technische universität dortmund

fakultät statistik

Statistical approaches for calculating alert concentrations from cytotoxicity and gene expression data

DISSERTATION
in Fullfilment of
the Requirements for the Degree of
Doktor der Naturwissenschaften

Submitted to the
Department of Statistics
of the
TU Dortmund University

by
Franziska Kappenberg

on
January 28, 2021

Referees:
Prof. Dr. Jörg Rahnenführer
JProf. Dr. Kirsten Schorning

Date of Oral Examination:
March 23, 2021

Danksagung

Mein großer Dank gilt vor allem meinem Doktorvater Prof. Dr. Jörg Rahnenführer, der mich in allen Phasen der Arbeit hervorragend begleitet und immer mit den richtigen Worten ermutigt hat. Er hat mir zum einen viele spannende Themen der Toxikologie nahe gebracht, und mir zum anderen die Freiheit für die Forschung an eigenen Ideen gelassen, wobei ich mir seiner Unterstützung und der Bereitschaft, statistische Fragestellungen zu diskutieren, jederzeit sicher sein konnte.

Ich möchte mich bei Prof. Dr. Jan G. Hengstler und Prof. Dr. Marcel Leist und ihren wissenschaftlichen Mitarbeiterinnen und Mitarbeitern für die gute Zusammenarbeit, die Bereitstellung echter Datensätze und das geduldige Erklären der biologischen Sachverhalte bedanken.

Weiterhin möchte ich mich bei allen Kolleginnen und Kollegen an der Fakultät Statistik für zahlreiche spannende Diskussionen und hilfreiche Anregungen bedanken. Meiner Zweitgutachterin JProf. Dr. Kirsten Schorning danke ich besonders für viele gute Hinweise und interessante Ausblicke bezüglich weiterer Anwendungsmöglichkeiten meiner Forschungsgebiete.

Bei meiner Familie möchte ich mich ganz herzlich dafür bedanken, dass sie im Verlauf meines gesamten Studiums immer für mich da waren.

Contents

1. Introduction

Dose-response or concentration-response studies are aimed at assessing the effects of the exposure of some condition, e.g. the treatment with a specific compound, on living beings or on cells. The central goal in toxicological assays, where concentrations are considered, is to find an *alert concentration*, where a pre-specified effect level is attained or exceeded by the response variable of interest. For this purpose, measurements are taken for several increasing concentrations of the specific compounds. Additionally, response values for the negative control, corresponding to the concentration 0, are usually measured. In the context of clinical studies, the equivalent is given by dose-finding studies, in which different doses of a compound and their effect on patients are considered.

Depending on the type of response data measured, different statistical tests can be conducted to compare results of the considered concentrations to the results obtained for the control. For a continuous response variable, the Dunnett procedure (Dunnett, 1955) and the Williams procedure (Williams, 1971) are two established methods. In both procedures, responses for multiple concentrations are simultaneously compared against the control, with the Williams procedure summarising several concentrations to determine trends in the data. When the response is given by binary data, i.e. proportions, the Cochran-Armitage test (Cochran, 1954; Armitage, 1955) allows statistical testing to find trends in the data that are similar to a set of pre-specified scores.

When considering the measured concentrations only, alert concentrations as the *no observed effect concentration* (NOEC), which is the highest of the measured concentrations where no significant effect can be observed, or the *lowest observed effective concentration* (LOEC), which is the lowest of the measured concentrations where a significant effect can be estimated, are calculated (Delignette-Muller et al., 2011). When, instead of the significant effect required by the LOEC, only an absolute effect is of interest, the corresponding concentration is called the *absolute lowest observed effective concentration* (ALOEC) (Grinberg, 2017). An obvious drawback of these alert concentrations is the limitation of potential alert concentrations to the set of measured concentrations.

In contrast to these situations, where the concentration or dose is assumed to be a qualitative factor, in modelling, it is assumed to be a quantitative variable. A parametric, usually non-linear function is assumed to describe the relationship between concentration and response. Typical functions to describe this relationship are, among others, given by the family of log-logistic functions, the family of log-normal functions, and the Weibull functions (e.g. Ritz et al., 2019, pp. 178-186). In this thesis, only the family of log-logistic functions, specifically the four-parameter log-logistic function, is considered.

In toxicology, typical continuous response variables are the viability of cells, and gene expression values. The viability of cells is usually measured and normalised to obtain percentages, thus asymptote values of 100% and 0% of the fitted function are plausible results. Additionally, it is typically known in advance whether the concentration-response effect shows an increasing or a decreasing pattern. Gene expression values, however, have no fixed values for upper and lower asymptote, or for the interval of covered values. The direction of the pattern is usually not known either. This leads to the fact

that, depending on the application, different definitions of the alert concentrations are required.

One family of alert concentrations that are often used, especially for cytotoxicity assays, is given by the *effective concentrations* (EC values). It is differentiated in absolute and relative EC values. For $\lambda \in (0, 100)$, the absolute EC value EC_λ indicates the concentration where the response attains the specific response value $\lambda\%$ or $100 - \lambda\%$, while the relative EC value $\text{EC}\lambda$ corresponds to the concentration where $\lambda\%$ of the maximal observed effect are attained (Ritz et al., 2019, pp. 173-174). Both values coincide in cases where the upper and lower asymptote of a fitted curve correspond to values of 100% and 0%, respectively. While the absolute EC values are considered especially for curves indicating viability, where the assumption of asymptotes corresponding to 100% and 0% are sensible, estimation of the relative EC values heavily depends on the actual observed values of the asymptotes.

A similar approach to the calculation of relative effective concentrations is the *benchmark dose* (BMD) methodology, dating back to Crump (1984). Calculation of the BMD is based on the definition of a benchmark risk (BMR), i.e. a small increase above the observed background risk. Depending on the specific scenario regarding the type of response data and the type of risk considered (e.g. excess risk or additional risk), the BMD is calculated as concentration where the curve attains a response value that is determined by a linkage of the BMR and the background risk. The lower limit of the confidence interval for the BMD then serves as estimate for the *point of departure*, which is defined as the lowest concentration where a response that differs from the background risk is observed (Jensen et al., 2019). Usually, the BMR is chosen by a low percentage, e.g. 10% above the normal response. Zeller et al. (2017) propose different methods to define the BMR based on the evaluation of historical control data. Basing the estimation of relevant alert concentrations on a curve instead of an individual concentration has the additional advantage that the entire information of the concentration-response profile is mathematically included in the estimation of the curve (Izadi et al., 2012).

In clinical dose-finding studies, a different alert is usually considered. The *minimum effective dose* (MED), dating back to Ruberg (1995), is defined as the smallest dose where the modelled response value exceeds the modelled response of the lowest concentration considered plus some biologically relevant threshold. Bretz et al. (2005) propose three estimators for the MED that take the confidence interval of the modelled response value and the biologically relevant threshold into account.

Three specific aspects of calculating alert concentrations from concentration-response data are considered in this thesis. The first topic is the handling of deviating control values in cytotoxicity assays. This describes the phenomenon that response values for the negative control deviate from the response values observed for the lowest measured concentrations. Hence, an upper asymptote is obtained that does not correspond to a viability of 100% when fitting a parametric curve to this data (Krebs et al., 2018). When the alert concentration of interest is given by absolute EC values, interpretation of these values becomes meaningless when the upper asymptote does not correspond to 100%, or calculation of the respective EC value may even become impossible. The extent of this problem is found to be relevant by an extensive literature review.

Four methods are proposed to deal with this problem, all pursuing the goal to obtain a curve whose upper asymptote, or in one exception the maximum value of the curve, corresponds to a viability of 100%. These methods include one method that is based on a re-normalisation procedure, one method where the upper asymptote is forced to attain a value of 100%, one method where the controls are completely omitted and one method where negative deviations are included in the modelling. The first three methods yield monotonously decreasing curves, but the fourth method may yield a curve that shows an increase before it is monotonously decreasing. These four methods are compared in a controlled simulation study, where the goal is the most precise estimation of different EC values. Results are interpreted with respect to the proportions of estimates that are acceptably close to the true underlying value, and with respect to the number of times each method yields the most precise estimate. Based on the results, a set of concrete recommendations, which method to use in which case, is derived.

In the second topic, gene expression data are considered. The observation-based alert concentration LOEC determines the lowest concentration where the fold change, i.e. the mean difference in gene expression, in comparison to the control, exceeds a pre-specified threshold. This alert concentration is well-established in this context. The ALOEC is the alternative when only absolute exceedance of the threshold is of interest. When moving to continuous, model-based alert concentrations, the *absolute lowest effective concentration* (ALEC) can be calculated as equivalent to the ALOEC (Jiang, 2013). In this thesis, a continuous equivalent to the LOEC, called *lowest effective concentration* (LEC) is proposed. For this alert concentration, a model-based statistical test, based on previous work by Grinberg (2017), is introduced and the alert concentration itself is determined making use of a search algorithm.

This yields four methods for calculating alert concentrations, two of which only consider the actually measured concentrations as potential alert concentrations, and two of which are based on fitting a curve, such that every positive concentration is a potential alert concentration. In each case, one of the methods considers absolute exceedance of the threshold only and the other additionally takes significance into account. These methods are compared in a simulation study covering several scenarios. Generally, the methods that are based on fitting a curve, ALEC and LEC, less drastically overestimate the underlying true alert concentrations. At the same time, the number of results that are lower than the true underlying alert concentration is not too high. In particular, all methods based on statistical testing maintain the respective significance level.

The third topic also deals with gene expression data. Concentration-response profiles, where the response is given by gene expression values, are measured simultaneously for thousands of genes in microarray analyses. Two approaches are presented that are aimed at sharing information across these genes and thus potentially improve the estimation of the parameter indicating the half-maximal effect. The idea to share information across genes is a relaxation of the approach of *common parameters*, where several curves are simultaneously estimated under the assumption that one or several parameters are equal for all curves (Feller et al., 2017). The first approach for information sharing is to employ a meta-analysis strategy for several fitted curves, as proposed by Jiang and Kopp-Schneider (2014). For each gene individually, the set of other genes to be included

into the meta-analysis is defined by similarity in terms of correlation. Considering the entire set of genes as potential similar genes, however, has the effect of adding noise rather than information to improve the fit. Thus, the set of potential similar genes is restricted to a small group of genes that share some biological properties. Still, in simulation studies, the meta-analysis approach does not show an improvement in the estimation.

The second approach is based on an empirical Bayes method. For each gene, a direct estimate of the parameter of interest is calculated. In essence, a weighted mean of this direct estimate and the mean of all estimates of the entire set of genes is calculated as result of the Bayes method. The weights are determined based on the standard error of the direct estimation and the variance of the estimates of the entire set. This method is assessed in three simulation studies that differ in their degree of similarity to a real situation on the one hand and the degree to which the required assumptions are fulfilled on the other hand. Generally, the Bayes method leads to an improvement of the mean squared errors in all situations, while the coverage probability of the obtained credible intervals is not decreased in comparison to the coverage probability of the confidence intervals for the direct estimate.

The topic about the handling deviating controls occurs in the context of cytotoxicity data, and the other two topics in the context of gene expression data. For all three topics, the different methods proposed in this thesis are compared in controlled simulation studies. Additionally, the methods are applied to real-data situations and interpreted with respect to the results from the simulation studies. One cytotoxicity and one gene-expression dataset are considered, where for both datasets, the respective response is measured for increasing concentrations of the compound valproic acid. Since in contrast to controlled simulation studies, the underlying true effect is not known in this case, application of the methods to the real datasets leads to results where the methods can only be compared with each other.

The methods and results regarding the handling of deviating controls are published in

> F. Kappenberg, T. Brecklinghaus, W. Albrecht, J. Blum, C. van der Wurp, M. Leist, J. G. Hengstler, and J. Rahnenführer. Handling deviating control values in concentration-response curves. *Archives of Toxicology*, 94(11):3787 – 3798, 2020.

The contributions of the authors, referred to by the respective first letter of first name and surname, are as follows:

- Formulating and discussing the problem: FK, CW, ML, JR
- Defining the set of criteria for the literature review: FK, TB, WA, JH, JR
- Performing the literature review: TB, WA, JB
- Analysing the literature review: FK
- Designing the simulation studies: FK, CW, JR
- Providing the real data: TB, JH

- Executing the simulation studies and the real data analysis: FK
- Interpreting all results: FK, ML, JH, JR
- Writing the initial version of the manuscript: FK
- Supervision of the project: JR
- All authors read, corrected and approved the manuscript.

The methods and results regarding the identification of alert concentrations are published in

> F. Kappenberg, M. Grinberg, X. Jiang, A. Kopp-Schneider, J. G. Hengstler, and J. Rahnenführer. Comparison of observation-based and model-based identification of alert concentrations from concentration-expression data. *Bioinformatics*, 2021. btab043.

The contributions of the authors, referred to by the respective first letter of first name and surname, are as follows:

- Formulating and discussing the problem: FK, MG, XJ, AKS, JR
- Providing the statistical basis for the new methods: MG, XJ, AKS, JR
- Deriving the statistical methods: FK, JR
- Designing the simulation studies: FK, MG, JR
- Providing the real data: JH
- Executing the simulation studies and the real data analysis: FK
- Interpreting all results: FK, JR
- Writing the initial version of the manuscript: FK
- Supervision of the project: JR
- All authors read, corrected and approved the manuscript.

This thesis is structured as follows: In Chapter 2, a detailed introduction for each of the three topics covered in this work is given. The underlying problem or challenge is explained, without getting too specific regarding the statistics, and background from already published literature is given. The statistical methods are explained heuristically and the specific goals with respect to the applications are presented. The used datasets and the biological background are introduced in Chapter 3. In particular, the Affymetrix GeneChip® data with the RMA pre-processing algorithm are presented.

The statistical methods used throughout this thesis are introduced in Chapter 4. This chapter starts with an introduction into the basics of concentration-response analysis, by introducing the Dunnett procedure, modelling concentration-response curves for the specific example of log-logistic models, and giving an overview over alert concentrations of interest. The specific methods used for the three topics addressed in this thesis are explained separately, beginning with the methods for handling deviating control values. Next, the four different alert concentrations compared in the second topic are introduced

and summarised while emphasising their similarities and differences. Finally, the two approaches for sharing information across genes are introduced together with the concept of plasmode simulation studies that are used for this topic. This chapter ends by giving a short overview over the software used for the analyses conducted in this thesis.

The results for the three topics considered in this thesis are presented in three separate chapters. In Chapter 5, the results regarding the handling of deviating control values are presented. Results comprise the results from the literature review, the setup of the simulation study, results of the simulation study, and, based on these results, a concrete set of recommendations which method to use in which case. Finally, the methods are applied to a real cytotoxicity dataset. Results regarding the identification of alert concentrations are summarised in Chapter 6. This chapter is structured into the three parts setup of the simulation study, results from the simulation study and the application to a real concentration-expression dataset.

The results regarding the sharing of information across genes are presented in Chapter 7. This chapter starts with an extensive descriptive analysis of a real dataset. Next, the simulation results for the meta-analysis are presented, followed by the simulation results for the empirical Bayes method. Finally, the meta-analysis and the Bayes method are both applied to a real dataset. This thesis is concluded by a summary of all the observed results in Chapter 8 and by an outlook how to extend the methods introduced here.

2. Objectives

Three different aspects of calculating alert concentrations from a concentration-response curve, fitted to toxicological in vitro data, are discussed in this thesis. These aspect are introduced in detail in this chapter. The first aspect is the occurrence and the handling of deviating control values. This is a biological phenomenon occurring in cytotoxicity assays, where response values of the negative control and the lowest tested concentrations differ from each other. The second topic deals with different methods of determining alert concentrations, mainly in the context of gene-expression data. An observation-based and a model-based approach are compared. The approaches are evaluated in two ways each, determining both an absolute alert concentration and a significant alert concentration. In the third topic, the goal is to exploit similarities in many concentration-gene expression profiles, measured for the same compound but a large set of genes, to achieve a higher-quality fit of a curve with respect to the parameter that corresponds to the half-maximal concentration.

2.1. Handling deviating control values

The problem of deviating control values in cytotoxicity assays was brought up by Krebs et al. (2018). The problem occurs in the situation of assays in which a cell function, such as viability, is assessed for a (negative) control and increasing concentrations of a compound. Typically, all response values are normalised with respect to the control in order to obtain percentages. The mean response value for the control then corresponds to a response of 100%. A concentration-response model is fitted to the data and based on this model, relevant concentrations such as effective concentrations (see Chapter 4.1.4) are calculated.

Problems may occur when the response values of the control and the response values for the lowest tested concentrations differ from each other. This situation is known as *deviating controls*. In that case, the upper asymptote of a fitted concentration-response curve may not correspond to 100%, since the responses for the lowest tested concentrations indicate a value smaller or larger than 100%. This leads to an inability to properly calculate and interpret the effective concentrations of interest.

In an extreme case, one might be interested in the concentration that corresponds to a viability of, say, 90%. Negatively deviating controls can then lead to the case in which an upper asymptote of a fitted model takes a value that is smaller than 90%, so that the concentration of interest cannot be calculated. However, the problem becomes clear already in less extreme cases: If due to deviating controls the upper asymptote does not correspond to a concentration of 100%, the intersection of the fitted curve with a pre-specified fixed response becomes meaningless.

An intuitive approach may be to force the upper asymptote to take a value of 100%, after normalising the data with respect to the response of the control values. A typical problem with this approach is illustrated in Figure 2.1, where a hypothetical example is shown: The viability of cells is assessed for a narrow series of concentrations and the responses

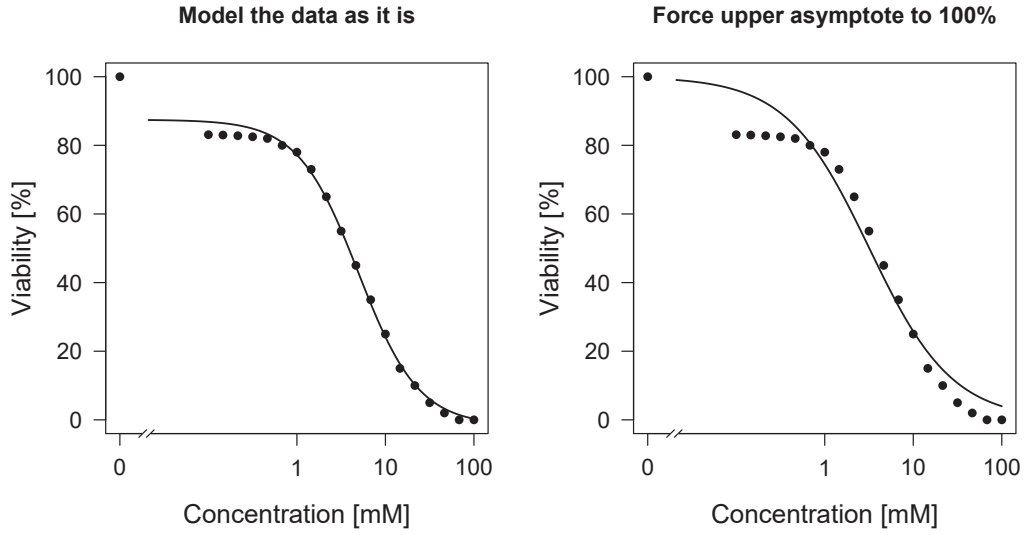

Figure 2.1: Hypothetical example illustrating a possible problem when forcing the upper asymptote of a fitted curve to 100%: In the left plot, the data is modelled as it is, and for higher concentrations, a good fit of the responses can be obtained. In the right plot, the upper asymptote is forced to take a value of 100%, which results in a poor fit also for higher concentrations.

of the lowest measured concentrations quite clearly define the upper asymptote. The response for the control, however, is (strongly) deviating from the values of the lowest measured concentrations.

In the left plot, a sigmoidal curve is fitted to the data and the value of the upper asymptote is only determined by the data. The asymptote of the fitted curve neither perfectly describes the responses for the lowest measured concentrations nor corresponds to a value of 100%. Still, for higher concentrations the curve offers a very good fit of the observed viabilities. In the right plot, however, the upper asymptote is forced to attain a value of 100%, while the other parameters (lower asymptote, inflection point and slope) are determined by the data. It becomes obvious that forcing the upper asymptote to attain a value of 100% has an effect on the entire course of the fitted curve, which in this example does not fit the data well. This method of forcing the upper asymptote to take a value of 100% places a lot of weight on the response values of the controls, as the data is normalised with respect to these values and then the asymptote is determined only on the basis of these values.

Looking at this quite extreme example, the next intuitive idea might be to simply omit the obviously deviating response value of the controls. In general, this is a difficult recommendation to give, as omitting the controls requires high-quality data for the measured concentrations. A loss in information is deliberately accepted when omitting the controls, and in some situations, information obtained by (slightly) deviating controls can still be valuable in receiving a higher-quality fit of the concentration-response curve.

Krebs et al. (2018) propose a solution to the problem of deviating controls that is based on a re-normalisation approach for the curve fitting: In a first step, it is visually decided whether the controls are to be kept or omitted. Afterwards, a model with a data-driven value of the upper asymptote is fitted, the data is re-normalised with respect to the upper asymptote and a new model is fitted in which the upper asymptote is forced to take a value of 100%. This proposal combines and extends the two intuitive ideas of omitting the controls or forcing the upper asymptote to take a value of 100%.

The problem of deviating controls, together with the three briefly introduced ideas of addressing this problem, is further illustrated in Figure 2.2. This figure makes use of a part of the real-data example that is explained in more detail in Chapter 5.5. A subset of the dataset, where the viability of cells treated with increasing concentrations plus a negative control of the compound valproic acid is observed, is shown there. The subset consists of five increasing concentrations with three replicates each. Four non-linear sigmoidal models are fitted to the data.

The top-left plot shows the data normalised to the mean of the response values for the control, with a model fitted to the data as it is. A slight deviation of the controls in the direction that their response value is larger than the response value of the two lowest concentrations can be observed. The three other plots show different methods of dealing with the deviating controls. The concentration of interest in this example is the concentration where the fitted curve attains a value of 80%, this value is called EC_{20} (see Chapter 4.1.4).

The top-right plot works with the re-normalisation procedure as proposed by Krebs et al. (2018), in which the data is normalised with respect to the upper asymptote after an initial fit. In the bottom-left plot, the upper asymptote is forced to attain a value of 100% and in the bottom-right plot, the control values are omitted from the model fit. These methods are explained in more detail in Chapter 4.2. A difference both in the course of the asymptote relative to the data points and in the concentration where a value of 80% is attained can be observed across the three plots. This difference illustrates the influence of the fitting procedure on the resulting curve and its derived parameters. Therefore, a comparison of possible methods in a controlled simulation study is required.

Since deviating controls can only be identified when the upper asymptote defined by the low- or no-toxicity concentrations is known, a careful experimental design is necessary. In cases where not enough measurements in a concentration range with no toxicity of the compound are available, the value of the upper asymptote can only be derived by the rest of the dataset and deviating controls cannot be identified with certainty. The biological reasons behind deviating controls in cases with a sensible experimental design are not entirely clear, but several guesses exist. In Kappenberg et al. (2020), the following list of reasons is stated:

- Random variation of data points due to experimental imperfections
- Errors during the performance of the experiments, e.g. in producing the stock solutions
- Variation in the concentration of solvents between samples

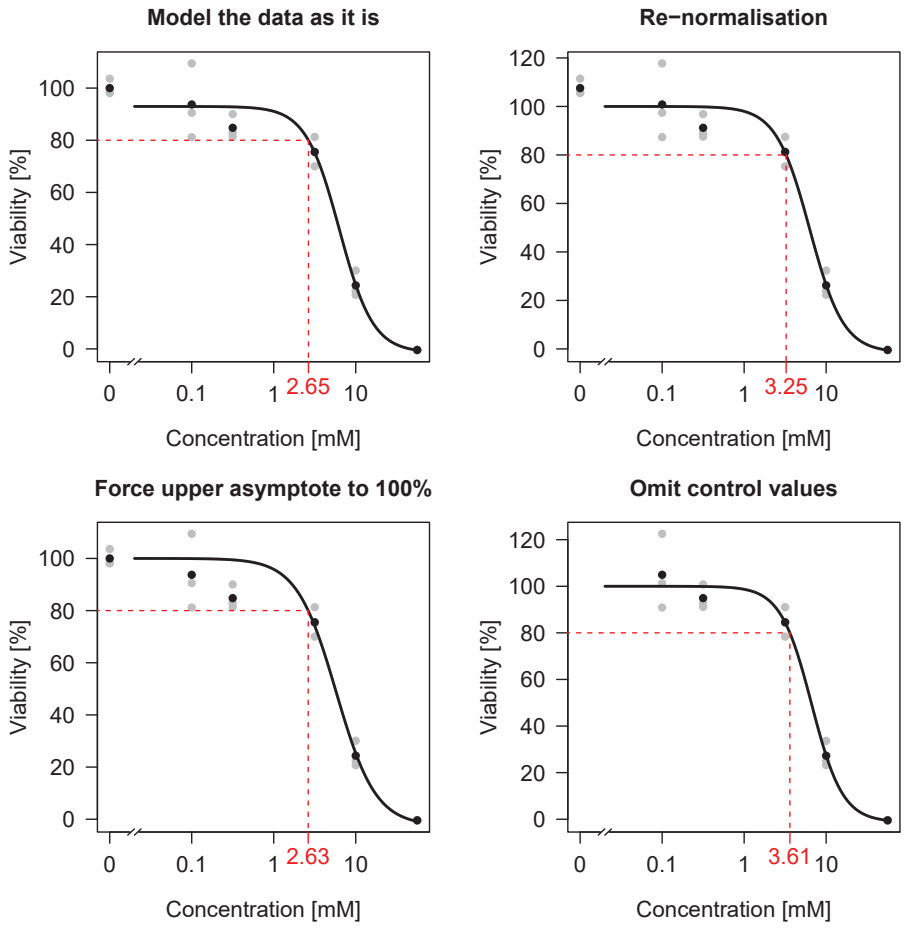

Figure 2.2: Exemplary presentation of the influence of the fitting procedure on the final curve and on derived parameters. The dataset is an excerpt of a real dataset in which the effect of increasing concentrations of VPA on the viability of cells is measured. Four different methods are used to fit a model to the data.

- Systematic deviation of endpoint readouts according to their position on culture plates or in analytical devices
- Systematic deviations due to the timing of sample preparation (e.g. incubation of cells, storage of solutions or during analysis, etc.)

In Krebs et al. (2018), the observation is made, based on a small study, that despite a frequent occurence of the deviating controls, this is not considered when fitting a model to the data. The study consisted of 100 posters that were assessed with regard to deviating controls at a toxicology meeting. In extension of this small study, an extensive literature review in three leading toxicological journals (Archives of Toxicology, Toxicological Sciences, Toxicology in Vitro) is conducted and presented in Chapter 5.1 of this thesis. This literature review serves to answer the questions how often this

phenomenon of deviating controls actually occurs in the published literature and how strong the deviations of the controls are in these cases.

Building on the proposal of Krebs et al. (2018) and the intuitive ideas how to handle the problem of deviating controls in the statistical analysis, four different methods are introduced in Chapter 4.2 and compared in a systematic simulation study (Chapter 5.3). Basic ideas of the four models are:

- Re-normalising the data after an initial fit, similar to the method described in Krebs et al. (2018)
- Forcing the upper asymptote to take a value of 100%
- Omitting the controls
- Taking the deviations into account when modelling the concentration-response curve

The goal is to derive a set of recommendations, which of the methods to use in which situation. These recommendations are explicitly stated in Chapter 5.4. All four methods are finally evaluated on a real cytotoxicity dataset, in which the effect of the compound valproic acid on the viability of cells is examined (Chapter 5.5). This dataset is introduced in detail in Chapter 3.2.1.

The results concerning the problems of deviating control values are published in Kappenberg et al. (2020). Analyses published there are extended by several aspects in this work: The literature review is analysed in more detail. In the simulation studies, additional simulation scenarios and an additional alert concentration are considered. Furthermore, an additional real cytotoxicity dataset is evaluated, and all analyses regarding real datasets are performed in more detail.

2.2. Identification of alert concentrations

An important aspect when modelling concentration-response curves is the determination of an alert concentration, i.e. the concentration where a specific level of the response is attained. In contrast to cytotoxicity data, where response values typically range from 100% to 0%, upper and lower limits of gene expression data are usually not pre-specified and may differ between genes. Therefore, specifying absolute response values that need to be attained is less straight-forward. Instead, response values depending on the actual values of the upper and lower limit need to be considered.

One possibility to define the concentration of interest is to consider the concentration where the *fold change*, i.e. the difference in gene expression between two concentrations, in comparison to the control exceeds a pre-specified level. A typical situation is the case where an effect level for the fold change is pre-specified and the smallest concentration is of interest, where this effect level is attained or exceeded when comparing to the control. In the context of concentration-gene expression data, considering the measured concentrations as potential alert concentrations is still more common than fitting a parametric curve to the data. A curve, however, allows any positive concentration

as alert concentration, in contrast to solely considering the often only few measured concentrations.

The two approaches of considering measured concentrations only or first fitting a parametric model are referred to as *observation-based* and *model-based* approaches, respectively. A further differentiation is given by whether an effect level only needs to be attained or needs to be significantly exceeded. This differentiation is referred to as *absolute exceedance* in contrast to *significant exceedance* of the effect level.

The observation-based approach where only absolute exceedance of the effect level is required can be determined by simply calculating the fold changes between the responses for each measured concentrations and the control. If additionally significance is required, some statistical testing needs to be carried out, for example a two-sample t-test or the Dunnett procedure. The latter is a common procedure in toxicology for simultaneous comparisons of several concentration versus a control and is introduced in detail in Chapter 4.1.1.

For the model-based approaches, each positive concentration is a potential alert concentration. For absolute exceedance of the effect level, the concentration is sought where the parametric curve attains the respective value. This concentration can be calculated via the inverse of the function or with numeric methods. When additionally significance is required, the approach is less straight-forward than for the observation-based methods. Grinberg (2017) introduced a first version of a test statistic together with a search algorithm, aimed at finding the model-based alert concentration that takes significance into account. In this work, the test statistic is introduced in a modified way.

In Grinberg (2017), a simulation study is conducted to compare the different methods for calculating alert concentrations briefly introduced above. A test based on the fitting of a log-logistic model is derived to find the concentration where the pre-specified effect level is significantly exceeded. Four different situations of concentration-gene expression profiles are considered. Results suggest that model-based alert concentrations are generally observed at lower concentrations than observation-based ones. For curves in which the pre-specified effect level is not exceeded or for curves with a clear sigmoidal shape, strictly better results are achieved with the model-based method. However, in situations where the concentration-gene expression profile is an unsaturated curve, the newly proposed method fails more often in yielding a valid estimate in comparison to classical observation-based approaches.

In this work, this simulation study is resumed and changed with respect to aspects regarding both the underlying methodology and the simulation itself. Concerning the methodology, the log-logistic model-based test is improved by incorporating a covariance term: In Grinberg (2017), a test statistic is introduced that aims at finding the concentration x where an effect level of λ is significantly exceeded, based on the assumption that the lower asymptote of the fitted curve attains a value of 0. The test statistic is based on the idea of calculating a confidence interval for the effect level calculated as a parametric function of the concentration. Only the variance of the fitted function at the specific concentration x is considered and the covariance between the fitted function at the concentration x and the concentration 0 is not taken into account.

This concept is broadened to any value for the lower asymptote c in finding a concentration x, where the difference of the function evaluated at x and the difference of the function evaluated at 0, yielding the response value c, significantly exceeds a pre-specified effect level. The covariance term between the function evaluated at both concentrations x and 0 cannot be neglected in general and is therefore included in calculating the test statistic.

The search algorithm introduced in Grinberg (2017), a version of the bisection algorithm, to find the smallest concentration x where the effect level is significantly exceeded is adopted for this work. The setup of the simulation study differs from the one conducted in Grinberg (2017) in several aspects:

- The true underlying curves of the simulation study are chosen slightly differently. In Grinberg (2017), four scenarios are considered. Only three of these scenarios are represented in this work and they are possibly slightly adjusted:
  - The first scenario corresponds to the 'null hypothesis', where the true underlying curve never exceeds the threshold. A similar scenario is chosen in this work as well, but with the inflection point of the curve at a higher concentration.
  - The second scenario is chosen in a way that the curve exceeds the threshold, but the upper limit is not reached within the range of considered concentrations. This scenario is adopted here as it is.
  - The third scenario describes the situation of a saturated sigmoidal curve. The basic idea of this curve is retained, but the specific parameters are chosen differently.
  - The fourth scenario forms a compromise between the first and the second scenario and is given by a curve with an upper asymptote close to the chosen threshold. No such scenario is considered in this work.
- For calculating the observation-based alert concentration that takes significance into account, Grinberg (2017) employs the `limma` methodology (Ritchie et al., 2015). This is an empirical Bayes approach in which combined information of all considered genes, typically several thousands, is used to adjust the individual variance estimates of genes. The goal of this work is to give recommendations on the choice of observation-based vs. model-based methods independently on the number of genes included in the dataset. Therefore, in this work, instead of the `limma` approach a simple t-test and the Dunnett procedure introduced in Chapter 4.1.1 are used for calculating observation-based alert concentrations.
- In Grinberg (2017), concentration-wise standard deviations are chosen from the set of observed standard deviations in the gene expression dataset from Krug et al. (2013), introduced in this work in Chapter 3.2.2. Analysis of this dataset shows some dependency of the standard deviation on the range of gene expression values. Therefore, instead of choosing the standard deviation individually for each simulated gene, in this analysis fixed standard deviations that depend on the range of the gene expression observed for each scenario are chosen.

- In Grinberg (2017), the simulation study is conducted considering $3, 6$ and 10 replicates per concentration, with the observation that increasing the sample size affects model-based estimates more than observation-based estimates (Grinberg, 2017). In this work, the simulation study is restricted to the case of 3 replicates per concentration.

The four different alert concentrations are introduced in detail in Chapter 4.3 with most emphasis on the new log-logistic model-based test (Chapter 4.3.3). These methods are compared in a simulation study whose setup is explained in Chapter 6.1 and the results are analysed in Chapter 6.2. The goal of this simulation study is the assessment of the quality of the estimated alert concentrations: The numbers of valid estimates, i.e. estimates that are in the range of considered concentrations, are compared between methods. Furthermore, the number of 'false positive' alerts, i.e. alerts found at lower concentrations than the alert concentration of the true underlying curve of the simulation study, are compared. Differences in the performance between the different true underlying curves considered in the simulation study are evaluated.

Finally, the methods are applied to a real gene expression dataset, where concentration-gene expression data for the compound valproic acid is collected (Chapter 6.3). This dataset is introduced in detail in Chapter 3.2.2.

The results concerning the identification of alert concentrations for concentration gene-expression data are published in Kappenberg et al. (2021). Analyses published there are extended by more details regarding the comparison of the LOEC based on the t-test and based on the Dunnett procedure in this work. Additionally, for the choice of the probe sets from the entire gene expression dataset, two different variants are calculated and compared in this thesis, while in the paper, only one variant is considered.

2.3. Information sharing across genes

Microarray technologies allow measurements of gene expression values for tens of thousands of genes simultaneously. Measuring gene expression values for several increasing concentrations of the same compound yields a concentration-response dataset. When fitting a parametric curve to such a dataset, the quality of the fit improves with an increasing number of concentrations and replicates. However, adding new replicates of already considered concentrations or even several replicates of new concentrations is very expensive. A statistical approach to obtain higher-quality fits of the parametric curves or of some parameters of the curves is to exploit similarities between genes. Certain aspects of the fits of several curves are pooled to improve each single fit. That means that some *information sharing* is conducted across genes.

The target aspect of each curve that is to be improved is the parameter corresponding to the concentration where the half-maximal effect is reached: The curves considered have a left-sided asymptote for concentration values tending towards zero and a right-sided asymptote for concentration values tending towards infinity. The parameter of interest indicates the concentration where exactly the center between the response values

corresponding to the two asymptotes is reached. This concentration value is a reasonable indicator for a relevant expression effect of the gene considered.

The first idea is to pool information of 'similar' genes in a meta-analytic way. The application of meta-analysis to concentration-response data is based on Jiang and Kopp-Schneider (2014), and is explained in this work in Chapter 4.4.2. Pooling by similarity, e.g. determined by high correlation values between genes, however, is not enough: A simulation study is conducted (Chapter 7.3.1), where the entire set of genes is considered as set of potential similar genes. Actual similar genes are determined by a high correlation score, and all genes with a correlation higher than a specific threshold are included in the meta-analysis to improve the estimate for a specific gene. Results show that neither the coverage probabilities of the resulting confidence interval nor the mean squared error between estimated and true underlying parameter are improved.

Therefore, in a next step, the biological similarities and therefore ideally similar properties of the concentration-gene expression profiles are exploited by considering only genes from a specific Gene Ontology group (GO group, see Chapter 3.3) as potential set of neighbours. In a simulation study, GO groups of different sizes and of different coherences in terms of the similarities of the genes are considered. These simulation studies are presented in Chapter 7.3.2.

The second idea is to consider the entire dataset to obtain the distribution of the parameters of the parametric curve. Based on this distribution, a weighted mean of actual estimated parameters and the entire distribution can be calculated. Thus, estimates are shrunk towards the most plausible value. This is achieved by employing an empirical Bayes method under normality assumptions, where a prior distribution of the parameter is calculated based on observations on the entire dataset. The posterior distribution is then also given by a normal distribution, where the posterior mean is a weighted mean of the observed value and the prior mean. This method is introduced in detail in Chapter 4.4.3.

The main idea of the method is to improve the estimation of the parameter corresponding to the half-maximal effect in terms of the difference to the true parameter. In cases where the concentration-response data of one single gene does not allow a high-quality fit of the curve and therefore much too large estimates result, shrinking the estimate toward the empirical mean value makes use of the knowledge about the true distribution to ensure a more plausible fit.

With the `limma` procedure (Ritchie et al., 2015), one empirical Bayes methodology is already established in the field of microarray analysis. The goal of the methodology is to simultaneously analyse comparisons between samples in high-dimensional gene-expression data. That means that it is aimed at analysing fold changes of genes, i.e. the difference in gene expression for different samples, while the methodology introduced in this work is aimed at giving statements about the concentration where some relevant change in expression can be observed.

The Bayesian shrinkage method is investigated in several simulation studies that are similar to each other in the basic ideas but differ in their resemblance to a real data situation. Main properties of the curves examined in these studies are derived from

the gene expression dataset introduced in Chapter 3.2.2. In the first simulation study, however, the underlying parameters of the curve are completely randomly sampled from normal distributions. For the second simulation study, parameters as observed in the real dataset are considered, but only after some normalisation is applied that ensures that the assumption of the normal distribution holds. In the third study, the parameters are taken from the real dataset as it is. The simulation studies with their results are presented in Chapter 7.4.

For both approaches, a deep understanding of the distribution of the parameters of the fitted curves is required. Therefore, in a first step, a detailed descriptive analysis of the parameters for all genes satisfying certain requirements of significance is conducted. This analysis is presented in Chapter 7.1. Since for the meta-analysis approach, GO-groups play a central role, their properties are also described in detail in Chapter 7.2.

Finally, the methods are applied to the real gene expression dataset, on which all simulation studies are based. The results are presented in Chapter 7.5.

3. Data and biological background

This chapter gives the biological background needed to understand the different types of datasets considered throughout this work. The respective biological background is explained as well as the methods to obtain a raw dataset. Some pre-processing is needed, such that the raw data is transformed to normalised data that can then be statistically analysed. These steps are first introduced in general for a cytotoxicity assay that results in viability measurements, thus also called viability assay, and for an Affymetrix microarray analysis that results in gene expression values.

After the general background, the specific datasets considered in this work are introduced. One cytotoxicity dataset and one gene expression dataset, both evaluating the effect of the compound valproic acid on cells, are described in detail.

Finally, a short introduction into the Gene Onotology is given, which is an initiative to represent and unify the meaning of genes with respect to their biological functions. The result is the Gene Ontology database, in which genes are structured according to their biological process, their molecular function or their cellular component.

In this work, the influencing variable considered always is a *concentration* instead of a *dose*. By concentration, the amount of a compound in a mixture that is applied to cells is described, while the dose is the total amount of a compound that is administered to tissue (Duffus et al., 2007). Data considered here are results from in vitro assays. The term *in vitro* (from lat. 'in glass') refers to a study or a toxicological assay conducted in the laboratory with tissue or cells, while *in vivo* (from lat. 'in the living body') refers to a study with living organisms (Duffus et al., 2007).

3.1. Different types of data

Different methods for assessing the cytotoxicity of a compound by measuring the viability of cells exist. One possible method is the Cell-Titer-Blue (CTB) assay, that works by measuring a fluorescent component and deriving the viability based on these fluorescence values. This assay is explained in more detail in Chapter 3.1.1.

The Affymetrix GeneChip® technology is a high-troughput microarray technology to simultaneously measure expression values for tens of thousands of genes. This technology is briefly introduced in Chapter 3.1.2, together with the biological basics needed for understanding gene expression data and with the normalisation procedure yielding the gene expression values in the form that is considered in this thesis.

3.1.1. Cytotoxicity data

The CTB assay is described here according to the Standard Operating Procedure 3A of Gu et al. (2018). Only an overview of the method is given here with omission of many details regarding proper preparation and handling of the cells, as this would exceed the scope of this work.

The basic idea of the CTB assay is to measure the fluorescence of cells to which the solvent control (i.e. none of the considered compound) and increasing concentrations of the considered compound are given. For this, the CTB medium containing the dark blue indicator dye resazurin is added to cells for measuring their metabolic capacity. Vital cells can reduce resazurin into the pink and highly fluorescent dye resofurin. Non-viable cells have a reduced capacity for the metabolization of resazurin. Therefore, lower fluorescence is observed. In the case of completely dead cells, no resazurin can be reduced to resofurin at all.

Preparation of the cells consists of seeding them in 96-well plates with 50000 cells per well and the beginning of treatment 16-20 hours later. Cells are exposed to the respective concentration of a compound for a pre-specified time. After this time, replacement of the medium is conducted where the CTB reagent is added to the new medium.

The cells with this reagent are incubated for about 3 hours at 37°C. After this incubation time, the fluorescent intensity is read out, in the specific example considered here using the Tecan Infinite M200 Pro plate reader (i- control software, version 1.7.1.12). This results in a fluorescence value for each well of the plate.

Together with the fluorescence values for the cells treated with the compound and then with the CTB medium, a background fluorescence value is calculated from CTB medium that has not been in contact with the cells. This background value is subtracted from each measurement, or, if several background fluorescence values were calculated, the mean of these values is subtracted from each measurement. The untreated cells are used as a reference corresponding to a viability to 100% and the data normalised in this way is ready for statistical analysis.

3.1.2. Affymetrix gene expression data

In order to understand the data obtained by Affymetrix GeneChip® technology, a basic understanding of the structure and the function of genes is necessary. Thus, this chapter starts by giving a short introduction into the biology behind gene expression data. Then the microarray technology is explained, and finally the pre-processing, necessary to transform raw data to gene expression values for statistical analysis, is introduced.

A gene is a segment of the DNA (deoxyribonucleic acid), storing genetic information that codes for a protein. The human genome consists of about 21000 genes that are separated in the DNA by noncoding segments. The DNA itself consists of long strands of nucleotides, each of which contains a sugar, a base and a phosphate group. Four different bases occur, they are called adenine (A), guanine (G), thymine (T), and cytosine (C). Two such strands are stabilized by hydrogen bonds between the bases, forming the double-helix structure of the DNA, with the sugar and the phosphate facing outwards and the bases facing inwards. The bases are always paired up in the same way, A is only paired with T and G is only paired with C. This is called *complementary base-pairing* and is an important feature of the DNA for expressing genetic information, as one strand can completely be replicated when knowing the other one (Fletcher and Hickey, 2012, pp. 1-6).

Genetic information stored in the DNA allows reproduction of cells and organisms. The process of transferring the genetic information from DNA for protein synthesis is called *gene expression*. RNA (ribonucleic acid) is a one-stranded chain of nucleotids that differ from the DNA nucleotids in the specific sugar. Additionally, the base thymine is replaced by uracil (U). Gene expression is a three-step procedure, with the steps RNA transcription, RNA processing, and translation (Fletcher and Hickey, 2012, pp. 6, 13).

In the first step, the process of transcription, the base sequence of the DNA is copied into a corresponding RNA sequence, yielding a mRNA (messenger RNA) transcript. This transcript is processed by removing non-coding segments. Finally, the mRNA is translated into proteins (Fletcher and Hickey, 2012, p. 16).

The Affymetrix GeneChip® is a high-density chip, called *microarray*, for the analysis of gene expression data. A gene expression microarray consists of sequences of 25 nucleotides, called *probes*, that are complementary to the gene that is targeted by the measurements. Usually, 11 to 20 of these probes correspond to one given gene, forming the so called *probe set*. In the Human Genome U133 Plus 2.0 chip used for the data analysed in this thesis, 54675 probe sets are used to analyse expression levels of more than 47000 transcripts (Bolstad, 2004; Affymetrix, 2003b).

Two pairs of probes can be found on a GeneChip. A *perfect match* (PM) is a probe that is exactly complementary to the sequence of interest. For each PM, there exist some partner probes that differ from the PM only by the 13th base, which is the middle of the sequence of nucleotids, called a *mismatch* (MM). Thus, no or at least weaker binding of the target sequences to the respective MM is achieved, allowing the quantification of background signal detected by the PM probe (Bolstad et al., 2003; Bolstad, 2004).

For details regarding the biological procedure of measuring gene expression data, the reader is referred to Bolstad (2004) and Affymetrix (2003a). The basic idea is to bring a transcribed version of the target mRNA in contact with the microarray, such that fragments bind to their respective complementary probes on the microarray. Fragments bound to the microarray emit some fluorescent light which is measured by a laser scanner, yielding intensity values.

Intensities of the single probe sets need to be transferred to expression values. This step is called *pre-processing*, and one of the algorithms used most often is the robust-multi array analysis (RMA) (Irizarry et al., 2003a). This algorithm yields an expression measure that is motivated by a log-scale linear additive model, where a transformation of the PM is given by the sum of $\log_2$-scale expression values found on the arrays, the log-scale effects for the respective probes and an error term. The transformation corresponds to background correction, normalisation and log-transformation.

Hence, the RMA algorithm is a three-step procedure with the steps of background correction, normalisation and summarising data to one value. Background correction is performed for each microarray individually. Irizarry et al. (2003b) show that the intuitive method of background correction, where MM are subtracted from PM, is not the best possible method. Instead, only PM is modelled by the sum of background noise and the signal of interest. By exploiting distribution assumption for both summands, the signal can be estimated.

Normalisation is needed for experiments with multiple microarrays, in order to deal with variation between microarrays that stems from the practical aspects of executing the experiments. In the RMA algorithm, quantile normalisation is performed. The goal of this procedure is to obtain normalised probe values, whose distributions are identical across microarrays. This method is motivated by quantile-quantile plots, in which sample quantiles of two distributions are plotted against each other. If the data points plotted there form a straight diagonal, the distributions are the same. Thus, quantile normalisation is performed by substituting the data from the original observation by a mean quantile, see Bolstad et al. (2003) for more details.

The final step is then again summarising the normalised intensity values to one expression value per probe set. This is achieved by employing a linear model as indicated above, that is, for a fixed probe set, specifically given by

$$T(\mathrm{PM})_{ij} = e_i + a_j + \varepsilon_{ij},$$

with T being the transformation that conducts background correction, normalisation and $\log_2$-transformation of the PM intensities. The index $i = 1, \ldots, I$ corresponds to the microarrays and $j = 1, \ldots, J$ refers to the j-th probe of the specific probe set considered. The $\log_2$-scale expression value of the probe set on microarray i is denoted by e_i and a_j denotes the probe effect for the j-th probe. ε_{ij} is an error term. The term e_i, estimated by a robust linear fitting procedure, is the corresponding gene expression value per probe of interest and referred to as RMA (Irizarry et al., 2003a).

A disadvantage of the RMA algorithm is that the final two steps are dependent on the set of microarrays normalised. Thus, when measurements for additional microarrays shall be pre-processed, the entire set of microarrays needs to be pre-processed again. Harbron et al. (2007) propose an extension of the RMA algorithm, implemented in the `R`-package `RefPlus`. The idea is to apply the RMA algorithm to a reference set of microarrays and to save the resulting parameters from the quantile normalisation and the final linear model. The microarrays of interest are then pre-processed using the parameters obtained from pre-processing the reference set.

The reference set is obtained by dividing the set of all microarrays into two sets: the reference set and the set of all other microarrays. In a first step, RMA is applied to the reference set, while remembering the quantiles and the probe effects estimated from the linear model. Then, the other microarrays are background-corrected individually. For the quantile normalisation, saved values of the quantiles are used and the final expression value of the probe set is calculated under the assumption that probe effects are equal for the reference set and the other microarrays (Harbron et al., 2007).

Often, when working with gene expression values, the *fold change* (FC) is the measure of interest. The FC is given by the difference of gene expression for one specific gene and two concentrations of interest. Usually, the FC between a concentration and the control, corresponding to concentration 0, is considered. Since the data considered here is $\log_2$-transformed, the logarithmised FC is considered: Denote by x^{Conc} and x^{Ctrl} the (not logarithmised, but potentially averaged across several replicates) gene expression values for a concentration Conc and the control Ctrl. Averaging is conducted using the

geometric mean, which corresponds to the arithmetic mean when considering logarithmised expression values. The logarithmised FC between the expression values for the two concentrations is then given by

$$\log_2(\text{FC}) = \log_2(x^{\text{Conc}}) - \log_2(x^{\text{Ctrl}}) = \log_2\left(\frac{x^{\text{Conc}}}{x^{\text{Ctrl}}}\right).$$

3.2. Datasets

The two datasets considered and explained in detail in the following two chapters are based on some response (viability or gene expression) of cells when treated with increasing concentrations of the compound *valproic acid* (VPA). VPA is a compound used in a clinical context among other for treatment of epilepsy. However, it is known that it triggers reproductive toxicity, specifically developmental neurotoxicity, in humans as well as in animals (Krug et al., 2013).

3.2.1. VPA cytotoxicity dataset

The dataset presented here was originally created for Kappenberg et al. (2020) and is analysed there as well.

For the VPA cytotoxicity dataset, a Cell-Titer-Blue assay was conducted as described in Chapter 3.1.1, using cells of the HepG2 cancer cell line. HepG2 cells are frozen cells coming from the liver tissue of a 15 year old Caucasian male that suffered from hepatocellular carcinoma[1]. These cells were cultivated in Dulbecco's Modified Eagle's Medium (DMEM) with 25 mM glucose. VPA (CAS number 99-66-1; Sigma Aldrich; product number: PHR1061-1G) was directly dissolved in the culture medium to generate the concentrations indicated below so that no solvent was required.

Viability of cells was measured for a negative control and for 12 increasing concentrations from 0.1 mM to 56.2 mM that differ approximately by a factor of $1.78 \approx 10^{1/4}$. The concentrations were chosen in order to obtain results with no toxicity for the lowest concentrations, and high toxicity, i.e. a viability of 0%, for the highest concentration. For each concentration, response values for three biological replicates with seven technical replicates each were measured. In total, this leads to 91 measurements for each of the three biological replicates. No observations are missing for this dataset.

3.2.2. VPA gene expression dataset

The dataset presented here was originally created in the context of a case study to investigate the development of human embryonic stem cells (hESC) to neuroectoderm (Krug et al., 2013). The study was carried out within the framework of the European Commission-funded research consortium ESNATS. This consortium targeted prediction

[1] `https://www.lgcstandards-atcc.org/products/all/HB-8065.aspx`, accessed on 24.8.2020, originally published in Knowles et al. (1980)

of toxicity of drug candidates (specifically VPA and methylmercury, which is not considered here) for the use of embryonic stem cell-based novel alternative tests.

The VPA dataset was created via the Affymetrix microchip technology, as presented in Chapter 3.1.2. The Affymetrix Human Genome U133 Plus 2.0 GeneChip was used, resulting in gene expression values for 54675 probe sets.

Cells were treated in vitro with VPA at the eight increasing concentrations $25, 150, 350, 450, 550, 650, 800$, and 1000 µM with three replicates each. Additionally, six replicates for untreated cells, serving as negative control, in the following also referred to as the concentration 0, were created. The data is pre-processed using the RMA algorithm with the `RefPlus` extension as introduced in Chapter 3.1.2. The same parameters as used by Krug et al. (2013) are used for pre-processing.

Grinberg (2017) notes that the samples corresponding to a concentration of 650 µM show a high variability when a principal component analysis based on the 100 probe sets with highest variance across all samples is conducted. As stated there, data quality can be improved by excluding these samples. Based on this observation, the three samples corresponding to a concentration of 650 µM are left out of all analyses in the present thesis. Thus, the considered VPA gene expression dataset consists of measurements for 54675 probesets, measured in 7 increasing concentrations and a negative control, leading to 27 samples in total.

3.3. Gene Ontology

An ontology is a system to summarise some data, while representing relationships between the single data points. The Gene Ontology Consortium (2000) developed a system to summarise functions of all known genes in three so-called *gene ontologies* (GO). The three ontologies are based on the biological processes, molecular functions or cellular components of genes and are independent from each other.

In this work, only the GO summarising biological processes is considered. The biological process refers to some biological goal, such as 'cell growth and maintenance' when considering a very broad term or 'translation' when considering more specific terms. Typically, in a process some transformation is conducted, such that initial product and end product differ. Since the GO summarises genes not only for humans, but for all animals, plants and fungi, the information about the genes does not refer to specific organs, but to the general processes only (Gene Ontology Consortium, 2000).

A GO is a directed acyclic graph with a hierarchical structure. In the case of biological process, each node corresponds to a process. The node consists of all genes annotated to the specific process. Nodes rather on top of the graph correspond to high-level biological processes, whereas nodes further down the graph correspond to more specific processes. All genes included in a specific node are also included in the parent node, thus the size of the nodes becomes smaller when the biological process becomes more specific. The nodes of the GO are referred to as *GO groups*.

4. Statistical Methods

In this chapter, the statistical methods used throughout the analyses of this thesis are presented. The first section deals with the basics of concentration-response analysis, specifically the modelling of concentration-response curves is introduced in detail. These results are needed for all the following methods. The following three sections introduce the special methods derived for the three objectives of this thesis, handling deviating control values, identification of alert concentrations, and information sharing across genes, as introduced in Chapter 2, in detail. In the final section, the software used for conducting the analyses is introduced.

4.1. Basics of concentration-response analyses

In the first part, the multiple comparison procedure by Dunnett is introduced. After that, the focus is laid on the modelling of curves: First the considered models are introduced. The following section explains the assumptions made on the data for fitting these curves and the numerical fitting process itself. An overview over potential alert concentrations derived from fitted curves is given and finally, the MCP-Mod methodology, which combines a multiple comparison step and a modelling step for several models at the same time, is shortly introduced.

4.1.1. The Dunnett procedure

The Dunnett procedure is a multiple-comparison approach to simultaneously test for significant differences between increasing concentrations of a compound and a negative control, while adjusting for multiplicity. It was initially proposed by Dunnett (1955) and is introduced here using Hothorn (2015, pp. 25-26) as additional source.

Let $p \in \mathbb{N}$ be the number of treatments to be compared to a negative control. Therefore, $p+1$ sets of observations are made with $n_i \in \mathbb{N}$ observations each for $i = 0, \ldots, p$. The index i refers to the different groups, with $i = 0$ referring to the control group and $i = 1, \ldots, p$ to the treatment groups. The observations X_{ij}, $i = 0, \ldots, p, j = 1, \ldots, n_i$ are assumed to be independently normally distributed with $\mathbb{E}[X_{ij}] = \mu_i$. Mean values of the observations X_{ij} in each treatment group i are denoted by $\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_p$ and the respective realisations are denoted by $\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_p$. The standard deviation σ is assumed to be equal for all groups. It is estimated in a pooled way across all groups, the estimate is denoted by S with realisation s.

Significance statements for the p differences $\mu_l - \mu_0$, $l = 1, \ldots, p$ are simultaneously calculated while correcting for multiplicity, i.e. the family-wise error rate is bounded by a pre-specified significance level α by incorporating the correlation between the differences. When considering the p one-sided null hypotheses $\mathrm{H}_0 : \mu_l - \mu_0 \leq 0$, the probability of having at least one type I error, i.e. a rejection of at least one true null hypothesis, is smaller than or equal to α. The procedure is introduced here only for the one-sided case where finding a significant increase in the response is of interest, the opposite direction works analogously.

The Dunnett procedure can be formulated as a multiple contrast test. The l-th contrast is given by $\sum_{i=0}^{p} c_{li}\bar{X}_i$ for $l = 1, \ldots, p$. The corresponding test statistics are given by the standardised contrasts and take the following form:

$$T_l = \frac{\sum_{i=0}^{p} c_{li}\bar{X}_i}{S\sqrt{\sum_{i=0}^{p} c_{li}^2/n_i}}$$

The specific contrast matrix corresponding to the Dunnett procedure is given by

$$C = (c_{li}) = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 \\ -1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{p\times(p+1)},$$

which leads to the following simplification of the test statistics:

$$T_l = \frac{\bar{X}_l - \bar{X}_0}{S\sqrt{1/n_0 + 1/n_l}}$$

Under the null hypothesis, the p-dimensional vector $(T_1, \ldots, T_p)^\top$ of test statistics follows a joint multivariate t-distribution with $\mathrm{df} = \sum_{i=0}^{p}(n_i - 1)$ degrees of freedom. The matrix $\boldsymbol{R} = (\rho_{ij}) \in \mathbb{R}^{p\times p}$ of correlations between the contrasts is given by

$$\rho_{ij} = \sqrt{\frac{1}{(1 + n_0/n_i)\,(1 + n_0/n_j)}}, \quad \text{for } 1 \le i, j \le p.$$

For a significance level $\alpha > 0$, the lower limits of simultaneous, one-sided $(1-\alpha)100\%$ confidence intervals are given by

$$\hat{\delta}_l^{\mathrm{low}} = \bar{x}_l - \bar{x}_0 - t_{p,1-\alpha}(\mathrm{df}, \boldsymbol{R})\, s \sqrt{\frac{1}{n_l} + \frac{1}{n_0}},$$

where $t_{p,1-\alpha}(\mathrm{df}, \boldsymbol{R})$ is the $(1-\alpha)$ quantile of the p-variate t-distribution with df degrees of freedom and correlation matrix $\boldsymbol{R}$. It is chosen in a way that if $q = t_{p,1-\alpha}(\mathrm{df}, \boldsymbol{R})$, then the probability that any of the absolute values of the observed test statistics, $|t_l|$, is larger than q is given by α.

4.1.2. The family of log-logistic models

The family of log-logistic models is one of the most commonly used model classes for fitting concentration-response curves, not only in the area of toxicological in vitro data, but also for ecotoxicology (van der Vliet and Ritz, 2013) or for modelling the hazard rate for the survival of a patient after a severe medical diagnosis, as in this case typically an initially increasing and later decreasing mortality can be observed (Nussbeck, 2014).

Given a concentration $x \geq 0$ and a parameter vector $\boldsymbol{\phi} = (\phi^{(b)}, \phi^{(c)}, \phi^{(d)}, \phi^{(e)})^\top$ with $\phi^{(e)} > 0$, the *four-parameter log-logistic model* (4pLL model) is defined by (e.g. Ritz et al., 2019, pp. 178-179)

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

$$= \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + (x/\phi^{(e)})^{\phi^{(b)}}}. \tag{2}$$

The parameters $\phi^{(c)}$ and $\phi^{(d)}$ correspond to the lower and the upper asymptotes of the model, where the assignment depends on the slope of the curve and on the sign of the slope parameter $\phi^{(b)}$. Typically, $\phi^{(c)}$ is the lower and $\phi^{(d)}$ the upper asymptote. Specifically, for $\phi^{(b)} > 0$ it holds

$$\lim_{x \to 0} f(x, \boldsymbol{\phi}) = \phi^{(d)} \qquad \text{and} \qquad \lim_{x \to \infty} f(x, \boldsymbol{\phi}) = \phi^{(c)}.$$

Equivalently, for $\phi^{(b)} < 0$ it holds

$$\lim_{x \to 0} f(x, \boldsymbol{\phi}) = \phi^{(c)} \qquad \text{and} \qquad \lim_{x \to \infty} f(x, \boldsymbol{\phi}) = \phi^{(d)},$$

see Appendix A.1 for the exact calculation.

The parameter $\phi^{(e)}$ corresponds to the concentration where the half-maximal effect (i.e. a response value of $\frac{\phi^{(d)} + \phi^{(c)}}{2}$) can be observed. For display of the 4pLL model on a logarithmic x-axis and $\phi^{(b)} \neq 0$, this parameter also corresponds to the inflection point of the curve. When the model is displayed on an untransformed x-axis, an inflection point is only present if $|\phi^{(b)}| > 1$, and the respective concentration is given by

$$x = \left(\frac{\phi^{(b)} - 1}{\phi^{(b)} + 1}\right)^{\frac{1}{\phi^{(b)}}} \phi^{(e)},$$

see Appendix A.2 for the detailed calculations.

$\phi^{(b)}$ is proportional to the actual slope at the concentration $\phi^{(e)}$, which is given by

$$-\frac{\phi^{(b)} \cdot \left(\phi^{(d)} - \phi^{(c)}\right)}{4\phi^{(e)}}.$$

When considering a logarithmic x-axis, the slope at concentration $\phi^{(e)}$ is given by

$$-\frac{1}{4} \cdot \phi^{(b)} \cdot \left(\phi^{(d)} - \phi^{(c)}\right),$$

which differs from the function above only by the factor $\phi^{(e)}$ in the denominator. For exact calculations of the slopes, see Appendix A.3.

Often, and especially for small datasets with less than 15 to 20 data points, the re-parametrisation $\phi^{(e)^*} := \log(\phi^{(e)})$ should be preferred (Ritz et al., 2019, p. 179).

Fixing one or two of the parameters to take a pre-specified value yields a model with only three or two parameters. In the context of response data corresponding to proportions, often the lower asymptote is fixed to take a value of 0, or the upper asymptote is fixed to take a value of 100% or 1, or even both. These models are then called *three-parameter log-logistic model* (3pLL model) or *two-parameter log-logistic model* (2pLL model), respectively.

The fits achieved with these models are monotonously decreasing or monotonously increasing and, if a logarithmic x-axis is considered, point symmetrical around the inflection point. The 2pLL model is equivalent to a logistic regression using a logit-link function with $\log(x)$ as the only explanatory variable (Ritz et al., 2019, p. 179).

In the literature, the 4pLL model is sometimes called differently and presented in different parametrisations, such as the sigmoidal Emax model (`sigEmax`, e.g. Bornkamp et al., 2009), parametrised by

$$f(x) = E_0 + \frac{E_{\max} x^h}{\left(\mathrm{EC50}^h + x^h\right)}, \tag{3}$$

with E_0 describing the effect for concentration 0, $E_{\max}$ describing the maximal effect, that is $\max_x (f(x) - E_0)$, EC50 describing the half-maximal effect with respect to E_0 and $E_{\max}$, and h describing the slope. It is also sometimes called the Hill-model, which dates back to Hill (1910), who introduced a model equivalent to the 2pLL model in the context of the binding of haemoglobin molecules. Equivalence of the `sigEmax` model to the 4pLL model presented in (2) can easily be shown (see Appendix A.4), and the slope parameter in these models is still sometimes called the *Hill parameter*.

The 4pLL model can be extended by a fifth parameter $\phi^{(f)}$ to result in an asymmetric, non-monotonous model. The *five-parameter log-logistic model* (5pLL model) was initially proposed by Finney (1979) and is given by

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

Another, often desired property of a concentration-response model is the ability to model a *hormesis effect*, that is a stimulating effect of the compound in low concentrations followed by the regular inhibition for higher concentrations (Calabrese and Baldwin, 2003). The consequence is a non-monotonic course of a concentration-response curve. In the case where inhibition leads to a decrease in the considered endpoint, the hormesis effect can be seen as an inverted U-shape, that is an increase in the curve before it decreases.

One possible model incorporating a hormesis effect is the Brain-Cousens model (Brain and Cousens, 1989), which results from the 4pLL model by incorporating a fifth parameter $\phi^{(f)}$ in the following way:

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

Cedergreen et al. (2005) recommend restricting the parameters $\phi^{(b)}$ and $\phi^{(f)}$ of the Brain-Cousens model to $\phi^{(b)} > 1$ and $\phi^{(f)} > 0$, as for a negative value of $\phi^{(f)}$ a U-shape instead of the desired inverted U-shape can be observed for the fitted model. For a value of $\phi^{(b)} < 1$, the limit of the fitted model for $x \to \infty$ approaches infinity as well. In their work it is however stated that these restrictions can in general not be observed in publications and thus will also not be made in this thesis.

In contrast to the 4pLL model, the Brain-Cousens model does not incorporate an explicit parameter for the concentration inhibiting the half-maximal effect, also there is no closed formula for calculating it. This concentration therefore needs to be computed numerically. Direct interpretations of the parameters $\phi^{(b)}$ and $\phi^{(e)}$ as in the 4pLL model are not possible, but $\phi^{(c)}$ and $\phi^{(d)}$ still correspond to the asymptotes of the model (Cedergreen et al., 2005).

Extensions of the Brain-Cousen model exist, e.g. a model proposed by Cedergreen et al. (2005) that replaces the term $\phi^{(f)}x$ in the Brain-Cousens model by $\phi^{(f)}\exp(-1/x^{\alpha})$ for some fixed $\alpha \geq 0$. In the present work, modelling a hormesis effect is restricted to the Brain-Cousens model.

4.1.3. The numerics of curve-fitting

For $x \geq 0$ a concentration and Y a response value corresponding to this concentration, where Y is subject to sampling variation, it is assumed that

$$\mathbb{E}[Y] = f(x, \boldsymbol{\phi}).$$

f is a function describing the relationship between the concentration and the response, e.g. one of the functions presented in Chapter 4.1.2. The function f is completely known, except for the model parameters $\boldsymbol{\phi}$ (Ritz et al., 2015).

Estimation of the model parameters requires assumptions regarding the distribution of the response values. Let $x_1, \ldots, x_n$ be the concentration values (whereby equal concentrations may occur) with corresponding response values $y_1, \ldots, y_n$. The response values are assumed to be observations of normally distributed random variables Y_i with mean $f(x_i, \boldsymbol{\phi})$ and equal variances σ^2. In this case, the parameters $\boldsymbol{\phi}$ are estimated using the nonlinear least squares method, which, for a normally distributed response variable, is equivalent to the maximum likelihood method. Estimates for $\boldsymbol{\phi}$ are obtained by minimising the following sum of squared errors:

$$\sum_{i=1}^{n} w_i^2 \left(y_i - f(x_i, \boldsymbol{\phi})\right)^2 \tag{4}$$

The w_i are weights that are specified for each application separately, e.g. when considering a proportion as response value, they are set to the number of observations for

the unique concentrations. Usually for normally distributed response data, and in this thesis as well, all of them are chosen equal to 1 (Ritz et al., 2019, p. 162).

The minimisation of the sum in (4) needs to be conducted by a numerical optimisation algorithm. Here, the Gauss-Newton algorithm is chosen for optimisation. An important step for the optimisation algorithm is the choice of appropriate starting values for the model parameters. This choice does not only affect whether convergence of the algorithm can be achieved, but also, whether a global or only a local optimum of parameters is found by the optimisation algorithm (Ritz et al., 2015).

The starting values are chosen either using information obtained from previous, similar experiments or in a data-driven way. For the 4pLL model in the parametrisation using $\phi^{(e)^*}$, the starting values for the parameters $\phi^{(c)}$ and $\phi^{(d)}$, describing the asymptotes, are calculated by taking the minimum and the maximum of all response values and adding or subtracting 0.001, respectively. For the parameters $\phi^{(b)}$ and $\phi^{(e)}$, the following linear model is fitted to the data:

$$\frac{\phi^{(d)} - Y}{Y - \phi^{(c)}} = \beta_0 + \beta_1 \log(x)$$

The parameters are then given by $\phi^{(b)} = \beta_1$ and $\phi^{(e)^*} = \exp(-\beta_0/\beta_1)$. In the Brain-Cousens model, the starting value for the additional parameter $\phi^{(f)}$ is equal to 1 (Ritz et al., 2015).

