Testing for similarity of binary efficacy-toxicity responses

Biostatistics (2022) **23**, 3, *pp.* 949–966
doi: 10.1093/biostatistics/kxaa058
Advance Access publication on March 19, 2021

Testing for similarity of binary efficacy–toxicity responses

KATHRIN MÖLLENHOFF*

Department of Mathematics, Ruhr-Universität Bochum, Universitätsstrasse 150, 44801 Bochum, Germany and Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, The Netherlands

kathrin.moellenhoff@rub.de

HOLGER DETTE

Department of Mathematics, Ruhr-Universität Bochum, Universitätsstrasse 150, 44801 Bochum, Germany

FRANK BRETZ

Statistical Methodology, Novartis Pharma AG, CH-4002 Basel, Switzerland

Summary

Clinical trials often aim to compare two groups of patients for efficacy and/or toxicity depending on covariates such as dose. Examples include the comparison of populations from different geographic regions or age classes or, alternatively, of different treatment groups. Similarity of these groups can be claimed if the difference in average outcome is below a certain margin over the entire covariate range. In this article, we consider the problem of testing for similarity in the case that efficacy and toxicity are measured as binary outcome variables. We develop a new test for the assessment of similarity of two groups for a single binary endpoint. Our approach is based on estimating the maximal deviation between the curves describing the responses of the two groups, followed by a parametric bootstrap test. Further, using a two-dimensional Gumbel-type model we develop methodology to establish similarity for (correlated) binary efficacy–toxicity outcomes. We investigate the operating characteristics of the proposed methodology by means of a simulation study and present a case study as an illustration.

Keywords: Binary data; Bootstrap; Dose response; Gumbel model; Logistic regression.

1. Introduction

A common problem in clinical drug development is the assessment of an investigational drug in two groups of patients, such as different age classes, gender, geographic regions, or different treatment groups (see Jhee *and others*, 2004; Otto *and others*, 2008) (among many others for clinical examples). A natural question is then whether the effect of the investigational drug is consistent across populations. Considering data with covariates, the effect of the investigational drug is described as a function of the covariates, and

*To whom correspondence should be addressed.

consistency across the two groups is claimed if these functions are similar in a suitable sense. Several authors use confidence bands for the difference between the response functions to construct such tests (see e.g., Liu *and others*, 2009; Gsteiger *and others*, 2011; Bretz *and others*, 2018). Alternatively, Dette *and others* (2018) and Möllenhoff *and others* (2018) propose more powerful tests by estimating a distance between the two functions, such as the squared integral of the difference or the maximal deviation between the functions. They claim similarity if the estimated distance is small.

All above approaches assume a single, continuous outcome. However, there are many situations where the efficacy of the drug to be investigated is defined by a binary outcome, such as tumor shrinkage or complete cure. In addition, many clinical trials involve the measurement of a second binary endpoint to assess the toxicity of the investigational drug, such as the occurrence of adverse events like fatigue or nausea. Hence the need arises to assess bivariate efficacy–toxicity outcomes which are likely to be correlated. Several methods for modeling multivariate binary outcomes have been proposed in the literature (see e.g., Glonek and McCullagh, 1995). Considering efficacy–toxicity outcomes, Murtaugh and Fisher (1990) and Heise and Myers (1996) investigate bivariate binary responses and derive optimal designs by fitting a bivariate logistic model and a Cox model to the data. Deldossi *and others* (2019) propose copulas to model the marginals and dependence structure of the outcomes separately. Further, Dragalin and Fedorov (2006) and Gaydos *and others* (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 (Zhang *and others*, 2006; Yin *and others*, 2006).

Different to the literature reviewed above, we investigate statistical tests to assess similarity of binary efficacy and toxicity responses for two groups of patients. Similarity can be claimed if the differences of both outcomes are below prespecified margins over the complete range of covariates. Accordingly, we first develop a new test of similarity for a single binary outcome. Second, we address similarity for bivariate binary (correlated) outcomes and develop a test for comparing simultaneously efficacy and toxicity outcomes among two populations. For this purpose, we use a two-dimensional Gumbel model (Gumbel, 1961) for bivariate logistic regression to model correlated bivariate binary endpoints. Our approach is based on a parametric bootstrap, which generates data under the constraint that the distances between the curves are precisely equal to the prespecified margins. We investigate finite sample properties and illustrate the procedures with a clinical trial example.

