CURVE FITTING WITH CONFIDENCE FOR PRECLINICAL DOSE–RESPONSE DATA

MATHIAS CARDNER

Abstract. In the preclinical stage of pharmaceutical drug development, when investigating the medicinal properties of a new compound, there are two important questions to address. The first question is simply whether the compound has a significant beneficial effect compared to vehicle (placebo) or reference treatments. The second question concerns the more nuanced dose–response relationship of the compound of interest. One of the aims of this thesis is to design an experiment appropriate for addressing both of these questions simultaneously. Another goal is to make this design optimal, meaning that dose-levels and sample sizes are arranged in a manner which maximises the amount of information gained from the experiment. We implement a method for assessing efficacy (the first question) in a modelling environment by basing inference on the confidence band of a regression curve. The verdicts of this method are compared to those of one-way ANOVA coupled with the multiple comparison procedure Dunnett's test. When applied to our empirical data sets, the two methods are in perfect agreement regarding which dose-levels have an effect at the 5% significance level. Through simulation, we find that our modelling approach tends to be more conservative than Dunnett's test in trials with few dose-levels, and vice versa in trials with many dose-levels. Furthermore, we investigate the effect of optimally designing the simulated trials, and also the consequences of misspecifying the underlying dose–response model during regression, in order to assess the robustness of the implemented method.

Key words: pharmacodynamic modelling; efficacy study; dose–response; optimal design; model-robust design; simultaneous confidence bands; bootstrap; multiple comparisons

UNIVERSITY OF GOTHENBURG

Date: January 21, 2015.

Thesis for the degree of Master of Science, written in the field of mathematical statistics as part of the two-year *Master's Programme in Mathematical Sciences* at the University of Gothenburg in Sweden. Please direct correspondence to `mathias@cardner.se`

Supervisor: Sofia Tapani, PhD, Senior Statistician, Discovery Statistics, AstraZeneca R&D in Mölndal.

Co-advisor: Rasmus Jansson Löfmark, PhD, CVMD iMed DMPK, AstraZeneca R&D in Mölndal.

Examiner: Professor Aila Särkkä, Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg.

Acknowledgements

I want to express my deep gratitude to my supervisor Sofia Tapani for her skilful guidance and insightful advice. Likewise, I am very thankful to my co-advisor Rasmus Jansson Löfmark for many stimulating discussions, and for giving me a better understanding of pharmacological concepts. I would also like to thank Marita and Janeli for fantastic company, interesting conversations, and resources both computational and literary.

Many important ideas in this thesis were inspired by a presentation which Marianne Månsson, Karin Nelander and Marita Olsson gave at Statistikerträffen in September of 2012. Furthermore, I have received highly valuable comments and suggestions from Pete Ceuppens, Brian Middleton and Marie South, which greatly enhanced the results of this thesis.

Finally, I would like to express my appreciation for the library at Chalmers University of Technology and its staff. The library has been a tremendous asset, without which I would likely not have gained access to the splendid sources of information upon which this thesis rests.

Contents

1. Introduction

1.1. **Background.** During the early development of a medicinal drug — when investigating whether a certain compound has some beneficial effect — there are traditionally two different strategies behind the statistical analyses performed [1]. The first strategy — which we shall refer to as an efficacy study — is designed to determine, with statistical significance, whether the compound actually does have a beneficial effect compared to control or reference treatments. This design typically has relatively few dose-levels, with many observations made at each level in order to yield statistical confidence in the results. The other strategy, which is model-based in nature, is geared towards gaining more nuanced information about the dose–response profile of the compound of interest [2]. This design typically has a larger number of dose-levels, spread out in such a way that we gain insight into the mechanistic (functional) form of the underlying phenomena, whilst sacrificing confidence by having fewer observations at each dose-level. It is important to keep in mind that the total number of observations is limited by available resources and, in the case of *in vivo* studies, by ethical considerations since animals are involved. The major benefit of this model-based approach is that it can be used to estimate appropriate dose-levels for future studies. If the target dose is incorrectly estimated, the recommended dose-level for subsequent trials may be set too high, potentially causing concerns of toxicity and safety. Alternatively, if the recommended dose-level is set too low, its beneficial effect may be too small to detect in a confirmatory phase [3].

The overarching goal of this project is to devise a strategy for a statistical analysis which addresses both of the above questions simultaneously. In other words, we aim to design an experiment which lends insight both into the efficacy and the dose–response relationship of a given compound. The most obvious reward is that of parsimony, since a successful synthesis of the traditionally disparate experiments would save resources — both financially and ethically. Moreover, it would be of value to experimenters who wish to gauge the mechanistic dose–response relationship whilst retaining statistical confidence in the results. Say for instance that a researcher is primarily interested in estimating the minimum effective dose (MED), defined as the smallest dose which yields a clinically relevant and statistically significant effect [4]. Then an experimental design could be optimised with respect to its ability of estimating the MED, in the sense of minimising the variance of this estimate [5]. The methodology derived from our investigations will be summarised in Section 6.2 as a guide for designing and analysing dose–response experiments.

Further nuance. It is worthwhile to make a few observations about the fundamental differences between the two types of strategies mentioned above. In the efficacy study, the dose-levels are considered to be qualitative, and the statistical analysis (e.g. ANOVA) requires comparatively few assumptions [1]. Conversely, in the model-based approach, assumptions must be made concerning the underlying model of dose–response, which then describes a functional relationship between the dose — now viewed as inherently quantitative — and the response. If the postulated model is representative of the true dose–response relationship, then this more nuanced knowledge can answer important questions. For instance, it can be used to estimate the dose which attains half the maximal effect (ED_{50}) or the minimum effective dose (MED) [3]. Thus, whereas inference from the efficacy study is limited to the dose-levels at which observations were made, the model-based approach can be used to interpolate the response within the range between the smallest and largest dose administered (including the vehicle treatment). This increased flexibility comes at the cost of more assumptions, which, if unrepresentative of the true phenomena, may yield misleading results.

Granted, when an experiment has been performed and the measurements have been gathered, there is nothing preventing us from using both of the above methodologies to analyse the data. Notice that it is the experimental *design* (i.e. the arrangement of dose-levels and their sample

sizes) which may make one of the methods more appropriate than the other. In light of this, an important objective of this thesis is to investigate how the allocation of experimental resources affects the inferences drawn from the different methodologies.

1.2. **Pharmacological background.** This thesis focuses on dose–response studies, in which a drug is administered at various dose-levels and the corresponding response is measured. The medicinally active substance is delivered in an inert substance (excipient) referred to as *vehicle.* In order to measure the null effect — i.e. to establish a baseline response — it is appropriate to include a treatment containing no active substance, but only the vehicle itself. This vehicle treatment is a control expected to have no effect, and as such, it can be thought of as a type of placebo. Since the vehicle treatment contains no trace of the active substance, it is appropriate to associate it with a dose-level of zero. Furthermore, the dose-levels are generally spread out more or less evenly on a logarithmic scale, in order to gauge the response across orders of magnitude. Hence it is often appropriate to use a logarithmic transformation on the dose scale when plotting the data. This however causes trouble with the vehicle, since the logarithm of zero is not (finitely) defined, meaning that it is not obvious how to handle the vehicle treatment. One possible solution is to adjust the response so that it becomes relative to the estimated baseline — for instance by uniformly subtracting the mean vehicle response. We have not done this in the upcoming figures, where the plots with a logarithmic dose scale simply omit the vehicle treatment. It will, however, play an important role in accounting for the variability in the mean vehicle response when we implement our method in Section 5.

1.3. **Reading guide.** As this thesis is written in the field of mathematical statistics applied to pharmacology, it is intended for two audiences with different backgrounds. Since the author's knowledge of pharmacology was virtually nonexistent at the beginning of this project, the thesis should hopefully be self-contained with respect to the pharmacological concepts involved. The same can however not be said of its mathematical and statistical contents, which will primarily appeal to those already acquainted with those fields. Therefore, we shall here offer a reading guide intended to outline the topics most relevant to the reader without a statistical background.

In Section 2 we describe empirical dose–response data sets which set the stage for the subsequent analysis. This leads to Section 2.2 in which we introduce common dose–response models. The literature review in Section 3 is the result of an extensive search for previous work addressing our problem, albeit rather tangentially. Nevertheless, the fruits of this search include articles and software which are highly useful for model-robust optimal design of dose–response trials.

Section 4 contains the mathematical and statistical methods we use throughout this thesis. It begins with the theory of optimal design of experiments, which, given some postulated dose–response model, determines the optimal allocation of experimental resources. In short, this theory tells us how many dose-levels to include, where to place them, and how many subjects to assign to each level. Granted, the resulting design will only be optimal from a mathematical point of view, and may suggest dose-levels which are inconceivable in practice. This is later addressed by software which rather lets the experimenter specify feasible candidate dose-levels beforehand, from which the most efficient allocation is chosen.

In Sections 4.2 and 4.3 we discuss regression and the construction of confidence bands around regression curves. This leads up to the key result in Section 5 in which we present our method for assessing efficacy under a modelling approach. The judgements made by this method are compared to those of one-way ANOVA coupled with the multiple comparison procedure Dunnett's test, which controls the family-wise error rate when comparing each dose-level against the vehicle treatment. (See Bretz et al. (2010) [6] for a detailed account of this test.) This is done through comprehensive simulations of dose–response trials, the findings of which are presented in Section 6 containing our results and discussion. Section 6.2 outlines a method for designing and analysing

dose–response studies using the `R` [7] package `DoseFinding` [8]. The conclusions in Section 7 give a summary of our findings, as well as elaborations on avenues of future work.

2. Data exploration and modelling

2.1. **Data.** Our investigations are supported by data sets containing empirical *in vivo* dose–response data for three different compounds, referred to as compounds A, B and C. For each compound, two response variables have been measured (from the same physical sample). The measured response quantities shall for each compound be referred to as response 1 and 2. Note, however, that these responses do not necessarily measure the same quantity for different compounds. For instance, response 1 of compound A may measure something different than does response 1 of compound B.

Figures 1–6 display boxplots and scatterplots of all six data sets. Notice that the dose-levels are logarithmically scaled in the scatterplots, meaning that the zero-dose vehicle treatment is not included. The data are analysed using one-way ANOVA coupled with Dunnett's test [6], which adjusts for multiplicity (see Section 4.3) when comparing each dose-level against vehicle. Throughout this thesis we perform Dunnett's test using the `R` [7] package `multcomp` [9]. We say that a dose-level is *active* if it shows a statistically significant effect compared to the vehicle treatment (at the given significance level, which in this thesis is set to 5%). The verdicts of Dunnett's test are recorded in the figure captions.

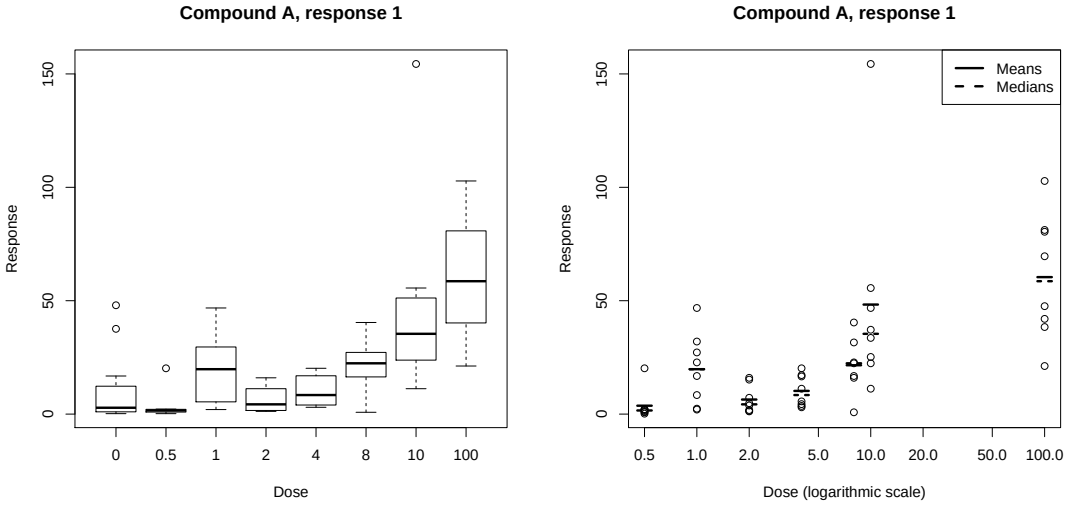

FIGURE 1. Boxplot (left) and scatterplot (right) of the observed dose–response data of compound A with respect to response 1. Dunnett's test indicates that the two highest dose-levels, 10 and 100, are active at the 5% significance level (both with p-values smaller than 0.03%). The data clearly suggest an increasing trend, with the possible exception of dose-level 1. The spread of the data at each dose-level is symmetric about its mean, which is generally quite close to its median. These statistics deviate most markedly at dose-level 10, due to a blatant outlier. Furthermore, there is evidence of heteroscedasticity, with larger variability at higher dose-levels — the exception being dose-level 1 whose data is quite variable.