Although $\log(0)$ is undefined and $\lim_{x\to 0} \log(x) = -\infty$, no addition of a very small number to the concentration value $x = 0$ is needed for fitting the curves. This is ensured by incorporating the asymptote values into the model-function, yielding well-defined response values for concentration 0 as well (Ritz et al., 2015).

In estimating the model parameters $\boldsymbol{\phi}$ some constraints are directly or indirectly made to the parameters: In the 4pLL model, the parameter $\phi^{(e)}$ corresponds to a specific concentration and therefore needs to take a value $\phi^{(e)} \geq 0$. By the choice of starting values, it is indirectly ensured that the parameter $\phi^{(b)}$ takes a negative value for increasing curves. In practice, however, in some cases it can be observed that $\phi^{(b)}$ is positive for increasing curves or negative for decreasing curves, and the typical assignment of the parameters $\phi^{(d)}$ and $\phi^{(d)}$ to upper and lower asymptote is swapped accordingly. The estimation procedure may be combined with any restriction of the range of one or several parameters if this is needed for any application-based reason (Ritz et al., 2015).

The covariance matrix $\hat{\Sigma} := \hat{\Sigma}(\hat{\boldsymbol{\phi}})$ of the estimated parameter values is obtained as the scaled inverse of the observed information matrix, i.e.

$$\begin{aligned} \hat{\Sigma} &= \hat{\sigma}^2 \left(\left\{ \frac{\partial^2 \hat{f}}{\partial \hat{\phi}_{p_1} \partial \hat{\phi}_{p_2}} \right\}_{p_1, p_2 \in \{1,\ldots,p\}} \right)^{-1} \\ &= \hat{\sigma}^2 \left(\frac{\partial^2 \hat{f}}{\partial \hat{\phi}_{p_1} \partial \hat{\phi}_{p_2}} \right)^{-1}. \end{aligned} \tag{5}$$

Diagonal entries of $\hat{\Sigma}$ are the estimated squared standard errors of the entries of parameter vector $\boldsymbol{\phi}$, i.e. $\hat{\Sigma}_{ii} = \hat{\text{var}}[\hat{\phi}_{p_i}] = \hat{\text{se}}(\hat{\phi}_{p_i})^2$, $p_i \in \{1, \ldots, p\}$. Here, in general, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$. The information matrix itself is approximated numerically (Ritz et al., 2019, p. 163). $\hat{\sigma}^2$ denotes the squared residual standard error. It is estimated as the residual standard error for the linear regression, which is typically estimated as

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

with p denoting the numbers of parameters in the model (without the intercept) and $\hat{y}_i = \hat{f}(x_i, \hat{\boldsymbol{\phi}})$ the fitted values. The denominator $n-p-1$ is used instead of n in order to achieve an unbiased estimate of the variance (Hastie et al., 2009, p. 47). In the case of a 4pLL model, this translates to

$$\hat{\sigma}^2 = \frac{1}{n-4} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i, \hat{\boldsymbol{\phi}}) \right)^2. \tag{6}$$

The $(1-\alpha)100\%$ confidence interval for parameter $\phi_p \in \{\phi^{(b)}, \phi^{(c)}, \phi^{(d)}, \phi^{(e)}, \phi^{(e)^*}\}$ is calculated as $\hat{\phi}_p \pm K \cdot \hat{\text{se}}(\hat{\phi}_p)$. In the case considered in this work, the response values are assumed to follow a normal distribution. Thus, the value of the constant K is chosen as the $(1-\alpha/2)$-quantile of a t-distribution. The degrees of freedom equal the number of response values minus the number of parameters, i.e. minus four in the case of a 4pLL model. As this calculation leads to intervals that are symmetric around the estimate, implausible lower bounds, such as negative values for parameter $\phi^{(e)}$, may occur. A simple approach to deal with this problem is to truncate lower limits at 0 (Ritz et al., 2019, p. 172).

The fitting methods presented here are implemented in the `R`-package `drc` (Ritz et al., 2015), which is used for all curve-modelling applications in this thesis.

4.1.4. Overview of alert concentrations

Different specific concentrations that correspond to some pre-specified property of the response value often are of interest after the fitting of a parametric curve. In the 4pLL model, one such concentration is already incorporated into the model function as the parameter $\phi^{(e)}$. This parameter corresponds to the concentration where the half-maximal effect can be observed, i.e. the concentration that corresponds to a response of $\frac{\phi^{(c)}+\phi^{(d)}}{2}$.

Broadening this concept to any other percentage of the maximal effect yields the so-called *relative effective concentrations*. The effective concentration corresponding to a response of $(100\gamma)\%$ for $0 < \gamma < 1$ is denoted by EC100γ and defined as the solution for

$$f\left(\text{EC}100\gamma, \boldsymbol{\phi}\right) = (1-\gamma) \lim_{x \to 0} f(x, \boldsymbol{\phi}) + \gamma \lim_{x \to \infty} f(x, \boldsymbol{\phi}),$$

which, for the models considered in this work, typically translates to

$$f\left(\mathrm{EC100}\gamma, \boldsymbol{\phi}\right) = \begin{cases} (1-\gamma)\phi^{(c)} + \gamma\phi^{(d)} & \text{for an increasing curve,} \\ (1-\gamma)\phi^{(d)} + \gamma\phi^{(c)} & \text{for a decreasing curve.} \end{cases}$$

In particular, the limits are thus always well-defined (Ritz et al., 2019, p. 173).

For cytotoxicity data, where viability of cells is considered, in theory the upper and lower limits of a modelled concentration-response curve should correspond to 100% and 0%, respectively. In practice, values of the upper asymptote might differ due to deviating controls (see Chapter 2.1), although this can be addressed by different methods (see Chapter 4.2). However, it may happen that the observed viability for the highest tested concentration does not yet correspond to a viability of 0% and testing of higher concentrations is not possible due to problems with the solubility of the considered compound in higher concentrations.

In these cases, where indirectly a lower asymptote of 0% is assumed, but cannot be observed, instead of relative effective concentrations, the *absolute effective concentrations* are preferred. For $0 < \lambda < 100$ and a decreasing curve, the effective concentration corresponding to an absolute response of $(100 - \lambda)\%$ viability is denoted by EC_λ and defined as the solution of

$$f\left(\mathrm{EC}_\lambda, \boldsymbol{\phi}\right) = 100 - \lambda.$$

It can only be calculated for $\lambda \in (\phi^{(c)}, \phi^{(d)})$ or $\lambda \in (\phi^{(d)}, \phi^{(c)})$, depending on the relation between the asymptote values (Ritz et al., 2019, p. 174). In this work, for brevity when considering absolute effective concentrations for several values of λ, they are referred to as *EC values*.

Absolute and relative effective concentrations coincide in the case of upper and lower asymptotes of 100% and 0%, respectively. In that case, the parameter $\phi^{(e)}$ also corresponds to the definition of the EC_{50} as given above.

For a 4pLL model, an absolute effective concentration is explicitly calculated using the inverse of the model function f:

$$\mathrm{EC}_\lambda = \exp\left(\phi^{(e)^*}\right)\left(\frac{\phi^{(d)} - (100-\lambda)}{(100-\lambda) - \phi^{(c)}}\right)^{1/\phi^{(b)}}$$

For the BC model, no such closed formula for calculating absolute effective concentrations exists, these concentrations therefore need to be calculated numerically using an optimisation algorithm. Confidence intervals for the EC values are obtained as for the absolute lowest effective concentration, as explained in Chapter 4.3.2.

In the context of gene expression data, usually no fixed value for the upper or lower asymptote of a modelling curve can be assumed and the range of the expression values may differ vastly. Therefore, neither the concept of relative nor that of absolute effective concentrations fits the requirements for an alert concentration perfectly. Instead, as one solution, the absolute lowest effective concentration can be considered. Further

details regarding this and other alert concentrations for gene expression data are given in Chapter 4.3.

Further different alert concentrations exist, for example the benchmark dose methodology, where the lowest concentration with a noticeable effect compared to the normal response is identified (Jensen et al., 2019). This and further alert concentrations are not within the scope of this work and are therefore not considered further.

4.1.5. The MCP-Mod methodology

The **M**ultiple **C**omparison **P**rocedure and **Mod**elling (MCP-Mod) approach is a two-step approach for analysing concentration-response data and was originally developed by Bretz et al. (2005) in the context of dose-finding studies in pharmaceutical drug development. These studies typically have two goals, the first being the *Proof of Concept* (PoC), in which it is shown that changes in the administered dose of a drug lead to the desired change in the considered endpoint. If this PoC is shown, a dose-finding step follows, in which a model is fitted to the data and some alert concentration is calculated from this model (Bornkamp et al., 2009). MCP-Mod combines both steps into a single procedure.

Several candidate models are assessed at once with the MCP-Mod procedure, forming the set of candidate models. Using contrast tests, each model in this set of candidate models is tested while adjusting for mulitplicity, more specifically while preserving the family-wise error rate. A PoC is established if at least one of the models yields a significant result. The final concentration-response profile and possible alert concentrations are calculated using either the 'best' model in some way or an average of all models that passed the PoC (Bornkamp et al., 2009). In the application considered in this thesis, only the calculation of a p-value for a single model, namely the 4pLL model, is of interest. Therefore, in the introduction of the MCP-Mod procedure, emphasis is put on the PoC step and the modelling step is only explained shortly for completeness.

The MCP-Mod procedure is introduced here according to Bretz et al. (2005). As the original applications are dose-finding studies for a Phase II clinical study, the notation suggests the presence of different dose-levels instead of concentrations. However, Duda (2019) shows that simultaneous application of the MCP-Mod methodology to a high-dimensional concentration-gene expression dataset yields valid results.

Consider increasing dose levels $d_2 < \ldots < d_k$ of a compound and the negative control denoted as d_1. The number of replicates per dose level is given by n_i, $i = 1, \ldots, k$ with the total number of patients, or replicates in the application considered here, given by $N = n_1 + \ldots + n_k$. The dose-response relationship, with Y_{ij} denoting the response, is then modelled as

$$Y_{ij} = \mu_i + \varepsilon_{ij} = f(d_i, \boldsymbol{\theta}) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

with $j = 1, \ldots, n_i$ denoting the number of the replicate within dose group i, and homogenous variances $\sigma^2 > 0$. $\boldsymbol{\theta}$ denotes the vector of model parameters and $\boldsymbol{\mu} = (\mu_1, \ldots \mu_k)$ the vector of unknown means per dose.

The first step of the MCP-Mod procedure is the MCP-step, in which the candidate models from set $\mathcal{M} = \{M_1, \ldots, M_M\}$ of pre-specified models are simultaneously assessed using multiple contrast tests. A PoC is established for those models where the test yields a significant result and they are further considered for modelling.

The equivalent model to the 4pLL model introduced before in the parametrisation of the MCP-Mod approach is the sigmoid Emax (`sigEmax`) model (see Appendix A.4, where equivalence of both models is shown). It is parametrised as introduced in formula (3). This model and all other possible models describing the dose-response or concentration-response relationship can be reformulated as

$$f(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(d, \boldsymbol{\theta}^0),$$

where θ_0 is the location parameter, θ_1 the scale parameter and $f^0(d, \boldsymbol{\theta}^0)$ is the standardized version of the model. In the case of a `sigEmax` model, this standardized model is specifically given as

$$f^0(d, \boldsymbol{\theta}^0) = \frac{d^h}{(\mathrm{EC50}^h + d^h)},$$

i.e. $\theta_0 = E_0$ and $\theta_1 = E_{\max}$.

In order to perform the contrast test, initial values for the paramter vector $\boldsymbol{\theta}^0$ are required, further called *guesstimates*. In the originally intended application of this method, namely dose-finding clinical studies of Phase II, prior knowledge of the dose-response behaviour based on biological properties or previous studies usually is available. Here, in the application to gene expression data, the same guesstimates are used for all genes.

The optimal contrast test assesses the null hypothesis $\mathrm{H}_0^m : \boldsymbol{c}_m^{\mathrm{T}} \boldsymbol{\mu}_m = 0$ versus the alternative hypotheses $\mathrm{H}_1^m : \boldsymbol{c}_m^{\mathrm{T}} \boldsymbol{\mu}_m \neq 0$, with m denoting the m-th model of the set of candidate models with mean response vector $\boldsymbol{\mu}_m = (\mu_{m1}, \ldots, \mu_{mk})$ and $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{mk})^{\mathrm{T}}$ the vector of contrast coefficients with $\sum_{i=1}^k c_{mi} = 0$.

A contrast for model m is calculated aiming at maximizing power under the assumption that $\boldsymbol{\mu}$ is the true model shape. Contrasts fulfilling this goal are given as

$$c_{mi} \propto n_i(\mu_{mi}^0 - \bar{\mu}_m^0), \qquad i = 1, \ldots, k,$$

where $\mu_{mi}^0 = f_m^0(d_i, \boldsymbol{\theta}_m^0)$ and $\bar{\mu}_m^0 = \sum_{i=1}^k {}^{n_i \mu_{mi}^0}/_n$. A unique solution is then obtained by standardisation, i.e. ${}^{c_m}/_{||c_m||}$, where $||c_m|| = \left(\sum_{i=1}^k c_{mi}^2\right)^{\frac{1}{2}}$.

The test statistics $T_m, m = 1, \ldots, M$ for the optimal contrasts as above are given by

$$T_m = \frac{\sum_{i=1}^k c_{mi} \bar{Y}_i}{S\sqrt{\sum_{i=1}^k {}^{c_{mi}^2}/_{n_i}}}, \qquad m = 1, \ldots, M.$$

$\bar{Y}_i$ is the sample mean at dose d_i and $S = \sum_{i=1}^k \sum_{j=1}^{n_i} {}^{(Y_{ij} - \bar{Y}_i)^2}/_{(N-k)}$ denotes the mean squared error. Under H_0 and the distribution assumptions, the vector of test statistics $(T_1, \ldots, T_M)^{\mathrm{T}}$ follows a central multivariate t-distribution with $N-k$ degrees of freedom and a correlation matrix $\boldsymbol{R}$ depending on sample sizes and the optimal contrast vectors.

The maximum of the observed test statistics, $t_{\max}$ is compared to $q_{1-\alpha}$, where $q_{1-\alpha}$ is the equicoordinate $(1-\alpha)$-quantile of the corresponding multivariate t-distribution. If $t_{\max} > q_{1-\alpha}$, a dose-response signal is present in the data. The set of models with established PoC is formed by all models with $t_m > q_{1-\alpha}$ The family-wise error rate is strictly controlled at level α using this approach.

Modelling is only performed if at least one model passes the MCP-step and a PoC is established. As in this thesis only the p-values obtained from the MCP-step are of interest, the specifics of the model selection, model averaging and the modelling itself are not explained here.

The MCP-Mod methodology is implemented in the `R`-package `DoseFinding` (Bornkamp, 2019).

4.2. Handling deviating control values

Here, four methods are proposed to deal with deviating control values when fitting a concentration-response curve to the data. All methods are based on the family of log-logistic models (see Chapter 4.1.2). Three of them pursue the goal to obtain a function fitted to the data with an upper asymptote that corresponds to 100%. In the fourth method, sometimes a maximum value is attained by the curve, and if this is the case, this maximum value shall correspond to 100% instead of the upper asymptote. The methods were first proposed in Kappenberg et al. (2020).

When applying these methods, it is assumed that data are available in raw format and only normalised with respect to background correction, but unnormalised with respect to the control values. The four methods and their acronyms used throughout this thesis are:

- **4pLL**
  - In a first step, a 4pLL-model is fitted to the original data. The value of the upper asymptote (usually $\phi^{(d)}$ for decreasing curves) is extracted.
  - All data points are normalised with respect to the value of the upper asymptote.
  - A new 4pLL model is fitted to the normalised data.
- **3pLL**
  - The data is normalised with respect to the mean response values of the controls.
  - A 3pLL model with a fixed value of 100% for the upper asymptote is fitted.
- **No Ctrl**
  - In this method, all control values are omitted from the analysis.
  - A 4pLL model is fitted to the original data without the controls. The value of the upper asymptote (usually $\phi^{(d)}$ for decreasing curves) is extracted.

  - All data points are normalised with respect to the value of the upper asymptote.
  - A new 4pLL model is fitted to the normalised data without the controls.

- **BC**
  - A Brain-Cousens model is fitted to the original data.
  - Depending on whether a hormesis effect can be observed, the data is normalised with respect to the upper asymptote of the fit (if no hormesis effect is present) or the maximal value of the fitted curve (if a hormesis effect is present).
  - A new Brain-Cousens model is fitted to the normalised data.

Additionally, the approaches are illustrated in Figure 4.1, where the steps of applying all four methods to a fictional viability assay are shown. For **4pLL**, **No Ctrl** and **BC**, the respective left plot shows the initial fit to the unnormalised data. The controls are omitted for **No Ctrl**. For **BC**, a hormesis effect can be observed. The respective values of the upper asymptote or of the maximal value, that are used for normalisation, are indicated in the plot. In the right plot, the final fit after normalisation of the data is shown together with the respective values of the EC_{20}. For **3pLL**, only the single fit to the already normalised data is shown.

In theory, the refit-step in the methods **4pLL**, **No Ctrl** and **BC** would not be needed: After normalising the data, only the values of the asymptotes $\phi^{(c)}$ and $\phi^{(d)}$ (and $\phi^{(f)}$ for **BC**) change according to the normalisation, while the values of the parameters $\phi^{(b)}$ and $\phi^{(e)}$ remain unchanged. In practice, it was observed that this is not always the case: Starting values of the fitting procedure are range-dependent (Ritz et al., 2015) and therefore, due to numerical issues having to do with the choice of starting values, slightly different results may be obtained.

BC is the only method that can model a non-monotonous concentration-response relationship and is therefore mostly aimed at handling negatively deviating controls. In these cases, response values for the lowest tested positive concentrations are larger than the response values, which can be modelled by incorporating a hormesis effect into the model. Instead of only finding a way to deal with deviating controls via some normalisation, this model actually incorporates the deviation into the model fit.

The other three models only allow monotonously decreasing curves. They differ in the relevance of the control: **No Ctrl** completely omits the response values for the controls, therefore they have no weight at all when modelling the data. Conversely, **3pLL** first normalises all data points with respect to the response values of the controls and then takes the mean of the normalised responses of the controls, i.e. 100%, as the true value of the upper asymptote. Therefore, the course of the fitted curve is crucially influenced by the response values of the controls. In between these extremes, the **4pLL** model normalises the data with respect to an asymptote that is determined by the response values of the controls and of other low concentrations. To some extent, a mean value of the response values of the controls and the low concentrations is therefore used for normalisation of the data and the controls are neither omitted nor relied upon entirely.

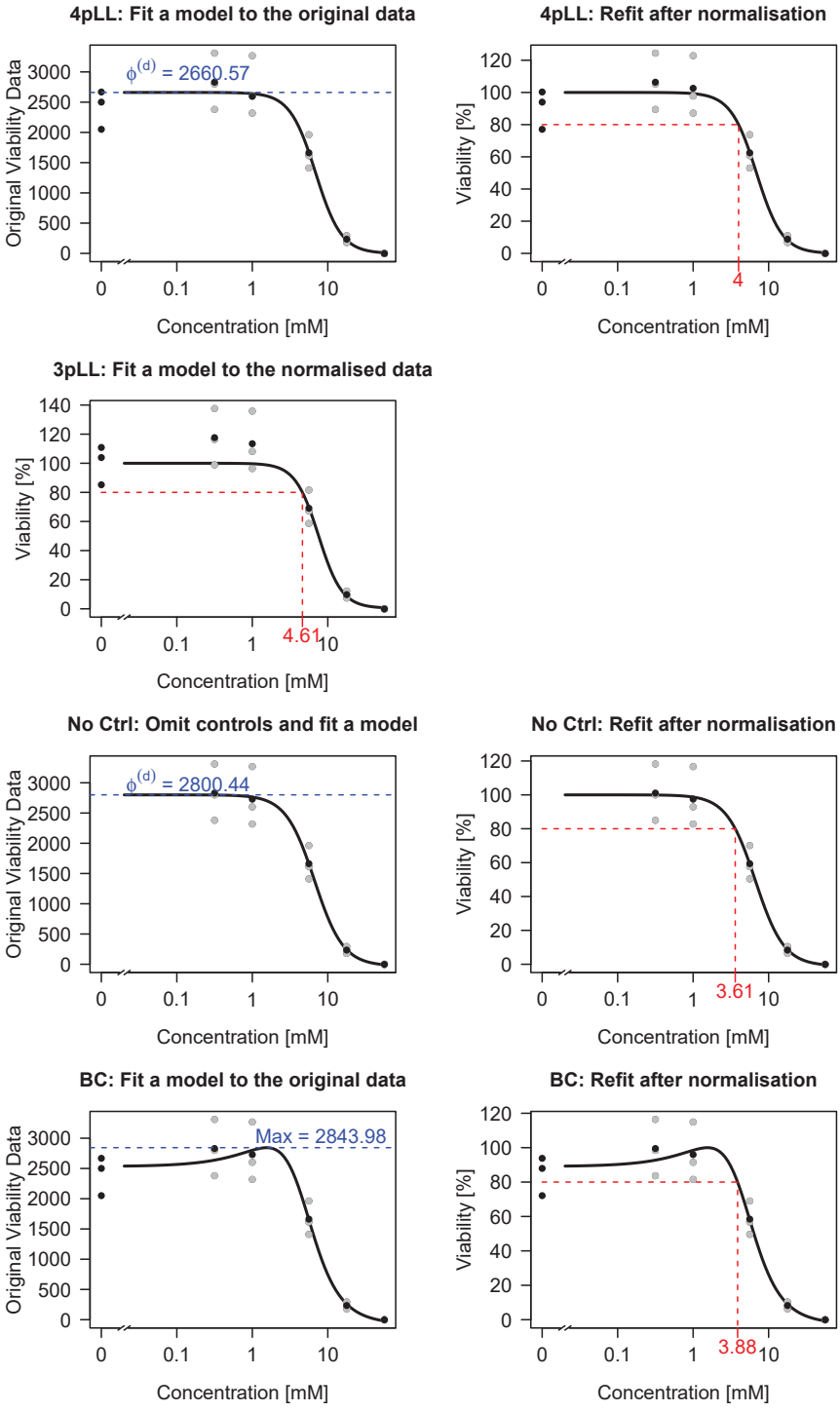

Figure 4.1: Graphical illustration of the approaches in all four proposed methods of dealing with deviating controls, with resulting estimates of the EC_{20}.

For the methods **4pLL**, **3pLL** and **No Ctrl**, a fit with an upper asymptote of 100% is the result of the procedure. For **BC**, this is only the case if no hormesis effect is present, otherwise instead the maximum value of the fitted curve corresponds to a response of 100% and the upper asymptote may take any value smaller than 100%.

4.3. Model-based and observation-based alert concentrations

For concentration-gene expression data, upper and lower limits of the fitted curves differ from gene to gene. Considering an absolute response value that needs to be attained in order to define an alert concentration is therefore not meaningful in this context. Instead, an absolute change in gene expression with respect to the left asymptote is of interest: The four methods presented in this chapter aim at determining a concentration where the fold change (FC) between this concentration and the control exceeds a pre-specified effect level $\lambda > 0$.

The methods are divided into two observation-based approaches, in which only the measured concentrations are potential alert concentrations, and two model-based approaches, where any concentration may be the alert concentration. Additionally, a differentiation in whether absolute or significant exceedance of the effect level λ is of interest is made. The acronyms of the four methods indicate which type of exceedance is considered: Both methods where only absolute exceedance is required contain the letter 'A' and both observation-based methods contain the letter 'O'.

The four methods are first described in detail, with most emphasis on the model-based approach that takes significance into account (Chapter 4.3.3). Then these four methods are summarised and the conditions needed to be fulfilled are compared and for easier understanding, they are displayed visually.

4.3.1. (Absolute) lowest observed effective concentration: ALOEC and LOEC

The observation-based alert concentrations are called the '**A**bsolute **L**owest **O**bserved **E**ffective **C**oncentration' (ALOEC) and the '**L**owest **O**bserved **E**ffective **C**oncentration' (LOEC). They are introduced here according to Grinberg (2017) with the difference, that significance testing is not conducted with the `limma` procedure as proposed there, but using a t-test or the Dunnett procedure.

The ALOEC is the smallest of the observed concentrations, where the difference in average gene expression between all samples for this concentration and the average gene expression for the negative control samples exceeds the threshold λ. For an increasing concentration-gene expression profile, this means that the FC needs to be larger than λ, and for a decreasing profile, the FC needs to be smaller than $-\lambda$. No significance testing is performed for this approach.

The LOEC additionally takes significance into account. Therefore, for increasing profiles, it is the smallest concentration where the difference in average gene expression

between all samples for this concentration and the average gene expression for the negative controls samples significantly exceeds λ. Accordingly, for decreasing profiles, this difference needs to be significantly smaller than $-\lambda$.

To test for significance, the easiest method is to perform a two-sample t-test for each concentration, taking the different variances of both samples into account, also called a Welch-test (Welch, 1947). Since several tests are performed, the problem of multiple testing arises. Therefore, a potential alternative is the Dunnett procedure (see Chapter 4.1.1), a methodology often used in the area of toxicology that adjusts for multiplicity.

In the general case, the direction of the gene expression profile (i.e. whether it is increasing or decreasing) is not known. For both alert concentrations, a reasonable restriction is to only calculate an alert concentration if the direction of the concentration gene-expression profile is unambiguous: If for one concentration-gene expression profile both a concentration where the difference to the control is (significantly) larger than λ and a concentration where the difference is (significantly) smaller than $-\lambda$ are found, no (A)LOEC is reported for the respective gene.

For the LOEC, two tests are performed per concentration to test in both directions. Both tests yield a p-value: The test assuming an increasing concentration-response pattern, where significant exceedance of λ is tested, yields the p-value called p_{inc} and the test assuming a decreasing concentration-response pattern yields p_{dec}. The final p-value is then calculated by $2 \cdot \min(p_{\text{inc}}, p_{\text{dec}})$.

4.3.2. Absolute lowest effective concentration: ALEC

The '**A**bsolute **L**owest **E**ffective **C**oncentration' (ALEC) is a measure for the toxicity of a test compound that depends on a regression function $y = f(x, \boldsymbol{\phi})$ fitted to the concentration-response data. For such a model function, the ALEC is defined as the concentration x where the fitted curve $f(x, \boldsymbol{\phi})$ attains a pre-specified effect level γ, i.e. $f(\text{ALEC}, \boldsymbol{\phi}) = \gamma$ (Jiang, 2013).

In the situation considered here, where the regression function is given by the 4pLL function, the ALEC is directly estimated from the inverse function by defining the function $h(\boldsymbol{\phi})$ as the inverse of the model function f. This is analogous to calculating an absolute effective concentration. Accordingly, the ALEC can only be calculated for an effect level γ that lies within the range of lower and upper asymptote of the function.

$$\begin{aligned} h(\hat{\boldsymbol{\phi}}) := h(\hat{\boldsymbol{\phi}}, \gamma) := \widehat{\text{ALEC}} &= f^{-1}\left(\gamma, \hat{\boldsymbol{\phi}}\right) \\ &= \hat{\phi}^{(e)} \left(\frac{\hat{\phi}^{(d)} - \gamma}{\gamma - \hat{\phi}^{(c)}} \right)^{1/\hat{\phi}^{(b)}} \end{aligned} \tag{7}$$

In the context of gene expression data as response value, upper and lower asymptotes of the fitted 4pLL curve differ from gene to gene. Therefore, no absolute value of the threshold can be specified in advance. Instead, the threshold of interest is composed of the value of the left asymptote of the curve, $f_0 := \hat{f}(0, \hat{\boldsymbol{\phi}})$ and the value λ that needs to

be exceeded in comparison to the asymptote. Again, a differentiation into increasing and decreasing profiles is required: For increasing profiles, the threshold γ is then defined by $\gamma = f_0 + \lambda$ and for decreasing profiles by $\gamma = f_0 - \lambda$. As only one of these values lies within the range of lower and upper asymptote of the fitted function, calculation of the ALEC is straightforward using the inverse formula.

To quantify uncertainties, Jiang (2013) derived several methods to calculate confidence intervals of the ALEC. One approach is to use the delta method (Vaart, 1998, p. 25) to approximate the variance of $h(\boldsymbol{\phi})$, yielding

$$\mathrm{Var}[h(\boldsymbol{\phi})] \approx \nabla h(\boldsymbol{\phi})^\top \Sigma \nabla h(\boldsymbol{\phi}), \tag{8}$$

with Σ denoting the covariance matrix of the parameter vector $\boldsymbol{\phi}$ and $\nabla h(\boldsymbol{\phi})$ the gradient of h with respect to the parameter vector $\boldsymbol{\phi}$:

$$\begin{aligned}\nabla h(\boldsymbol{\phi}) &= \left(\tfrac{\partial h(\boldsymbol{\phi})}{\partial \phi^{(b)}}, \tfrac{\partial h(\boldsymbol{\phi})}{\partial \phi^{(c)}}, \tfrac{\partial h(\boldsymbol{\phi})}{\partial \phi^{(d)}}, \tfrac{\partial h(\boldsymbol{\phi})}{\partial \phi^{(e)}}\right)^\top \\ &= h(\boldsymbol{\phi}) \begin{pmatrix} -\frac{1}{\phi^{(b)}} \log\left(\frac{\phi^{(d)}-\gamma}{\gamma-\phi^{(c)}}\right) \\ \frac{1}{\phi^{(b)}\left(\gamma-\phi^{(c)}\right)} \\ \frac{1}{\phi^{(b)}\left(\phi^{(d)}-\gamma\right)} \\ \frac{1}{\phi^{(e)}} \end{pmatrix}\end{aligned}$$

See Jiang (2013) and Grinberg (2017) for a detailed derivation. By plugging in the estimated parameter vector $\hat{\boldsymbol{\phi}}$ and the estimated covariance matrix $\hat{\Sigma}$ into formulas (7) and (8), the estimated value $\widehat{\mathrm{ALEC}}$ with corresponding variance $\widehat{\mathrm{var}}[\widehat{\mathrm{ALEC}}]$ are obtained (Jiang, 2013).

The $(1-\alpha)100\%$ confidence interval for the ALEC is then given by

$$\widehat{\mathrm{ALEC}} \pm t_{(1-\alpha/2),\nu} \sqrt{\hat{\mathrm{var}}[\widehat{\mathrm{ALEC}}]},$$

with $t_{(1-\alpha/2),\nu}$ being the $(1-\alpha/2)$-quantile of a t-distribution with ν degrees of freedom. In the case of a 4pLL model, $\nu = n - 4$ with n denoting the number of data points.

A problem with this method is that lower confidence limits may attain negative values. As the ALEC is a concentration and can therefore only take positive values, this can lead to implausible results and a method for determining the confidence interval for the ALEC based on the variance on the logarithm of the ALEC may therefore be more appropriate. Again, the delta method is used for approximating the variance of log(ALEC). Then, a $(1-\alpha)100\%$ confidence interval CI for the ALEC can be obtained by back transformation (Jiang, 2013):

$$\log(\text{ALEC}) = \log(h(\boldsymbol{\phi})) = \log(\phi^{(e)}) + \frac{1}{\phi^{(b)}} \log\left(\frac{\phi^{(d)} - \gamma}{\gamma - \phi^{(c)}}\right)$$

$$\text{Var}[\log(h(\boldsymbol{\phi}))] \approx \nabla \log(h(\boldsymbol{\phi}))^\top \Sigma \nabla \log(h(\boldsymbol{\phi}))$$

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

Further methods, such as the profile likelihood method or different Bootstrap approaches, are compared to the two methods introduced above in a simulation study in Jiang (2013). Results show that Bootstrap-based confidence intervals have a low coverage probability and are therefore not an appropriate approach for determining confidence intervals. Both delta-method based intervals performed almost identically and sufficiently good. So although the profile likelihood method performed better especially in situations with missing concentrations in the lower range of the curve, in this work, for simplicity, calculation of the confidence interval for the ALEC is based on the logarithmic confidence interval based on the delta method.

4.3.3. Lowest effective concentration: LEC

The model-based approach that additionally takes significance into account yields the '**L**owest **E**ffective **C**oncentration' (LEC) as alert concentration. In order to determine this concentration, a test procedure is required to assess whether the difference of function values of a fitted curve $f(\cdot, \boldsymbol{\phi})$ for a concentration x and the concentration 0 significantly exceeds a pre-specified threshold λ. This test is first introduced in a general form where the difference of function values for any two concentrations $x_1 > 0$ and $x_2 > 0$ are considered, and then applied to the special case of finding the LEC. A bisection algorithm from Grinberg (2017) is presented, which is a search algorithm to actually determine the LEC making use of the newly introduced test statistic. The new test statistic is published in Kappenberg et al. (2021) and is based on a first version of this test from Grinberg (2017).

In the general case, for two concentrations x_1 and x_2 with $x_1 \geq 0$, $x_2 \geq 0$, and $x_1 \neq x_2$, and $f(\cdot, \boldsymbol{\phi})$ a 4pLL modell with parameter vector $\boldsymbol{\phi}$, the hypotheses of interest are given by

$$\text{H}_0 : f(x_1, \boldsymbol{\phi}) = f(x_2, \boldsymbol{\phi}),$$
$$\text{H}_1 : f(x_1, \boldsymbol{\phi}) \neq f(x_2, \boldsymbol{\phi}).$$

The observed test statistic is formulated as

$$t_{\text{4pLL}} := t_{\text{4pLL}}(x_1, x_2, \hat{\boldsymbol{\phi}}) = \frac{\hat{f}\left(x_1, \hat{\boldsymbol{\phi}}\right) - \hat{f}\left(x_2, \hat{\boldsymbol{\phi}}\right)}{\sqrt{\hat{\text{var}}\left[\hat{f}\left(x_1, \hat{\boldsymbol{\phi}}\right) - \hat{f}\left(x_2, \hat{\boldsymbol{\phi}}\right)\right]}}. \tag{9}$$

The variance term in (9) can be re-written as

$$\begin{aligned}&\operatorname{Var}\left[\hat{f}\left(x_1,\hat{\boldsymbol{\phi}}\right)-\hat{f}\left(x_2,\hat{\boldsymbol{\phi}}\right)\right]\\&\quad=\operatorname{Var}\left[\hat{f}(x_1,\hat{\boldsymbol{\phi}})\right]+\operatorname{Var}\left[\hat{f}(x_2,\hat{\boldsymbol{\phi}})\right]-2\cdot\operatorname{Cov}\left[\hat{f}(x_1,\hat{\boldsymbol{\phi}}),\hat{f}(x_2,\hat{\boldsymbol{\phi}})\right].\end{aligned}$$

As $\hat{f}(x_1,\hat{\boldsymbol{\phi}})$ and $\hat{f}(x_2,\hat{\boldsymbol{\phi}})$ are highly correlated, the covariance term in calculating the variance of the difference does not vanish and needs to be taken into account. Using the delta rule, the variance term in (9) is approximated as follows:

$$\begin{aligned}\operatorname{Var}\left[\hat{f}\left(x_1,\hat{\boldsymbol{\phi}}\right)-\hat{f}\left(x_2,\hat{\boldsymbol{\phi}}\right)\right] &= \operatorname{Var}\left[\hat{f}(x_1,\hat{\boldsymbol{\phi}})\right]+\operatorname{Var}\left[\hat{f}(x_2,\hat{\boldsymbol{\phi}})\right]\\&\quad-2\cdot\operatorname{Cov}\left[\hat{f}(x_1,\hat{\boldsymbol{\phi}}),\hat{f}(x_2,\hat{\boldsymbol{\phi}})\right]\\&\approx \nabla\hat{f}(x_1,\hat{\boldsymbol{\phi}})^{\mathrm{T}}\Sigma\nabla\hat{f}(x_1,\hat{\boldsymbol{\phi}})+\nabla\hat{f}(x_2,\hat{\boldsymbol{\phi}})^{\mathrm{T}}\Sigma\nabla\hat{f}(x_2,\hat{\boldsymbol{\phi}})\\&\quad-2\cdot\nabla\hat{f}(x_1,\hat{\boldsymbol{\phi}})^{\mathrm{T}}\Sigma\nabla\hat{f}(x_2,\hat{\boldsymbol{\phi}})\end{aligned}\tag{10}$$

Σ denotes the covariance matrix of the parameters and $\nabla f(x,\boldsymbol{\phi})$ denotes the gradient of f with respect to the parameter vector $\boldsymbol{\phi}$, see Grinberg (2017) for a detailed derivation:

$$\nabla f(x,\boldsymbol{\phi})=\left(\tfrac{\partial f(x,\boldsymbol{\phi})}{\partial\phi^{(b)}},\tfrac{\partial f(x,\boldsymbol{\phi})}{\partial\phi^{(c)}},\tfrac{\partial f(x,\boldsymbol{\phi})}{\partial\phi^{(d)}},\tfrac{\partial f(x,\boldsymbol{\phi})}{\partial\phi^{(e)}}\right)^{\top}$$

$$=\begin{pmatrix}-\dfrac{\left(\phi^{(d)}-\phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2}\\[2ex]1-\left[\dfrac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}\right]\\[2ex]\dfrac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}\\[2ex]\dfrac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2}\end{pmatrix}$$

Under the null hypothesis, it asymptotically holds that the test statistic is normally distributed with mean 0 and variance 1. Therefore the null hypothesis is rejected at level α if the observed value $t_{4\mathrm{pLL}}$ exceeds $z_{1-\alpha/2}$ or is smaller than $z_{\alpha/2}$. z_q denotes the q-quantile of the standard normal distribution.

When applying this test to the situation where the goal is to find the LEC as alert concentration, only one concentration $x>0$ is considered in comparison to the concentration 0. Two null hypotheses are formulated, for an increasing and for a decreasing

concentration-gene expression profile. For a pre-defined threshold $\lambda > 0$ they are as follows:

$$\mathrm{H}_0 : f(x, \boldsymbol{\phi}) - f(0, \boldsymbol{\phi}) \leq \lambda \quad \text{for an increasing curve} \tag{11}$$
$$\mathrm{H}_0 : f(x, \boldsymbol{\phi}) - f(0, \boldsymbol{\phi}) \geq -\lambda \quad \text{for a decreasing curve} \tag{12}$$

If for the gene of interest the direction of the curve is known beforehand from the biological background, only the respective hypothesis needs to be tested. The assumptions regarding the distribution of the two test statistics presented below remain the same as in the general case.

The test statistic for an increasing curve is given by

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

The corresponding p-value is calculated as $1 - \Phi(t_{4\mathrm{pLL};\ \mathrm{inc}})$ with Φ denoting the distribution function of the standard normal distribution.

The test statistic for a decreasing curve is analogously given by

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

with corresponding p-value calculated as $\Phi(t_{4\mathrm{pLL};\ \mathrm{dec}})$.

Since in general, the direction of the curve is not known in advance, both test statistics are calculated independently and their respective p-values are determined. A two-sided p-value is then calculated as

$$2 \cdot \min\left(1 - \Phi(t_{4\mathrm{pLL};\ \mathrm{inc}}),\ \Phi(t_{4\mathrm{pLL};\ \mathrm{dec}})\right). \tag{13}$$

In the special case considered for finding the LEC, for $\phi^{(b)} > 0$ the left asymptote is given by $\phi^{(d)}$, and for $\phi^{(b)} < 0$ the left asymptote is given by $\phi^{(c)}$. The corresponding values of the gradient $\nabla f(0, \boldsymbol{\phi})$ simplify to

$$\nabla f(0, \boldsymbol{\phi}) = (0, 0, 1, 0)^\top \qquad \text{for } \phi^{(b)} > 0,$$
$$\nabla f(0, \boldsymbol{\phi}) = (0, 1, 0, 0)^\top \qquad \text{for } \phi^{(b)} < 0.$$

This means that in the limit of $x \to 0$, only parameter $\phi^{(c)}$ or $\phi^{(d)}$, respectively, has an influence on the function value. The exact calculation is given in Appendix A.5. The second summand in the variance term (10) therefore simplifies to Σ_{33} for $\phi^{(b)} > 0$ and to Σ_{22} for $\phi^{(b)} < 0$. Diagonal entries of the covariance matrix Σ correspond to the squared standard error of the estimated coefficients $\boldsymbol{\phi}$, i.e. $\hat{\Sigma}_{22} = \hat{\mathrm{se}}(\hat{\phi}^{(c)})^2$ and $\hat{\Sigma}_{33} = \hat{\mathrm{se}}(\hat{\phi}^{(d)})^2$.

A search algorithm is required to find the LEC as the smallest concentration $x > 0$ in the tested concentration range where testing any or both hypotheses (11) and (12) yields

a significant result. A significant result is identified by a p-value smaller than $\alpha = 0.05$. In the data example considered here, the tested concentration range goes from 0 µM to 1000 µM. The search algorithm presented here is a version of a bisection algorithm and was published by Grinberg (2017).

As a first step, a p-value as in (13) is calculated for the highest concentration in the range of considered concentrations and compared to the prespecified significance level α. Since the 4pLL curve that is the basis for this alert concentration is monotonously increasing or decreasing, this test result already determines whether the LEC exists: If this test for exceedance of the effect level λ does not yield a significant result even for the highest concentration, no LEC can be determined. Otherwise, the bisection algorithm is conducted.

The first interval is limited by the lowest and highest concentration considered. The first concentration considered is the mean concentration of this interval. A p-value as in (13) is calculated and compared to α. If the p-value is smaller than α, then the concentration considered is not the smallest concentration fulfilling the criteria for the LEC, therefore the parameter space is restricted to the lower half of the considered interval. In the other case, where the p-value is larger than α, significant exceedance of the threshold is only achieved for higher concentrations and the parameter space is restricted to the upper half of the interval.

This procedure of conducting the 4pLL model based test for the respective mean of the considered interval is repeated until the length of the remaining interval is smaller than a small pre-specified threshold $\varepsilon > 0$. The last concentration for which the test was conducted is then taken as the LEC.

4.3.4. Summary of all four alert concentrations

To give a definitive overview how the four alert concentrations introduced in the previous sections differ from each other and how they are referred to throughout this thesis, the four methods are summarised in Table 4.1. The letter 'O' in the names ALOEC

Table 4.1: Comparison of the four methods for estimating alert concentrations from concentration-gene expression data. The cutoff critera that either a foldchange value is exceeded (FC) or that additionally it is significantly exceeded (FC & p-value) are indicated in the rows. The columns indicate the methods for estimating fold changes, either using a t-test / the Dunnett procedure or a 4pLL model.

	Observation-based	Model-based
	***t*-test / Dunnett**	**4pLL**
FC	ALOEC	ALEC
FC & *p*-value	LOEC	LEC

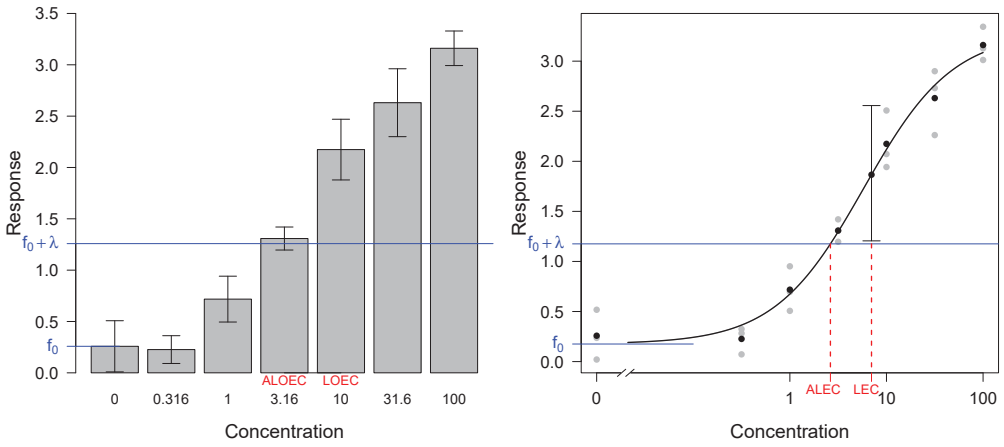

Figure 4.2: Hypothetical example illustrating the four different alert concentrations for concentration gene-expression data.

and LOEC indicates that in comparison to ALEC and LEC, only the observed concentrations are potential alert concentrations. The letter 'A' in the names ALOEC and ALEC indicates that in comparison to LOEC and LEC, only absolute exceedance of the threshold without assessing the significance is required.

The four methods are finally visualised in Figure 4.2. In this plot, a hypothetical example of concentration-response data is visualised in two ways: Left, concentration-wise means of the response are depicted in a barplot and additionally, the standard deviation is indicated. The response value for the control is denoted as f_0, therefore the concentrations are of interest, where $f_0 + \lambda$ is absolutely and significantly exceeded, yielding the ALOEC and the LOEC respectively, as indicated.

On the right side, a 4pLL model is fitted to the data. Denoting the value of the lower asymptote by f_0, the ALEC is the concentration where the curve attains the value of $f_0 + \lambda$. The display of the LEC is to be understood rather heuristically: While the plot shows a confidence interval for a concentration x that completely lies above the response given by $f_0 + \lambda$, in the test statistics introduced in the previous chapter the variance of the difference between the function evaluated at concentration x and at concentration 0 is of interest.

4.4. Information sharing across genes

Two methods are presented, how information are shared across genes in order to improve the estimation of the parameter $\phi^{(e)*}$. This improvement is assessed with respect to the difference to the true underlying parameter in a simulation situation and with respect to the coverage probabilities of the associated confidence intervals.

The first method is the concept of *meta-analyses*, that is introduced in the general case and in the specific application of concentration-response curves in Chapter 4.4.2. The

second method is an empirical Bayes approach, where parameters are shrunk based on a normal-normal model. This approach is introduced together with the basics of Bayesian statistics in Chapter 4.4.3. All methods again are based on the assumption of a sigmoidal relationship between concentration and gene expression that can be modelled by a 4pLL model.

Both methods are evaluated by simulation studies in different scenarios. Regarding the performance of the methods, biological structures represented in relationships between the four parameters of the 4pLL model play a major part. Therefore, a fully theoretical simulation study may miss out on important observations and is an unsuitable basis for giving recommendations about the best approach to share information across genes. Instead, a simulation study based on real data is conducted. This approach is called *plasmode* and is introduced in the following Chapter 4.4.1.

4.4.1. Plasmode simulation study

The basic idea of using plasmode data is to base the simulation study on a real data situation while at the same time manipulating the data in a way that true effects are known. Vaughan et al. (2009) define a plasmode dataset by stating three conditions that need to be fulfilled: The dataset needs to be "the result of a real biological process", instead of being simulated by a computer only and is modified or constructed in a way that "at least some aspect of the 'truth' of the data generating process is known" (both from Vaughan et al., 2009).

In this work, the analyses regarding the information sharing across genes are performed based on the VPA gene expression dataset introduced in detail in Chapter 3.2.2. For the plasmode simulation studies, a subset of probe sets from this dataset is selected. The values from an initial fit of the 4pLL model are considered as true underlying values. That way, relationships between the values of the parameters are retained. Based on these parameters, new concentration-response data are simulated. The methods introduced in the following chapters are then evaluated on these data, leading to results applicable to actual data and not only to theoretic situations. The simulation studies constructed based on this idea are each introduced in more detail in Chapters 7.3 and 7.4.

4.4.2. Summarising parameters using meta-analysis

A meta-analysis is a statistical procedure often applied in the context of clinical studies that allows the combination of several studies that are aimed at answering the same question, thus yielding a pooled estimate with a narrower confidence interval. There are several approaches possible to combine the data, e.g. via p-values or via the effect of interest. The latter version is considered in this work, where the effect of interest is the parameter $\phi^{(e)}$, or ${\phi^{(e)}}^*$.

In the general case, let $\hat{\theta}_i$ for $i = 1, \ldots, k$ be the estimated effect of the i-th experiment, usually the i-th study considered. The basic assumption is the normality of θ_i, i.e. $\theta_i \sim \mathcal{N}(\theta, \tau^2 + \sigma_i^2)$, with θ denoting the true underlying mean, τ^2 denoting the true

variance between experiments and σ_i^2 the true variance within the experiment (Hartung and Knapp, 2001b).

A distinction is made between the case in which $\tau^2 = 0$, i.e. the experiments are homogeneous, and the case in which $\tau^2 > 0$, i.e. there is additional heterogeneity between the experiments. The first case is called *fixed effects model* and the second case *random effects model* (Hartung and Knapp, 2001b).

The estimation of the parameter θ is based on a weighted mean of the single estimates $\hat{\theta}_i$, where the weights depend on estimates $\hat{\sigma}_i^2$ and, in the random effects model only, on $\hat{\tau}_i^2$. A variety of methods for estimating τ^2 exist. The estimator by DerSimonian and Laird (1986) is used in this work and is explained later explicitly in the context considered here. The application of the general meta-analysis to the specific situation of summarising parameters from a 4pLL model is presented according to Jiang and Kopp-Schneider (2014). Specifically, summarising parameters $\phi^{(e)}$ estimated from several curves is introduced there.

Consider the situations with k datasets with concentration-response data to each of which a 4pLL model is fitted. The estimate of $\phi^{(e)}$ obtained by each of the models is denoted by $\phi_i^{(e)}$ for $i = 1, \ldots, k$. Jiang and Kopp-Schneider (2014) propose the use of a random-effects model in order to explain both the sampling variance and the heterogeneity between the experiments leading to the different datasets. Under the assumption of independence for a set $\phi_i^{(e)}$, $i = 1, \ldots, k$ of estimates, the model is formulated as

$$\phi_i^{(e)} = \theta_i + \varepsilon_i = \mu + \mu_i + \varepsilon_i, \tag{14}$$

with θ_i denoting the true value of the parameter $\phi^{(e)}$ in the i-th experiment and ε_i denoting the normally distributed, heteroscedastic sampling error $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. μ denotes the true average of the estimates and μ_i is the random-effect term that represents the deviation of θ_i from the true value μ. It is assumed that $\mu_i \sim \mathcal{N}(0, \tau^2)$, i.e. the unexplained variability of θ_i is purely random. τ^2 is the variance indicating the variability of the true estimates.

Under the assumption of independence of μ_i and ε_i, it holds $\phi_i^{(e)} \sim \mathcal{N}(\mu, \tau^2 + \sigma_i^2)$. The goal of the meta-analysis is the estimation of μ and the standard error $\mathrm{se}(\mu)$ in model (14), leading to a pooled estimate of parameter $\phi^{(e)}$. The estimation of μ is conducted using a weighted least squares estimator. The meta-analysis model (14) is rewritten as

$$\phi_i^{(e)} = \mu + \varepsilon_i^*, \tag{15}$$

with $\varepsilon_i^* \sim \mathcal{N}(0, \tau^2 + \sigma_i^2)$. Weights w_i are defined as $w_i = 1/(\tau^2 + \sigma_i^2)$, where σ_i^2 is the squared estimated standard error of $\phi^{(e)}$. The best linear unbiased estimator for μ in model (15) is then given by

$$\hat{\mu} = \frac{\sum_{i=1}^k \hat{w}_i \phi_i^{(e)}}{\sum_{i=1}^k \hat{w}_i}, \tag{16}$$

with $\mathrm{Var}[\hat{\mu}] = 1/\sum_{i=1}^k \hat{w}_i \phi_i^{(e)}$, with $\hat{w}_i$ denoting the estimate of w_i (Jiang and Kopp-Schneider, 2014).

As the weight w_i is dependent on the unknown parameter τ^2, a two-step approach is required for estimation of μ in formula (15), where in the first step, τ^2 is estimated and then based on this estimate, the value of μ is estimated by $\hat{\mu}$. Jiang and Kopp-Schneider (2014) propose several variants to estimate τ^2, while in this work only the version by DerSimonian and Laird (1986) is considered. Let $m_i = {}^1\!/\!_{\hat{\sigma}_i^2}$ and $\overline{\phi^{(e)}}_m = \frac{\sum_{i=1}^k m_i \phi_i^{(e)}}{\sum_{i=1}^k m_i}$. The proposed estimate of τ^2 in model (15) is given by

$$\hat{\tau}^2 = \max\left\{0,\ \frac{\sum_{i=1}^k m_i \left(\phi_i^{(e)} - \overline{\phi^{(e)}}_m\right)^2 - (k-1)}{\sum_{i=1}^k m_i - \sum_{i=1}^k {m_i}^2 / \sum_{i=1}^k m_i}\right\}. \tag{17}$$

Jiang and Kopp-Schneider (2014) propose four variants of calculating confidence intervals for the estimate $\hat{\mu}$. The first three take the classic form $\hat{\mu} \pm q_{1-\alpha/2}\hat{\text{se}}(\hat{\mu})$, where α is the pre-specified effect level and $q_{1-\alpha/2}$ the upper ${}^\alpha\!/\!_2$-quantile of some distribution that is specified later. The standard error of $\hat{\mu}$ is estimated by $\hat{\text{se}}(\hat{\mu}) = \sqrt{\frac{1}{\sum_{i=1}^k \hat{w}_i}}$.

Under the assumption of asymptotic normality for $\hat{\mu}$, the choice of $q_{1-\alpha/2}$ as $z_{1-\alpha/2}$, i.e. the upper ${}^\alpha\!/\!_2$-quantile of the standard normal distribution yields an approximate $100(1-\alpha)\%$ confidence interval. The coverage probability may be improved using a t-distribution instead, as the assumption of asymptotic normality may not be fulfilled for a small number k of experiments and additionally, uncertainty in the estimation of τ^2 may affect the estimation of $\hat{\text{se}}(\hat{\mu})$, as this depends on τ^2 via $\hat{w}_i$. Therefore, for the second and the third variant, choosing t-distributions with $k-1$ and $k-2$ degrees of freedom, respectively, is proposed (Jiang and Kopp-Schneider, 2014).

The fourth variant is based on a modified Wald statistic proposed by Hartung and Knapp (2001a,b), where the corresponding $100(1-\alpha)\%$ confidence interval is calculated as

$$\hat{\mu} \pm t_{k-1,1-\alpha/2}\sqrt{\frac{\sum_{i=1}^k \hat{w}_i \left(\phi_i^{(e)} - \hat{\mu}\right)^2}{(k-1)\sum_{i=1}^k \hat{w}_i}},$$

with $t_{k-1,1-\alpha/2}$ denoting the upper ${}^\alpha\!/\!_2$-quantile of the t-distribution with $k-1$ degrees of freedom (Jiang and Kopp-Schneider, 2014).

All four variants are compared in different simulation scenarios by Jiang and Kopp-Schneider (2014). The results show that the confidence interval based on the normal distribution performs worst in terms of coverage probability and the fourth variant, which is based on the modified Wald statistic, performs best.

Although the meta-analysis is presented here for the parameter $\phi^{(e)}$ itself, it is applied to $\phi^{(e)*}$ in this work, as for the size of the dataset considered, the assumption of normality is rather fulfilled for this re-parametrised variant.

4.4.3. Shrinkage of parameters using an empirical Bayes method

Before the shrinkage of parameters itself is explained, a short introduction into Bayesian statistics is given. Specifically, the concept of empirical Bayes statistics is introduced, where prior distributions are estimated from the observed data itself. Main emphasis is put on the normal-normal model, which is used throughout the analyses. The application of these methods to the situation considered in this work, where shrinkage of the parameter $\phi^{(e)^*} = \log(\phi^{(e)})$ is of interest, is explained in detail. As in one of the simulation studies presented in this work, normalised values of the parameter $\phi^{(e)^*}$ are examined, the concept of quantile normalisation is explained as well.

The fundament of Bayesian statistics is *Bayes' theorem*. Let A and B be two different events with $\mathbb{P}[B] > 0$. Then Bayes' theorem states

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\,\mathbb{P}[A]}{\mathbb{P}[B]},$$

where $\mathbb{P}[A|B]$ and $\mathbb{P}[B|A]$ denote the conditional probability of A, given that B is true and vice versa. This theorem allows calculation of the probability of an event given an other event, when among other the probability of the reverse condition is known.

In the context of Bayesian inference, let $\boldsymbol{X} = (X_1, \ldots, X_n), n \in \mathbb{N}$ be the random variable leading to observations $(x_1, \ldots, x_n)$. The density function $f(\boldsymbol{X}, \boldsymbol{\theta})$, called *likelihood function*, depends on the fixed, but unknown value of the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p), p \in \mathbb{N}$. The goal of Bayesian inference is the estimation of the parameter vector $\boldsymbol{\theta}$ that is considered to be a random variable. The specification of a prior distribution $\pi(\boldsymbol{\theta})$ is required in order to be able to apply Bayes' rule (Reich and Ghosh, 2019, p. 21).

Under the assumptions described above, a posterior distribution of $\boldsymbol{\theta}$, conditional on $\boldsymbol{X}$ is given by

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \propto f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{18}$$

i.e. the posterior distribution is proportional to the product of likelihood function and prior distribution (Reich and Ghosh, 2019, p. 21).

The intuition is that first, the prior distribution captures uncertainty about the parameters before any data is observed. In formula (18), the posterior distribution captures uncertainty that remains after both accounting for the acutal observed data and the prior knowledge (Reich and Ghosh, 2019, p. 21).

To describe uncertainty of the estimated posterior parameter, in the context of Bayes statistics, *credible intervals* are calculated. A $(1-\alpha)100\%$ credible interval is any interval (l, u) with the property that $\mathbb{P}[l < \boldsymbol{\theta} < u|\boldsymbol{X}] = 1 - \alpha$. Infinitely many such intervals can be calculated from a given posterior distribution. The easiest way, which is also employed in this work, is to set $l = z_{\alpha/2}$ and $u = z_{1-\alpha/2}$, with z_q denoting the $q\%$-quantile of the posterior distribution. Regarding the interpretation of a credible interval, for a given prior and the observed data, the $(1-\alpha)100\%$ credible interval (l, u) covers the true value of the parameter $\boldsymbol{\theta}$ with a probability of $(1-\alpha)100\%$ (Reich and Ghosh, 2019, p. 26).

One important step in Bayesian analyses is the choice of a prior distribution. Depending on this distribution, more or less weight is put on the prior or on the observed data. If knowledge about the parameters, e.g. from previous experiments, is present, the prior should reflect this knowledge. If, however, no such knowledge is available, then the prior should be *uninformative*, e.g. by choosing a prior that has a uniform distribution. Additionally, the functional form of the prior influences the functional form of the resulting posterior distribution.

In this work, an *empirical prior* is chosen. This is a prior that makes use of the data to estimate the actual prior distribution, e.g. by plugging in estimated parameters from the dataset to the specific functional form of a distribution. Despite the problem with this approach, where the data is considered twice in estimating the posterior distribution, this is a useful approach especially for high-dimensional data sets as considered here (Reich and Ghosh, 2019, pp. 62-63).

Certain combinations of likelihood functions and prior distributions lead to closed forms of the posterior distributions that can be analytically calculated and do not need to be simulated. A pair of prior distribution and likelihood function is called *conjugate*, if prior and resulting posterior distribution stem from the same family of distributions (Reich and Ghosh, 2019, p. 42). One specific example of such a conjugate pair of prior and likelihood function, the normal-normal model, is used in this work.

Let $\boldsymbol{X}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2)$ with the prior distribution $\pi(\boldsymbol{\theta})$ given by $\boldsymbol{\theta} \sim \mathcal{N}(\mu, \tau^2)$. Denote by $x \in \mathbb{R}$ the observed value of $\boldsymbol{X}$. Then the posterior distribution $p(\boldsymbol{\theta}|x)$ from equation (18) is given by

$$\boldsymbol{\theta}|x \sim \mathcal{N}\left(\frac{\tau^2 x + \sigma^2 \mu}{\tau^2 + \sigma^2}, \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}\right), \tag{19}$$

see Appendix A.6 for a detailed calculation.

In the specific application considered here, the parameter ${\phi^{(e)}}^*$ is assumed to be normally distributed with mean $\boldsymbol{\theta}$ and variance σ^2. A large number n of genes is considered simultaneously, yielding estimates ${\hat{\phi}^{(e)}}^*_1, \ldots, {\hat{\phi}^{(e)}}^*_n$, based on which the parameters μ and τ^2 of the prior distribution are calculated.

The first possibility to estimate these parameters is to use the maximum-likelihood estimation

$$\hat{\mu}_{\text{ML}} = \widehat{\text{mean}}({\hat{\phi}^{(e)}}^*_i) \quad \text{and} \quad \hat{\tau}^2_{\text{ML}} = \widehat{\text{var}}({\hat{\phi}^{(e)}}^*_i).$$

The second possibility is to use robust measures

$$\hat{\mu}_{\text{rob}} = \widehat{\text{median}}({\hat{\phi}^{(e)}}^*_i) \quad \text{and} \quad \hat{\tau}^2_{\text{rob}} = (1.4826 \cdot \widehat{\text{MAD}}({\hat{\phi}^{(e)}}^*_i))^2.$$

MAD denotes the median absolute deviation, and multiplication with the factor 1.4826 ensures consistency for normally distributed data, i.e. convergence to the true variance for increasing sample sizes. The parameter σ^2 is individually estimated for each gene as the squared standard error of ${\phi^{(e)}}^*$, calculated as presented in equation (6).

To ensure that the normality assumption for ${\phi^{(e)}}^*$ holds, one simulation scenario is considered where the distribution of this parameter for the set of genes considered is made

equal to a normal distribution. Specifically, this means that the original data are manipulated to form the normalised dataset with normally distributed values of $\phi^{(e)^*}$. This is achieved by applying *quantile normalisation*, a procedure that was originally developed for the normalisation of gene expression data (Bolstad et al., 2003). A quantile-quantile-plot (qq-plot) is considered, in which sample quantiles and theoretical quantiles of the normal distribution are plotted against each other.

Specifically, quantile normalisation for a set of observations of $\phi^{(e)^*}$ is conducted by calculating the theoretical quantiles in a qq-plot corresponding to the sample quantiles. Then the values of the sample quantiles are set to the respective values of the theoretical quantiles, yielding values that follow a standard normal distribution. Multiplying with the standard distribution of the originally observed set of $\phi^{(e)^*}$ and adding the mean value finally yields a dataset that follows a normal distribution with the same mean and standard deviation as the originally observed sample.

4.5. Software

All analyses in this thesis are conducted using the statistical programming software `R` (R Core Team, 2020), version 4.0.0. For curve-fitting and the calculation of p-values based on the MCP-Mod approach, the packages `drc` (Ritz et al., 2015) and `DoseFinding` (Bornkamp, 2019) are used. The Dunnett procedure is conducted using the function `glht` from the `multcomp`-package (Hothorn et al., 2008). GO groups are calculated using the package `topGO` (Alexa and Rahnenführer, 2020), and meta-analyses are conducted using the package `metafor` (Viechtbauer, 2010). Plots are created using basic `R`-functions and the package `ggplot2` (Wickham, 2016), making use of the package `gridExtra` (Auguie, 2017).

5. Handling deviating control values

In the context of deviating control values for cytotoxicity data, first the results from a literature review aiming at giving an overview of the extent of the problem in real published data are presented in Chapter 5.1. Four approaches for handling deviating controls, called **4pLL**, **3pLL**, **No Ctrl**, and **BC**, are introduced in Chapter 4.2. These methods are compared in a controlled simulation study that is based on several scenarios regarding the choice of concentrations considered. Additionally, different standard deviations of the replicates and deviations of the controls are assumed. The setup of the simulation study is presented in Chapter 5.2 and results from the simulation study are shown in Chapter 5.3. Specific recommendations, which method to use in which situation, are derived from the results and explicitly stated in Chapter 5.4. Finally, the four methods compared in the simulation study are applied to a real cytotoxicity dataset. The results are shown in Chapter 5.5.

Most results for the literature review, the simulation study and the real data study presented here are published in Kappenberg et al. (2020). Especially the recommendations are clearly expressed there. Results from Kappenberg et al. (2020) are extended by several aspects in this work: The literature review is analysed in more detail. In the simulation studies, additional simulation scenarios and an additional alert concentration are considered. Furthermore, an additional real cytotoxicity dataset is evaluated, and all analyses regarding real datasets are performed in more detail.

Many of the plots shown in this chapter are also already published in Kappenberg et al. (2020) and are only slightly adjusted regarding notation or the division into different figures. The Figures already published in the same or a similar form are: Figures 5.1 to 5.10, 5.12, B.5 to B.7, B.20 to B.22, B.32, and B.33.

5.1. Literature review

A literature review was conducted to investigate the frequency and the extent of the problem of deviating controls[2]. Three leading international toxicological journals, namely 'Archives of Toxicology' (ArchTox, all issues from 2016 to 2018), 'Toxicological Sciences' (ToxSci, all issues from 2017 to 2018) and 'Toxicology in Vitro' (ToxVitro, all issues from 2015 to 2017) were chosen as the basis of the review. The goal of the review was to answer the following two questions:

- How often does the problem of deviating controls occur?
- How strong are the deviations in these cases?

Additionally, further information about published and modelled concentration-response data was collected, including information about the number of concentrations, the order of magnitude of the standard deviations for each concentration, and about the chosen

[2]The literature review was conducted by W. Albrecht, T. Brecklinghaus and J. Blum, co-authors of Kappenberg et al. (2020).

models fitted to the data. The research was restricted to viability assays, where viability was defined in a broad sense including (mostly mitochondrial) activity, motility, contraction, or mitotic activity.

A set of necessary criteria was defined for curves resulting from these assays to be included in the analysis set. These criteria are:

- A concentration-response model is fitted to the data.
- Measurements for a negative control are available.
- Measurements for at least four positive concentrations (i.e. additional to the control) are available.
- When neglecting the control, the response values are monotonously decreasing with increasing concentration.
- For at least two of the concentrations other than control, no effect can be observed, i.e. the difference between the corresponding response values is smaller than 10% of the response value for the lowest concentration.
- For every concentration, at least three replicate values are available, regardless of whether these are technical or biological replicates.

For curves fulfilling these criteria, the average values of the controls and the value of an upper asymptote when omitting the controls were both looked up in the corresponding publication or estimated from the plots. Deviation of the controls was calculated in percent based on a value of 100% for the asymptote. Let `Control` be the average of all individual control values and `Fit` be the value of the upper asymptote for very small concentrations when omitting the controls. Then the deviation Δ is calculated by

$$\Delta = \frac{\texttt{Control} - \texttt{Fit}}{\texttt{Fit}} \cdot 100. \tag{20}$$

Additional information collected for each curve includes the response values and the standard deviation of the replicates for each concentration, as well as the number of replicates and the type of model fitted to the data. The standard deviation was determined by estimating the values from the plotted data. For response values measured in percent values, if a standard deviation is smaller than 1, the value is set to 1. An average standard deviation for an entire curve is calculated as the median of the standard deviations for all concentrations except the control and is denoted by $\hat{\sigma}_{\text{med}}$. If a standard error is plotted instead of a standard deviation, the standard deviation is calculated by multiplying the standard error with the square root of the number of replicates. For some curves, it is not explicitly stated in the paper whether standard deviations or standard errors are used as measure of dispersion. These curves are still analysed regarding the deviation of the controls, but omitted from analyses regarding the standard deviation of the concentration-wise response values.

In total, 2199 papers were reviewed. Table 5.1 summarises key figures of the literature review: The total number of papers published in the respective timespan per journal is

indicated as well as the number of papers in which modelled curves of concentration-response data are presented. This is followed by the number of papers with at least one curve fulfilling the criteria. The total number of curves in the respective journal and the number of curves fulfilling the criteria are indicated as well as the number of curves for which it is explicitly stated whether the standard deviation or the standard error is plotted in the respective figure.