In a recent publication, Möllenhoff *and others* (2020) investigate the situation where some of the parameters of the models used to describe the dose-response relationship coincide (e.g., the placebo effect). They investigate continuous endpoints from two populations, propose a parametric bootstrap and demonstrate that using such additional information leads to more efficient statistical inference. In order to apply the methods proposed in this article to similar settings with additional toxicity outcomes, we will also extend our methodology to models with shared parameters and illustrate their use with the aforementioned clinical trial example.

2. COMPARING CURVES FOR BINARY OUTCOMES

In this section, we introduce a model-based approach for comparing the responses between two groups assuming binary outcomes. We consider models with covariates and assume for simplicity a one-dimensional covariate, although the proposed methodology applies more broadly. For both groups, we choose the covariate space as a dose range $\mathcal{D}$ and assume that the groups are investigated at k_ℓ dose levels $d_{\ell,1}, \ldots, d_{\ell,k_\ell}$, $\ell = 1, 2$, where the index $\ell = 1, 2$ is the group indicator. More precisely the dose range is given by all dose levels between the lowest and highest dose across both two groups, that is $\mathcal{D} = \left[\min_{\ell=1,2} d_{\ell,1}, \max_{\ell=1,2} d_{\ell,k_\ell}\right]$. Often $\min_{\ell=1,2} d_{\ell,1} = 0$ is the placebo or zero-dose control group. The highest dose level $\max_{\ell=1,2} d_{\ell,k_\ell}$ is often determined in earlier trials investigating the tolerability or safety

of a compound. Note that clinical trials typically randomize patients to a few fixed dose levels, which have to be determined in advance, such that k_ℓ is often in the range of 4–6 (Bretz *and others*, 2005).

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 in group ℓ. We use the indicators $Y_{\ell,i,j} = 1$, if a patient responds to the treatment and $Y_{\ell,i,j} = 0$ otherwise. Therefore $Y_{\ell,i,j}$ follows a Bernoulli distribution with parameter $p_\ell(d_{\ell,i})$ modeling the probability of success in group ℓ with dose level $d_{\ell,i}$, $i = 1, \ldots, k_\ell$, $\ell = 1, 2$. The response probability of the jth patient allocated to dose level $d_{\ell,i}$ in group ℓ is given by the monotone function

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

where η_ℓ^E is a known distribution function determined by the parameters β_ℓ, γ_ℓ. Hence, the function $\eta_\ell^E(d, \beta_\ell, \gamma_\ell)$ models the response probability over the entire dose range. Note that model (2.1) uses subscripts ℓ and i on the dose in order to be consistent with the typical clinical trial setting of randomizing patients to a few fixed dose levels, as discussed above.

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

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

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

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

and corresponding maximum likelihood estimates (MLE) are obtained by maximizing the function (2.2). In order to investigate the difference in efficacy between the two groups we consider the maximal deviation between the two curves in (2.1) and test the hypotheses

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

where ϵ^E denotes a prespecified margin measuring the degree of similarity for efficacy. The latter has to be carefully chosen in advance by clinical experts and depends on the application.

The following algorithm provides a bootstrap test for the hypotheses (2.3). It is derived by adapting the methodology developed in Dette *and others* (2018) to binary data.

Algorithm 2.1 (parametric bootstrap for testing similarity 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

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

(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

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

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

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

We discretize the dose range $\mathcal{D}$ to get a feasible optimization problem by fixing r nodes $d_1, \ldots, d_r$ and using a smooth approximation of the maximum,

$$\max(d_1, \ldots, d_r) \approx \lambda \log \sum_{i=1}^{r} \exp \tfrac{d_i}{\lambda} \text{ for } \lambda \to 0,$$

in (2.4). We solve the constrained optimization problem by using the augmented Lagrangian minimization algorithm as implemented with the auglag() function in the R package alabama (Varadhan, 2014).