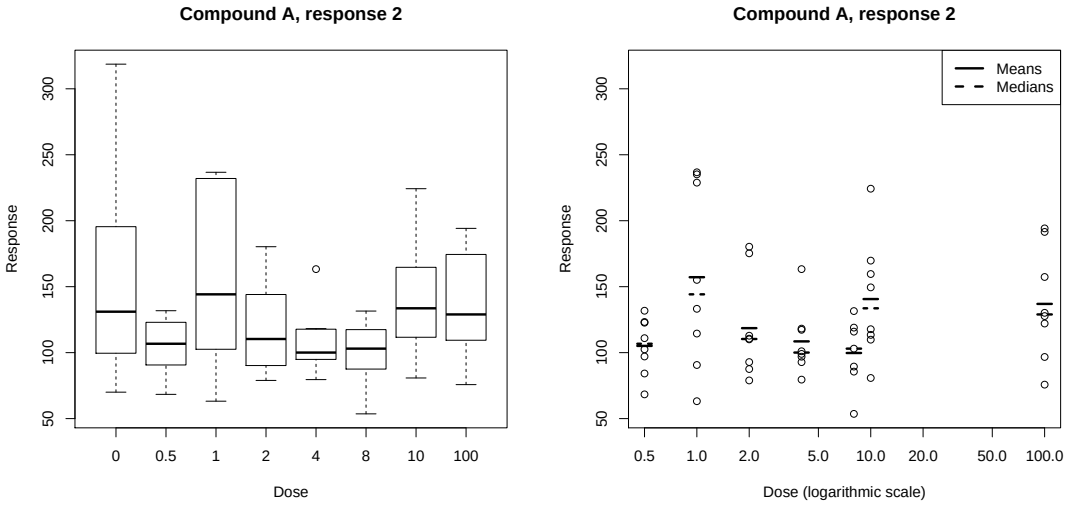

FIGURE 2. Boxplot (left) and scatterplot (right) of the observed dose–response data of compound A with respect to response 2. Dunnett's test declares no activity at significance level 5% (nor at 10%). The data indicate no apparent trend. At each dose-level, the observations are spread symmetrically about the mean, which tends to be slightly higher than the median. The heteroscedasticity follows no simple pattern, though again the variability at dose-level 1 is conspicuously large.

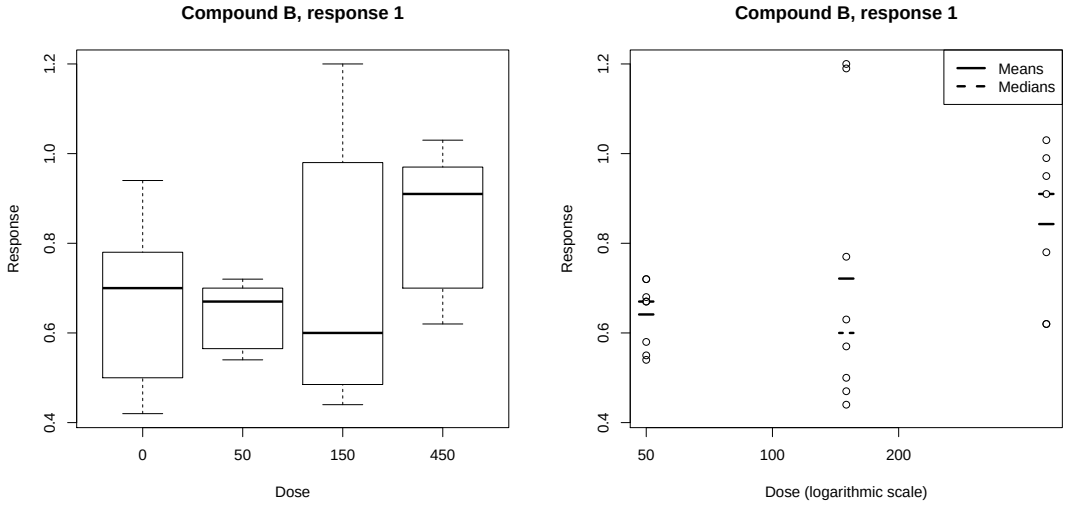

FIGURE 3. Boxplot (left) and scatterplot (right) of the observed dose–response data of compound B with respect to response 1. Dunnett's test declares no activity at significance level 5% (nor at 10%). The data indicate no monotonic trend. The observations are quite skewed, though not uniformly in the same direction, which disqualifies a monotone transformation. Heteroscedasticity appears erratic, with high variability in the vehicle treatment and two extreme observations at dose-level 150.

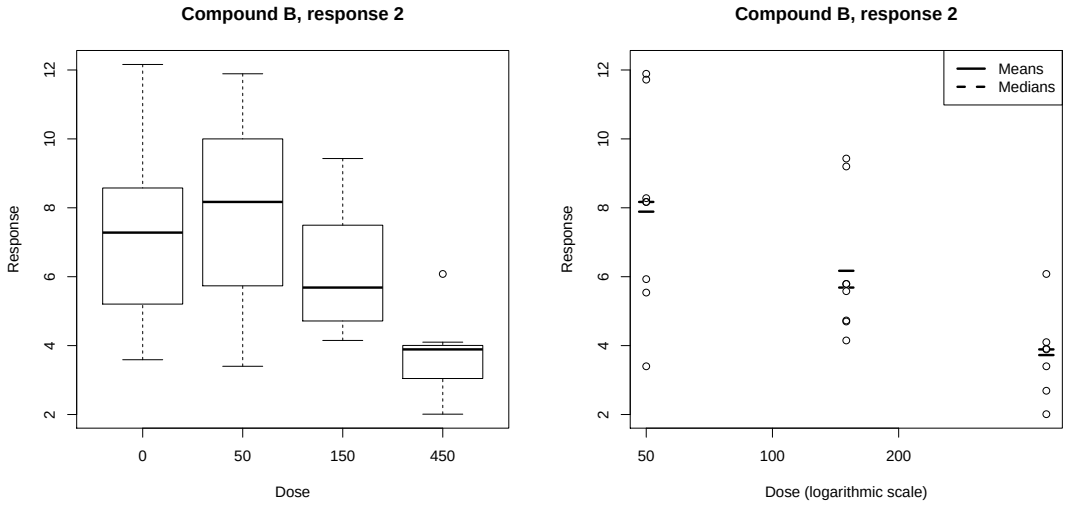

FIGURE 4. Boxplot (left) and scatterplot (right) of the observed dose–response data of compound B with respect to response 2. Dunnett's test indicates that the highest dose-level, 450, is active at the 5% significance level (with a p-value of 3%). Aside from the vehicle treatment, there is evidence of a decreasing trend. The observations are fairly symmetrical around the means. Heteroscedasticity is present in that variability appears to decrease with higher dose-levels.

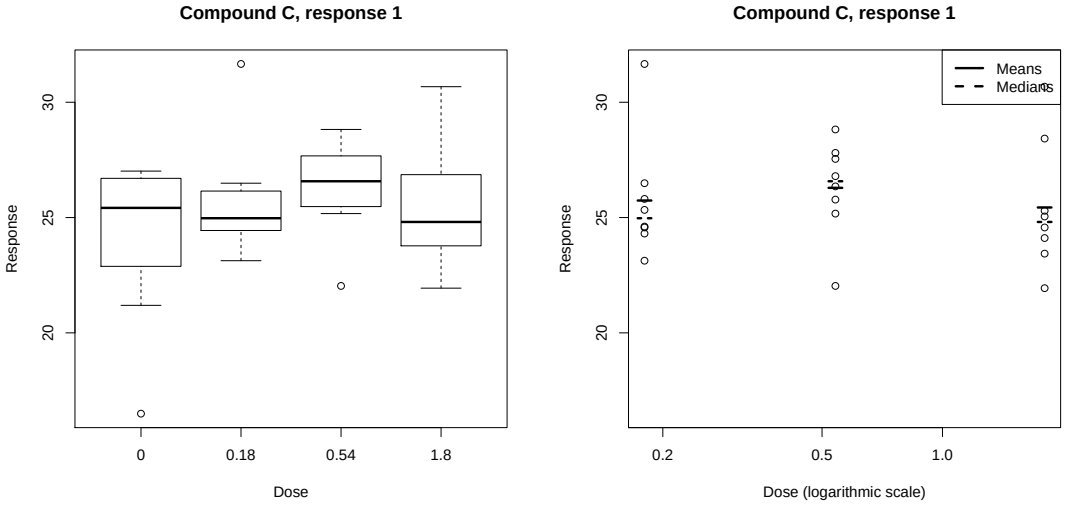

FIGURE 5. Boxplot (left) and scatterplot (right) of the observed dose–response data of compound C with respect to response 1. Dunnett's test declares no activity at significance level 5% (nor at 10%). The data suggest no apparent trend. The observations are quite symmetrical about the means, with no obvious heteroscedasticity, barring a few outliers.

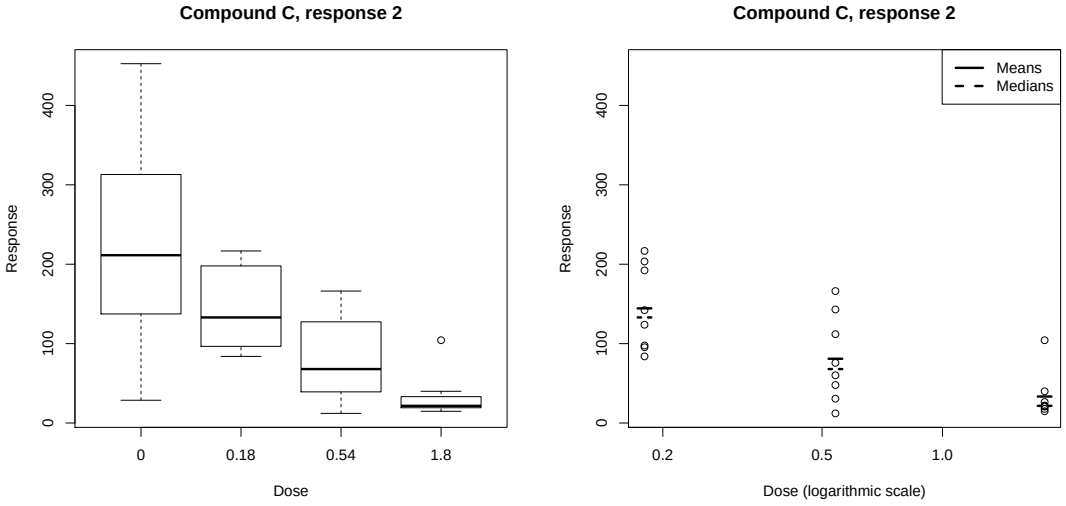

FIGURE 6. Boxplot (left) and scatterplot (right) of the observed dose–response data of compound C with respect to response 2. Dunnett's test indicates that the two highest dose-levels, 0.54 and 1.8, are active at the 5% significance level (both with p-values smaller than 0.4%). The data clearly indicate a decreasing trend. The observations are slightly positively skewed, perhaps indicating the propriety of a monotone transformation. Indeed, a logarithmic transformation of the response (which, for the sake of brevity, is not shown) does yield a dramatically more symmetrical spread around the means. It also alleviates the ostensible heteroscedasticity, with high variability in the vehicle treatment, as opposed to the remarkably low variability at dose-level 1.8. The logarithmically transformed data are however nearly homoscedastic.

2.2. **Mathematical models.** Throughout this thesis we will consider the functional relationship between an explanatory (predictor) variable x and a response variable y. Denoting the response function by η we will write $y = \eta(x, \theta)$ where θ is a vector containing all parameters of the given response function.

The $E_{\max}$ model and its derivatives. In pharmacodynamic modelling of dose–response relationships, one of the most common choices is the so-called $E_{\max}$ model [10] which can be motivated by using a biological interpretation [11]. Its response, in the case of an increasing dose–response profile, is given by

$$\eta(x, E_0, E_{\max}, \mathrm{ED}_{50}) = E_0 + \frac{E_{\max} \cdot x}{\mathrm{ED}_{50} + x},$$

where E_0 is the response at dose $x = 0$ (i.e. the vehicle response), $E_{\max}$ is the maximum effect induced by the drug, and ED_{50} is the smallest dose which attains half the maximal effect. Notice that $E_{\max}$ is the change in response *relative* to the baseline E_0, meaning that the absolute maximal response is $R_{\max} = E_0 + E_{\max}$.