Even when adjusting for the shorter time period considered, the fewest papers with curves were found in ToxSci and the most in ToxVitro. The most curves in total were observed for ArchTox with 440 curves fulfilling the criteria, but only 266 come with an explicit statement about the measure of dispersion. The number of curves per paper is in the range from 1 to 204, with 8 papers containing only one curve. The median value of curves per paper fulfilling all criteria is 6 and the mean value is 17.04, with standard deviation 33.31. The three papers with the most curves fulfilling all criteria contain 204, 91, and 57 such curves.

All curves considered have in common that the response values of the controls are normalised to correspond to a response of 100%. The only exception is given by curves from Gu et al. (2018, ArchTox), where the **4pLL** method as explained in Chapter 4.2 was applied in the curve fitting procedure. Thus, the values of the asymptote are guaranteed to attain 100% and the controls are normalised accordingly.

Results from the literature review do not allow a precise analysis regarding the choice of models in published literature. From observation of the plotted curves, it becomes clear that models yielding non-linear sigmoidal curves are the most popular choice. However, in many publications a clear statement regarding the chosen model is missing, and only the software used for model fitting is stated. Often, this software is GraphPad Prism (GraphPad Software) in various versions, which allows fitting of several different models. Among the clear statements regarding the choice of models, the family of log-logistic models occurs most frequently. Specifically, the 4pLL model or a constraint version, in

Table 5.1: Key figures of the literature review summarising the total number of papers in the respective timespans for the three journals considered. Additionally, the number of papers and curves fulfilling different criteria are stated.

	ArchTox (2016 - 2018)	ToxSci (2017 - 2018)	ToxVitro (2015 - 2017)
Total number of papers	810	592	797
Number of papers with curves	31	7	37
Number of papers with at least one curve fulfilling the criteria	15	6	26
Number of curves	702	65	345
Fulfilling the criteria	440	56	213
Number of curves with indicated measure of dispersion	266	56	202

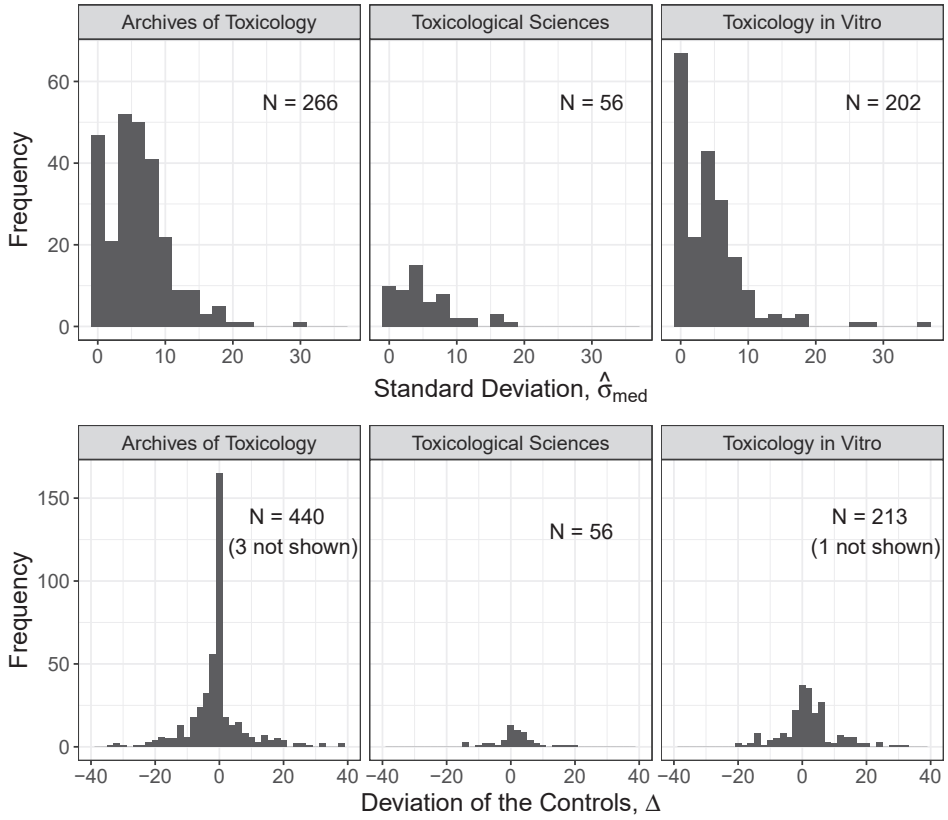

Figure 5.1: Histograms of the estimated standard deviation $\hat{\sigma}_{\text{med}}$ (top) and the estimated deviation of the controls Δ (bottom) in the literature review.

which the upper or both the upper and the lower asymptote are fixed to 100% and 0%, respectively, are often chosen.

Only curves with at least four positive concentrations in addition to the control are considered for analysis. The maximum number of concentrations observed is 18. Out of all 709 considered curves, only for 30 curves 10 or more concentrations are measured. Most often (for 187 curves), 6 concentrations are considered, followed by $5, 8$ and 7 concentrations with 151, 149 and 137 curves, respectively.

The observed values of the deviation Δ and the standard deviation $\hat{\sigma}_{\text{med}}$ are summarised for each journal individually by histograms in Figure 5.1. The standard deviation is in the range between 0 and 10 for 85% (ArchTox), 88% (ToxSci) and 91% (ToxVitro) of the curves and between 0 and 20 for 99%, 100% and 99% of the curves, respectively.

A deviation of the control is considered to be essentially negligible if $|\Delta| \leq 2$. This occurs only for 47% (ArchTox), 38% (ToxSci) and 31% (ToxVitro) of the curves. For 80%, 88% and 79% of the curves, respectively, it holds that $|\Delta| \leq 10$. This observation is also used for the design of the simulation study (Chapter 5.2), in which the considered deviation of the controls varies in this range. Except for four curves (three in ArchTox, one in ToxVitro), for all values of the deviation Δ it holds that $|\Delta| \leq 40$.

The direction of the deviation of the controls differs across the three journals. A deviation is defined to be negative if $\Delta < -2$ and positive if $\Delta > 2$. For ArchTox, negative deviations occur for 34% and positive deviations for 19% of the curves. For ToxSci, negative deviations occur for 21% and positive deviations for 41% of the curves, and for ToxVitro, negative deviations occur for 23% and positive deviations for 46% of the curves.

All in all, results from the literature review show that the problem of deviating control values occurs in a notable number of curves. Positively and negatively deviating controls are observed about equally often, and values of the deviations as calculated in formula (20) are smaller than 10 in approximately 80% of cases. Median values of the concentration-wise standard deviation of the replicates are mostly smaller than 10.

5.2. Setup of the simulation study

A controlled simulation study is conducted to compare the four methods introduced in Chapter 4.2 in different scenarios where deviating controls occur. The goal of the methods is to yield the best possible estimate of the effective concentrations EC_{10}, EC_{20} and EC_{50} that are of interest in many toxicological applications. Therefore, for each simulated curve, the four methods are applied, the EC values are calculated, and they are compared with the known true EC value of the underlying curve used for simulation.

The shape of the true concentration-response curve is based on the real dataset introduced in Chapter 3.2.1. A 4pLL curve is chosen as underlying models with parameters $\phi^{(b)} = 1.462$, $\phi^{(c)} = 0$, $\phi^{(d)} = 100$, and $\phi^{(e)} = 4.22$. Since the upper and lower asymptotes take the values 100% and 0%, respectively, the value of the parameter $\phi^{(e)}$ coincides with the EC_{50}. The corresponding curve is displayed in Figure 5.2. The EC_{10}, EC_{20} and EC_{50} are indicated in this curve and take the values $0.94, 1.63$ and 4.22, respectively.

Different scenarios regarding the choice of concentrations, where the viability is measured, are considered. The three main scenarios considered consist of 5 concentrations and the concentration 0 as control value each, with 3 replicates per concentration. These scenarios are subsequently labelled ‘easy’, ‘medium’ and ‘difficult’ and are visualised in Figure 5.3. Main properties of these scenarios are:

‘Easy’: The concentration values cover the entire range of the curve. Especially the upper asymptote is well-covered by concentrations, with two concentrations in the range of no or low toxicity. One concentration corresponds to a viability of approximately 60% and the two highest concentration are in a range of high toxicity.

‘Medium’: Even for the lowest measured positive concentration, the viability has already dropped by 10%, so no concentration in the range of no toxicity exists. Two concentrations are in the middle range of the curve and two concentrations in a range of high toxicity, corresponding to viabilities smaller than 10%.

‘Difficult’: The responses for all five concentration values are in the range between 80% and 10% viability, such that neither the upper nor the lower asymptote are covered.

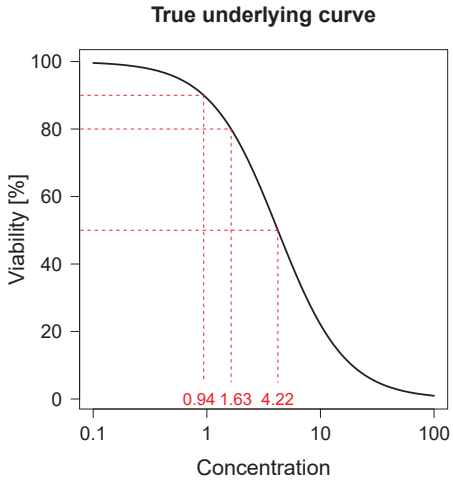

Figure 5.2: True underlying 4pLL model of the simulation study with indicated values of EC_{10}, EC_{20} and EC_{50}.

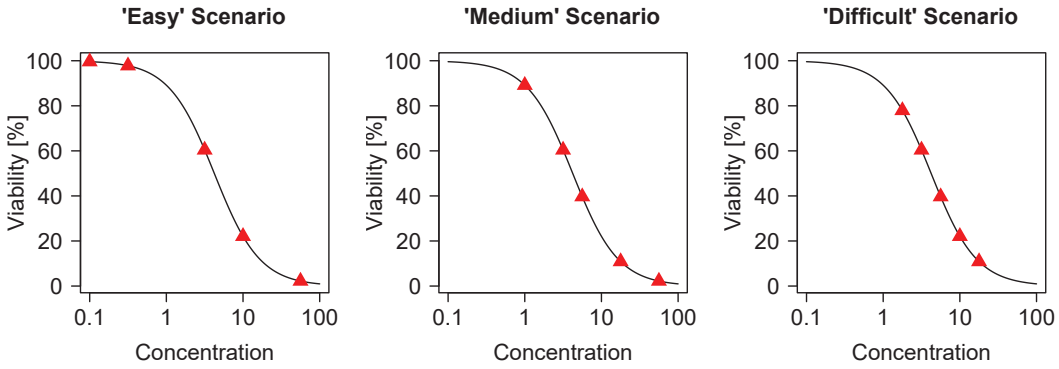

Figure 5.3: The three main scenarios 'easy', 'medium' and 'difficult' for the simulation study. The red triangles indicate the concentrations where the viability is measured, together with the corresponding response value based on the true underlying curve.

In addition to these scenarios, three scenarios consisting of more or less than five concentrations are included in the analysis. The scenario with the most concentrations consists of 12 equidistant concentrations (on log-scale) that cover the entire range of the curve. The second scenario consists of seven concentrations, three of which are in the no- or low-toxicity range of the curve, two in the middle range and two in the high-toxicity range. The last scenario considered consists only of four concentrations, two in the no-toxicity range and two in the medium range, such that the range of high toxicity is not covered. These scenarios are shown in Figure B.1 in Appendix B.1.

For the simulation of datasets, three replicates for each of the concentrations in the respective scenario are independently drawn from a normal distribution with mean $\mu = 0$.

An equal value of the standard deviation σ is chosen across all concentrations, with four different values of σ, specifically $\sigma \in \{2, 4, 8, 12\}$, considered. A deviation Δ is added to the three response values of the control, with $\Delta \in \{-10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10\}$.

This leads to 44 parameter combinations per scenario. For each parameter combination and each scenario, 5000 datasets are simulated. Four concentration-response curves are fitted to each dataset, using each of the four methods **4pLL**, **3pLL**, **No Ctrl** and **BC**. The EC values EC_{10}, EC_{20}, and EC_{50} are estimated from these fitted curves and compared to the known true EC values of the underlying curve.

5.3. Results from the simulation study

The simulation study is analysed in two different ways: First, the proportions of estimated EC values that are in an *acceptable range* around the true EC value are calculated and compared across the methods. The specific definition of the acceptable range depends on the EC value considered and is explained in more detail below. Second, the method with the smallest absolute difference between estimated and true EC value is determined. This method is subsequently referred to as the 'winner'. Determination of the winner method is restricted to those iterations of the simulation where at least one method leads to an acceptable result.

Both analyses thus require a definition of the acceptable range. An estimate of the respective EC value is considered to be acceptable, if it does not differ from the true EC value by more than a pre-specified, fixed factor. This factor defines an interval around the true EC value which is the acceptable range. Note that a factor is used for defining the acceptable range instead of an additive term as the concentrations are considered on log-scale and in this way, the acceptable range is a symmetric interval (again on log-scale) around the true value.

The choice of the factor differs for the three different EC values. The EC_{10} and EC_{20} are much more influenced by a potential deviation of the controls and by the design of the corresponding scenario than the EC_{50}. Therefore, better results are generally expected for the EC_{50}. To retain comparability between the three EC values and to avoid a perfect proportion of acceptable results for the EC_{50} in all cases considered, a smaller factor and resulting from that a narrower acceptable range is chosen for the EC_{50} than for EC_{10} and EC_{20}.

The factors are chosen in a way that ensures about 3000 acceptable results even in the 'difficult' scenario for the largest value of the standard deviation, $\sigma = 12$, for very small deviations, i.e. $|\Delta| \leq 2$. Specifically, a factor of 1.3 is chosen for EC_{10} and EC_{20}, leading to acceptable ranges of $[0.72, 1.22]$ and $[1.25, 2.12]$ around the true values $EC_{10} = 0.94$ and $EC_{20} = 1.63$. For the EC_{50}, a smaller factor of 1.1 is chosen, leading to an acceptable range of $[3.84, 4.64]$ around the true value $EC_{50} = 4.22$.

Proportions of acceptable estimates for the EC_{20} in the three scenarios 'easy', 'medium' and 'difficult' are shown in Figures 5.4, 5.5 and 5.6. Corresponding plots for EC_{10} and EC_{20} and for the additional three scenarios are shown in Figures B.2 to B.16 in Appendix B.1. Each cell of one plot corresponds to one combination of the parameters

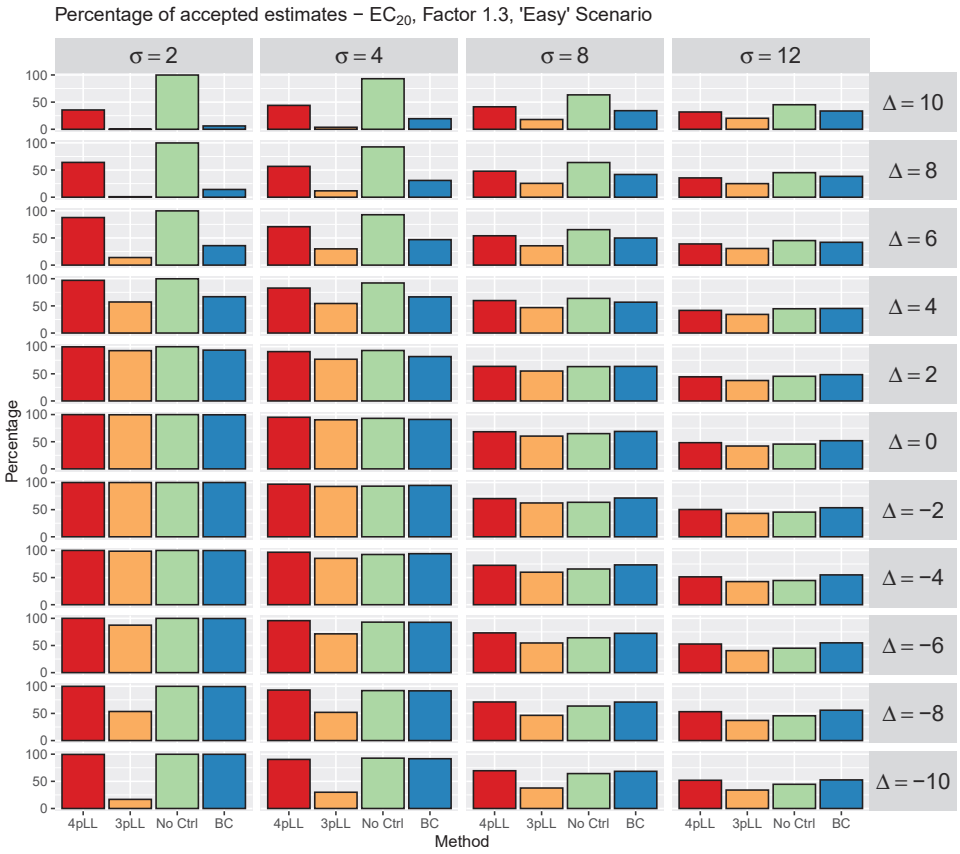

Figure 5.4: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{20} in the 'easy' scenario. Columns correspond to the different standard deviations σ and rows to the deviations of the controls Δ. Each cell corresponds to one combination of the simulation parameters σ and Δ and shows, from left to right, the results for **4pLL**, **3pLL**, **No Ctrl** and **BC**. The factor defining the acceptable range is chosen as 1.3.

σ and Δ, with the standard deviation in the columns, increasing from left to right and the deviation of the controls in the rows, with decreasing values from top to bottom and no deviation in the middle row. The bars indicate the percentage of acceptable estimate for the four methods, from left to right **4pLL**, **3pLL**, **No Ctrl**, **BC** for the 5000 iterations of the simulation per cell.

Results are first described for the EC_{20} for the three main scenarios as shown in Figures 5.4, 5.5 and 5.6. Then, EC_{10} and EC_{50} are considered and finally description of the results is broadened to the three further scenarios.

In the 'easy' scenario (Figure 5.4), in the range of essentially negligible deviations, i.e. $|\Delta| \leq 2$, all methods perform similar. Overall, the methods **4pLL** and **No Ctrl** achieve the highest proportions of acceptable estimates. Especially for large $\Delta > 6$, **No Ctrl** also outperforms **4pLL**. For not or negatively deviating controls, i.e. $\Delta \leq 2$, results for

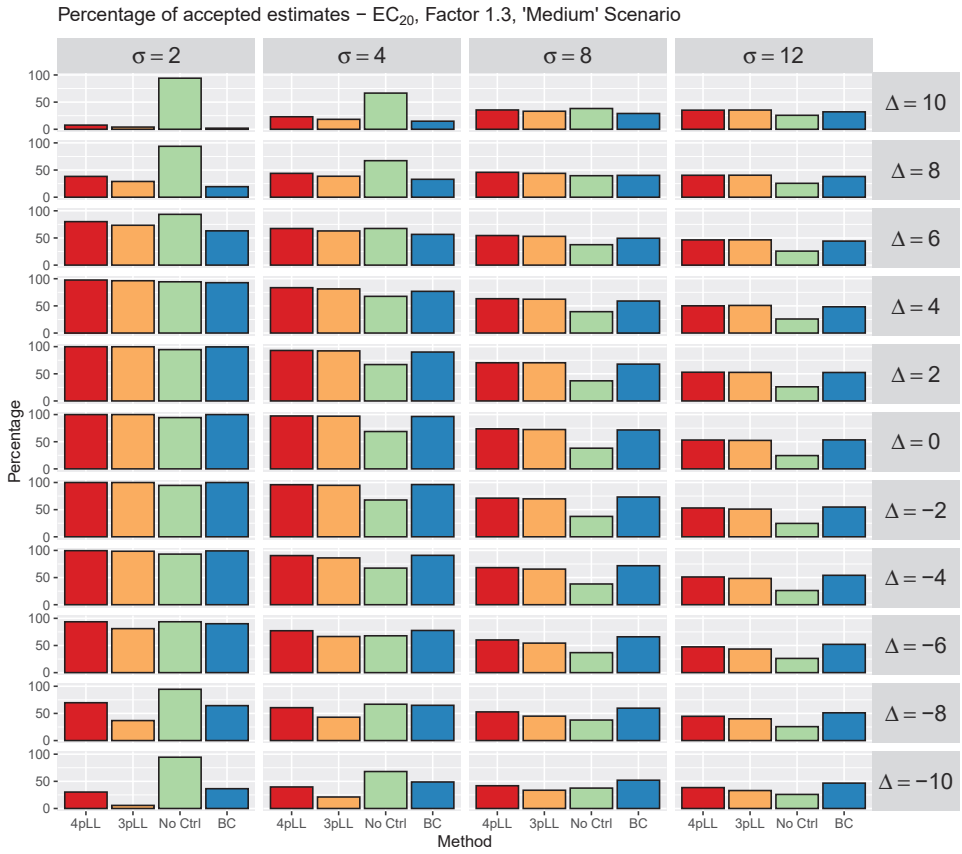

Figure 5.5: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{20} in the 'medium' scenario and are structured as explained in Figure 5.4.

BC are competitive with those of **4pLL** and **No Ctrl**. The method **3pLL** performs clearly worse than all of the other methods, especially for large absolute values of Δ, where only a very small percentage of estimates is acceptable. Results for **No Ctrl** are not influenced by different values of Δ, since responses for control values are omitted for this method. For increasing values of the standard deviation σ, the overall percentage of acceptable estimates decreases. While for $\sigma = 2$, about 100% of the estimates for **No Ctrl** are acceptable, for $\sigma = 12$ this percentage becomes less than 50%, with comparable results for the other methods.

Results for the 'medium' scenario (Figure 5.5) show a stronger influence of the standard deviation σ on the percentage of acceptable results for **No Ctrl**. While **No Ctrl** leads to acceptable estimates in almost all iterations for $\sigma = 2$, this percentage strongly decreases as σ increases, and for $\sigma \geq 8$, **No Ctrl** only outperforms the other methods in the one combination of parameters $\sigma = 8$ and $\Delta = 10$. For $|\Delta| \leq 6$, the three methods **4pLL**, **3pLL** and **BC** perform similar and for $\sigma \geq 4$ in this range of Δ, they outperform **No Ctrl**. For small $\sigma \leq 4$ and large deviations with $|\Delta| \geq 8$, **No Ctrl** still outperforms the other methods.

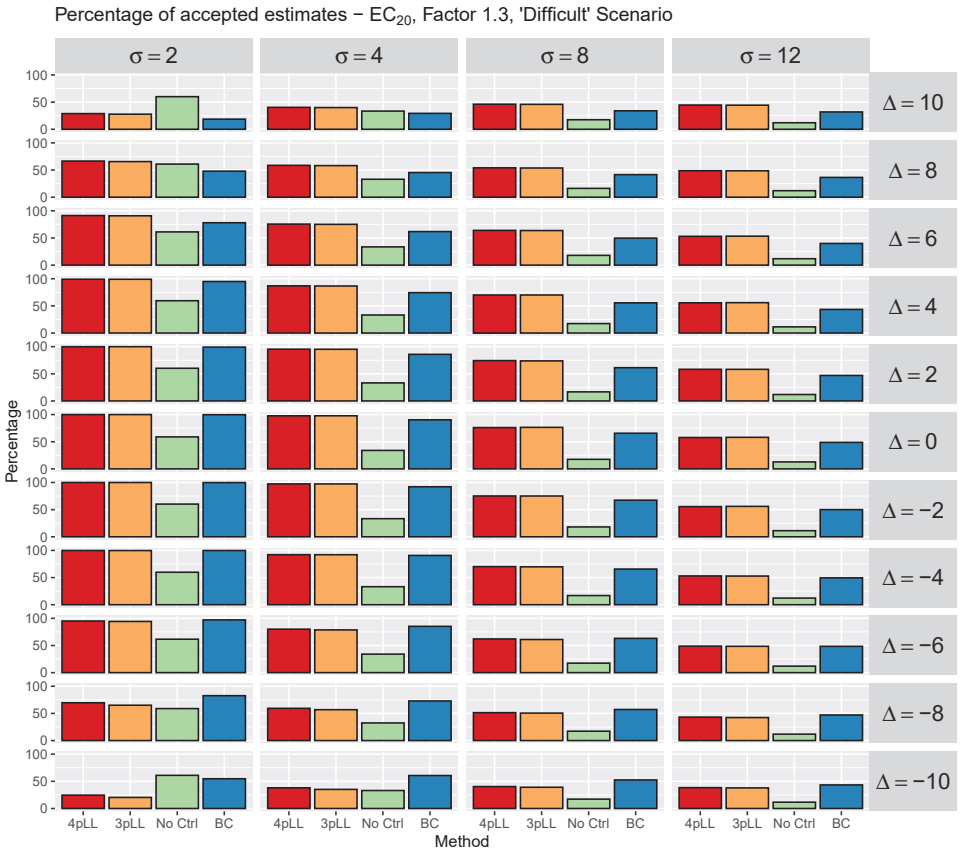

Figure 5.6: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{20} in the 'difficult' scenario and are structured as explained in Figure 5.4.

In the 'difficult' scenario (Figure 5.6), the results are in general very similar to those from the 'medium' scenario, with an overall worse performance of **No Ctrl**, which only leads to acceptable results most often for $\sigma = 2$ and $\Delta = 10$ or $\Delta = -10$. The performance of **3pLL**, **4pLL** and **BC** is basically the same for $|\Delta| \leq 6$ across all values of σ. For large $\Delta > 6$, however, **BC** performs slightly worse and for strongly negatively deviating controls with $\Delta < -6$, it leads to acceptable results more often than **4pLL** and **3pLL**.

When considering the EC_{10} instead of EC_{20} (Figures B.2, B.3, B.4), the most striking difference is the performance of **BC**, which is much worse than before in all three of the main scenarios, especially for small values of σ. Overall, the percentage of acceptable estimates is lower for EC_{10} than for EC_{20} in all three scenarios, but the comparability of the methods **4pLL**, **3pLL** and **No Ctrl** remains as described above.

The results for the EC_{50} are shown in Figures B.5, B.6 and B.7. It can be observed that overall, the percentage of acceptable estimates slightly decreases in comparison to the results for EC_{10} and EC_{20}. Due to the different factors defining the acceptable range, this observation is not meaningful. It is more relevant to note that for the comparison

of the methods, similar conclusions can be drawn as in the previous cases. The only exception is **BC** in the 'easy' scenario, which leads to an acceptable result more often also for $\Delta > 2$, although **BC** is designed to mainly handle negatively deviating controls.

Results for the three further scenarios introduced in Figure B.1 are only briefly reported here. In the scenario with 12 concentrations (Figures B.8, B.11, B.14), the main observations for all three alert concentrations are that **4pLL** and **No Ctrl** perform similarly well with high percentages of acceptable estimates. **3pLL** leads to acceptable estimates notably often only for $|\Delta| \leq 2$. The only difference between the alert concentration is again formed by the method **BC**, which leads to clearly worse results for the EC_{10} in this scenario but is performing similarly to the other methods for EC_{20} and EC_{50}.

The same basic observations can be made for the scenario with 7 concentrations (Figures B.9, B.12, B.15), including the results for **BC**. In some way surprising results are observed for the scenario with only 4 concentrations (Figures B.10, B.13, B.16). Again, **No Ctrl** is by definition not affected by the deviations of the controls and, at least for $\sigma \leq 4$, leads to acceptable results considerably often. While for EC_{10} and EC_{20}, for $\Delta > 2$, the three other methods never or only very seldom result in an acceptable estimate, **BC** performs extremely well for negative values of Δ. In particular, acceptable results are observed more often than in the 'easy' scenario, respectively. For EC_{50}, this result can also be observed, but for this alert concentration, all methods perform well for positive values of Δ.

As the individual fitted curves are not assessed, no clear explanation of this effect regarding **BC** can be given. However, in the real data application (Chapter 5.5) it can be observed that this method sometimes leads to biologically implausible results. One possible explanation for the results obtained here is that the missing concentration in the high-toxicity range in comparison to the 'easy' scenario allows more flexible, but biologically implausible modelling of the Brain-Cousens curve, which actually results in a better estimate of the respective effective concentration. In any case, fitting a Brain-Cousens curve, comprising 5 parameters, to a dataset consisting of only four concentrations plus a control is numerically difficult and therefore not recommended. Thus, these results need to interpreted with immense caution and should not lead to the conclusion that **BC** actually performs best in these cases.

For the second analysis, the number of winners for the EC_{20} in the 'easy', 'medium' and 'difficult' scenario are shown in Figures 5.7, 5.8, and 5.9. Corresponding plots for EC_{10} and EC_{50} and for the additional three scenarios are shown in Figures B.17 to B.31 in Appendix B.1. The 'winner' is the method yielding the smallest absolute difference between estimated and true EC value, with the difference calculated on log-scale. Again, each cell corresponds to one combination of σ and Δ, sorted in the same way as explained above. Additionally, the number in each cell states the number of iterations in which at least one method (i.e. at least the winner) leads to an acceptable estimate of the true EC value. Only these iterations are considered for determining a winner.

As for the analysis regarding the percentage of acceptable estimates, results are first described for the EC_{20} for the three main scenarios as shown in Figures 5.7, 5.8 and 5.9.

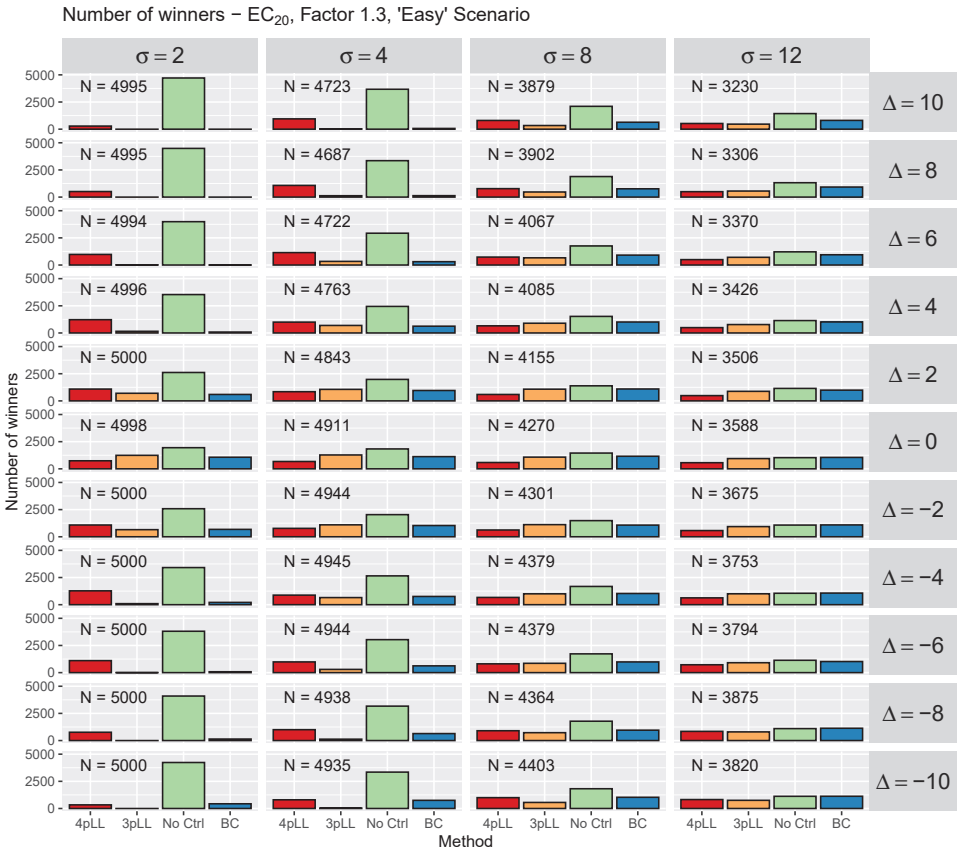

Figure 5.7: Number of times each method is the winner, i.e. leads to the smallest absolute difference between true and estimated EC value. Results are shown here for the EC_{20} in the 'easy' scenario. Columns correspond to the different standard deviations σ and rows to the deviations of the controls Δ. Each cell corresponds to one combination of the simulation parameters σ and Δ and shows, from left to right, the results for **4pLL**, **3pLL**, **No Ctrl** and **BC**. The number in each cell indicates the number of simulation iterations where at least one method yields an acceptable result, with the factor defining such a result chosen as 1.3.

Then, EC_{10} and EC_{50} are considered, and finally the description of the results is extended to the three further scenarios.

In the 'easy' scenario (Figure 5.7), it can clearly be seen that **No Ctrl** is the winning method most often, especially for small $\sigma \leq 4$. But also for $\sigma > 4$ and large deviations with $|\Delta| > 6$, **No Ctrl** again is the winner notably more often than the other methods. The number of iterations in which the respective methods are the winner are more equally distributed for $\sigma > 4$ and $|\Delta| \leq 6$. A remarkable observation is that **No Ctrl** is the best method even for $|\Delta| \leq 2$, where the controls do not deviate and are therefore expected to help in obtaining a good fit. For $\sigma = 2$, **4pLL** also leads to the best estimate considerably often, while **3pLL** and **BC** fail more often. **BC** is only competitive in comparison to **No Ctrl** for $\sigma = 12$ and $\Delta < 6$.

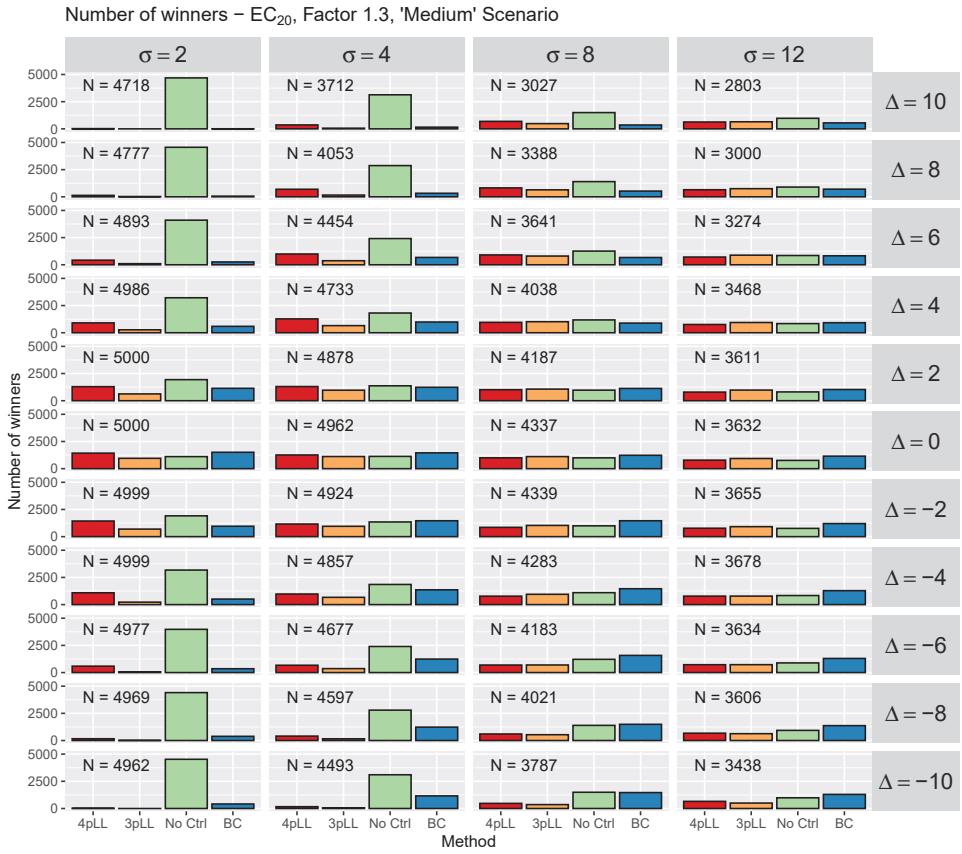

Figure 5.8: Number of times each method is the winner, i.e. leads to the smallest absolute difference between true and estimated EC value. Results are shown here for the EC_{20} in the 'medium' scenario and are structured as explained in Figure 5.7.

Results for the 'medium' scenario (Figure 5.8) are very similar to the 'easy' scenario. Differences can be seen in the performance of the method **BC**, which is the winning method here most often for $\sigma \geq 8$ and $\Delta \leq 0$. The method **3pLL** performs worst again, and for small $|\Delta| \leq 4$, **4pLL** is also competitive. In particular, for $\sigma = 2$ and $\Delta = 0$, in contrast to the 'easy' scenario, no longer **No Ctrl** but **4pLL** and **BC** are the winning methods most often.

In the 'difficult' scenario (Figure 5.9), results are similar for $\sigma \geq 8$ across all values of Δ: **No Ctrl** only very rarely is the best method and overall, the other three methods are the winner similarly often, with **BC** being slightly better than **4pLL** and **3pLL** for $\Delta < 0$. Only for $\sigma \leq 4$, $\Delta \geq 6$ and for $\sigma = 2$, $\Delta \leq -8$, **No Ctrl** remains the best method. For moderate values of Δ, **4pLL** is competitive to **BC**, and **BC** clearly dominates the other methods in the other cases.

When considering the EC_{10} instead of the EC_{20}, in all three main scenarios very similar results can be observed (Figures B.17, B.18, B.19). Only a slight difference can be seen with regard to **4pLL**, which performs slightly better in comparison to **BC** in the cases

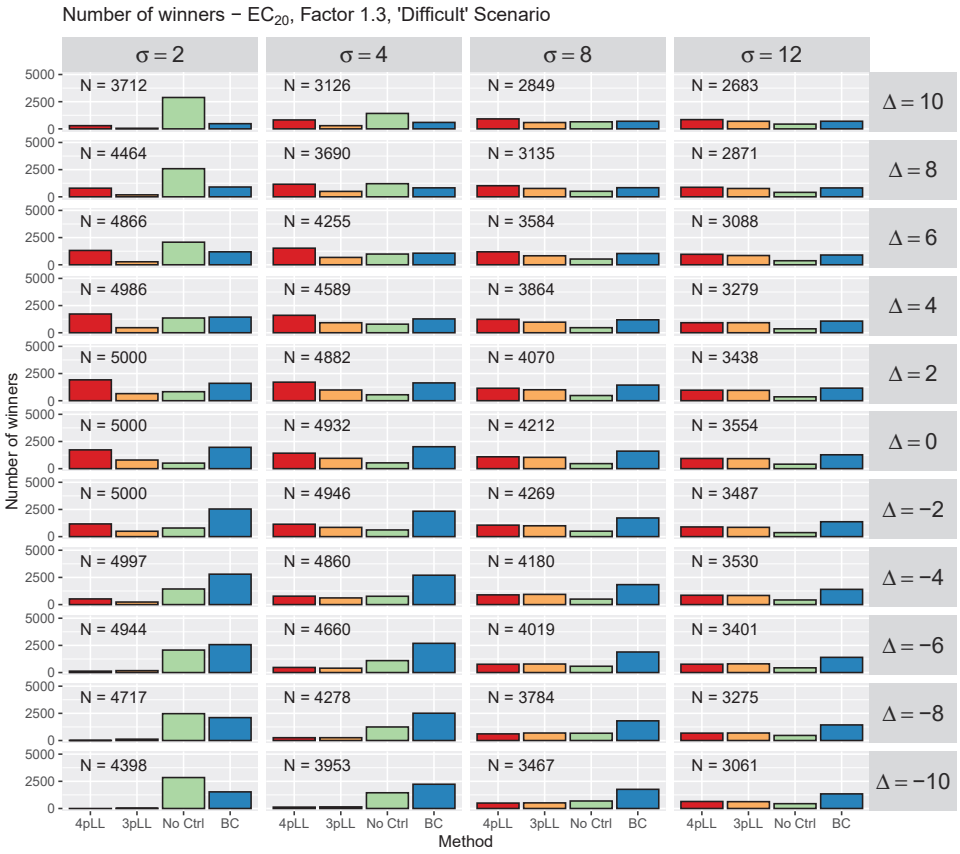

Figure 5.9: Number of times each method is the winner, i.e. leads to the smallest absolute difference between true and estimated EC value. Results are shown here for the EC_{20} in the 'difficult' scenario and are structured as explained in Figure 5.7.

with negatively deviating controls. For the EC_{50}, an increase in the number of times that **3pLL** is the winner is noticeable (Figures B.20, B.21, B.22). Especially in the 'difficult' and in the 'medium' scenario for $|\Delta| \leq 2$, **3pLL** is the winner most often, followed by **No Ctrl** (mainly in the 'medium' scenario) and **BC** (mainly in the 'difficult' scenario). In comparison to results for EC_{10} and EC_{20}, **4pLL** is the winner only very rarely.

In the scenarios with 12 concentrations (Figures B.23, B.26, B.29) and with 7 concentrations (Figures B.24, B.27, B.30), similar results with **No Ctrl** as the winning method most often for EC_{10} and EC_{20} can be observed. These methods are followed by **4pLL** for $\sigma \leq 4$ and **3pLL** for $\sigma \geq 8$. For the EC_{50}, **No Ctrl** less obviously dominates the other methods, with **BC** as winning method for $\Delta < 0$ more often than for EC_{10} and EC_{20}. Still, **BC** is the winning method less often than **No Ctrl**. For $|\Delta| \leq 2$, **3pLL** is competitive to **No Ctrl** and it is the winner most often for these Δ and $\sigma \geq 6$.

A differentiation between results for EC_{10} and EC_{20} and results for EC_{50} is also necessary for the scenario with 4 concentrations (Figures B.25, B.28, B.31). For EC_{10} and EC_{20}, clearly **BC** leads to the best results for $\Delta < 0$ and **No Ctrl** to the best results for

$\Delta > 2$. For the EC_{50}, as observed before in other scenarios, **3pLL** performs similarly well as **No Ctrl** for $0 \leq \Delta \leq 6$. For $\Delta < 0$, **No Ctrl** is the winner more often than for EC_{10} and EC_{20}. For $\sigma = 2$ it even is the winner most often, followed by **BC**.

The analyses with regard to the number of winners have one main point of criticism: Consider the case in which a first method only slightly dominates a second method in most of the cases, but this second method is far better than the first one in the remaining minority of cases. Using the second method would then be preferred over using the first one, but the analysis only considering the number of winners would suggest otherwise. Therefore, a cautious interpretation of this analysis is required and recommendations, which method to use in which case, should mainly follow the results from the analysis regarding the percentage of acceptable estimates.

5.4. Recommendations

Main results from the simulation study are summarised here. Most emphasis is put on the three main scenarios 'easy', 'medium', and 'difficult' that are chosen to represent frequently occurring scenarios in real-data studies. Based on these results, a set of recommendations is given, which of the four methods **4pLL**, **3pLL**, **No Ctrl** and **BC** should be used in which scenario. These recommendations are explicitly stated in an algorithmic procedure. General results summarising the performance of the four methods for the other scenarios are briefly stated as well.

- In the 'easy' scenario, **No Ctrl** performs best, both when considering the percentage of acceptable estimates and the number of winners. The method **4pLL** is competitive when considering acceptable estimates, especially for smaller values of the deviation Δ. However, for small standard deviations and large values of Δ, **No Ctrl** is clearly better.
- In the 'medium' scenario, a more strict distinction between the analysis regarding the number of acceptable results and the analysis regarding the number of winners is required: While the latter indicates a good performance of **No Ctrl**, the former shows that especially for larger values of σ, **No Ctrl** leads to acceptable results least often. Only for EC_{50} and $|\Delta| \leq 4$, **3pLL** is competitive with respect to the number of winners.
- In the 'difficult' scenario, for moderate values of the deviation, **4pLL** is competitive. Only for large positive values of Δ, **No Ctrl** leads to better results, while for large negative values of Δ, **BC** performs better. Again only when considering the EC_{50}, for moderate values of Δ, **3pLL** is also competitive.
- In scenarios with more than five concentrations, all methods except **3pLL** perform similarly well with respect to the percentage of acceptable estimates. When considering the number of winners, **No Ctrl** performs best.
- In the scenario with only four concentrations, for negative deviations of the controls, **BC** seems to lead to the best results, and for positive deviations, **No Ctrl** performs best.

A problem with giving concrete guidelines based on these simulation results, that work with specific values of the standard deviation σ and the deviation of the controls Δ, is that the true scenario is unknown in real-data situations. Especially in situations where no high-quality fit of the upper asymptote is possible using only the concentrations that are not the control, a reliable estimation of Δ is not straightforward possible. Estimation of the concentration-wise standard deviation, however, is easily possible.

Only approximate methods for assigning a real dataset to one of the three main scenarios can be given. In the 'easy' scenario, two no-effect concentrations are available, based on which a determination of the value of the upper asymptote without the controls and an estimation of the deviation are possible. This is no longer the case in the 'medium' and the 'difficult' scenario. These scenarios can be distinguished from another by the fact that the two highest measured concentrations correspond to a high toxicity in the 'medium' scenario, while still a clear decrease in viability for the highest two measured concentrations is observable in the 'difficult' scenario. In the 'difficult' scenario, an observation of response values for the controls that are only slightly larger than response values of the lowest measured concentrations is an indicator for negatively deviating control values.

All recommendations regarding the choice of which method to use are based on the assumption that the true concentration-response relationship can be described by a monotonously decreasing sigmoidal-shaped curve. In general for these cases, **4pLL** works well with respect to the estimation of all three examined EC values. If the response values of the low concentrations allow a high-quality fit of the asymptote, the replicates have low variances and the controls are (strongly) deviating, then **No Ctrl** clearly leads to good results. If the concentrations measured cover only a medium range of the response and values are missing both in the range of low and of high toxicity, **4pLL** and **BC** lead to the best results and should therefore both be considered. Strictly speaking, in these cases, the assay should be repeated with more appropriate concentrations of the compound of interest, covering the entire range of viability. However, if this is not possible, the mentioned methods still allow for the best possible estimation of the respective EC value. In these cases, a plausibility check for **BC** is required, as this method may lead to biologically implausible results in the case of positively deviating controls.

Although it is the best method considerably often for the EC_{50} in the three main scenarios, the use of the method **3pLL** is strongly discouraged. It performs much worse when considering the percentage of acceptable estimates for EC_{10} and EC_{20} in comparison to the other three methods. In the scenarios where **3pLL** is the best method most often, **4pLL** and **BC** lead to acceptable results equally often.

In Kappenberg et al. (2020), an algorithmic procedure is proposed as practical guideline which method to use in which case. The procedure is based on the three scenarios 'easy', 'medium' and 'difficult' as analysed in the simulation study, and it relies on the same assumptions regarding the shape of the underlying relationship as stated above. Figure 5.10 summarises this algorithmic procedure.

- Normalise all response values with respect to the mean response values of the controls.
- Estimate the standard deviation σ by $\hat{\sigma}_{\text{med}}$, which is the median of the estimated standard deviations for all concentrations.
- Estimate the relative difference d of the response values between the two lowest concentrations (e.g. if the lowest tested concentration gives a value of 98% and the second lowest concentration gives a value of 88%, then d would be $100 - \frac{88}{98} \times 100 = 10.2\%$).
- Use the **4pLL** model, with the followings three exceptions:

  1. EASY case: When there are at least two concentrations in the no-effect range (identified by a small value of d, e.g. $d \leq 5$):
     (a) Estimate the deviation of the controls Δ: If `Control` represents the average response values of all individual controls and `Asymp` represents the mean of the response values observed at the lowest two concentrations, then the deviation is $\Delta =$ `(Control - Asymp)/Asymp` $* 100$.
     (b) If $\hat{\sigma}_{\text{med}}$ ist small, e.g. $\hat{\sigma}_{\text{med}} \leq 8$, and Δ ist large, e.g. $\Delta \geq 6$, use the method **No Ctrl**.
  2. MEDIUM case: When there are less than two concentrations in the no-effect range (identified by a large value of d, e.g. $d > 5$), but there are two concentrations in the high-toxicity range (identified by small response values, e.g. below 12%), and $\hat{\sigma}_{\text{med}}$ is very small, e.g. $\hat{\sigma}_{\text{med}} \leq 2$, then use **No Ctrl**.
  3. DIFFICULT case: When there are less than two concentrations in the no-effect range (identified by a large value of d, e.g. $d > 5$) and less than two concentrations in the high-toxicity range, use **4pLL** or **BC**.
     Note that in this case the correct solution would be to repeat the experiment with further low concentrations, as Δ cannot be estimated in this situation. If **BC** is chosen, the plausibility of the fit needs to be checked visually. If an implausible fit is obtained with **BC**, then choose **4pLL** instead.

Figure 5.10: Algorithmic procedure summarising which method to use when fitting concentration-response curves to toxicological data from viability assays with potential deviations of the negative controls. Figure slightly modified from Kappenberg et al. (2020).

5.5. Application to a real dataset

All four methods are applied to a real dataset that is introduced in detail in Chapter 3.2.1. Viability of cells is measured for 12 concentration and a negative control for three biological replicates (subsequently called 'donors') with seven technical replicates per concentration. The complete dataset with a 4pLL model fitted to each of the donors (Don1, Don2 and Don3) separately is shown in Figure 5.11.

From these complete datasets, subsets corresponding to the three main scenarios in terms of chosen concentrations and number of replicates are taken. As the concentra-

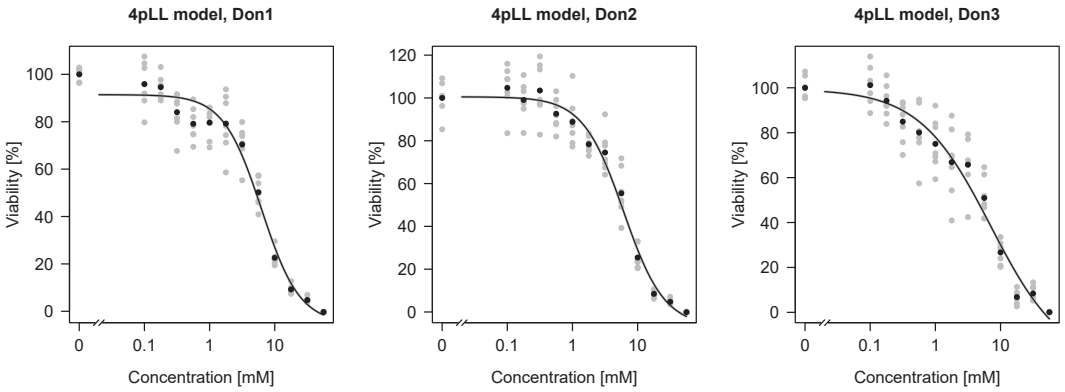

Figure 5.11: Complete dataset measuring viability of cells treated with the compound VPA, with 4pLL models fitted to each of the donors separately. Grey dots indicate the individual measurements and black dots the concentration-wise mean values.

tions and parameters of the simulation study are chosen based on the concentrations and the general profile of this dataset, taking the concentrations corresponding to the main scenarios is straightforward. Three replicates per concentration are obtained by randomly sampling three out of the seven replicates available for each concentration.

The concentration-response curves in Figure 5.11 suggest slightly positively deviating controls for Don1 and slightly negatively deviating controls for Don2. For Don3, the responses measured for the lowest concentrations do not allow a high-quality fit of the upper asymptote. Thus, no clear statement regarding potential deviations of the controls can be made. For further analysis, only Don1 and Don2 are considered.

Figure 5.12 shows curves obtained by applying the four methods **4pLL**, **3pLL**, **No Ctrl**, and **BC** to the dataset for Don1, when choosing the subset that resembles the 'easy' scenario. Corresponding plots for the 'medium' and the 'difficult' scenarios are shown in Figures B.32 and B.33 in Appendix B.1. In all plots, estimates of the EC_{20} are indicated by red lines. Differences in the normalisation procedures are illustrated well in the four plots in Figure 5.12. All four plots have in common that the upper asymptote corresponds to a value of 100%. For the methods **4pLL** and **BC**, the mean value of the responses for the lowest positive measured concentration lies almost exactly on the curve. For **3pLL**, this same mean value lies below the asymptote with a distance of about 5% and for **No Ctrl**, the mean value lies above the asymptote.

In the 'medium' and 'difficult' scenario, the different normalisation for the method **No Ctrl** in contrast to the other three methods is striking: While in the 'medium' scenario, for **4pLL**, **3pLL** and **BC**, the mean response value for the lowest tested concentration is well below the fitted curve, for **No Ctrl** this concentration decisively influences the value of the asymptote. Therefore, the mean value is much closer to the fitted curve. In the 'difficult' scenario, the lowest concentration approximately coincides with the EC_{20} for **4pLL**, **3pLL** and **BC**, while for **No Ctrl**, the second lowest concentration corresponds to a viability of about 80%.

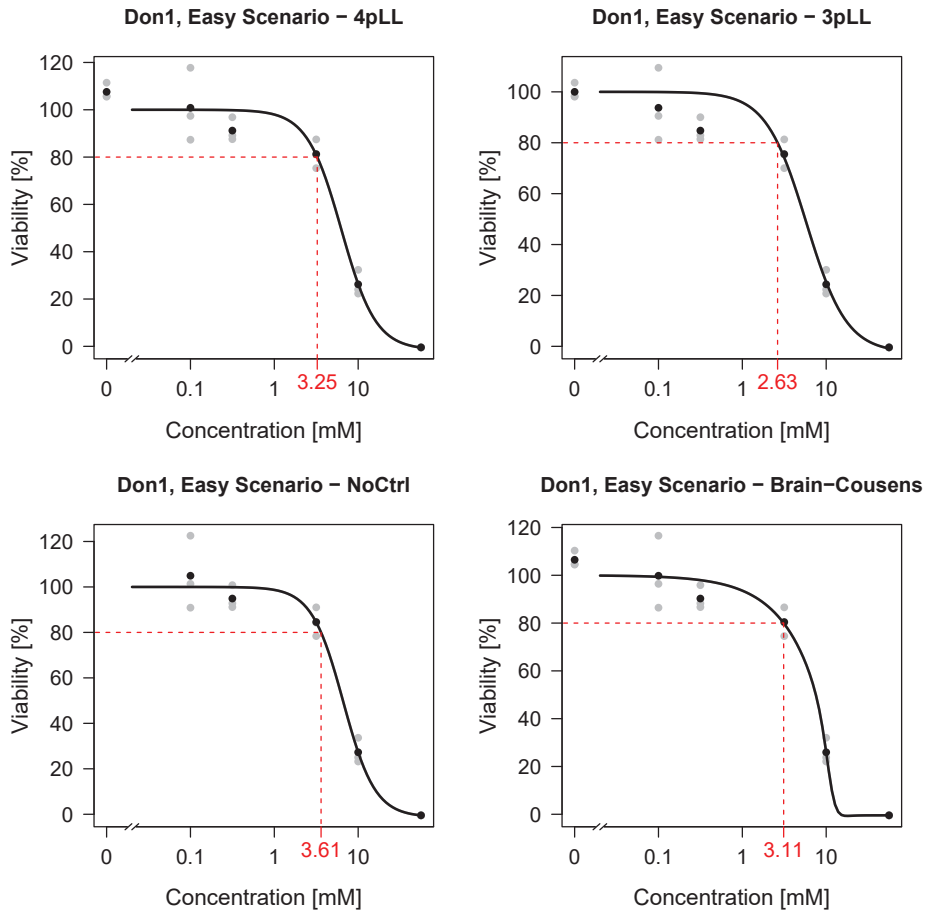

Figure 5.12: Application of the four methods to the original dataset, Don1, resembling the 'easy' scenario. The EC_{20} obtained by each of the four methods is indicated by red lines.

In addition, in the two scenarios 'medium' and 'difficult', implausible behaviour of the curve fitted with the **BC** method can be observed: At first, the curve is monotonously decreasing without modelling a hormesis effect. Then, around the highest tested concentration, the curve starts increasing again, yielding asymptote values even higher than 100%. This behaviour is very implausible from a biological point of view, and the reason why in the recommendations, a visual check of the fitted curves is strongly recommended.

Concentration-wise standard deviations of the three replicates differ slightly across the four methods. This is due to the different normalisation procedures that lead to different response values. Median values across the concentrations in the 'easy' scenario are around 5.3 for **4pLL**, **No Ctrl** and **BC**, and they are slightly smaller for **3pLL**. For the 'medium' scenario, median standard deviations range from 5.2 (**No Ctrl**) to 6.0 (**4pLL**) and for the 'difficult' scenario, median standard deviations are very similar for **4pLL**, **3pLL** and **No Ctrl** with values around 5.8, with a larger value of 6.7 for **No Ctrl**.

It can be seen in Figure 5.12 that the estimated values of the EC_{20} differ considerably, with **3pLL** yielding the smallest value of 2.63 and **No Ctrl** yielding the largest value of 3.61. All resulting EC values, i.e. EC_{10}, EC_{20} and EC_{50}, in the 'easy' scenario for all four methods are summarised in Table 5.2. For **4pLL**, **3pLL** and **No Ctrl**, upper and lower limits of 95% confidence intervals are shown. For **BC**, calculation of the confidence limits is not easily possible, as optimisation procedures are conducted when estimating EC values from curves that do not show a clear hormesis effect. Corresponding results for 'medium' and 'difficult' scenario are given in Tables C.1 and C.2 in Appendix C.

In the 'easy' scenario, for EC_{20} and EC_{50}, **3pLL** yields the smallest estimate, whereas the estimate from **BC** is slightly smaller than that of **3pLL** for the EC_{10}. For the EC_{50}, however, **BC** leads to the largest estimate, while for EC_{10} and EC_{20}, **No Ctrl** leads to the largest estimate. The estimated results differ at most by a factor of 1.63 (EC_{10}), 1.37 (EC_{20}) and 1.34 (EC_{50}). These differences become larger for the 'medium' and 'difficult' scenario, where estimated results differ by at most $3.25, 2.25$ and 1.26, respectively ('medium' scenario) and $2.21, 1.74$ and 1.24, respectively ('difficult' scenario). Overall, the most similar results across the four methods in all three scenarios are obtained for the EC_{50}. In the 'difficult' scenario, only results for **No Ctrl** strongly differ from those for the other three methods that lead to very similar results, only differing by a very small factor.

For the 'easy' and the 'medium' scenario, the lengths of the confidence intervals are decreasing for increasing value of $\lambda \in \{10, 20, 50\}$ for the calculation of EC_λ. In the 'difficult' scenario, for **4pLL** and **3pLL**, the confidence interval for EC_{50} is wider than those for EC_{10} and EC_{20}. The broadest confidence interval observed for any of the estimates is yielded by **No Ctrl** for EC_{10} with a length of 10.7, while for EC_{20} and EC_{50}, only lengths of 4.6 and 2.1 are observed.

True underlying EC values of the curve are not known, therefore the quality of fit cannot be evaluated based on the comparison of the estimated with the true EC values. Instead, the sum of the squared differences between the fitted curve and the response values for all replicates of all concentrations, except the control, are considered. These values are summarised in Table 5.3 for all three main scenarios.

Table 5.2: EC_{10}, EC_{20}, and EC_{50} values together with corresponding limits of 95% confidence intervals for the four methods in a real data study resembling the 'easy' scenario for Don1.

	EC_{10}			**EC_{20}**			**EC_{50}**		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
4pLL	2.22	1.35	3.66	3.25	2.24	4.71	6.23	4.77	8.13
3pLL	1.68	0.88	3.21	2.63	1.70	4.08	5.62	4.16	7.60
No Ctrl	2.54	1.50	4.28	3.61	2.43	5.37	6.58	5.02	8.63
BC	1.56			3.11			7.53		

Table 5.3: Sum of squared differences between fitted curve and response values for all replicates of all concentrations, except the controls, in the three main scenarios for Don1.

	4pLL	**3pLL**	**No Ctrl**	**BC**
'Easy'	887.00	1342.66	864.17	828.00
'Medium'	818.72	909.72	605.21	682.04
'Difficult'	510.25	507.71	667.19	463.59

In the 'easy' scenario, **BC** leads to the smallest value, followed by **No Ctrl** and **4pLL**. The value for **3pLL** is much larger than values for the three other methods. For the 'medium' scenario, **No Ctrl** leads to the best result and again, **3pLL** yields the largest value. Only for the 'difficult' scenario, **No Ctrl** yields the largest value and not **3pLL**, which yields a value similar to that of **4pLL**. **BC** performs best in this case.

The recommendations from Figure 5.10 suggest the use of **4pLL** or **No Ctrl**, depending on the estimated values of $\hat{\sigma}_{\text{med}}$ and Δ. The observed value for $\hat{\sigma}_{\text{med}}$ is about 5.3, and from Figure 5.12 a notable deviation between the imaginary asymptote derived from the responses of the two lowest positive concentrations and the response of the controls can be seen. In that case, the recommended method is **No Ctrl**, resulting in comparatively large estimates of the effective concentrations, but the lowest value for the sum of squared differences between the fitted curve and the response values.

In the 'medium' scenario as seen here, use of **4pLL** or, for a low median standard deviation, **No Ctrl** is suggested. Although the standard deviation is not as low as recommended in the algorithm, **No Ctrl** seems to lead to the best result in this case. In the 'difficult' scenario, **BC** or **4pLL** are recommended. The visual check of the curve fitted by the method **BC** does indeed suggest a biologically implausible result. This coincides with the additional knowledge about the entire curve, instead of only the subset chosen to represent the 'difficult' scenario, in this case: Don1 rather is an example of positively deviating controls, a scenario for which **BC** is not suitable. Therefore, **4pLL** should be chosen, and the sum of squared differences for this method is sufficiently low.

In addition to Don1, for which Figure 5.11 suggests slightly positive deviations of the controls, Don2 is considered, where rather slightly negative deviations are present. Figure 5.13 shows fits obtained by applying the four methods to a part of the dataset for Don2, resembling the 'easy' scenario. Corresponding plots for the 'medium' and the 'difficult' scenarios are shown in Figures B.34 and B.35 in Appendix B.1.

In contrast to the results for Don1, a clear hormesis effect is modelled by **BC** in the 'easy' scenario. In the 'medium' and 'difficult' scenarios on the other hand, **BC** also yields a curve that is monotonously decreasing in (almost) the entire range of concentrations considered. In the 'easy' scenario, the different normalisation procedures can be recognized when comparing the mean responses for the two lowest concentrations to the value of the upper asymptote: This value corresponds to a viability of 100% for all methods except for **BC**, where the curve is normalised in a way that the maximal value instead of the asymptote corresponds to a viability of 100%. For **4pLL** and **3pLL**, mean

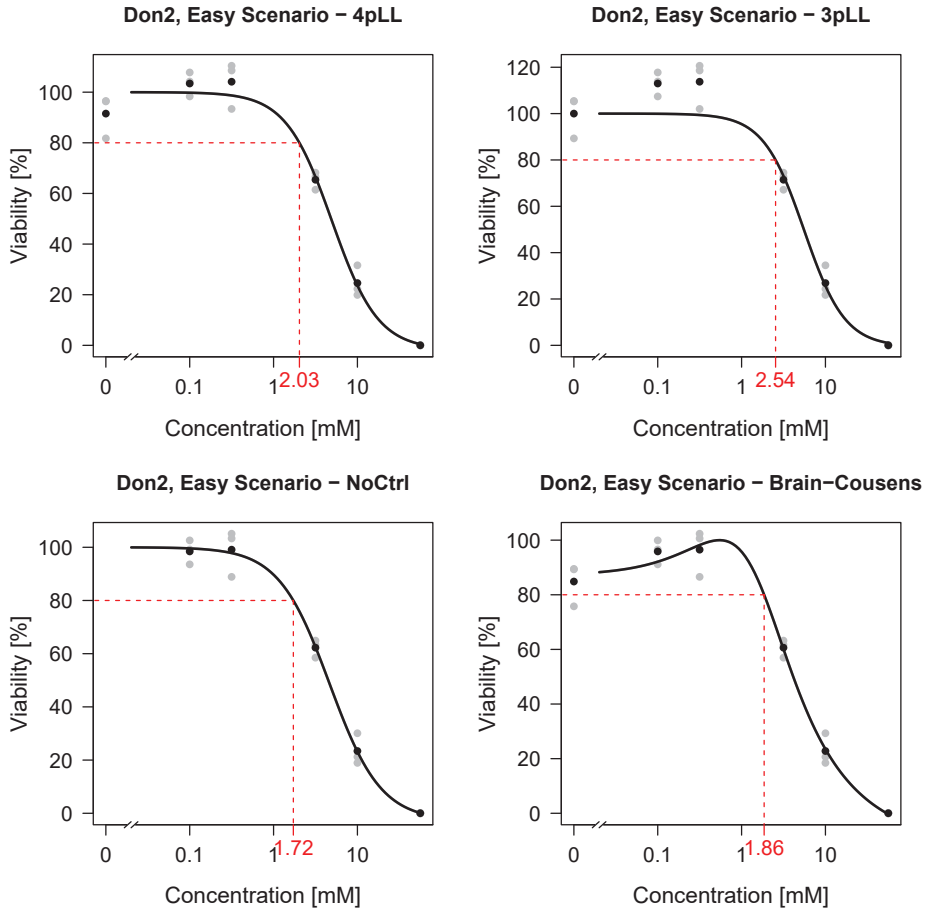

Figure 5.13: Application of the four methods to the original dataset, Don2, resembling the 'easy' scenario. The EC_{20} obtained by each of the four methods is indicated in red.

values of the responses for the two lowest concentrations lie above the asymptote. For **No Ctrl** and **BC**, those response values lie almost exactly on the curve, but the profiles of the curves differ.

In the 'medium' scenario, hardly any visual difference can be perceived between the curves. In the 'difficult' scenario, **4pLL**, **3pLL** and **BC** lead to similar curves, when the profiles are only compared up to the highest tested concentrations. The profile of the curve for **BC** differs from the ones for **4pLL** and **3pLL** in the range of high toxicity, with more plausible results for **BC**. The normalisation for **No Ctrl** strongly differs from the other three methods.

As explained above, values of the concentration-wise standard deviations depend on the method used and are calculated here for all concentrations without the control. Median values across the concentrations in the 'easy' scenario range from 4.4 (**BC**) to 5.2 (**3pLL** and **No Ctrl**). In the 'medium' scenario, **4pLL**, **3pLL** and **BC** correspond to values

Table 5.4: EC_{10}, EC_{20} and EC_{50} values together with corresponding limits of 95% confidence intervals for the four methods in a real data study resembling the 'easy' scenario for Don2.

	$\mathbf{EC_{10}}$			$\mathbf{EC_{20}}$			$\mathbf{EC_{50}}$		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
4pLL	1.22	0.72	2.07	2.03	1.39	2.97	4.86	3.61	6.54
3pLL	1.61	0.93	2.79	2.54	1.70	3.82	5.53	3.84	7.96
No Ctrl	0.98	0.58	1.66	1.72	1.20	2.47	4.49	3.59	5.63
BC	1.33			1.86			4.13		

of 3.7 and **No Ctrl** to 6.4. This pattern is repeated in the 'difficult' scenario with values around 6.2 for **4pLL**, **3pLL** and **BC** and a value of 7.9 for **No Ctrl**.

Estimates for all EC values with corresponding confidence intervals in the 'easy' scenario are summarised in Table 5.4. Corresponding results for the 'medium' and the 'difficult' scenario are summarised in Tables C.3 and C.4 in Appendix C. In the 'easy' scenario, except for the EC_{50}, **No Ctrl** always leads to the smallest estimate and for the EC_{50}, it leads to the second smallest estimate after **BC**. For all three EC values, **3pLL** leads to the largest estimate, but overall the results do not differ strongly between the methods. The maximal factors by with the methods differ are 1.64 (EC_{10}), 1.48 (EC_{20}) and 1.34 (EC_{50}).

Differences between the four methods are in a similar range for the 'medium' scenario (maximal factors of 1.52, 1.25 and 1.06, respectively), but with **No Ctrl** always leading to the maximal estimate here. Only in the 'difficult' scenario, estimates differ more strongly between the methods: Again, **No Ctrl** always leads to the largest estimate and the other three methods to very similar estimates. The estimates differ at most by factors 3.02, 2.09 and 1.22 respectively. In all three scenarios, differences between methods are smallest for estimating the EC_{50}.

As for Don1, generally a decrease in the width of the confidence interval for an increase of $\lambda \in \{10, 20, 50\}$ can be observed. An exception is given by the methods **4pLL** and **3pLL** in the 'difficult' scenario, where confidence intervals for the EC_{50} are much broader than for EC_{10} and EC_{20}.