(3) Proceed as follows:

(i) Generate bootstrap data under the null hypothesis (2.3) by creating binary data specified by the parameters $\left(\hat{\bar{\beta}}_\ell, \hat{\bar{\gamma}}_\ell\right)$, $\ell = 1, 2$. More precisely, calculate $\eta_\ell^E(d_{\ell,i}, \hat{\bar{\beta}}_\ell, \hat{\bar{\gamma}}_\ell)$, $i = 1, \ldots, k_\ell$, $\ell = 1, 2$ yielding the probabilities $p(d_{\ell,i})$ 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

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

(iii) Repeat the steps (i) and (ii) n_{boot} times to generate replicates $\hat{\Delta}_1^{E*}, \ldots, \hat{\Delta}_{n_{boot}}^{E*}$ of $\hat{\Delta}^{E*}$. Let $\hat{\Delta}_{(1)}^{E*} \leq \ldots \leq \hat{\Delta}_{(n_{boot})}^{E*}$ denote the corresponding order statistic. The estimator of the α-quantile of the distribution of $\hat{\Delta}^*$ is defined by $\hat{\Delta}_{(\lfloor n_{boot}\alpha \rfloor)}^{E*}$. Reject the null hypothesis (2.3) and decide for similarity if

$$\hat{\Delta}^E < \hat{\Delta}_{(\lfloor n_{boot}\alpha \rfloor)}^{E*}. \tag{2.6}$$

Alternatively, calculate the p-value $\hat{F}_{n_{boot}}^E(\hat{\Delta}^E) = \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} I\{\hat{\Delta}_i^{E*} \leq \hat{\Delta}^E\}$ and reject the null hypothesis (2.3) if $\hat{F}_{n_{boot}}^E(\hat{\Delta}^E) < \alpha$ for a prespecified significance level α, where $\hat{F}_{n_{boot}}^E$ denotes the empirical distribution function of the bootstrap sample.

Both, the bootstrap quantile $\hat{\Delta}^{E*}_{(\lfloor n_{boot}\alpha \rfloor)}$ and the p-value $\hat{F}^{E}_{n_{boot}}(\hat{\Delta}^{E})$, depend on the number of bootstrap replicates n_{boot} and the margin ε^E given in the hypotheses (2.3), but we do not reflect the latter dependence in our notation.

The test proposed in Algorithm 2.1 has asymptotic level α and is consistent. More precisely, $\hat{\Delta}^{E*}_{(\lfloor n_{boot}\alpha \rfloor)} \to \hat{q}_\alpha$ 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

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

and that under H_1^E

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

These results follow from the well-known fact that under suitable regularity conditions the MLE converges weakly to a normal distribution (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's information matrix corresponding to group ℓ. The weak convergence (2.9) is the essential ingredient to apply the proof of Dette *and others* (2018) to the situation considered in this article and (2.7) and (2.8) follow. These arguments provide the validity of the test (2.6) for large sample sizes. For moderate sample sizes the quality of the approximation depends on the model under consideration (including the parameters), see Section 4 for a numerical investigation.

3. Tests for similarity of efficacy–toxicity responses

3.1. *The Gumbel model for efficacy–toxicity outcomes*

We now extend the approach of Section 2 to correlated bivariate binary outcomes. We consider the bivariate Gumbel model (see e.g., Murtaugh and Fisher, 1990; Heise and Myers, 1996) based on the bivariate logistic function derived by Gumbel (1961),

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

Note that the marginal distributions are logistic and that the parameter ν represents the dependence of U and V, where $\nu = 0$ corresponds to independent margins and in this case, two separate logistic models for efficacy and toxicity can be fitted separately to the data. The approach in this article can also be applied to other parametric two-dimensional distributions. In the Supplementary material available at *Biostatistics* online, we investigate an alternative distribution with logistic margins and a different dependence structure.

