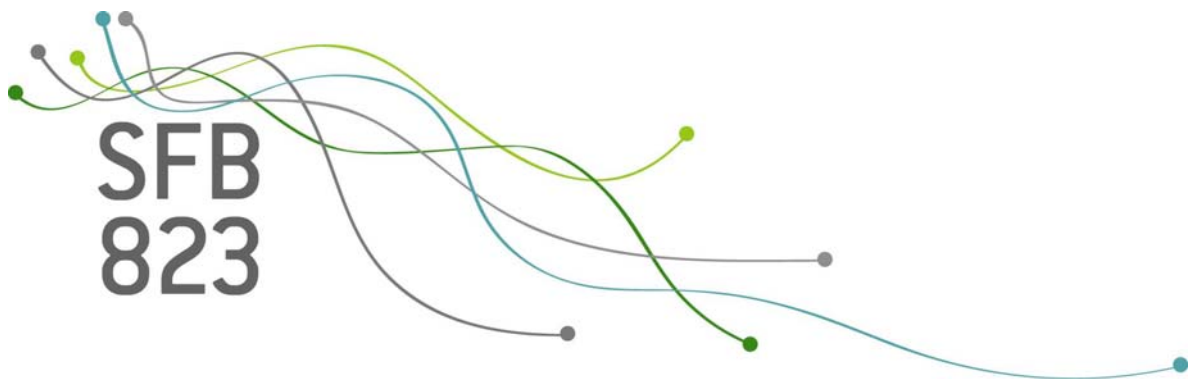# Equivalence tests for binary efficacy-toxicity responses

Kathrin Möllenhoff, Holger Dette,
Frank Bretz

Equivalence tests for binary efficacy-toxicity responses

Kathrin Möllenhoff[1], Holger Dette[1] , Frank Bretz[2]

[1] Department of Mathematics, Ruhr-Universität Bochum, Germany,

[2]Novartis Pharma AG, CH-4002 Basel, Switzerland,

October 19, 2019

Abstract

Clinical trials often aim to compare a new drug with a reference treatment in terms of efficacy and/or toxicity depending on covariates such as, for example, the dose level of the drug. Equivalence of these treatments can be claimed if the difference in average outcome is below a certain threshold over the covariate range. In this paper we assume that the efficacy and toxicity of the treatments are measured as binary outcome variables and we address two problems. First, we develop a new test procedure for the assessment of equivalence of two treatments over the entire covariate range for a single binary endpoint. Our approach is based on a parametric bootstrap, which generates data under the constraint that the distance between the curves is equal to the pre-specified equivalence threshold. Second, we address equivalence for bivariate binary (correlated) outcomes by extending the previous approach for a univariate response. For this purpose we use a 2-dimensional Gumbel model for binary efficacy-toxicity responses. We investigate the operating characteristics of the proposed approaches by means of a simulation study and present a case study as an illustration.

Keywords and Phrases: binary data, dose response, logistic regression, Gumbel model, bootstrap

1 Introduction

Equivalence tests are used in clinical drug development to assess similarity of a test treatment with a reference treatment. Considering continuous data with covariates, the effect of a drug

is described as a function of the covariates and equivalence is claimed is these function are in some sense similar. Several authors use confidence bands for the difference between the response functions to construct equivalence tests (see, for example Liu et al. (2009); Gsteiger et al. (2011); Bretz et al. (2018)). Alternatively, Dette et al. (2018); Möllenhoff et al. (2018) proposed more powerful tests by estimating a distance between the two functions, such as the the squared integral of the difference or the maximal deviation between the curves. They claim equivalence if the estimated distance is small. All these approaches assume continuous outcomes. However, in some situations drug efficacy is measured using a binary outcome (for some examples see Chow and Liu (1992); Cox (2018)). For example, a patient is considered to be a responder, that is the efficacy response is 1, if the drug effect is as desired. This can be for example the shrinkage of a tumor or the curing of any disease. Equivalence tests have been proposed in these settings by, for example, Nam (1997) and Chen et al. (2000), who derive methodology for comparing the treatments in response probabilities. These authors investigate different types of test statistics and perform sample size determination in several situations but they do not include any covariates such as, for example, the dose.

Many clinical trials involve the measurement of a second endpoint (e.g. to assess toxicity) and hence bivariate outcomes are considered which are likely to be correlated, see for example Murtaugh and Fisher (1990); Heise and Myers (1996); Thall and Cook (2004); Dragalin and Fedorov (2006) and Gaydos et al. (2006). This is, for instance, the case when observing efficacy and toxicity of a drug. The toxicity response is 1 if a side-effect (e.g. fatigue or nausea) is observed. Several methods for modelling multivariate binary outcomes have been proposed, see for example, Glonek and McCullagh (1995). Considering efficacy-toxicity responses, Murtaugh and Fisher (1990) and Heise and Myers (1996) investigate bivariate binary responses and derive optimal designs for this situation by fitting the data to a bivariate logistic model and a Cox model (see also Cox (2018)). Deldossi et al. (2019) propose Copula functions to model these types of outcomes. Further, Thall and Cook (2004) and Dragalin and Fedorov (2006) develop adaptive designs for identifying the optimal safe dose. Finally several authors investigate the modeling and design of phase I/II dose-finding trials incorporating bivariate outcomes using Bayesian methods, see for instance Nebiyou Bekele and Shen (2005); Zhang et al. (2006); Yin et al. (2006).

Different to the literature reviewed above, we investigate equivalence tests with The purpose to assess similarity of a reference and a test treatment for efficacy and toxicity. Equivalence can only be claimed if the differences of *both* outcomes are below prespecified thresholds over the complete range of covariates. Accordingly, we first develop a new test for assessing equivalence in case of a single binary endpoint over the range of covariates. Second, we address equivalence for bivariate binary (correlated) outcomes and develop an equivalence test for comparing simultaneously efficacy and toxicity of a reference to a test treatment. For this purpose we use

a 2-dimensional Gumbel model (see Gumbel (1961)) for bivariate logistic regression to model correlated bivariate binary endpoints. Our approach is based on a non-standard parametric bootstrap, which generates data under the constraint that the distances between the curves are precisely equal to the thresholds. We investigate finite sample properties and illustrate the procedures with a clinical trial example.

2 Comparing curves for binary outcomes

In this section we introduce a model-based approach for the investigation of equivalence between the efficacy of two treatments assuming binary endpoints. We consider models with covariates and assume for simplicity a one-dimensional covariate, although the proposed methodology applies more broadly. For both treatment groups we choose the covariate space as a (log-transformed) dose range $\mathcal{D}$ and assume that treatments are conducted at k_ℓ dose levels $d_{\ell,1}, \ldots, d_{\ell,k_\ell}$, $\ell = 1, 2$, where the index $\ell = 1$ corresponds to the reference and $\ell = 2$ to the test treatment. At dose level $d_{\ell,i}$ we observe $n_{\ell,i}$ patients, $i = 1, \ldots, k_\ell$. Let $Y_{\ell,i,j}$ denote the binary outcome for the jth patient allocated to the ith dose level receiving treatment ℓ. If a patient responds to the drug, we have $Y_{\ell,i,j} = 1$, otherwise $Y_{\ell,i,j} = 0$. More precisely, $Y_{\ell,i,j}$ follows a Bernoulli distribution with parameter $p_\ell(d_{\ell,i})$ modelling the probability of success under treatment ℓ with dose level $d_{\ell,i}$, $i = 1, \ldots k_\ell$, $\ell = 1, 2$. We use regression techniques to model the dose-response relationship. More precisely, the probability of the jth patient allocated to dose level $d_{\ell,i}$ responding to treatment ℓ is given by

$$p_\ell(d_{\ell,i}) = \mathbb{P}(Y_{\ell,i,j} = 1 \mid d_{\ell,i}) = \eta_\ell^E(d_{\ell,i}, \beta_\ell, \gamma_\ell), \ \ell = 1, 2, \tag{2.1}$$

where η_ℓ^E is a known distribution function determined by parameters β_ℓ, γ_ℓ. Hence the curve $\eta_\ell^E(d, \beta_\ell, \gamma_\ell)$ models the probability of efficacy over the entire dose range.