Again, sum of squared differences between the fitted curves and the response values for all replicates of all concentrations except for the control are calculated in the three main scenarios. Values of these squared differences are summarised in Table 5.5. In the 'easy' scenario, **3pLL** clearly leads to the largest sum of squared differences, which is about five times as large as the sum of squared differences for **No Ctrl**, yielding the smallest result in this scenario. Results in the 'medium' and 'difficult' scenario are quite similar across the three methods **4pLL**, **3pLL** and **No Ctrl** with the smallest sum of squared error obtained in both scenarios by **BC**.

Following the recommendations from Figure 5.10, in the 'easy' scenario the medium value for the standard deviation is smaller than 8, while at the same time a notable value of Δ

Table 5.5: Sum of squared differences between fitted curve and response values for all replicates of all concentrations, except the controls, in the three main scenarios for Don2.

	4pLL	**3pLL**	**No Ctrl**	**BC**
'Easy'	457.31	1530.96	300.64	330.13
'Medium'	447.55	448.03	463.86	398.86
'Difficult'	713.10	704.61	753.47	547.25

can be observed. Therefore, the recommended method to use is **No Ctrl**. The standard deviation for the 'medium' scenario is not smaller than 2, so the recommendation is to use the **4pLL** model, which leads to very similar estimates as **3pLL** and specifically **BC** as well, which is the model with the smallest sum of squared errors. In the 'difficult' scenario, **4pLL** or **BC** are suggested, and the visual check shows that **BC** leads to a plausible concentration-response curve with slightly larger estimates than **4pLL** and at the same time a smaller sum of squared errors.

6. Identification of alert concentrations

In Chapter 4.3, four methods for calculating alert concentrations from concentration-gene expression data are introduced. The four alert concentrations are called ALOEC, LOEC, ALEC, and LEC. The LOEC can be calculated in two different ways, based on the two-sample t-test and on the Dunnett procedure. These methods are compared in a controlled simulation study with three underlying scenarios describing three different concentration-gene expression profiles with respective true ALEC values. In Chapter 6.1, the setup of the simulation study with the choice of sensible standard deviations for the simulated gene expression values is explained in detail. Then the results for the simulation study in the different scenarios are presented in Chapter 6.2. Finally, the four methods for calculating alert concentrations are applied to a real concentration-gene expression dataset and compared. The results are shown in Chapter 6.3.

The results concerning the identification of alert concentrations for concentration-gene expression data are published in Kappenberg et al. (2021). Analyses presented there are extended here by more details regarding the comparison of the two variants of the LOEC. Additionally, for the choice of the probe sets from the entire gene expression dataset, two different variants are calculated and compared here, while in the publication, only one variant is considered.

Many of the plots shown in this thesis are also published in Kappenberg et al. (2021), with only slight adjustments regarding the titles or notation. The figures published there are Figures 6.1, 6.3 to 6.5, 6.8, B.36 to B.38, B.41, and B.42.

6.1. Setup of the simulation study

Properties of the true underlying curves in the simulation study are derived from the real concentration-gene expression dataset introduced in Chapter 3.2.2, where gene expression values for 54675 probe sets were measured for a negative control and 7 increasing concentrations of the compound VPA. In the simulation study, three different scenarios corresponding to the true underlying concentration-response profiles are considered. All three scenarios are based on 4pLL models. They are shown in Figure 6.1.

For all curves, the lower asymptote corresponds to a response value of 0. Thus, the gene expression value that needs to be attained for the true underlying ALEC is equal to the critical effect level λ. The value of λ that needs to be attained or significantly exceeded to yield the respective alert concentration is chosen to represent a fold change (FC) of 1.5. Since data in the VPA gene expression dataset, which is supposed to be resembled by the simulation study, are $\log_2$-transformed, a FC of 1.5 corresponds to the critical effect level $\lambda = \log_2(1.5) \approx 0.585$.

The range of concentrations considered is the interval $[0, 1000]$ with response values evaluated for the concentrations $(0, 25, 150, 350, 450, 550, 800, 1000)$, where concentration 0 refers to the negative control. Three replicates per concentration are considered, yielding $n = 24$ data points in total. Specifically, the parameters and the respective ALEC values in the three simulation scenarios are chosen as follows:

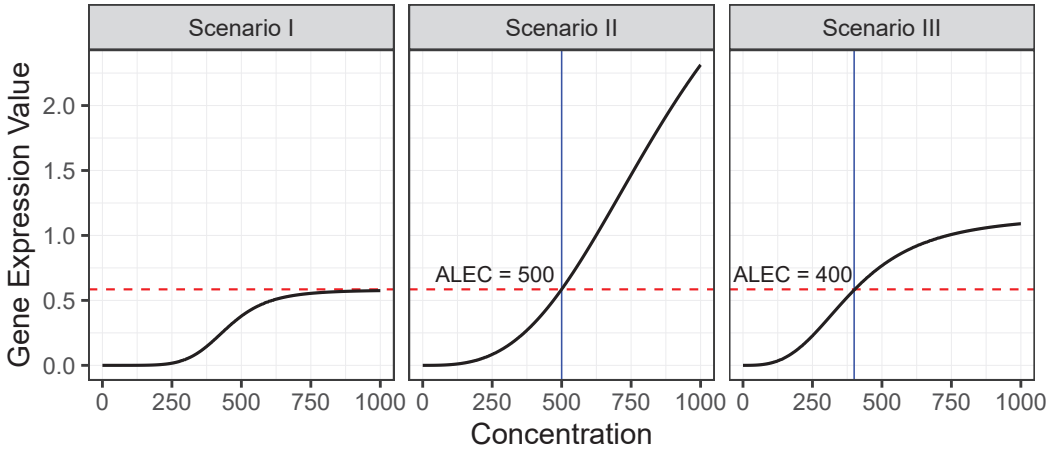

Figure 6.1: Visualisation of the three scenarios used as true underlying curves for the simulation study. The threshold that needs to be (significantly) exceeded is given by $\lambda = \log_2(1.5) \approx 0.585$ and is indicated by a red line. In Scenarios II and III, the value of the true underlying ALEC is indicated by a blue line.

Scenario I: The parameters of the 4pLL model are

$$\phi^{(b)} = -6,\ \phi^{(c)} = 0,\ \phi^{(d)} = 0.58,\ \phi^{(e)} = 450,$$

yielding a curve that never crosses the threshold, i.e. a true ALEC cannot be calculated. Therefore, this scenario serves as null hypothesis for the methods requiring significance. Only for approximately 5% of the simulated genes, calculation of LOEC and LEC should yield a valid result when choosing a significance level of $\alpha = 0.05$.

Scenario II: The parameters of the 4pLL model are

$$\phi^{(b)} = -3,\ \phi^{(c)} = 0,\ \phi^{(d)} = 4,\ \phi^{(e)} = 900,$$

yielding a curve that clearly exceeds the threshold with a true ALEC value of 500. However, the curve is not saturated in the range of concentrations considered. The parameter $\phi^{(e)}$, which corresponds to the concentration where the half-maximal effect is attained, takes the high value of 900.

Scenario III: The parameters of the 4pLL model are

$$\phi^{(b)} = -3,\ \phi^{(c)} = 0,\ \phi^{(d)} = 1.16,\ \phi^{(e)} = 400,$$

yielding a curve that also clearly crosses the threshold with a true ALEC value of 400. This curve is saturated, i.e. the response values for the higher concentrations tend towards the value of the upper asymptote.

The left asymptote of the true underlying curves of all three scenarios attains a value of 0. However, since the observed value of the lower asymptote or the observed response value

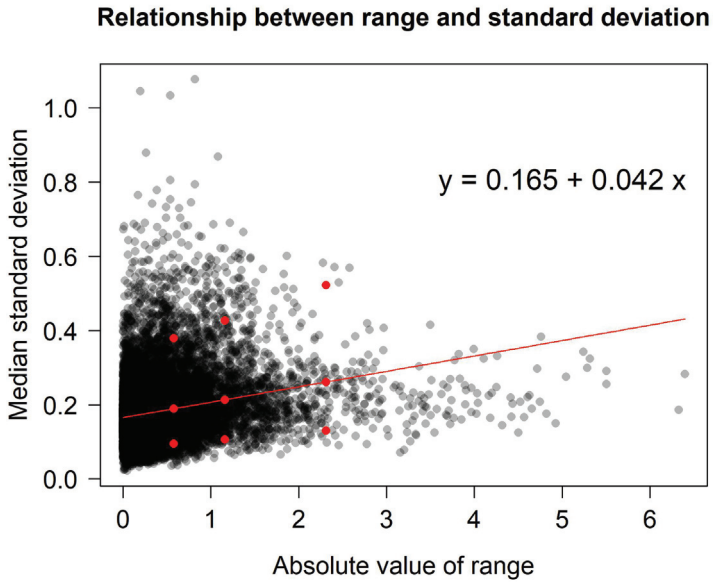

Figure 6.2: Graphical display of the relationship between absolute values of the observed ranges and the median values of concentration-wise standard deviations for a sample of 20000 probe sets from the VPA gene expression dataset. The specific linear model is stated in the plot. The ranges and resulting standard deviations are indicated by red dots.

at concentration 0 for the simulated gene is considered, these scenarios can be generalised to any value of the lower asymptote. Although only increasing profiles are considered in the three scenarios, all analyses are conducted in a two-sided way. Thus, generalisation to situations where the true direction of the profiles is not known in advance, is possible. In Scenarios II and III, where a true ALEC is present, this ALEC does not coincide with any observed concentration. Therefore, in Scenario II, the optimal result for the observation-based methods is 550, and in Scenario III, the optimal result is 450.

For each concentration $x \geq 0$, gene expression data is sampled from a normal distribution with $f(x, \boldsymbol{\phi})$ as mean value. f is a 4pLL model and $\boldsymbol{\phi}$ the parameter vector from the respective Scenario I, II or III. Analysis of the real dataset shows that concentration-wise standard deviations positively correlate with the range of the expression values. For a random sample of 20000 probe sets from the VPA gene expression dataset, ranges of the gene expression profiles are calculated as FC between the highest concentration 1000 µM and the control. The median of concentration-wise standard deviations is calculated for each probe set. Then, a linear model with intercept is fitted to the data, with the median standard deviation as dependent variable and the absolute value of the range as regressor. The relationship between range and standard deviation with indicated linear model is shown in Figure 6.2. For very high ranges, this linear model does not explain the relationship between ranges and standard deviations well, but for smaller ranges the linear model seems to be well-fitting.

The three scenarios considered correspond to ranges of 0.58 (Scenario I), 2.31 (Scenario II) and 1.16 (Scenario III). Note that only the actual observed range, when each curve is considered up to its highest measured concentration, is calculated, not the range of the entire curve given by $|\phi^{(d)} - \phi^{(c)}|$. The corresponding estimated standard deviations are 0.189, 0.261 and 0.213. These are referred to as 'medium' standard deviations ('medium' SD). Additional standard deviations are achieved from multiplying the 'medium' values with the factor 0.5 (yielding 'small' values $(0.095, 0.131, 0.107)$) and the factor 2 (yielding 'large' values $(0.379, 0.522, 0.427)$). These additional situations are referred to as 'small' SD and 'large' SD. The values corresponding to these situations are still observed remarkably often in Figure 6.2 and are thus sensible choices as well.

For each scenario and each of the respective standard deviations, the simulation procedure is repeated 1000 times. This yields nine datasets comprising concentration-gene expression profiles for 1000 simulated genes and 24 concentrations, respectively. All four alert concentrations are calculated for each gene, whereby for the LOEC both the t-test procedure and the Dunnett procedure are performed. The LEC follows from the iterative algorithm based on the newly proposed 4pLL-test.

6.2. Results from the simulation study

Main interest of the simulation study lies in the comparison of the different alert concentrations obtained by the four different methods with respect to their accuracy in estimating the true alert concentration. Some simulated genes have to be excluded from the analysis due to numerical reasons: In some cases, estimation of the covariance matrix Σ yields negative diagonal entries. Since these entries correspond to variances of the parameter estimators, negative values for these variances are not meaningful and are an indicator of numerical difficulties instead. Calculation and interpretation of the 4pLL-model based test is then severely impaired, therefore the genes are excluded from further analysis.

In the situation with 'small' standard deviation, the number of excluded genes is 0 in Scenario I, 1 in Scenario II and 0 in Scenario III. When considering a 'medium' standard deviation, 14, 4 and 2 genes are excluded, respectively and in the situation with 'large' standard deviation, 112, 8 and 16 genes are excluded. As explained in Chapter 4.3.1, in the observation-based approaches, alert concentrations are only calculated if the direction of the concentration-response profile is unambiguous. In the case of ambiguous profiles in the sense defined there, neither LOEC nor ALOEC are calculated. This does not impair possible results for the LEC and the ALEC. Throughout the analysis, the term 'valid estimate' describes an unambiguously resulting alert concentration that lies within the range of considered concentration, i.e. in the interval from 0 to 1000 in the situations considered here.

Main results of the simulation study are summarised graphically in Figures 6.3 (situation with 'small' SD), 6.4 (situation with 'medium' SD) and 6.5 (situation with 'large' SD). For the results shown in these plots, the t-test is used for determining the LOEC. Corresponding results obtained by using the Dunnett procedure for determining the LOEC are shown in Figures B.36, B.37 and B.38 in Appendix B.2.

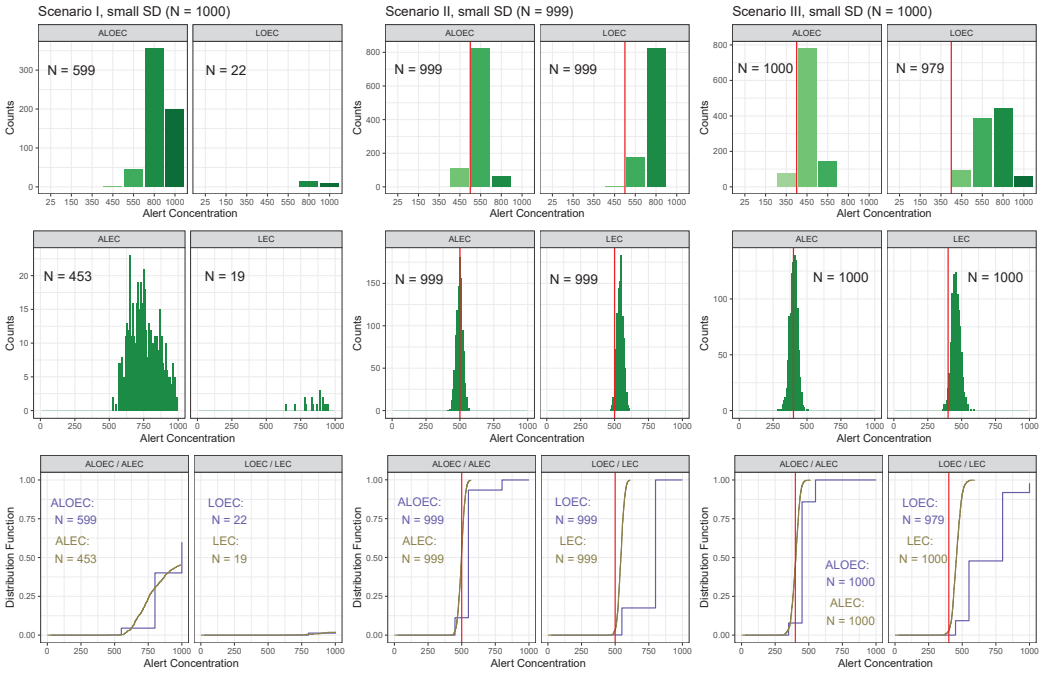

Figure 6.3: Results of the simulation study for 'small' SD. Columns correspond to scenarios and are divided into the criteria **FC** (left) and **FC & *p*-value** (right). The top row depicts the observation-based methods, the middle row the model-based methods and the bottom row shows empirical distribution functions for both methods. True underlying values of the ALEC are indicated by red lines for Scenarios II and III. 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 after exclusion of genes with negative diagonal entries of the covariance matrix.

The figures are structured as follows: Each figure is divided into three columns, which represent from left to right Scenario I, II and III. Then, each column is subdivided into two further columns, where in the first of these columns, the AL(O)EC is displayed and in the second column, the L(O)EC. ALOEC and LOEC, i.e. the observation-based alert concentrations, are summarised in the top row for the three scenarios, respectively. ALEC and LEC, i.e. the model-based alert concentrations, are summarised in the middle row. All results are summarised together in the bottom row, where distribution functions for a direct comparison between both measures ALOEC/ALEC and LOEC/LEC are depicted. The number in the title of each of the three columns indicates the numbers of genes considered after excluding the genes with negative entries of the covariance matrix.

Specifically, the Figures 6.3 - 6.5 and Figures B.36 - B.38 that correspond to the same choice of standard deviation differ only in the method for assessing the significance when calculating the LOEC. Therefore, only the two panels of the plot per scenario corresponding to the LOEC or the comparison of LOEC and LEC (top right and bottom right for each scenario, respectively) are different.

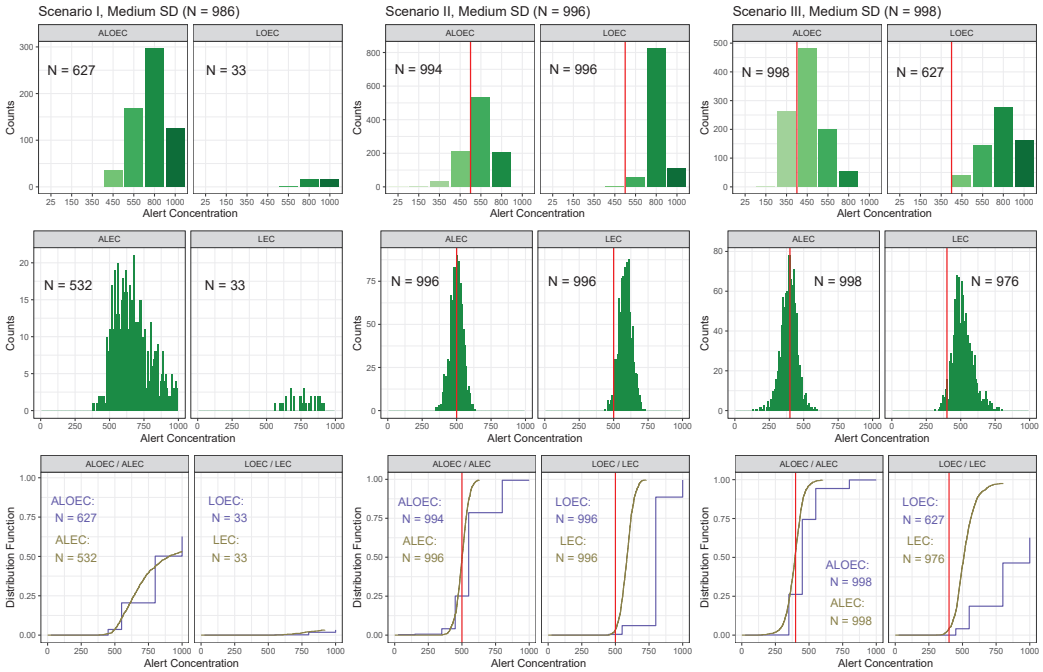

Figure 6.4: Results of the simulation study for 'medium' SD, with the same structure as Figure 6.3.

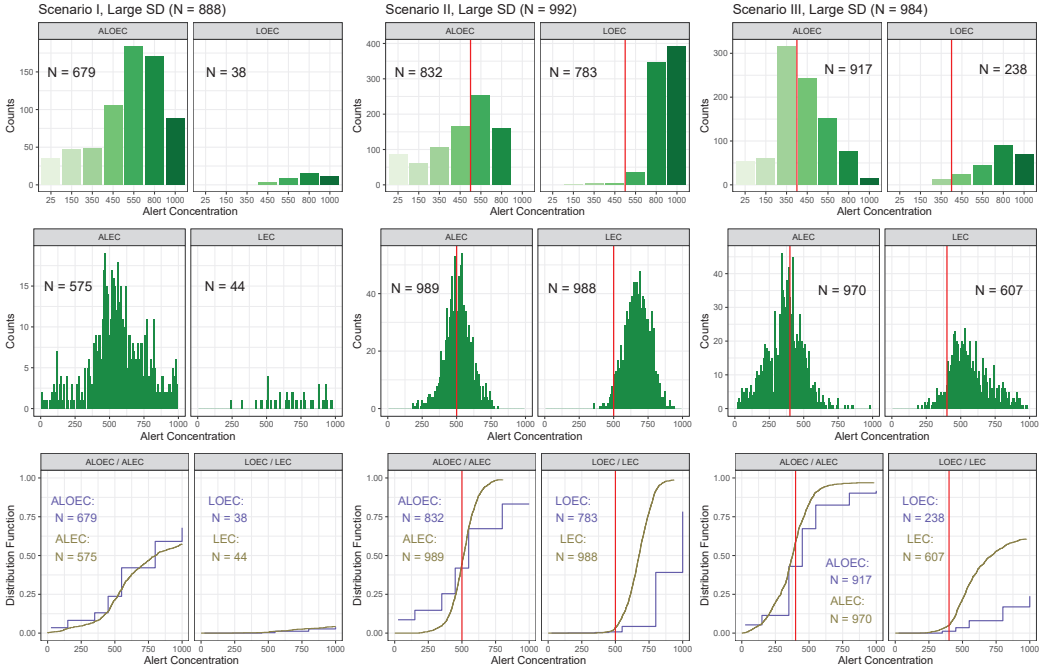

Figure 6.5: Results of the simulation study for 'large' SD, with the same structure as Figure 6.3.

Table 6.1: Summary statistics for the distributions of the ALEC and the LEC. The total number of alerts (n), the median (Med) and the standard deviation (Standard Dev.) are presented for small, medium and large values of the standard deviation in each of the three scenarios.

	n			Med			Standard Dev.		
	Small	Medi.	Large	Small	Medi.	Large	Small	Medi.	Large
				ALEC					
Scen. I	453	532	575	745.3	651.6	553.5	105.0	131.0	207.8
Scen. II	999	996	989	498.0	500.7	510.9	23.7	45.9	95.7
Scen. III	1000	998	970	403.8	396.0	373.0	28.9	61.1	136.1
				LEC					
Scen. I	19	33	44	889.3	768.3	687.3	88.1	98.0	194.2
Scen. II	999	996	988	542.3	585.5	674.1	23.3	46.8	92.8
Scen. III	1000	976	607	456.9	507.4	543.6	32.3	73.3	147.0

Additionally, summary statistics of the key figures for the modelling-based alert concentrations are summed up in Table 6.1. For the ALEC (top part of the table) and the LEC (bottom part of the table), the total number of valid alerts, the median value of the alerts and the standard deviation of the alerts are presented for 'small', 'medium' and 'large' values of the standard deviation in each of the three scenarios considered.

Based on the figures and the tables, it can be seen that the total number of valid alerts differs from the number of considered genes. For ALOEC and LOEC, this may come from excluding ambiguous genes or simply because for none of the concentrations, the FC (significantly) exceeds the pre-defined threshold λ. For ALEC and LEC, the obvious reasons are that the upper asymptote does not exceed the threshold or an estimate is outside of the range of considered concentrations. However, another possible reason is that the curve-fit does not converge due to numerical issues and thus, no resulting curve can be calculated.

A true ALEC can only be calculated for Scenarios II and III, since in Scenario I the effect level of $\lambda = 0.585$ is never reached by the curve. Every alert concentration identified by any of the methods is therefore a *false positive* in this scenario. In Scenarios II and III, an alert concentration is considered to be false positive if it is smaller than the true ALEC value. Total numbers of false positive alerts for all three scenarios for absolute exceedance of the threshold (top rows) and significant exceedance of the threshold (bottom rows) are summarised in Table 6.2.

Main results from these figures and tables are presented for each scenario separately.

Scenario I: In Scenario I, all alerts are false positive alerts. The number of alerts for the methods assessing absolute exceedance of the threshold only ranges from 453 ($\widehat{\text{ALEC}}$ for 'small' SD) to 679 ($\widehat{\text{ALOEC}}$ for 'large' SD), with generally fewer false positives for the $\widehat{\text{ALEC}}$ than for the $\widehat{\text{ALOEC}}$. For the methods taking significance into account, the

Table 6.2: Total numbers of false positive alerts, i.e. estimates below the true ALEC value, and in Scenario I, all identified alerts. The first three rows correspond to the cutoff criterion where an alert is identified when the FC is reached exactly. Since no testing is performed in these cases, the differentiation in t-test and Dunnett is meaningless. The last three rows correspond to significant exceedance of the threshold.

	Scenario I			**Scenario II**			**Scenario III**		
	t-test	Dunn.	4pLL	t-test	Dunn.	4pLL	t-test	Dunn.	4pLL
				AL(O)EC					
Small	599		453	112		541	78		455
Medium	627		532	251		491	262		536
Large	679		575	419		444	430		587
				L(O)EC					
Small	22	8	19	1	0	34	0	0	28
Medium	33	12	33	5	0	35	0	0	42
Large	38	4	44	8	0	36	12	1	50

number of false positive alerts accounts for less than 5% of all simulated genes, thus all methods maintain the significance level of $\alpha = 0.05$. For these three methods, the $\widehat{\text{LOEC}}$ based on the Dunnett procedure yields the fewest alerts, while the $\widehat{\text{LOEC}}$ based on the t-test and the $\widehat{\text{LEC}}$ based on the 4pLL-test yield comparatively many alerts.

The summary statistics show that median values of the $\widehat{\text{ALEC}}$ are smaller, the larger the SD considered is, while the standard deviation of the alert concentrations becomes larger with larger SD. This can also be seen in the histograms for the $\widehat{\text{ALEC}}$. While for 'small' SD, the smallest alert concentration takes a value of about 500, for the 'medium' SD the smallest alert concentration takes a value of about 375 and for the 'large' SD even a value of 0.

Since an estimated $\widehat{\text{ALEC}}$ close to 0 seems implausible at first glance, the courses of the three simulated genes yielding the smallest $\widehat{\text{ALEC}}$ values for the 'large' SD are shown in Figure 6.6. All three courses are very similar: The mean response value for the control takes a value of approximately -0.5. The response for the lowest measured concentration and all following concentrations are at the same level of about 0.5. Thus a very steep increase of the curve between the concentration 0 and the concentration 25 is present, leading to the observed very small values of the $\widehat{\text{ALEC}}$. Only for the third example gene, a valid $\widehat{\text{LEC}}$ can be calculated, with a value of 318. The $\widehat{\text{ALOEC}}$ of all three genes is 25, but a $\widehat{\text{LOEC}}$ based on the t-test can again only be calculated for the third gene, yielding a result of 800. No $\widehat{\text{LOEC}}$ based on the Dunnett procedure can be calculated for either of the genes.

The same structure of results as for the $\widehat{\text{ALEC}}$ can be observed in the barplots depicting the $\widehat{\text{ALOEC}}$: For the 'small' SD, most alerts are observed at the concentration 800, followed by 1000 and only very few alerts for lower concentrations. For 'medium' SD,

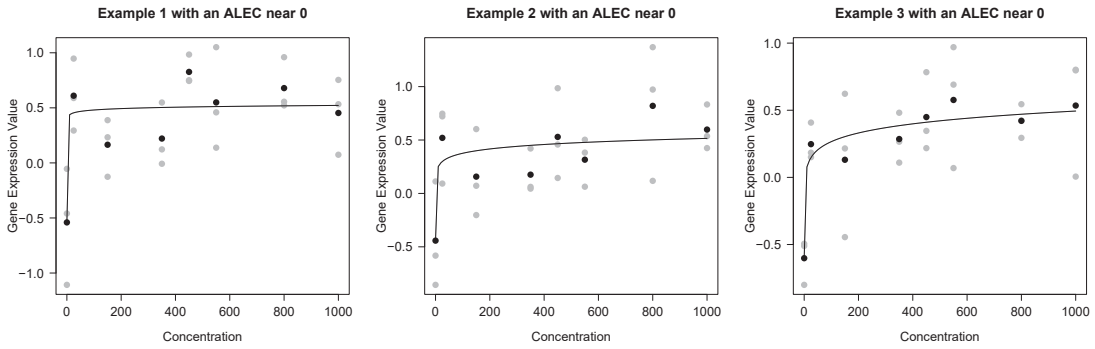

Figure 6.6: Three simulated genes with very low values of the ALEC. Grey dots show the individual simulated response values per concentration and black dots depict the concentration-wise means of the responses.

still most alerts are observed at the concentration 800, but followed by the concentration 550 and only then the concentration 1000. Finally, for 'large' SD, the $\widehat{\text{ALOEC}}$ yields most alerts at the concentration 550 but all in all, the observed alert concentrations cover the entire set of possible values. Detailed comparison of the methods that additionally assess the significance is not meaningful due to the very small sample sizes.

The distribution functions of $\widehat{\text{ALOEC}}$ and $\widehat{\text{ALEC}}$ have a very similar profile in terms of their endpoint, their slope, and the concentrations where the respective curves start to increase for all three possible values of the SD. Again, comparison of the distribution functions of $\widehat{\text{LOEC}}$ and $\widehat{\text{LEC}}$ is not possible due to the small sample size.

Scenario II: For Scenario II, the true value of the ALEC is given by 500. About half of all simulated genes, with the largest number for the 'small' SD, lead to a false positive alert for the $\widehat{\text{ALEC}}$. This number is far higher than the number of false positive alerts for $\widehat{\text{ALOEC}}$ in the cases of 'small' and 'medium' SD and still higher by 25 for 'large' SD. For the $\widehat{\text{LOEC}}$ based on the t-test, less than 10 genes lead to false positive alerts and for the $\widehat{\text{LOEC}}$ based on the Dunnett procedure, no false positive alert is observed. For the $\widehat{\text{LEC}}$, the number of false positive alerts is between 33 and 36 and is therefore below the significance level of 5%.

For 'small' and 'medium' SD, except for 2 genes in the estimation of the $\widehat{\text{ALOEC}}$ for 'medium' SD, all genes lead to valid estimates of $\widehat{\text{ALOEC}}$, $\widehat{\text{ALEC}}$, both versions of $\widehat{\text{LOEC}}$ and $\widehat{\text{LEC}}$. In the situation with 'large' SD, this is not the case: While for the model-based methods, all genes except for 2 or 3 yield a valid estimate, only 832 valid estimates are obtained for the $\widehat{\text{ALOEC}}$ and only 783 and 848 for the t-test and Dunnett procedure based $\widehat{\text{LOEC}}$, respectively.

The median values of the $\widehat{\text{ALEC}}$ for all three situations regarding the SD are very close to the true value of 500, with larger standard deviations of the alert concentrations in the situations with larger SD. Median values of $\widehat{\text{LEC}}$ are consistently larger than those of $\widehat{\text{ALEC}}$ with an increasing difference as SD increases, but with very comparable values of

the standard deviations. These observations are supported by the histograms for $\widehat{\text{ALEC}}$ and $\widehat{\text{LEC}}$: The histograms for the $\widehat{\text{ALEC}}$ are centered around the true value indicated by the red line, while for the $\widehat{\text{LEC}}$ a bias towards higher values can be observed. For larger values of the SD, the range of observed values becomes larger: For 'small' SD, the $\widehat{\text{ALEC}}$ ranges from about 410 to about 570. For 'medium' SD, it ranges from about 350 to about 630 and even from about 180 to about 800 for 'large' SD.

For the $\widehat{\text{ALOEC}}$ and all three values of SD, most alerts are observed at the concentration 550, which is the smallest concentration that is larger than the true value of the ALEC. Thus, this is the best possible result for this method. In none of the three situations regarding the SD, any observation for an $\widehat{\text{ALOEC}}$ of 1000 is made. The distribution of the observed alerts across the possible measured concentrations becomes broader for larger SD, spanning all measured concentrations from 25 to 800 for 'large' SD.

For the $\widehat{\text{LOEC}}$, most alerts yield a value of 800 for 'small' and 'medium' SD and a value of 1000 for 'large' SD, irrespective of the method used for calculation of the $\widehat{\text{LOEC}}$. A direct comparison of the values of $\widehat{\text{LOEC}}$ using the t-test and using the Dunnett procedure is given by Table 6.3. In the situation with 'small' SD, alert concentrations mostly coincide for both methods. Only in 29 cases does the Dunnett procedure yield a smaller concentration, and in 95 cases the t-test yields a smaller alert than the respective other method. For 'medium' SD, more than 800 observations coincide with a value of 800, which is observed in total in 950 cases for the Dunnett procedure. For the t-test, 100 of these genes yield the higher value of 1000 and 42 yield smaller values. For 'large' SD, the alerts are more spread out, mainly across the alerts 800, 1000 and also 'No Alert'. The alert concentrations of more than half of the genes still coincide at the three mentioned values.

The distribution functions for 'small' and 'medium' SD show that model-based alerts generally take lower values than observation-based alerts, and this discrepancy becomes larger for the methods taking significance into account. This observation holds for the comparison of $\widehat{\text{LOEC}}$ and $\widehat{\text{LEC}}$ in the situation with 'large' SD as well. However, for $\widehat{\text{ALOEC}}$ and $\widehat{\text{ALEC}}$, first there are more observation-based alerts for lower concentrations, then the distribution functions intersect and more model-based alerts are observed starting at concentration 500.

Scenario III: The true value of the ALEC is given by 400 in Scenario III. For the $\widehat{\text{ALOEC}}$, the number of false positive alerts ranges from 78 for 'small' SD to 430 for 'large' SD, and the number of false positive alerts is even higher with numbers from 455 to 587 for the $\widehat{\text{ALEC}}$. For the methods taking significance into account, again the number of false positive alerts accounts for less than five percent of all simulated genes. The $\widehat{\text{LOEC}}$ in both versions yields false positive alerts in the situation with 'large' SD only.

Summary statistics for $\widehat{\text{ALEC}}$ and $\widehat{\text{LEC}}$ show that median values of the $\widehat{\text{ALEC}}$ are smaller than the true value of 400 with a difference that increases with increasing SD. Median values of the $\widehat{\text{LEC}}$ are again larger and increasing with increasing SD. For 'large' SD, only 607 simulated genes yield a valid estimate of the $\widehat{\text{LEC}}$, while in all other cases, none

Table 6.3: Comparison of the alert concentrations $\widehat{\text{LOEC}}$ based on the t-test (rows) and on the Dunnett procedure (columns) for Scenario II. 'NA' indicates the case in which no valid alert concentration can be determined.

		LOEC Dunnett							
		'small' SD				**'medium' SD**			
		550	800	1000	NA	550	800	1000	NA
	150	0	0	0	0	0	0	0	0
	350	0	0	0	0	0	0	0	0
LOEC	450	1	0	0	0	2	3	0	0
***t*-test**	550	80	94	0	0	17	39	0	0
	800	29	795	0	0	10	808	7	0
	1000	0	0	0	0	0	100	10	0
	NA	0	0	0	0	0	0	0	0

		LOEC Dunnett			
		'large' SD			
		550	800	1000	NA
	150	0	0	0	1
	350	0	3	0	0
LOEC	450	0	3	1	0
***t*-test**	550	6	23	4	2
	800	0	232	99	17
	1000	0	76	275	41
	NA	0	27	99	83

or only few genes yield no alert concentration. The histograms of $\widehat{\text{ALEC}}$ and $\widehat{\text{LEC}}$ show the same structure as described for Scenario II for 'small' and 'medium' SD. For 'large' SD, apart from the different number of valid alert concentrations in the case of the $\widehat{\text{LEC}}$, the histograms are comparatively broader with heavier tails.

For the observation-based methods for calculating the alert concentration, 450 would be the best result. This is also the alert concentration that is obtained most often by the $\widehat{\text{ALOEC}}$ for the situations with 'small' and 'medium' SD. For 'large' SD, the concentration observed most often is 350, but then followed by 450, 550 and 800. In this situation, all possible concentrations are observed as $\widehat{\text{ALOEC}}$ for at least 15 simulated genes, respectively. The results span across the concentrations 150 to 800 for the 'medium' SD and only across 350, 450, and 550 for 'small' SD. The $\widehat{\text{LOEC}}$ vastly overestimates the true value of the ALEC, with 800 as most frequent result. In the situation with 'medium' SD, the second most alerts are also calculated for the concentration 1000. However, the number of valid estimates is only 627 or 607 in this situation and even smaller with 238 and 143 estimates only for 'large' SD.

Table 6.4: Comparison of the alert concentrations $\widehat{\text{LOEC}}$ based on the t-test (rows) and on the Dunnett procedure (columns) for Scenario III. 'NA' indicates the case in which no valid alert concentration can be determined.

		LOEC Dunnett									
		'small' SD					**'medium' SD**				
		450	550	800	1000	NA	450	550	800	1000	NA
	350	0	0	0	0	0	0	0	0	0	0
	450	34	40	19	0	0	12	8	17	2	2
LOEC	550	14	277	92	2	0	1	72	39	20	13
***t*-test**	800	3	101	333	4	0	4	12	170	34	58
	1000	0	9	42	9	0	1	6	31	79	46
	NA	0	2	18	1	0	0	10	42	47	272

		LOEC Dunnett					
		'large' SD					
		350	450	550	800	1000	NA
	350	1	0	3	2	1	5
	450	0	4	1	4	1	13
LOEC	550	0	1	12	3	6	23
***t*-test**	800	0	0	4	33	7	45
	1000	0	0	1	6	20	42
	NA	0	0	8	14	11	713

A direct comparison of both variants of the $\widehat{\text{LOEC}}$ is given in Table 6.4. For 'small' SD, most of the alerts coincide at the same concentrations, and about equally many genes yield a smaller alert for the Dunnet variant and for the t-test variant, respectively. For 'medium' SD, the alerts in the direct comparison are more scattered, but still with most alerts coinciding at the same concentration or deviating by only one level. However, some genes with a low alert concentration in either of the variants do not yield a valid alert in the other variant. In the situation with 'large' SD, most observations coincide at yielding no alert. For the Dunnett procedure, apart from the coinciding genes, about 130 additional genes do not yield a valid alert, and the respective alerts for these genes in the t-test variant scatter across all concentrations from 350 to 1000.

All in all, the observations regarding the comparison of t-test and Dunnett procedure for estimating the $\widehat{\text{LOEC}}$ for both Scenario II and III show that not one of these methods is generally more conservative. In most cases, the alerts coincide, but in the other cases none of these methods structurally yields the smaller results, even though in the Dunnett procedure it is accounted for multiple testing.

Comparison of the distribution functions for $\widehat{\text{ALEC}}$ and $\widehat{\text{ALOEC}}$ shows generally slightly smaller alerts for the $\widehat{\text{ALEC}}$. For the comparison of $\widehat{\text{LEC}}$ and $\widehat{\text{LOEC}}$, this difference is

Table 6.5: Coverage Probabilities of the 95% CIs for the ALEC in Scenario II and III. Only those CI, whose length is less than or equal to 1000, are taken into account. The number of these CI is indicated as well.

	Scenario II		**Scenario III**	
	n	CP	n	CP
'Small' SD	999	0.83	1000	0.83
'Medium' SD	996	0.83	998	0.84
'Large' SD	979	0.86	920	0.77

larger. Especially for 'medium' and 'large' SD, the different numbers of valid estimates is again illustrated by these distribution functions.

To summarise these results, the model-based approaches ALEC and LEC perform better than the observation-based approaches ALOEC and LOEC. Fewer false positive alerts are calculated by the model-based approaches in Scenario I, where the underlying curve is chosen such that it does not exceed the critical threshold. In Scenarios II and III, where a true ALEC value can be calculated based on the true underlying curve, model-based alert concentrations less drastically overestimate this true value and estimates are overall closer to the true value of the ALEC, as can be seen by the steepness of the distribution functions. A drastic difference especially for Scenario III is given by the number of valid estimates yielded by the model-based methods, where mostly for almost all simulated genes an alert can be calculated, and the observation-based approaches, where far fewer genes yield a valid estimate.

In addition to the analyses presented above, 95% confidence intervals (CIs) are calculated for the ALEC in the log-transformed version introduced in Chapter 4.3.2. Based on these CIs, coverage probabilities (CPs) are calculated as the proportion of cases in which the true value of the ALEC lies inside the CI. These CPs can only be calculated for Scenarios II and III, since for Scenario I, no true value of the ALEC is given. For the calculation of the CPs, only CIs that are not wider than 1000 are taken into account. The CPs for all three situations regarding the SD are summarised in Table 6.5. Most of these CPs yield results between 0.83 and 0.86, with the exception of Scenario III, 'large' SD, where the CP takes only a value of 0.77. Generally, these values are quite low, as one would expect a result of approximately 0.95 for the 95% CIs considered here.

6.3. Application to a real dataset

The four methods for calculating alert concentrations are applied to the real dataset introduced in Chapter 3.2.2. Briefly summarised, this dataset consists of (adequately pre-processed) gene expression values, measured for 54675 probe sets in 7 increasing concentrations with three replicates each and a negative control with six replicates. In order to reduce the dimension of the considered dataset, only those probe sets are considered that have a significant change in response value for at least one concentration: An `anova` procedure is applied to each of the probe sets and only those probe sets are

further considered that lead to an unadjusted p-value smaller than 0.001. This way, only 9460 of the initial 54675 probe sets remain in the dataset.

An alternative to applying `anova` is the application of the MCP-Mod methodology as explained in Chapter 4.1.5 and considering only those probe sets for which a PoC can be established. For comparability, the same level $\alpha = 0.001$ is chosen. The set of candidate models only consists of the `sigEmax` model in this context, with chosen guesstimate priors of EC50 = 450 and $h = 5.117$. These values are based on the assumptions that the half-maximal effect might be observed at concentration 450, and already 95% of the maximal effect is observed at concentration 800. The parameter h is then calculated from this pair of assumptions. These heuristic guesstimates that are equal for all genes are used, since for the many thousand probe sets considered simultaneously, individual guesstimates are not practicable.

Applying the MCP-Mod procedure with the guesstimate specified above yields 15306 probe sets with a p-value smaller than 0.001, with 9188 in the overlap of the probe sets selected by `anova` and MCP-Mod. This leaves only 272 probe sets considered significant by `anova` and not by MCP-Mod and 6118 probe sets vice versa. Out of the 6118 probe sets pre-selected only by the MCP-Mod approach, estimation of the covariance matrix yields negative diagonal entries for 215 probe sets. For the remaining 5903 probe sets, an $\widehat{\text{ALOEC}}$ can only be calculated for 2265 of them. For the $\widehat{\text{LOEC}}$, this number is even higher, with 5746 and 5888 probe sets that do not yield a valid $\widehat{\text{LOEC}}$ for the two variants, respectively. Similar results can be observed for the model-based versions with 3905 probe sets without a valid $\widehat{\text{ALEC}}$ and even 5497 probe sets without a valid $\widehat{\text{LEC}}$. Due to this very high number of probe sets without valid alert concentrations, no additional information is gained by choosing the larger set of pre-selected probe sets for further analysis. The following analyses are therefore restricted to the set of probe sets pre-selected by the `anova` procedure.

Again, all probe sets with diagonal entries for the estimated covariance matrix are excluded from further analysis. This applies to 286 probe sets, keeping 9174 out of the 9460 probesets in the analysis. The $\widehat{\text{ALOEC}}$ yields 7074 valid estimates and the $\widehat{\text{ALEC}}$ 6811. For the methods taking significance into account, the number of valid estimates is smaller with 4126 and 3648 valid estimates for the $\widehat{\text{LOEC}}$ (t-test and Dunnett-based, respectively) and 4929 valid estimates for the $\widehat{\text{LEC}}$.

Since for this dataset the true underlying values of the ALEC are not known, univariate considerations of the resulting alert concentrations are less meaningful than the direct comparison of observation- and model-based methods with or without assessing significance, respectively. Figure 6.7 shows the relationship between estimated $\widehat{\text{ALOEC}}$ and $\widehat{\text{ALEC}}$ for the probe sets from the VPA dataset: Values of the $\widehat{\text{ALEC}}$ alert concentration, stratified by corresponding values of the $\widehat{\text{ALOEC}}$, are summarised by boxplots. The $\widehat{\text{ALOEC}}$ is also indicated by a red dot in each boxplot for easier comparison of observation-based and model-based alert concentrations. Numbers in the bottom row indicate the number of probe sets with the corresponding value of the alert concentration $\widehat{\text{ALOEC}}$, numbers in the top row indicate the number of genes with a corresponding $\widehat{\text{ALOEC}}$ but without a valid result for the $\widehat{\text{ALEC}}$.

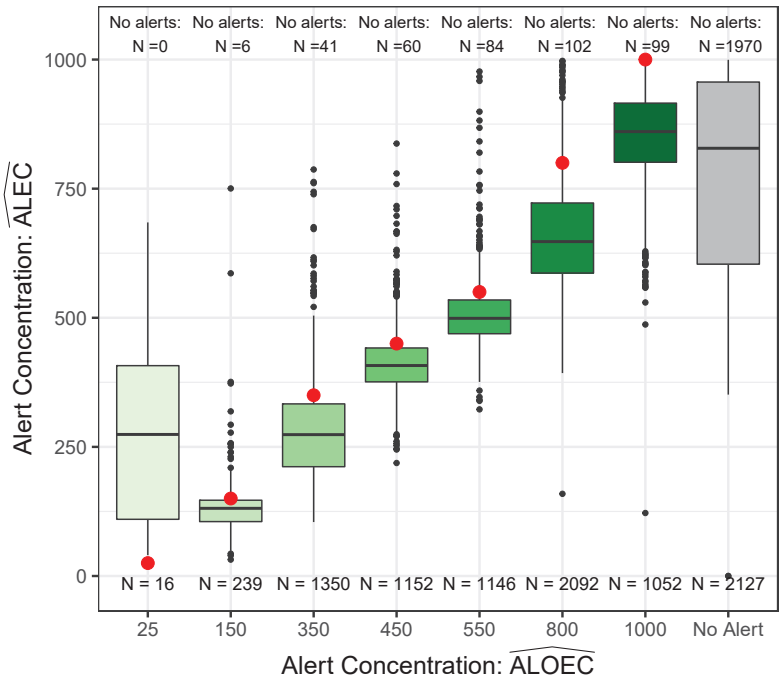

Figure 6.7: Results of the analysis of the VPA dataset for methods considering absolute exceedance of the threshold. The boxplots summarise values of the $\widehat{\text{ALEC}}$ values, stratified by corresponding $\widehat{\text{ALOEC}}$ values, which are also visualized by red dots. Numbers in the bottom row indicate total numbers of ALOEC alerts, and numbers in the top row indicate cases with ALEC alert outside the permitted range, i.e. each boxplot comprises alert concentrations for 'bottom number - top number' probe sets, e.g. for an $\widehat{\text{ALOEC}}$ of 350 this corresponds to $1350 - 41 = 1309$ probe sets.

Except for the smallest $\widehat{\text{ALOEC}}$ of 25, the boxes summarising the $\widehat{\text{ALEC}}$ are below the indicated points, meaning that in more than 75% of the cases, the model-based approach yields lower alert concentrations than the observation-based approach. The difference between $\widehat{\text{ALOEC}}$ and $\widehat{\text{ALEC}}$ becomes larger with larger value of the $\widehat{\text{ALOEC}}$. For an $\widehat{\text{ALOEC}}$ of 800, about 88% of the observed values of the $\widehat{\text{ALEC}}$ take a value smaller than 800, and for an $\widehat{\text{ALOEC}}$ of 1000, all observed values of the $\widehat{\text{ALEC}}$ are smaller. Thereby, the number of invalid alert concentrations is not larger here than for an $\widehat{\text{ALOEC}}$ of 800.

Of the 2127 probe sets yielding no valid alert concentration for the $\widehat{\text{ALOEC}}$, 1970 do not yield a valid $\widehat{\text{ALEC}}$ either. The quartiles of the $\widehat{\text{ALEC}}$ concentrations for the remaining 157 probe sets are given by 603 and 957. For the $\widehat{\text{ALOEC}}$s 550, 800 and 1000, the number of probe sets without a valid $\widehat{\text{ALEC}}$ ranges from 84 to 102, with less missing alert concentrations for smaller $\widehat{\text{ALOEC}}$. Most $\widehat{\text{ALOEC}}$s are observed at concentration 800, followed by 450, 550 and 1000, which have similar numbers. Quartiles of the distribution of all $\widehat{\text{ALEC}}$s are 374 and 694 with a median value of 523, so that all in all, the $\widehat{\text{ALEC}}$ yields smaller values than the $\widehat{\text{ALOEC}}$.

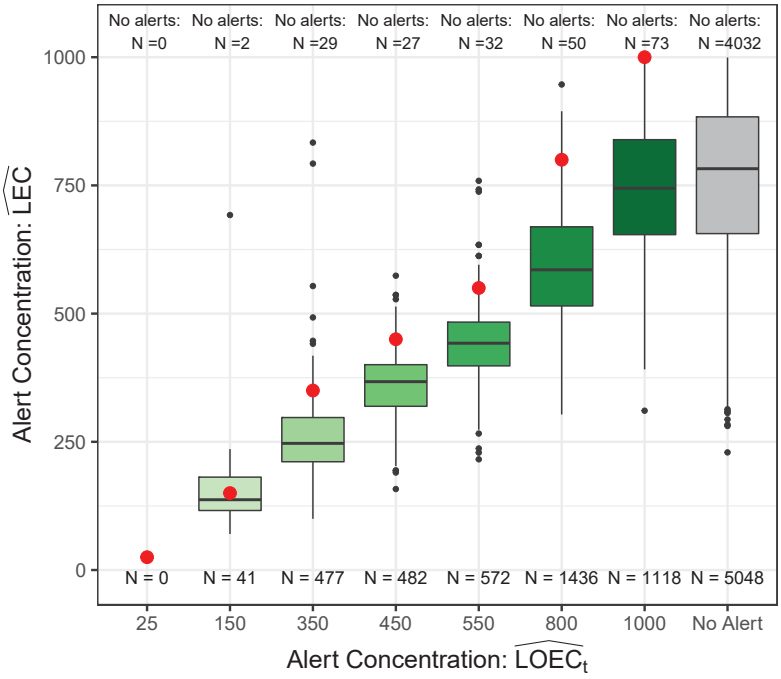

Figure 6.8: Results of the analysis of the VPA dataset for $\widehat{\text{LOEC}}$ based on the t-test and $\widehat{\text{LEC}}$. The structure of the plot is the same as Figure 6.7.

Additionally, the results are considered separately for increasing and decreasing probe sets. The direction of the concentration-gene expression profile is determined based on the direction of the fitted curve, yielding 4949 probe sets with increasing and 4225 probe sets with decreasing profiles. Figures B.39 and B.40 in Appendix B.2 show the same boxplots as introduced above, once only for the increasing and once only for the decreasing probe sets. Generally, the results are very similar to the ones described for the entire set of probe sets, with about equally many probe sets yielding $\widehat{\text{ALOEC}}$s between 450 and 1000, respectively. For decreasing probe sets, the concentrations obtained by the $\widehat{\text{ALEC}}$ are comparatively slightly larger than for increasing probe sets, but all in all the same statements as for the entire dataset can be made.

The same analyses are conducted for the methods taking significance into account. Figure 6.8 shows the same boxplots as introduced above for the comparison of $\widehat{\text{LEC}}$ and the $\widehat{\text{LOEC}}$ based on the t-test, and Figure 6.9 the corresponding comparison for the $\widehat{\text{LOEC}}$ based on the Dunnett procedure. The first observation is that the number of probe sets that do not yield a valid estimate for either of the $\widehat{\text{LOEC}}$s is far higher than for the $\widehat{\text{ALOEC}}$. Out of the approximately 5000 or 5500 probe sets not yielding an estimate for the respective $\widehat{\text{LOEC}}$, (slightly) more than 1000 still yield a valid estimate for the $\widehat{\text{LEC}}$, with estimates spread quite broadly across the range of considered concentrations.

Again, the concentration 800 is the one where most alerts can be observed for the $\widehat{\text{LOEC}}$. Quartiles of the $\widehat{\text{LEC}}$ are given by 433 and 746 with a median of 587. All of these values

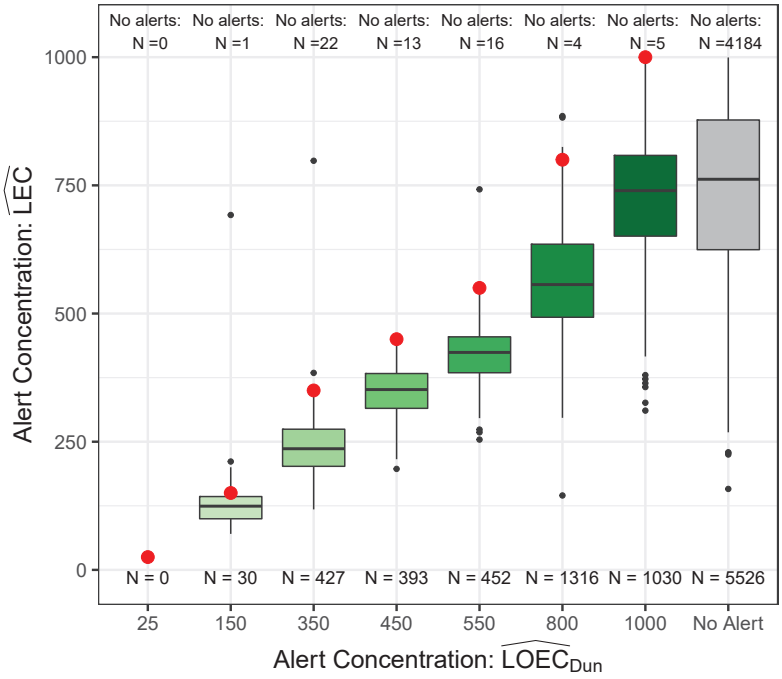

Figure 6.9: Results of the analysis of the VPA dataset for $\widehat{\text{LOEC}}$ based on the Dunnett procedure and $\widehat{\text{LEC}}$. The structure of the plot is the same as Figure 6.7.

are larger by about 60 than the corresponding values for the $\widehat{\text{ALEC}}$. Generally, in comparison with the respective $\widehat{\text{LOEC}}$, the $\widehat{\text{LEC}}$ is far smaller. For the $\widehat{\text{LOEC}}$ based on the Dunnett procedure, except for the concentration 150, not only the box of the boxplot for the respective $\widehat{\text{LEC}}$s, but the box and the whiskers are below the displayed red dot that indicates the corresponding $\widehat{\text{LOEC}}$. For the $\widehat{\text{LOEC}}$ based on the t-test, the red dot approximately lies at the center of the upper whisker.

Analogously to the results for the $\widehat{\text{ALEC}}$ and the $\widehat{\text{ALOEC}}$, Figures B.41 to B.44 in Appendix B.2 show the respective boxplots for both variants of the $\widehat{\text{LOEC}}$ separated into probe sets with increasing and with decreasing profiles. For the $\widehat{\text{LOEC}}$ based on the t-test and on the Dunnett procedure, in the subset of probe sets with decreasing profile, only very few alerts are observed at concentrations $25, 150$ and 350. For increasing profiles, more observations are made for the concentrations 150 and 350. No probe sets results in a $\widehat{\text{LOEC}}$ of 25, regardless of the direction of the profile. Apart from these observations, results for these probe sets separated into the direction of the curve are analogous to the ones for the entire dataset.

7. Information sharing across genes

Two methods for information sharing, meta-analysis and an empirical Bayes method, are introduced in Chapters 4.4.2 and 4.4.3. In simulation studies, estimating the parameter $\phi^{(e)^*}$ from a 4pLL model making use of these methods is compared to the direct estimation of $\phi^{(e)^*}$. These simulation studies heavily depend on the structure and biological properties of the underlying true dataset. Thus, plasmode datasets are used to capture the biological properties. A plasmode dataset is obtained by using real data that is manipulated in a way that the true effects are known, while retaining the basic properties of the real dataset (see Chapter 4.4.1). The underlying VPA dataset, introduced in Chapter 3.2.2, is used as basis for these plasmode simulation studies. This dataset is first described in detail. Results regarding the parameter distribution for fitted 4pLL curves are presented in Chapter 7.1 and results regarding the GO groups are presented in Chapter 7.2.

First, two simulation studies for the meta-analysis approach are conducted and presented in Chapter 7.3. The first simulation study is based on the entire set of considered genes, and the setup and the results are summarised in Chapter 7.3.1. In the second simulation study, biological similarities are taken into account by considering individual GO groups. The setup and the results are presented in Chapter 7.3.2. The simulation studies to investigate the empirical Bayes approach are presented in Chapter 7.4. The general setup of the simulation studies is explained there as well. In the three Chapters 7.4.1, 7.4.2, and 7.4.3, results from the three specific simulation studies that are based on different underlying datasets are shown. The datasets are increasingly similar to the real VPA dataset while moving further away from fulfilling all assumptions.

Finally, the methods are applied to the real VPA dataset where no simulation takes place. Results are shown in Chapter 7.5, with the results for the meta-analysis, where only two GO groups from the dataset are considered, shown in Chapter 7.5.1, and the results for the empirical Bayes method shown in Chapter 7.5.2.

7.1. Descriptive analysis of the parameter distributions for a real dataset

The VPA dataset comprises gene expression results, measured for 54675 probe sets in the seven increasing concentrations $25, 150, 350, 450, 550, 800$, and 1000 µM in three replicates each, and for the negative control with six replicates. A 4pLL model is fitted to each probe set, resulting in a parameter vector $\boldsymbol{\phi} = (\phi^{(b)}, \phi^{(c)}, \phi^{(d)}, \phi^{(e)^*})$ for each probe set. Additionally, p-values are calculated for each probe set making use of the MCP-Mod procedure. The set of candidate models consists only of the `sigEmax` model with chosen priors of 450 as EC50 and $h = 5.117$, analogously to Chapter 6.3. Both an upward and a downward profile are assessed, leading to two p-values per probe set.

Out of the 54675 probe sets in the dataset, no model fit can be achieved for 29 probe sets that are removed for further analysis, leaving 54646 probe sets in the analysis. Calculation of MCP-Mod based p-values is possible for each of the probe sets. Results are first presented univariately for each parameter, and in a second step, bivariate relationships are analysed.

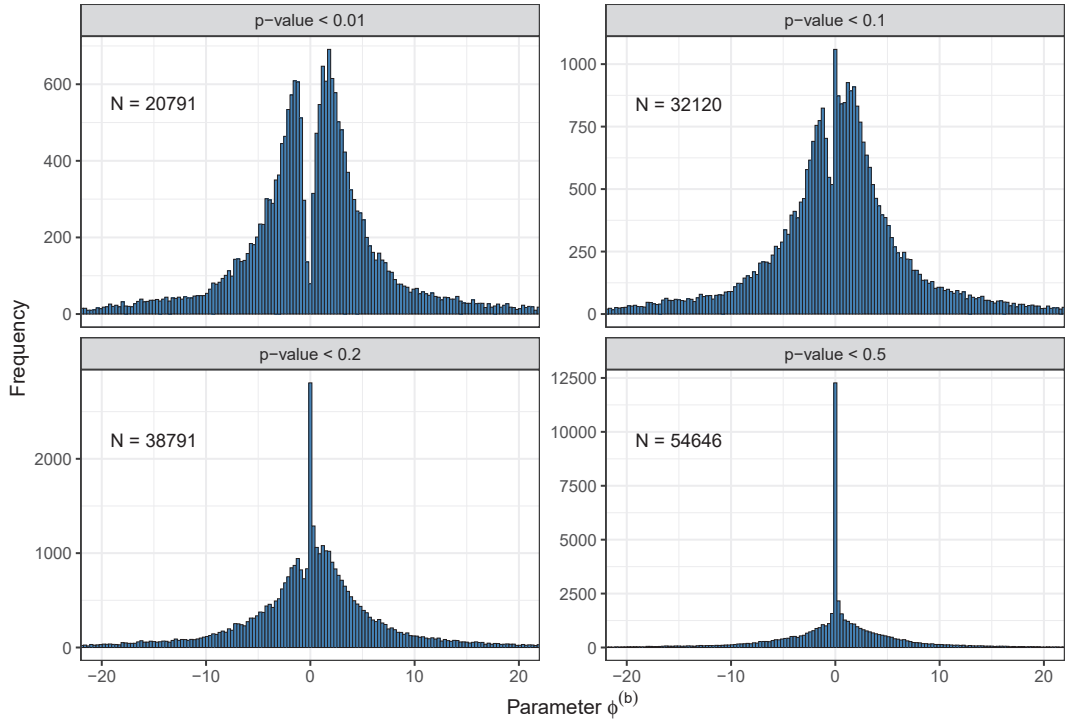

Figure 7.1: Histograms of $\phi^{(b)}$ for 4 different cutoffs of the MCP-Mod based p-values.

First, the parameter $\phi^{(b)}$ is considered. This parameter indicates the slope of the curve, but is only proportional to the actual value of the slope. Still, a value of $\phi^{(b)}$ close to zero indicates a curve that is not very steep, and therefore usually covers only a small interval of expression values in the range of considered concentrations. Histograms in Figure 7.1 show the distribution of the parameter $\phi^{(b)}$ for all probesets where the p-value for the upward or the downward direction is smaller than a specific cutoff. The cutoffs are chosen as 0.01, 0.1, 0.2 and 0.5.

For the small cutoff of 0.01, 20791 probe sets yield a significant result, for a cutoff of 0.1, 32120 probe sets yield a significant results, and 38791 probe sets yield a significant results for a cutoff of 0.2. All probe sets yield a p-value smaller than 0.5 for increasing or decreasing profiles. The striking property of the distribution of $\phi^{(b)}$ in the case of the smallest cutoff is the crater around the value of 0. This indicates that this value is only very rarely observed in probe sets for which a PoC at level 0.01 is established. For increasing p-values, however, the number of probe sets with $\phi^{(b)} \approx 0$ drastically increases, up to the point where in the histogram, the other observations seem negligible.

Figure B.45 in Appendix B.3 shows histograms of parameter $\phi^{(b)}$ for all probe sets, where only the test aimed at finding descending profiles (left) or aimed at finding increasing profiles (right) yields p-values smaller than the cutoffs 0.01 or 0.2, respectively. For basically all probe sets corresponding to decreasing profiles, it holds $\phi^{(b)} > 0$ and for basically all probe sets corresponding to increasing profiles, it holds $\phi^{(b)} < 0$, although a

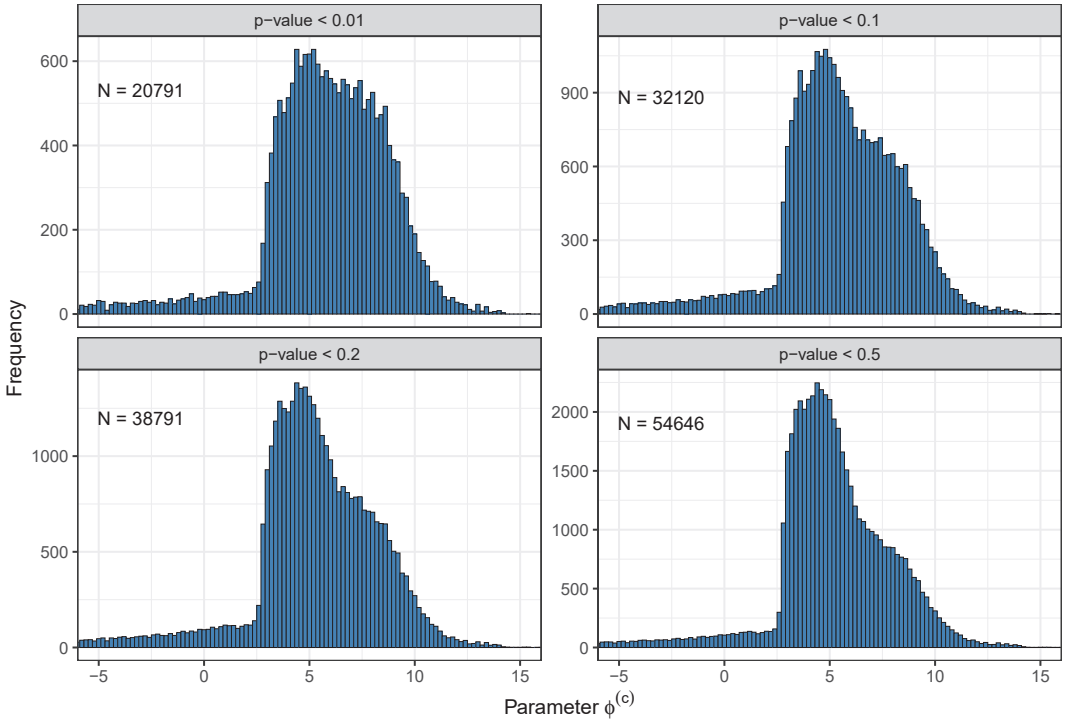

Figure 7.2: Histograms of $\phi^{(c)}$ for 4 different cutoffs of the MCP-Mod based p-values.

few exceptions exist. The number of probe sets with $\phi^{(b)} \approx 0$ is comparable for increasing and decreasing profiles when considering a p-value of 0.2.

Corresponding histograms for parameter $\phi^{(c)}$ are shown in Figure 7.2 and Figure B.46 in Appendix B.3. Most values can be observed in a range between 3 and 12. Differences between the four cutoff values for the p-values can mainly be seen in the comparison of values smaller than 6 and larger than 6: For low p-values, similar numbers can be observed. For increasing p-values, the ratio of probe sets with $\phi^{(c)} < 6$ increases in comparison to the ratio of probe sets with $\phi^{(c)} > 6$. Values larger than 15 are basically never observed, while many values smaller than 0 occur. Considering increasing profiles only, no or almost no negative values of the parameter $\phi^{(c)}$ are observed, depending on the chosen p-value. Negative values only occur for decreasing profiles. Apart from these observations, the distribution of parameter $\phi^{(c)}$ differs only little between probe sets with increasing and with decreasing profiles.

Histograms for parameter $\phi^{(d)}$ are shown in Figure 7.3 and Figure B.47 in Appendix B.3. Most observed values are in a range between 3 and 12, while basically no values smaller than 2 are observed, but substantially many values larger than 15. The left-sided slope of the histograms become steeper with increasing p-value. Analogously to the results for parameter $\phi^{(c)}$, large values for $\phi^{(d)}$ are only observed for increasing profiles, while basically no values larger than 15 are observed for decreasing profiles. The distributions of $\phi^{(d)}$ in the mainly observed range between 3 and 12 are very similar for increasing and decreasing profiles when considering the same cutoff values.

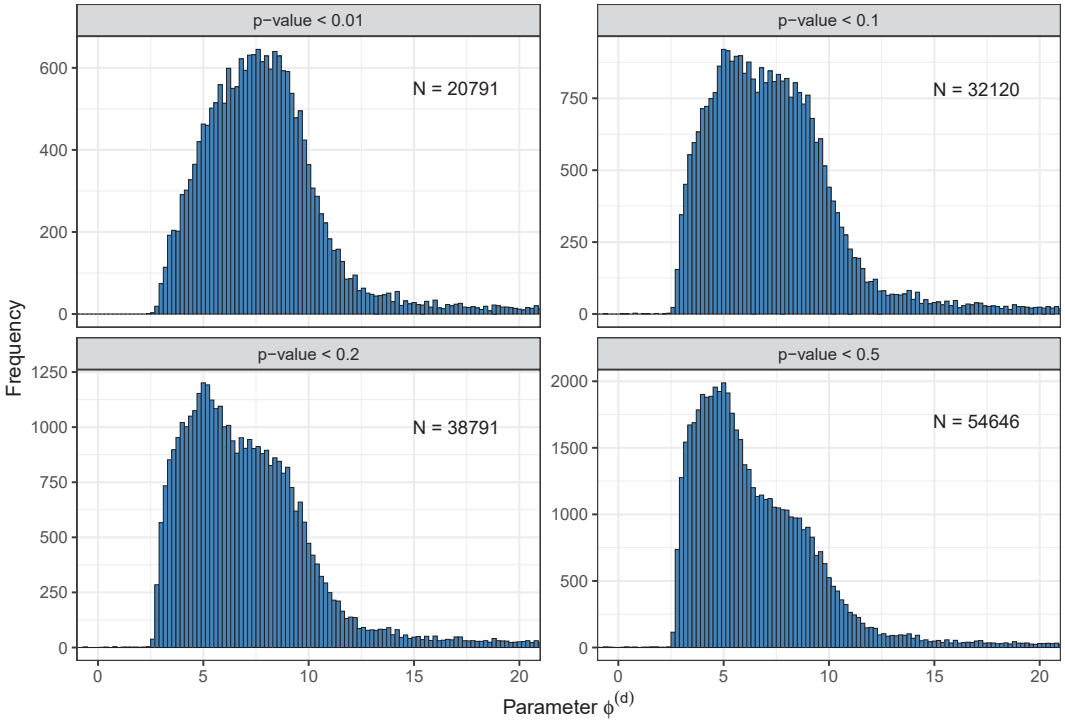

Figure 7.3: Histograms of $\phi^{(d)}$ for 4 different cutoffs of the MCP-Mod based p-values.