The $E_{\max}$ model in the case of a decreasing dose–response relationship is analogous, with the response function

$$\eta(x, E_0, I_{\max}, \mathrm{ID}_{50}) = E_0 - \frac{I_{\max} \cdot x}{\mathrm{ID}_{50} + x},$$

where $I_{\max}$ is the maximum inhibition induced by the drug, and ID_{50} is the smallest dose which attains half the minimal effect. Henceforth, the inhibitory model will be included in the model with an increasing profile by allowing the parameter $E_{\max}$ to be negative.

The $E_{\max}$ model can be extended by raising the dose x and ED_{50} to a power h, which yields the *sigmoid* $E_{\max}$ model given by

$$\eta(x, E_0, E_{\max}, \mathrm{ED}_{50}, h) = E_0 + \frac{E_{\max} \cdot x^h}{\mathrm{ED}_{50}^h + x^h}.$$

The so-called Hill coefficient h is proportional to the slope of η at ED_{50} — see Gabrielsson and Weiner (2006) [11]. In order to distinguish the two models, the former is sometimes referred to as the hyperbolic $E_{\max}$ model [10].

Perhaps at the expense of biological interpretation, the sigmoid $E_{\max}$ model can be more succinctly expressed (by simplifying the quotient) as

$$\eta(x, E_0, E_{\max}, \mathrm{ED}_{50}, h) = E_0 + \frac{E_{\max}}{1 + (\mathrm{ED}_{50}/x)^h}.$$

This equivalent form of the sigmoid $E_{\max}$ model is referred to as the four-parameter logistic model [12]. Another equivalent form — useful for numerical purposes, or when applying a logarithmic transformation to the dose — is the logistic model

$$\eta(x, E_0, E_{\max}, \mathrm{ED}_{50}, h) = E_0 + \frac{E_{\max}}{1 + \exp[(\log \mathrm{ED}_{50} - \log x)h]}.$$

Additional dose–response models. Without giving biological interpretations, we shall also employ three dose–response models included in the `R` [7] package `DoseFinding` [8] (see the next Section 3 for more information about this package). These comprise an exponential model

$$\eta(x, E_0, E_1, \delta) = E_0 + E_1(\exp(x/\delta) - 1),$$

a linear-in-log model

$$\eta(x, E_0, \delta, \mathrm{offset}) = E_0 + \delta \log(x + \mathrm{offset}),$$

and a beta model

$$\eta(x, E_0, E_{\max}, \delta_1, \delta_2, \mathrm{scale}) = E_0 + E_{\max}\mathrm{B}(\delta_1, \delta_2)\Big(\frac{x}{\mathrm{scale}}\Big)^{\delta_1}\Big(1 - \frac{x}{\mathrm{scale}}\Big)^{\delta_2},$$

where $B(\cdot\,,\cdot)$ is the beta function and "scale" is a fixed dose-scaling parameter. Figure 7 displays standardised schematic plots conveying the typical shape of the above mentioned models. In particular, this figure reveals the ability of the beta model to capture a non-monotone dose–response profile.

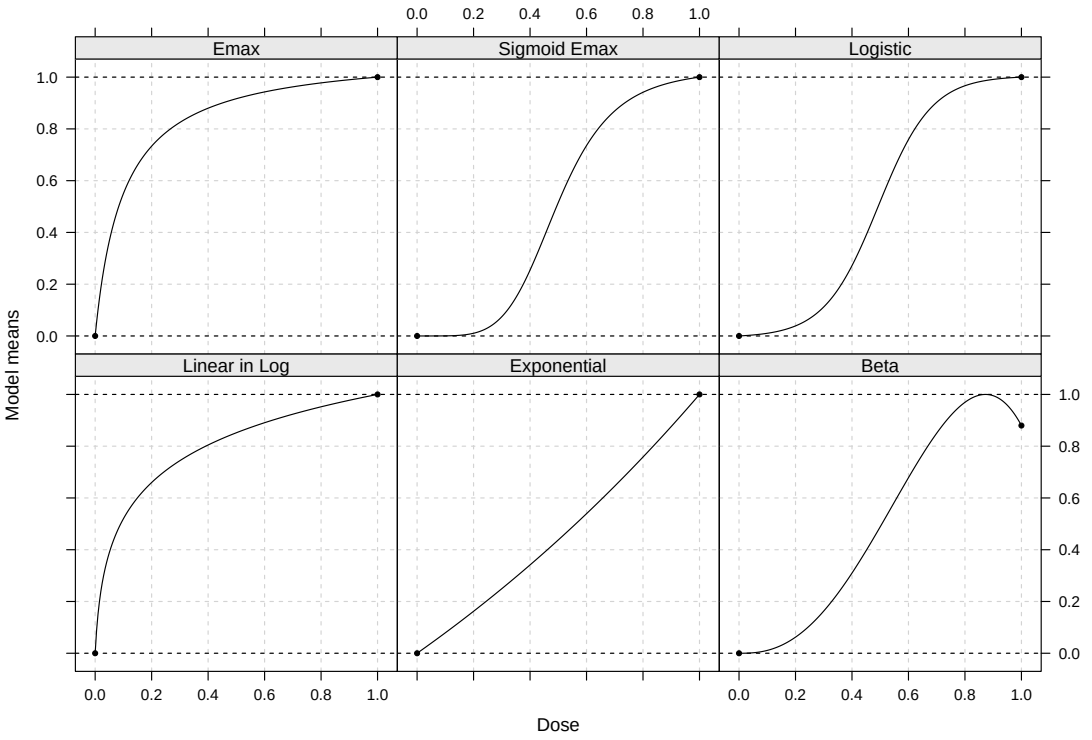

FIGURE 7. Standardised schematic plots of common dose–response models. All models are monotonic except for the beta model, which can accommodate a non-monotone dose–response profile. The graphic was created using the `R` [7] package `DoseFinding` [8].

3. Literature review

As mentioned above, the aim of this thesis is to combine an analysis of efficacy and a model-based approach. In the former analysis, when investigating the effect of various dose-levels (and vehicle) we traditionally utilise multiple comparison procedures to control the family-wise error rate. Conversely, the strength of the model-based approach is its potential to give more nuanced information about the underlying dose–response relationship. We have followed the work of Frank Bretz, José Pinheiro, Björn Bornkamp and Holger Dette (with specific references to follow).

Bretz et al. (2005) [1] devised a strategy for combining multiple comparison techniques with dose–response modelling, introduced in their so-called *MCP-Mod* methodology. The term is an abbreviation of "*m*ultiple *c*omparison *p*rocedures with *mod*elling techniques". Whilst this may sound very similar to the objective of this thesis, the multiple comparisons of which they speak pertain to model selection — not dose-levels. Their method can be used to select from a collection of candidate models (with specified parameters) the one which most aptly describes a given set

of dose–response data. The fundamental idea is to develop, for each model, a hypothesis test based on contrasts of the means of the responses at the various dose-levels (including the vehicle treatment). The contrasts for each model are chosen in such a way that if that particular model is correct, the chance of rejecting the null hypothesis (of no effect) is maximised. After establishing a critical value for adjusting the family-wise error rate of the multiple hypothesis tests, the models with significant test statistics can be identified from the candidate set. Subsequently, the winning model is chosen to be the one whose test statistic yielded the smallest p-value (adjusted for multiplicity). This leads to the main advantage of their approach, namely that the winning model can be used to make inferences beyond the original dose-levels. For instance, it can be used in dose-finding — that is, to estimate which dose ought to yield a given response. Constructing a confidence band around the curve corresponding to the winning model allows us to judge whether a given dose has a statistically significant effect compared to vehicle. They have implemented this work in the `R` [7] package `MCPMod` [13]. Their framework was expounded upon by Pinheiro et al. (2006) [2] who developed methods for power and sample size calculations under this methodology.

It is important to note, however, that their method is concerned with model selection *a posteriori* whereas we are interested in optimal design of experiments *a priori*. In other words, *MCP-Mod* can determine which model best describes a given data set, but our task is to design an experiment before having seen the resulting data. (Indeed, Pinheiro et al. (2006) [2] identified this as an avenue of further research.) Nevertheless, their work will prove beneficial also for our purposes, and the functionality of `MCPMod` was subsequently subsumed into the `R` package `DoseFinding` [8] written by the same authors. This latter package expands upon the *MCP-Mod* framework by including routines for optimal design, the theory of which was developed by Dette et al. (2008) [5]. In the cited article they show (via simulation) that locally optimal designs developed for common dose–response models — the exponential, linear-in-log, and $E_{\max}$ models — are moderately robust with respect to misspecification of the model parameters, but that they are far more sensitive to misspecification of the regression model itself. Their attempt to circumvent this problem is reminiscent of *MCP-Mod,* but rather than focusing on model selection, their analysis is geared toward optimal design of experiments. More concretely, their strategy is based on considering a collection of relevant candidate models, and choosing the design which maximises the minimum efficiency of the designs tailored to the members of that family. Thus, rather than designing for a single model (which may have been misspecified) they search for a design which is adequate for *all* candidate models. This makes the design more robust against model misspecifications.

The methodologies discussed thus far allow for the inclusion of a vehicle treatment, but Dette et al. (2014) [14] describe a method which in addition incorporates an active control. This may be a reference treatment — for instance a compound by a competitor — against which the compound under investigation is to be compared. It may also be a medical condition which the drug is meant to alleviate. Time has not permitted us to investigate this scenario, but it nevertheless provides a compelling avenue for further research.

4. Theory

4.1. **Optimal design of experiments.** In this section we shall briefly outline the theory of optimal experimental design needed for the purposes of this thesis. We will follow Fedorov and Leonov (2013) [12] in the formulation of the theory, but much inspiration has also been drawn from Atkinson et al. (2007) [15]. The following will necessarily be technically involved, but loosely speaking, optimal design theory is concerned with determining where to make measurements — and how many to make at each point — in order to gain as much information as possible from the experiment. This latter condition will depend on the criterion of optimality, but one common choice is the D-criterion, in which points of measurement are selected in such a way that, given some postulated model, the variance of its estimated parameters is minimised. For our purposes, when fitting for instance an $E_{\max}$ model, a D-optimal design ensures that the estimates of E_0, $E_{\max}$ and ED_{50} will be as precise as possible — provided that the underlying dose–response relationship can be accurately represented by an $E_{\max}$ model.

Consider an experiment whose response function $\eta(x, \theta)$ depends on the parameter vector $\theta \in \mathbb{R}^m$ containing m parameters. The predictor x belongs to the design region $\mathcal{X} \subseteq \mathbb{R}^k$. (For the purposes of this thesis, $k = 1$ since x is a univariate measure of dose.) For $i = 1, \dots, n$ let there be r_i replications made at design point (dose-level) $x_i \in \mathcal{X}$ and thus $N = \sum_{i=1}^{n} r_i$ observations in total. An *exact design*

$$\xi_N = \begin{Bmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{Bmatrix} \quad \text{where } p_i = r_i/N$$

is a probability measure on $\mathcal{X}$. Thus a design $\xi_N = \{x_i, p_i\}_{i=1}^n$ is simply a collection of support points $x_i \in \mathcal{X}$ and their corresponding weights p_i (which sum to unity).

A design is referred to as exact if the weight of the ith support point is $p_i = r_i/N$. A *continuous design* $\xi = \{x_i, p_i\}_1^n$ is one in which each p_i can be any real number in $[0, 1]$ under the constraint that $\sum_i p_i = 1$. In practice, all designs are of course exact (since the number r_i of measurements must necessarily be an integer). However, the ensuing optimisation problem in the discrete case is much more complex than its continuous counterpart. Therefore, in order to harness the power of convex optimisation theory, we shall focus on continuous designs. When having found the optimal continuous design ξ for some problem, this can be trivially approximated to yield the optimal exact design ξ_N.

The linear response model. The purpose of optimal design of experiments is to strategically determine the design points $x_i \in \mathcal{X}$ and their corresponding weights p_i. The optimal allocation of design points x_i will of course depend on the underlying model, so we must make an assumption about the response function $\eta(x, \theta)$. We shall begin by investigating the linear model $\eta(x, \theta) = \theta^\top f(x)$ where $\theta = (\theta_1, \dots, \theta_m)^\top$ contains the m parameters and $f(x) = [f_1(x), \dots, f_m(x)]^\top$ contains the m base functions. (Thus the model is linear with respect to the parameters, whereas the base functions need not be linear.) The motivation for this apparent limitation is that when we eventually consider nonlinear models, these will be linearised using their first-order Taylor approximations. (See Section 4.1 for an elaboration on this simplification.)