Common examples of (2.1) include the logistic regression model $\mathbb{P}(Y_{\ell,i,j} = 1 \mid d_{\ell,i}) = \frac{1}{1+e^{-\beta_\ell - \gamma_\ell \cdot d_{\ell,i}}}$ and the probit regression model $\mathbb{P}(Y_{\ell,i,j} = 1 \mid d_{\ell,i}) = \Phi(\beta_\ell + \gamma_\ell \cdot d_{\ell,i})$, $\ell = 1, 2$, where Φ is the distribution function of the standard normal distribution (see for example Long and Freese (2006)). Assuming independent observations, the likelihood of the observed data in treatment group $\ell = 1, 2$ is

$$\begin{aligned}
\mathcal{L}_\ell(\beta_\ell, \gamma_\ell | y_{\ell,1,1}, \ldots, y_{\ell,k_\ell,1}, \ldots, y_{\ell,k_\ell,n_{\ell,k_\ell}}) &= \prod_{i=1}^{k_\ell} \prod_{j=1}^{n_{\ell,i}} p_\ell(d_{\ell,i})^{y_{\ell,i,j}} (1 - p_\ell(d_{\ell,i}))^{(1-y_{\ell,i,j})} \\
&= \prod_{i=1}^{k_\ell} p_\ell(d_{\ell,i})^{\zeta_{\ell,i}} (1 - p_\ell(d_{\ell,i}))^{n_{\ell,i} - \zeta_{\ell,i}},
\end{aligned}$$

where $\zeta_{\ell,i} := \sum_{j=1}^{n_{\ell,i}} y_{\ell,i,j}$, $i = 1, \ldots, k_\ell$, $\ell = 1, 2$. Taking the logarithm yields

$$
\begin{aligned}
l_\ell(\beta_\ell, \gamma_\ell) &:= \log \mathcal{L}(\beta_\ell, \gamma_\ell | y_{1,1}, \ldots, y_{k,1}, \ldots, y_{k,n_k}) \\
&= \sum_{i=1}^{k_\ell} \zeta_{\ell,i} \log p_\ell(d_{\ell,i}) + (n_{\ell,i} - \zeta_{\ell,i}) \log\left(1 - p_\ell(d_{\ell,i})\right)
\end{aligned}
\tag{2.2}
$$

and corresponding Maximum-Likelihood-estimates (MLE) are obtained by maximizing the function (2.2).

In order to investigate the difference in efficacy between the reference and the test treatment we consider the maximal deviation between the two curves in (2.1) and define the equivalence hypotheses by

$$
H_0^E : \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \beta_1, \gamma_1) - \eta_2^E(d, \beta_2, \gamma_2) \right| \geq \varepsilon^E \text{ vs. } H_1^E : \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \beta_1, \gamma_1) - \eta_2^E(d, \beta_2, \gamma_2) \right| < \varepsilon^E,
\tag{2.3}
$$

where ε^E denotes a prespecified margin of equivalence in efficacy between the two curves, which has to be carefully chosen by clinicians in advance. The choice of these thresholds is a common issue and there are no general recommendations available. However, according to guidelines (U.S. Food and Drug Administration (2003)) equivalence margins between 0.1 and 0.2, that is a deviation of the two products between 10% to 20%, seem to be reasonable.

The following algorithm provides a bootstrap test for the hypotheses (2.3), which keeps its nominal level, say α, and is consistent. It is derived by adapting the methodology developed in Dette et al. (2018) to binary data.

Algorithm 2.1. (parametric bootstrap for testing equivalence of binary outcomes)

(1) Calculate the MLE $(\hat{\beta}_\ell, \hat{\gamma}_\ell)$, $\ell = 1, 2$, by maximizing for each group the log-likelihood given in (2.2). The test statistic is obtained by

$$
\hat{d}^E := \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \hat{\beta}_1, \hat{\gamma}_1) - \eta_2^E(d, \hat{\beta}_2, \hat{\gamma}_2) \right|.
$$

(2) Define estimators of the parameters β_ℓ, γ_ℓ, $\ell = 1, 2$, so that the corresponding curves fulfill the null hypothesis (2.3), that is

$$
\left(\hat{\bar{\beta}}_\ell, \hat{\bar{\gamma}}_\ell \right) =
\begin{cases}
(\hat{\beta}_\ell, \hat{\gamma}_\ell) & \text{if } \hat{d}^E \geq \varepsilon \\
(\bar{\beta}_\ell, \bar{\gamma}_\ell) & \text{if } \hat{d}^E < \varepsilon
\end{cases}
\quad \ell = 1, 2,
$$

where $(\bar{\beta}_1, \bar{\gamma}_1)$ and $(\bar{\beta}_2, \bar{\gamma}_2)$ minimize the same objective function as defined in (2.2), but under the constraint

$$
d^E = \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \beta_1, \gamma_1) - \eta_2^E(d, \beta_2, \gamma_2) \right| = \varepsilon^E.
\tag{2.4}
$$

4

We discretize the dose range $\mathcal{D}$ to get a feasible optimization problem by fixing r nodes $d_1, \ldots, d_r$ and use the smooth approximation (as $\lambda \to 0$)

$$\max_{i=1}^{r} d_i \approx \lambda \log \sum_{i=1}^{r} \exp \frac{d_i}{\lambda}$$

for the calculation of the maximum in (2.4). Finally the optimization procedure is performed by running the auglag() function implemented in the R package alabama by Varadhan (2014). The algorithm implemented in this function is based on the augmented Lagrangian minimization algorithm, which is typically used for solving constrained optimization problems.

(3) Proceed as follows:

(i) Generate bootstrap data under the null hypothesis (2.3), that is, create binary data specified by the parameters $(\hat{\hat{\beta}}_\ell, \hat{\hat{\gamma}}_\ell)$, $\ell = 1, 2$. More precisely, calculate $\eta_\ell^E(d_{\ell,i}, \hat{\hat{\beta}}_\ell, \hat{\hat{\gamma}}_\ell)$, $i = 1, \ldots, k_\ell$, $\ell = 1, 2$ yielding the probabilities of success $p(d_{\ell,i})$ at each dose level $d_{\ell,i}$.

(ii) From the bootstrap data calculate the MLE $(\hat{\beta}_\ell^*, \hat{\gamma}_\ell^*)$ as in step (1) and the test statistic

$$\hat{d}^{E*} = \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \hat{\beta}_1^*, \hat{\gamma}_1^*) - \eta_2^E(d, \hat{\beta}_2^*, \hat{\gamma}_2^*). \right| \tag{2.5}$$

(iii) Repeat the steps (i) and (ii) n_{boot} times to generate replicates $\hat{d}_{\infty,1}^{E*}, \ldots, \hat{d}_{\infty,n_{boot}}^{E*}$ of $\hat{d}^{E*}$. Let $\hat{d}^{E*(1)} \le \ldots \le \hat{d}^{E*(n_{boot})}$ denote the corresponding order statistic. The estimator of the α-quantile of the distribution of $\hat{d}^*$ is defined by $\hat{d}^{E*(\lfloor n_{boot}\alpha \rfloor)}$. Reject the null hypothesis (2.3), if
$$\hat{d}^E < \hat{d}^{E*(\lfloor n_{boot}\alpha \rfloor)}. \tag{2.6}$$

Further we obtain the p-value by $\hat{F}_{n_{boot}}(\hat{d}^E)$, where $\hat{F}_{n_{boot}}(x) = \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} I\{\hat{d}_{\infty,i}^{E*} \le x\}$ denotes the empirical distribution function of the bootstrap sample.

Note that the bootstrap quantile $\hat{d}^{E*(\lfloor n_{boot}\alpha \rfloor)}$ depends on the number of bootstrap replicates n_{boot} and the threshold ε^E given in the hypotheses (2.3), but we do not reflect this dependence in our notation. The test proposed in Algorithm 2.1 has asymptotic level α and is consistent. More precisely, note that $\hat{d}^{E*(\lfloor n_{boot}\alpha \rfloor)} \to \hat{q}_\alpha$ as $n_{boot} \to \infty$, where $\hat{q}_\alpha$ denotes the α-quantile of the distribution of the statistic (2.5). It can then be shown that under H_0^E

$$\limsup_{n_1, n_2 \to \infty} \mathbb{P}_{H_0^E}(\hat{d}^E < \hat{q}_\alpha) \le \alpha \tag{2.7}$$

and that under H_1^E

$$\lim_{n_1,n_2\to\infty} \mathbb{P}_{H_1^E}(\hat{d}^E < \hat{q}_\alpha) = 1. \tag{2.8}$$

These results follow from the well-known fact that under suitable conditions of regularity the MLE converge weakly to a normal distribution (see Bradley and Gart (1962)), that is

$$\sqrt{n_\ell}\left((\hat{\beta}_\ell, \hat{\gamma}_\ell) - (\beta_\ell, \gamma_\ell)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_\ell^{-1}), \ \ell = 1, 2, \tag{2.9}$$

where the asymptotic variance-covariance matrix I_ℓ is the Fisher Information Matrix corresponding to treatment group ℓ. The weak convergence (2.9) is the essential ingredient to apply the proof of Dette et al. (2018) to the situation considered in this paper and (2.7) and (2.8) follow.

3 Equivalence tests for efficacy-toxicity responses

3.1 The Gumbel model for efficacy-toxicity outcomes

In this section we extend the approach of Section 2 to equivalence tests for correlated bivariate binary outcomes. We consider the bivariate Gumbel model (see for example Murtaugh and Fisher (1990); Heise and Myers (1996)) based on the bivariate logistic function derived by Gumbel (1961), which is given by

$$F_{U,V}(u, v) = \frac{1}{1 + e^{-u}} \frac{1}{1 + e^{-v}} \cdot \left(1 + \frac{\nu e^{-u-v}}{(1 + e^{-u})(1 + e^{-v})}\right). \tag{3.1}$$

Note that the marginal distributions are logistic and that the parameter $\nu \in (-1, 1)$ represents the dependence of U and V. In particular the case $\nu = 0$ corresponds to independent margins and in this case two separate logistic models for efficacy and toxicity can be fitted separately to the data.