Finally, the distribution of parameter ${\phi^{(e)}}^*$ is shown in Figure 7.4, with the corresponding histograms of the increasing and decreasing profiles separately in Figure B.48 in Appendix B.3. The histograms for the observations divided into increasing and decreasing probe sets seem virtually identical to the ones for the entire set of probe sets, therefore no further emphasis is put on these plots.

For further analysis, ${\phi^{(e)}}^*$ is the most important parameter, as this is the parameter that indicates the alert concentration of interest for the application of information sharing considered here. Since the parameter ${\phi^{(e)}}^* = \log(\phi^{(e)})$ is considered here, the actual concentration is obtained from ${\phi^{(e)}}^*$ by applying the exponential function for back transformation. The largest concentration considered in the VPA dataset is 1000 µM. For a curve with observed half-maximal effect at this concentration, it holds ${\phi^{(e)}}^* = \log(1000) = 6.91$. Further measured concentrations and their respective values of ${\phi^{(e)}}^*$, together with multiples of the maximal tested concentration up to a factor of 5, with their respective values ${\phi^{(e)}}^*$, are summarised in Table 7.1. Values of ${\phi^{(e)}}^*$ that are larger than 10 or even 12 are observed for all four cutoffs of the p-value, corresponding to unrealistic high half-maximal effect concentrations, which are not meaningful for biological interpretation of the concentration-gene expression profiles. Most observations, however, are in the range between 3 and 10, with the lower boundary decreasing for increasing cutoffs of the p-value.

The distribution of ${\phi^{(e)}}^*$ seems to be bimodal for small cutoffs and trimodal for a cutoff of 0.5. The by far highest peak of the histograms is centred around values slightly

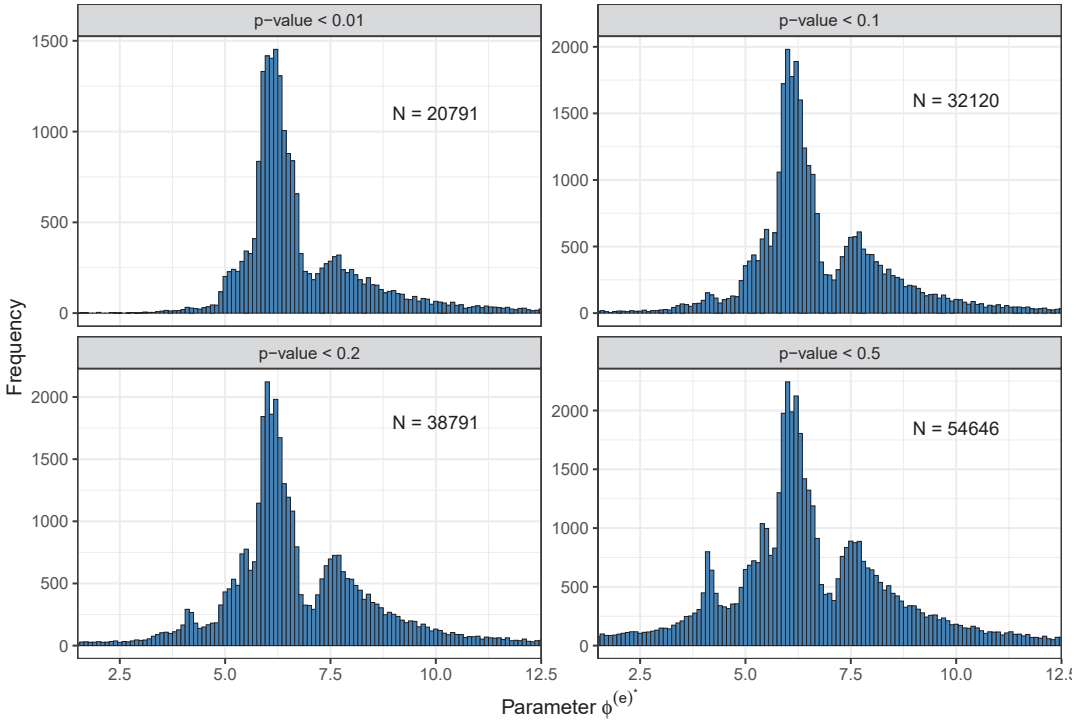

Figure 7.4: Histograms of $\phi^{(e)*}$ for 4 different cutoffs of the MCP-Mod based p-values.

larger than 6 and is followed by a low point at around 7. This low point corresponds to curves with half-maximal effect concentrations around the highest measured concentration, which apparently is only seldom observed. A second, smaller peak is observed at values slightly smaller than 8, and, only for large cutoff values, a third, narrow peak is observed at values around 4. For small cutoffs, the distribution of $\phi^{(e)*}$ observed is right-skewed, but appears more symmetric for large cutoffs.

In terms of bivariate analyses, only the relationship between $\phi^{(b)}$ and $\phi^{(e)*}$ as well as the relationship between $\phi^{(c)}$ and $\phi^{(d)}$ are examined, as no or almost no relation between the other pairs of parameters can be observed. The relations are only shown for one fixed cutoff value of 0.01, since for larger cutoffs and the resulting larger sample size, the relationships are no longer recognizable in the plots. Figure 7.5 shows the relationships between these parameters for the cutoff 0.01.

The relationship between $\phi^{(b)}$ and $\phi^{(e)*}$ is symmetric, while the specific properties of the distributions as seen in the histograms are recognizable. For very large or small values

Table 7.1: Measured concentrations of the VPA dataset and multiples of the maximal measured concentration together with the respective value of $\phi^{(e)*}$.

Conc.	25	150	350	450	550	800	1000	2000	3000	4000	5000
$\phi^{(e)*}$	3.22	5.01	5.86	6.11	6.31	6.68	6.91	7.60	8.01	8.29	8.52

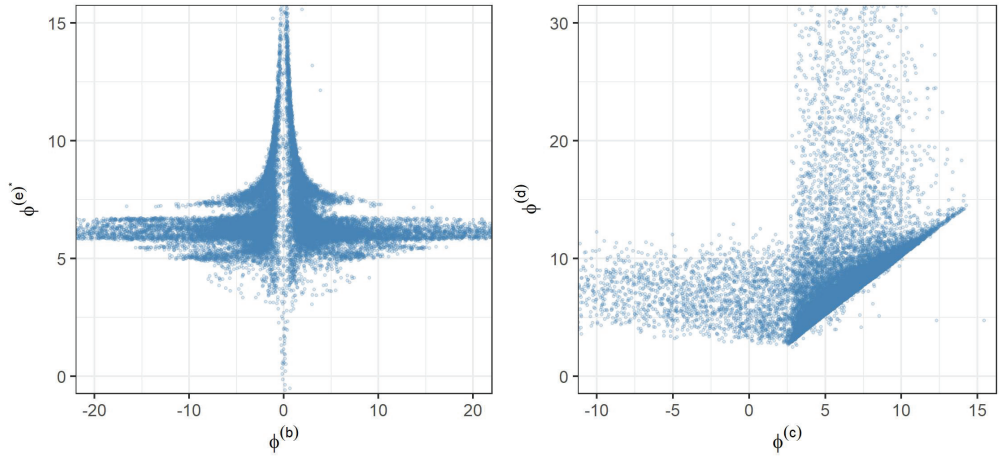

Figure 7.5: Relationships between parameters $\phi^{(b)}$ and $\phi^{(e)*}$ (left) and between $\phi^{(c)}$ and $\phi^{(d)}$ (right) for all probe sets with a p-value smaller than 0.01.

of $\phi^{(e)*}$, corresponding values of $\phi^{(b)}$ are close to 0. The largest values of $\phi^{(b)}$ correspond to values of $\phi^{(e)*}$ around 6. In the plot depicting $\phi^{(c)}$ and $\phi^{(d)}$, most observations lie in an area shaped like a triangle. Apart from very few exceptions, $\phi^{(d)}$ is larger than $\phi^{(c)}$. Extreme values of one of the parameters typically correspond to values of the other parameter that are in the range of values observed most often.

The results shown here are based on the MCP-Mod p-values, since these values allow stratification into increasing and decreasing profiles. However, the following simulation studies are based on the set of probe sets already considered in Chapter 6.3, where 9460 probe sets are selected from the entire dataset. These probe sets yield a p-value smaller than 0.001 when individually applying `anova`, without adjusting for multiplicity. Additional to this requirement for statistical significance, the biological relevance is assessed. Of the 9460 selected probe sets, only those that result in a valid ALOEC (Chapter 4.3.1), i.e. that cover a range of at least 0.585, are considered further. This leaves 7191 probe sets to be considered for further analysis. Figure 7.6 shows histograms of the distributions of the four parameters $\phi^{(b)}, \phi^{(c)}, \phi^{(d)}$ and $\phi^{(e)*}$, when fitting 4pLL models to these 7191 selected probe sets individually. The main observations about the distributions are the same as before, such that no detailed description is given here.

7.2. Descriptive analysis of the GO groups for a real dataset

Out of the 7191 probe sets that are chosen according to the criteria statistical significance and biological relevance, 5775 probe sets can be annotated to a GO group and can thus be used in the analysis. For the resulting GO graph, only groups with a minimum size of 15 are considered. The final graph contains 3807 nodes and 8360 edges between these nodes. The sizes of the GO groups are summarised in the histogram in Figure 7.7. Only GO groups with sizes up to 100 are considered there. The number of groups with a specific size decreases as the size increases, with most observations for size 15.

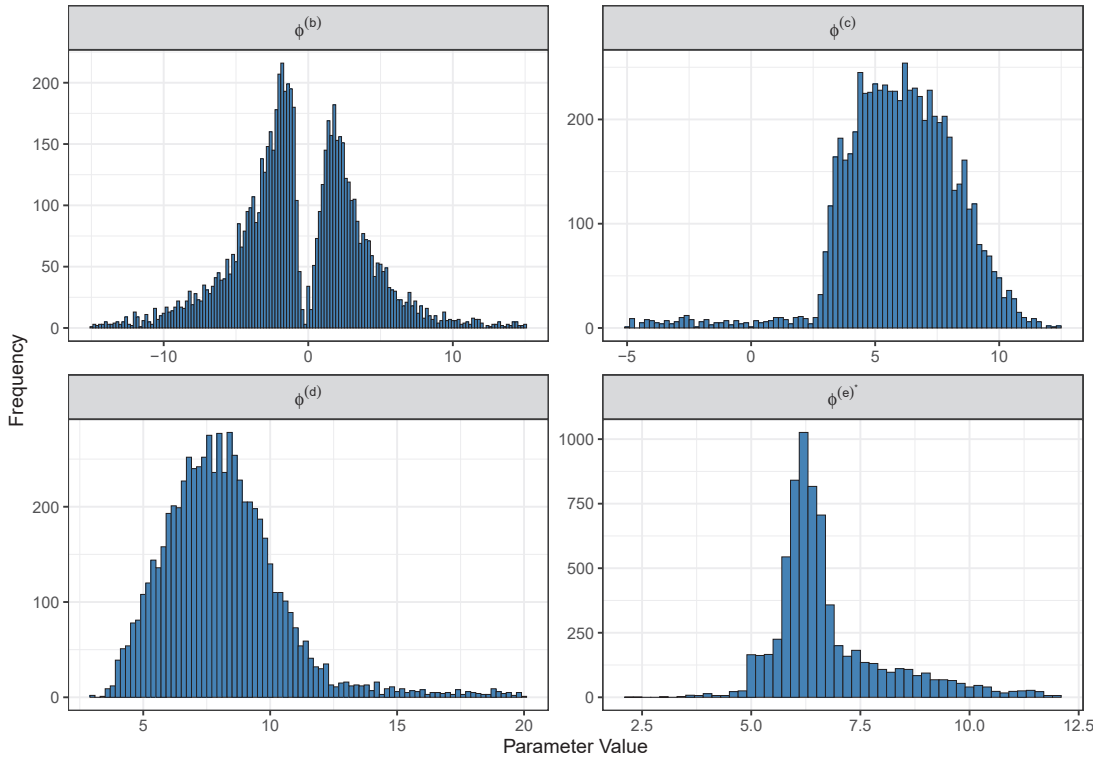

Figure 7.6: Histograms of the four parameters $\phi^{(b)}, \phi^{(c)}, \phi^{(d)}$ and $\phi^{(e)^*}$, when considering only those 7191 probesets that fulfil both the criterium of statistical significance and biological relevance.

For analyses taking biological similarities into account, only selected GO groups are considered. These GO groups are chosen from the sets of all groups containing 15 or 30 probe sets, with a total number of 56 and 147 groups, respectively. Three groups with size 15 and three groups with size 30 are chosen. They are chosen in a way that one group consists of rather similar probe sets, one group consists of probe sets that are rather different from each other and the third group lies in between these extremes, respectively.

To assess for similarity, a *correlation score* is calculated for each GO group as follows: For each probe set, concentration-wise means of the three replicates (six for the control) are calculated. Pairwise correlations, using Pearson's correlation coefficient, between the resulting vectors of mean expression values are calculated. The correlation score for the entire group is the mean value of these pairwise correlations, without considering the correlation results of each probe set with itself. Histograms of the correlation scores obtained this way are shown in Figure 7.8. For both sizes of the groups considered, positive or only slightly negative correlation scores are observed. The largest values observed are higher for groups of size 15, but the lowest values observed are similar.

For each of the sizes, three groups are chosen for the simulation study taking biological similarities into account. One group is chosen with the respective highest correlation

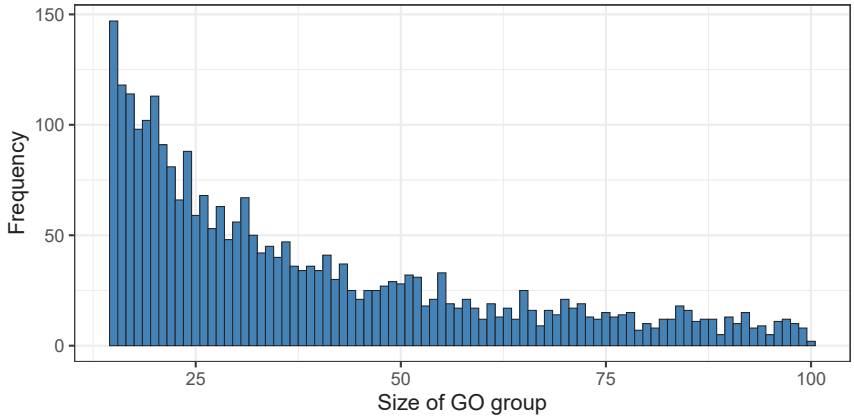

Figure 7.7: Histograms of the sizes of the GO groups when considering groups between sizes of 15 and 100 only.

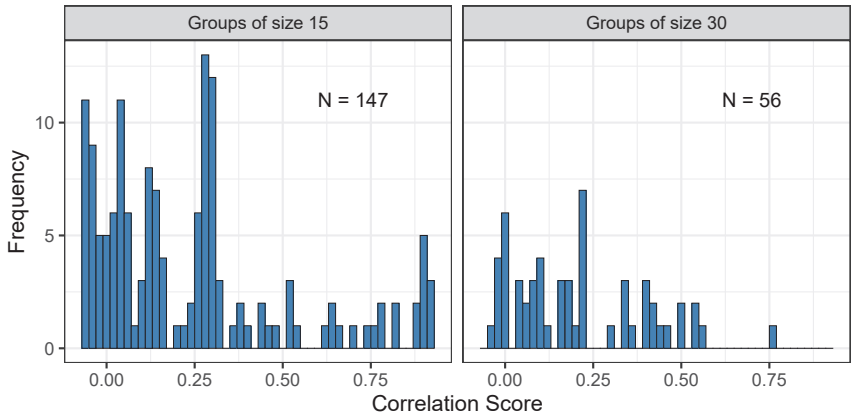

Figure 7.8: Histograms of the correlation scores for all GO groups of size 15 (left) and size 30 (right).

score, one with a medium value of the correlation score, and one group with the respective lowest correlation score is chosen. Additionally, a fourth group of size 15 and of size 30 is randomly sampled from all available probe sets, respectively. The eight groups obtained this way, together with their respective correlation score and the biological process are summarised in Table 7.2.

The specific courses of the probe sets in the selected GO groups are shown in Figures B.49 and B.50 in Appendix B.3. These plots show fitted curves for the 15 or 30 probe sets, respectively, together with the half-maximal effect concentration, corresponding to parameter $\phi^{(e)}$. This concentration is only indicated if it is in the range of concentrations considered, i.e. if $\phi^{(e)} < 1000$. Different patterns of the distribution of $\phi^{(e)}$ can be observed across the groups: sometimes, many similar values are attained and sometimes, only very few of the values are actually shown in the plot.

Table 7.2: Summary of the eight GO groups chosen for the simulation study taking biological similarities into account. The name of the chosen group, the correlation score observed for the probe sets in this group and the corresponding biological process are stated. Four groups of size 15 are chosen and four groups of size 30, whereby on group, respectively, consists of randomly sampled probe sets.

Name	Corr. Score	Biological process
GO groups of size 15		
GO:0001916	0.923	Positive regulation of T cell mediated cytotoxicity
GO:0030255	0.408	Macrophage differentiation
GO:0045601	−0.068	Regulation of endothelial cell differentiation
Random group	−0.050	–
GO groups of size 30		
GO:0034110	0.751	Regulation of homotypic cell-cell adhesion
GO:0097006	0.220	Regulation of plasma lipoprotein particle levels
GO:0048596	−0.031	Embryonic camera-type eye morphogenesis
Random group	−0.029	–

7.3. Summarising parameters using meta-analysis

Two different types of simulations study are conducted for the meta-analysis approach for information sharing. The first simulation study is based on the entire set of considered probe sets at once. Similar probe sets, where similarity is based on correlation values, are included in a meta-analysis and pooled estimates of ${\phi^{(e)}}^*$ are calculated. The second type of simulation study takes biological similarities into account. Instead of the entire set of probe sets, only subsets of probe sets that are in the same GO group are considered as potential candidates for a meta-analysis.

7.3.1. Simulation study based on an entire plasmode dataset

The entire set of 7191 probe sets, chosen as described above while taking statistical significance and biological relevance into account, is considered for this simulation study. The four parameters obtained by a first fit of a 4pLL model to each of the probe sets form the set of true underlying probe sets, called 'gene' from now on for simplicity. The following procedure is conducted for each simulation run, with 1000 simulation runs in total.

The first step is to evaluate the 4pLL models defined by the set of underlying parameters at the concentrations of the VPA dataset, i.e. $0, 25, 150, 350, 450, 550, 800$ and 1000 µM. Normally distributed noise with mean 0 and standard deviation 0.1 is added in three replicates to each of the concentrations, with exception of concentration 0, where six replicates are considered. This yields a concentration-response dataset with 27 observations for each of the 7191 genes. A 4pLL model is again fitted to the concentration-response data for each gene. All parameter values and the corresponding standard errors

are extracted, while the main interest lies on the parameter $\phi^{(e)*}$. Using the t-distribution as explained in Chapter 4.1.3, confidence intervals for $\phi^{(e)*}$ are calculated for each gene.

The next step is to determine similarities between genes via their pairwise correlation values. For each of the genes, concentration-wise means of the response values for the three replicates (or six, for concentration 0) are calculated. Pairwise correlation values between genes, using Pearsons's correlation coefficient, are determined.

A meta-analysis is calculated for each gene separately. All genes whose correlation value with the gene under consideration is at least 0.995 form the potential set of genes included in the meta-analysis. Especially the gene under consideration itself is included into this set as well. Genes, to which no model can be fitted or calculation of the standard error of $\hat{\phi^{(e)}}^*$ is impossible, are excluded from the set. In the meta-analysis, a pooled estimate of the individual values of $\hat{\phi^{(e)}}^*$ is calculated as explained in Chapter 4.4.2. Up to four confidence intervals are calculated, whereby the t-distribution-based confidence intervals can only be calculated in situations where at least two or three genes form the set considered for the meta-analysis. The number of genes included in the meta-analysis is stored as additional information. The resulting datasets summarise all the measured aspects of the simulation study for 7191 genes in 1000 simulation runs.

The first step is to consider the mean squared error (MSE) between the underlying true value of parameter $\phi^{(e)*}$ and the respective estimate, calculated directly from the model fit or calculated by the meta-analysis. For some genes, curve fitting or the calculation of standard errors of the estimated parameters is impossible in some simulation runs due to numerical problems. For 7149 out of the 7191 genes, calculation of $\phi^{(e)*}$ is possible in all simulation runs, and for the other 42 genes, at most 15 simulation runs yield no result. More missing estimates are observed for the meta-analysis, with only 5511 genes with a complete dataset and up to 305 missing values for single genes. However, this is only rarely observed, all in all 6870 genes have 20 or less missing values from all 1000 simulation runs.

For the gene-wise calculation of the MSE, only those simulation runs are considered where for both the direct estimate and the meta-analysis a result is obtained. MSEs for the direct estimate and for the meta-analysis based estimate are compared in Figure 7.9. Each dot indicates one gene, and the colouring is based on the performance of the direct estimation and the meta-analysis based estimation: Genes marked in red yield very small MSEs, smaller than 0.2, for both methods. This set consists of 3957 genes in total, i.e. more than half of the total number of genes considered. The set of blue genes consists of 1448 genes, where the MSE based on the meta-analysis is larger than or equal to the MSE from the direct estimation. For 1786 genes, marked in green, estimation of $\phi^{(e)*}$ leads to a smaller MSE for the meta-analysis than for the direct estimation.

Therefore, for about one quarter of the genes, results become worse when applying the meta-analysis in terms of the MSE, and the number of genes for which results improve is not satisfactorily higher than that. These observations are further substantiated when considering coverage probabilities (CPs) of the direct estimate in comparison to CPs obtained from the four variants of confidence intervals for the meta-analysis. Histograms of the CPs are shown in Figures B.51 (direct estimate) and B.52 (meta-analysis) in

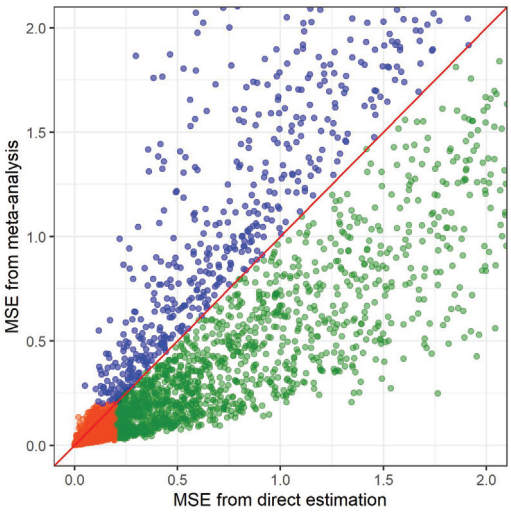

Figure 7.9: MSE of the direct estimate and the estimate based on the meta-analysis. Each dot in the plot indicates one gene, for which the MSEs are calculated based on up to 1000 simulation runs. The colouring is based on the resulting MSEs, with red dots indicating that both MSEs are smaller than 0.2, blue dots indicating that the meta-analysis based MSE yields higher results than the direct estimate and green dots vice versa.

Appendix B.3. The different number of observations for the meta-analysis based CPs stem from the requirements for the number of genes included in the meta-analysis for calculation of the t-distributions. Only genes where confidence intervals can be calculated for at least 900 simulation runs are included in the analysis.

For the 95% confidence intervals calculated here, the CPs for each gene should attain values of approximately 0.95. However, for the direct estimate, many values are smaller than 0.95, with the smallest observed values only slightly larger than 0.6. For the four variants of confidence intervals based on the meta-analysis, results are even worse, with peaks of the histograms around 0.4 and far more low than high values. Direct comparisons of the CPs from the direct estimate and the four variants based on the meta-analysis are shown in Figure 7.10. Again, only those genes are considered where confidence intervals can be calculated for at least 900 simulation runs. Only those simulation runs are considered where confidence intervals for both compared methods, respectively, can be determined.

For none of the four methods does the CP become better for any gene when using the meta-analysis. Direct comparisons between the four methods need to be interpreted with caution since the number of genes that are included in these plots differ vastly. Confidence intervals based on the t-distribution can only be calculated for genes where at least 2 or 3 genes are included in the meta-analysis, whereas the confidence interval based on the normal distribution can also be calculated when only one gene is in the meta-analysis. These genes, where no gene is similar enough to be included in the meta-analysis, lead to similar confidence intervals for the direct estimate and the meta-analysis

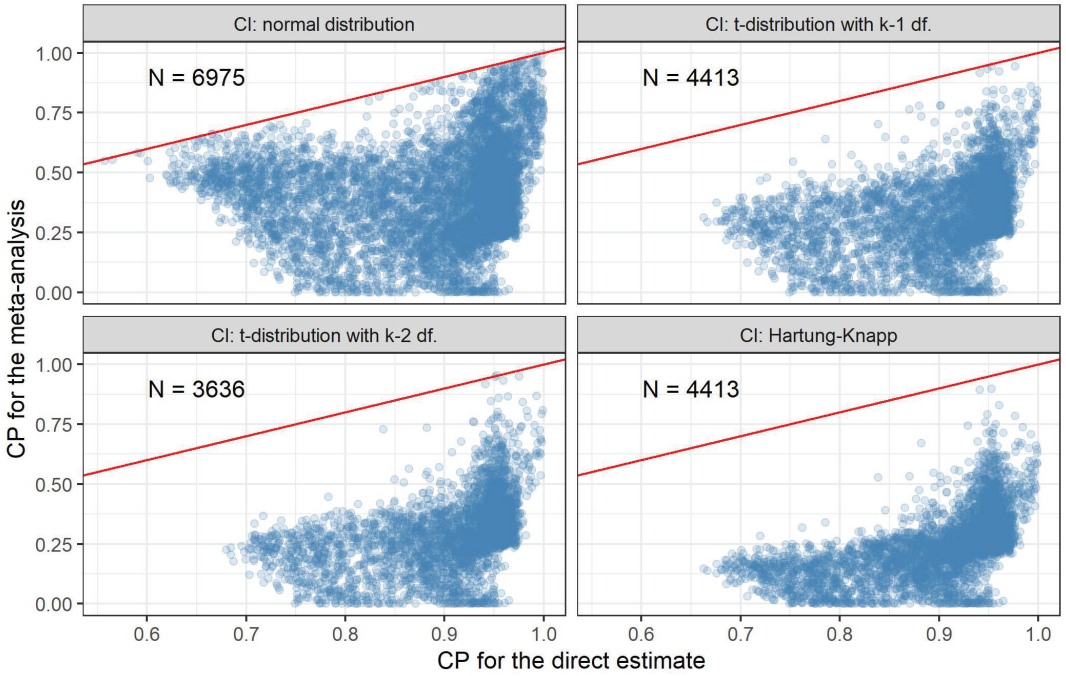

Figure 7.10: Pairwise comparison of the CP for confidence intervals calculated of the direct estimate with coverage probabilities from each of the four variants for calculating confidence intervals from the meta-analysis. Only genes, where the respective confidence interval can be calculated in at least 900 simulation runs, are considered.

estimate. This explains the apparent better performance of the normal-distribution based confidence interval in comparison to the other three confidence intervals.

In order to further demonstrate the deterioration of the meta-analysis based CPs for increasing size of the meta-analysis, for each gene, the median number of genes included in the meta-analysis across simulation runs is calculated. These median sizes are plotted against the CPs for confidence intervals from the meta-analysis, using the normal distribution. For comparison, median sizes are plotted against the coverage probabilities for confidence intervals from the direct estimation as well. Results are shown in Figure 7.11, with results for the direct estimates in the top and the meta-analysis based CPs in the bottom plot.

The distribution of the CP for the direct estimate is distributed between 1.0 and 0.6 for very small median sizes of the meta-analysis. For increasing sizes, the spread of the observed CP becomes narrower, where the values always scatter around 0.95. In particular, this means that most of the observations with a small CP are from genes for which there are only few or no similar genes. For the meta-analysis based CPs, however, a relationship to the median size can be observed: Similar to the direct estimate, observations are scattered in a range between 0.5 and 1.0 for low median sizes of the meta-analysis. For increasing median sizes, the corresponding CPs become smaller and are even approximately 0 relatively often from a median size of 50 and upwards.

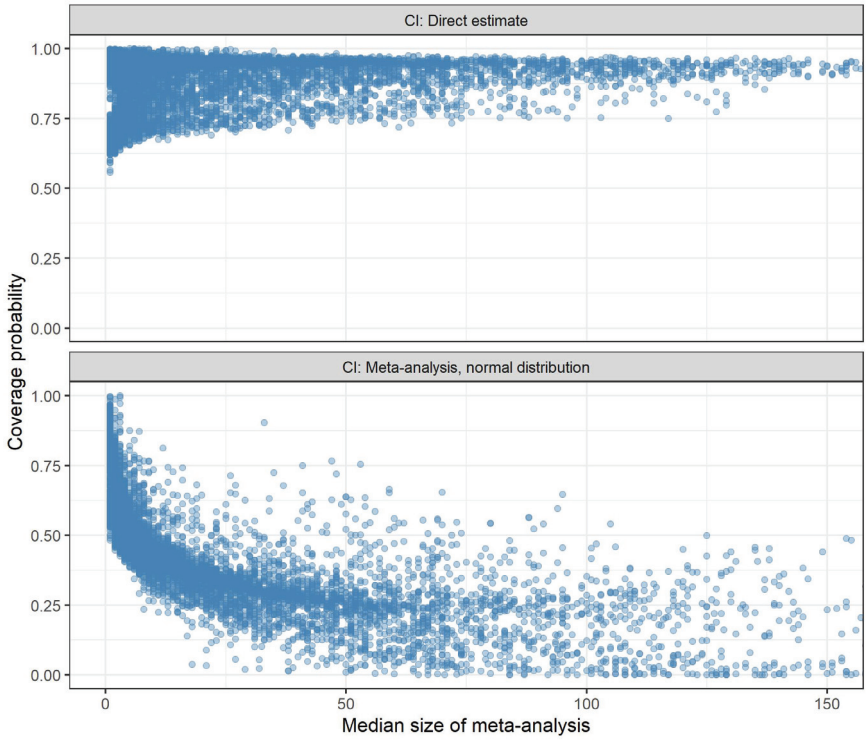

Figure 7.11: Median sizes of the meta-analyses plotted against the coverage probabilities from the directly estimated confidence interval (top) and the confidence interval based on the meta-analysis, using a normal distribution (bottom).

All in all, these analyses show that simply relying on similarities between genes without accounting for any biological reasoning behind these similarities does not improve the estimation of $\phi^{(e)*}$ when conducting a meta-analysis including all similar genes. The MSE is improved only little more often than it is worsened by conducting the meta-analysis. CPs of resulting confidence intervals for the meta-analysis are clearly lower than those for the direct estimate. A relationship can be observed between the sizes of the meta-analyses and the meta-analysis based CPs, specifically, CPs are smaller for larger sizes.

7.3.2. Simulation studies based on GO groups

Based on the observations from the previous section, biological similarities between genes are taken into account when looking for potential similar genes. Instead of considering the entire set of 7191 genes at once, only specific GO groups are selected. Potential similar genes to be included into the meta-analyses thus are only selected from the genes annotated to the same GO group as the gene under consideration.

In this simulation study, again, similarity between genes is determined by their respective correlation. Eight GO groups previously introduced (Chapter 7.2) are considered: Four

contain 15 genes each and four contain 30 genes each. The four GO groups of one size are divided into one group with a high correlation score, one group with a medium value of the correlation score and one group with a low correlation score. Additionally, one fictional group of the respective size is randomly sampled from all 7191 genes. The simulation study is conducted as explained in the previous section for each GO group separately. The main difference is the much smaller set of genes that are potential similar genes to be included into the meta-analysis. Furthermore, the cutoff of the pairwise correlation that needs to be exceeded in order to include a gene into the meta-analysis is varied between 0.915 and 0.995 in steps of 0.2.

The MSE for each individual gene is considered as the first step. This is assessed first for a fixed correlation cutoff of 0.955. Only those simulation runs are included in the calculation of the MSE where the direct estimate and the meta-analysis approach both yield a valid result. No genes are excluded from the analysis here. The MSEs of the respective 15 genes in the four chosen GO groups of size 15 are shown in Figure 7.12 and the MSEs of the respective 30 genes in the four chosen GO groups of size 30 in Figure 7.13. In the group of size 15 with a high correlation score, the values of the MSE are smaller when conducting the meta-analysis. For medium and small correlation scores, four out of the 15 probe sets considered lead to a (much) larger MSE when conducting the meta-analysis, respectively, but the MSE is very small for both methods for notably many probe sets. Assessment of the MSEs in the randomly sampled GO group is made difficult by one probe set leading to very large values, but for the other probe sets, the meta-analytis approach actually performs better than the direct estimation.

For all four GO groups of size 30, however, the MSEs are smaller when conducting a meta-analysis for several probe sets. Only for some probe sets with rather small MSE for the direct estimate, this MSE is improved by the meta-analysis approach. In the randomly sampled GO group of size 30, the MSE is improved for many probe sets by the meta-analysis method. The performance of the respective methods in the eight groups considered does not follow a clear pattern: While for the groups of size 15, the meta-analytis method performs comparatively best in the group with the highest correlation score, it also performs better in the randomly sampled group than in the two groups with medium and small correlation score. Also for the groups of size 30, the meta-analysis leads to the most improved results in the randomly sampled group.

While these MSEs are calculated for a fixed value of the correlation cutoff, in the next step, the MSEs are calculated for varied values of the correlation cutoff. The MSE for the direct estimate is not directly influenced by the change in the cutoff: Only in the scenario where a higher cutoff leads to more missing values for the meta-analysis in more simulation runs, the MSE for the direct estimate also minimally changes. However, these changes are so small that they are neglected here. Thus, only the changes in the MSE for the meta-analysis approach for different correlation cutoffs to be exceeded are illustrated. These resulting MSEs for all four GO groups of size 15 are shown in Figure 7.14. The corresponding plot for all GO groups of size 30 is shown in Figure B.53 in Appendix B.3.

In the GO group with high correlation score, the MSE becomes larger with increasing cutoff values, i.e. the more genes are included in the meta-analysis, the better the resulting MSE. For the GO groups with medium and small correlation score, the reverse

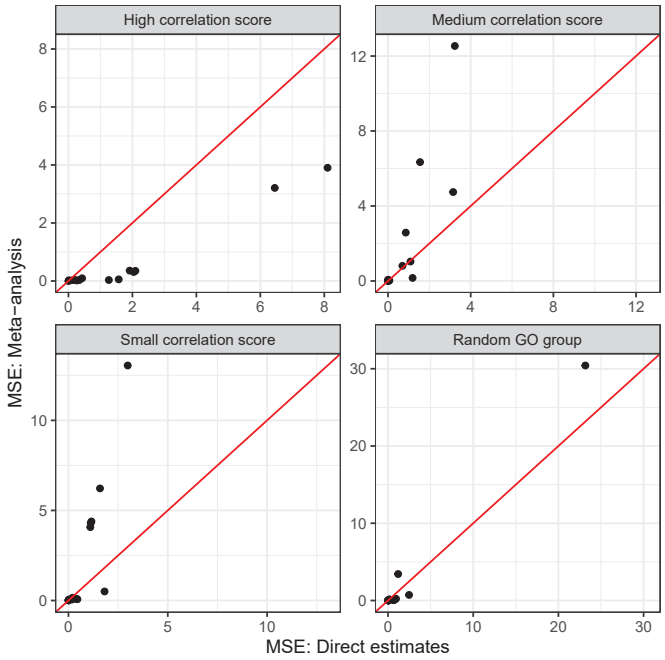

Figure 7.12: Comparison of MSEs obtained from the direct estimate and the meta-analysis estimate for each probe set in all four GO groups of size 15 considered.

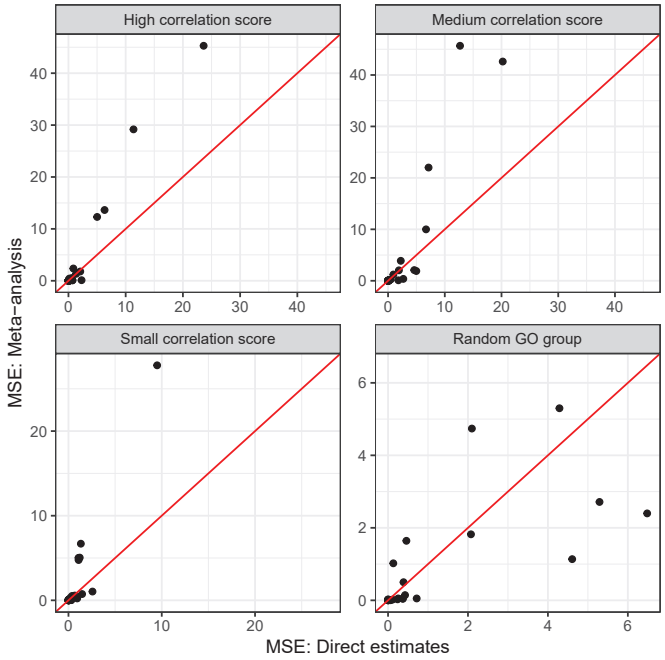

Figure 7.13: Comparison of MSEs obtained from the direct estimate and the meta-analysis estimate for each probe set in all four GO groups of size 30 considered.

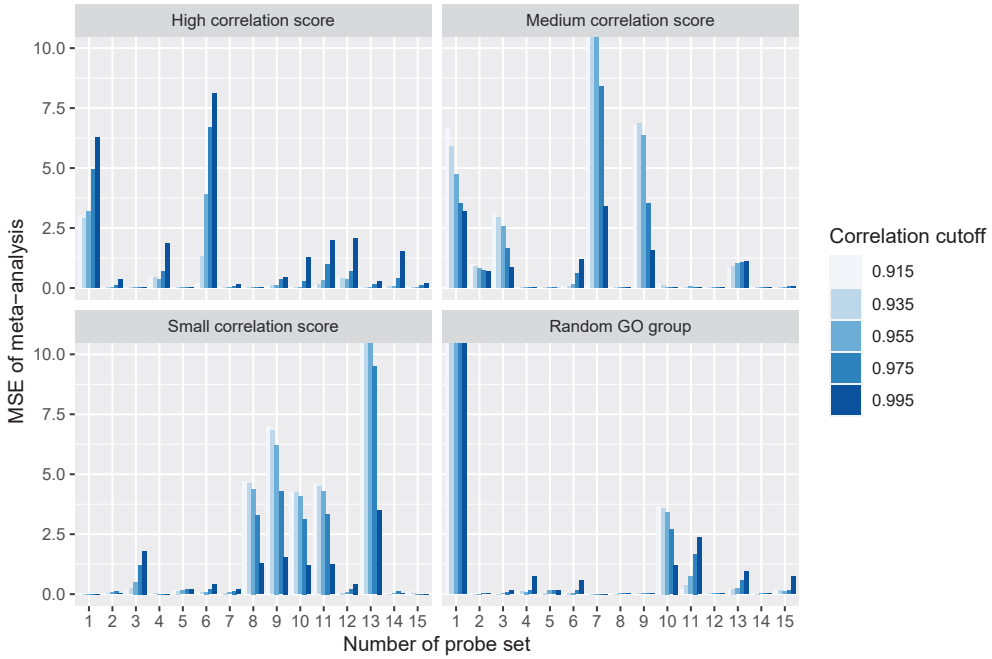

Figure 7.14: Different MSEs of the meta-analysis method for the four GO groups of size 15 when changing the correlation cutoff to be exceeded. The colour of the respective bar indicates the correlation cutoff to be exceeded and the height of the bar the resulting MSE for the meta-analysis approach.

relationship can be observed for most genes. However, especially for the genes leading to a small MSE, the same observation as for the group with high correlation score holds. For the GO groups of size 30, no clear pattern regarding the improvement or worsening of the MSE can be observed.

The next step is the analysis of the CP. For the meta-analysis, again four different types of confidence intervals can be calculated. Three of these types depend on sizes of meta-analyses of at least 2 or 3, leading to many missing values. Furthermore, the results are very similar to the results for the confidence intervals based on the normal distribution, thus, only these results are shown here. The results for the four GO groups of size 15 are summarised in Figure 7.15. Results for the GO groups of size 30 are very similar and not explicitly shown here. While the CPs for the confidence intervals based on the direct estimate are mostly in a reasonable range between 0.8 and 1.0, very small values down to a probability of 0 can be observed for the meta-analysis. Thus, also when taking biological considerations into account, the meta-analysis method performs far worse than the direct estimate in terms of CP.

The final step is to directly consider the sizes of the meta-analyses and to set them into relation with the MSEs from the meta-analysis approach. The size is the number of genes that are included in the meta-analysis. This is done only for the GO groups of size 15 with a medium and a high correlation score since the group with a small score and the randomly sampled group show far smaller sizes of the meta-analysis overall.

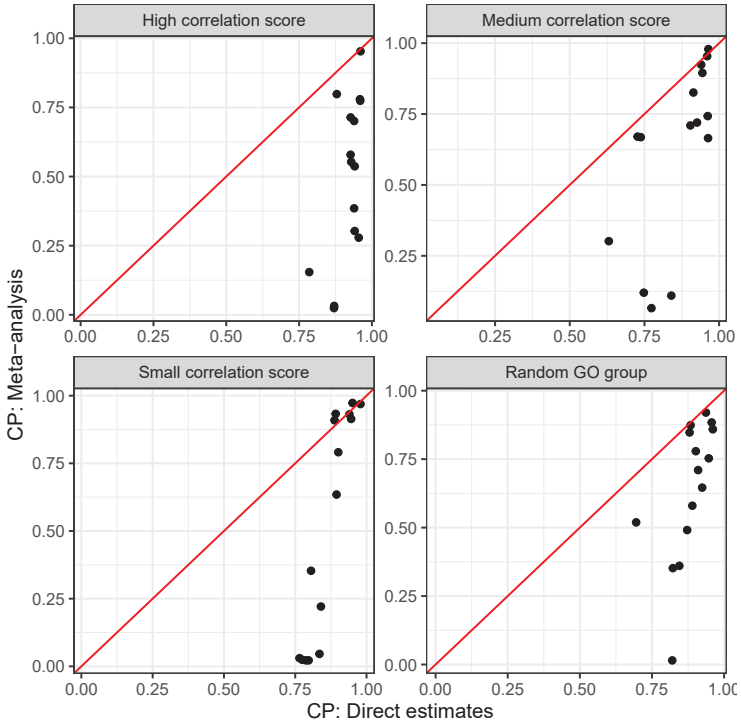

Figure 7.15: Comparison of the coverage probabilities for the confidence interval based on the direct estimate and the normal-distribution based confidence interval for the GO groups of size 15.

For a fixed correlation cutoff of 0.955, the gene-wise sizes of the meta-analyses in all simulation runs considered are shown by boxplots in Figure 7.16. Colours of the boxes indicate the corresponding value of the MSE.

A direct result from the choice of GO groups is the larger size of meta-analyses in the group with high correlation score. In this group, for most genes the median size of meta-analyses across all simulation runs is 8 or higher. However, the range of observed sizes is very broad for most genes, indicating some simulation runs where almost no genes are included in the meta-analysis and some simulation runs where almost all genes are included. The exception from this observation is the eighth gene, where the resulting meta-analyses always only consist of this gene itself. Except for two genes with comparatively low sizes and the already discussed eighth gene, the MSEs are low. Only for the two genes with smaller sizes of the meta-analyses, the MSE is larger. This coincides with the observations from Figure 7.14, where a smaller cutoff, i.e. more genes in the meta-analysis, also coincides with a smaller MSE.

For the GO group with median correlation score, the median sizes of the meta-analyses are consistently smaller than 5, and the observed range of sizes per gene is smaller overall. No clear pattern in the relationship between size of meta-analysis and MSE can be observed with some gene, e.g. the eighth and the 15th, leading to a small MSE

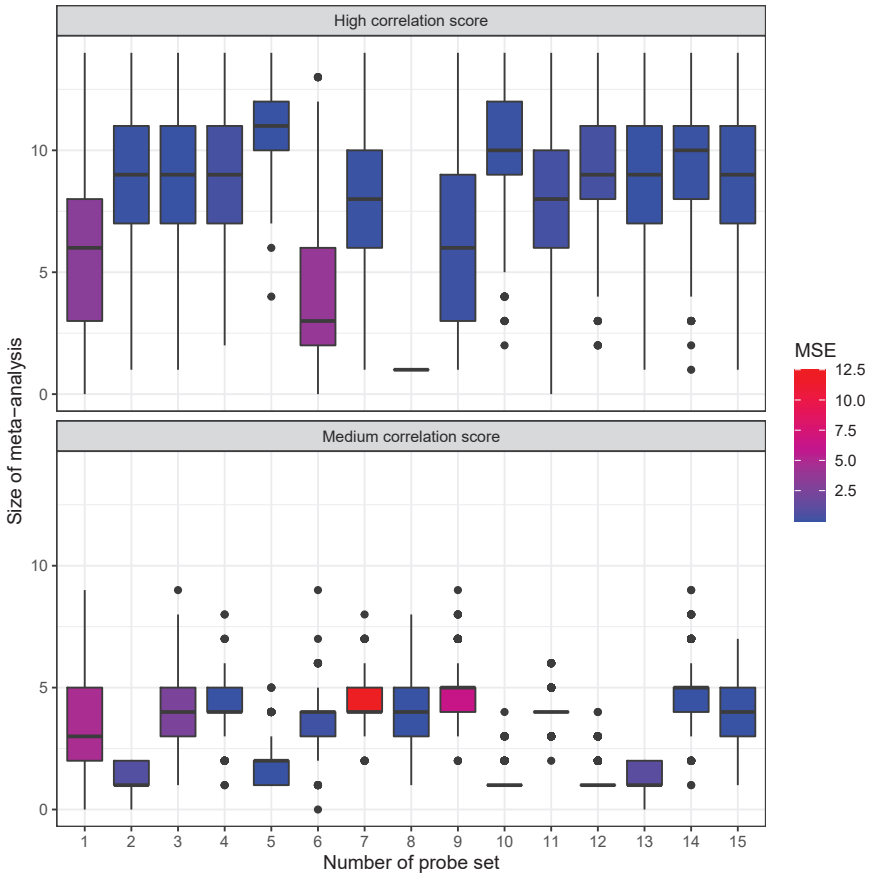

Figure 7.16: Boxplots of the sizes of the meta-analysis for each probe set individually for the GO groups of size 15 with high (top) and medium (bottom) correlation score. The colour of the boxes indicate the value of the MSE for the respective gene when using the meta-analysis method. Note that the genes in both groups are not the same.

and other gene, e.g. the third or the ninth leading to much larger values despite similar location and size of the boxes.

To summarise, taking biological similarities into account does not reliably lead to better results. While the MSE is indeed improved in some of the chosen GO groups when conducting the meta-analysis, it is not always clear from the choice of GO group, why this is the case. For some groups, increasing the size of meta-analyses is beneficial regarding the resulting MSE, in other groups, it is disadvantageous. This leads to the presumption that improvement of the MSE is randomly obtained and not based on biological features of the genes. The CP is far worse for almost all genes considered in comparison to the CP based on the confidence intervals obtained from the direct estimate. Thus, biological similarities defined by GO groups are not sufficient to base information sharing using meta-analysis on in order to improve results for the estimation of the parameter $\phi^{(e)*}$.

7.4. Shrinkage of parameters using an empirical Bayes method

For the empirical Bayes method for information sharing across genes, three simulation studies are conducted. The methodology is equal for all studies, the only difference is the underlying set of genes on which the simulated datasets are based. First, a completely synthetic dataset is used, then a normalised version of the VPA dataset is used, and finally, the simulation study is conducted for the true VPA dataset. In the following, for simplicity, probe sets are called 'genes' again. More details about the format of the dataset are given in the respective section.

The increasingly realistic choice of datasets is conducted to obtain results both in situations where all assumptions are fulfilled and in situations that are closest to reality. The main assumption is given by the normal distribution of parameter $\phi^{(e)*}$. To recall the distribution of this parameter, Figure 7.17 shows a histogram of the respective parameter values. While again the set of 7191 genes that are both statistically significant and biologically relevant are the basis of this histogram, only those genes are considered where $\phi^{(e)*} > 0$ and $\phi^{(e)*} < \log(2000) = 7.6$. Thus, genes with inflection points at unreasonably high or low concentrations are excluded from this display, leaving 5719 genes. Additionally, curves for estimated normal distributions are indicated in the plot. The red curve corresponds to the normal distribution estimated using the maximum-likelihood (ML) approach. The blue curve corresponds to the normal distribution estimated using the robust approach, making use of the median and the MAD (Chapter 4.4.3).

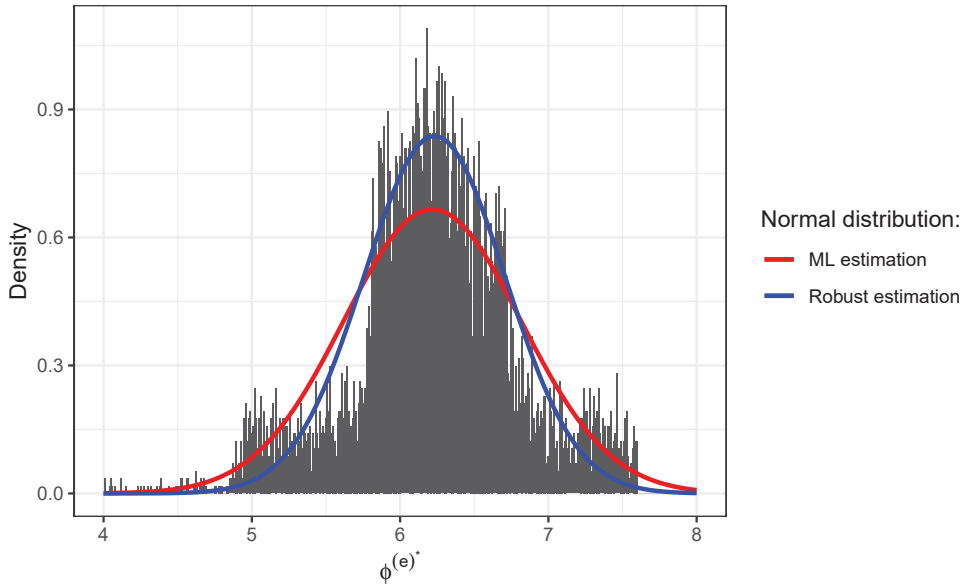

Figure 7.17: Histogram of the values of parameter $\phi^{(e)*}$ where $\phi^{(e)*} > 0$, truncated at $\log(2000) = 7.6$ for genes that are statistically significant and biologically relevant. The density of estimated normal distributions are added for the ML approach (red) and the robust approach (blue).

It can be observed that the symmetry in the histogram corresponds to the symmetry of the estimated normal distributions. The density for the mean value, however, is much too small for the ML estimation and still a little too small for the robust estimation. The tails of the actual distribution as seen in the histogram and the estimated normal distributions do not fit well. In the range closer to the mean, the density of the observed values is too low, and for values of $\phi^{(e)^*}$ further away from the mean, the density of the observed values is too high.

The set of genes considered in the respective simulation study is given in the form of the four parameters of the 4pLL model. The simulation is conducted as follows: For each simulation run, separately, the first step is to simulate a concentration-response dataset. This is achieved by evaluating a 4pLL function with the parameters of each gene at the concentrations of the VPA gene-expression study, i.e. $0, 25, 150, 350, 450, 550, 800,$ and 1000 µM. Normally distributed noise with mean 0 and standard deviation 0.1 is added to each of the concentrations in three replicates, with exception of the control, where six replicates are added. This leads to a dataset with 27 observations for each of the genes considered.

The next step is to fit a 4pLL model to each gene. All parameter values together with corresponding standard errors are extracted, with main interest in the parameter $\phi^{(e)^*}$. Prior mean and prior variance for the empirical Bayes procedure are estimated directly from the dataset. This is done in two variants: In the first variant, the prior mean and variance are calculated as sample mean and sample variance of the estimates of $\phi^{(e)^*}$, i.e. with the ML approach, for all genes in one simulation run. In the second variant, instead of sample mean and sample variance, the median and the squared MAD are used. Note that whenever values of the MAD are shown, these are the values after multiplication with the factor 1.4826 as defined in Chapter 4.4.3. The posterior means and posterior variances are calculated for each gene individually, making use of the prior values as explained above, the specific estimate of $\phi^{(e)^*}$, and its standard error.

This procedure is conducted for 1000 simulation runs. The main goal of the simulation study is a small MSE, where the estimates of the 1000 simulation run for the direct estimation and the Bayesian estimation are compared to the true underlying value of $\phi^{(e)^*}$ for each gene. As a second goal, the CPs of the direct confidence interval and the credible interval from the Bayesian approach are compared.

7.4.1. Simulation study based on a synthetic dataset

The first simulation study is based on a completely synthetic dataset, i.e. the four parameters determining a 4pLL model are randomly sampled for each gene individually. All in all, 5000 genes are simulated, while retaining the basic features of the distributions of the parameters from the VPA dataset.

The parameter $\phi^{(b)}$ is simulated using a normal distribution with mean 0 and standard deviation 4. To account for the fact that only genes with a notable slope shall be included in the simulation studies, all observations of $\phi^{(b)}$ in the range from -0.2 to 0.2 are re-sampled according to the same normal distribution. In the bivariate analysis from

Chapter 7.1, a correlation between parameters $\phi^{(c)}$ and $\phi^{(d)}$ can be observed. Therefore, these parameters are simulated from a bivariate normal distribution with mean values 5 and 8 and the following covariance matrix:

$$\begin{pmatrix} 5 & 2.7 \\ 2.7 & 4 \end{pmatrix}$$

These values are based on the covariance between parameters $\phi^{(c)}$ and $\phi^{(d)}$ as observed in the real dataset. Finally, the parameter $\phi^{(e)*}$ is simulated using a normal distribution with mean 6.2 and standard deviation 0.9.

Histograms showing the distributions of the four parameters are shown in Figure B.54 in Appendix B.3. The relationship between parameters $\phi^{(c)}$ and $\phi^{(d)}$ is shown in Figure B.55, where the gene-wise observations are plotted against each other. It can be observed that, although the triangle shape is not reproduced here, in the vast majority of cases, $\phi^{(d)}$ is larger than $\phi^{(c)}$, analogously to the original VPA dataset.

First, the priors from the Bayes analysis, using both the ML and the robust estimates, are assessed. Histograms of the priors observed in all 1000 simulation runs are shown in Figure 7.18. The distribution of the robust estimated prior parameters (bottom) is much narrower than the distribution of the ML estimations (top), especially for the measures of dispersions. Empirical mean priors mostly take values between 6.0 and 6.6, corresponding to concentrations of approximately 400 and 735. The empirical prior MAD always takes a value around 1, while the smallest observed empirical prior standard deviations start at 2 and values even larger than 10 are observed.

Since estimation of the posterior distributions depends on the standard error of the direct estimation, the posterior distribution can only be calculated in the case that this standard error can be calculated. A total of 523 genes for which estimation of the posterior distribution is not possible in more than 200 simulation runs are excluded from further analysis, leaving 4477 genes in the analysis. For calculation of MSE and CPs, only those simulation runs are taken into account where an estimate is obtained for the direct estimation and the Bayesian estimations.

The next step is to examine the MSE across all 1000 simulation runs for the 4477 remaining simulated genes. These MSEs are calculated once for the estimation stemming directly from the fitted curve and once for the Bayesian estimates based on ML estimation. The comparison of these MSEs is shown in the left plot of Figure 7.19. The colours in the plot are chosen as follows: If the MSE from the Bayesian estimation is smaller than the MSE from the direct estimation, divided by 1.1, the corresponding dot is coloured green. If the Bayesian MSE is larger than the direct MSE multiplied by 1.1, the colour is black. The dots in between these margins are blue. Additionally, all dots where both MSEs are smaller than 0.2, i.e. where the MSEs are essentially negligible, are coloured in red.

All in all, only 2 dots are coloured black, 38 dots are coloured blue, 1530 dots are coloured green and the remaining 2907 dots are coloured red. This means that estimation of $\phi^{(e)*}$ is improved by using the Bayes method with ML estimation for 1530 genes, while results

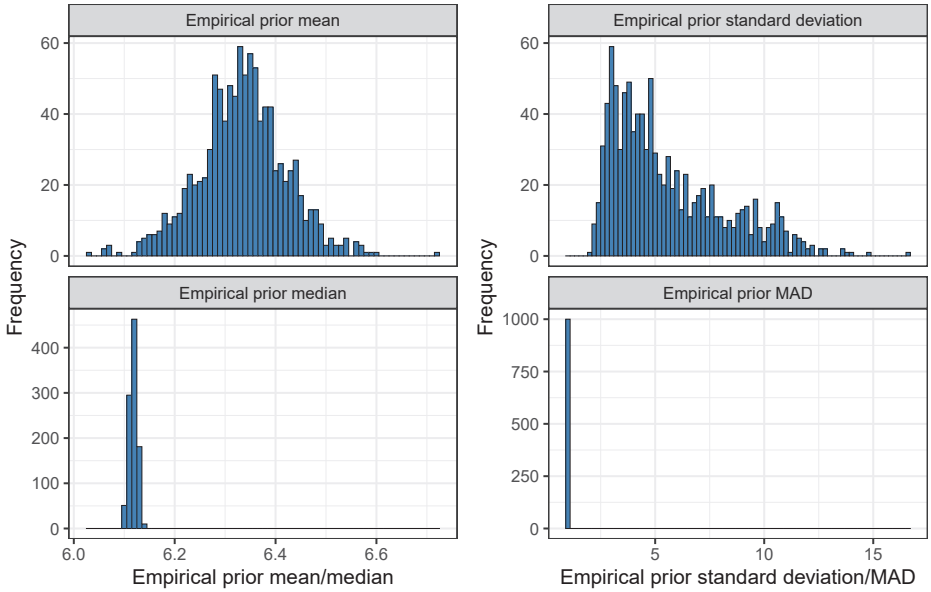

Figure 7.18: Histogram of the prior estimates for the empirical Bayes method to estimate the parameter $\phi^{(e)*}$. Results for ML estimation are shown at the top, results for robust estimation at the bottom of the plot.

are worse only for 2 genes. To find patterns in the entire set of genes, indicating whether an improvement of the fit can be achieved, the true underlying parameters $\phi^{(b)}$ and $\phi^{(e)*}$ are plotted against each other, shown in the right plot of Figure 7.19. Genes with improved estimation when using the Bayes methodology have very small values of $\phi^{(b)}$ or rather large values of $\phi^{(e)*}$. The blue dots, indicating genes where applying the Bayes method does not lead to better results than the direct estimate, can only be found for values of $\phi^{(e)*}$ in the range between 7 and 9.

The final step of the analysis is to consider the CPs of the resulting confidence or credible intervals. As for the MSE, only those simulation runs are considered where the direct estimate and the Bayes methods allow calculation of the respective interval. A histogram of the CP for the confidence intervals obtained by the direct estimate is shown in Figure 7.20 (left). Most observations scatter around 0.95, indicated by a red vertical line, which is the probability that should be achieved for the 95% confidence intervals considered. However, observations go as low as approximately 0.6.

A direct comparison of the CPs for the direct estimate and for the credible intervals obtained by the Bayes method using ML estimation is shown in the right plot of Figure 7.20. The dots plotted there are coloured as in Figure 7.19. It can be observed that the CPs are almost equal for both methods. Higher CPs are observed for genes with very small MSEs, coloured in red, and lower CPs for genes with larger MSEs, coloured in green.

Corresponding results for the Bayes method based on the robust estimation are shown in Figures B.56 and B.57 in Appendix B.3. Briefly, the robust estimation performs worse

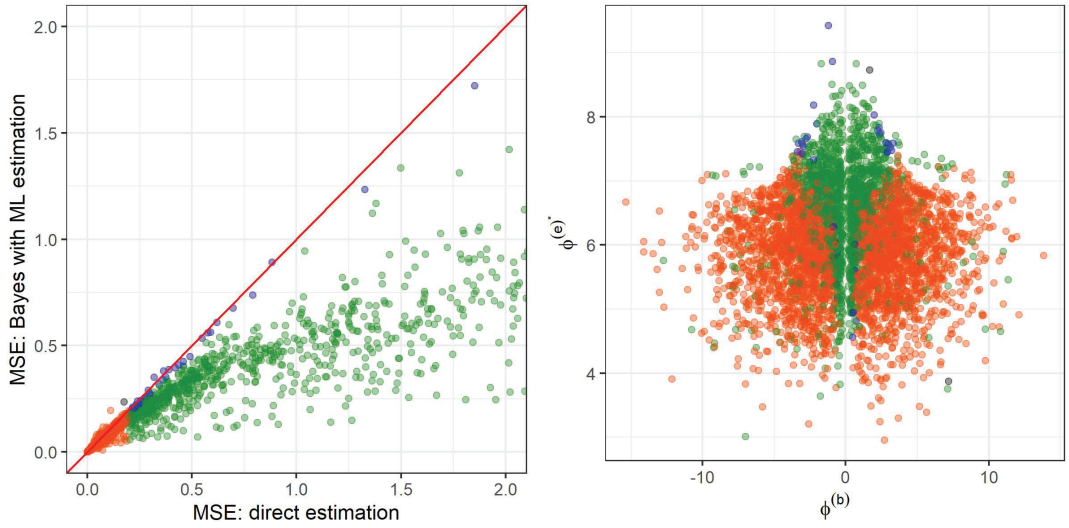

Figure 7.19: Left: Comparison of MSE for the direct and the Bayes estimate, based on ML estimation. If the Bayesian MSE is smaller than the direct MSE divided by 1.1, the corresponding dot is coloured green, and the colour is black if the Bayesian MSE is larger than the direct MSE multiplied with 1.1. The dots in between are blue. All dots where both MSEs are smaller than 0.2 are coloured in red. Right: True underlying parameters $\phi^{(b)}$ and $\phi^{(e)^*}$ plotted against each other and coloured according to the comparison of MSEs. These are the results for the synthetic dataset.

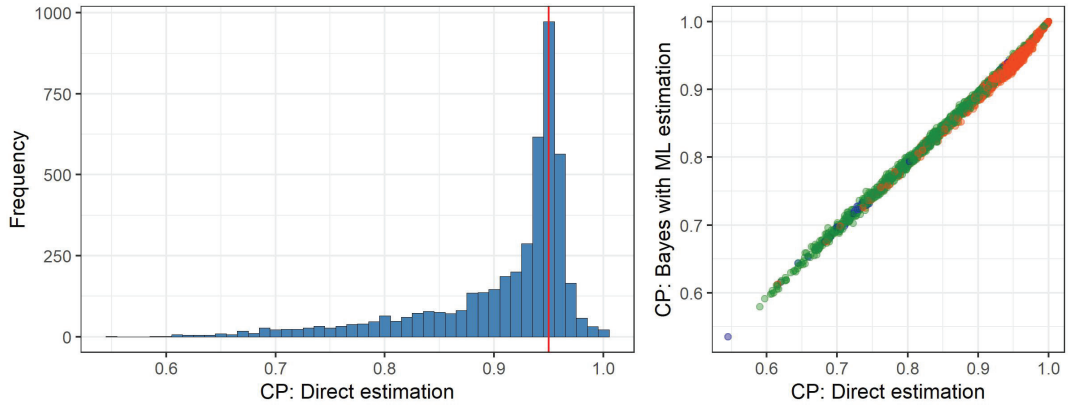

Figure 7.20: Left: Histogram showing the CPs for the confidence intervals obtained by the direct estimate. Right: Comparison of these CPs with the CPs for the credible intervals obtained by the Bayes method. Colours are the same as in Figure 7.19. These are the results for the synthetic dataset.

than the ML estimation: For 1423 genes, the MSE for the Bayes estimate is smaller than the MSE for the direct estimate, but for 181 genes, it is worse. The MSE of the Bayes method for those genes, where an improvement can be observed, is far smaller in comparison to the MSE of the direct estimate. The genes where the Bayes method

does not lead to better results are mainly those with rather large or rather small true underlying values of ${\phi^{(e)}}^*$ and corresponding values of $\phi^{(b)}$ that are not close to 0. CPs for the direct and the Bayes estimate are similar to each other, but with a larger dispersion than observed for the ML method. In few cases, the CP observed for the Bayes method is far lower, with observations at 0 even, than the one for the direct estimate.

To briefly summarise results from the simulation study for the synthetic dataset, it can be stated that in terms of MSE, the Bayes method leads to an improvement in comparison to the direct estimate. This improvement does not come at the cost of lower CPs of the credible interval. Instead, they are almost equal to the CPs of the confidence intervals obtained by the direct estimate. In general, the ML estimation of the prior distribution leads to better results than the robust estimation.

7.4.2. Simulation study based on a normalised plasmode dataset

The second simulation study is based on a set of true underlying parameters that closely resembles the VPA dataset, while still ensuring that the assumption of normality for parameter ${\phi^{(e)}}^*$ is fulfilled. The set of 7191 genes, chosen according to statistical significance and biological relevance, as introduced before, is again considered. Similar to Figure 7.17 only those genes are considered where ${\phi^{(e)}}^* < \log(2000)$. Quantile normalisation is applied to all 5723 observations of parameter ${\phi^{(e)}}^*$, leading to normalised values that follow a normal distribution with the same mean and variance as the original values. The other three parameters remain unchanged. The simulation study is conducted as described before.

Histograms of the distribution of the prior parameters, shown in Figure B.58 in Appendix B.3, are very similar to those from the synthetic dataset. As for the synthetic dataset, genes with more than 200 missing estimates are excluded from further analyses. This is true for 373 genes, leaving 5350 genes in the analysis. Again, only those simulation runs are taken into account where an estimate is obtained for both the direct estimation and the Bayesian estimations.

The first step again is the comparison of the MSE obtained by the direct estimation and by the Bayes method using the ML estimation. The MSEs are shown in Figure 7.21 (left), whereby the colours are chosen by the same rules as explained for the synthetic dataset. 25 dots are coloured in black, corresponding to genes where the Bayes method performs worse than the direct estimate. On the other hand, for 1851 genes, the Bayes method performs notably better. 42 dots are coloured in blue and the remaining 3432 dots are coloured red.