We consider the probabilistic model $Y_{ij} = \theta^\top f(x_i) + \varepsilon_{ij}$ whose errors ε_{ij} are uncorrelated with zero mean and homoscedastic variance σ^2. The least-squares estimator

$$\hat{\theta}_N := \arg\min_\theta \sum_{i=1}^{n} \sum_{j=1}^{r_i} [Y_{ij} - \eta(x_i, \theta)]^2$$

can be explicitly found via the *normal equations*

$$\mathbf{M}(\xi_N)\theta = \mathcal{Y}$$

where $\mathbf{M}(\xi_N) = \sigma^{-2}\sum_{i=1}^{n} r_i f(x_i) f^\top(x_i)$ and $\mathcal{Y} = \sigma^{-2}\sum_{i=1}^{n} r_i \bar{Y}_i f(x_i)$ with $\bar{Y}_i = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij}$. Provided that $\mathbf{M}(\xi_N)$ is invertible, it follows that

$$\hat{\theta}_N = \mathbf{M}^{-1}(\xi_N)\mathcal{Y}.$$

Furthermore, it can be shown [12] that $\operatorname{Var}\hat{\theta}_N = \mathbf{M}^{-1}(\xi_N)$. Thus, if our goal is to reduce the variability of the parameter estimator $\hat{\theta}$, it is natural to seek to minimise (in some sense) its variance-covariance matrix $\mathbf{M}^{-1}(\xi_N)$.

Under normal theory, a confidence region for θ with coverage probability $1-\alpha$, where $\alpha \in (0,1)$, is given by

$$\{\theta \in \mathbb{R}^m : (\theta - \hat{\theta}_N)^\top \mathbf{M}(\xi_N)(\theta - \hat{\theta}_N) \leqslant \chi^2_{m,\alpha}\},$$

where $\chi^2_{m,\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with m degrees of freedom. The volume of this ellipsoid is proportional to $\sqrt{\det \mathbf{M}^{-1}(\xi_N)}$. This illustrates the sensibility in focusing on minimising the determinant $\det \mathbf{M}^{-1}(\xi_N)$ of $\mathbf{M}^{-1}(\xi_N)$ (since $\sqrt{\cdot}$ is monotonic). In fact, this is precisely the criterion of the celebrated [16] D-optimality, upon which our analysis shall rest.

The main optimisation problem. To recapitulate, an exact design is said to be D-optimal if it minimises $\det \mathbf{M}^{-1}(\xi_N)$ across all exact designs ξ_N with N support points. For a general optimality criterion Ψ, the optimisation problem is to find $\arg\min_{\xi_N} \Psi[\mathbf{M}(\xi_N)]$ where Ψ is a (scalar-valued) functional. Under the D-criterion, $\Psi[\mathbf{M}(\xi_N)] = \det \mathbf{M}^{-1}(\xi_N)$.

As previously mentioned, the discrete optimisation problem

$$\xi_N^* = \arg\min_{\xi_N} \Psi[\mathbf{M}(\xi_N)]$$

is computationally expensive. However, its continuous counterpart may readily be solved using the theory of convex optimisation [12]. We shall therefore consider as our main optimisation problem

$$\xi^* = \arg\min_{\xi} \Psi[\mathbf{M}(\xi)]$$

where optimisation is performed over the set of all probability measures on the design region $\mathcal{X}$.

The nonlinear response model. Suppose that $Y_{ij} = \eta(x_i, \theta) + \varepsilon_{ij}$ where η is nonlinear (in the parameters) and the errors ε_{ij} are uncorrelated with zero mean and homoscedastic variance σ^2. The least-squares estimator is defined by $\hat{\theta}_N := \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} r_i[\bar{Y}_i - \eta(x_i,\theta)]^2$ where Θ is a compact subset of $\mathbb{R}^m$ and $\bar{Y}_i = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij}$. Generally there is no closed-form expression for $\hat{\theta}_N$, so a numerical approach must be employed. Furthermore, $\hat{\theta}_N$ is a biased estimator of the true parameter value θ_t, but under mild assumptions [12] it is strongly consistent — that is, $\hat{\theta}_N$ converges almost surely to θ_t as $N \to \infty$.

As alluded to above, we shall transfer the nonlinear problem to the linear framework by considering its Taylor expansion about the true parameter value θ_t. Thus we linearise η by

$$\eta(x,\theta) \approx \eta(x,\theta_t) + (\theta - \theta_t)^\top g(x,\theta_t),$$

where

$$g = \frac{\partial \eta}{\partial \theta} = \left[\frac{\partial \eta}{\partial \theta_1}, \ldots, \frac{\partial \eta}{\partial \theta_m}\right]^\top$$

is the gradient of η with respect to θ (and the arguments (x,θ) have been omitted for brevity).

It could well be argued that linearisation is too crude an approximation, and that higher order terms ought to be included in the analysis. However, Fedorov and Leonov question whether it is worthwhile to use significantly more complicated methods, considering that the postulated model is merely a simplification of the real phenomena [12]. They point out that there are more pressing concerns afoot, such as the assumption of homoscedastic noise, or indeed the assumption of the underlying model itself, which, if inappropriate, can have dire consequences [5]. Thus, whereas

it is admittedly convenient to settle for a linear approximation, we believe that it is a matter of subordinate concern.

The linear approximation leads to normal equations containing the information matrix

$$\mathbf{M}(\xi_N, \theta_t) = \sigma^{-2} \sum_{i=1}^{n} r_i g(x_i, \theta_t) g^\top(x_i, \theta_t),$$

which now (unlike in the purely linear case) depends on the true parameter value θ_t. Hence, the corresponding (continuous) optimal design $\xi^*(\theta_t) := \arg\min_\xi \Psi[\mathbf{M}(\xi, \theta_t)]$ also depends on the unknown parameter value θ_t. This important contrast between optimal designs in the linear case (in which they do *not* depend on the parameter) and the nonlinear case (in which they *do* depend on the parameter) led Chernoff (1953) [17] to term the latter *locally* optimal designs. In other words, an optimal design for a nonlinear response model will be optimal for a specific parameter value, but not necessarily for others. In this sense it is local with the respect to the parameter.

Now, provided that the above linearisation is valid, we find that $\operatorname{Var} \hat{\theta}_N \approx \mathbf{M}^{-1}(\xi, \theta_t)$. This depends on the true parameter value θ_t, but since $\hat{\theta}_N$ is strongly consistent (under mild assumptions), we feel justified in making the final approximation $\operatorname{Var} \hat{\theta}_N \approx \mathbf{M}^{-1}(\xi, \hat{\theta}_N)$. Recall that under the D-criterion, the goal is to minimise the determinant of this variance-covariance matrix.

Ill-conditioned problems. It is quite possible for the information matrix of a given model to be singular (or nearly so) at some point in the design space. Consider for instance the sigmoid $E_{\max}$ model

$$\eta(x, E_0, E_{\max}, \mathrm{ED}_{50}, h) = E_0 + \frac{E_{\max}}{1 + (\mathrm{ED}_{50}/x)^h}.$$

It is straightforward to analytically calculate that

$$\frac{\partial \eta}{\partial h} = \frac{E_{\max} (\mathrm{ED}_{50}/x)^h}{\left(1 + (\mathrm{ED}_{50}/x)^h\right)^2} (\ln x - \ln \mathrm{ED}_{50}).$$

The factor $(\ln x - \ln \mathrm{ED}_{50})$ ensures that the partial derivative $\frac{\partial \eta}{\partial h}$ approaches and attains zero as the dose x tends to ED_{50}. Consequently, the information matrix will be singular at this point (and near-singular in its neighbourhood). Pázman [18] and Silvey [19] have developed remedies for singular information matrices, though we shall not expound upon their detail. Instead we shall make use of the expertly written software in the `R` package `DoseFinding` [8] which circumvents this issue by utilising generalised inverses and singular value decomposition.

D- and ED_p-optimal designs for select models. In the above we have mentioned D-optimal designs, which minimise the variance of the parameter estimator $\hat{\theta}$. In some cases, however, only a subset of the parameter vector is of primary interest. For instance, if designing for a model in the $E_{\max}$ family, we may be chiefly concerned with getting a precise estimate of ED_{50}, whereas the variability of the estimates of E_0 and $E_{\max}$ is of subordinate concern. In such a case we may employ an ED_{50}-optimal design, which minimises the variance of the estimator of the dose which attains half the maximal effect. To make this notion slightly more general, for each $p \in (0, 1)$ let ED_p be the dose which attains $100p\%$ of the maximal effect $E_{\max}$. A design which minimises the variance of the estimator of ED_p is said to be ED_p-optimal. This is a special case of a c-optimal design, which minimises the variance of the best (uniform minimum variance) linear unbiased estimator of a given linear combination of the model parameters in θ.

Dette et al. (2010) [20] showed that both D- and ED_p optimal designs for the $E_{\max}$, exponential and linear-in-log models (see Section 2.2) are supported by precisely three design points — two of which lie at the boundary of the design space. The weights assigned to the points will vary depending on the optimality criterion (D or ED_p) but the locations of the points are the same.

It may be surprising that optimal designs for these models include so few dose-levels. Intuitively, we would probably prefer to have more dose-levels, spread out across the design space. This would likely aid assessment of the dose–response profile. However, when designing for a specific model, the only consideration is to find the allocation of measurements which is optimal for that particular model. Intuition is perhaps more cautious and wary about committing so fully to the model assumption, which is the topic of the upcoming section.

Efficient designs robust against model misspecifications. As discussed above, optimal designs depend on the assumption of the underlying model. (This is true for both linear and nonlinear models, but in the latter case the design also depends on the model parameters.) Thus a design which is optimal for a certain model is generally not optimal for another. As we mentioned in the literature review (Section 3) Dette et al. (2008) [5] showed that c-optimal designs for common dose–response models are sensitive to misspecifications of the underlying model. In the same article, they develop a method for constructing designs which are robust against such model misspecifications. Rather than constructing a design which is optimal for any single model, their method yields a design which is appropriate for a collection of candidate models. This collection, which must be specified beforehand by the experimenter, should contain all models likely to capture the underlying phenomena, but preferably no others.

We shall describe the gist of this method, but all details can be found in the article by Dette et al. (2008) [5]. The benchmark used for judging the aptitude of a given design is its *efficiency.* Under a general Ψ-criterion of optimality, the Ψ-efficiency of a design ξ is defined by

$$\text{eff}_\Psi(\xi, \theta) = \frac{\Psi(\xi^*(\theta), \theta)}{\Psi(\xi, \theta)}$$

where $\xi^*(\theta)$ is a locally optimal design (which, being local, depends on the parameter θ). Recall that the optimal design minimises the functional Ψ, meaning that the efficiency of a given design ξ lies between 0 and 1. Say that we under the D-criterion have a design ξ whose efficiency $\text{eff}_\Psi(\xi, \theta) = 50\%$. Then the optimal design $\xi^*(\theta)$ would estimate the parameters with the same precision as ξ using only half the number of measurements [5]. Alternatively, $\xi^*(\theta)$ would yield a more precise confidence region for the parameter estimator $\hat{\theta}$.

Now, since a locally optimal design depends on the underlying model assumption, so does its efficiency. Thus, for a given design we must consider its efficiency with respect to each candidate model. Dette et al. define a local *maximin* optimal design to be a design which maximises the minimum efficiency across all candidate models. In practice, this design must be found numerically, and they have incorporated software for this in the package `DoseFinding` [8] via the function `optDesign`. This software also includes a Bayesian counterpart for constructing adaptive designs under multistage trials. We shall, however, focus on the minimax procedure. Incidentally, the minimax procedure also provides the opportunity to include prior information, namely by weighting the candidate models according to their postulated relevance.

4.2. **Nonlinear regression.** The function `fitMod` in the `R` package `DoseFinding` [8] is very convenient for performing nonlinear least-squares regression for common dose–response models. For instance, if we have empirical dose–response data stored in the vectors `x` and `Y`, respectively, we can fit an $E_{\max}$ model to the data by calling `fitMod(dose = x, resp = Y, model = "emax")`. This command will return an object whose information can be retrieved using generic `R` functions such as `coef` (for extracting the estimated model parameters) or `vcov` (for computing the variance-covariance matrix of the estimated parameters).

