project Number 031L0119) and was partially supported by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, Project P2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation: Project Number 427806116). The authors would like to thank the anonymous reviewers for their numerous helpful comments.


CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The original data that support the findings of this study are openly available in ArrayExpress as stated by Krug et al. (2013) (<https://doi.org/10.1007/s00204-012-0967-3>).

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Julia C. Duda  <https://orcid.org/0000-0002-3894-0124>

Franziska Kappenberg  <https://orcid.org/0000-0001-8066-5333>

Jörg Rahnenführer  <https://orcid.org/0000-0002-8947-440X>

REFERENCES

- Abelson, R. P., & Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis general theory and the case of simple order. *The Annals of Mathematical Statistics*, 34(4), 1347–1369.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.
- Bornkamp, B., Bretz, F., Dette, H., & Pinheiro, J. (2011). Response-adaptive dose-finding under model uncertainty. *The Annals of Applied Statistics*, 5(2B), <https://doi.org/10.1214/10-aos445>
- Bornkamp, B., Pinheiro, J., & Bretz, F. (2009). MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies. *Journal of Statistical Software*, 29(7), <https://doi.org/10.18637/jss.v029.i07>
- Bretz, F., Pinheiro, J. C., & Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3), 738–748.
- Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12–19.
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.
- Cotariu, D., Zaidman, J. L. (1991). Developmental toxicity of valproic acid. *Life Sciences*, 48(14), 1341–1350.

- European Medicines Agency (2015). *Qualification opinion of mcp-mod as an efficient statistical methodology for model-based design and analysis of phase II dose finding studies under model uncertainty*. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/draft-qualification-opinion-mcp-mod-efficient-statistical-methodology-model-based-design-analysis_en.pdf.
- Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S., & Martin, M. T. (2016). tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics*, *btw680*, <https://doi.org/10.1093/bioinformatics/btw680>
- Food and Drug Administration. (2016). *Determination letter*. <https://www.fda.gov/media/99313/download>.
- Genton, P., Semah, F., & Trinka, E. (2006). Valproic acid in epilepsy. *Drug Safety*, *29*(1), 1–21.
- House, J. S., Grimm, F. A., Jima, D. D., Zhou, Y.-H., Rusyn, I., & Wright, F. A. (2017). A pipeline for high-throughput concentration response modeling of gene expression for toxicogenomics. *Frontiers in Genetics*, *8*, 168.
- Kappenberg, F., Grinberg, M., Jiang, X., Kopp-Schneider, A., Hengstler, J. G., & Rahnenführer, J. (2021). Comparison of observation-based and model-based identification of alert concentrations from concentration–expression data. *Bioinformatics*, *37*(14), 1990–1996.
- Krug, A. K., Kolde, R., Gaspar, J. A., Rempel, E., Balmer, N. V., Meganathan, K., Vojnits, K., Baquié, M., Waldmann, T., Ensenat-Waser, R., Evans, R. M., Julien, S., Kortenkamp, A., Hescheler, J., Hothorn, L., Bremer, S., van Thriel, C., Krause, K.-H., Hengstler, J. G., Rahnenführer, J., Leist, M., & Sachinidis, A. (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: A transcriptomics approach. *Archives of Toxicology*, *87*(1), 123–143.
- O’Quigley, J., Iasonos, A., & Bornkamp, B. (2017). *Handbook of methods for designing, monitoring, and analyzing dose-finding trials*. Handbooks of Modern Statistical Methods, CRC Press.
- Pinheiro, J., Bornkamp, B., & Bretz, F. (2006). Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics*, *16*(5), 639–656.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Schorning, K., Bornkamp, B., Bretz, F., & Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, *35*(22), 4021–4040.
- Thomas, N., Sweeney, K., & Somayaji, V. (2014). Meta-analysis of clinical dose–response in a large drug development portfolio. *Statistics in Biopharmaceutical Research*, *6*(4), 302–317.
- Ting, N. (2006). *Dose finding in drug development*. Springer Science & Business Media.
- Wheeler, M. W., & Bailer, A. J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics*, *16*(1), 37–51.
- Xun, X. & Bretz, F. (2017). The MCP-Mod methodology: Practical considerations and the dose finding r package. In J. O’Quigley, A. Iasonos, & B. Bornkamp (Eds.), *Handbook of methods for designing, monitoring and analyzing dose-finding trials* (pp. 205–227). Chapman and Hall/CRC.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Duda, J.C., Kappenberg, F., & Rahnenführer, J. (2022). Model selection characteristics when using MCP-Mod for dose–response gene expression data. *Biometrical Journal*, *64*, 883–897. <https://doi.org/10.1002/bimj.202000250>