We make the same assumptions as in the univariate case and further let $Y = (Y^E, Y^T) \in \{0, 1\}^2$ denote the bivariate outcome for a patient allocated to the dose level d, where Y^E denotes the efficacy and Y^T the toxicity response. We follow Murtaugh and Fisher (1990) and formulate the model by deriving the four cell probabilities

$$p_{00}(d) := \mathbb{P}(Y^E = 0, Y^T = 0| d) = 1 - \frac{1}{1+e^{-u_1(d)}} - \frac{1}{1+e^{-u_2(d)}} + \frac{1}{1+e^{-u_1(d)}}\frac{1}{1+e^{-u_2(d)}}$$
$$+ \frac{\nu e^{-u_1(d)-u_2(d)}}{(1+e^{-u_1(d)})^2(1+e^{-u_2(d)})^2},$$

$$p_{01}(d) := \mathbb{P}(Y^E = 0, Y^T = 1| d) = \frac{1}{1+e^{-u_2(d)}} - \frac{1}{1+e^{-u_1(d)}}\frac{1}{1+e^{-u_2(d)}} - \frac{\nu e^{-u_1(d)-u_2(d)}}{(1+e^{-u_1(d)})^2(1+e^{-u_2(d)})^2},$$

$$p_{10}(d) := \mathbb{P}(Y^E = 1, Y^T = 0| d) = \frac{1}{1+e^{-u_1(d)}} - \frac{1}{1+e^{-u_1(d)}}\frac{1}{1+e^{-u_2(d)}} - \frac{\nu e^{-u_1(d)-u_2(d)}}{(1+e^{-u_1(d)})^2(1+e^{-u_2(d)})^2},$$

$$p_{11}(d) := \mathbb{P}(Y^E = 1, Y^T = 1| d) = \frac{1}{1+e^{-u_1(d)}}\frac{1}{1+e^{-u_2(d)}} + \frac{\nu e^{-u_1(d)-u_2(d)}}{(1+e^{-u_1(d)})^2(1+e^{-u_2(d)})^2}. \tag{3.2}$$

6

Here, $u_1(d) = \beta_1 + \gamma_1 \cdot d$ and $u_2(d) = \beta_2 + \gamma_2 \cdot d$ denote the transformed doses for efficacy and toxicity, respectively (see Heise and Myers (1996)). Consequently, the Gumbel model is determined by the 5-dimensional parameter $\theta := (\beta_1, \gamma_1, \beta_2, \gamma_2, \nu) \in \mathbb{R}^5$. The individual curves for efficacy and toxicity are obtained by the marginal probabilities

$$\eta^E(d, \theta) := \mathbb{P}(Y^E = 1 | d) = p_{11}(d) + p_{10}(d) = \frac{1}{1 + e^{-u_1(d)}},$$

$$\eta^T(d, \theta) := \mathbb{P}(Y^T = 1 | d) = p_{11}(d) + p_{01}(d) = \frac{1}{1 + e^{-u_2(d)}}. \tag{3.3}$$

Note that for simplicity we do not display the dependence on θ in the cell probability functions (3.2). We further denote by $\eta(d, \theta) := \left(\eta^E(d, \theta), \eta^T(d, \theta)\right)$ the vector of bivariate response probabilities at dose d. Note that the correlation parameter ν is part of the model but not displayed explicitly. We also note that the restrictions on ν depend on the other model parameters $\beta_1, \gamma_1, \beta_2, \gamma_2$ such that all cell probabilities in (3.2) vary between 0 and 1 for all doses $d \in \mathcal{D}$. Because the correlation of Y^E and Y^T is given by

$$corr(Y^E, Y^T | d) = \frac{\nu}{(e^{u_1(d)/2} + e^{-u_1(d)/2})(e^{u_2(d)/2} + e^{-u_2(d)/2})} \tag{3.4}$$

the upper bound of ν is at most 4.

For the estimation of the model parameters we use again MLE. Therefore the likelihood for one observation $y = (y^E, y^T) \in \{0, 1\}^2$ modelled by the Gumbel model is given by

$$\mathcal{L}(\theta | y) = p_{11}(d)^{y^E y^T} p_{01}(d)^{(1-y^E)y^T} p_{10}(d)^{y^E(1-y^T)} p_{00}(d)^{(1-y^E)(1-y^T)}. \tag{3.5}$$

3.2   The test procedure

Now assume that we have two groups with bivariate (efficacy/ toxicity) outcomes corresponding to the new ($\ell = 1$) and reference ($\ell = 2$) treatment and we want to compare two treatment groups with respect to their efficacy and toxicity response.

Let $Y_{\ell,i,j} = (Y^E_{\ell,i,j}, Y^T_{\ell,i,j}) \in \{0, 1\}^2$ denote the bivariate outcome for the jth patient allocated to the ith dose level $d_{\ell,i}$ of treatment group ℓ. We observe the data $Y_{\ell,i,j} = (Y^E_{\ell,i,j}, Y^T_{\ell,i,j})$ and denote by

$$\zeta^{\ell,i}_{pq} := \sum_{j=1}^{n_{\ell,i}} I\{(y^E_{\ell i j}, y^T_{\ell i j}) = (p, q)\}$$

the number of responses with outcome (p, q) at dose level $d_{\ell,i}$ in group $\ell = 1, 2$, $i = 1, \ldots, k_\ell$. We use the Gumble model as introduced in Section 3.1. According to (3.5) the likelihood of

the Gumbel model for group ℓ is given by

$$\mathcal{L}_\ell(\theta_\ell | y_{\ell,1,1}, \ldots, y_{\ell,1,n_{\ell,1}}, \ldots, y_{\ell,k_\ell,n_{\ell,k_\ell}})$$

$$= \prod_{i=1}^{k_\ell} \prod_{j=1}^{n_{\ell,i}} p_{11}(d_{\ell,i})^{y_{\ell ij}^E y_{\ell ij}^T} p_{01}(d_{\ell,i})^{(1-y_{\ell ij}^E)y_{\ell ij}^T} p_{10}(d_{\ell,i})^{y_{\ell ij}^E(1-y_{\ell ij}^T)} p_{00}(d_{\ell,i})^{(1-y_{\ell ij}^E)(1-y_{\ell ij}^T)}$$

$$= \prod_{i=1}^{k_\ell} p_{11}(d_{\ell,i})^{\zeta_{11}^{\ell,i}} p_{01}(d_{\ell,i})^{\zeta_{10}^{\ell,i}} p_{10}(d_{\ell,i})^{\zeta_{10}^{\ell,i}} p_{00}(d_{\ell,i})^{\zeta_{00}^{\ell,i}}.$$

Taking the logarithm yields

$$l_\ell(\theta_\ell) := \log \mathcal{L}_\ell(\theta_\ell | y_{\ell,1,1}, \ldots, y_{\ell,1,n_{\ell,1}}, \ldots, y_{\ell,k_\ell,n_{\ell,k_\ell}})$$

$$= \sum_{i=1}^{k_\ell} \zeta_{11}^{\ell,i} \log p_{11}(d_{\ell,i}) + \zeta_{01}^{\ell,i} \log p_{01}(d_{\ell,i}) + \zeta_{10}^{\ell,i} \log p_{10}(d_{\ell,i}) + \zeta_{00}^{\ell,i} \log p_{00}(d_{\ell,i}) \quad (3.6)$$

and the estimate $\hat{\theta}_\ell$ for the parameter θ_ℓ of the Gumbel model is obtained by maximizing this function over the parameter space ($\ell = 1, 2$). Note that the model estimates $\hat{\beta}_{\ell,1}$, $\hat{\gamma}_{\ell,1}$ are the same as the ones obtained by maximizing the likelihood function in the univariate case (2.2) if $\nu = 0$.

Let

$$\eta_\ell(d, \theta_\ell) = \left(\eta_\ell^E(d, \theta_\ell), \eta_\ell^T(d, \theta_\ell) \right) = \left(\frac{1}{1 + e^{-\beta_{\ell,1} - \gamma_{\ell,1} \cdot d}}, \frac{1}{1 + e^{-\beta_{\ell,2} - \gamma_{\ell,2} \cdot d}} \right)^T$$

denote the vector of efficacy and toxicity curves for group $\ell = 1, 2$. We now want to ensure that claiming equivalence of both treatment groups guarantees that both, efficacy and toxicity response, do not deviate more than a certain prespecified threshold $\varepsilon = (\varepsilon^E, \varepsilon^T)$. Consequently the global hypotheses are given by

$$H_0 : \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \theta_1) - \eta_2^E(d, \theta_2) \right| \geq \varepsilon^E \text{ or } \max_{d \in \mathcal{D}} \left| \eta_1^T(d, \theta_1) - \eta_2^T(d, \theta_2) \right| \geq \varepsilon^T \quad (3.7)$$

against the alternative

$$H_1 : \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \theta_1) - \eta_2^E(d, \theta_2) \right| < \varepsilon^E \text{ and } \max_{d \in \mathcal{D}} \left| \eta_1^T(d, \theta_1) - \eta_2^T(d, \theta_2) \right| < \varepsilon^T. \quad (3.8)$$

This problem can be solved by simultaneously testing the individual hypotheses

$$H_0^E : \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \theta_1) - \eta_2^E(d, \theta_2) \right| \geq \varepsilon^E \text{ vs. } H_1^E : \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \theta_1) - \eta_2^E(d, \theta_2) \right| < \varepsilon^E \quad (3.9)$$

and

$$H_0^T : \max_{d \in \mathcal{D}} \left| \eta_1^T(d, \theta_1) - \eta_2^T(d, \theta_2) \right| \geq \varepsilon^T \text{ vs. } H_1^T : \max_{d \in \mathcal{D}} \left| \eta_1^T(d, \theta_1) - \eta_2^T(d, \theta_2) \right| < \varepsilon^T. \quad (3.10)$$

As the global null in (3.7) is the union of H_0^E and H_0^T we can apply the Intersection-Union-Principle (see Berger (1982)). Only if both individual null hypotheses in (3.9) and (3.10) can

be rejected, the global null in (3.7) is rejected and equivalence of the two responses can be claimed. Each of the two individual tests for (3.9) and (3.10) is performed by extending the parametric bootstrap approach in Algorithm 2.1, as described in the following algorithm.