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

$$
\begin{aligned}
p_{00}(d) &:= \mathbb{P}(Y^E = 0, Y^T = 0 \mid d) = 1 - \tfrac{1}{1+\mathrm{e}^{-u_1(d)}} - \tfrac{1}{1+\mathrm{e}^{-u_2(d)}} + \tfrac{1}{1+\mathrm{e}^{-u_1(d)}}\tfrac{1}{1+\mathrm{e}^{-u_2(d)}} \\
&\qquad + \frac{\nu \mathrm{e}^{-u_1(d)-u_2(d)}}{(1+\mathrm{e}^{-u_1(d)})^2(1+\mathrm{e}^{-u_2(d)})^2}, \\
p_{01}(d) &:= \mathbb{P}(Y^E = 0, Y^T = 1 \mid d) = \tfrac{1}{1+\mathrm{e}^{-u_2(d)}} - \tfrac{1}{1+\mathrm{e}^{-u_1(d)}}\tfrac{1}{1+\mathrm{e}^{-u_2(d)}} - \tfrac{\nu \mathrm{e}^{-u_1(d)-u_2(d)}}{(1+\mathrm{e}^{-u_1(d)})^2(1+\mathrm{e}^{-u_2(d)})^2}, \\
p_{10}(d) &:= \mathbb{P}(Y^E = 1, Y^T = 0 \mid d) = \tfrac{1}{1+\mathrm{e}^{-u_1(d)}} - \tfrac{1}{1+\mathrm{e}^{-u_1(d)}}\tfrac{1}{1+\mathrm{e}^{-u_2(d)}} - \tfrac{\nu \mathrm{e}^{-u_1(d)-u_2(d)}}{(1+\mathrm{e}^{-u_1(d)})^2(1+\mathrm{e}^{-u_2(d)})^2}, \\
p_{11}(d) &:= \mathbb{P}(Y^E = 1, Y^T = 1 \mid d) = \tfrac{1}{1+\mathrm{e}^{-u_1(d)}}\tfrac{1}{1+\mathrm{e}^{-u_2(d)}} + \tfrac{\nu \mathrm{e}^{-u_1(d)-u_2(d)}}{(1+\mathrm{e}^{-u_1(d)})^2(1+\mathrm{e}^{-u_2(d)})^2}.
\end{aligned} \tag{3.2}
$$

Here, $u_1(d) = \beta_1 + \gamma_1 d$ and $u_2(d) = \beta_2 + \gamma_2 d$ denote the transformed doses for efficacy and toxicity, respectively (see Heise and Myers, 1996). Consequently, the Gumbel model is determined by the five-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

$$
\begin{aligned}
\eta^E(d,\theta) &:= \mathbb{P}(Y^E = 1 \mid d) = p_{11}(d) + p_{10}(d) = \frac{1}{1+\mathrm{e}^{-u_1(d)}}, \\
\eta^T(d,\theta) &:= \mathbb{P}(Y^T = 1 \mid d) = p_{11}(d) + p_{01}(d) = \frac{1}{1+\mathrm{e}^{-u_2(d)}}.
\end{aligned} \tag{3.3}
$$

For simplicity we do not display the dependence on θ in the cell probability functions (3.2). We further denote by $\eta(d,\theta) := \big(\eta^E(d,\theta), \eta^T(d,\theta)\big)$ the vector of bivariate response probabilities at dose d. Note that the correlation parameter ν is part of the model but not displayed explicitly. In order to guarantee that all cell probabilities in (3.2) lie between 0 and 1 for all doses $d \in \mathcal{D}$, the lower bound on ν is always -1, whereas the upper bound depends on the other model parameters $\beta_1, \gamma_1, \beta_2$, and γ_2 and is at most 4. As derived by Murtaugh and Fisher (1990), the correlation of Y^E and Y^T is given by

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

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

We now compare the two groups with respect to their efficacy and toxicity outcomes. 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 group ℓ and define by $z^{\ell,i}_{pq} := \sum_{j=1}^{n_{\ell,i}} I\{(y^E_{\ell ij}, y^T_{\ell ij}) = (p,q)\}$ 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 Gumbel model from Section 3.1.

According to (3.5) the likelihood of the Gumbel model for group ℓ is given by

$$\begin{aligned}
&\mathcal{L}_\ell(\theta_\ell | y_{\ell,1,1}, \ldots, y_{\ell,1,n_{\ell,1}}, \ldots, y_{\ell,k_\ell,n_{\ell,k_\ell}}) \\
&\quad = \prod_{i=1}^{k_\ell} \prod_{j=1}^{n_{\ell,i}} p_{11}(d_{\ell,i})^{y^E_{\ell ij} y^T_{\ell ij}} p_{01}(d_{\ell,i})^{(1-y^E_{\ell ij}) y^T_{\ell ij}} p_{10}(d_{\ell,i})^{y^E_{\ell ij}(1-y^T_{\ell ij})} p_{00}(d_{\ell,i})^{(1-y^E_{\ell ij})(1-y^T_{\ell ij})} \\
&\quad = \prod_{i=1}^{k_\ell} p_{11}(d_{\ell,i})^{z^{\ell,i}_{11}} p_{01}(d_{\ell,i})^{z^{\ell,i}_{10}} p_{10}(d_{\ell,i})^{z^{\ell,i}_{10}} p_{00}(d_{\ell,i})^{z^{\ell,i}_{00}}.
\end{aligned}$$