Analogously, we can fit a sigmoid $E_{\max}$ model to the data by calling `fitMod(dose = x, resp = Y, model = "sigEmax")`. As discussed in Section 2.2 the sigmoid $E_{\max}$ model augments the ordinary $E_{\max}$ model by the inclusion of a slope parameter h, which gives the sigmoid model more

flexibility. However, this additional parameter (which enters the model in a nonlinear fashion) also makes it more difficult to estimate the parameters of the sigmoid model. An attempt to address this issue is to simply fix some of the model parameters; most commonly E_0 and $E_{\max}$. The parameter E_0 can be gauged from the vehicle response, since this is indeed a realisation of the baseline response. The parameter $E_{\max}$ is not quite as straightforward to estimate, but if we interpret $E_{\max}$ as the maximum effect seen within the dose-range — as opposed to the maximum effect as the dose tends to infinity — then this parameter may be sensibly estimated from the range of the observed responses.

The function `fitMod` does not support fixation of parameters in the regression model. It would be natural to instead use the `R` function `nls` to perform nonlinear least squares, but its built-in algorithms (e.g. Gauss–Newton) often encounter issues of convergence with the sigmoid $E_{\max}$ model. To remedy this, we wish to utilise the Levenberg–Marquardt algorithm, which is more robust since it mixes the Gauss–Newton algorithm with the method of steepest descent. The `R` package `minpack.lm` (where "lm" stands for "Levenberg–Marquardt") provides this functionality. It is most conveniently utilised via the wrapper function `nlsLM` whose syntax is very similar to that of the ordinary `nls` function, and which supports generic `R` functions such as `coef` and `vcov`. Since `nlsLM` requires the user to specify the regression model, we are free to fix any of its parameters. For instance, when fitting a sigmoid $E_{\max}$ model, we may fix E_0 and $E_{\max}$ based on prior knowledge (such as a pilot study). This lets the nonlinear regression focus on estimating ED_{50} and the slope parameter h, which in turn yields more precise parameter estimates — of course at the cost of having fixed some of them.

4.3. **Construction of confidence bands.** When fitting a regression curve to data, it is important to assess the level of confidence in the fit. The role played by a confidence interval for a single quantity can be extended to a confidence *band* serving the same purpose for a curve. This concept can most clearly be illustrated in the case of simple linear regression. Suppose that we make N observations $\{(x_i, Y_i)\}_{i=1}^N$ under the probabilistic model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ whose errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently. Notice that each x_i is considered fixed and that each Y_i is random by virtue of the random error ε_i. The parameter estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively, are based on ordinary least squares. In Figure 8 we have fitted a regression line to perturbed data generated from the true model $y = 2 + 3x$. The curved dashed lines constitute a pointwise 95% confidence band around the regression line. It is constructed by computing at each x_i a 95% confidence interval for the fitted value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Then the upper confidence limits for each fitted value are joined by straight lines in order to outline the upper curve of the confidence band — and analogously for the lower confidence limits. The curvature is caused by the fact that an observation Y_i at a predictor value x_i far from the mean $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ will exert greater leverage on the fitted value $\hat{Y}_i$, and therefore increase its variance, which in turn will widen the confidence interval around $\hat{Y}_i$. More concretely [21] the variance of the ith fitted value is

$$\operatorname{Var} \hat{Y}_i = \left[\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^N (x_k - \bar{x})^2}\right]\sigma^2,$$

where N is the total number of observations. Thus the variance of the ith fitted value $\hat{Y}_i$ increases with the distance of x_i from the mean $\bar{x}$.

It is important to note that confidence bands are intended to capture the *fitted values* $\hat{Y}_i$, i.e. the mean profile of the regression curve. It does *not* give information about the region in which new observations are likely to appear. This task is instead fulfilled by *prediction bands* which are wider than the corresponding confidence bands, due to the increased uncertainty introduced by the variance of the new observation. For instance, suppose that we wish to predict the response $Y^\star$ at some point $x^\star$. Since the errors are centred around zero, our best estimate of $Y^\star$ is its

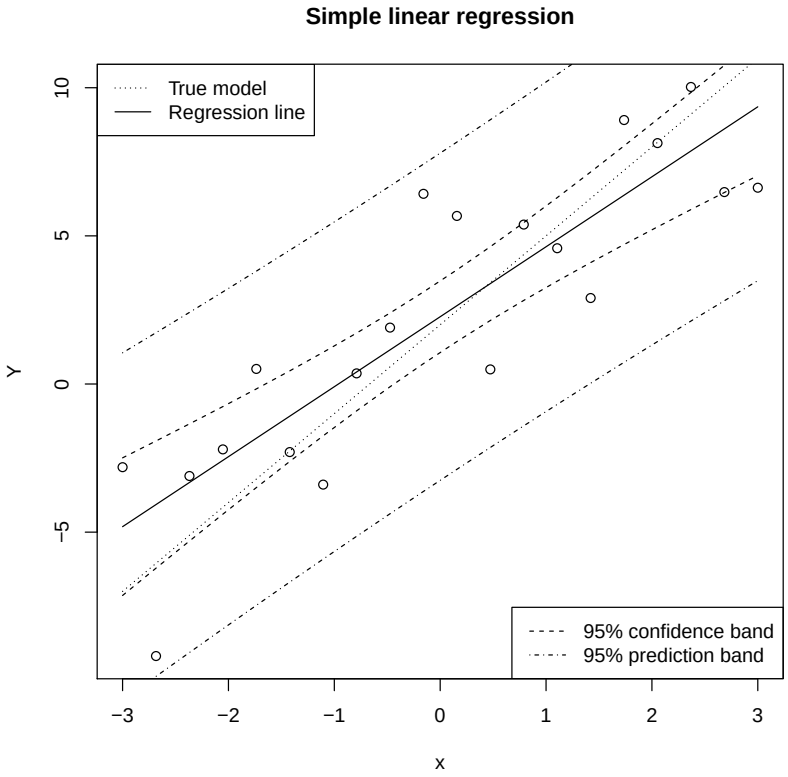

FIGURE 8. Simple linear regression with 95% pointwise confidence and prediction bands. The confidence band is designed to capture the true model, which in this case has been accomplished. In contrast, the prediction band is designed to capture the variability of individual observations.

mean $\hat{Y}^\star = \hat{\beta}_0 + \hat{\beta}_1 x^\star$. It can be shown [21] that

$$\operatorname{Var} \hat{Y}^\star = \left[1 + \frac{1}{N} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2}\right]\sigma^2.$$

In Figure 8 we have constructed a 95% prediction band alongside the 95% confidence band to illustrate this difference. On average, 95% of new observations ought to fall within the wider region delimited by the prediction band. The confidence band, however, is designed to capture, with some level of confidence, the true model. We say "*some* level of confidence" because pointwise confidence bands do not in general attain the nominal confidence level of the individual confidence intervals from which they are constructed, due to the problem of multiple testing [22].

The problem of multiple comparisons. Suppose that we generate a pointwise confidence band by joining the upper and lower limits of all $100(1-\alpha)\%$ confidence intervals constructed around the fitted values of N points. Then at each given point, the probability that the corresponding confidence interval fails to cover the true response is α. However — presuming that the confidence intervals are independent of each other — the probability that a confidence interval at *some* point will fail to cover the true response (called the family-wise error rate) is $\text{FWER} = 1 - (1-\alpha)^N$.

Writing $\alpha = 1 - (1 - \alpha)$ and noting that $(1 - \alpha) < 1$ tells us that FWER $> \alpha$ for $N > 1$. For typical values of α and N, the family-wise error rate is dramatically larger than α. For instance, if $\alpha = 0.05$ and $N = 10$, then FWER $\approx 40\%$ which is eight times as large as the nominal type I error rate of 5%. Admittedly, these computations rely on the assumption of independence, which is violated when fitting a smooth curve since its continuity introduces a dependency structure between the confidence intervals surrounding the fitted values of neighbouring points. Nevertheless, the coverage level of the pointwise confidence band will generally be lower than the nominal confidence level $1 - \alpha$ of the constituent confidence intervals [22]. To remedy this problem, we can attempt to construct *simultaneous* confidence bands, which are adjusted for multiplicity in order to attain the nominal confidence level. In the upcoming sections we shall investigate both approaches.

According to Hall and Horowitz (2013) [23] pointwise confidence bands are, despite their undercoverage, ubiquitous amongst practitioners and in the literature. In the cited article they present a bootstrap method for controlling the asymptotic coverage level of a pointwise confidence band for a specified proportion of the predictor values. A pointwise confidence band constructed using their methodology will, in other words, have a $1-\alpha$ coverage level for a $1-\zeta$ proportion of the predictor values, where α and ζ are specified in advance. Their method is a highly compelling alternative to pointwise and simultaneous bands, and it works very well for data whose predictor space is densely populated. However, since it is based on nonparametric regression via a local linear estimator, it appears to have difficulties with discrete dose-levels which are separated by more than the estimated bandwidth. We have tried using both an Epanechnikov kernel and a Gaussian kernel, but to no avail. We also tried appropriating the method for parametric regression (at the risk of violating the assumptions upon which its theoretical foundation depends) but without success. Hence we shall not pursue this method further. Instead, we will focus on constructing both pointwise and simultaneous confidence bands under normal theory, and also, in the former case, using a bootstrap approach.

First-order Taylor approximation for estimation of variance. In order to construct a confidence band for a nonlinear function $\eta(x, \theta)$ we must at each point x gauge the variance of $\eta(x, \theta)$. Thus we consider the predictor point x to be fixed, with all uncertainty about η coming from the parameter estimator. In the present section it is important to distinguish between the parameter θ and its true unknown value, which, as before, is denoted θ_t.

Following Casella and Berger [24] we let $\theta = (\theta_1, \ldots, \theta_m)^\top$ be a random vector with mean $\mathbb{E}\theta = \theta_t = (\theta_{t,1}, \ldots, \theta_{t,m})^\top$. Suppose that $\eta(x, \theta)$ is a real-valued differentiable function. Treating x as fixed, the first-order Taylor approximation of η about θ_t is

$$\eta(x, \theta) \approx \eta(x, \theta_t) + \sum_{i=1}^{m} \frac{\partial \eta}{\partial \theta_i}(x, \theta_t)(\theta_i - \theta_{t,i}).$$

Using vector notation, this can be written as

$$\eta(x, \theta) \approx \eta(x, \theta_t) + (\theta - \theta_t)^\top \frac{\partial \eta}{\partial \theta}(x, \theta_t), \tag{1}$$

where $\frac{\partial \eta}{\partial \theta} = (\frac{\partial \eta}{\partial \theta_1}, \ldots, \frac{\partial \eta}{\partial \theta_m})^\top$ is the gradient of η with respect to θ. Moreover, since $\mathbb{E}[\theta_i - \theta_{t,i}] = 0$ it follows that $\mathbb{E}\eta(x, \theta) \approx \eta(x, \theta_t)$. Both of these approximations will be used in the first two

steps of the following derivation of the variance of η at x with respect to θ.

$$\operatorname{Var}\eta(x,\theta) \approx \mathbb{E}[\eta(x,\theta)-\eta(x,\theta_t)]^2 \approx \mathbb{E}\Big[\sum_{i=1}^{m}\frac{\partial\eta}{\partial\theta_i}(x,\theta_t)(\theta_i-\theta_{t,i})\Big]^2$$
$$=\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial\eta}{\partial\theta_i}(x,\theta_t)\frac{\partial\eta}{\partial\theta_j}(x,\theta_t)\operatorname{Cov}(\theta_i,\theta_j)$$

Using matrix notation — and denoting by $\mathbf{V}$ the variance-covariance matrix of the random vector θ — this can be written

$$\operatorname{Var}\eta(x,\theta) \approx \frac{\partial\eta}{\partial\theta}(x,\theta_t)^\top\mathbf{V}\frac{\partial\eta}{\partial\theta}(x,\theta_t).$$

This approximation will be utilised in the upcoming section.

Simultaneous confidence bands under normal theory. Gsteiger et al. (2011) [22] describe a method for constructing simultaneous confidence bands for nonlinear mixed-effects models under normal theory. This can easily be applied to our nonlinear fixed-effects model $Y_{ij}=\eta(x_{ij},\theta)+\varepsilon_{ij}$ where we now add an assumption of normality by supposing that $\varepsilon_{ij}\sim\mathcal{N}(0,\sigma^2)$ independently.