Algorithm 3.1. (parametric bootstrap for testing equivalence for bivariate binary outcomes)

(1) Calculate the MLE $\hat{\theta}_\ell = (\hat{\beta}_{\ell,1}, \hat{\gamma}_{\ell,1}, \hat{\beta}_{\ell,2}, \hat{\gamma}_{\ell,2}, \hat{\nu}_\ell)$, $\ell = 1, 2$, by maximizing the log-likelihood given in (3.6) for each group. The test statistics are obtained by

$$\hat{d}^E = d^E(\hat{\theta}_1, \hat{\theta}_2) = \max_{d \in \mathcal{D}} \left| \eta_1^E(d, \hat{\theta}_1) - \eta_2^E(d, \hat{\theta}_2) \right|$$

and

$$\tilde{d}^T = d^T(\hat{\theta}_1, \hat{\theta}_2) = \max_{d \in \mathcal{D}} \left| \eta_1^T(d, \hat{\theta}_1) - \eta_2^T(d, \hat{\theta}_2) \right|$$

(2) For each individual test for (3.9) and (3.10) we perform a constrained optimization as described in Algorithm 2.1, yielding estimates $\hat{\tilde{\theta}}_\ell$, $\ell = 1, 2$. Note that this procedure is done separately for each individual test because the constraints and hence the generation of the bootstrap data differ. Thus we generate bootstrap data for each individual test separately and obtain replicates $\hat{d}_{\infty,1}^{E*}, \ldots, \hat{d}_{\infty,n_{boot}}^{E*}$ for the comparison of the efficacy curves and $\hat{d}_{\infty,1}^{T*}, \ldots, \hat{d}_{\infty,n_{boot}}^{T*}$ for the comparison of the toxicity curves. Let $\hat{d}^{E*(1)} \leq \ldots \leq \hat{d}^{E*(n_{boot})}$ and $\hat{d}^{T*(1)} \leq \ldots \leq \hat{d}^{T*(n_{boot})}$ denote the corresponding order statistics and let $\hat{d}^{E*(\lfloor n_{boot}\alpha \rfloor)}$ and $\hat{d}^{T*(\lfloor n_{boot}\alpha \rfloor)}$ denote the corresponding empirical level α quantiles.

(3) Reject the global null hypothesis (3.7) if

$$\hat{d}^E < \hat{d}^{E*(\lfloor n_{boot}\alpha \rfloor)} \quad \text{and} \quad \tilde{d}^T < \tilde{d}^{T*(\lfloor n_{boot}\alpha \rfloor)}. \tag{3.11}$$

Note that according to the Intersection-Union-Principle we use the α-quantile and there is no need of adjusting the level of the two individual tests. The technical difficulty of the implementation of this algorithm consists in generating bivariate correlated binary data in Step (2), which is explained in more detail in the following section.

3.3 Generation of bivariate correlated binary data

The bootstrap test described in Algorithm 3.1 requires the simulation of bivariate binary data. Due to the dependency of the outcomes this is a technical difficulty investigated by several authors (see for example Emrich and Piedmonte (1991); Lunn and Davies (1998); Leisch et al. (1998) among many others). We used the algorithm developed by Emrich and Piedmonte (1991), as implemented with the function generate.binary in the R package MultiOrd (see Amatya and Demirtas (2015)). For this purpose, we use expression (3.4) for the correlation

and the marginal distributions in (3.3) to generate the data at each dose level as long as the correlation does not exceed the boundaries specified by the model parameter θ given by

$$\max\left(-\sqrt{\frac{p_1(d)p_2(d)}{(1-p_1(d))(1-p_2(d))}}, -\sqrt{\frac{(1-p_1(d))(1-p_2(d))}{p_1(d)p_2(d)}}\right) \leq corr(Y^E, Y^T \mid d) \qquad (3.12)$$

and

$$corr(Y^E, Y^T \mid d) \leq min\left(\sqrt{\frac{p_1(d)(1-)p_2(d))}{(1-p_1(d))p_2(d)}}, \sqrt{\frac{(1-p_1(d))p_2(d)}{p_1(d)(1-p_2(d))}}\right). \qquad (3.13)$$

Here, $p_1(d) = \eta^E(d, \theta_1)$ and $p_2(d) = \eta^T(d, \theta_2)$ denote the marginal probabilities of efficacy and toxicity, respectively. These restrictions have to be fulfilled at each dose in order to guarantee that a joint distribution of Y^E and Y^T can exist. We impose these inequality constraints in the optimization step in addition to the constraint described in (2.4) such that the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ generate a distribution and bootstrap data can be obtained.

4 Finite sample properties

We now investigate the finite sample properties of the two tests based on Algorithms 2.1 and 3.1. Following Murtaugh and Fisher (1990) we consider the (log-transformed) dose range $\mathcal{D} = [-3, 3]$ and 7 dose levels $-3, -2, \ldots, 2, 3$. We assume 7, 14, 21, 28 and 50 patients per dose level and group, that is $n_{\ell,i} = 7$, 14, 21, 28 and 50 for $\ell = 1, 2$, $i = 1, \ldots 7$, resulting in $n_\ell = 49, 98, 147, 196$ and 350, $\ell = 1, 2$. The significance level is $\alpha = 0.05$ throughout. Following Nam (1997) and Chen et al. (2000) we assume three different equivalence thresholds, 0.1, 0.15 and 0.2. All simulations are performed using 1000 simulation runs and $n_{boot} = 400$ bootstrap replications. The binary data are generated as described in Section 3.3. We set $\nu = 0$ for the univariate case.

4.1 Univariate efficacy outcomes

We consider a logistic regression in (2.1). The reference model is specified by $(\beta_{1,1}, \gamma_{1,1}) = (0, 1)$ yielding $\eta_1^E(d, \beta_{1,1}, \gamma_{1,1}) = \frac{1}{1+e^{-d}}$. We choose the parameters $(\beta_{2,1}, \gamma_{2,1})$ of the second model as

$$\eta_2^E(d, \beta_{2,1}, \gamma_{2,1}) = \frac{1}{1 + e^{-\beta_{2,1} - \gamma_{2,1}d}},$$

such that the maximum deviations d^E between the two efficacy curves η_1^E and η_2^E are 0, 0.05, 0.1, 0.15, 0.2 and 0.3, attained at the doses 1.11, 0.99, 0.78, 0.65 and 0.26, respectively. This leads to the configurations

$$(\beta_{2,1}, \gamma_{2,1}) = (0, 1), \ (\beta_{2,1}, \gamma_{2,1}) = (0.1, 1.2), \ (\beta_{2,1}, \gamma_{2,1}) = (0.2, 1.4),$$
$$(\beta_{2,1}, \gamma_{2,1}) = (0.4, 1.6), \ (\beta_{2,1}, \gamma_{2,1}) = (0.6, 1.9) \text{ and } (\beta_{2,1}, \gamma_{2,1}) = (1.3, 2.1). \qquad (4.1)$$

10

Note that for $\eta_1^E = \eta_2^E$ the difference between the curves is zero at all doses. Figure 1 displays the reference curve η_1^E and the curve η_2^E determined by the parameters described in (4.1).
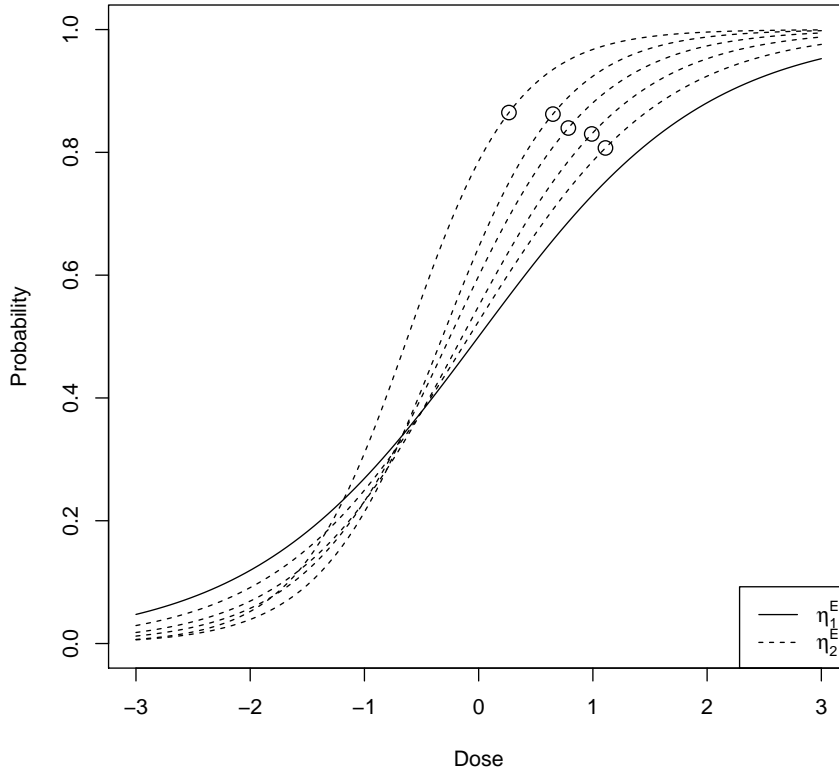