APPENDIX

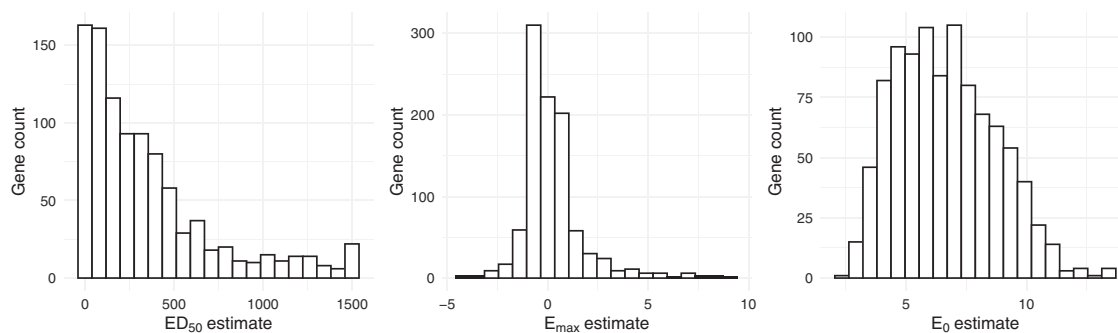


FIGURE A1 Barplots of parameter estimates of the E_{\max} model for genes where it was selected as winner model by the AIC

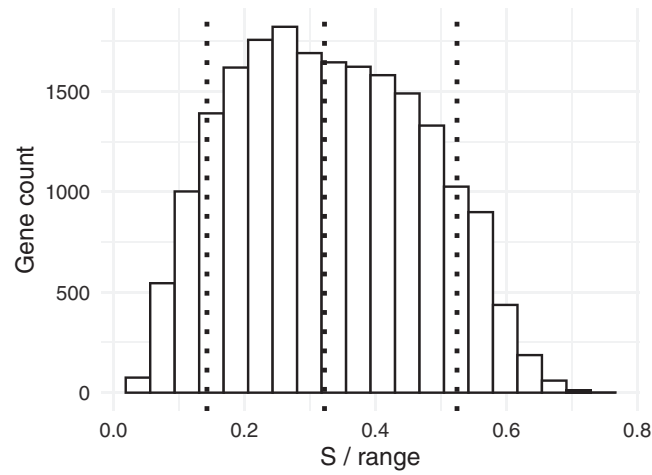


FIGURE A2 Barplot of S to range ratio for genes with an FDR adjusted signifMCP in Analysis I. The vertical dotted lines are at $q(0.1) = 0.1427$, $q(0.5) = 0.3222$, and $q(0.9) = 0.5245$

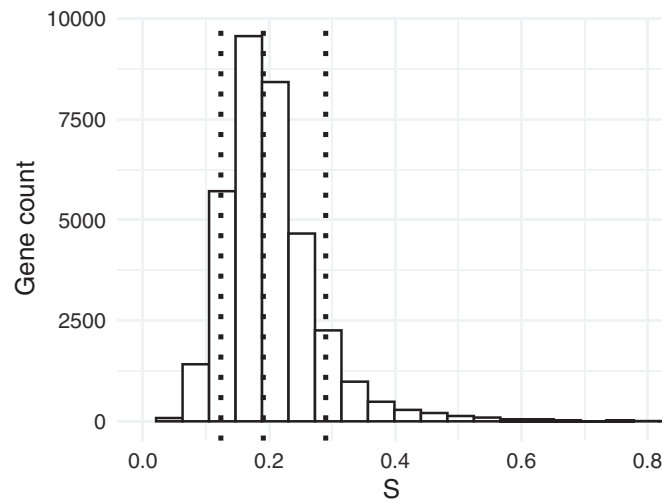


FIGURE A3 Barplot of S for nonsignificant genes in Analysis I. The vertical dotted lines are at $q_{\text{null}}(0.1) = 0.1236$, $q_{\text{null}}(0.5) = 0.1909$, and $q_{\text{null}}(0.9) = 0.2898$