Provided that the maximum likelihood estimate $\hat\theta$ is close to the true parameter value θ_t, the Taylor linearisation in Equation (1) above can be used to yield that

$$\eta(x,\hat\theta)-\eta(x,\theta_t) \approx (\hat\theta-\theta_t)^\top\frac{\partial\eta}{\partial\theta}(x,\hat\theta),$$

where we have substituted $\hat\theta$ for θ_t in the argument of the gradient $\frac{\partial\eta}{\partial\theta}$ based on the assumption of its closeness to the true value θ_t. The assumption of normality implies that $\eta(x,\hat\theta)-\eta(x,\theta_t)$ is approximately $\mathcal{N}\big(0,\frac{\partial\eta}{\partial\theta}(x,\hat\theta)^\top\mathbf{V}\frac{\partial\eta}{\partial\theta}(x,\hat\theta)\big)$, where $\mathbf{V}$ is the variance-covariance matrix of $\hat\theta$. Thus, in order to construct a simultaneous confidence band for the curve over the predictor (covariate) region $\mathfrak{X}:=[x_{\min},x_{\max}]$ we need only calibrate the critical value d so that

$$\mathbb{P}\big(\eta(x,\theta_t)\in\eta(x,\hat\theta)\pm d\sqrt{\frac{\partial\eta}{\partial\theta}(x,\hat\theta)^\top\mathbf{V}\frac{\partial\eta}{\partial\theta}(x,\hat\theta)}:x\in\mathfrak{X}\big)=1-\alpha,$$

where $1-\alpha$ is the nominal confidence level and $\alpha\in(0,1)$. Note that setting d to be the $(1-\alpha)$-quantile of the standard normal distribution yields a *pointwise* confidence band with confidence level $(1-\alpha)$. At first glance it may not be obvious that the variability of the data influences the width of the confidence band, but it does so implicitly through the parameter estimate $\hat\theta$ and its variance-covariance matrix $\mathbf{V}$.

In their paper, Gsteiger et al. describe two methods for determining the critical value d. The first method is based on an asymptotic χ^2 distribution, and leads to setting d to the square root of the $(1-\alpha)$-quantile of the χ^2_m distribution with m degrees of freedom, where m is the number of parameters in θ. The second method relies on iterated simulation from a multivariate normal distribution, and it generally yields smaller critical values d than does its approximation-based counterpart.

In the same paper, Gsteiger et al. carried out an extensive simulation study in order to assess the empirical coverage of the confidence bands constructed using critical values determined by either of the two methods. With $\alpha=0.05$ and five distinct dose-levels, they found that the approximation-based method tends to be conservative for sample sizes larger than 25. In other words, the approximation-based confidence bands tend to have an empirical coverage greater than the nominal 95% confidence level (of course at the cost of increased width and thereby less precision). In contrast, the simulation-based confidence bands generally undercover with respect to the nominal level. Specifically, their empirical coverage ranges (in this study) from 89% to 93% for sample sizes between 10 and 50. In summary, the simulation-based method will tend to

yield narrower bands which are likely to undercover, whereas the approximation-based method will generally yield excessively wide and conservative bands.

Since we will run extensive simulations of dose–response trials, it would be too computationally demanding to use the simulation-based method for calibrating the critical value d. We have therefore chosen to use the approximation-based method, which tends to be conservative for large sample sizes.

A simple bootstrap method for pointwise confidence bands. The following method is suggested by Hastie et al. (2009) [25]. Suppose that we make N observations $\{(x_i, Y_i)\}_{i=1}^N$ and that a model $\eta(x, \theta)$ is fitted to the data, yielding the parameter estimate $\hat{\theta}$. Next we compute the residuals $\tilde{\varepsilon}_i = Y_i - \eta(x_i, \hat{\theta})$ and the centred residuals $\hat{\varepsilon}_i = \tilde{\varepsilon}_i - \frac{1}{N}\sum_{i=1}^N \tilde{\varepsilon}_i$. The latter will be used during the following bootstrap procedure, which will be iterated for each $j \in [1, B] \cap \mathbb{Z}$ where $B \geqslant 1000$.

(1) For $i = 1, \ldots, N$ resample $Y_i^* = \eta(x_i, \hat{\theta}) + \varepsilon_i^*$ where ε_i^* is sampled *with* replacement from the centred residuals $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N\}$.
(2) Fit $\eta(x, \theta)$ to the resampled data, yielding the parameter estimate $\hat{\theta}_j^*$.
(3) Save $\hat{\theta}_j^*$.

This procedure will yield B curves, each of which has been fitted to some resampling of the original data. Next we construct a pointwise confidence band by determining empirical $(1 - \alpha)$ coverage intervals for each predictor point. In other words, for each x in the predictor region $\mathfrak{X} := [x_{\min}, x_{\max}]$ we consider the set $\{\eta(x, \hat{\theta}_j^*) : 1 \leqslant j \leqslant B\}$ of the B fitted values at that particular point. The empirical $(\alpha/2)$- and $(1 - \alpha/2)$-quantiles of this set are chosen as the lower and upper limits of the confidence band at this point. Notice that this procedure is a simulation-based realisation of the definition of a pointwise confidence band.

Examples. Figure 9 displays a sigmoid $E_{\max}$ model (selected on the basis of Akaike's information criterion) which has been fitted to the dose–response data of compound A with respect to response 1. Also included are 95% confidence bands constructed using the methods described above. The bands based on normal theory are constructed in the same manner, only using different critical values to determine the width of the band. As expected, the simultaneous confidence band is wider than its pointwise counterpart, in order to adjust for multiplicity. It is however encouraging to see that the pointwise confidence band generated using bootstrap agrees rather well with the pointwise confidence band based on normal theory. They deviate most in the regions where the regression curve changes most rapidly, which is where curve fitting is most challenging.

In this case we had no difficulty fitting the regression curve and its confidence band to the data. However, the sigmoid $E_{\max}$ model can sometimes be difficult to fit for numerical reasons, yielding unwieldy confidence bands which are much too wide to be useful. In such a situation we can fix some of the model parameters (as described in Section 4.2) in order to get more precise estimates of the remaining parameters. Alternatively, we can in such a situation rely more heavily on the confidence bands based on bootstrap, since they suffer no such numerical problems.

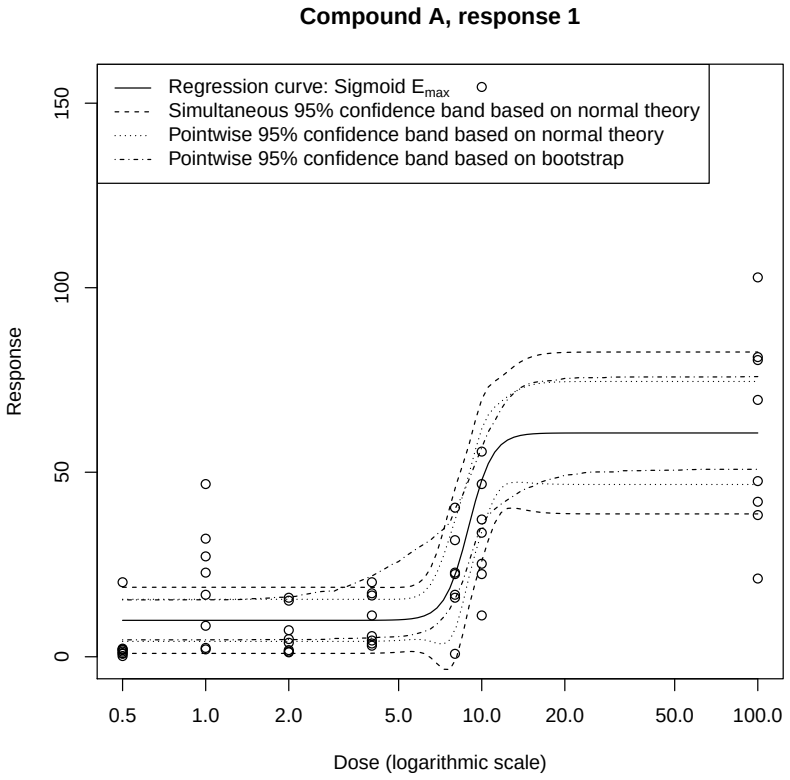

FIGURE 9. Regression curve of a sigmoid $E_{\max}$ model fitted to the dose–response data of compound A with respect to response 1. Three 95% confidence bands have been constructed using the methods described above.

5. Implemented method

We are now ready to address the key question of this thesis, namely the synthesis of an efficacy study and pharmacodynamic modelling of dose–response data. Concretely, this is attained by using a simultaneous confidence band around a regression curve (adjusted by subtracting the mean vehicle response) to declare whether a given dose in the administered dose-range is active. A dose is said to be *active* if it shows a statistically significant effect compared to the vehicle treatment (at the given significance level, which in this thesis is set to 5%). It is crucial that we construct a confidence band around the *difference* between the regression curve and the mean vehicle response. Doing so assures that we account for the variability in the former as well as the latter.

ALGORITHM 1 (Model-based approach to assess efficacy of a given dose). **This is the key result of this thesis.**

(1) Use least-squares regression to fit a model $\eta(x, \hat{\theta})$ to dose–response data.
(2) Subtract (uniformly) the mean vehicle response $\bar{V}$ from this curve to yield the difference $\eta(x, \hat{\theta}) - \bar{V}$.

(3) Assuming that the fitted value $\eta(x, \hat{\theta})$ and the mean vehicle response $\bar{V}$ are uncorrelated, the variance of their difference is $\operatorname{Var}\eta(x, \hat{\theta}) + \operatorname{Var}\bar{V}$. The Taylor approximation in Section 4.3 can be used to estimate $\operatorname{Var}\eta(x, \hat{\theta})$, and an unbiased estimate of $\operatorname{Var}\bar{V}$ is given by $\frac{1}{n(n-1)}\sum_{i=1}^{n}(V_i - \bar{V})^2$, where $V_1, \ldots, V_n$ are the n vehicle responses [24].
(4) Specify a significance level $\alpha \in (0, 1)$ and use the method in Section 4.3 to construct a simultaneous $100(1-\alpha)\%$ confidence band around the difference $\eta(x, \hat{\theta}) - \bar{V}$. A common choice, used throughout this thesis, is $\alpha = 5\%$.
(5) Declare a given dose x_0 to have a significant effect compared to vehicle (at the specified significance level) if the confidence band around $\eta(x_0, \hat{\theta}) - \bar{V}$ at this point x_0 does *not* contain zero.

Figure 10 shows an illustration of the above method, using the empirical dose–response data pertaining to response 1 of compound A. Since the confidence band is adjusted for multiplicity, it is legitimate to use it for simultaneous inference about all doses in the administered dose-range.

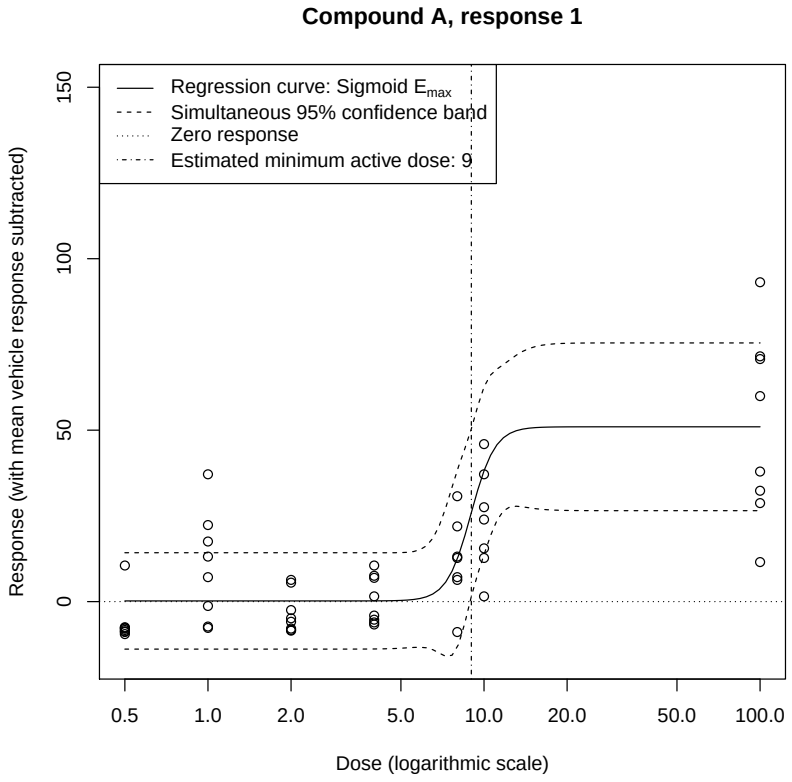

FIGURE 10. An illustration of the fundamental idea behind judging the activity of doses under the modelling approach. The accuracy of the vehicle-adjusted regression curve is assessed by the dashed confidence band. A dose is deemed active precisely when the confidence band around its fitted value does not cover the zero response. In the case shown in this figure, where the lower limit of the confidence band does not return below zero after first exceeding it, any dose higher than the estimated minimum active dose is deemed active.