Figure 1: *The reference efficacy curve* $\eta_1^E(d, \beta_{1,1}, \gamma_{1,1}) = \frac{1}{1+e^{-d}}$ *(solid line) and the curves* $\eta_2^E(d, \beta_{2,1}, \gamma_{2,1}) = \frac{1}{1+e^{-\beta_{2,1}-\gamma_{2,1}d}}$ *(dashed lines) for different choices of the parameters* $(\beta_{2,1}, \gamma_{2,1})$, *as defined in* (4.1). *The scenarios correspond to a maximum absolute deviation* $d^E = 0, 0.05, 0.1, 0.15, 0.2, 0.3$ *attained at the open dots (from right to left).*

Table 1 displays the simulated type I error rates of the bootstrap test (2.6) for the equivalence of efficacy responses with margins $\varepsilon^E = 0.1, 0.15, 0.2$. The numbers in bold face indicate the scenarios where simulations have been run on the margin of the null, that is $d^E = \varepsilon^E = 0.1, 0.15$ and 0.2. Note that the configuration $d^E = 0.15$ and $\varepsilon^E = 0.2$ falls under the alternative (as $d^E < \varepsilon^E$) and is therefore omitted from Table 1. We conclude that the test controls its level in all cases under consideration. The approximation of the level is very precise at the margin of the null hypothesis (that is, $d^E = \varepsilon^E$) and this accuracy increases with increasing sample sizes. Moreover, in the interior of the null hypothesis (that is $d^E \geq \varepsilon^E$) the number of rejections is close to zero in all scenarios, indicating that the type I error rate is well below α in these cases.

Table 2 displays the power of the test (2.6). We conclude that for sufficiently large sample sizes the procedure has reasonable power. For instance, for $n_{\ell,i} = 28$, $i = 1, \ldots 7$, $\ell = 1, 2$, the maximum power attained at $d^E = 0$ is 0.785 for an equivalence threshold of $\varepsilon^E = 0.2$. For larger sample sizes of 50 patients per dose level, the test achieves more than 80% power, namely 0.803 for $\varepsilon^E = 0.15$ and 0.976 for $\varepsilon^E = 0.2$. In general we observe that the power increases with increasing sample sizes. Note that the case $d^E = \varepsilon^E = 0.1$ falls under the null hypothesis and results are therefore shown in Table 1.

$n_{\ell,i}$	$(\beta_{2,1}, \gamma_{2,1})$	d^E	$\varepsilon^E = 0.1$	$\varepsilon^E = 0.15$	$\varepsilon^E = 0.2$
7	(1.3,2.1)	0.3	0.005	0.014	0.018
	(0.6,1.9)	0.2	0.020	0.022	**0.055**
	(0.4,1.6)	0.15	0.036	**0.037**	-
	(0.2,1.1)	0.1	**0.060**	-	-
14	(1.3,2.1)	0.3	0.002	0.001	0.003
	(0.6,1.9)	0.2	0.005	0.014	**0.055**
	(0.4,1.6)	0.15	0.020	**0.038**	-
	(0.2,1.1)	0.1	**0.042**	-	-
21	(1.3,2.1)	0.3	0.000	0.000	0.002
	(0.6,1.9)	0.2	0.004	0.007	**0.052**
	(0.4,1.6)	0.15	0.010	**0.042**	-
	(0.2,1.1)	0.1	**0.036**	-	-
28	(1.3,2.1)	0.3	0.000	0.000	0.001
	(0.6,1.9)	0.2	0.000	0.012	**0.062**
	(0.4,1.6)	0.15	0.008	**0.040**	-
	(0.2,1.1)	0.1	**0.036**	-	-
50	(1.3,2.1)	0.3	0.000	0.000	0.000
	(0.6,1.9)	0.2	0.002	0.011	**0.057**
	(0.4,1.6)	0.15	0.006	**0.052**	-
	(0.2,1.1)	0.1	**0.034**	-	-

Table 1: *Simulated type I error rates of the bootstrap test (2.6) for the equivalence of efficacy responses. Bold numbers indicate simulations at the margin of the null hypothesis.*

4.2 Bivariate efficacy-toxicity outcomes

We now consider bivariate efficacy-toxicity outcomes using a Gumbel model for both treatment groups as defined in Section 3.1. The reference model is defined by the parameter

$$\theta_1 = (\beta_{1,1}, \gamma_{1,1}, \beta_{1,2}, \gamma_{1,2}, \nu_1) = (0, 1, 0, 0.5, \nu_1) \tag{4.2}$$

and we assume two different levels of dependence representing a moderate ($\nu_1 = 1$) and a rather strong dependence ($\nu_1 = 3$) between the efficacy and toxicity outcomes. According to (3.4), the correlation of Y_1^E and Y_1^T at dose $d \in \mathcal{D}$ is given by

$$corr(Y_1^E, Y_1^T \mid d) = \frac{\nu_1}{(e^{d/2} + e^{-d/2})(e^{d/4} + e^{-d/4})}, \tag{4.3}$$

$n_{\ell,i}$	$(\beta_{2,1}, \gamma_{2,1})$	d^E	$\varepsilon^E = 0.1$	$\varepsilon^E = 0.15$	$\varepsilon^E = 0.2$
	(0.2,1.4)	0.1	-	0.082	0.090
7	(0.1,1.2)	0.05	0.058	0.076	0.145
	(0,1)	0	0.076	0.171	0.232
	(0.2,1.4)	0.1	-	0.137	0.247
14	(0.1,1.2)	0.05	0.075	0.142	0.391
	(0,1)	0	0.101	0.226	0.418
	(0.2,1.4)	0.1	-	0.166	0.434
21	(0.1,1.2)	0.05	0.090	0.344	0.547
	(0,1)	0	0.134	0.356	0.603
	(0.2,1.4)	0.1	-	0.203	0.474
28	(0.1,1.2)	0.05	0.103	0.367	0.690
	(0,1)	0	0.179	0.470	0.785
	(0.2,1.4)	0.1	-	0.303	0.729
50	(0.1,1.2)	0.05	0.184	0.640	0.905
	(0,1)	0	0.363	0.803	0.976

Table 2: *Simulated power of the bootstrap test* (2.6) *for the equivalence of efficacy responses.*

which ranges from 0.08 to 0.25 for $\nu_1 = 1$ and 0.25 to 0.75 for $\nu_1 = 3$. Note that the highest correlation is always attained at the dose level 0. The left panel of Figure 2 displays the probability of efficacy without toxicity response, that is $\mathbb{P}(Y^E = 1, Y^T = 0 \mid d) = p_{10}(d)$. The right panel displays the correlation for different choices of ν in dependence of the dose. In order to investigate different situations under the null and the alternative, we vary the parameters of the second model resulting in seven scenarios for each choice of ν_1; see Table 3. We assume the same correlations as for the reference model, that is $\nu_2 = \nu_1$. As an illustration, we show the efficacy and toxicity curves for three scenarios and $\nu_1 = 1$ in Figure 3.

	θ_1	θ_2	$d = (d^E, d^T)$
	$(0, 1, 0, 0.5, \nu_2)$	$(0, 1, 0, 0.5, \nu_2)$	$(0, 0)$
Alternative	$(0, 1, 0, 0.5, \nu_2)$	$(0.1, 1.2, 0.1, 0.6, \nu_2)$	$(0.05, 0.05)$
	$(0, 1, 0, 0.5, \nu_2)$	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	$(0.1, 0.1)$
	$(0, 1, 0, 0.5, \nu_2)$	$(0.4, 1.6, 0.4, 0.8, \nu_2)$	$(0.15, 0.15)$
Null hypothesis	$(0, 1, 0, 0.5, \nu_2)$	$(0, 1, 0.4, 0.8, \nu_2)$	$(0, 0.15)$
	$(0, 1, 0, 0.5, \nu_2)$	$(0.6, 1.9, 0.5, 1, \nu_2)$	$(0.2, 0.2)$
	$(0, 1, 0, 0.5, \nu_2)$	$(0, 1, 0.5, 1, \nu_2)$	$(0, 0.2)$

Table 3: *Different scenarios corresponding to the null hypothesis* (3.7) *and the alternative* (3.8).