The right plot of Figure 7.21 shows the true underlying parameters $\phi^{(b)}$ and ${\phi^{(e)}}^*$ coloured according to the comparison of MSEs. Since in this simulation study the true underlying set is closer to the real datasets, some artefacts influencing the distribution of the colours can be observed. As in the previous analyses, almost all genes with a values of $\phi^{(b)}$ close to 0 are coloured in green. The entire set of genes with ${\phi^{(e)}}^* \approx 7$, approximately corresponding to the highest measured concentration 1000, is coloured green. Additionally, many genes with ${\phi^{(e)}}^* \approx 4.6$, corresponding to a concentration of 100, are

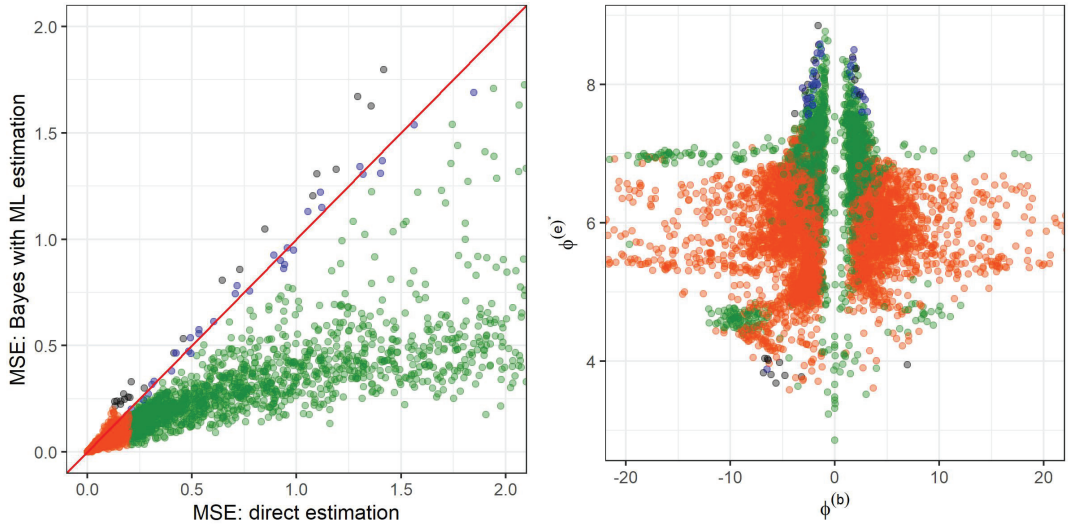

Figure 7.21: Left: Comparison of MSE for the direct estimate and the Bayes estimate based on the ML estimation. Right: True underlying parameters $\phi^{(b)}$ and $\phi^{(e)^*}$ plotted against each other and coloured according to the comparison of MSEs. These are the results for the normalised dataset.

coloured green. Black and blue dots occur for the largest and the smallest values of $\phi^{(e)^*}$ where $\phi^{(b)}$ is not close to 0.

The CPs from the confidence intervals obtained by the direct estimate and the comparison of CPs for the direct estimate and the Bayes method are shown in Figure 7.22, whereby the colours stem from Figure 7.21. Most of the observed probabilities are around 0.95, but again values as low as 0.6 are observed. The comparison shows the very similar results in terms of CPs, that also cluster as described in the previous section with regards to the colours.

Corresponding results for the Bayes estimation using the robust estimated priors are shown in Figures B.59 and B.60 in Appendix B.3. As for the synthetic dataset, the results are notably worse than for the ML estimation. Especially for genes with true underlying values of $\phi^{(e)^*} > 7$ or $\phi^{(e)^*} < 4.8$, the Bayes method often leads to worse results, in terms of MSE and CP.

7.4.3. Simulation study based on an entire plasmode dataset

The final simulation study is conducted with the original parameters obtained directly from the VPA dataset as set of underlying genes. Again, the set of 7191 genes fulfilling the criteria of statistical significance and biological relevance is considered. Only those genes with $\phi^{(e)^*} > \log(2000)$ and the four genes with $\phi^{(e)^*} < 0$ are excluded from the analysis, leaving 5719 genes in the analysis. The simulation study is conducted as described above.

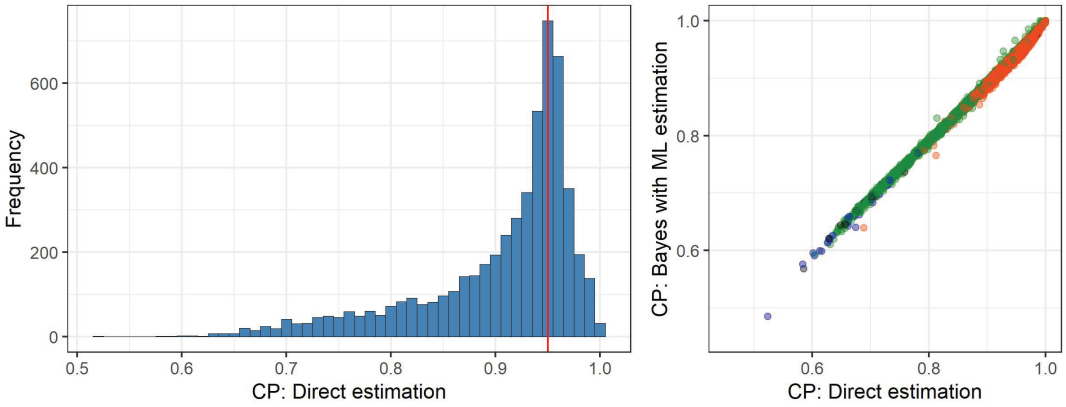

Figure 7.22: Left: Histogram showing the coverage probabilities for the confidence intervals obtained by the direct estimate. Right: Comparison of this coverage probability with the coverage probability for the credible intervals obtained by the Bayes method. Colours are the same as in Figure 7.19. These are the results for the normalised dataset.

The priors are shown in Figure B.61 in Appendix B.3. The main difference to the priors for the synthetic and the normalised dataset is that the measures of dispersion take comparatively smaller values. Only 117 genes are excluded from the analyses for this dataset due to more than 200 missing values, leaving 5602 genes in the analysis. For calculation of MSE and CPs, only those simulation runs are taken into account where an estimate is obtained for all methods.

Comparison of the MSEs is shown in Figure 7.23, together with the true underlying parameters $\phi^{(b)}$ and $\phi^{(e)^*}$, coloured as described before. For 35 genes, the Bayes method leads to a larger MSE than the direct estimate. However, also the MSE based on the Bayes method is still comparatively low for these genes. These genes correspond to low values of $\phi^{(e)^*}$ about 4 or high values of approximately 7.6. Still, for 1687 genes, the MSE is improved when conducting the Bayes procedure. In comparison, this improvement is even greater than in the two previous simulation studies. No difference can be observed for 11 genes and the remaining 3869 genes, coloured in red, lead to very low MSEs for both methods. The patterns in the data are the same as observed before.

Also the results regarding the CPs, shown in Figure 7.24 first for the direct estimate only and then for the comparison of direct estimate and Bayes method, are very similar to the ones observed before. The lowest observed CPs are even higher than for the previous simulation studies. On the other hand, for very few genes, the Bayes method leads to lower CPs than the direct estimate. The dots with the largest differences are coloured green, i.e. the Bayes method performs better than the direct estimate in terms of MSE, so that the credible intervals resulting from the Bayes method seem to be too narrow rather than structurally biased.

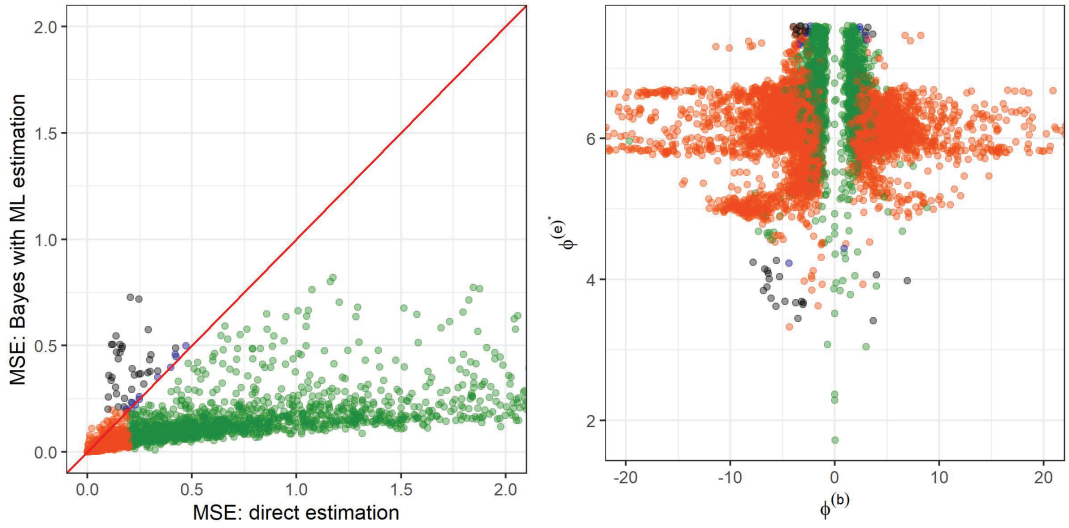

Figure 7.23: Left: Comparison of MSE for the direct estimate and the Bayes estimate based on the ML estimation. Right: True underlying parameters $\phi^{(b)}$ and $\phi^{(e)^*}$ plotted against each other and coloured according to the comparison of MSEs. These are the results for the original plasmode dataset.

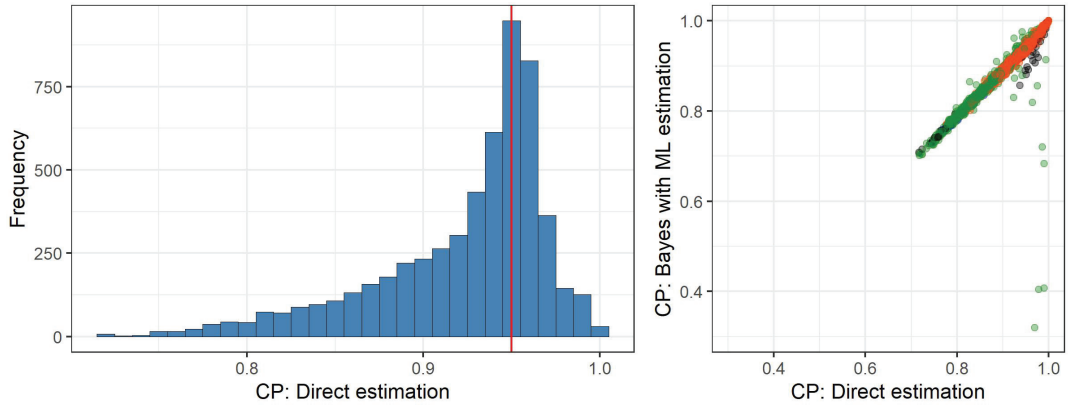

Figure 7.24: Left: Histogram showing the coverage probabilities for the confidence intervals obtained by the direct estimate. Right: Comparison of this coverage probability with the coverage probability for the credible intervals obtained by the Bayes method. Colours are the same as in Figure 7.19. These are the results for the original plasmode dataset.

Corresponding results, again very similar to the results observed before, for the Bayesian method based on robust estimation of the priors are shown in Figures B.62 and B.63 in Appendix B.3.

The three simulation studies presented above are increasingly closer to the real underlying dataset and at the same time moving further away from the assumptions. The

first simulation study based on the synthetic dataset shows the least genes with a lower MSE for the Bayesian method in comparison to the direct estimate. However, for this method, most genes have to be excluded from the analysis due to the inability to estimate the standard error of the estimate and thus preventing the calculation of the Bayes method. The normalised and the original dataset, retaining more features of the real VPA datasets, also allow for drastic improvements of the MSE when using the Bayes method for the vast majority of genes, while only for very few genes, the MSE becomes larger. The improvement in the simulation study based on the original plasmode dataset is even greater than for the other two simulation studies. This underlines the importance of accurately capturing the actual concentration-gene expression profiles observed in real data. These improvements of the MSE do not yield a decrease in the CPs of the resulting credible intervals in comparison to the confidence intervals.

7.5. Application to a real dataset

The meta-analysis and the Bayes method are applied to the real VPA gene expression dataset that is also used as basis for the simulation studies presented above. In contrast to the simulation studies, no new response data is simulated based on the previously fitted parameter values. Instead, the concentration-response data as originally measured is used. Since the true value of $\phi^{(e)*}$ is not known in this case, only comparisons of the different estimates can be made. Some probe sets and their respective estimates are examined individually to see the specific influence of the other probe sets or the empirical prior distribution on the final estimate in detail.

7.5.1. Meta-analysis for a real dataset

The meta-analysis is assessed in detail for the two GO groups of size 15 that yield a high and a medium correlation score. These two groups are 'GO:0001916' and 'GO:0030225'. In the group with high correlation score, with the exception of one probe set, all pairwise correlations are higher than 0.90 and for far more than half of the comparisons even higher than 0.95. Only for one probe set, all pairwise correlations with the other probe sets are between 0.5 and 0.8. In the group with medium correlation, two probe sets are strongly negatively correlated with the other probe sets. The rest of the pairwise correlations are slightly smaller than the one in the other group, with a few values larger than 0.85 and only very few correlations larger than 0.95.

The meta-analysis approach is conducted as explained before for these two datasets, each comprising 15 probe sets, separately. A cutoff of 0.955 to determine the probe sets to be included in the meta-analysis is used. Direct comparisons of the direct estimates and the meta-analysis estimates obtained for both GO groups are shown in Figure 7.25.

In the group with high score, only one probe set yields the same estimate for the meta-analysis and the direct estimate. This is the probe set with the lowest correlation to each of the other probe sets, thus no other probe set is included in this meta-analysis. For the remaining 14 probe sets, the meta-analysis estimate takes values around 5.5, while

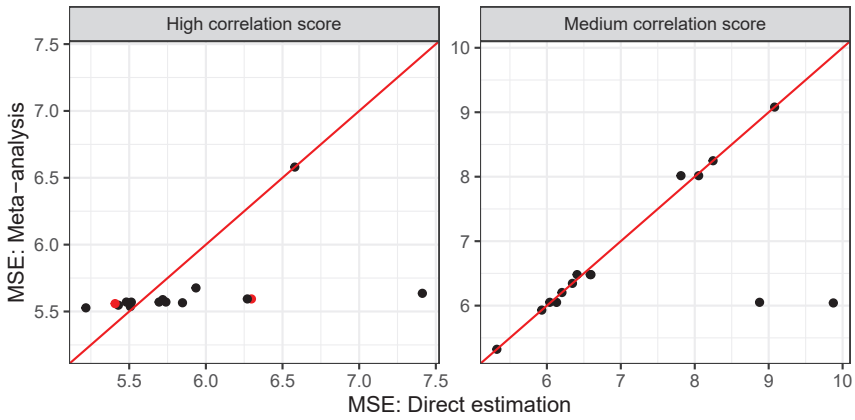

Figure 7.25: Comparison of the estimates obtained using the direct estimate and the meta-analysis method for the GO group with a high correlation score (left) and the GO group with a medium correlation score (right). Two probe sets that are examined in further detail are marked in red.

the direct estimates range from 5.0 to 7.0. At least 7 probe sets are included in each meta-analysis, and sizes go up to 14. For the GO group with medium score, however, most estimates are the same or almost the same for the direct estimate and the meta-analysis. Only for two probe sets, notably smaller values are estimated when using the meta-analysis in comparison to the direct estimate. The sizes of the meta-analyses are very small with a maximum size of 4 probe sets and only the size of 1 for 6 probe sets, i.e. no additional probe sets are included in the meta-analysis.

The two probe sets corresponding to the red dots from Figure 7.25 are examined in more detail. First the probe set is considered where the direct estimate yields a larger value than the meta-analysis based estimate. The fitted curve for this probe set together with the fitted curve of the 8 additional probe sets included in the meta-analysis are shown in Figure 7.26 (left). In red, the concentration corresponding to the direct estimate of $\phi^{(e)*}$ is indicated, which takes a value of 543. The concentration from the meta-analysis, with a value of 269, is much smaller.

All individual estimates of $\phi^{(e)*}$ together with their respective confidence intervals and the resulting estimate from the meta-analysis are summarised in a forest-plot in Figure 7.26 (right). The probe set under consideration is marked with an asterisk. The estimates for all additional probe sets in the meta-analysis are smaller than the estimate for the probe set under consideration. This can also be seen from the fitted curves since most of them start increasing at smaller concentrations and are already saturated in the range of concentrations considered. In comparison, the estimate for the considered probe set is subject to much uncertainty, as indicated by the large confidence interval. Thus, this estimate has less influence on the result of the meta-analysis than the estimates from probe sets with a smaller standard error. This results in the far lower estimate for the meta-analysis than for the direct estimate.

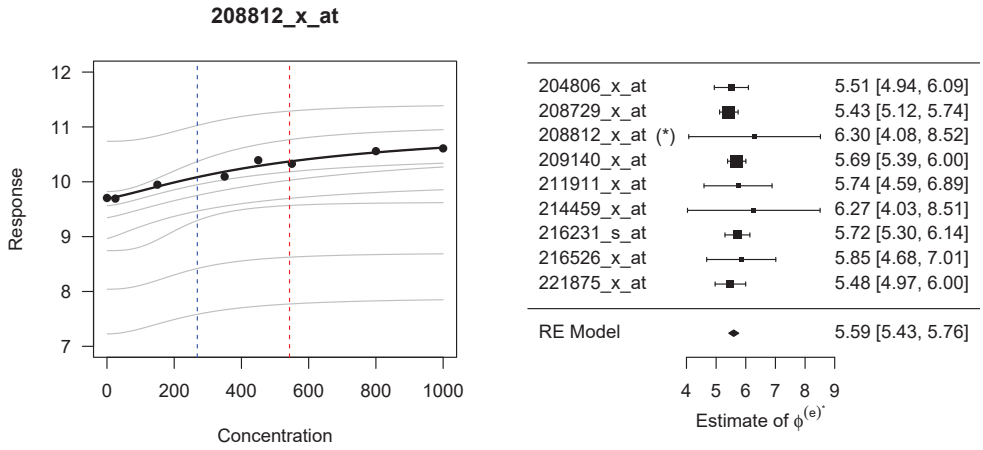

Figure 7.26: Left: Fitted concentration-response curve of the probe set considered with concentration-wise means indicated by dots. In grey, all additional probe sets included in the meta-analysis are plotted. The red line indicates the concentration corresponding to the direct estimate of $\phi^{(e)^*}$ and the blue line the concentration corresponding to the meta-analysis estimate. Right: Summary of the estimates for all probe sets included in the meta-analysis together with the 95% confidence interval based on the normal distribution, and the final random-effects model result. The original probe set considered is indicated by $(*)$.

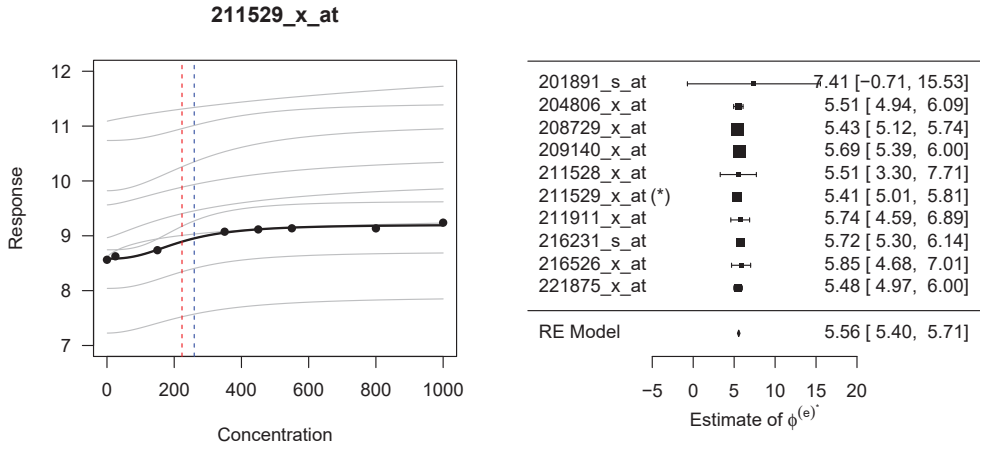

Figure 7.27: Examination of a specific example probe set with the same structure as Figure 7.26.

Corresponding results for the second probe set considered are shown in Figure 7.27. The concentrations corresponding to the direct estimate and the meta-analysis estimate are close to each other with values of 223 and 260, respectively. Ten probe sets in total

are included in the meta-analysis, with all of them yielding larger estimates than the considered probe set. The largest estimate observed for this set of probe sets shows a very large confidence interval and thus has only very little influence on the final result. The other estimates are closer to each other and show narrower confidence intervals. The direct estimate of the considered probe set is subject to little uncertainty, thus it strongly influences the meta-analysis estimate.

7.5.2. Shrinkage for a real dataset

Finally, the Bayes methodology is applied to the 5719 probe sets considered in the simulation study based on the original plasmode dataset in Chapter 7.4.3. Here, the original dataset is used, consisting of 27 observations for each probe set. A 4pLL curve is fitted for each probe set. The resulting distribution of parameter $\phi^{(e)^*}$ is the same as shown in Figure 7.17. Based on this distribution, the priors for both variants of the Bayes method are calculated: The ML estimation results in the prior parameters $\hat{\mu}_{\text{ML}} = 6.218$ and $\hat{\tau}^2_{\text{ML}} = 0.361$. The robust estimation results in $\hat{\mu}_{\text{rob}} = 6.232$ and $\hat{\tau}^2_{\text{rob}} = 0.226$, thus the median is slightly larger than the mean and the squared MAD is smaller than the variance. Using these prior parameters, the Bayes procedure is conducted for each probe set individually.

A direct comparison of the estimates obtained using the 4pLL curve directly and applying the Bayes method is shown in Figure 7.28. The shrinkage of the parameters towards the prior mean value, indicated by a red line, can clearly be observed there: For probe sets whose direct estimate is smaller than the respective prior mean, the Bayes estimate is larger or equal, but never smaller, and for probe sets whose direct estimate is larger than the respective prior mean, the Bayes estimate is smaller or equal, but never larger than the direct estimate. The larger values tend to be shrunken more than the smaller values, indicating more uncertainty in the estimation of $\phi^{(e)^*}$ when this parameter attains a large value.

In the next step, the lengths of the confidence intervals and credible intervals for each probe set are directly compared. The lengths of the respective intervals are plotted against each other in Figure 7.29, for the ML and the robust estimation method both. For fixed prior values across an entire dataset, as is the case here, the length of the credible interval only depends on the standard error of the estimate $\phi^{(e)^*}$. Analogously, for equal sizes of the dataset, the length of the credible interval also depends on this standard error only. This explains the functional relationship observed between the two lengths. In general, the credible intervals are far narrower than the confidence intervals, with maximum lengths observed at approximately 1.5 and 1.2 for the Bayes method, while lengths larger than 10 are observed for the confidence intervals.

As final step, three probe sets are examined individually in more detail. The three example probe sets are chosen in a way that for one probe set, the direct estimate is larger than the prior mean and thus the Bayes estimate is smaller, for one probe set, the opposite holds , and for the third probe set, almost no difference between the direct and the Bayes estimate can be observed. The choice of the three examples is illustrated

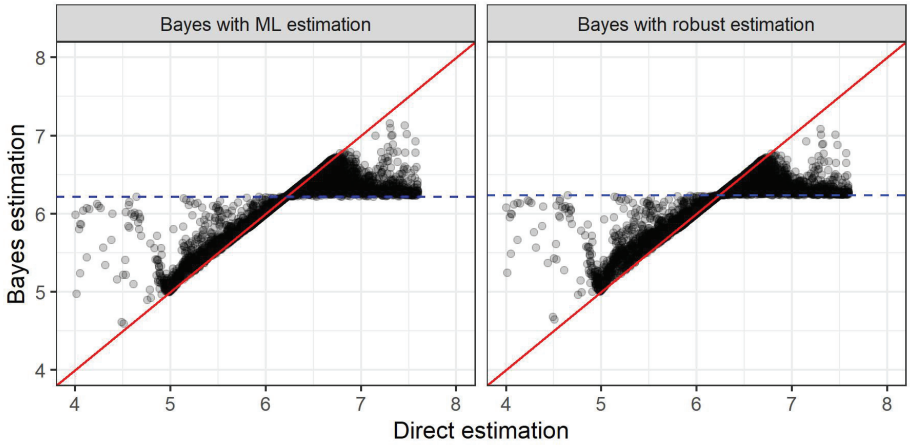

Figure 7.28: Comparison of the estimates obtained by the direct estimate and the Bayes method, with priors estimated using the ML method (left) and priors estimated using the robust method (right). The blue line indicates the respective prior mean of the normal distribution.

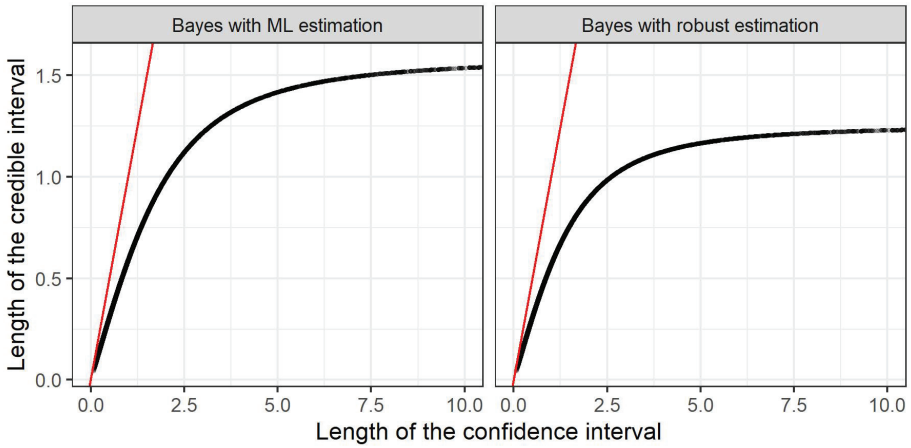

Figure 7.29: Comparison of the widths of the confidence intervals obtained by the direct estimation with the width of the credible interval obtained by the Bayes estimation. Results of the ML estimation are shown left and results of the robust estimation right.

in Figure 7.30 (top left) which shows the comparison of the estimates as already shown in Figure 7.28, but with the three chosen probe sets highlighted in red.

The complete concentration-response profiles together with the fitted 4pLL curves, the concentration corresponding to the direct estimate and the Bayes estimate using the ML method are shown in the other three plots of Figure 7.30. The first probe set (top right) has an increasing concentration-response profile. The fitted curve is not yet saturated in the range of concentrations considered, thus the concentration where the half-maximal

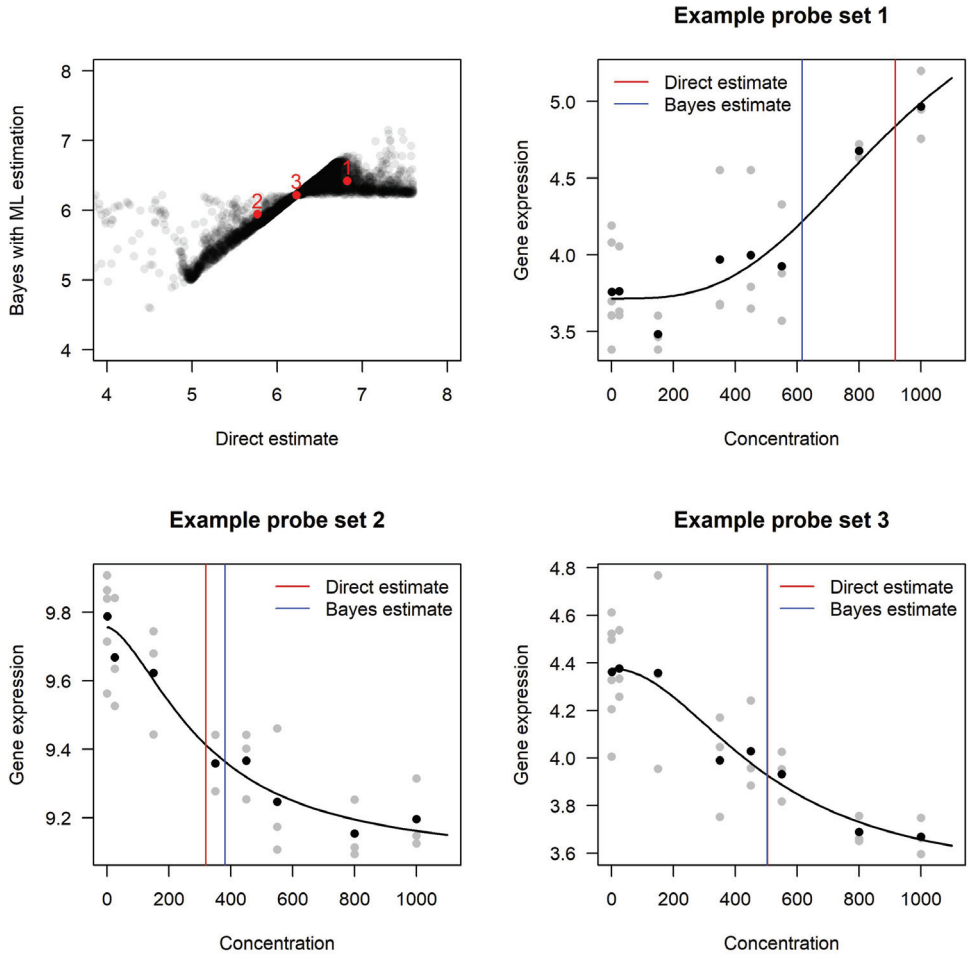

Figure 7.30: Top left: Comparison of direct estimate and Bayes estimate of $\phi^{(e)*}$. The red points indicate the three example probe sets chosen for more detailed examination. The other three plots show the concentration-response dataset together with the fitted curve for the three examples chosen. The red vertical line indicates the concentration corresponding to the direct estimate and the blue line the concentration corresponding to the Bayes estimate when using the ML method.

effect is achieved takes the relatively high value of approximately 920. This corresponds to an estimate of $\hat{\phi^{(e)}}^* = 6.821$ with a squared standard error of 0.701, which is quite large in comparison to the prior variance. Thus, the prior mean has more influence on the resulting Bayes estimate than the observation itself, leading to an estimate that corresponds to a concentration of approximately 620. This value seems more plausible when only considering the response values of the actually measured concentrations.

The second probe set considered (bottom left) shows a decreasing concentration-response profile with a right-sided asymptote that corresponds to a value only slightly lower than

the response for the highest concentration considered, i.e. the curve is almost saturated in the range of concentrations considered. The direct estimate of $\hat{\phi^{(e)}}^* = 5.764$ corresponds to a concentration of approximately 320, and the squared standard error takes a value of 0.234, so it is smaller than $\hat{\tau}^2_{\text{ML}}$, leading to more influence of the actual observed value for the calculation of the Bayes estimate. The Bayes estimate takes a value of 5.943, corresponding to the concentration of 380 and is larger than the originally observed value, but smaller than the prior mean.

The third and final probe set considered (bottom right) also has a decreasing profile. The direct estimate of $\hat{\phi^{(e)}}^* = 6.223$ almost corresponds to the prior mean of 6.218, also the squared standard error of 0.352 is almost equal to the prior variance of 0.361. Thus, the Bayes estimate, taking a value of 6.220, is only slightly higher than the observed value. These values correspond to a half-maximal effect concentration of approximately 500.

8. Conclusion and Discussion

In this thesis, three different statistical aspects for calculating alert concentrations from concentration-response data were considered. All three aspects were motivated from an application-oriented perspective and from problems arising in day-to-day toxicological work. Two types of concentration-response data were considered, namely cytotoxicity data, where the response is given by viability of cells, and gene expression data. For all three aspects, different statistical approaches were introduced and compared in controlled simulation studies in order to give recommendations regarding the best approach for analysing the specific type of data. The different methods were also applied to real datasets and results were compared to each other.

All three aspects were centred around the fitting of concentration-response data with a parametric curve. Specifically, the 4pLL model was used, which yields a monotonous, sigmoidal curve. The model function depends on four parameters, two for the upper and lower asymptote, respectively, one parameter that is proportional to the slope and one parameter that corresponds to the concentration where the half-maximal effect can be observed. Different alert concentrations, indicating a biological relevant and/or statistical significant effect of the compound of interest can be derived from these curves. The three aspects examined here are aimed at improving the estimation of these alert concentrations in the specific situations considered.

The first aspect was the problem of *deviating controls*, occurring in the context of cytotoxicity data. The term 'deviating controls' describes a situation in which the response values for the replicates of the negative control and the response values for the lowest tested positive concentrations differ from each other. Thus, when fitting a 4pLL model to this data that is normalised with respect to the control values, a curve is obtained whose upper asymptote does not correspond to a viability of 100%. Alert concentrations of interest are the EC values, where for $\lambda \in (0, 100)$, the EC_λ corresponds to the concentration at which the fitted curve attains a value of $100 - \lambda\%$. However, if the upper asymptote does not correspond to a value of 100%, EC values are meaningless in interpretation.

Four different methods that deal with the problem of deviating controls were introduced. In brief, three methods are based on the 4pLL model. For the first method the data are re-normalised based on the upper asymptote of a preliminary fit, for the second method the upper asymptote is forced through the value of 100% when normalisation is conducted with respect to the controls, and for the third method the controls are completely omitted. The fourth model makes use of the Brain-Cousens (BC) function, where a hormesis effect can be modelled. In this case, deviating values of the controls are considered in the model instead of dealing with them by normalisation of all values.

The extent of the problem of deviating controls was assessed by conducting an extensive literature review in three leading toxicological journals. Approximately 2200 papers were searched for viability curves fulfilling a set of criteria, and the respective curves were assessed for deviating controls. In this manner, 709 curves were chosen for analysis, 524 of which could additionally be analysed with respect to standard deviations. The

standard deviation was smaller than 10 in the majority of cases and smaller than 20 in almost all cases. Positively and negatively deviating controls occurred equally often with non-negligible deviations (i.e. deviations larger than 2 or smaller than -2) in far more than half of the cases and even deviations larger than 10 or smaller than -10 reasonably often. Thus it was shown that this is a problem occurring frequently in published data.

To really grasp the extent of this problem, however, a review in unpublished data would be necessary. The worst cases of deviating controls are probably not published, thus a bias in the results is plausible. Furthermore, the deviations and the standard deviations were mostly estimated visually, leading to additional sources of variation. Access to raw data would yield much more precise estimations of the deviations.

The four methods introduced for dealing with this problem were compared in a controlled simulation study. In this study, several scenarios were examined that differed in terms of the number of concentrations considered and the choice of the concentrations themselves. Additionally, different standard deviations of the three replicates per concentrations and different values of the deviations of the controls were considered. Main results are that the method where the upper asymptote is forced through the value of 100% performs worst in terms of the quality of the estimation of EC values. Omitting the controls very often leads to very good results, especially for large deviations of the controls, and in situations where the upper asymptote can be estimated well based on the remaining concentrations. The re-normalisation procedure is also competitive, especially when no or only few concentrations are measured in a range where no decrease in viability can be observed. The BC-based method yields good results as well when controls are negatively deviating, but needs to be visually checked for plausibility in case of positively deviating controls.

Based on the observations from the simulation study, an algorithm was proposed that helps to decide which method to use. This algorithm was summarised in Figure 5.10. The major drawback of the results of the simulation study is the need to know the specific situation that is reflected by a given curve in order to decide which method to use. However, without measuring response values for several low concentrations for which the compound does not yet inhibit viability, estimation of the deviation of the controls is not possible. This problem was already addressed by the algorithm, where typical properties of the situations that can actually be observed from a real-data curve are used for decision making.

One further possibility to address this issue is to average several curves obtained from the four different methods introduced for dealing with deviating controls. Instead of selecting only one curve based on which the EC value is calculated, a weighted mean of curves can be calculated, thus also leading to a weighted mean of EC values as final result. This procedure is called *model averaging* in contrast to the model selection performed here. The weights can be based on some quality criterion of the curve, e.g. the information criteria AIC and BIC. Including knowledge about the typically observed deviations of the controls is possible via prior information included in the calculation of the weights (Link and Barker, 2006). Inference for these model-averaged derived parameters while controlling the type I error rate is also possible, at least for sufficiently high sample size

(Jensen and Ritz, 2015). Ongoing research based on calculating the weights via AIC and no further prior information shows first promising results.

In the second topic, gene expression data was considered. An alert concentration of interest in this context is the concentration, where the fold change (FC, i.e. the difference of mean response values for two concentrations) in comparison to the control value exceeds a given threshold λ. Established alert concentrations are given by the ALOEC (absolute lowest observed effective concentration) and the LOEC (lowest observed effective concentration). These alert concentrations only take the measured concentrations into account and determine the lowest concentration, where the FC is exceeded (ALOEC) or significantly exceeded (LOEC). In order to assess the significance for the LOEC, a two-sample t-test or the Dunnett procedure can be employed.

Results from the (A)LOEC stem from the discrete set of measured concentrations only. In contrast, fitting a parametric curve to the data and determining alert concentrations based on this curve allows for any positive concentration as alert concentration. Thus, two model-based concentrations are examined: The ALEC (absolute lowest effective concentration) is the continuous equivalent to the ALOEC, and it is calculated as the concentration where the fitted curve attains the response value $f_0 \pm \lambda$, with f_0 denoting the asymptote of the fitted curve for the concentration tending towards 0. The LEC (lowest effective concentration) correspondingly is the continuous equivalent to the LOEC and is determined as the concentration where the fitted curve significantly exceeds the response value $f_0 \pm \lambda$. A statistical test based on the 4pLL model was derived that, together with a version of the bisection method as search algorithm, allows determination of the LEC. Methods introduced in this thesis are extensions from the methods introduced in the same context by Grinberg (2017).

The four different methods to calculate alert concentrations, with (A)LOEC also called 'observation-based' and (A)LEC 'model-based' methods, were compared in a controlled simulation study. Three different concentration-response profiles were considered as true underlying curves, with one scenario representing the null situation where the threshold is never exceeded and thus no alert concentration can be calculated. Each of these three scenarios was assessed in three variants regarding the choice of standard deviation for the simulated datasets. The LOEC was calculated using both the t-test and the Dunnett procedure.

In all cases considered, the model-based methods performed better than the observation-based methods, in terms of overestimating the true underlying alert concentrations less drastically while at the same time not yielding too many false positive results, i.e. results that underestimate the true alert concentrations. In the scenario where no alert concentration could be calculated, for all methods taking significance into account the proportion of false positive alerts was smaller than the significance level of 5%. While the observation-based methods, especially the LOEC, vastly overestimated the true alert concentration, the model-based methods lead to narrow distributions around the true value. Generally, model-based estimates lead to smaller results than observation-based estimates. Especially for the situations of large variances, more valid estimates were obtained using the model-based methods.

Comparison of the LOECs based on the t-test and the Dunnett procedure showed that both methods lead to very similar results with no structure behind larger or smaller results for one or the other method. Additionally to these results, coverage probabilities for the confidence intervals obtained by the ALEC were calculated. These probabilities were mostly in the same range for all situations considered, but lower than would be desired for 95% confidence intervals that were calculated.

An advantage of the model-based approaches clearly is the independence of the measured concentration levels, as any positive concentration is allowed as result. Moreover, in contrast to alert concentrations as the EC50 that heavily depend on the values of the asymptotes, (A)LEC values can also be estimated reliably in the case of an incomplete concentration-response dataset. In such a situation, the right-sided asymptote can still be extrapolated by fitting a curve, but in doing so, it is accepted that the asymptote may be slightly biased. Such a biased estimation has only little impact on the (A)LEC since the threshold λ that is to be exceeded is pre-specified and does not depend on the values of the asymptotes.

The distribution of the test statistic under the null hypothesis for the newly developed test based on the 4pLL model is only asymptotically a standard normal distribution. In typical toxicological applications, the sample size is too small to allow asymptotic statements. Thus, other methods for calculating p-values to reach a test decision should be examined, for example resampling-based approximate distributions.

The third topic also dealt with gene expression data, mainly for the specific case of microarray data where response values are measured for tens of thousands of genes simultaneously. The goal was to find methods that allow the *sharing of information* regarding the model parameters of fitted 4pLL curves, thus leading to improved estimations of the model parameters. The parameter indicating the log-transformed concentration corresponding to the half-maximal effect of the curve, $\phi^{(e)^*}$, was chosen as alert concentration of interest. This is a reasonable indicator for relevant change in the observed gene expression.

Two different statistical approaches were examined: In the first approach, the assumption of a normal distribution of parameter $\phi^{(e)^*}$ was exploited to perform a meta-analysis. For a specific gene, all genes that are similar to this gene in terms of correlation are included in the meta-analysis. Then a random-effects model, making use of the DerSimonian-Laird estimator for the heterogeneity, is calculated based on the estimates of $\phi^{(e)^*}$ and the respective standard errors for all genes considered.

The second approach is an empirical Bayes procedure. A 4pLL model is fitted to each of the genes in the analysis. The parameter $\phi^{(e)^*}$ is assumed to be normally distributed. The mean of this normal distribution is assumed to follow an a-priori normal distribution as well. The mean and the variance for this prior distribution are empirically estimated from the observed estimates for $\phi^{(e)^*}$. This can be done using the maximum-likelihood estimators or using median and MAD as robust estimators. Under the assumptions stated above, the posterior distribution of the parameter for each gene, given the specific estimate, again follows a normal distribution and can explicitly be calculated. With

this method, essentially a shrinkage of the estimated parameter value $\phi^{(e)*}$ towards the observed mean of all estimated values is performed.

The simulation studies conducted for this topic were all based on true underlying plasmode datasets. A dataset where gene expression at seven increasing doses and a negative control of the compound valproic acid (VPA) was measured for 54675 probe sets was used as basis of the simulation studies. A smaller set of probe sets was chosen according to the criteria of statistical significance and biological relevance. Parameters of 4pLL models fitted to this set of probe sets were used for the simulation studies, in which datasets were simulated on the basis of these parameters, together with normally distributed noise. Thus, biological properties in the dataset where retained in the simulation scenarios.

First, a simulation study based on a large set of probe sets to examine the performance of the meta-analysis approach was conducted. Targets to assess the performance of this approach in comparison to directly estimating the parameter $\phi^{(e)*}$ for each probe set separately were the MSE and the coverage probability (CP). For about equally many probe sets, the MSE was smaller or larger, respectively, when comparing meta-analysis and the direct estimate. The CP was much worse for the meta-analysis than for the direct estimate, regardless of which specific method for calculating the confidence interval for the meta-analysis was used.

The negative results for this simulation study were expected. In this analysis, more or less random noise was added to each probe set, which in most cases lead to biased and thus worse results than the direct comparison. Thus, simply adding information based on similarity of genes is not a good approach.

The next step was to restrict the set of probe sets from which potential additional information are gathered. This was done by considering GO groups, i.e. groups of probe sets that contribute to the same biological process. Eight GO groups were considered in total, four of size 15 and four of size 30, with one group with similar probe sets in terms of correlation, one group with a medium similarity, one group with only little similarity and one randomly sampled group, respectively.

Results are ambiguous: In some groups, the MSE could be improved by conducting the meta-analyses, and in some groups, the MSE was worse. This did not follow a clear pattern. No clear link between the sizes of the meta-analyses and the results could be observed: In some cases, allowing more probe sets into the meta-analysis lead to improved values of the MSE, in some cases, to worse values. The CP was worse for the meta-analysis based confidence intervals than for the direct estimates in all cases examined.

Therefore, the meta-analysis approach in the way applied in this thesis does not lead to promising results. A major drawback of the GO group based method would anyway be given by the fact that all probe sets are assigned to several GO groups. In each GO group, different probe sets would be included in the meta-analysis, thus leading to different meta-analysis based estimates of $\phi^{(e)*}$, depending on the choice of the specific GO group. However, an extension of this method to find one summary value of $\phi^{(e)*}$ for a specific GO group and not the probe sets themselves could be promising.

Similarity between probe sets was assessed here using the correlations of concentration-wise means of gene expression only. Considering different criteria would be interesting: One possibility is the clustering of probe sets according to their concentration-gene expression profiles. The order-restricted information criterion-based cluster algorithm 'ORICC' (Liu et al., 2009) and the order-restricted inference for ordered gene expression data, 'ORIOGEN', (Peddada et al., 2005) are two order-restricted clustering methods that are specifically aimed at clustering genes according to their concentration-response profiles. These methods can either be used as a filtering step, additional to the biological similarities, to reduce the set of probe sets that are potential similar probe sets, or to establish the set of probe sets considered in the meta-analysis.

The Bayes method was analysed for three situations of true underlying datasets that on the one hand are more and more similar to the VPA dataset, but on the other hand are moving away from fulfilling the assumption of normality. In the first simulation study, a completely simulated dataset was used as basis. The MSE was improved by the Bayesian method for almost all synthetic probe sets considered, and CPs of the confidence interval for the direct estimate and the credible interval for the Bayesian method were almost equal in all cases.

In the second dataset, the parameters from the VPA dataset were used, and only probe sets where $\phi^{(e)^*}$ did not exceed log(2000) were considered. A quantile normalisation was applied to the values of parameter $\phi^{(e)^*}$, thus ensuring the assumption of normality. Only for very few probe sets, the MSE did not improve when applying the Bayes procedure in comparison to the direct estimate, and the CP was almost equal for both methods, again.

Lastly, the original parameter estimates from the VPA dataset were used, whereby only probe sets where $\phi^{(e)^*}$ did not exceed $\log(2000) = 7.60$ and $\phi^{(e)^*}$ was larger than 0 were considered. When using these parameter estimates as true underlying set of probe sets, again an improvement in the MSE could be observed for most probe sets. Only for the very low number of 35 probe sets, the MSE was worse. The true underlying values of $\phi^{(e)^*}$ for these probe sets was either very low or very high, such that the shrinkage performed by the Bayes method was counter-productive here. For the other probe sets, however, the Bayes method performed far better than the direct estimate in terms of MSE and not worse than the direct estimate in terms of CP.

Results regarding the Bayes method stated above referred to the variant where the prior distribution is estimated using maximum likelihood estimates. When estimating the prior distribution using robust measures, results were consistently worse, both in terms of MSE and CP. The variance used in the Bayes method was smaller for the robust estimation, thus leading to higher influence of the empirical prior mean in comparison to the observed values. Especially for probe sets with rather large or rather small values of $\phi^{(e)^*}$, this yields a worsened MSE.

In the simulation studies, the set of true underlying probe sets was restricted to situations with values of $\phi^{(e)^*}$ that are not unreasonably large in comparison to the range of concentrations considered. Such large values usually indicate curves that are very difficult to estimate numerically. In real-data situations, identifying such probe sets in

advance to exclude them from further fitting is not easily possible. Large estimated values of $\phi^{(e)^*}$ could even heavily influence the estimation of the prior distribution, thus leading to biased results.

The Bayes method could be improved by choosing a more appropriate prior distribution that better fits the actual distribution of $\phi^{(e)^*}$. As summarised in Figure 7.17, the tails of the actual distribution are heavier than the tails of the estimated normal distributions. One approach would be the use of a mixture of two normal distribution as prior distribution. The resulting posterior distribution would be a mixture of normal distributions as well (Reich and Ghosh, 2019, p. 56). This allows the direct calculation of the posterior, instead of using algorithms like Markov Chain Monte Carlo methods for simulating the posterior distributions.

In the VPA dataset, gene expression values are measured for the concentrations $25, 150, 350, 450, 550, 800$, and 1000 µM. These concentrations are not equidistant when considering the concentrations on log-scale, as is done for the 4pLL model. Although this dataset has a high overall quality and many replicates measured in narrow concentrations, it is not necessarily the ideal dataset to base simulation studies on for any model where concentrations are considered on log-scale. Especially the results for the information sharing heavily depend on the specific structure of the dataset and the interaction between the substance under consideration and the measured concentrations. Broadening the results for alert concentrations in the context of gene expression data to more datasets would be interesting, however, high-quality datasets with enough measurements are very expensive and time-consuming to generate.

The TG-Gates database (Igarashi et al., 2015) comprises extensive toxicological data for more than 170 compounds. Among others, microarray data for the same Affymetrix GeneChip® as used for the VPA dataset can be found there. This data, however, is only measured for three different concentrations but for a narrow series of incubation times. This data can be used to perform the same analyses regarding information sharing as for the VPA dataset using the time instead of the concentration as independent variable.

The three topics considered in this thesis are based on the 4pLL model, and all simulation studies are based on this model as true underlying model. Sensitivity analyses, in the sense that different models are used as true underlying models, while the methods are conducted based on the 4pLL model as presented here, would lead to insights about the performance of the methods in more general cases. Beyond that, a generalisation of the methods to other frequently used models in toxicology, e.g. Weibull models or log-normal models (e.g. Ritz et al., 2019, pp. 183-186), would be a sensible extension.

All analyses were conducted under the assumption of homogeneity, i.e. equal variances across concentrations. Especially for viability assays, variances of the replicates are often observed to be decreasing in ranges of high toxicity, i.e. for decreasing viability. Model fitting is also possible in the case of heteroscedasticity or in the case that the residuals do not follow a normal distribution (Calderazzo et al., 2019). The problem of heteroscedasticity can also be addressed by using robust methods to estimate the standard errors of the parameters. One example for such estimators are the sandwich variance estimators (Zeileis, 2006).

For many toxicological assays, and at least the viability assay considered in this thesis is no exception, the choice of considered concentrations is based on the convention of using equidistant concentrations. A practical goal, however, would be to achieve a high precision in estimating the target parameter, while at the same time minimising the required number of measurements. Statistical design theory gives the background to create such an optimal experimental design. Holland-Letz and Kopp-Schneider (2015) give concrete guidance on the calculation of such optimal designs, among others in the case of log-logistic curves and even provide an online-tool to calculate such designs or compare specific designs with the optimal one.

In conclusion, it can be said that the combination of curve-fitting with typical toxicological procedures is an exciting field with many statistical questions still unanswered. High-dimensional data from gene expression experiments offer opportunities to identify many properties of substances already by conducting in vitro analyses. At the same time, the statistical handling of these data is a challenge. The three topics covered in this thesis provide answers to dealing with different tasks and the proposed methods can be the starting point for many other investigations.

References

Affymetrix. GeneChip® Expression Analysis. Technical Manual, 2003a. rev 4.0 edition.

Affymetrix. Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays. Technical Report, 2003b. rev 2.0 edition.

A. Alexa and J. Rahnenführer. *topGO: Enrichment Analysis for Gene Ontology*, 2020. R package version 2.40.0.

P. Armitage. Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, 11 (3):375–386, 1955.

B. Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.3.

B. M. Bolstad. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, 2004.

B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

B. Bornkamp. *DoseFinding: Planning and Analyzing Dose Finding Experiments*, 2019. URL `https://CRAN.R-project.org/package=DoseFinding`. R package version 0.9-17.

B. Bornkamp, J. Pinheiro, and F. Bretz. MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies. *Journal of Statistical Software*, 29(7):1–23, 2009.

P. Brain and R. D. Cousens. An equation to describe dose responses where there is stimulation of growth at low dose. *Weed research*, 29:93–96, 1989.

F. Bretz, J. C. Pinheiro, and M. Branson. Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. *Biometrics*, 61(3):738–748, 2005.

E. J. Calabrese and L. A. Baldwin. Hormesis: The Dose-Response Revolution. *Annual Review of Pharmacology and Toxicology*, 43(1):175–197, 2003.

S. Calderazzo, D. Tavel, M.-G. Zurich, and A. Kopp-Schneider. Model-based estimation of lowest observed effect concentration from replicate experiments to identify potential biomarkers of in vitro neurotoxicity. *Archives of Toxicology*, 93:2635–2644, 2019.

N. Cedergreen, C. Ritz, and J. C. Streibig. Improved empirical models describing hormesis. *Environmental Toxicology and Chemistry*, 24(12):3166 – 3172, 2005.

W. Cochran. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics*, 10 (4):417–451, 1954.

K. S. Crump. A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology*, 4(5):854 – 871, 1984.

M.-L. Delignette-Muller, C. Forfait, E. Billoir, and S. Charles. A new perspective on the Dunnett procedure: Filling the gap between NOEC/LOEC and EC_x concepts. *Environmental Toxicology and Chemistry*, 30(12):2888–2891, 2011.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177 – 188, 1986.

J. C. Duda. Model Selection and Model Averaging of Dose-Response Gene Expression Data with MCP-Mod. Master's thesis, TU Dortmund University, 2019.

J. H. Duffus, M. Nordberg, and D. M. Templeton. Glossary of terms used in toxicology, 2nd edition (IUPAC Recommendations 2007). *Pure and Applied Chemistry*, 79(7): 1153–1344, 2007.

C. W. Dunnett. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, 50(272):1096 – 1121, 1955.

C. Feller, K. Schorning, H. Dette, G. Bermann, and B. Bornkamp. Optimal Designs for Dose Response Curves with Common Parameters. *The Annals of Statistics*, 45(5): 2102–2132, 2017.

D. J. Finney. Bioassay and the Practice of Statistical Inference. *International Statistical Review*, 47(1):1–12, 1979.

H. Fletcher and I. Hickey. *BIOS Instant Notes in Genetics*. CRC Press LLC, London, 4 edition, 2012.

O. Forster. *Analysis 1*. Springer Spektrum, Wiesbaden, 12 edition, 2016.

Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

GraphPad Software. *GraphPad Prism*. San Diego, California USA. URL `www.graphpad.com`.

M. Grinberg. *Statistical analysis of concentration-dependent high-dimensional gene expression data*. PhD thesis, Department of Statistics at TU Dortmund University, 2017.

X. Gu, W. Albrecht, K. Edlund, F. Kappenberg, J. Rahnenführer, M. Leist, W. Moritz, P. Godoy, C. Cadenas, R. Marchan, T. Brecklinghaus, L. T. Pardo, J. V. Castell, I. Gardner, B. Han, J. G. Hengstler, and R. Stoeber. Relevance of the incubation period in cytotoxicity testing with primary human hepatocytes. *Archives of Toxicology*, 92(12):3505–3515, 2018.

C. Harbron, K.-M. Chang, and M. C. South. RefPlus: an R package extending the RMA Algorithm. *Bioinformatics*, 23(18):2493–2494, 2007.

J. Hartung and G. Knapp. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12):1771–1782, 2001a.

J. Hartung and G. Knapp. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24):3875 – 3889, 2001b.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, second edition edition, 2009.

A. V. Hill. The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curve. *The Journal of Physiology*, 40(Suppl.):iv – vii, 1910.

T. Holland-Letz and A. Kopp-Schneider. Optimal experimental designs for dose-response studies with continuous endpoints. *Archives of Toxicology*, 89:2059 – 2068, 2015.

L. A. Hothorn. *Statistics in Toxicology Using R*. CRC Press, Boca Raton, 2015.

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, and H. Yamada. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic acids research*, 43(Database issue):D921–D927, 2015.

R. A. Irizarry, B. M. Bolstand, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):1 – 8, 2003a.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249 – 264, 2003b.

H. Izadi, J. E. Grundy, and R. Bose. Evaluation of the Benchmark Dose for Point of Departure Determination for a Variety of Chemical Classes in Applied Regulatory Settings. *Risk Analysis*, 32(5):830–835, 2012.

S. M. Jensen and C. Ritz. Simultaneous Inference for Model Averaging of Derived Parameters. *Risk Analysis*, 35(1):68 – 76, 2015.

S. M. Jensen, F. M. Kluxen, and C. Ritz. A Review of Recent Advances in Benchmark Dose Methodology. *Risk Analysis*, 39(19):2295 – 2315, 2019.

X. Jiang. *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, 2013.

X. Jiang and A. Kopp-Schneider. Summarizing EC50 estimates from multiple dose-response experiments: A comparison of a meta-analysis strategy to a mixed-effects model approach. *Biometrical Journal*, 56(3):493 – 512, 2014.

F. Kappenberg, T. Brecklinghaus, W. Albrecht, J. Blum, C. van der Wurp, M. Leist, J. G. Hengstler, and J. Rahnenführer. Handling deviating control values in concentration-response curves. *Archives of Toxicology*, 94(11):3787 – 3798, 2020.

F. Kappenberg, M. Grinberg, X. Jiang, A. Kopp-Schneider, J. G. Hengstler, and J. Rahnenführer. Comparison of observation-based and model-based identification of alert concentrations from concentration-expression data. *Bioinformatics*, 2021. btab043.

B. B. Knowles, C. C. Howe, and D. P. Aden. Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis b surface antigen. *Science*, 209(4455): 497–499, 1980.

A. Krebs, J. Nyffeler, J. Rahnenführer, and M. Leist. Normalization of data for viability and relative cell function curves. *ALTEX- Alternatives to animal experimentation*, 35 (2):268–271, 2018.

A. K. Krug, R. Kolde, J. A. Gaspar, E. Rempel, N. V. Balmer, K. Meganathan, K. Vojnits, M. Baquié, T. Waldmann, R. Ensenat-Waser, S. Jagtap, R. M. Evans, S. Julien, H. Peterson, D. Zagoura, S. Kadereit, D. Gerhard, I. Sotiriadou, M. Heke, K. Natarajan, M. Henry, J. Winkler, R. Marchan, L. Stoppini, S. Bosgra, J. Westerhout, M. Verwei, J. Vilo, A. Kortenkamp, J. Hescheler, L. Hothorn, S. Bremer, C. van Thriel, K.-H. Krause, J. G. Hengstler, J. Rahnenführer, M. Leist, and A. Sachinidis. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Archives of Toxicology*, 87(1):123–143, 2013.

W. A. Link and R. J. Barker. Model Weights and the Foundations of Multimodel Inference. *Ecology*, 87(10):2626–2635, 2006.

T. Liu, N. Lin, N. Shi, and B. Zhang. Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments. *BMC Bioinformatics*, 10(146), 2009.

F. W. Nussbeck. Log-logistic models. In A. C. Michalos, editor, *Encyclopedia of Quality of Life and Well-Being Research*, pages 3686–3689. Springer Netherlands, Dordrecht, 2014.

S. D. Peddada, S. Harris, J. Zajd, and E. Harvey. ORIOGEN: order restricted inference for ordered gene expression data. *Bioinformatics*, 21(20):3933–3934, 2005.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`. Version 4.0.0.

B. J. Reich and S. K. Ghosh. *Bayesian Statistical Methods*. Chapman and Hall / CRC, New York, 2019.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

C. Ritz, F. Baty, J. C. Streibig, and D. Gerhard. Dose-Response Analysis Using R. *PLOS ONE*, 10(12), 2015.

C. Ritz, S. M. Jensen, D. Gerhard, and J. C. Streiberg. *Dose-Response Analysis Using R.* Chapman and Hall / CRC, New York, 2019.

S. J. Ruberg. Dose response studies. I. Some design considerations. *Journal of Biopharmaceutical Statistics*, 5(1):1–14, 1995.

A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

L. van der Vliet and C. Ritz. Statistics for Analyzing Ecotoxicity Test Data. In J.-F. Férard and C. Blaise, editors, *Encyclopedia of Aquatic Ecotoxicology*, pages 1081–1096. Springer Netherlands, Dordrecht, 2013.

L. K. Vaughan, J. Divers, M. Padilla, D. T. Redden, H. K. Tiwari, D. Pomp, and D. B. Allison. The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Computational statistics & data analysis*, 53(5):1755–1766, 2009.

W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016.

D. A. Williams. A Test for Differences between Treatment Means When Several Dose Levels are Compared with a Zero Dose Control. *Biometrics*, 27(1):103 – 117, 1971.

A. Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software, Articles*, 16(9):1–16, 2006.

A. Zeller, G. Duran-Pacheco, and M. Guérard. An appraisal of critical effect sizes for the benchmark dose approach to assess dose–response relationships in genetic toxicology. *Archives of Toxicology*, 92(12):3799 – 3807, 2017.

List of Figures

List of Tables

A. Calculations

The 4pLL model in its two parametrisations as introduced in Chapter 4.1.2 used for the calculations in the following sections is given as

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

$$= \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + (\frac{x}{\phi^{(e)}})^{\phi^{(b)}}}. \tag{2}$$

In the following, first the values of the left-sided and right-sided asymptotes of a 4pLL model depending on the sign of the parameter $\phi^{(b)}$ are calculated. Then the slope is calculated, both when considering an untransformed and a log-transformed x-axis. The equivalence of the parameterization of the `sigEmax`-model from the MCP-Mod approach to the 4pLL model is shown and finally, the gradient $\nabla f(0, \boldsymbol{\phi})$, again depending on the sign of the parameter $\phi^{(b)}$, is calculated. In the final section of this Appendix, the posterior in a normal-normal model is calculated. This is independent of the 4pLL function.

A.1. Calculation of the limits of a 4pLL model

The limits of the function $f(x, \boldsymbol{\phi})$ depend on the sign of the parameter $\phi^{(b)}$. First, let $\phi^{(b)} > 0$. Then $\lim_{x\to 0}\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}} = 0$, s.t.

$$\phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to 0}{\longrightarrow} 0}} \quad \overset{x\to 0}{\longrightarrow} \quad \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1} \quad = \quad \phi^{(d)}.$$

It also follows that $\lim_{x\to\infty}\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}} = \infty$, s.t.

$$\phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to\infty}{\longrightarrow} \infty}} \quad \overset{x\to\infty}{\longrightarrow} \quad \phi^{(c)}.$$

Analogously for $\phi^{(b)} < 0$ it holds $\lim_{x\to 0}\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}} = \infty$ and $\lim_{x\to\infty}\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}} = 0$, s.t.

$$\phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to 0}{\longrightarrow}\infty}} \quad \overset{x\to 0}{\longrightarrow} \quad \phi^{(c)} \qquad \text{and}$$

$$\phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to\infty}{\longrightarrow} 0}} \quad \overset{x\to\infty}{\longrightarrow} \quad \phi^{(d)}.$$

A.2. Calculation of the inflection point of a 4pLL model

The inflection points are calculated for the original scale and the logarithmic x-axis, both. Since the parameters $\phi^{(c)}$ and $\phi^{(d)}$ only change the location of the curve on the y-axis, they do not influence the inflection point. Thus, without loss of generality, it is assumed that $\phi^{(d)} = 1$ and $\phi^{(c)} = 0$, yielding the function

$$f(x, \boldsymbol{\phi}) = \frac{1}{1 + \left(\exp(\phi^{(b)}(\log(x) - \phi^{(e)^*})\right)}, \tag{3}$$

with $\phi^{(e)^*} = \log(\phi^{(e)})$.

For the first calculation, the logarithmic x-axis is considered, i.e. with $\tilde{x} := \log(x)$, it needs to be shown that $\phi^{(e)^*}$ is the inflection point of $f(\tilde{x}, \boldsymbol{\phi})$. The necessary criterion to be shown is that the second derivative of $f(\tilde{x}, \boldsymbol{\phi})$ evaluated at $\phi^{(e)^*}$ equals 0, while the third derivative evaluated at the same concentration is unequal to 0.

The derivatives are only stated here and not calculated in detail, this can be done using the basic differentiation rules. All throughout the calculations, $\phi^{(b)} \neq 0$ is assumed. For $\phi^{(b)} = 0$, the resulting curve would be flat and thus no inflection point would be present.

$$f'(\tilde{x}, \boldsymbol{\phi}) = -\frac{\phi^{(b)} \exp\left(\phi^{(b)}(\tilde{x} - \phi^{(e)^*})\right)}{\left(1 + \exp\left(\phi^{(b)}(\tilde{x} - \phi^{(e)^*})\right)\right)^2}$$

$$f''(\tilde{x}, \boldsymbol{\phi}) = -\frac{{\phi^{(b)}}^2 \exp\left(\phi^{(b)}(\tilde{x} - \phi^{(e)^*})\right)}{\left(1 + \exp\left(\phi^{(b)}(\tilde{x} - \phi^{(e)^*})\right)\right)^2} + \frac{2{\phi^{(b)}}^2 \exp\left(2\phi^{(b)}(\tilde{x} - \phi^{(e)^*})\right)}{\left(1 + \exp\left(\phi^{(b)}(\tilde{x} - \phi^{(e)^*})\right)\right)^3}$$

$$f'''(\tilde{x}, \boldsymbol{\phi}) = -\frac{{\phi^{(b)}}^3 \exp\left(\phi^{(b)}(\tilde{x} - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}(\tilde{x} - {\phi^{(e)}}^*)\right)\right)^2} + \frac{6{\phi^{(b)}}^3 \exp\left(2\phi^{(b)}(\tilde{x} - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}(\tilde{x} - {\phi^{(e)}}^*)\right)\right)^3} - \frac{6{\phi^{(b)}}^3 \exp\left(3\phi^{(b)}(\tilde{x} - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}(\tilde{x} - {\phi^{(e)}}^*)\right)\right)^4}$$

Insert ${\phi^{(e)}}^*$ into the second and third derivative:

$$\begin{aligned}
f''({\phi^{(e)}}^*, \boldsymbol{\phi}) &= -\frac{{\phi^{(b)}}^2 \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)\right)^2} + \frac{2{\phi^{(b)}}^2 \exp\left(2\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)\right)^3} \\
&= -\frac{{\phi^{(b)}}^2}{4} + \frac{2{\phi^{(b)}}^2}{8} = 0 \\
f'''({\phi^{(e)}}^*, \boldsymbol{\phi}) &= -\frac{{\phi^{(b)}}^3 \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)\right)^2} + \frac{6{\phi^{(b)}}^3 \exp\left(2\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)\right)^3} \\
&\quad - \frac{6{\phi^{(b)}}^3 \exp\left(3\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)}{\left(1 + \exp\left(\phi^{(b)}({\phi^{(e)}}^* - {\phi^{(e)}}^*)\right)\right)^4} \\
&= -\frac{{\phi^{(b)}}^3}{4} + \frac{6{\phi^{(b)}}^3}{8} - \frac{6{\phi^{(b)}}^3}{16} \\
&= \frac{2{\phi^{(b)}}^3}{16} \neq 0
\end{aligned}$$

This shows that $\phi^{(e)}$ is the inflection point of the function (3) and thus also of the general 4pLL model. To calculate the inflection point of function (3) when the x-axis is non-logarithmic, derivatives of the untransformed functions are calculated. The second derivative is set to 0 and the resulting concentration is considered to be the inflection point. A specification of the third derivative is omitted here. It is required that $|\phi^{(b)}| > 1$ for the calculation of an inflection point for the non-logarithmic x-axis, otherwise the calculation in the last step would yield a non-real result.

$$f'(x, \boldsymbol{\phi}) = -\frac{\phi^{(b)} x^{\phi^{(b)}-1} \left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left(1 + \left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^2}$$

$$f''(x,\boldsymbol{\phi}) = -\frac{(\phi^{(b)}-1)\phi^{(b)}x^{\phi^{(b)}-2}\left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^2} + \frac{2{\phi^{(b)}}^2 x^{2\phi^{(b)}-2}\left(\frac{1}{\phi^{(e)}}\right)^{2\phi^{(b)}}}{\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^3}$$
$$= \frac{-(\phi^{(b)}-1)\phi^{(b)}x^{\phi^{(b)}-2}\left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}}\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right) + 2{\phi^{(b)}}^2 x^{2\phi^{(b)}-2}\left(\frac{1}{\phi^{(e)}}\right)^{2\phi^{(b)}}}{\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^3}$$

To find the inflection point, the numerator of the last fraction needs to be equal to 0. After multiplication of the numerator with $\frac{{\phi^{(e)}}^{\phi^{(b)}}}{\phi^{(b)}x^{\phi^{(b)}-2}}$, it follows:

$$-(\phi^{(b)}-1)\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right) + 2\phi^{(b)}x^{\phi^{(b)}}\left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}} \stackrel{!}{=} 0$$
$$\Rightarrow \quad x^{\phi^{(b)}}\left(-\phi^{(b)}\left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}} + \left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}} + 2\phi^{(b)}\left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}}\right) \stackrel{!}{=} \phi^{(b)}-1$$
$$\Rightarrow \quad x^{\phi^{(b)}} \stackrel{!}{=} \frac{\phi^{(b)}-1}{(\phi^{(b)}+1)\left(\frac{1}{\phi^{(e)}}\right)^{\phi^{(b)}}}$$

Thus, it follows that the concentration for the inflection point is given by

$$x = \left(\frac{\phi^{(b)}-1}{\phi^{(b)}+1}\right)^{\frac{1}{\phi^{(b)}}}\phi^{(e)}.$$

A.3. Calculation of the slope of a 4pLL model

For calculating the slope of the 4pLL function at concentration $\phi^{(e)}$, the first step is to calculate the general form of the derivative. The parametrisation of $f(x,\boldsymbol{\phi})$ from (2) is used for this calculation:

$$\frac{d}{dx}f(x,\boldsymbol{\phi}) = \frac{-\left(\phi^{(d)}-\phi^{(c)}\right)\cdot\phi^{(b)}\cdot\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}-1}\cdot\frac{1}{\phi^{(e)}}}{\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^2}$$
$$= \frac{-\left(\phi^{(d)}-\phi^{(c)}\right)\cdot\frac{\phi^{(b)}}{\phi^{(e)}}\cdot\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}-1}}{\left(1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^2}$$

Then, the derivative is evaluated at $x = \phi^{(e)}$ what leads to the final result:

$$\begin{aligned}\frac{d}{dx} f(x, \boldsymbol{\phi})\Big|_{x=\phi^{(e)}} &= \frac{-\left(\phi^{(d)} - \phi^{(c)}\right) \cdot \frac{\phi^{(b)}}{\phi^{(e)}} \cdot \left(\frac{\phi^{(e)}}{\phi^{(e)}}\right)^{\phi^{(b)}-1}}{\left(1 + \left(\frac{\phi^{(e)}}{\phi^{(e)}}\right)^{\phi^{(b)}}\right)^2} \\ &= -\frac{\phi^{(b)} \cdot \left(\phi^{(d)} - \phi^{(c)}\right)}{4\phi^{(e)}}\end{aligned}$$

Calculating the slope of $f(x, \boldsymbol{\phi})$ at concentration $\phi^{(e)}$ for a logarithmic x-axis is equivalent to calculating the slope of the transformed function $f(\tilde{x}, \boldsymbol{\phi})$ with $\tilde{x} = \log(x)$ at the concentration $\phi^{(e)^*} = \log(\phi^{(e)})$, so that the transformed function from the parametrisation in (1) takes the following form:

$$f(\tilde{x}, \boldsymbol{\phi}) = \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \exp\left\{\phi^{(b)}\left[\tilde{x} - \phi^{(e)^*}\right]\right\}}$$

First, the general form of the derivative is calculated:

$$\frac{d}{d\tilde{x}} f(\tilde{x}, \boldsymbol{\phi}) = \frac{-\left(\phi^{(d)} - \phi^{(c)}\right) \cdot \phi^{(b)} \cdot \exp\left\{\phi^{(b)}\left[\tilde{x} - \phi^{(e)^*}\right]\right\}}{\left(1 + \exp\left\{\phi^{(b)}\left[\tilde{x} - \phi^{(e)^*}\right]\right\}\right)^2}$$

Then, the derivative is evaluated at $\tilde{x} = \phi^{(e)^*}$ what leads to the final result:

$$\begin{aligned}\frac{d}{d\tilde{x}} f(\tilde{x}, \boldsymbol{\phi})\Big|_{\tilde{x}=\phi^{(e)^*}} &= \frac{-\left(\phi^{(d)} - \phi^{(c)}\right) \cdot \phi^{(b)} \cdot \exp\left\{\phi^{(b)}\left[\phi^{(e)^*} - \phi^{(e)^*}\right]\right\}}{\left(1 + \exp\left\{\phi^{(b)}\left[\phi^{(e)^*} - \phi^{(e)^*}\right]\right\}\right)^2} \\ &= \frac{-\left(\phi^{(d)} - \phi^{(c)}\right) \cdot \phi^{(b)} \cdot 1}{(1+1)^2} \\ &= -\frac{1}{4} \cdot \phi^{(b)} \cdot \left(\phi^{(d)} - \phi^{(c)}\right)\end{aligned}$$

A.4. Equivalence of the `sigEmax`-model from the MCP-Mod approach and the 4pLL model

The `sigEmax` model from the MCP-Mod approach is parametrised as

$$f(x, \boldsymbol{\theta}) = E_0 + \frac{E_{\max} x^h}{(\mathrm{EC50}^h + x^h)},$$

with E_0 describing the effect for concentration 0, $E_{\max}$ describing the maximal effect, i.e. $\max_x (f(x) - E_0)$, EC50 describing the half-maximal effect with respect to E_0 and $E_{\max}$, and $h > 0$ describing the slope (Bornkamp et al., 2009).