As a comparison, we perform one-way ANOVA on the same data set, and apply Dunnett's test [6] in order to compare each dose-level against the vehicle treatment. According to this test, only the two highest dose-levels, 10 and 100, are deemed active at the 5% significance level. This is in agreement with the conclusion drawn from the modelling approach, since the minimum active dose is estimated to lie between dose-levels 8 and 10. Hence this example is ideal in the sense that both methods agree on which of the (discrete) administered dose-levels are active. In addition, the modelling approach provides information about where between the highest inert dose-level and the lowest active dose-level we begin to see a signal. The reliability of this information of course depends on the validity of the model assumptions (as we have previously stressed). Nevertheless, this example concretely illustrates the benefit of the quantitative model, which allows inference to be extended from the administered dose-levels to the entire dose-range. It is however highly relevant to ask whether the above example is representative of the typical situation, and we will address this question via simulations presented in Section 6.1.

6. Results and discussion

For each of the empirical data sets described in Section 2, our implemented method from Section 5 agrees perfectly with Dunnett's test regarding which dose-levels are active at the 5% significance level. This is encouraging, since we want our method to concur with the authoritative Dunnett's test on (discrete) administered dose-levels. We will use simulation to investigate whether this amount of agreement between the two methods is representative.

6.1. **Simulations.** We shall assess the typical behaviour of the model-based approach (MOD) described in Section 5 by simulating dose–response studies. As a frame of reference, we will compare the inferences drawn from MOD to those made by the multiple comparison procedure (MCP) Dunnett's test [6]. In order to avoid tailoring the results to a specific scenario, all parameters will be drawn from probability distributions. We make extensive use of the (scale-parametrised) gamma distribution $\Gamma(k, \theta)$ whose mean is $k\theta$ (see e.g. Casella and Berger [24]). In each simulated trial, the minimum dose is drawn from $\Gamma(1, 1)$ with mean 1, and the maximum dose is drawn from $\Gamma(11, 13)$ with mean 143. The desired number of dose-levels are then spaced evenly on a logarithmic scale in this dose-range, and a zero-dose vehicle treatment is included. A total of 40 subjects are allotted evenly amongst the groups, and in the case of a remainder, we randomly select the dose-levels which receive an additional subject. Recall that the main goal of this thesis is to investigate how the *experimental design* affects the inferences drawn from MOD compared to those of MCP. Therefore we are interested in the consequences of varying the number of dose-levels in the simulated trials. However, the total number of subjects is always fixed at 40, in order to prevent this quantity from influencing the results. Consequently, as the number of dose-levels increases, the number of individuals at each level must decrease.

In the following we will investigate the inferences drawn from the two methods MOD and MCP. Concretely, this means that we consider their respective judgements of whether a certain dose-level in a given trial is active (i.e. that it shows a significant effect compared to the vehicle treatment). Since the methods are different in nature, we expect them to disagree some of the time. For instance, the dose-level under consideration may be declared active by MOD but inert (not active) by MCP. A natural way of comparing the methods against each other is to investigate which method is more liberal or conservative than the other. By saying, for a certain dose-level in a given trial, that one method is more *liberal* than the other, we simply mean that the former declares the given dose-level to be active whereas the latter does not. By *conservative* we mean the opposite. It is important to note that this does not mean that the former method is making an error. Rather, it is merely a way to describe more clearly the disagreement between the two methods.

Simulations based on the $E_{\max}$ model. In this section we will use the $E_{\max}$ model both to generate dose–response data and to fit a regression curve to the simulated data. The model parameters used for simulation are $E_0 \sim \Gamma(3, 5)$ and $E_{\max} \sim \Gamma(5, 7)$, whilst ED_{50} is drawn uniformly between the minimum and maximum dose. An $E_{\max}$ model with these parameters is employed to generate data, with noise added according to $\mathcal{N}(0, \sigma^2)$ where $\sigma = \frac{1}{4}E_{\max}$. (This choice of standard deviation is based upon the article by Dette et al. (2008) [5].) A typical simulated trial will have a minimum dose of 1, a maximum dose of 143, an E_0 of 15, an $E_{\max}$ of 35, and an ED_{50} of 72. To each simulated trial, an $E_{\max}$ model is fitted to the perturbed dose–response data, and the procedure described in Section 5 is performed. In other words, a 95% simultaneous confidence band is used to determine active doses according to the modelling approach. This judgement is compared against that made by Dunnett's test regarding the discrete "administered" dose-levels.

The above procedure is iterated 1000 times, which has been shown to yield high precision in the results at a reasonable computational expense. In each trial we record which dose-levels are

declared active by MOD and MCP, respectively, and average over the 1000 trials to get an estimate of the proportion of dose-levels deemed active by either method. Since we are interested in how the design affects the results, we run simulations with the total number of dose-levels (including the vehicle group) ranging from 3 through 10. The results are displayed in Figure 11.

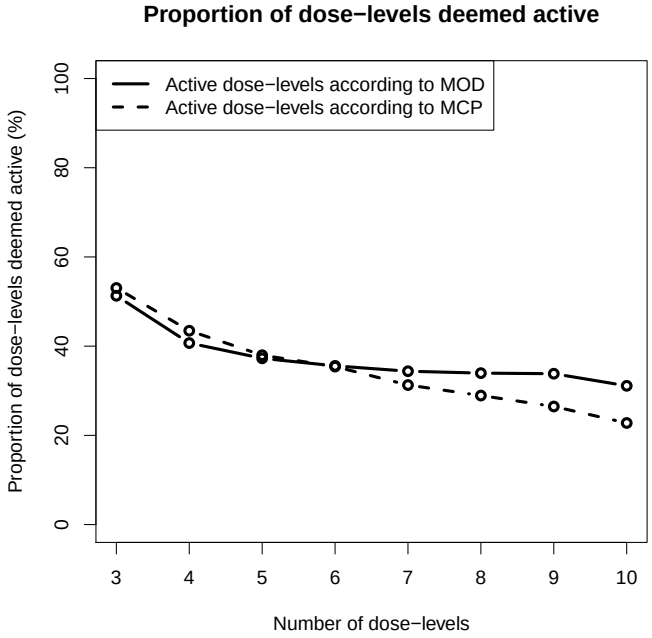

FIGURE 11. Proportion of dose-levels declared active at the 95% significance level by the modelling approach (MOD) and the multiple comparison procedure (MCP). Each estimate is based on 1000 simulated dose–response trials with the indicated number of dose-levels. Since the total number of subjects is fixed (at 40), the number of individuals per dose-level must decrease as the number of dose-levels increases.

We find that MOD tends to be more conservative than MCP in designs with fewer than six dose-levels, but that this relationship is reversed in designs with more than six dose-levels.

Simulations with model misspecification. In order to assess the consequences of misspecifying the underlying model we now simulate trials from a sigmoid $E_{\max}$ model, but fit an ordinary (hyperbolic) $E_{\max}$ model to the data. The parameters for data generation follow the same distributions as before, with the addition of the slope parameter $h \sim \mathcal{N}(0, 3^2)$. This distribution of h has been chosen in order to yield slope parameters which are spread quite far from ± 1, since the models coincide in those cases. Figure 12 shows the results of 1000 such simulated trials. Note that MCP does not include any assumption of the underlying dose–response model, wherefore it is unaffected by its misspecification. Thus the dashed line pertaining to MCP can be used as a reference for comparison with simulations under a valid model assumption, shown in Figure 11 above.

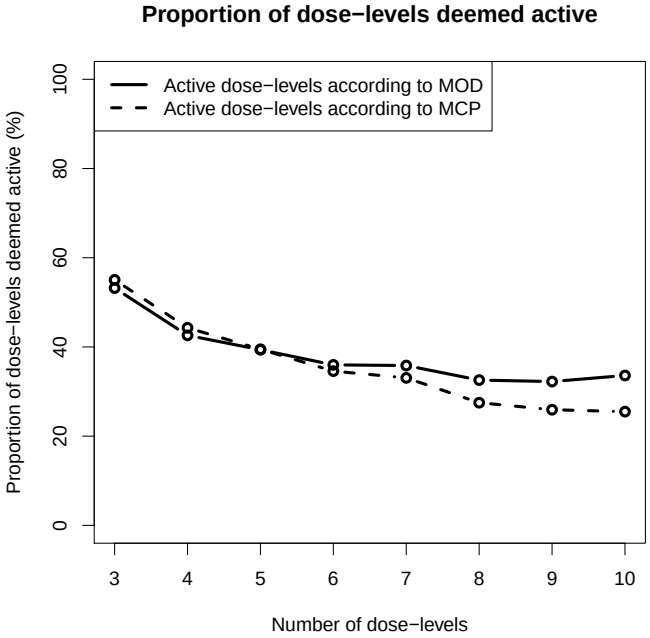

FIGURE 12. Proportion of dose-levels declared active at the 95% significance level by MCP and MOD under model misspecification. Each estimate is based on 1000 simulated dose–response trials with the indicated number of dose-levels. Since the total number of subjects is fixed (at 40), the number of individuals per dose-level must decrease as the number of dose-levels increases.

Simulating optimal designs. Consider that MOD is most liberal when its confidence band is as narrow as possible (since such a band is most prone not to cover the zero response). This occurs when the parameter estimates are as precise as possible, which they are in a D-optimal design. In order to connect optimal design to comparisons between MOD and MCP, we will now investigate the consequences of optimally designing the simulated trials. Thus we modify the procedure above by allocating dose-levels and their sample sizes in a D-optimal fashion. We shall let our model assumption be valid in these simulations, and use an $E_{\max}$ model both for data generation and regression. More concretely, we use the function `optDesign` of the `R` package `DoseFinding` [8] to choose a D-optimal design for the $E_{\max}$ model. However, this function requires that we supply a collection of candidate dose-levels, amongst which `optDesign` determines the most efficient allocation. Hence `optDesign` will never suggest a dose-level not included amongst the candidates. Rather, it will return the optimal allocation of observations across the candidate dose-levels, some of which may well be assigned a zero weight, thereby excluding them from the design. As we shall see momentarily, this will generally lead to a discrepancy between the nominal and the actual number of dose-levels of an optimal design. The nominal number is simply the number of candidate dose-levels, whereas the actual number is the number of dose-levels selected by `optDesign`.

For each (nominal) number of dose-levels $(3, 4, \ldots, 10)$ we simulate 1000 dose–response data sets from an $E_{\max}$ model as in the previous section. In each such trial, we let the candidate dose-levels be spread out evenly on a logarithmic scale. The actual dose-levels, chosen amongst

these candidates, are then used when simulating a dose–response trial. The verdicts of MOD and MCP, respectively, are displayed in Figure 13.

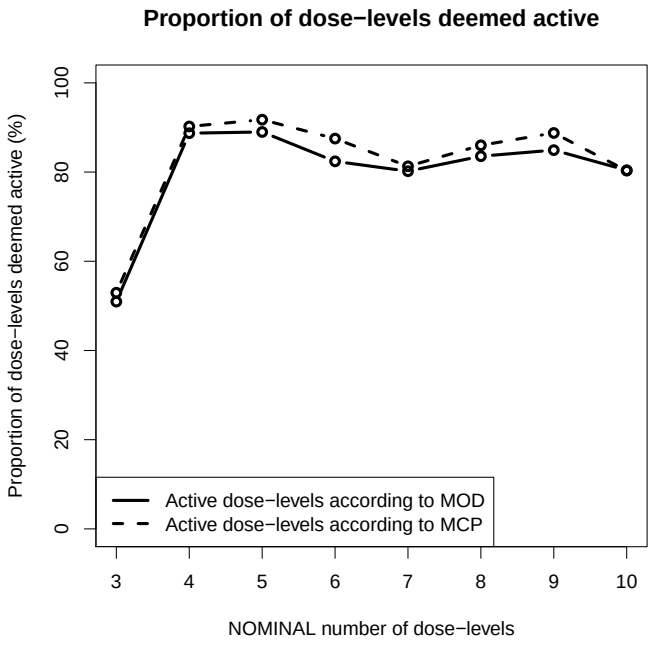

FIGURE 13. Proportion of dose-levels declared active at the 95% significance level by MOD and MCP. Each estimate is based on 1000 simulated dose–response trials which have been D-optimally designed with a valid $E_{\max}$ model assumption. It is important to note that the number of dose-levels are *nominal* — not actual. In fact, the true number of dose-levels is almost always 3.

See Section 7 for our conclusions about all the above results.