For the Type I error rate simulations we counted the number of individual and simultaneous rejections of both null hypotheses in (3.9) and (3.10), allowing us to reject the global null hypothesis in (3.7). All simulation results are displayed in Tables 4 and 5, where the numbers in brackets correspond to the proportion of rejections for the individual tests on efficacy and toxicity. For the sake of brevity we assume only two different thresholds $\varepsilon = (\varepsilon^E, \varepsilon^T) = (0.15, 0.15)$ and $(0.2, 0.2)$, thus allowing for a deviation of 15% and 20%, respectively, for efficacy
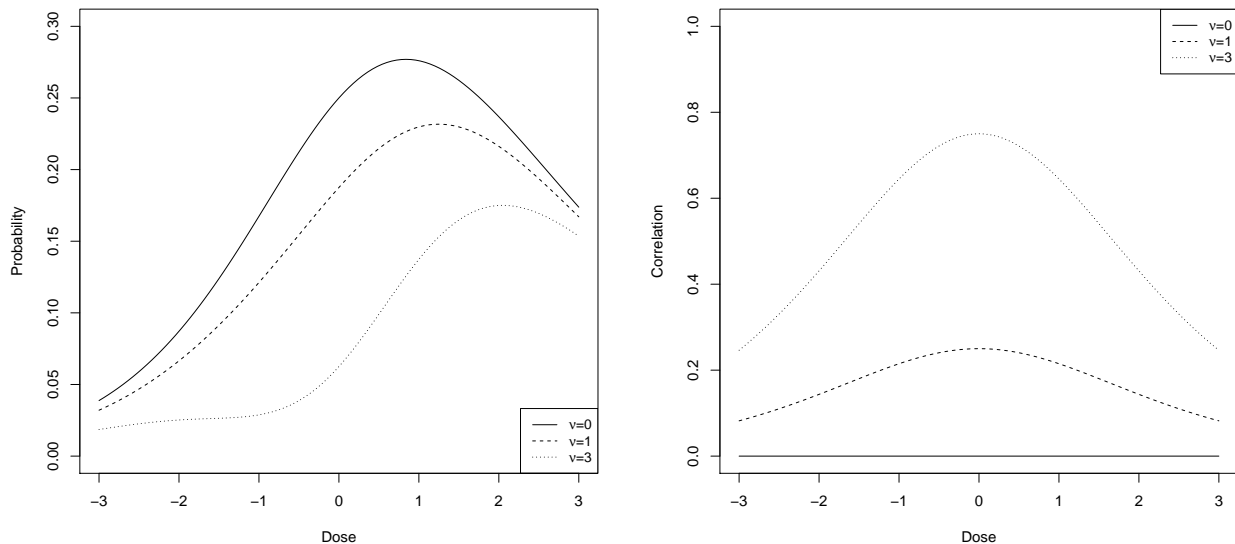
Figure 2: *Left panel: Probability $\mathbb{P}(Y^E = 1, Y^T = 0) = p_{10}(d)$ in dependence of the dose for the reference model (4.2) for different choices of the correlation parameter ν. Right panel: Correlation of efficacy and toxicity response for different choices of ν in dependence of the dose.*

and toxicity in order to claim equivalence. In general, we observe that the global bootstrap test according to Algorithm 3.1 is rather conservative as the Type I error rates are very small. For example, for $n_{\ell,i} = 21$, $\nu_1 = \nu_2 = 1$ and $\varepsilon = (0.2, 0.2)$ the individual proportions of rejection are 0.041 for efficacy and 0.050 for toxicity, whereas the Type I error rate for the global test is 0.005, which is well below the nominal level. This is a common feature of the Intersection-Union-Principle for the problem of testing bioequivalence in multivariate responses (see, for example Berger and Hsu (1996)).

In general, we conclude that the individual tests on efficacy and toxicity yield rejection probabilities that are very close to 0.05 when simulating on the margin of the global null hypothesis (that is $d = \varepsilon$) and hence the global Type I error rates are well below α in these cases. However, there are some scenarios where the Type I error rate is too large when $\nu_1 = \nu_2 = 3$. For instance, we observe a proportion of rejections of the global null hypothesis of 0.129 for $n_{\ell,i} = 28$, $\varepsilon = (0.2, 0.2)$ and $d = (0, 0.2)$. Note that the values of ν_1 and ν_2 do not influence the curves obtained by the marginal densities in (3.3) and hence do not directly impact the proportions of rejections obtained for the individual tests. However, the choice of ν_ℓ affects the estimation of the parameter θ_ℓ of the Gumbel model, which explains the different results for the individual tests for $\nu_\ell = 1$ and $\nu_\ell = 3$, $\ell = 1, 2$, resulting in higher Type I error rates for the global test in settings with $\nu_\ell = 3$. For example, the choice $d = (0.2, 0)$ and $\varepsilon = (0.2, 0.2)$ corresponds to the
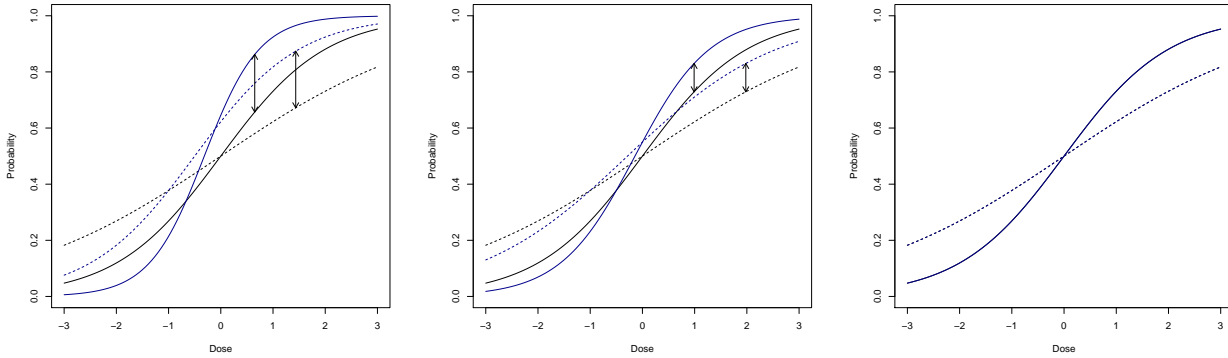
14

Figure 3: *Efficacy curves (solid lines) and toxicity curves (dashed lines) derived in* (3.3). *The black lines correspond to the reference model, the blue lines to the second model, specified by θ_2. The scenarios shown correspond to a maximum absolute deviation (indicated by the arrows) of $d^E = d^T = 0.2$, 0.1 and 0 (from left to right).*

global null hypothesis (3.7), as $d^E = 0.2 = \varepsilon^E$. However, due to the fact that we are far under the alternative for toxicity ($d^T = 0$) and due to the high correlation ($\nu_\ell = 3$) we observe a Type I error rate inflation for the individual test on efficacy and consequently for the global test as well, for all sample sizes. There are two reasons causing this effect: on the one hand the high correlation results in difficulties to estimate the curves properly, even for large sample sizes. On the other hand, the asymptotic distribution of the maximum absolute deviation of the two curves is different and more complex in case of $d^T = 0$, which also affects the test results; see Dette et al. (2018) for further numerical and theoretical details on this issue.

A similar argument also holds for the power results shown in Table 5. It turns out that the global test achieves reasonable power for sufficiently large sample sizes. For example a maximum power (always attained at $d = (0,0)$) of 0.817 is achieved for the global test for a choice of $n_{\ell,i} = 28$, $\nu_1 = \nu_2 = 3$ and $\varepsilon = (0.2, 0.2)$. For a lower threshold, that is $\varepsilon = (0.15, 0.15)$, the maximum power is smaller, but still increasing with growing sample sizes, reaching for instance 0.830 for $n_{\ell,i} = 50$ and $\nu_1 = \nu_2 = 3$.