To show the equivalence between the `sigEmax` and the 4pLL model, it needs to be distinguished into the case where $\phi^{(b)} < 0$ and where $\phi^{(b)} < 0$. Without loss of generality, it is assumed that $\phi^{(d)} > \phi^{(c)}$.

In the first scenario it holds $\phi^{(b)} < 0$, which corresponds to an increasing curve under the aforementioned assumptions. Setting $\phi^{(b)} := -h$, $\phi^{(c)} := E_0$, $\phi^{(d)} := E_0 + E_{\max}$ and $\phi^{(e)} := \mathrm{EC50}$ yields

$$\begin{aligned}
f(x, \boldsymbol{\theta}) &= E_0 + \frac{E_{\max} x^h}{(\mathrm{EC50}^h + x^h)} \\
&= E_0 + \frac{E_{\max}}{1 + \left(\frac{\mathrm{EC50}}{x}\right)^h} \\
&\overset{\wedge}{=} \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \left(\frac{\phi^{(e)}}{x}\right)^{-\phi(b)}} \\
&= \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \left(\frac{x}{\phi^{(e)}}\right)^{\phi(b)}} \\
&= f(x, \boldsymbol{\phi})
\end{aligned}$$

in the parametrisation (2).

In the second scenario with $\phi^{(b)} > 0$, this corresponds to a decreasing curve for $\phi^{(d)} > \phi^{(c)}$. Setting $\phi^{(b)} := h$, $\phi^{(d)} := E_0$, $\phi^{(c)} := E_0 + E_{\max}$ and $\phi^{(e)} := \mathrm{EC50}$ yields

$$\begin{aligned}
f(x, \boldsymbol{\theta}) &= E_0 + \frac{E_{\max} x^h}{1 + \left(\frac{\mathrm{EC50}}{x}\right)^h} \\
&\overset{\wedge}{=} \phi^{(d)} + \frac{\phi^{(c)} - \phi^{(d)}}{1 + \left(\frac{\phi^{(e)}}{x}\right)^{\phi(b)}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\phi^{(d)} + \phi^{(d)} \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}} + \phi^{(c)} - \phi^{(d)}}{1 + \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}}} \\
&= \frac{\phi^{(d)} \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}} + \phi^{(c)} \left(1 + \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}}\right) - \phi^{(c)} \left(\frac{(\phi^{(e)}}{x}\right)^{\phi^{(b)}}}{1 + \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}}} \\
&= \phi^{(c)} + \frac{\left(\phi^{(d)} - \phi^{(c)}\right) \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}}}{1 + \left(\frac{\phi^{(e)}}{x}\right)^{\phi^{(b)}}} \\
&= \phi^{(c)} + \frac{\phi^{(d)} - \phi^{(c)}}{1 + \left(\frac{x}{\phi^{(e)}}\right)^{\phi(b)}} \\
&= f(x, \boldsymbol{\phi})
\end{aligned}
$$

in the parametrisation (2). Note that here, $E_{\max} < 0$.

The standardised version of the `sigEmax` model, $f_0(x, \boldsymbol{\theta})$ and the corresponding parametrisation as introduced for the 4pLL model are given as

$$
\begin{aligned}
f^0(x, \boldsymbol{\theta}^0) &= \frac{x^h}{\mathrm{EC50}^h + x^h} \\
&= \frac{1}{1 + \left(\frac{\mathrm{EC50}}{x}\right)^h} \\
&\overset{\wedge}{=} \frac{1}{1 + \left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}},
\end{aligned}
$$

that means that the location parameter θ_0 is given by $\theta_0 = \phi^{(c)}$ and the scale parameter θ_1 is given by $\theta_1 = \phi^{(d)} - \phi^{(c)}$.

A.5. Calculation of $\nabla f(0, \boldsymbol{\phi})$

The following calculations show that

$$\begin{aligned} \nabla f(0, \boldsymbol{\phi}) &= (0, 0, 1, 0)^{\mathrm{T}} \qquad \text{for } \phi^{(b)} > 0, \\ \nabla f(0, \boldsymbol{\phi}) &= (0, 1, 0, 0)^{\mathrm{T}} \qquad \text{for } \phi^{(b)} < 0 \end{aligned}$$

where

$$\nabla f(x, \boldsymbol{\phi}) = \left(\tfrac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(b)}}, \tfrac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(c)}}, \tfrac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(d)}}, \tfrac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(e)}} \right)^{\mathrm{T}}$$

$$= \begin{pmatrix} -\dfrac{\left(\phi^{(d)} - \phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} \\ 1 - \left[\dfrac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}\right] \\ \dfrac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}} \\ \dfrac{\phi^{(b)}\left(\phi^{(d)} - \phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} \end{pmatrix}$$

As $\log(\frac{x}{\phi^{(e)}})$ cannot be calculated for $x = 0$, instead the limit $\lim_{x \to 0} \nabla f(x, \boldsymbol{\phi})$ is considered for each partial derivative individually. To calculate the limits, in some cases *L'Hôpital's rule* will be employed, that is introduced here according to Forster (2016, p. 190).

Theorem (L'Hôpital's rule)**.** *Let $f, g : I \to \mathbb{R}$ two differentiable functions on the open interval $I = (a, b)$ with $-\infty \leq a < b \leq \infty$. Assume $\forall x \in I, g'(x) \neq 0$ and the limit $\lim_{x \searrow a} \frac{f'(x)}{g'(x)} := c \in \mathbb{R}$ exists. Then:*

1. *If $\lim_{x \searrow a} g(x) = \lim_{x \searrow a} f(x) = 0$, then $\forall x \in I$ it holds $g(x) \neq 0$ and*

$$\lim_{x \searrow a} \frac{f(x)}{g(x)} = c.$$

2. *If $\lim_{x \searrow a} g(x) = \pm\infty$, then $\exists x_0$ with $a < x_0 < b$ s.t. $\forall x \geq x_0, g(x) \neq 0$ and*

$$\lim_{x \searrow a} \frac{f(x)}{g(x)} = c.$$

First, the gradient $\nabla f(0, \boldsymbol{\phi})$ is calculated for $\phi^{(b)} > 0$, i.e. with $\lim_{x \to 0} f(x, \boldsymbol{\phi}) = \phi^{(d)}$.

- $\frac{\partial f(x, \boldsymbol{\phi})}{\partial \phi^{(b)}} = -\frac{\left(\phi^{(d)} - \phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2}$

  First, consider the denominator:

$$\lim_{x \to 0}\left[1 + \underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\substack{x \to 0 \\ \longrightarrow 0}}\right]^2 = 1$$

  Then consider the nominator while momentarily dismissing the term $\left(\phi^{(d)} - \phi^{(c)}\right)$, using L'Hôpital's rule:

$$\begin{aligned}
\lim_{x \to 0}\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}} &= \lim_{x \to 0} \frac{\log\left(\frac{x}{\phi^{(e)}}\right)}{\left(\frac{x}{\phi^{(e)}}\right)^{-\phi^{(b)}}} \\
&= \lim_{x \to 0} \frac{\frac{\mathrm{d}}{\mathrm{d}x}\log\left(\frac{x}{\phi^{(e)}}\right)}{\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x}{\phi^{(e)}}\right)^{-\phi^{(b)}}} \\
&= \lim_{x \to 0} \frac{\frac{1}{x}}{-\phi^{(b)}\left(\frac{x}{\phi^{(e)}}\right)^{-\phi^{(b)}-1}} \\
&= \lim_{x \to 0} -\frac{1}{\phi^{(b)}}\left(\phi^{(e)}\right)^{\phi^{(b)}-1} \cdot \underbrace{x^{\phi^{(b)}}}_{\substack{x \to 0 \\ \longrightarrow 0}} \\
&= 0 \qquad (4)
\end{aligned}$$

  All in all it follows:

$$\lim_{x \to 0} -\frac{\left(\phi^{(d)} - \phi^{(c)}\right)\overbrace{\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}^{\substack{x \to 0 \\ \longrightarrow 0}}}{\left[1 + \underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\substack{x \to 0 \\ \longrightarrow 0}}\right]^2} = -\frac{\left(\phi^{(d)} - \phi^{(c)}\right) \cdot 0}{(1+0)^2} = 0.$$

- $\frac{\partial f(x, \boldsymbol{\phi})}{\partial \phi^{(c)}} = 1 - \left[\frac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}\right]$

The limit for $x \to 0$ is given by:

$$\lim_{x \to 0} 1 - \left[\frac{1}{1 + \underbrace{\left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}_{\overset{x \to 0}{\longrightarrow} 0}} \right] = 1 - \frac{1}{1+0} = 0$$

- $\frac{\partial f(x, \boldsymbol{\phi})}{\partial \phi^{(d)}} = \frac{1}{1 + \left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}$

  The limit for $x \to 0$ is given by:

$$\lim_{x \to 0} \frac{1}{1 + \underbrace{\left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}_{\overset{x \to 0}{\longrightarrow} 0}} = \frac{1}{1+0} = 1$$

- $\frac{\partial f(x, \boldsymbol{\phi})}{\partial \phi^{(e)}} = \frac{\phi^{(b)} \left(\phi^{(d)} - \phi^{(c)} \right) \left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}{\phi^{(e)} \left[1 + \left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}} \right]^2}$

  The limit for $x \to 0$ is given by:

$$\lim_{x \to 0} \frac{\phi^{(b)} \left(\phi^{(d)} - \phi^{(c)} \right) \overbrace{\left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}^{\overset{x \to 0}{\longrightarrow} 0}}{\phi^{(e)} \left[1 + \underbrace{\left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}_{\overset{x \to 0}{\longrightarrow} 0} \right]^2} = \frac{\phi^{(b)} \left(\phi^{(d)} - \phi^{(c)} \right) \cdot 0}{\phi^{(e)} (1+0)^2} = 0$$

All in all it follows that for $\phi^{(b)} > 0$, $\nabla f(0, \boldsymbol{\phi}) = (0, 0, 1, 0)^{\mathrm{T}}$.

Then, the gradient $\nabla f(0, \boldsymbol{\phi})$ is calculated for $\phi^{(b)} < 0$, i.e. with $\lim_{x \to 0} f(x, \boldsymbol{\phi}) = \phi^{(c)}$.

- $\frac{\partial f(x, \boldsymbol{\phi})}{\partial \phi^{(b)}} = -\frac{\left(\phi^{(d)} - \phi^{(c)} \right) \left(\log\left(\frac{x}{\phi^{(e)}} \right) \right) \left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}}}{\left[1 + \left(\frac{x}{\phi^{(e)}} \right)^{\phi^{(b)}} \right]^2}$

The limit for $x \to 0$ is given by:

$$-\frac{\left(\phi^{(d)}-\phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} \overset{(*)}{>} -\frac{\left(\phi^{(d)}-\phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left[\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2},$$

$$\lim_{x\to 0} -\frac{\left(\phi^{(d)}-\phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\left[\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} = \lim_{x\to 0} -\frac{\left(\phi^{(d)}-\phi^{(c)}\right)\left(\log\left(\frac{x}{\phi^{(e)}}\right)\right)}{\left[\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]}$$

From (4) it is known that this converges towards 0 as $x \to 0$. Since all quotients are strictly smaller than 0 and with $(*)$ holding as $\phi^{(e)} > 0$, it then follows that the limit of the derivative is given by 0 as well.

- $\frac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(c)}} = 1 - \left[\frac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}\right]$

  The limit for $x \to 0$ is given by:

$$\lim_{x\to 0} 1 - \left[\frac{1}{1+\underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to 0}{\longrightarrow}\infty}}\right] = 1 - 0 = 1$$

- $\frac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(d)}} = \frac{1}{1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}$

  The limit for $x \to 0$ is given by:

$$\lim_{x\to 0} \frac{1}{1+\underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to 0}{\longrightarrow}\infty}} = 0$$

- $\frac{\partial f(x,\boldsymbol{\phi})}{\partial \phi^{(e)}} = \frac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2}$

The limit for $x \to 0$ is given by:

$$\frac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} \overset{(*)}{<} \frac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2},$$

$$\lim_{x\to 0}\frac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} = \lim_{x\to 0}\frac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)}{\phi^{(e)}\left[\underbrace{\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}_{\overset{x\to 0}{\longrightarrow}\infty}\right]} = 0$$

$$\Rightarrow \lim_{x\to 0}\frac{\phi^{(b)}\left(\phi^{(d)}-\phi^{(c)}\right)\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}}{\phi^{(e)}\left[1+\left(\frac{x}{\phi^{(e)}}\right)^{\phi^{(b)}}\right]^2} = 0,$$

since all quotients considered are strictly larger than 0 and with $(*)$ holding as $\phi^{(e)} > 0$.

A.6. Calculation of the posterior in a normal-normal model

Let $\boldsymbol{X}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2)$ with the prior distribution $\pi(\boldsymbol{\theta})$ given by $\boldsymbol{\theta} \sim \mathcal{N}(\mu, \tau^2)$. Then the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ from equation (19) in Chapter 4.4.3 is given by

$$\boldsymbol{\theta}|x \sim \mathcal{N}\left(\frac{\tau^2 x + \sigma^2\mu}{\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right).$$

According to formula (18) it holds

$$p(\boldsymbol{\theta}|x) = \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

Start by calculating $f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$:

$$\begin{aligned} f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\tau^2}}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left[\frac{(\boldsymbol{\theta}-\mu)^2}{\tau^2}+\frac{(x-\boldsymbol{\theta})^2}{\sigma^2}\right]^2\right) \\ &\overset{(*)}{=} \frac{1}{\sqrt{2\pi}\sqrt{2\pi\tau^2\sigma^2}}\exp\left(-\frac{1}{2}\frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left[\boldsymbol{\theta}-\frac{\tau^2 x+\sigma^2\mu}{\tau^2+\sigma^2}\right]^2\right)\exp\left(-\frac{1}{2}\frac{(\mu-x)^2}{\tau^2+\sigma^2}\right) \end{aligned}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}}\exp\left(-\frac{1}{2}\frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left[\boldsymbol{\theta}-\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2}\right]^2\right)}_{\text{Density of a } \mathcal{N}\left(\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2},\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right) \text{ distribution}}$$
$$\times \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}}\exp\left(-\frac{1}{2}\frac{(\mu-x)^2}{\tau^2+\sigma^2}\right)$$

Thus it holds:

$$\begin{aligned}
p(\boldsymbol{\theta}|x) &= \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\
&= \frac{\frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}}\exp\left(-\frac{1}{2}\frac{(\mu-x)^2}{\tau^2+\sigma^2}\right)}{\frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}}\exp\left(-\frac{1}{2}\frac{(\mu-x)^2}{\tau^2+\sigma^2}\right)} \\
&\quad \times \frac{\frac{1}{\sqrt{2\pi\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}}\exp\left(-\frac{1}{2}\frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left[\boldsymbol{\theta}-\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2}\right]^2\right)}{\underbrace{\int \frac{1}{\sqrt{2\pi\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}}\exp\left(-\frac{1}{2}\frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left[\boldsymbol{\theta}-\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2}\right]^2\right)\mathrm{d}\boldsymbol{\theta}}_{=1}} \\
&= \frac{1}{\sqrt{2\pi\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}}\exp\left(-\frac{1}{2}\frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left[\boldsymbol{\theta}-\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2}\right]^2\right),
\end{aligned}$$

which is the density of a $\mathcal{N}\left(\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2},\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right)$ distribution, what was to be shown.

It remains to show that $(*)$ holds. Therefore the following equation needs to be proved:

$$\frac{(\boldsymbol{\theta}-\mu)^2}{\tau^2}+\frac{(x-\boldsymbol{\theta})^2}{\sigma^2} = \frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left(\boldsymbol{\theta}-\frac{\tau^2x+\sigma^2\mu}{\tau^2+\sigma^2}\right)^2+\frac{(\mu-x)^2}{\tau^2+\sigma^2}$$

Start on the left side:

$$\begin{aligned}
\frac{(\boldsymbol{\theta}-\mu)^2}{\tau^2}\frac{\sigma^2}{\sigma^2}+\frac{(x-\boldsymbol{\theta})^2}{\sigma^2}\frac{\tau^2}{\tau^2} &= \frac{\boldsymbol{\theta}^2(\tau^2+\sigma^2)-2\boldsymbol{\theta}\mu\sigma^2+\mu^2\sigma^2-2\boldsymbol{\theta}x\tau^2+x^2\tau^2}{\tau^2\sigma^2} \\
&= \underbrace{\frac{\boldsymbol{\theta}^2(\tau^2+\sigma^2)}{\tau^2\sigma^2}}_{(1)}-\underbrace{\frac{2\boldsymbol{\theta}\mu\sigma^2+2\boldsymbol{\theta}x\tau^2}{\tau^2\sigma^2}}_{(2)}+\underbrace{\frac{x^2\tau^2+\mu^2+\sigma^2}{\tau^2\sigma^2}}_{(3)}
\end{aligned}$$

Continue on the right side:

$$
\begin{aligned}
&\frac{1}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\left(\boldsymbol{\theta}-\frac{\tau^2 x+\sigma^2\mu}{\tau^2+\sigma^2}\right)^2+\frac{(\mu-x)^2}{\tau^2+\sigma^2}\\
&=\frac{\tau^2+\sigma^2}{\tau^2\sigma^2}\left(\frac{\boldsymbol{\theta}(\tau^2+\sigma^2)-(\tau^2 x+\sigma^2\mu)}{\tau^2+\sigma^2}\right)^2+\frac{(\mu^2-2\mu x+x^2)(\tau^2\sigma^2)}{(\tau^2+\sigma^2)(\tau^2)(\sigma^2)}\\
&=\frac{1}{\tau^2\sigma^2}\frac{1}{\tau^2+\sigma^2}\Big(\boldsymbol{\theta}(\tau^2+\sigma^2)-2\boldsymbol{\theta}(\tau^2+\sigma^2)(\tau^2 x+\sigma^2\mu)+(\tau^2 x+\sigma^2\mu)\Big)\\
&\quad+\frac{\mu^2\tau^2\sigma^2-2\mu x\tau^2\sigma^2+x^2\tau^2\sigma^2}{(\tau^2+\sigma^2)(\tau^2)(\sigma^2)}\\
&=\underbrace{\frac{\boldsymbol{\theta}^2(\tau^2+\sigma^2)}{\tau^2\sigma^2}}_{(1)}-\underbrace{\frac{2\boldsymbol{\theta}\mu\sigma^2+2\boldsymbol{\theta}x\tau^2}{\tau^2\sigma^2}}_{(2)}+\underbrace{\frac{\tau^4x^2+2\tau^2x\sigma^2\mu+\sigma^4\mu^2-2\tau^2\sigma^2\mu x+\tau^2\sigma^2x^2}{(\tau^2+\sigma^2)(\tau^2\sigma^2)}}_{=\frac{x^2\tau^2(\tau^2+\sigma^2)+\mu^2\sigma^2(\tau^2+\sigma^2)}{(\tau^2+\sigma^2)(\tau^2\sigma^2)}=\frac{x^2\tau^2+\mu^2+\sigma^2}{\tau^2\sigma^2}=(3)}
\end{aligned}
$$

This completes the calculation of the posterior distribution for the given likelihood and prior distribution.

B. Figures

B.1. Handling deviating control values

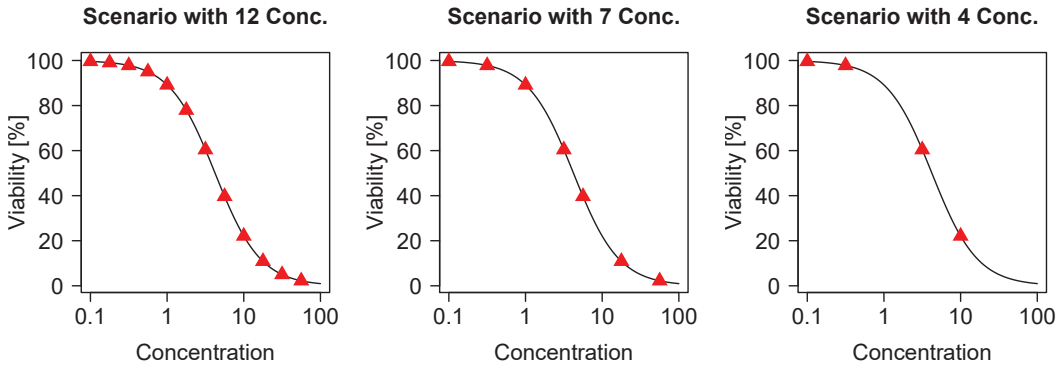

Figure B.1: Three additional scenarios considered in the simulation study with 12, 7 and 4 concentrations, respectively. The true underlying curve is the same as presented in Figure 5.2.

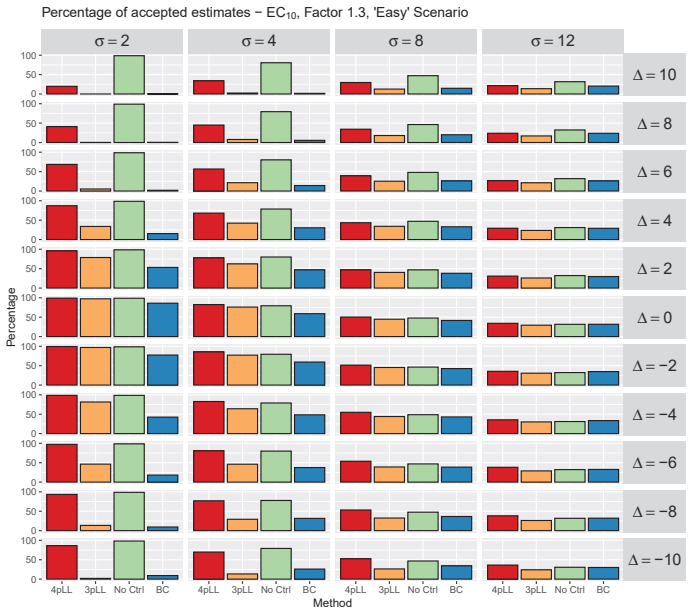

Figure B.2: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{10} in the 'easy' scenario and are structured as explained in Figure 5.4.

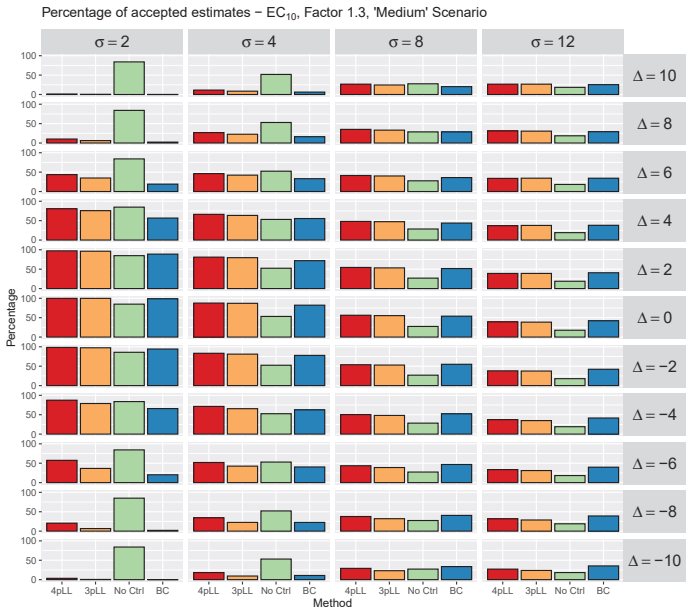

Figure B.3: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{10} in the 'medium' scenario and are structured as explained in Figure 5.4.

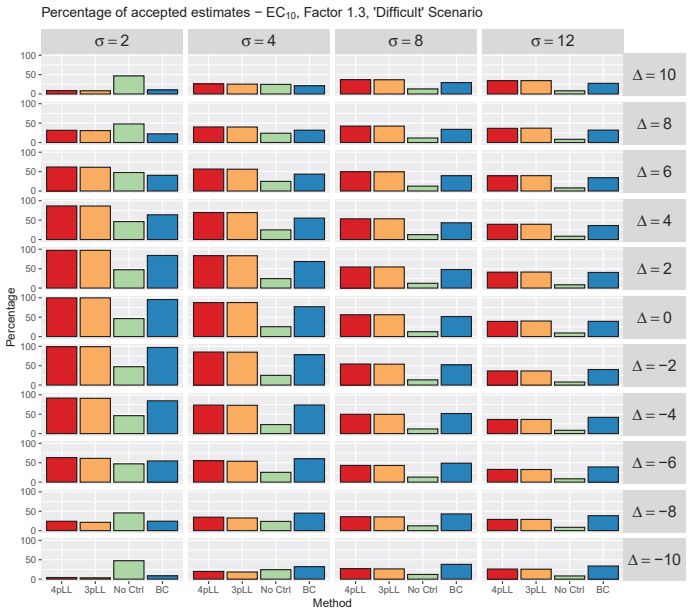

Figure B.4: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{10} in the 'difficult' scenario and are structured as explained in Figure 5.4.

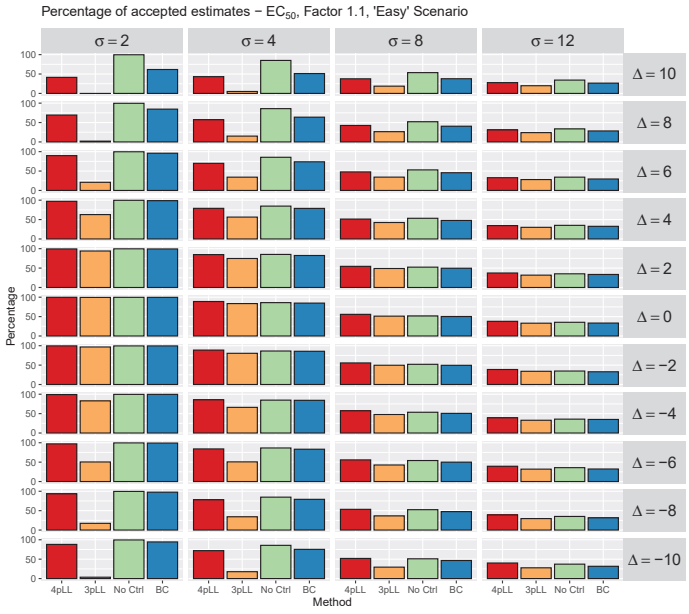

Figure B.5: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{50} in the 'easy' scenario and are structured as explained in Figure 5.4 with the exception of the factor defining the acceptable range, which is chosen as 1.1 here.

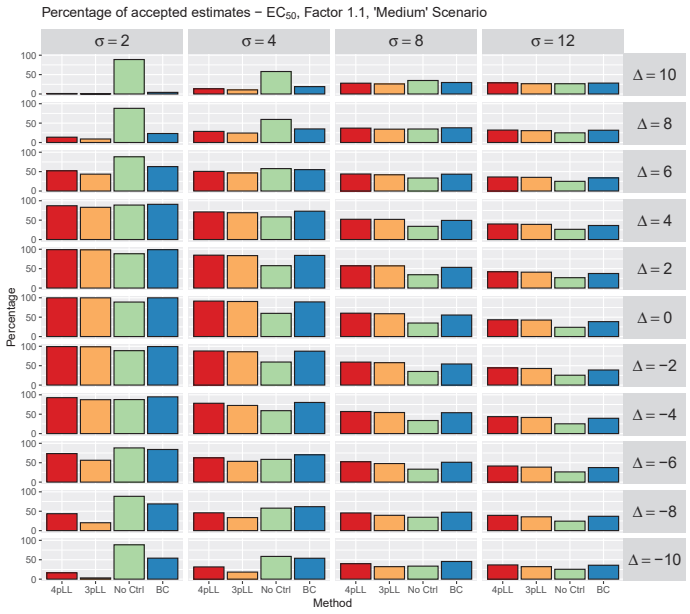

Figure B.6: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{50} in the 'medium' scenario and are structured as explained in Figure 5.4 with the exception of the factor defining the acceptable range, which is chosen as 1.1 here.

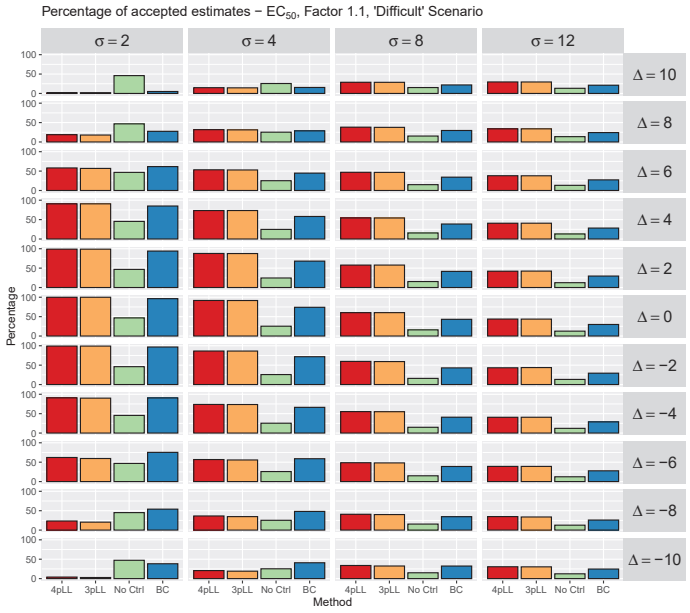

Figure B.7: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{50} in the 'difficult' scenario and are structured as explained in Figure 5.4 with the exception of the factor defining the acceptable range, which is chosen as 1.1 here.

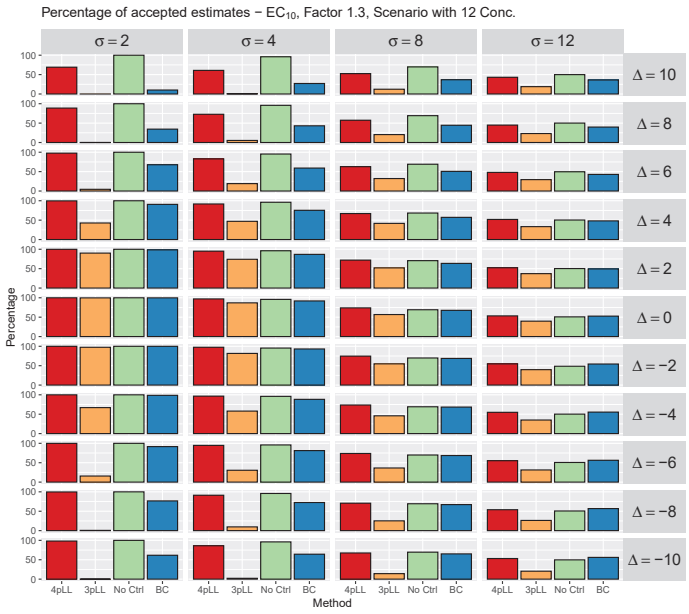

Figure B.8: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{10} in the scenario with 12 concentrations and are structured as explained in Figure 5.4.

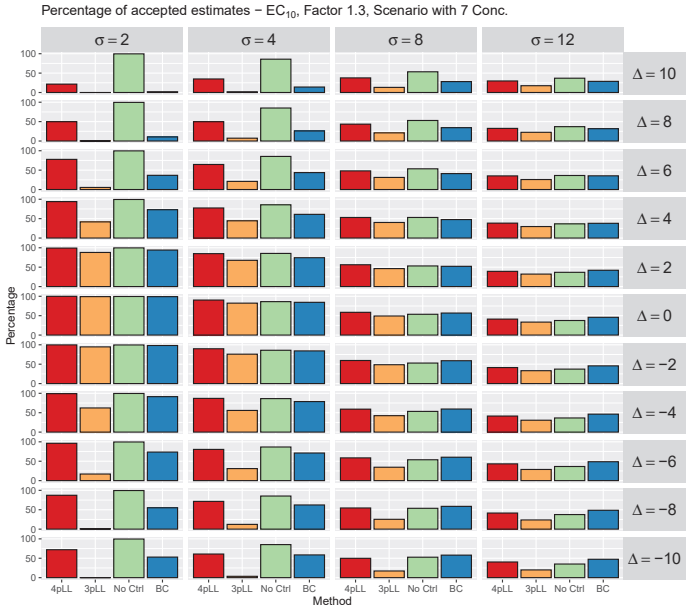

Figure B.9: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{10} in the scenario with 7 concentrations and are structured as explained in Figure 5.4.

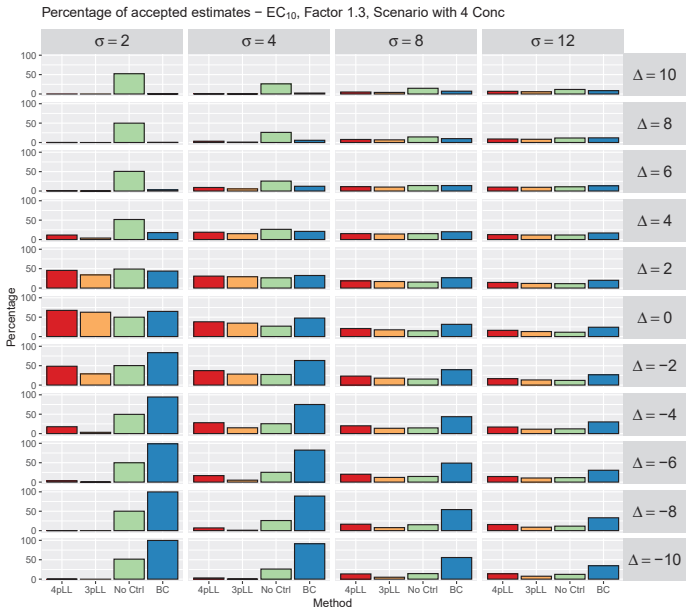

Figure B.10: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{10} in the scenario with 4 concentrations and are structured as explained in Figure 5.4.

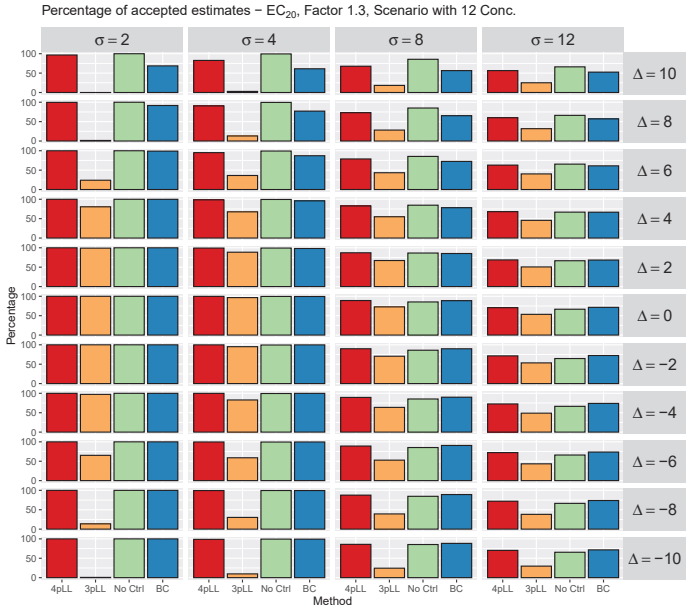

Figure B.11: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{20} in the scenario with 12 concentrations and are structured as explained in Figure 5.4.

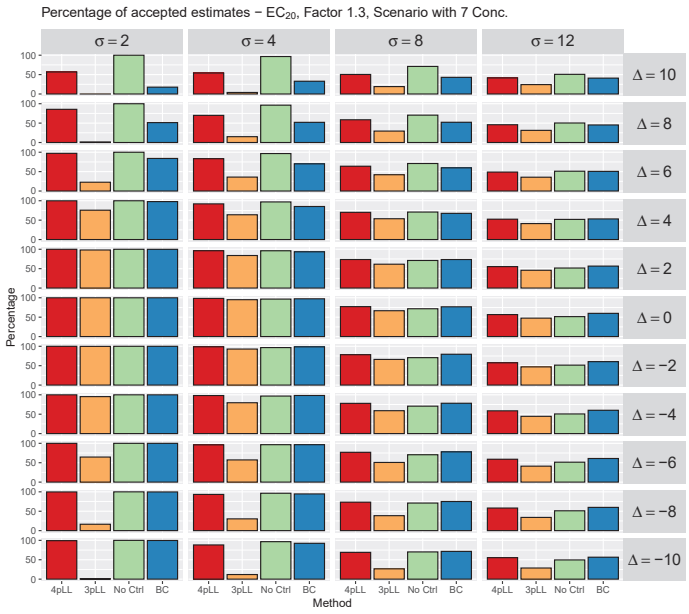

Figure B.12: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{20} in the scenario with 7 concentrations and are structured as explained in Figure 5.4.

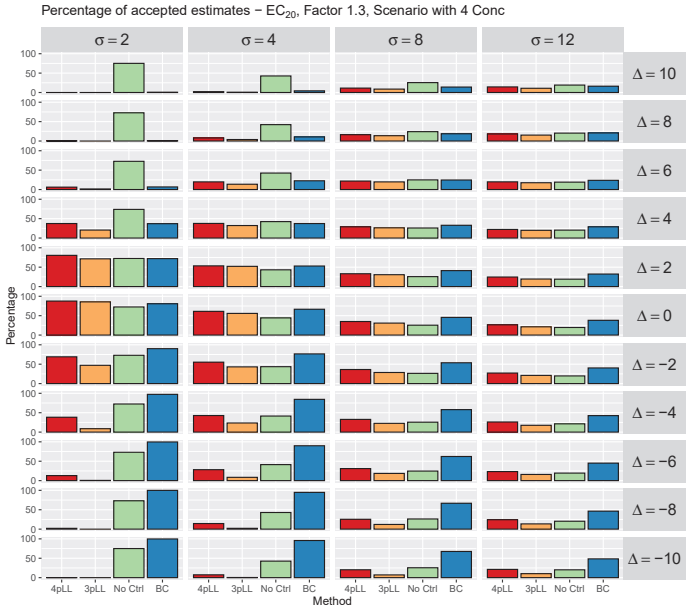

Figure B.13: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{20} in the scenario with 4 concentrations and are structured as explained in Figure 5.4.

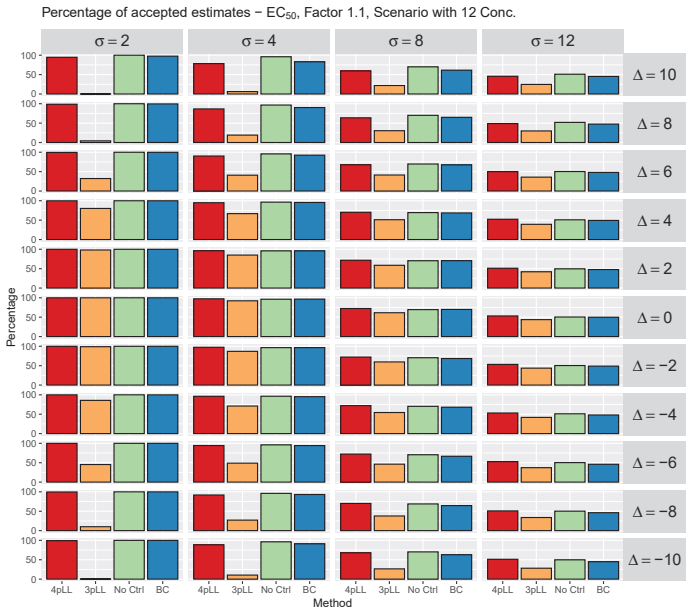

Figure B.14: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{50} in the scenario with 12 concentrations and are structured as explained in Figure 5.4 with the exception of the factor defining the acceptable range, which is chosen as 1.1 here.

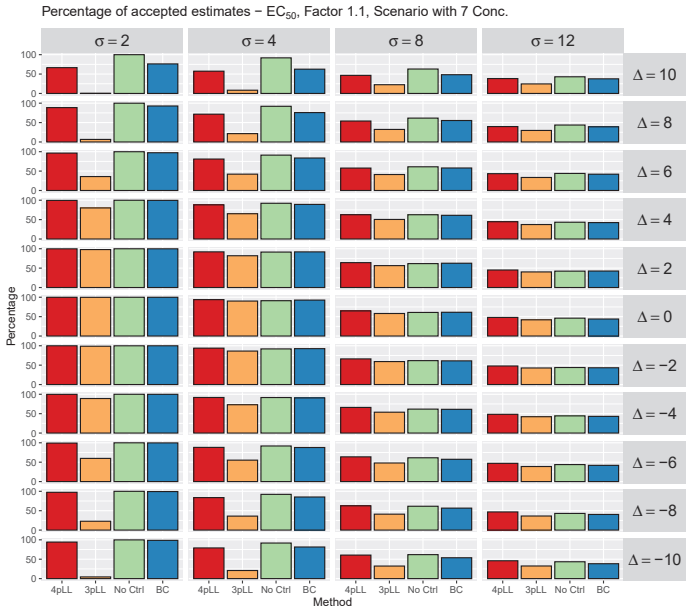

Figure B.15: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{50} in the scenario with 7 concentrations and are structured as explained in Figure 5.4 with the exception of the factor defining the acceptable range, which is chosen as 1.1 here.

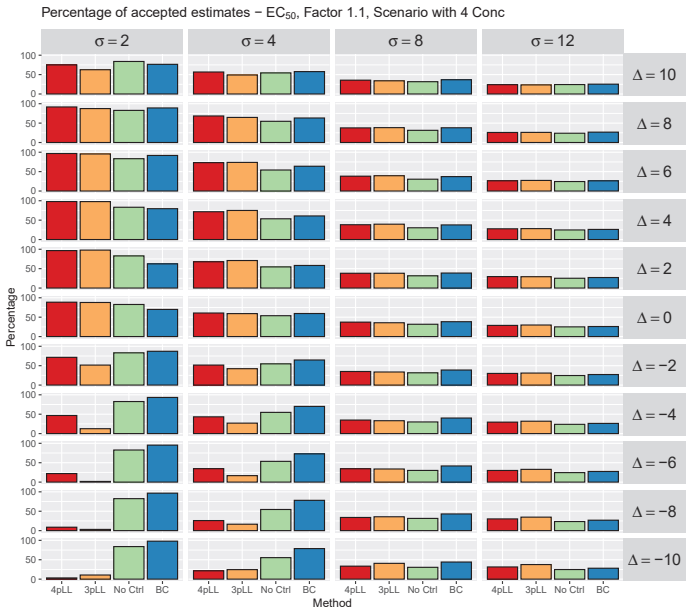

Figure B.16: Percentages of acceptable estimates for the 5000 iterations of the simulation study. Results are shown here for the EC_{50} in the scenario with 4 concentrations and are structured as explained in Figure 5.4 with the exception of the factor defining the acceptable range, which is chosen as 1.1 here.

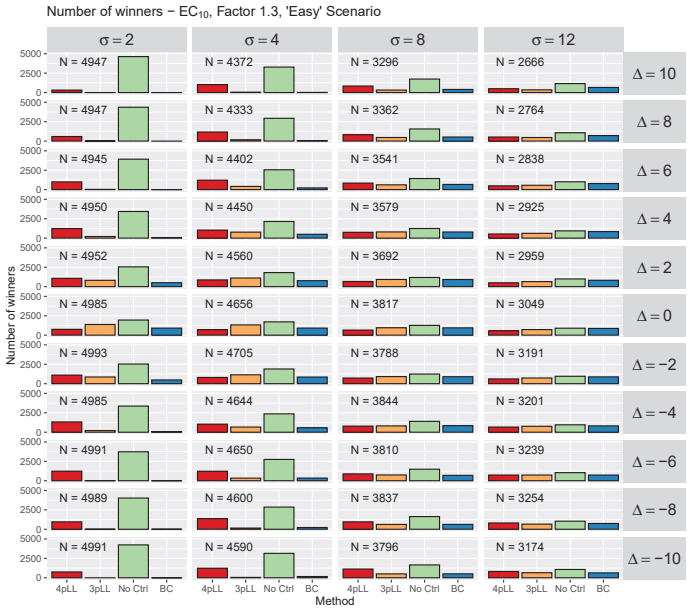

Figure B.17: Number of times each method is the winner. Results are shown here for the EC_{10} in the 'easy' scenario and are structured as explained in Figure 5.7.

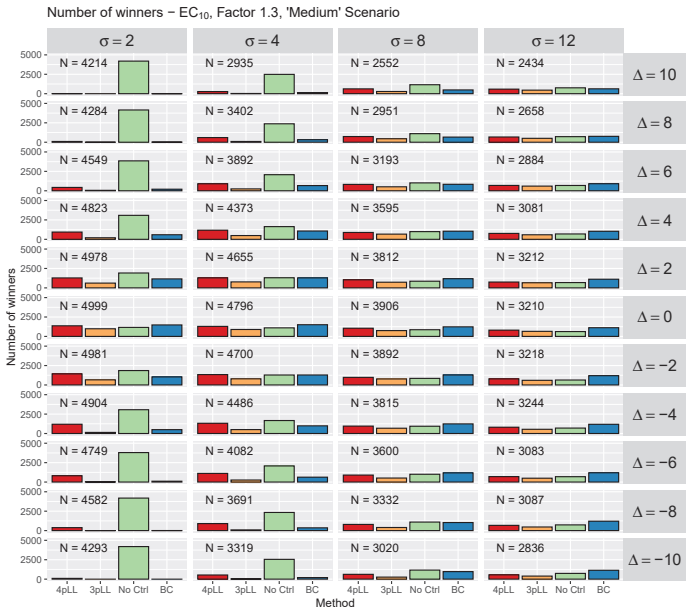

Figure B.18: Number of times each method is the winner. Results are shown here for the EC_{10} in the 'medium' scenario and are structured as explained in Figure 5.7.

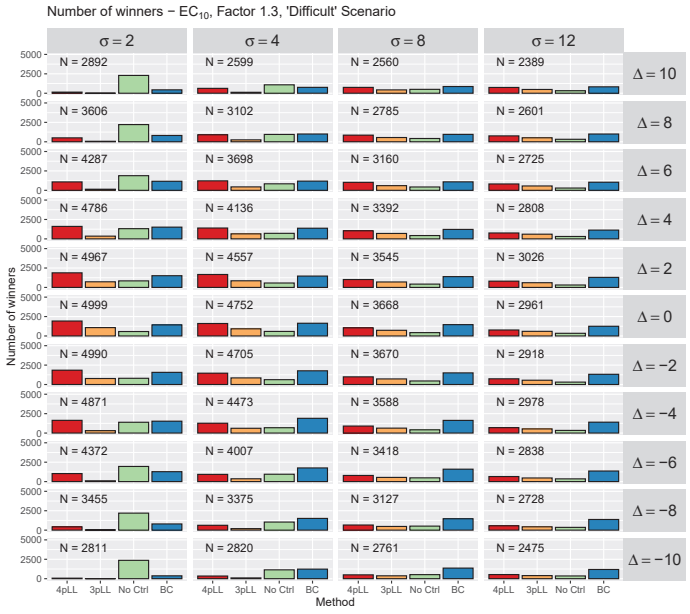

Figure B.19: Number of times each method is the winner. Results are shown here for the EC_{10} in the 'difficult' scenario and are structured as explained in Figure 5.7.

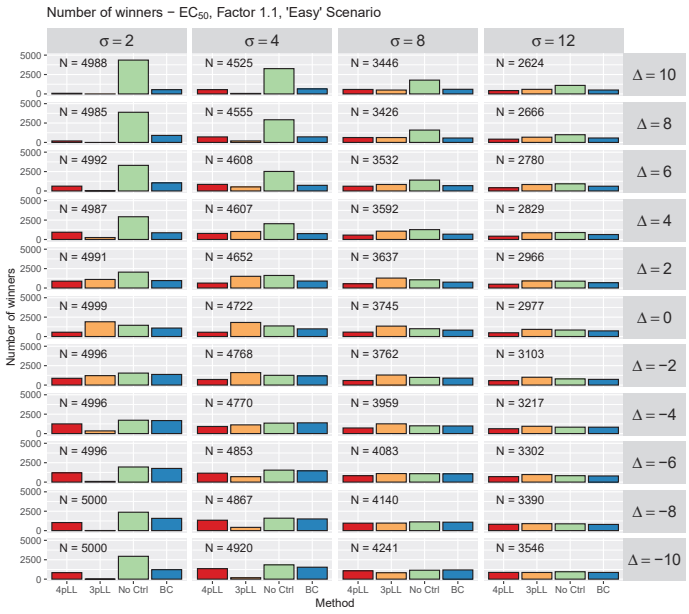

Figure B.20: Number of times each method is the winner. Results are shown here for the EC_{50} in the 'easy' scenario and are structured as explained in Figure 5.7 with the exception of the factor defining the acceptable rangen, which is chosen as 1.1 here.

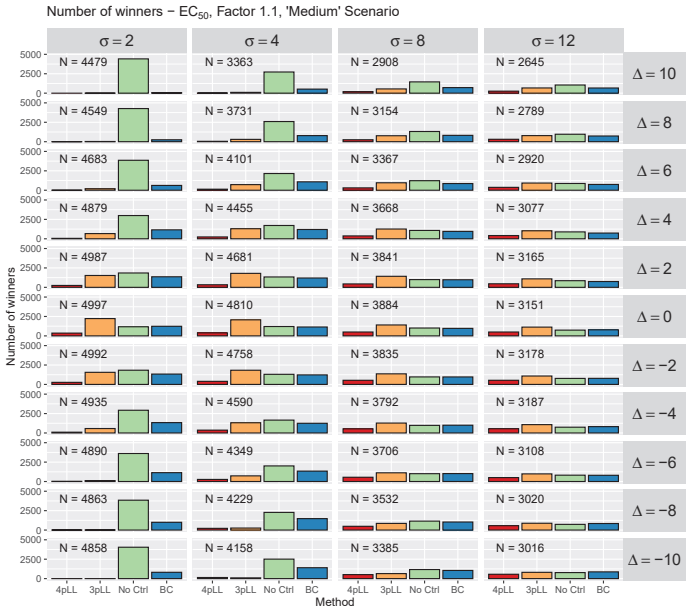

Figure B.21: Number of times each method is the winner. Results are shown here for the EC_{50} in the 'medium' scenario and are structured as explained in Figure 5.7 with the exception of the factor defining the acceptable rangen, which is chosen as 1.1 here.

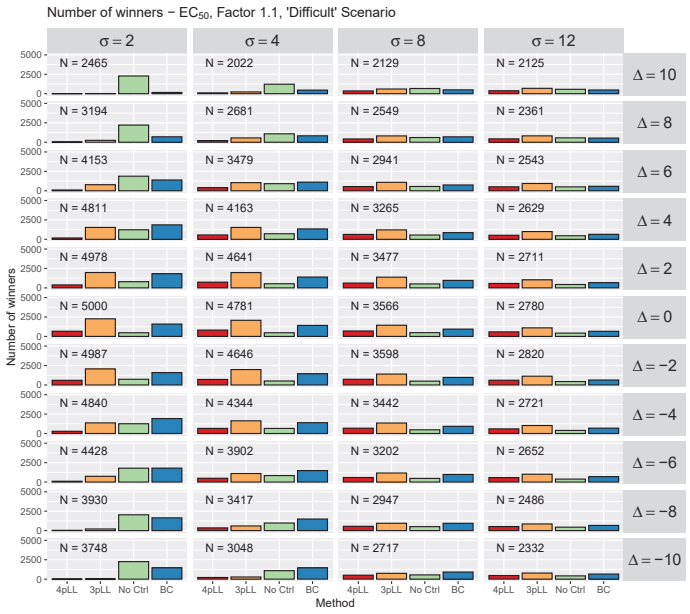

Figure B.22: Number of times each method is the winner. Results are shown here for the EC_{50} in the 'difficult' scenario and are structured as explained in Figure 5.7 with the exception of the factor defining the acceptable rangen, which is chosen as 1.1 here.

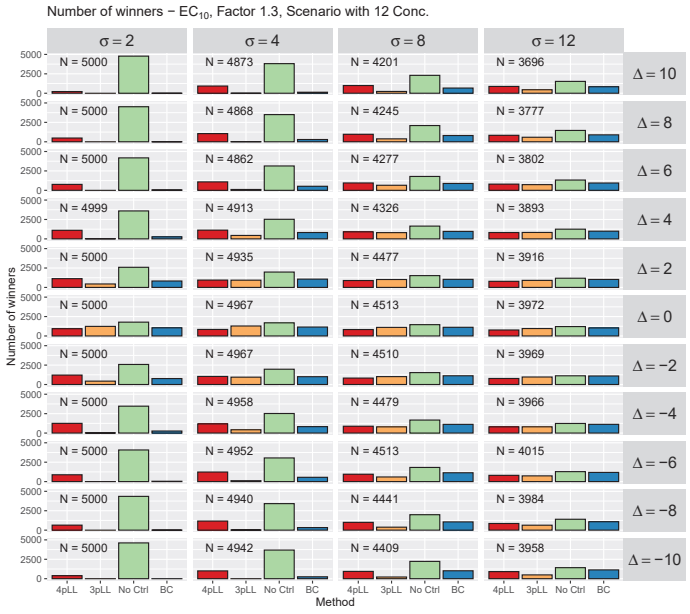

Figure B.23: Number of times each method is the winner. Results are shown here for the EC_{10} in the scenario with 12 concentrations and are structured as explained in Figure 5.7.

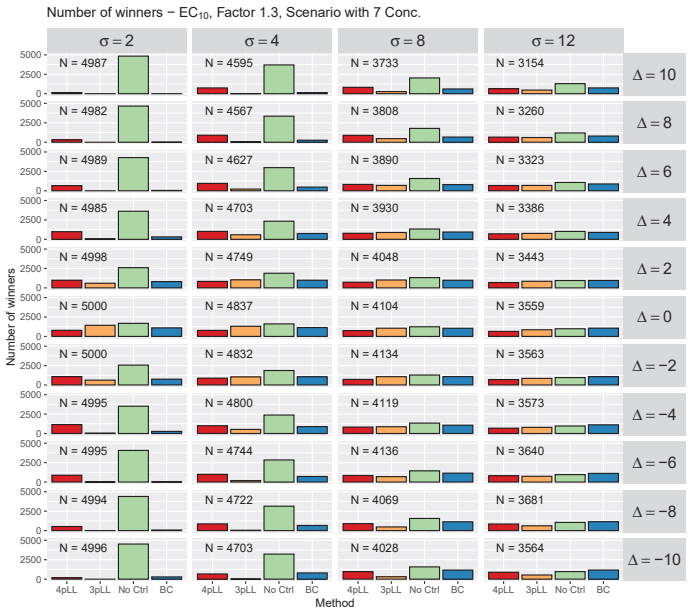

Figure B.24: Number of times each method is the winner. Results are shown here for the EC_{10} in the scenario with 7 concentrations and are structured as explained in Figure 5.7.

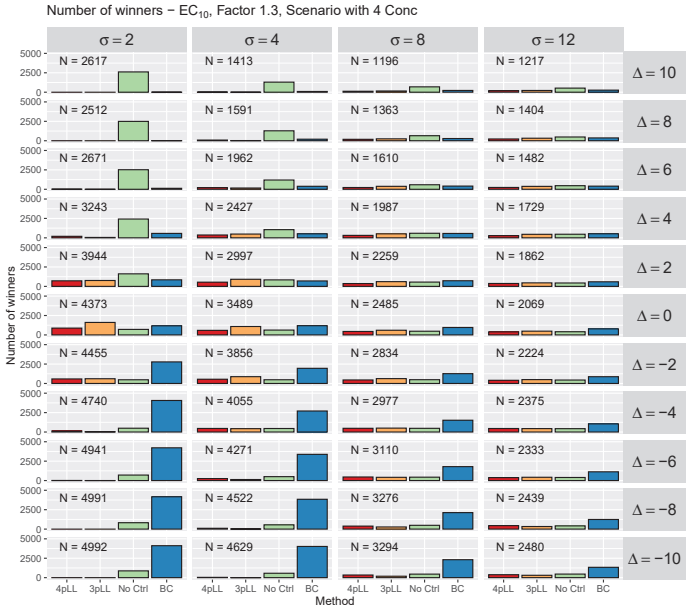

Figure B.25: Number of times each method is the winner. Results are shown here for the EC_{10} in the scenario with 4 concentrations and are structured as explained in Figure 5.7.

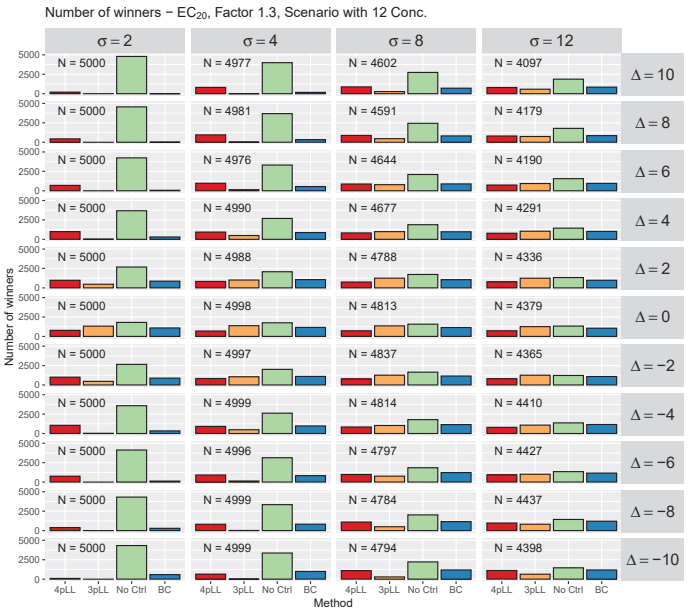

Figure B.26: Number of times each method is the winner. Results are shown here for the EC_{20} in the scenario with 12 concentrations and are structured as explained in Figure 5.7.

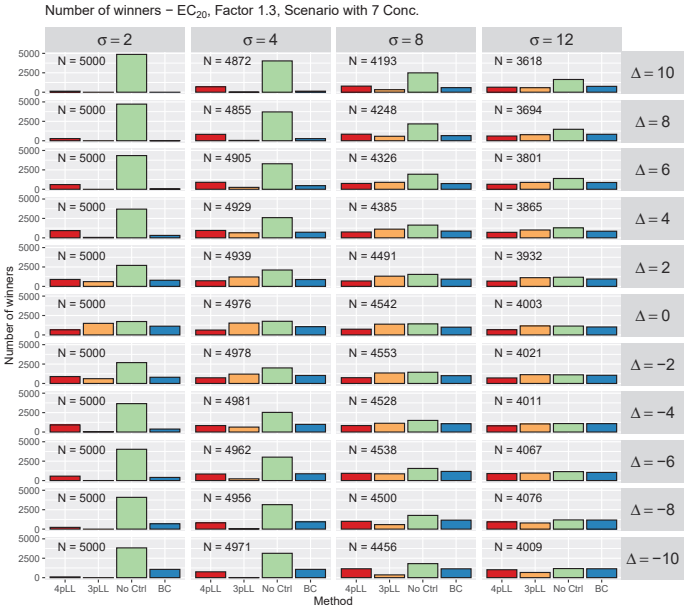

Figure B.27: Number of times each method is the winner. Results are shown here for the EC_{20} in the scenario with 7 concentrations and are structured as explained in Figure 5.7.

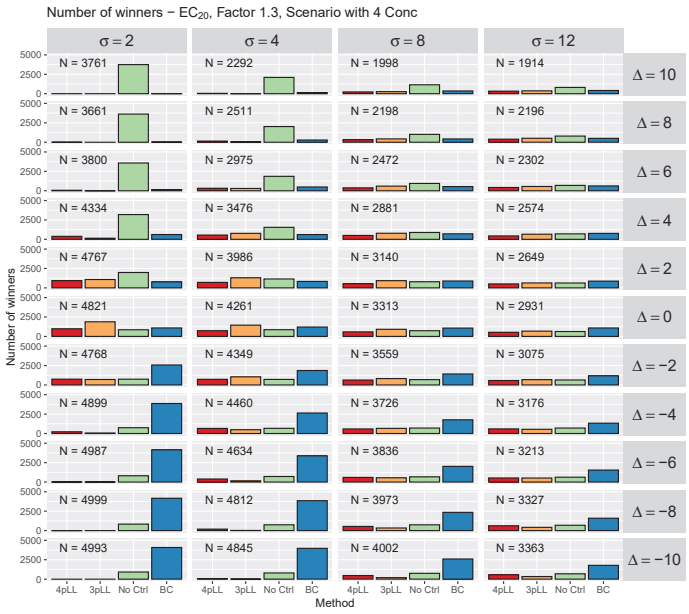

Figure B.28: Number of times each method is the winner. Results are shown here for the EC_{20} in the scenario with 4 concentrations and are structured as explained in Figure 5.7.

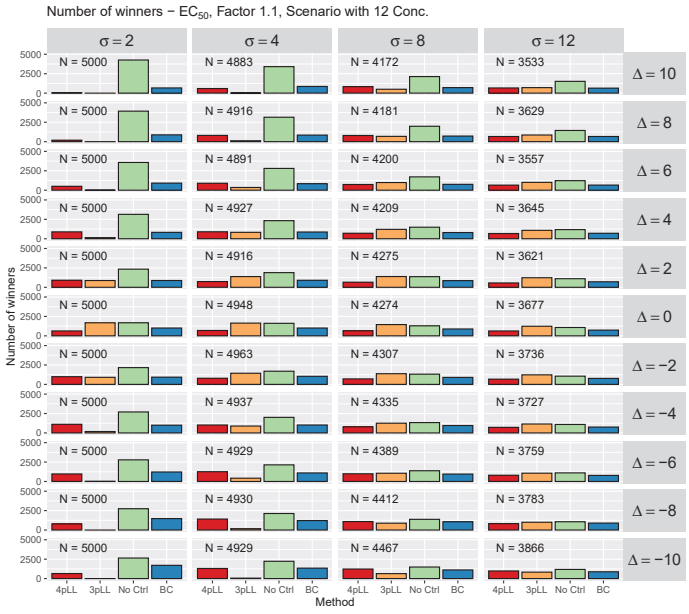

Figure B.29: Number of times each method is the winner. Results are shown here for the EC_{50} in the scenario with 12 concentrations and are structured as explained in Figure 5.7 with the exception of the factor defining the acceptable rangen, which is chosen as 1.1 here.

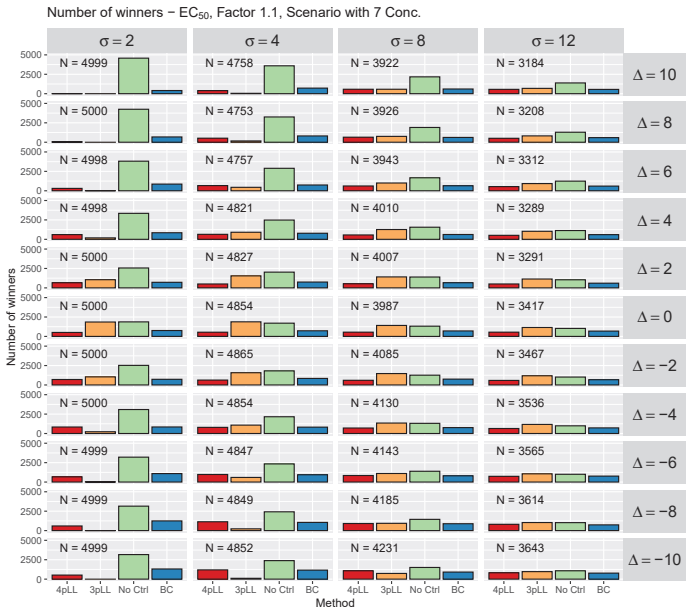

Figure B.30: Number of times each method is the winner. Results are shown here for the EC_{50} in the scenario with 7 concentrations and are structured as explained in Figure 5.7 with the exception of the factor defining the acceptable rangen, which is chosen as 1.1 here.