6.2. **White book.** In this section we shall describe how to use the methods in the `R` [7] package `DoseFinding` [8] for design and analysis of dose–response experiments. This software supports the D-criterion of optimality, which maximises the precision of the estimated model parameters. It also offers support to make the design optimal for estimating the minimum effective dose (MED) defined as the smallest dose which yields a clinically relevant and statistically significant effect [4]. A design which is optimal in the latter sense is said to be MED-optimal. It is important to note that the clinically relevant effect Δ must be specified by an expert before the statistical analysis is performed. For instance, an experimenter might only be interested in whether a compound can increase the response by at least $\Delta = 0.4$ units over the vehicle response — regardless of whether a smaller effect is found to be statistically significant. It is of course possible that no clinically relevant effect exists — and also that it does but without being detectable due to high variability in the measurements or insufficient sample size [1]. However, it is only the MED-optimal design which actually requires Δ to be specified. The D-optimal design operates without it, since its only objective is to minimise the variance of the estimator of the model parameters.

Algorithm 2 (Robust optimal design of a dose–response experiment). The following functions are contained in the `R` package `DoseFinding` [8] whose help pages offer more detailed information.

(1) Nominate the candidate models which are conjectured to capture the true dose–response relationship. Only include models which are likely to be relevant, and rank each model with a probability corresponding to your belief that it most aptly describes the underlying relationship. These probabilities must sum to unity, and in the absence of prior knowledge they may be set to $1/m$, where m is the number of candidate models. Notice that the parameters of the nonlinear part of each model must be (numerically) specified.
(2) Suggest a collection of feasible dose-levels. The recommended dose-levels will be chosen amongst this set, though allocated in an optimal way.
(3) Use `optDesign` to compute the optimal dose-levels and their corresponding weights. Use `rndDesign` to efficiently round the optimal design to attain an integer number of measurements (possibly zero) to be made at each dose-level.

Algorithm 3 (Analysis of a dose–response experiment). The following functions are contained in the `R` package `DoseFinding` [8] whose help pages offer more detailed information.

(1) Fit each candidate model (from Algorithm 2) to the empirical data using `fitMod`.
(2) Use `MCPMod` to determine which model (fitted in the previous step) most aptly describes the empirical data. If there are few active dose-levels, or if the models are highly correlated (and thus difficult to distinguish), `MCPMod` may fail to identify the most appropriate model. If so, model selection can for instance be based on Akaike's information criterion. Use the winning model for all further analysis and inference.
(3) Use the method in Section 5 to construct a simultaneous confidence band around the vehicle-adjusted regression curve of the winning model. A dose is declared to yield a significant effect compared to vehicle if and only if the confidence band around its fitted response does not cover zero.

7. Conclusions

In this thesis we implement a method for assessing efficacy in a model-based setting (see Section 5). This method is based on constructing a confidence band around a vehicle-adjusted regression curve fitted to dose–response data. We declare a dose to be active if and only if the confidence band at its fitted value does not cover zero. A dose-level is said to be *active* if it shows a statistically significant effect compared to the vehicle treatment (at the given significance level, which in this thesis is set to 5%). In order to evaluate the inferential properties of this method, we simulate dose–response trials in which its conclusions are compared to those of ANOVA coupled with Dunnett's multiple comparison procedure. We consider the latter to be an authoritative benchmark regarding which of the (discrete) administered dose-levels are active in a given trial. The main advantage of the model-based procedure is its capacity to declare activity (on a continuous scale) *in between* administered dose-levels. We find that to a large extent, the two methods draw similar inferences concerning administered dose-levels. There are however systematic discrepancies, which may be elucidated by considering the nature of either method.

When comparing our model-based implementation (MOD) to ANOVA coupled with Dunnett's multiple comparison procedure, we find that the former tends to be more conservative relative to the latter in designs with fewer than six logarithmically spaced dose-levels, but that the converse is true in designs with more than six logarithmically spaced dose-levels. These results are displayed in Figure 11. This makes sense intuitively, because a design with few dose-levels and many individuals per level benefits Dunnett's test, since there are few comparisons for which to adjust and many subjects per group to gain high precision. However, it is not ideal for the regression and confidence band in MOD, which rather benefit from a larger number of dose-levels spread out across the dose-range. This is more advantageous for curve fitting, giving more precision in the estimated parameters. In turn, this yields narrower confidence bands, which are more likely not to contain zero. It should be noted that both methods are punished as the number of dose-levels increases. In the case of Dunnett's test, this is because the procedure has to adjust for more comparisons with fewer individuals per dose-level, which yields less statistical confidence at each level, thereby raising the bar for detecting a signal. In the case of MOD it is because the estimate of the mean vehicle response becomes more variable as the vehicle group is reduced, which yields wider, more conservative, confidence bands.

In order to investigate the effect of misspecifying the underlying dose–response model, we simulate trials in which we fit an ordinary $E_{\max}$ model to data generated by a sigmoid $E_{\max}$ model. The results are displayed in Figure 12. We see no systematic effect of this model misspecification on the proportion of dose-levels deemed active by the model-based procedure.

Now consider that MOD is most liberal when its confidence band is as narrow as possible (since such a band is most prone not to cover the zero response). This occurs when the parameter estimates are as precise as possible, which they are in a D-optimal design. Increasing the number of dose-levels cannot yield narrower confidence bands than those of a D-optimal design. When optimally designing the simulated trials with a valid $E_{\max}$ model assumption, both methods are dramatically more prone to declare dose-levels active. These results are displayed in Figure 13. In the case of MOD, the leap in declarations of activity between three and four nominal dose-levels may at its root be caused by the fact that a larger nominal number of dose-levels will yield a finer, more densely spaced set of candidate dose-levels. As a result, the design algorithm will be able to choose dose-levels closer to those of the true D-optimal design, which will yield higher precision in the parameter estimates. This in turn will lead to narrower confidence bands, which, finally, yield more liberal judgements of activity. However, we witness a virtually identical leap in declarations of activity by Dunnett's test, which may well benefit from a more carefully chosen design, but on which the remaining chain of events has no bearing.

At first glance, it may also appear that optimally designing the simulated trials eliminates the previously observed trend that declarations of activity decrease as the nominal number of dose-levels increases. This is however an illusion, caused by the fact that the design algorithm generally selects only three dose-levels, irrespective of the nominal number of candidate dose-levels with which it is provided. As mentioned in Section 4.1, this can be shown [20] to be the correct number of dose-levels for any D-optimal design under the $E_{\max}$ model. A natural continuation of these investigations would be to combine model misspecification with optimal design. Moreover, this offers an opportunity to assess model-robustness, namely by including several candidate models and utilising the model-robust optimal designs described in Section 6.2 of this thesis.

Future work. Our model-based method (Section 5) is designed to detect doses which show a significant effect compared to the vehicle treatment. However, as mentioned in Section 3, it may also be of interest to compare the compound under investigation against a reference treatment (e.g. a competing compound). Dette et al. (2014) [14] have developed optimal designs for such studies, and their work may serve as a source of inspiration for ideas of how to appropriate our method to accommodate this scenario.

When implementing our method in Section 5 we made the assumption that the fitted values of the regression curve and the mean vehicle response are uncorrelated. This may not be valid, since the vehicle responses are used during regression. Investigating this correlation would perhaps yield more faithful results. However, the pointwise confidence bands generated by the bootstrap agree quite well with those constructed under this assumption, which indicates that it may be of subordinate concern. On a related note, when estimating the variance of the mean vehicle response, we do not use the pooled sample variance across all dose-levels. This may be advisable in a heteroscedastic case, but in a homoscedastic case, doing so would likely yield a better estimate.

In this thesis we have only mentioned Dunnett's test as a benchmark against which to compare our modelling approach. This is because Dunnett's test is designed to compare the response of each dose-level to that of the vehicle treatment, which is the qualitative analogue of what we do when checking whether the confidence band contains the zero response. There are of course other multiple comparison procedures which may be appropriate, and we have for instance experimented with Williams' trend test. This is designed to test for a monotonic dose–response profile, and it involves borrowing strength from neighbouring dose-levels in order to do so. It could be argued that this is a more appropriate analogue to our modelling approach, since our candidate regression models are (all but one) monotonic. However, we do not use our modelling procedure to test for the existence of a trend. For instance, even though the regression curve is constricted to be monotonic, its confidence band is not. If the parameter estimates are highly variable, a dose may be inert even though it is higher than the minimum active dose. In the case of an increasing dose–response profile, this occurs when the lower limit of the confidence band returns below the zero response after first having exceeded it. As mentioned above, this is because we are investigating individual dose-levels rather than the existence of a trend, which is the reason for our preference of Dunnett's test. Nevertheless, it would be a natural continuation of our work to extend the model-based procedure to test for the existence of a monotonic dose–response trend. If for instance the confidence band around the smallest (non-vehicle) dose does not overlap with that around the largest dose, then we would be confident that a dose–response trend exists. However, this is almost certainly an excessively cautious criterion, but it is not obvious how to moderate it whilst still retaining confidence in the result. Doing so may well present a stimulating challenge.

References

[1] F. Bretz, J. C. Pinheiro, and M. Branson, "Combining multiple comparisons and modeling techniques in dose–response studies," *Biometrics*, vol. 61, no. 3, pp. 738–748, 2005.

[2] J. Pinheiro, B. Bornkamp, and F. Bretz, "Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures," *Journal of Biopharmaceutical Statistics*, vol. 16, no. 5, pp. 639–656, 2006.

[3] F. Bretz, H. Dette, and J. C. Pinheiro, "Practical considerations for optimal designs in clinical dose finding studies," *Statistics in Medicine*, vol. 29, no. 7-8, pp. 731–742, 2010.

[4] S. J. Ruberg, "Dose response studies I. Some design considerations," *Journal of Biopharmaceutical Statistics*, vol. 5, no. 1, pp. 1–14, 1995.

[5] H. Dette, F. Bretz, A. Pepelyshev, and J. Pinheiro, "Optimal designs for dose-finding studies," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1225–1237, 2008.

[6] F. Bretz, T. Hothorn, and P. Westfall, *Multiple comparisons using R*. CRC Press, 2010.

[7] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[8] B. Bornkamp, J. Pinheiro, and F. Bretz, *DoseFinding: planning and analyzing dose finding experiments*, 2014. R package version 0.9-12.

[9] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008.

[10] J. Macdougall, "Analysis of dose–response studies—$E_{\max}$ model," in *Dose Finding in Drug Development* (N. Ting, ed.), Statistics for Biology and Health, pp. 127–145, Springer New York, 2006.

[11] J. Gabrielsson and D. Weiner, *Pharmacokinetic & Pharmacodynamic Data Analysis: Concepts and Applications*. Stockholm: Swedish Academy of Pharmaceutical Sciences, 4th ed., 2006.

[12] V. V. Fedorov and S. L. Leonov, *Optimal Design for Nonlinear Response Models*. Chapman & Hall/CRC Biostatistics Series, Taylor & Francis, 2013.

[13] B. Bornkamp, J. Pinheiro, and F. Bretz, "MCPMod: An R package for the design and analysis of dose-finding studies," *Journal of Statistical Software*, vol. 29, no. 7, pp. 1–23, 2009.

[14] H. Dette, C. Kiss, N. Benda, and F. Bretz, "Optimal designs for dose finding studies with an active control," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 265–295, 2014.

[15] A. C. Atkinson, A. N. Donev, and R. Tobias, *Optimum Experimental Designs, with SAS*. Oxford statistical science series, Oxford University Press, 2007.

[16] V. Fedorov, "Optimal experimental design," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 581–589, 2010.

[17] H. Chernoff, "Locally optimal designs for estimating parameters," *The Annals of Mathematical Statistics*, pp. 586–602, 1953.

[18] A. Pázman, "Computation of the optimum designs under singular information matrices," *The Annals of Statistics*, vol. 6, no. 2, pp. 465–467, 1978.

[19] S. D. Silvey, "Optimal design measures with singular information matrices," *Biometrika*, vol. 65, no. 3, pp. 553–559, 1978.

[20] H. Dette, C. Kiss, M. Bevanda, and F. Bretz, "Optimal designs for the $E_{\max}$, log-linear and exponential models," *Biometrika*, vol. 97, no. 2, pp. 513–518, 2010.

[21] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Applied Regression Analysis: A Research Tool*. Springer Texts in Statistics, Springer, 2nd ed., 1998.

[22] S. Gsteiger, F. Bretz, and W. Liu, "Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses," *Journal of Biopharmaceutical Statistics*, vol. 21, no. 4, pp. 708–725, 2011.

[23] P. Hall and J. Horowitz, "A simple bootstrap method for constructing nonparametric confidence bands for functions," *The Annals of Statistics*, vol. 41, no. 4, pp. 1892–1921, 2013.

[24] G. Casella and R. L. Berger, *Statistical Inference*. Pacific Grove, California: Duxbury, 2nd ed., 2002.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Stanford, California: Springer, 2nd ed., 2009.