5   Case study

To illustrate the proposed methodology, we consider an example that is inspired by a recent consulting project of one of the authors. A nonsteroidal anti-inflammatory drug is to be investigated for its ability to attenuate dental pain after the removal of two or more impacted third molar teeth. Dental pain is a common and inexpensive setting for analgesic proof of concept, recruitment being fast and the end-point being available within a few hours. It is

$\varepsilon = (\varepsilon^E, \varepsilon^T)$	$n_{\ell,i}$	θ_2	$d = (d^E, d^T)$	$\nu_\ell = 1$	$\nu_\ell = 3$
(0.15, 0.15)	7	$(0.4, 1.6, 0.4, 0.8, \nu_2)$	(0.15, 0.15)	0.004 (0.051/0.050)	0.021 (0.078/0.063)
		$(0, 1, 0.4, 0.8, \nu_2)$	(0, 0.15)	0.009 (0.142/0.056)	0.029 (0.180/0.087)
	14	$(0.4, 1.6, 0.4, 0.8, \nu_2)$	(0.15, 0.15)	0.005 (0.052/0.056)	0.006 (0.041/0.044)
		$(0, 1, 0.4, 0.8, \nu_2)$	(0, 0.15)	0.007 (0.212/0.051)	0.025 (0.256/0.071)
	21	$(0.4, 1.6, 0.4, 0.8, \nu_2)$	(0.15, 0.15)	0.005 (0.032/0.055)	0.011 (0.042/0.050)
		$(0, 1, 0.4, 0.8, \nu_2)$	(0, 0.15)	0.029 (0.364/0.056)	0.049 (0.395/0.091)
	28	$(0.4, 1.6, 0.4, 0.8, \nu_2)$	(0.15, 0.15)	0.004 (0.036/0.044)	0.014 (0.055/0.065)
		$(0, 1, 0.4, 0.8, \nu_2)$	(0, 0.15)	0.017 (0.648/0.044)	0.064 (0.610/0.098)
	50	$(0.4, 1.6, 0.4, 0.8, \nu_2)$	(0.15, 0.15)	0.004 (0.050/0.047)	0.035 (0.096/0.071)
		$(0, 1, 0.4, 0.8, \nu_2)$	(0, 0.15)	0.053 (0.831/0.062)	0.112 (0.866/0.128)
(0.2, 0.2)	7	$(0.6, 1.9, 0.5, 1, \nu_2)$	(0.2, 0.2)	0.002 (0.061/0.038)	0.021 (0.062/0.064)
		$(0, 1, 0.5, 1, \nu_2)$	(0, 0.2)	0.018 (0.209/0.060)	0.029 (0.253/0.077)
	14	$(0.6, 1.9, 0.5, 1, \nu_2)$	(0.2, 0.2)	0.007 (0.033/0.042)	0.009 (0.042/0.046)
		$(0, 1, 0.5, 1, \nu_2)$	(0, 0.2)	0.015 (0.417/0.038)	0.053 (0.465/0.088)
	21	$(0.6, 1.9, 0.5, 1, \nu_2)$	(0.2, 0.2)	0.005 (0.041/0.050)	0.018 (0.060/0.057)
		$(0, 1, 0.5, 1, \nu_2)$	(0, 0.2)	0.025 (0.431/0.034)	0.074 (0.612/0.093)
	28	$(0.6, 1.9, 0.5, 1, \nu_2)$	(0.2, 0.2)	0.008 (0.068/0.072)	0.021 (0.070/0.050)
		$(0, 1, 0.5, 1, \nu_2)$	(0, 0.2)	0.044 (0.796/0.055)	0.129 (0.817/0.144)
	50	$(0.6, 1.9, 0.5, 1, \nu_2)$	(0.2, 0.2)	0.009 (0.076/0.067)	0.043 (0.103/0.083)
		$(0, 1, 0.5, 1, \nu_2)$	(0, 0.2)	0.053 (0.968/0.059)	0.223 (0.972/0.267)

Table 4: *Simulated Type I error rates of the global bootstrap test* (3.11) *for different choices of* ν_ℓ, $\ell = 1, 2$. *The numbers in brackets show the proportion of rejections for the individual tests according to the hypotheses* (3.9) *and* (3.10).

common to measure the pain intensity on an ordinal scale at baseline and several times after the administration of a single dose. The pain intensity difference from baseline (PID), averaged over several hours after drug administration, may then be compared with a clinical relevance threshold to create a binary success variable for efficacy. In this particular setting, side effects such as nausea and sedation after dosing were anticipated, resulting in a binary toxicity variable whether the patient experienced any such adverse events. As approved analgesics with an identified dosing range and a known dose-response relationship for tolerability are available, the objective of the study at hand was to demonstrate equivalence with a marketed product for the bivariate efficacy-toxicity outcome in a proof of concept setting.

This was a randomized double-blind parallel group trial with a total of 300 patients being allocated to either placebo or one of four active doses coded as 0.05, 0.20, 0.50, and 1 (for the new treatment) and 0.10, 0.30, 0.60, and 1 (for the marketed product), resulting in $n = 30$ per group (assuming equal allocation). To maintain confidentiality, the actual doses have been scaled to lie within the $[0, 1]$ interval. Since the study has not been completed yet, we use a hypothetical data set to illustrate the proposed methodology.

We fit two Gumbel models as defined in Section 3.1 to the data, one for the marketed product ($\ell = 1$) and one for the new product ($\ell = 2$). The estimated model parameters are

$$\hat{\theta}_1 = (-0.971, 2.254, -2.497, 1.806, -0.030), \ \hat{\theta}_2 = (-1.585, 2.963, -2.162, 1.287, 1.003), \quad (5.1)$$

16

$\varepsilon = (\varepsilon^E, \varepsilon^T)$	$n_{\ell,i}$	θ_2	$d = (d^E, d^T)$	$\nu_\ell = 1$	$\nu_\ell = 3$
(0.15, 0.15)	7	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.005 (0.076/0.088)	0.020 (0.089/0.092)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.010 (0.117/0.109)	0.047 (0.168/0.142)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.015 (0.120/0.133)	0.061 (0.179/0.148)
	14	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.008 (0.144/0.119)	0.042 (0.137/0.116)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.027 (0.156/0.159)	0.097 (0.236/0.204)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.040 (0.234/0.182)	0.152 (0.296/0.263)
	21	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.023 (0.153/0.157)	0.088 (0.197/0.189)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.067 (0.270/0.271)	0.207 (0.387/0.326)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.126 (0.380/0.308)	0.259 (0.462/0.363)
	28	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.029 (0.204/0.161)	0.150 (0.262/0.266)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.113 (0.319/0.353)	0.344 (0.536/0.473)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.230 (0.502/0.441)	0.437 (0.646/0.526)
	50	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.103 (0.313/0.332)	0.281 (0.426/0.416)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.401 (0.615/0.624)	0.650 (0.792/0.727)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.678 (0.811/0.827)	0.830 (0.943/0.856)
(0.2, 0.2)	7	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.010 (0.127/0.113)	0.060 (0.176/0.169)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.037 (0.199/0.133)	0.067 (0.236/0.177)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.050 (0.230/0.186)	0.099 (0.264/0.218)
	14	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.067 (0.197/0.239)	0.191 (0.354/0.322)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.144 (0.351/0.353)	0.249 (0.495/0.409)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.198 (0.442/0.407)	0.324 (0.560/0.439)
	21	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.137 (0.317/0.365)	0.311 (0.476/0.453)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.286 (0.532/0.535)	0.495 (0.683/0.578)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.418 (0.676/0.613)	0.601 (0.828/0.666)
	28	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.253 (0.483/0.478)	0.460 (0.634/0.574)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.451 (0.637/0.706)	0.723 (0.870/0.775)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.650 (0.798/0.791)	0.817 (0.942/0.843)
	50	$(0.2, 1.4, 0.2, 0.7, \nu_2)$	(0.1, 0.1)	0.511 (0.702/0.700)	0.745 (0.853/0.804)
		$(0.1, 1.2, 0.1, 0.6, \nu_2)$	(0.05, 0.05)	0.826 (0.906/0.910)	0.964 (0.999/0.966)
		$(0, 1, 0, 0.5, \nu_2)$	(0, 0)	0.961 (0.979/0.980)	0.985 (1.000/0.987)

Table 5: *Simulated power of the global bootstrap test* (3.11) *for different choices of* ν_ℓ, $\ell = 1, 2$. *The numbers in brackets show the proportion of rejections for the individual tests according to the hypotheses* (3.9) *and* (3.10).

see Figure 4 for the corresponding efficacy and toxicity curves.

The maximum distances are $\hat{d}^E = 0.106$ and $\hat{d}^T = 0.039$, attained at dose 0.08 and the maximum dose 1, respectively. We perform an equivalence test at a significance level of $\alpha = 0.05$, as defined in Algorithm 3.1, for three different sets of hypotheses as we vary the equivalence thresholds $\varepsilon = (\varepsilon^E, \varepsilon^T)$ in (3.9) and (3.10). Table 6 displays the critical values obtained by $n_{boot} = 1000$ bootstrap replications for the different choices of ε.