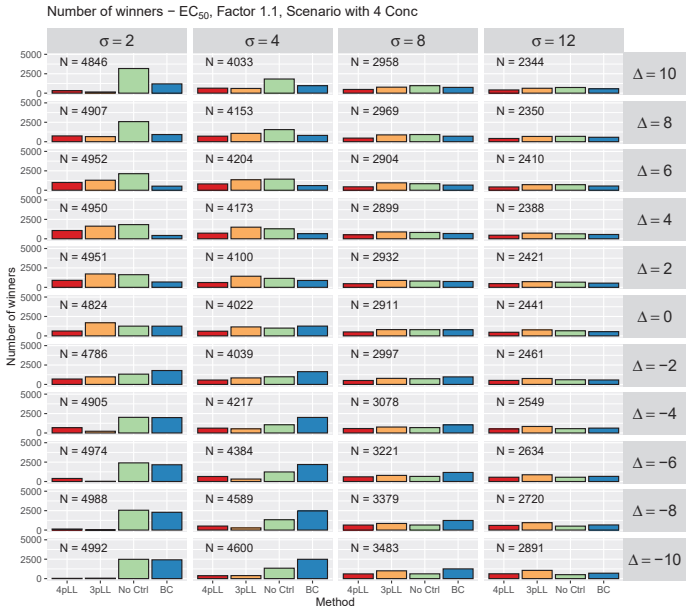

Figure B.31: Number of times each method is the winner. Results are shown here for the EC_{50} in the scenario with 4 concentrations and are structured as explained in Figure 5.7 with the exception of the factor defining the acceptable rangen, which is chosen as 1.1 here.

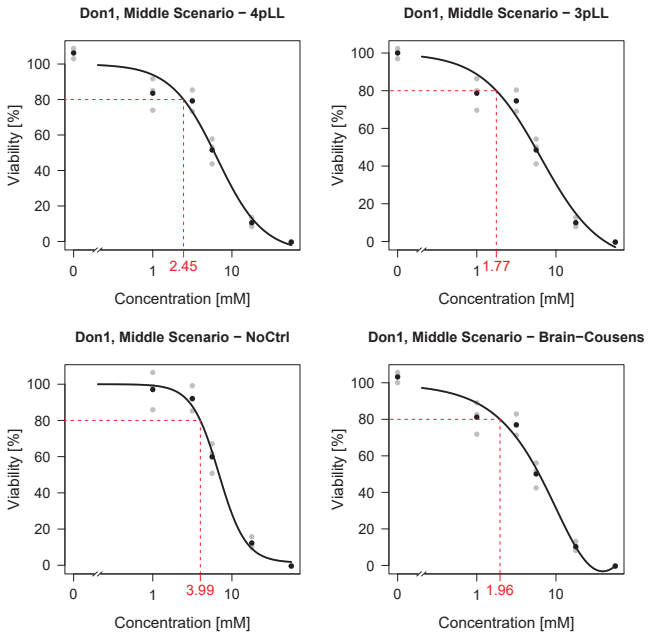

Figure B.32: Application of the four methods to the original dataset, Don1, resembling the 'medium' situation.

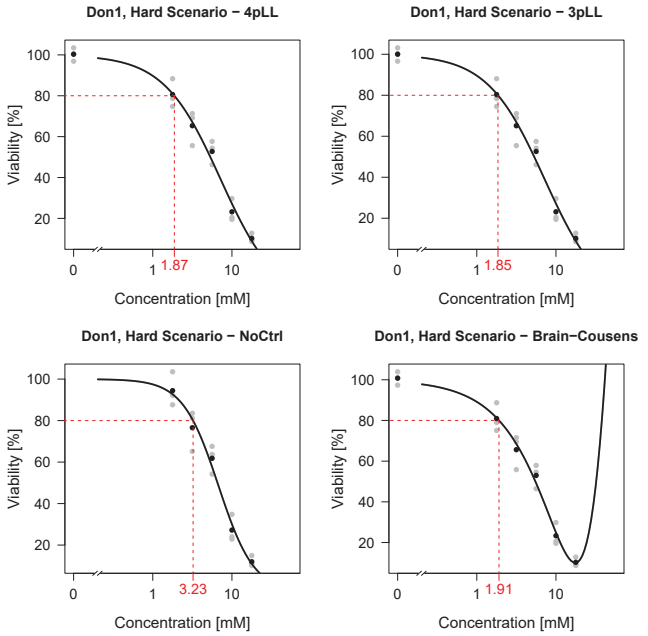

Figure B.33: Application of the four methods to the original dataset, Don1, resembling the 'difficult' situation.

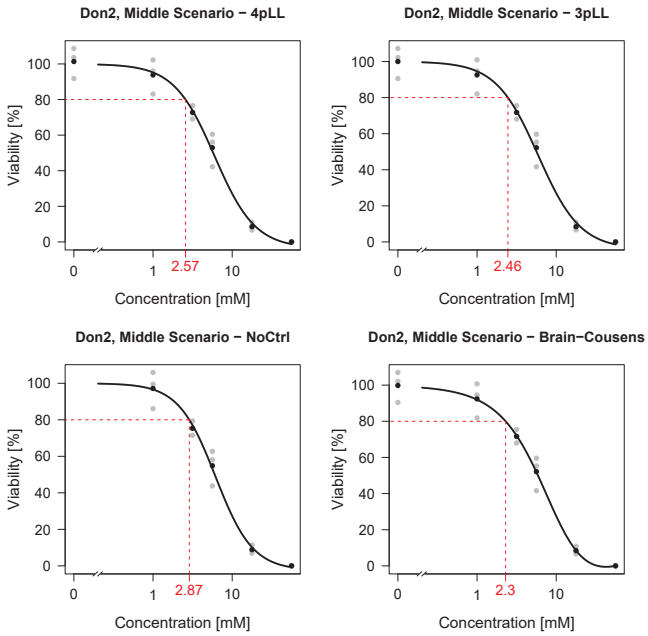

Figure B.34: Application of the four methods to the original dataset, Don2, resembling the 'medium' situation.

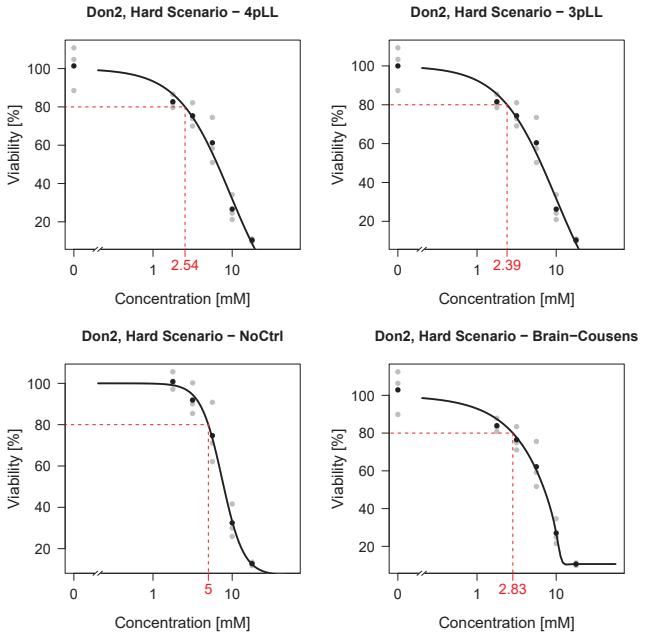

Figure B.35: Application of the four methods to the original dataset, Don2, resembling the 'difficult' situation.

B.2. Identification of alert concentrations

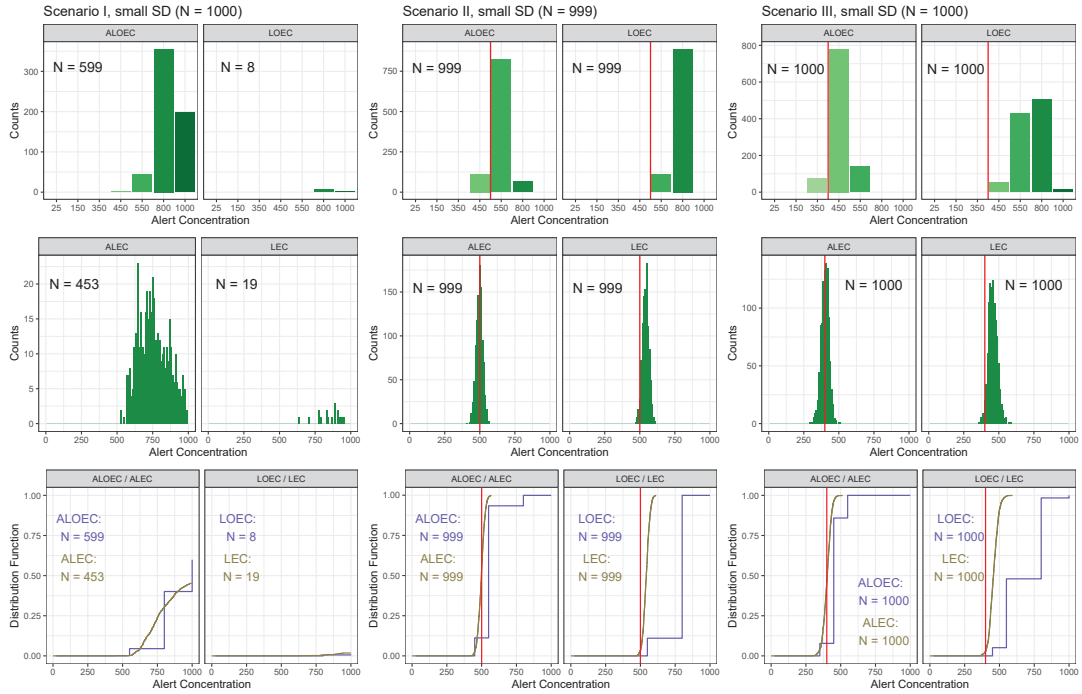

Figure B.36: Results of the simulation study for 'small' SD, with the same structure as Figure 6.3, when using the Dunnett procedure for the LOEC.

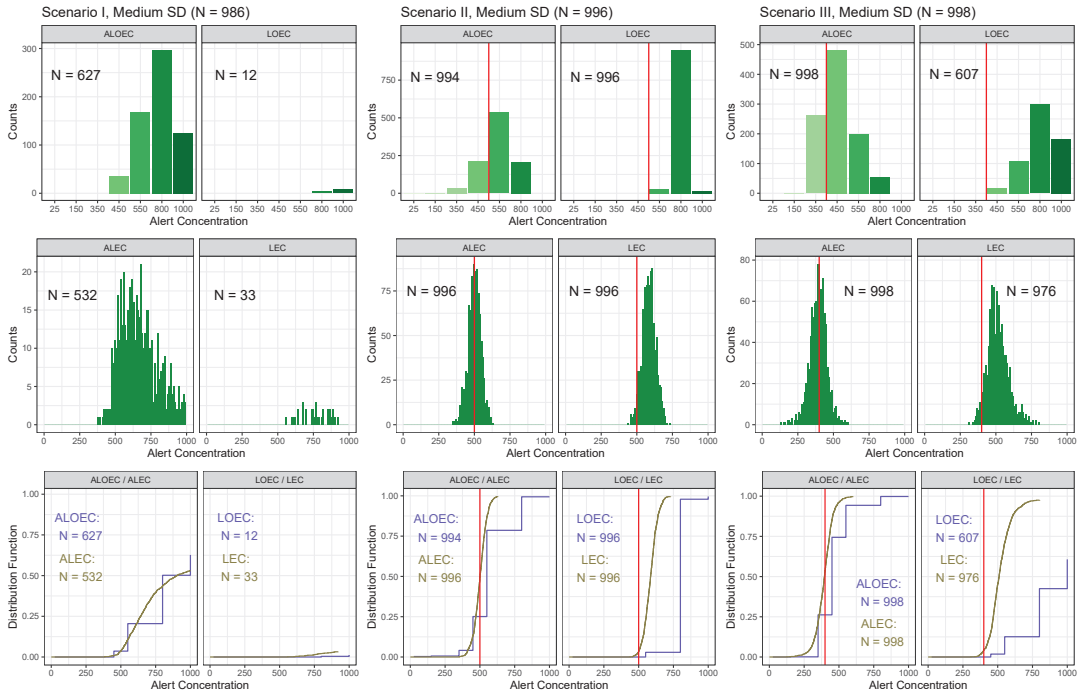

Figure B.37: Results of the simulation study for 'medium' SD, with the same structure as Figure 6.3, when using the Dunnett procedure for the LOEC.

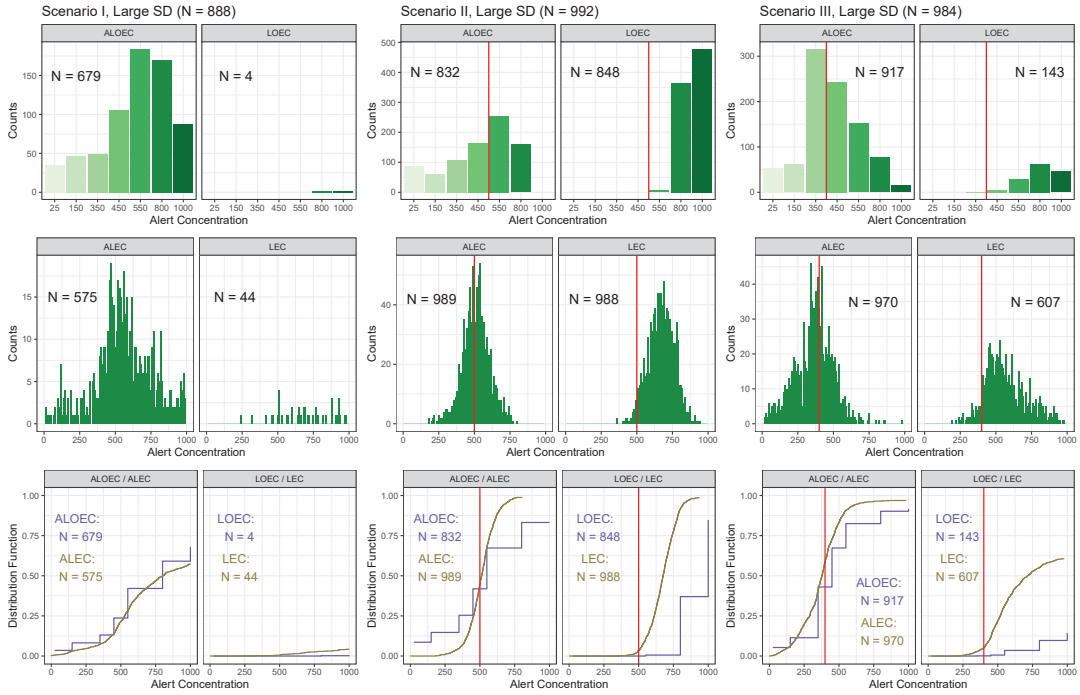

Figure B.38: Results of the simulation study for 'large' SD, with the same structure as Figure 6.3, when using the Dunnett procedure for the LOEC.

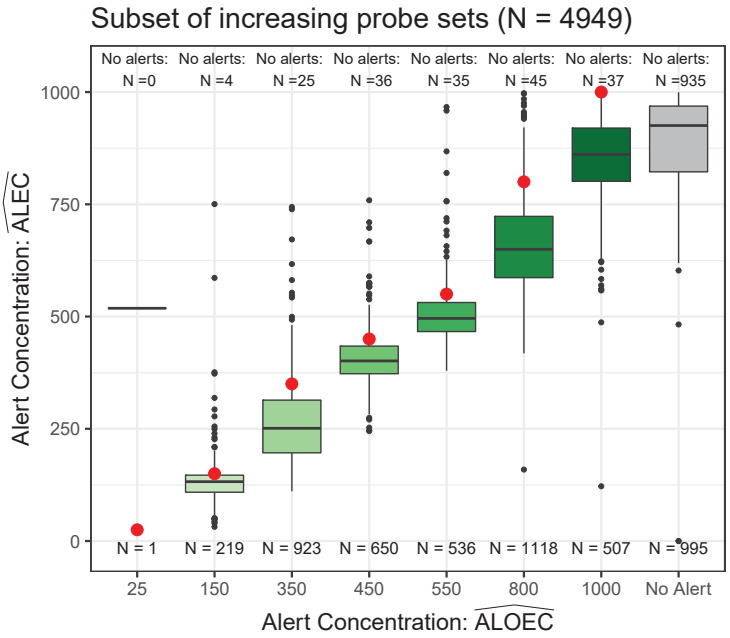

Figure B.39: Results of the analysis of the VPA dataset for methods considering absolute exceedance of the threshold, only for increasing probe sets. The structure of the plot is the same as Figure 6.7.

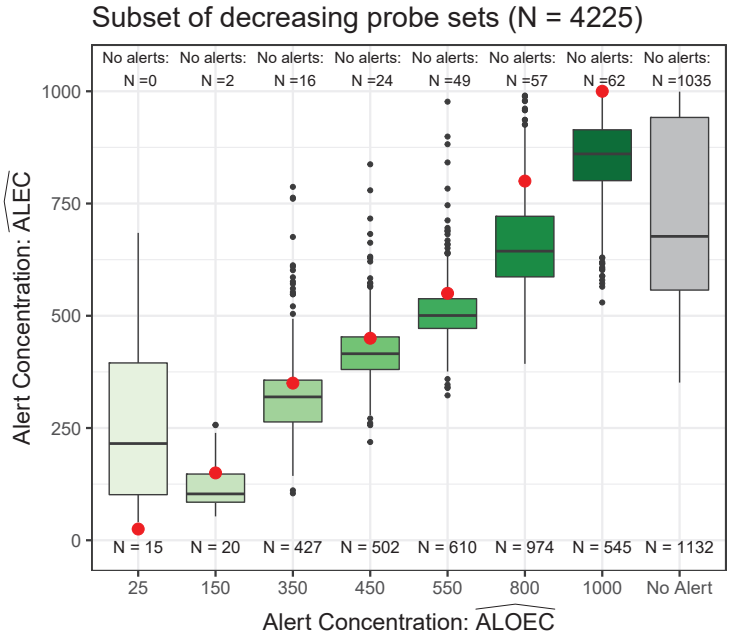

Figure B.40: Results of the analysis of the VPA dataset for methods considering absolute exceedance of the threshold, only for decreasing probe sets. The structure of the plot is the same as Figure 6.7.

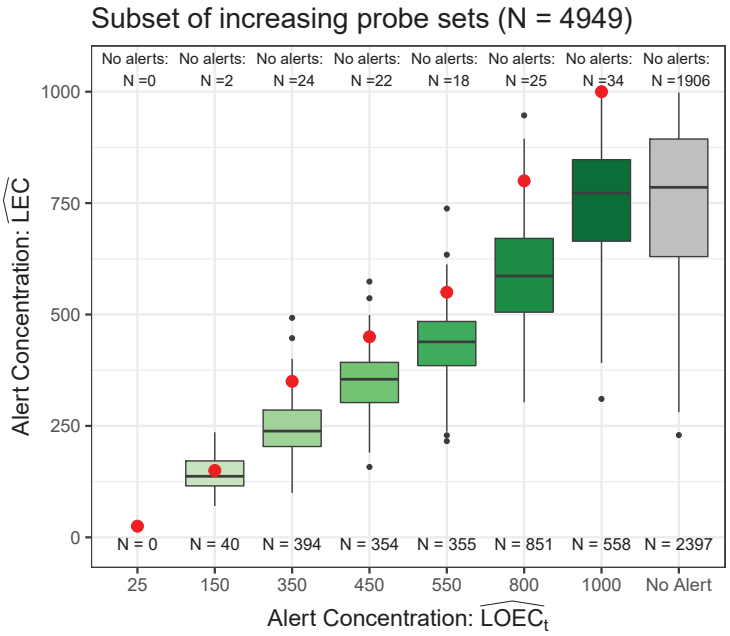

Figure B.41: Results of the analysis of the VPA dataset for $\widehat{\text{LOEC}}$ based on the t-test and $\widehat{\text{LEC}}$, only for increasing probe sets. The structure of the plot is the same as Figure 6.7.

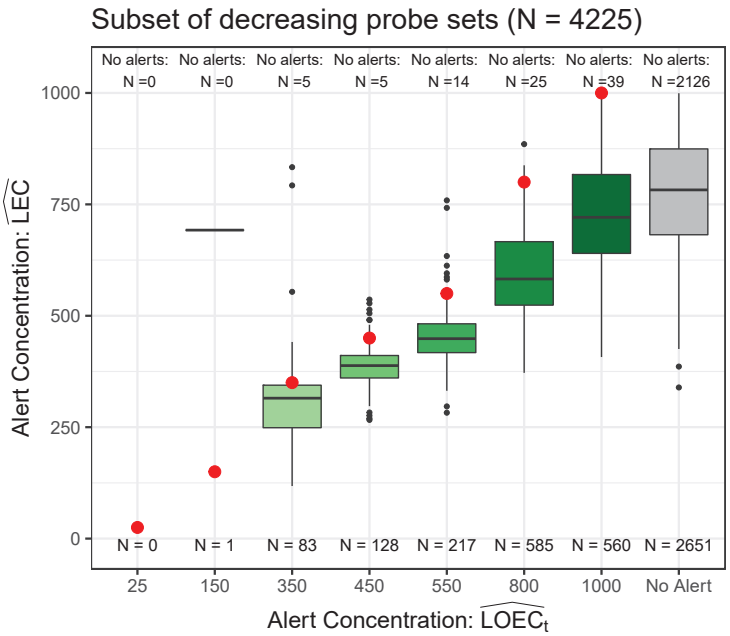

Figure B.42: Results of the analysis of the VPA dataset for $\widehat{\text{LOEC}}$ based on the t-test and $\widehat{\text{LEC}}$, only for decreasing probe sets. The structure of the plot is the same as Figure 6.7.

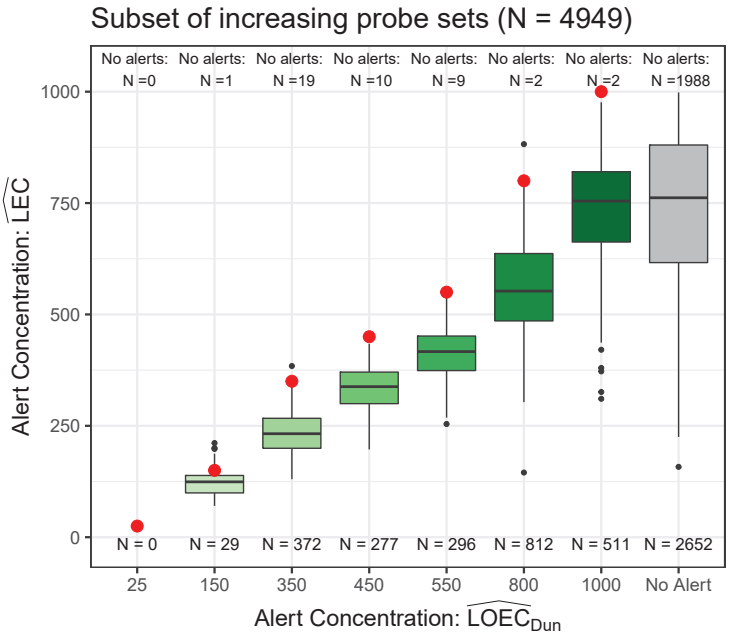

Figure B.43: Results of the analysis of the VPA dataset for $\widehat{\text{LOEC}}$ based on the Dunnett procedure and $\widehat{\text{LEC}}$, only for increasing probe sets. The structure of the plot is the same as Figure 6.7.

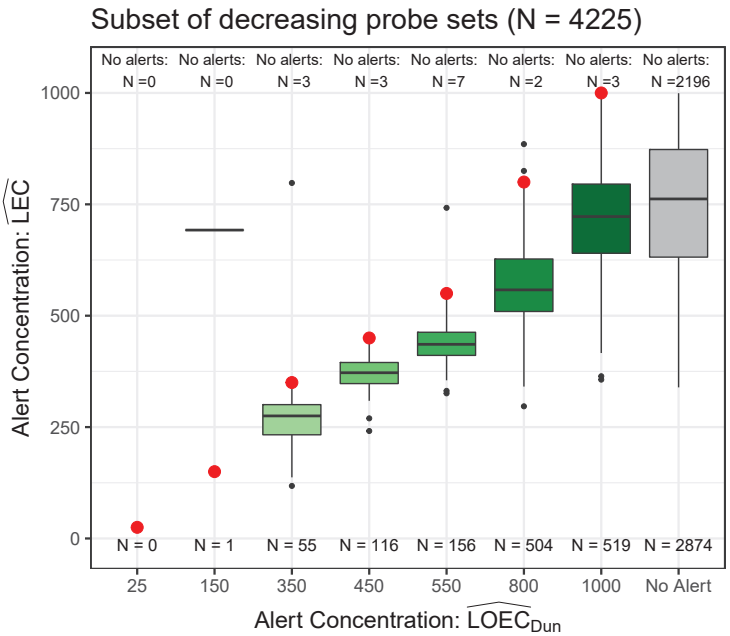

Figure B.44: Results of the analysis of the VPA dataset for $\widehat{\text{LOEC}}$ based on the Dunnett procedure and $\widehat{\text{LEC}}$, only for decreasing probe sets. The structure of the plot is the same as Figure 6.7.

B.3. Information sharing across genes

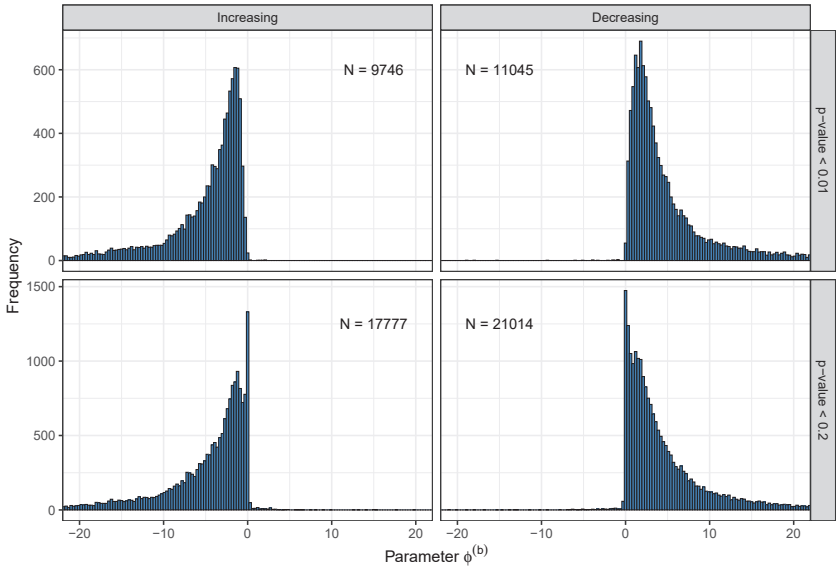

Figure B.45: Histograms of parameter $\phi^{(b)}$ for 2 different cutoffs of the MCP-Mod based p-values for increasing profiles (left) and decreasing profiles (right).

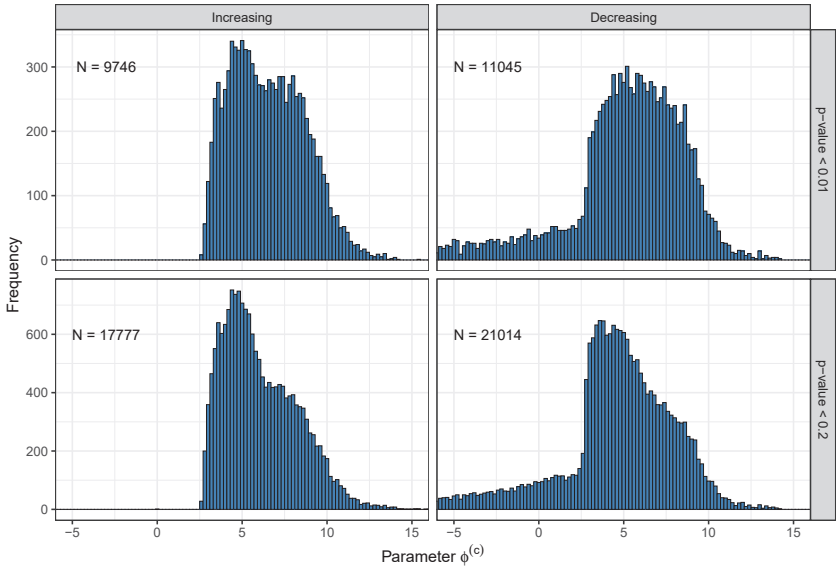

Figure B.46: Histograms of parameter $\phi^{(c)}$ for 2 different cutoffs of the MCP-Mod based p-values for increasing profiles (left) and decreasing profiles (right).

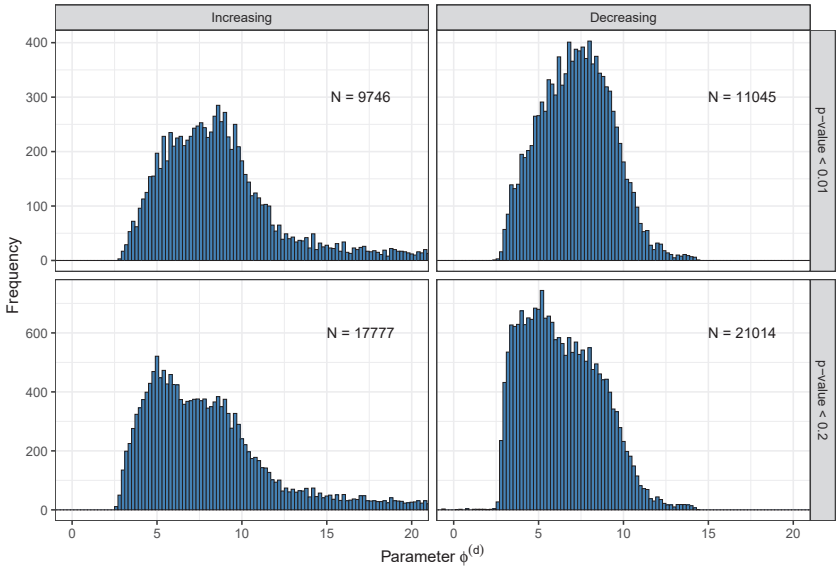

Figure B.47: Histograms of parameter $\phi^{(d)}$ for 2 different cutoffs of the MCP-Mod based p-values for increasing profiles (left) and decreasing profiles (right).

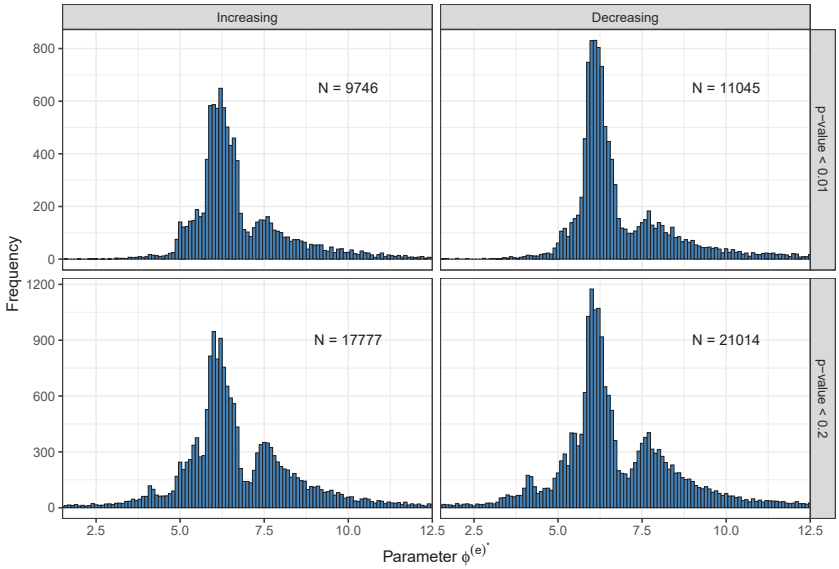

Figure B.48: Histograms of parameter $\phi^{(e)*}$ for 2 different cutoffs of the MCP-Mod based p-values for increasing profiles (left) and decreasing profiles (right).

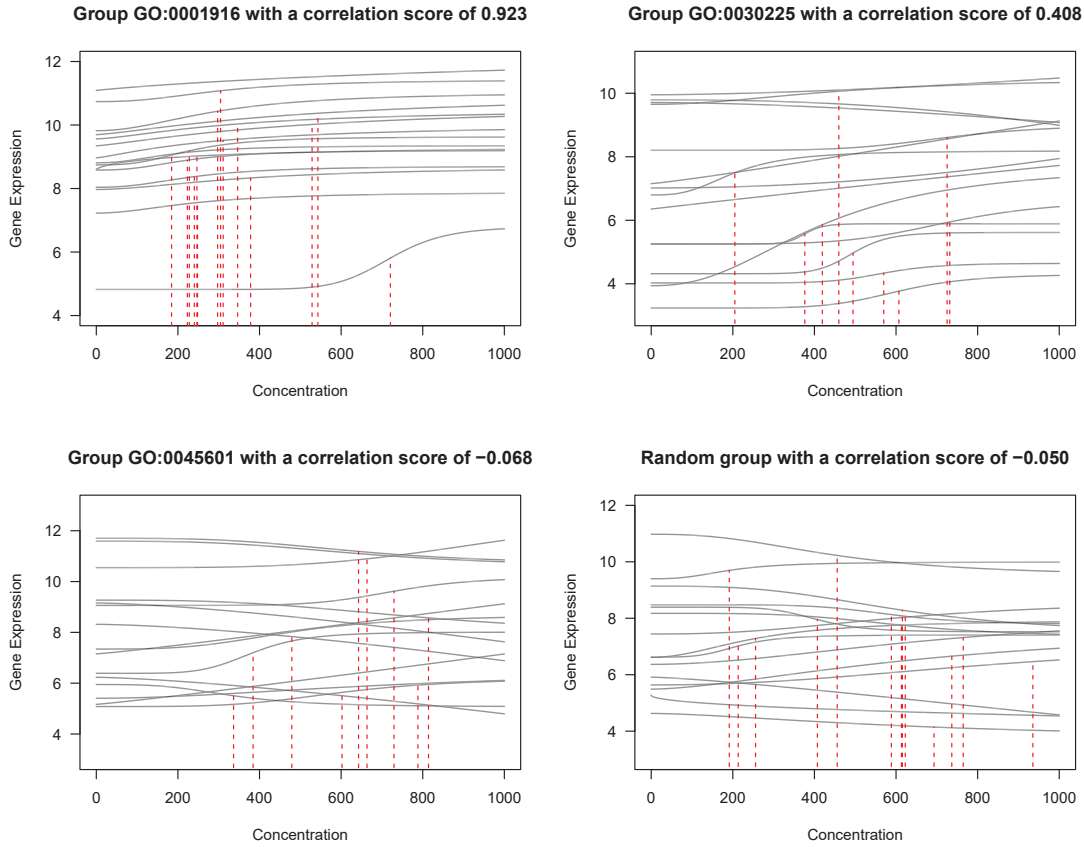

Figure B.49: Fitted curves for the 15 genes, respectively, for four GO-groups selected specifically. Red dotted lines indicates the concentration where the half-maximal effect is observed.

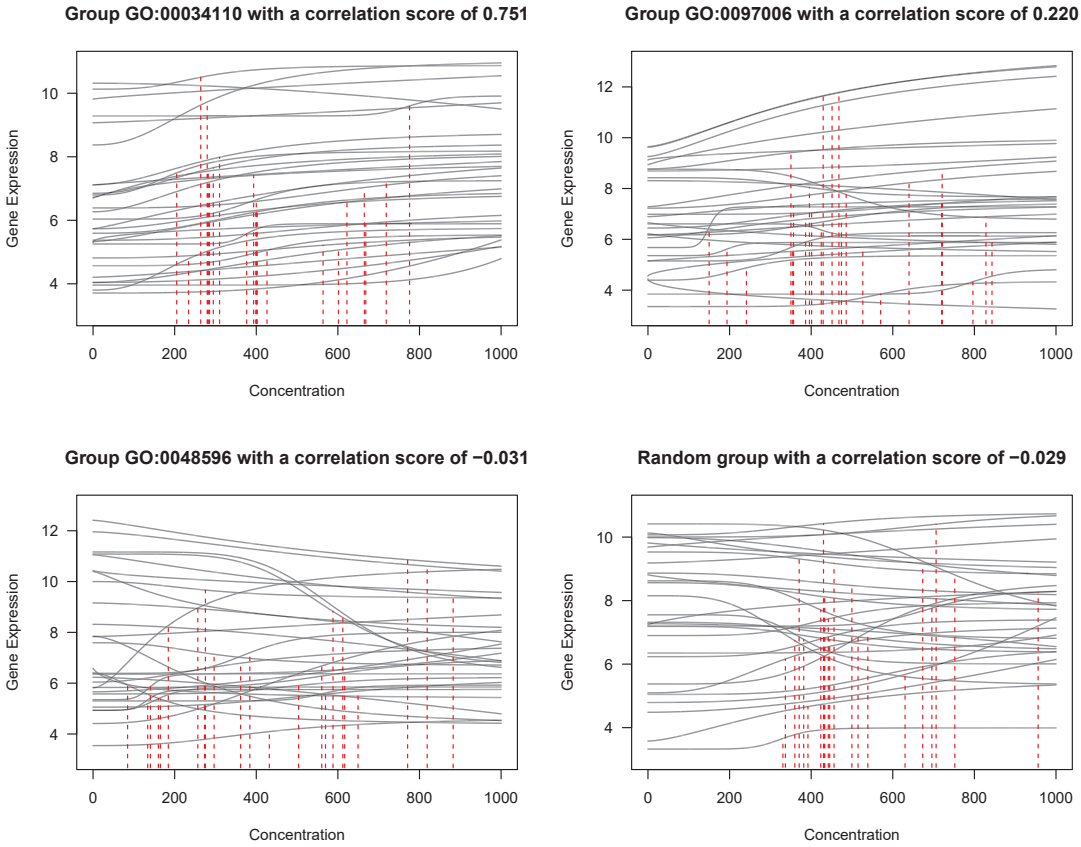

Figure B.50: Fitted curves for the 30 genes, respectively, for four GO-groups selected specifically. Red dotted lines indicates the concentration where the half-maximal effect is observed.

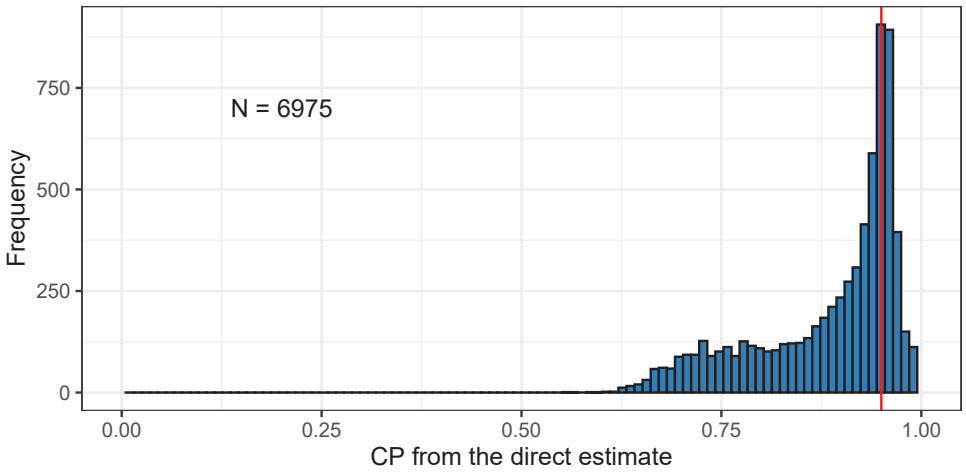

Figure B.51: Histogram of the coverage probability for the direct estimation of parameter $\phi^{(e)*}$ for the simulation study based on the large set of probe sets..

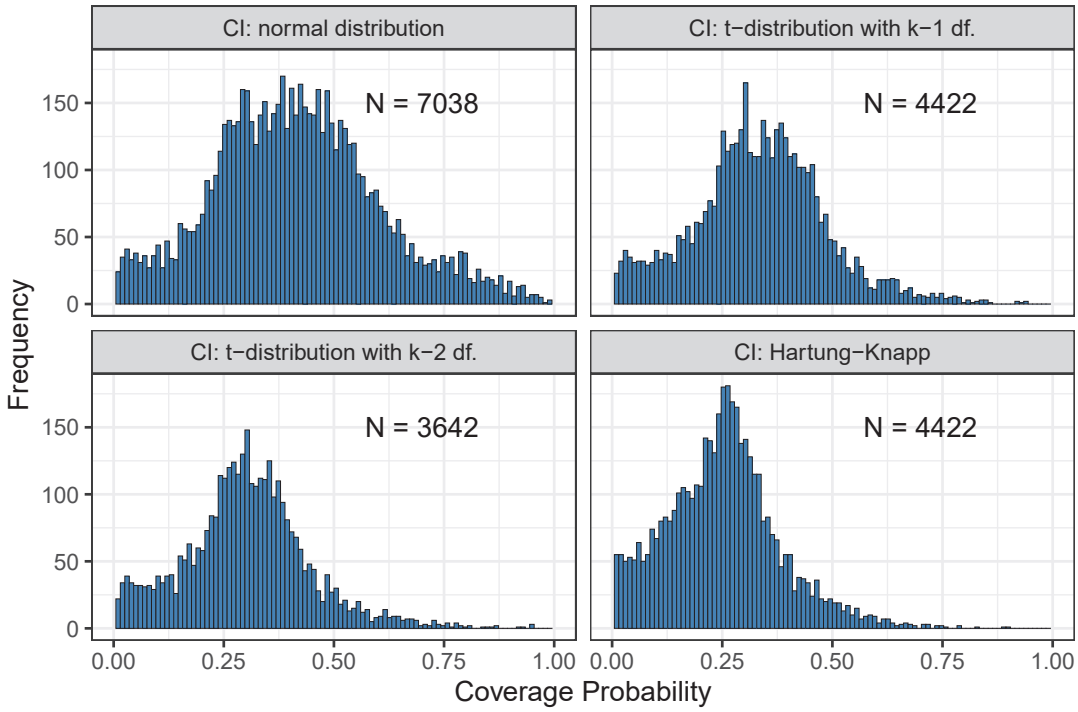

Figure B.52: Histogram of the coverage probability for the four variants of calculating confidence intervals from a meta-analysis.

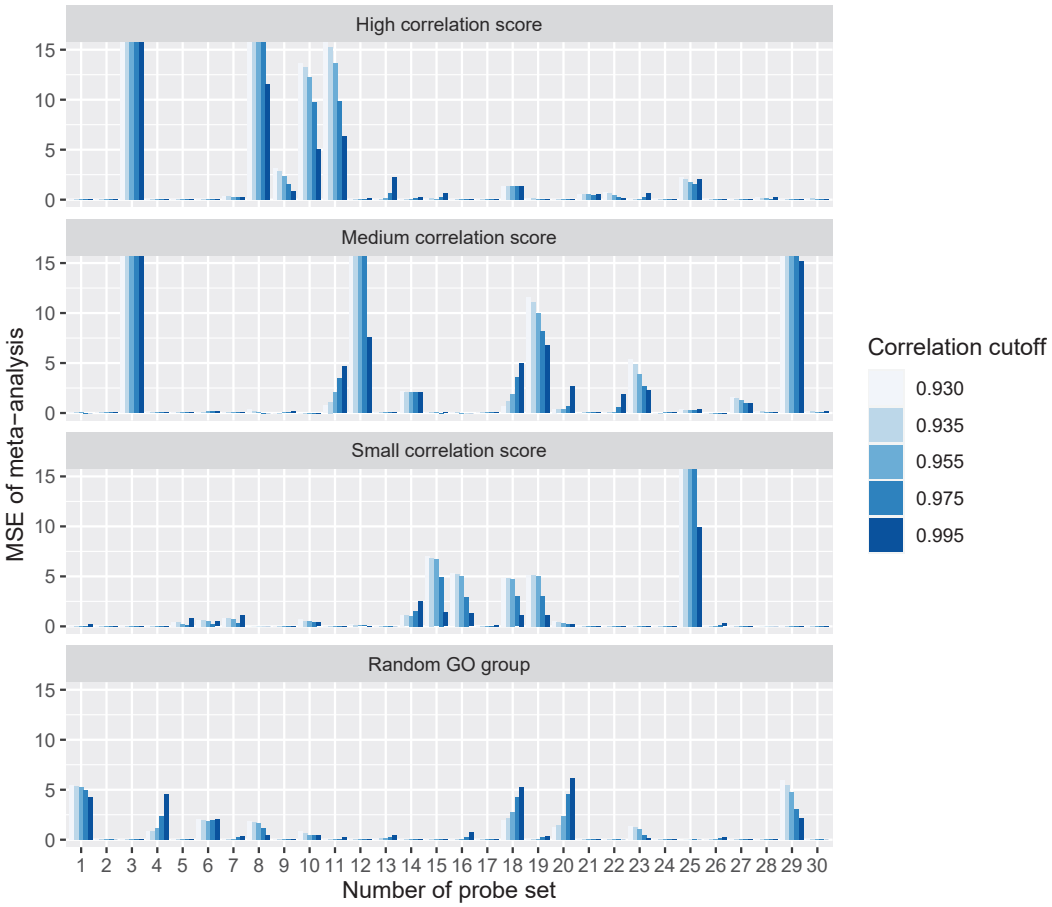

Figure B.53: Different MSEs of the meta-analysis method for the four GO-groups of size 30 when changing the correlation cutoff to be exceeded. The color of the respective bar indicates the correlation cutoff to be exceeded and the height of the bar the resulting MSE for the meta-analysis approach.

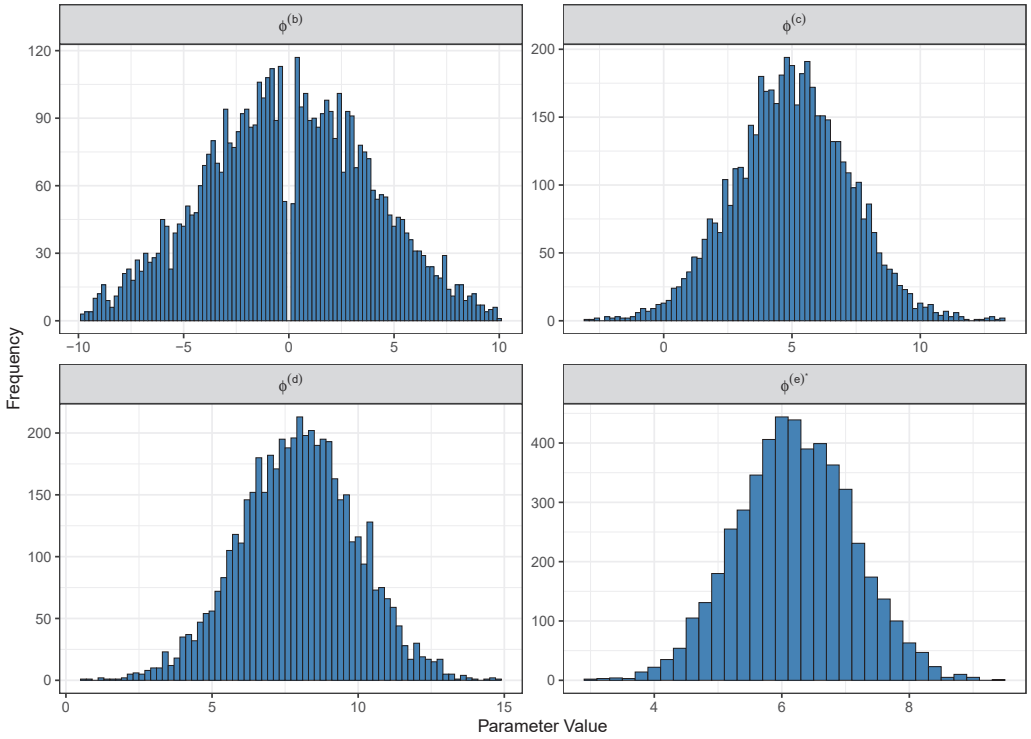

Figure B.54: Histograms of the four simulated parameters for the synthetic dataset.

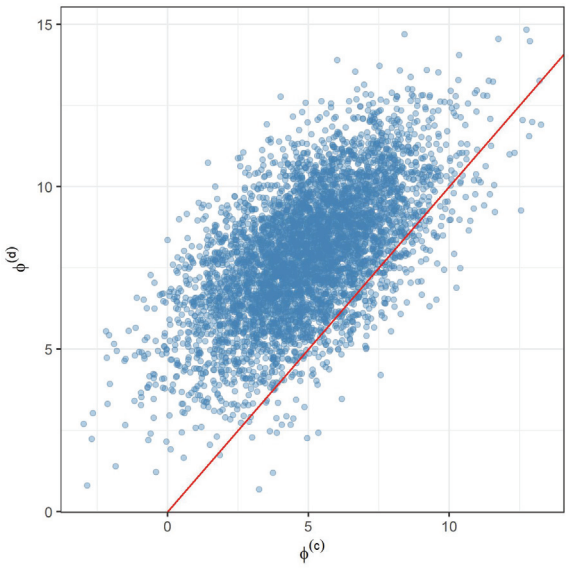

Figure B.55: Relationship between $\phi^{(c)}$ and $\phi^{(d)}$ in the synthetic dataset.

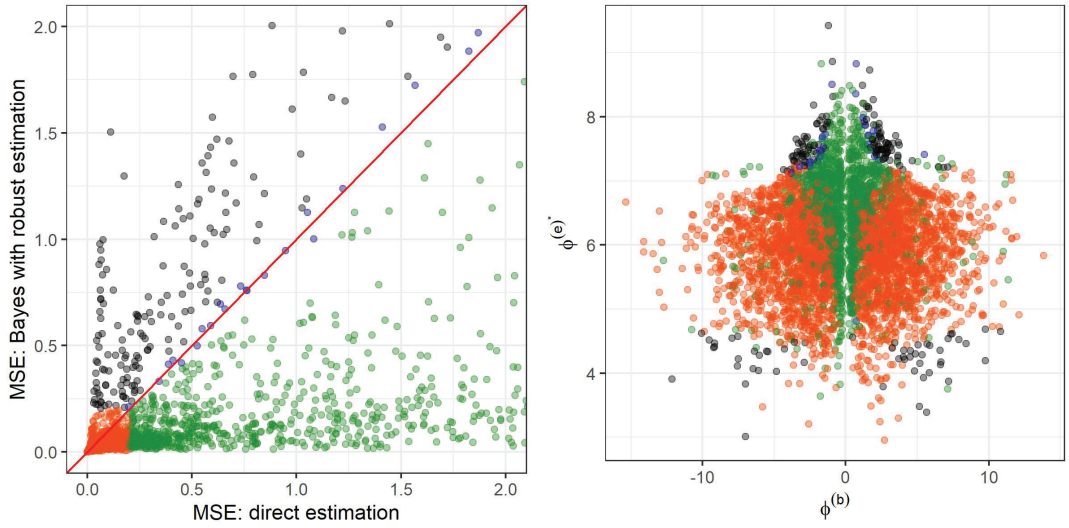

Figure B.56: Left: Comparison of MSE for the direct estimate and the Bayes estimate based on the robust estimation. Right: True underlying parameters $\phi^{(b)}$ and $\phi^{(e)^*}$ plotted against each other and colored according to the comparison of MSEs.

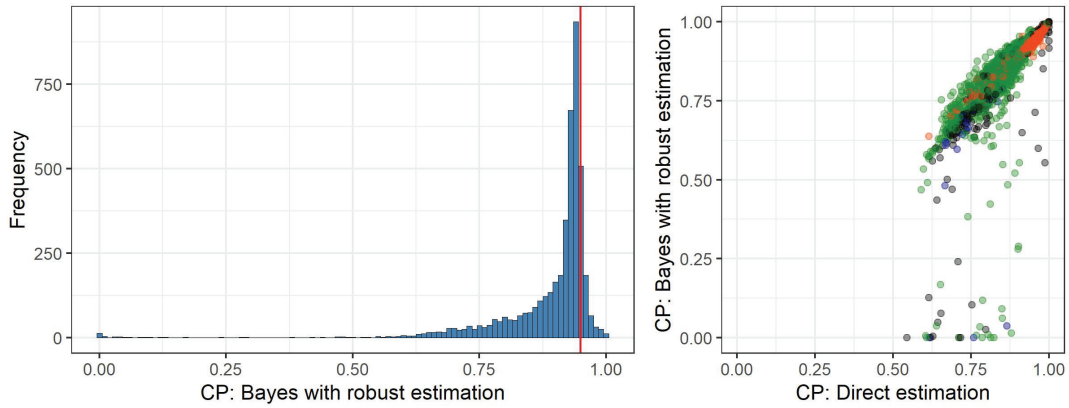

Figure B.57: Left: Histograms of the coverage probability for the credible intervals based on the robust estimation. Right: Comparison of the coverage probabilities from the direct estimate and the coverage probability shown in the left. Colors are chosen according to the comparison of MSE from Figure B.56.

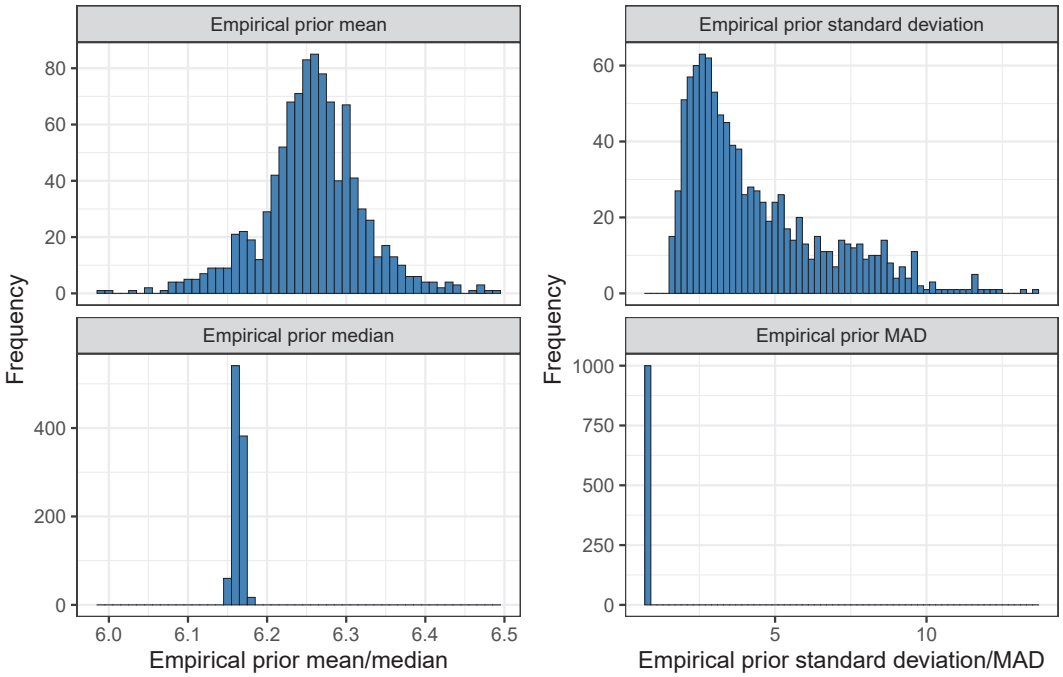

Figure B.58: Histogram of the prior estimates for the empirical Bayes method to estimate the parameter $\phi^{(e)*}$. Results for ML estimation are shown in the top, results for robust estimation in the bottom of the plot.

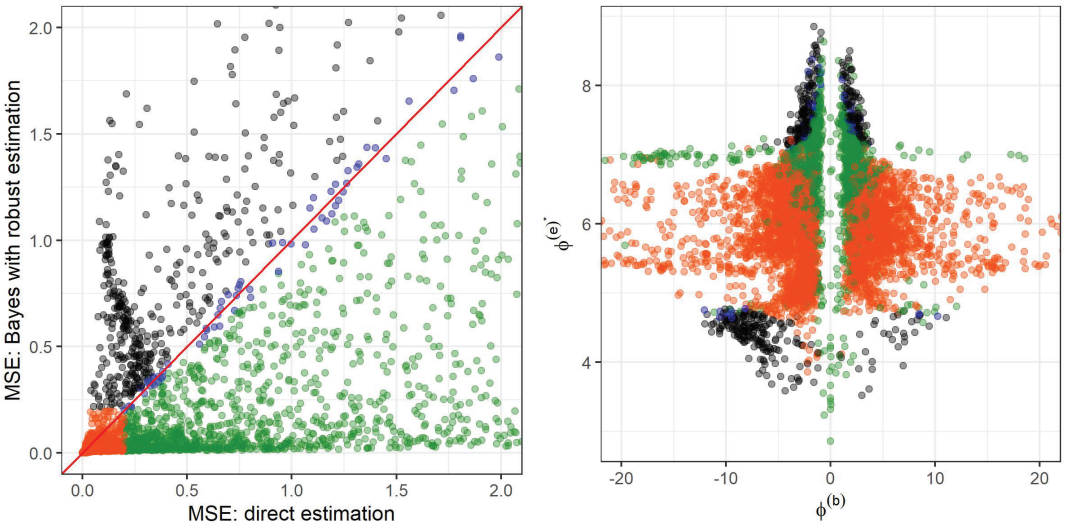

Figure B.59: Left: Comparison of MSE for the direct estimate and the Bayes estimate based on the robust estimation. Right: True underlying parameters $\phi^{(b)}$ and $\phi^{(e)*}$ plotted against each other and colored according to the comparison of MSEs.

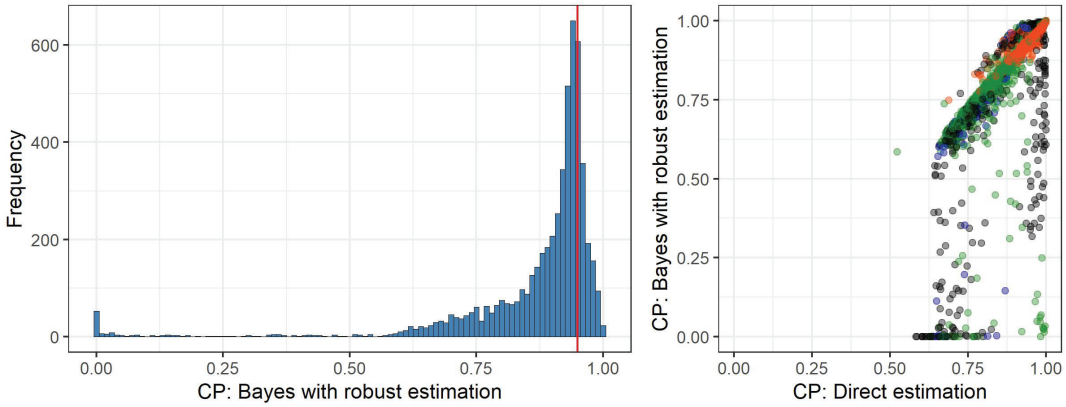

Figure B.60: Left: Histograms of the coverage probability for the credible intervals based on the robust estimation. Right: Comparison of the coverage probabilities from the direct estimate and the coverage probability shown in the left. Colors are chosen according to the comparison of MSE from Figure B.56.

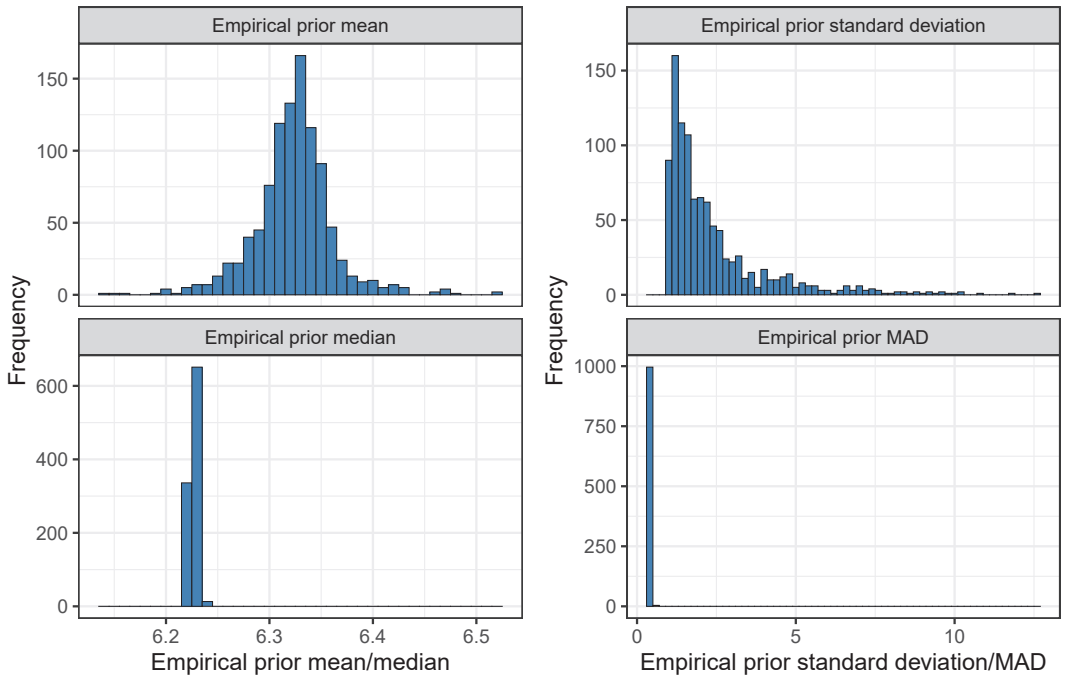

Figure B.61: Histogram of the prior estimates for the empirical Bayes method to estimate the parameter $\phi^{(e)*}$. Results for ML estimation are shown in the top, results for robust estimation in the bottom of the plot.

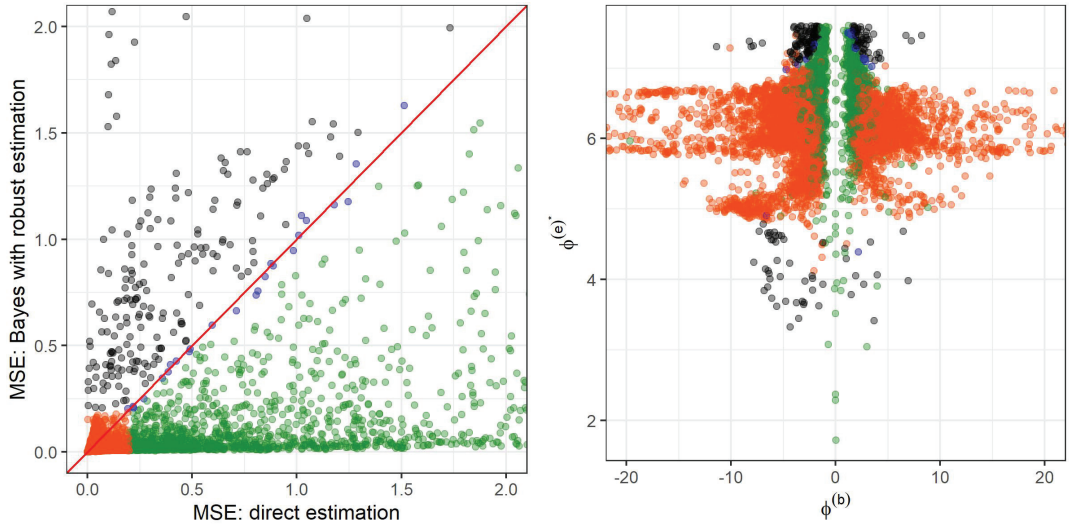

Figure B.62: Left: Comparison of MSE for the direct estimate and the Bayes estimate based on the robust estimation. Right: True underlying parameters $\phi^{(b)}$ and $\phi^{(e)^*}$ plotted against each other and colored according to the comparison of MSEs.

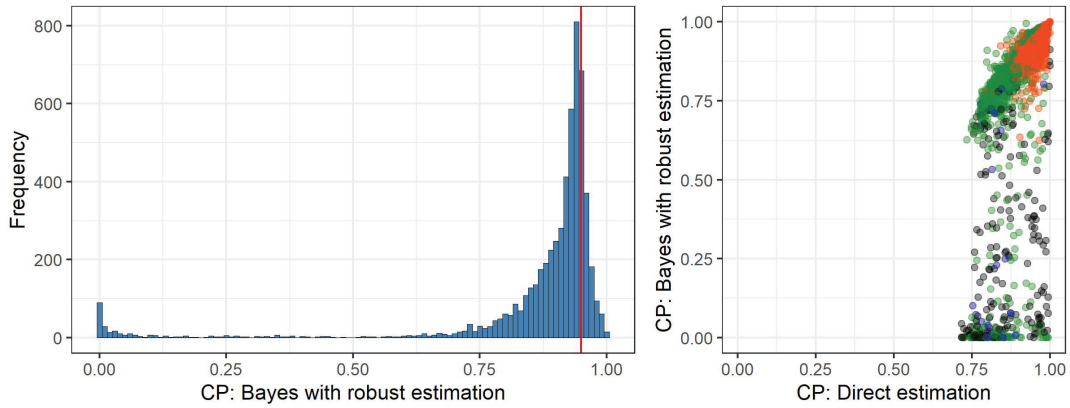

Figure B.63: Left: Histograms of the coverage probability for the credible intervals based on the robust estimation. Right: Comparison of the coverage probabilities from the direct estimate and the coverage probability shown in the left. Colors are chosen according to the comparison of MSE from Figure B.56.

C. Tables: Handling deviating control values

Table C.1: EC_{10}, EC_{20} and EC_{50} values together with corresponding limits of 95% confidence intervals for the four methods in a real data study resembling the 'medium' scenario for Don1.

	EC_{10}			**EC_{20}**			**EC_{50}**		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
4pLL	1.43	0.58	3.54	2.45	1.37	4.37	6.02	4.42	8.21
3pLL	0.89	0.48	1.65	1.77	1.20	2.59	5.45	3.78	7.85
No Ctrl	2.90	1.92	4.39	3.99	3.08	5.16	6.90	5.24	9.08
BC	1.00			1.96			6.13		

Table C.2: EC_{10}, EC_{20} and EC_{50} values together with corresponding limits of 95% confidence intervals for the four methods in a real data study resembling the 'difficult' scenario for Don1.

	EC_{10}			**EC_{20}**			**EC_{50}**		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
4pLL	0.98	0.60	1.59	1.87	1.29	2.73	5.34	2.87	9.93
3pLL	0.96	0.69	1.34	1.85	1.37	2.50	5.31	2.90	9.72
No Ctrl	2.12	0.65	6.93	3.23	1.50	6.95	6.57	4.53	9.53
BC	1.00			1.91			5.50		

Table C.3: EC_{10}, EC_{20} and EC_{50} values together with corresponding limits of 95% confidence intervals for the four methods in a real data study resembling the 'medium' scenario for Don2.

	EC_{10}			**EC_{20}**			**EC_{50}**		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
4pLL	1.60	1.00	2.55	2.57	1.87	3.53	5.76	4.63	7.16
3pLL	1.50	1.06	2.11	2.46	1.96	3.07	5.65	4.62	6.91
No Ctrl	1.86	1.07	3.24	2.87	1.95	4.24	6.02	4.79	7.57
BC	1.22			2.30			5.88		

Table C.4: EC_{10}, EC_{20} and EC_{50} values together with corresponding limits of 95% confidence intervals for the four methods in a real data study resembling the 'difficult' scenario for Don2.

	$\mathbf{EC_{10}}$			$\mathbf{EC_{20}}$			$\mathbf{EC_{50}}$		
	Est.	Lower	Upper	Est.	Lower	Upper	Est.	Lower	Upper
4pLL	1.40	0.74	2.63	2.54	1.52	4.25	6.51	2.67	15.91
3pLL	1.28	0.84	1.94	2.39	1.45	3.95	6.40	2.51	16.30
No Ctrl	3.87	2.54	5.91	5.00	3.76	6.64	7.83	6.27	9.79
BC	1.41			2.83			7.09		

Eidesstattliche Erklärung

Hiermit erkläre ich, Franziska Kappenberg, dass ich die vorliegende Arbeit mit dem Titel ‘Statistical approaches for calculating alert concentrations from cytotoxicity and gene expression data’ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Dissertation ist bisher keiner anderen Fakultät vorgelegt worden. Ich erkläre, dass ich bisher kein Promotionsverfahren erfolglos beendet habe und dass keine Aberkennung eines bereits erworbenen Doktorgrads vorliegt.

Ort, Datum	Franziska Kappenberg