We now test the global null hypothesis (3.7) against the alternative (3.8). For $\varepsilon = (0.2, 0.2)$ we have $\hat{d}^E = 0.106 < 0.111 = \hat{q}^E_{0.05}$ and $\hat{d}^T = 0.039 < 0.054 = \hat{q}^T_{0.05}$. According to (3.9) and (3.10), we can therefore reject (3.7) at level $\alpha = 0.05$ and for $\varepsilon = (0.2, 0.2)$. However, we cannot reject (3.7) for the other choices of ε. For example, $\hat{d}^E = 0.106 > 0.073 = \hat{q}^E_{0.05}$ for $\varepsilon^E = 0.1$ and we cannot reject (3.7) according to (3.9).
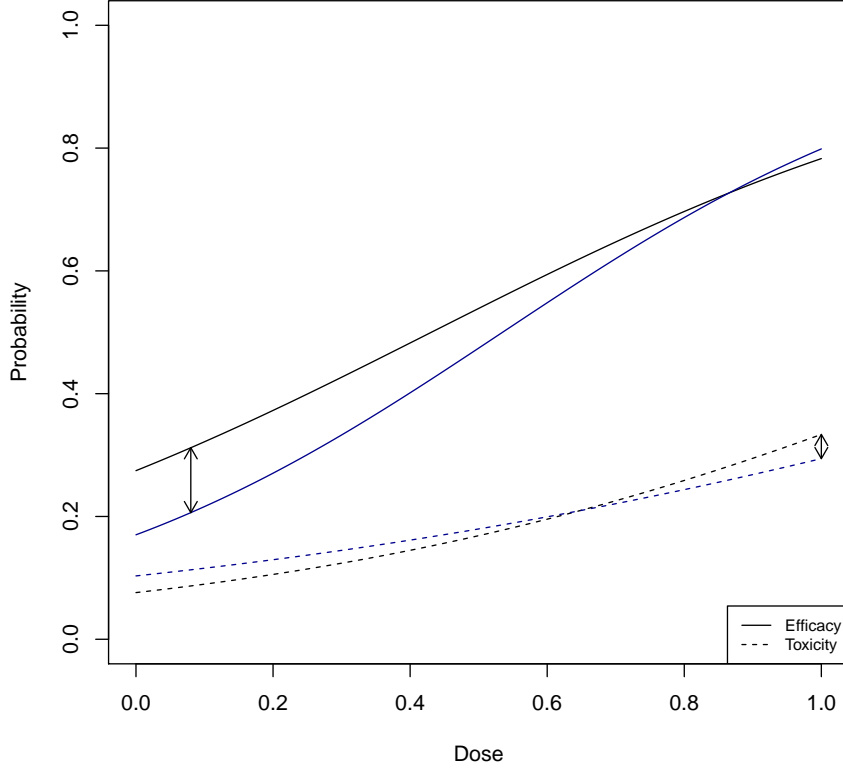
Figure 4: *Efficacy and toxicity curves corresponding to the fitted Gumbel model* (5.1). *The black lines correspond to the marketed product, the blue lines to the new product respectively, where solid lines display the efficacy and dashed lines the toxicity response. The arrows indicate the maximum absolute distances.*

Quantile	$\varepsilon = (0.1, 0.1)$	$\varepsilon = (0.15, 0.15)$	$\varepsilon = (0.2, 0.2)$
$\hat{q}_{0.05}^{E}$	0.042 (0.327)	0.073 (0.160)	0.111 (0.040)
$\hat{q}_{0.05}^{T}$	0.026 (0.094)	0.037 (0.061)	0.054 (0.030)

Table 6: *Critical values of the two individual bootstrap tests on the hypotheses* (3.9) *and* (3.10) *for three different equivalence thresholds ε. The numbers in brackets correspond to the p-values of the individual tests.*

We obtain the same conclusions based on the observed p-values reported in brackets in Table 6. These p-values were obtained from the empirical distribution functions of the bootstrap sample according to Step (iii) of Algorithm 2.1. In general, we reject the null hypothesis (3.7) at level α if the maximum of the two individual p-values for (3.9) and (3.10) is smaller than or equal to α. In our example, this only holds for $\varepsilon = (0.2, 0.2)$ since the individual p-values are given

by $\hat{F}_{n_{boot}}^E(\hat{d}^E) = 0.04$ and $\hat{F}_{n_{boot}}^T(\hat{d}^T) = 0.03$ such that $\max(0.03, 0.04) = 0.04 < 0.05 = \alpha$.

6   Conclusions and discussion

In the first part of this paper we investigated a single efficacy response given by a binary outcome and derived a test procedure for the equivalence of the corresponding dose-response curves, which can be modelled, for instance, by a parametric logistic regression or a probit model. We developed a parametric bootstrap test and decide for equivalence if the maximum deviation between the estimated dose response profiles is sufficiently small. We also considered the situation of an additional second toxicity endpoint to model the joint efficacy-toxicity responses. For this purpose we assumed efficacy and toxicity to be observed simultaneously resulting in bivariate (correlated) binary outcomes and used a Gumbel model to fit the data. The bootstrap test was extended to this situation by combining two individual tests through the Intersection-Union-Principle.

In the second part of this paper we investigated the operating characteristics by means of an extensive simulation study. We demonstrated that the resulting procedures control their level and achieve reasonable power. The choice of the equivalence threshold ε has a major impact on the performance of the test. The explicit choice has to be made on an individual basis and under consideration of clinical experts.

In certain settings the efficacy or toxicity responses are not modelled by binary outcomes, but rather by a continuous response. In case of two continuous outcomes, Fedorov and Wu (2007) considered normally distributed correlated responses which are dichotomized due to binary utility and the methodology proposed in this paper can be adapted to the situation considered by these authors. A further interesting situation occurs in case of mixed outcomes, where one of the response variables is continuous and the other a binary one. Modelling these types of responses is a challenging problem and not much work has been done on this topic in the literature. Tao et al. (2013) investigated this situation by modelling these multiple endpoints by a joint model constructed with archimedean copula. An equivalence test for these types of outcomes is an interesting topic which we leave for future research.

7   Software

Software in the form of R code together with a sample input data set and complete documentation is available online at `https://github.com/kathrinmoellenhoff/Efficacy_Toxicity`.

jekt T1) of the German Research Foundation (DFG).

References

Amatya, A. and Demirtas, H. (2015). Multiord: An r package for generating correlated ordinal data. *Communications in Statistics-Simulation and Computation*, 44(7):1683–1691.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.*, 11(4):283–319.

Bradley, R. A. and Gart, J. J. (1962). The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214.

Bretz, F., Möllenhoff, K., Dette, H., Liu, W., and Trampisch, M. (2018). Assessing the similarity of dose response and target doses in two non-overlapping subgroups. *Statistics in Medicine*, 37(5):722–738.

Chen, J. J., Tsong, Y., and Kang, S.-H. (2000). Tests for equivalence or noninferiority between two proportions. *Drug Information Journal*, 34(2):569–578.

Chow, S.-C. and Liu, P.-J. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, New York.

Cox, D. R. (2018). *Analysis of binary data*. Routledge.

Deldossi, L., Osmetti, S. A., and Tommasi, C. (2019). Optimal design to discriminate between rival copula models for a bivariate binary response. *TEST*, 28(1):147–165.

Dette, H., Möllenhoff, K., Volgushev, S., and Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, 113:711–729.

Dragalin, V. and Fedorov, V. (2006). Adaptive designs for dose-finding based on efficacy–toxicity response. *Journal of Statistical Planning and Inference*, 136(6):1800–1823.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304.

Fedorov, V. V. and Wu, Y. (2007). Dose finding designs for continuous responses and binary utility. *Journal of Biopharmaceutical Statistics*, 17(6):1085–1096.

Gaydos, B., Krams, M., Perevozskaya, I., Bred, F., Liu, Q., Gallo, P., Berry, D., Chuang-Stein, C., Pinheiro, J., and Bedding, A. (2006). Adaptive dose-response studies. *Drug information journal*, 40(4):451–461.

Glonek, G. F. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):533–546.

Gsteiger, S., Bretz, F., and Liu, W. (2011). Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. *Journal of Biopharmaceutical Statistics*, 21(4):708–725.

Gumbel, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association*, 56(294):335–349.

Heise, M. A. and Myers, R. H. (1996). Optimal designs for bivariate logistic regression. *Biometrics*, pages 613–624.

Leisch, F., Weingessel, A., and Hornik, K. (1998). On the generation of correlated artificial binary data.

Liu, W., Bretz, F., Hayter, A. J., and Wynn, H. P. (2009). Assessing non-superiority, non-inferiority of equivalence when comparing two regression models over a restricted covariate region. *Biometrics*, 65(4):1279–1287.

Long, J. S. and Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

Lunn, A. D. and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika*, 85(2):487–490.

Möllenhoff, K., Dette, H., Kotzagiorgis, E., Volgushev, S., and Collignon, O. (2018). Regulatory assessment of drug dissolution profiles comparability via maximum deviation. *Statistics in Medicine*, 37(20):2968–2981.

Murtaugh, P. A. and Fisher, L. D. (1990). Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Communications in Statistics-Theory and Methods*, 19(6):2003–2020.

Nam, J.-m. (1997). Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics*, pages 1422–1430.

Nebiyou Bekele, B. and Shen, Y. (2005). A bayesian approach to jointly modeling toxicity and biomarker expression in a phase i/ii dose-finding trial. *Biometrics*, 61(2):343–354.

Tao, Y., Liu, J., Li, Z., Lin, J., Lu, T., and Yan, F. (2013). Dose-finding based on bivariate efficacy-toxicity outcome using archimedean copula. *PloS one*, 8(11):e78805.

Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. *Biometrics*, 60(3):684–693.

U.S. Food and Drug Administration (2003). Guidance for industry: bioavailability and bioequivalence studies for orally administered drug products-general considerations. *Food and Drug Administration, Washington, DC*. available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070124.pdf.

Varadhan, R. (2014). Constrained nonlinear optimization. R package version 2011.9-1, available at http://cran.r-project.org/web/packages/alabama/index.html.

Yin, G., Li, Y., and Ji, Y. (2006). Bayesian dose-finding in phase i/ii clinical trials using toxicity and efficacy odds ratios. *Biometrics*, 62(3):777–787.

Zhang, W., Sargent, D. J., and Mandrekar, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in medicine*, 25(14):2365–2383.