Taking the logarithm yields

$$\begin{aligned}
l_\ell(\theta_\ell) &:= \log \mathcal{L}_\ell(\theta_\ell | y_{\ell,1,1}, \ldots, y_{\ell,1,n_{\ell,1}}, \ldots, y_{\ell,k_\ell,n_{\ell,k_\ell}}) \\
&= \sum_{i=1}^{k_\ell} z^{\ell,i}_{11} \log p_{11}(d_{\ell,i}) + z^{\ell,i}_{01} \log p_{01}(d_{\ell,i}) + z^{\ell,i}_{10} \log p_{10}(d_{\ell,i}) + z^{\ell,i}_{00} \log p_{00}(d_{\ell,i}) \qquad (3.6)
\end{aligned}$$

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 in the case of independence, that is $\nu_\ell = 0$, the parameter 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).

Let

$$\eta_\ell(d, \theta_\ell) = \big(\eta^E_\ell(d, \theta_\ell), \eta^T_\ell(d, \theta_\ell)\big) = \Big(\frac{1}{1 + \mathrm{e}^{-\beta_{\ell,1} - \gamma_{\ell,1} d}}, \frac{1}{1 + \mathrm{e}^{-\beta_{\ell,2} - \gamma_{\ell,2} d}}\Big)^T$$

denote the vector of efficacy and toxicity curves for group $\ell = 1, 2$. Claiming similarity of both groups, we want to ensure that the efficacy and toxicity responses do not deviate by more than a prespecified margin $\epsilon = (\epsilon^E, \epsilon^T)$. Consequently, we test the global null hypothesis

$$H_0 : \max_{d \in \mathcal{D}} \big|\eta^E_1(d, \theta_1) - \eta^E_2(d, \theta_2)\big| \geq \epsilon^E \text{ or } \max_{d \in \mathcal{D}} \big|\eta^T_1(d, \theta_1) - \eta^T_2(d, \theta_2)\big| \geq \epsilon^T \qquad (3.7)$$

against the alternative

$$H_1 : \max_{d \in \mathcal{D}} \big|\eta^E_1(d, \theta_1) - \eta^E_2(d, \theta_2)\big| < \epsilon^E \text{ and } \max_{d \in \mathcal{D}} \big|\eta^T_1(d, \theta_1) - \eta^T_2(d, \theta_2)\big| < \epsilon^T. \qquad (3.8)$$

This problem can be solved by simultaneously testing the individual hypotheses

$$H^E_0 : \max_{d \in \mathcal{D}} \big|\eta^E_1(d, \theta_1) - \eta^E_2(d, \theta_2)\big| \geq \epsilon^E \text{ vs. } H^E_1 : \max_{d \in \mathcal{D}} \big|\eta^E_1(d, \theta_1) - \eta^E_2(d, \theta_2)\big| < \epsilon^E \qquad (3.9)$$

and

$$H^T_0 : \max_{d \in \mathcal{D}} \big|\eta^T_1(d, \theta_1) - \eta^T_2(d, \theta_2)\big| \geq \epsilon^T \text{ vs. } H^T_1 : \max_{d \in \mathcal{D}} \big|\eta^T_1(d, \theta_1) - \eta^T_2(d, \theta_2)\big| < \epsilon^T. \qquad (3.10)$$

As the global null in (3.7) is the union of H^E_0 and H^T_0 we can apply the intersection union principle (Berger, 1982). That is, we reject the global null in (3.7) and claim similarity only if both individual null hypotheses in (3.9) and (3.10) are rejected. Each of the two individual tests in (3.9) and (3.10) is performed by extending the parametric bootstrap approach in Algorithm 2.1, as described below.

Algorithm 3.1 (parametric bootstrap for testing for similarity of 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

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

and

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

(2) For each individual test in (3.9) and (3.10) we perform a constrained optimization as described in Algorithm 2.1, yielding estimates $\hat{\tilde{\theta}}_\ell$, $\ell = 1, 2$. This procedure is done separately for each individual test because the constraints differ. Although the constraints are only imposed on the marginal densities which do not contain the dependence parameters ν_ℓ, they appear in the likelihood function to be maximized under the constraints. Consequently the constrained estimates of the parameters ν_ℓ are usually different from the unconstrained estimates. We generate bootstrap data for each individual test separately and obtain replicates $\hat{\Delta}_1^{E*}, \ldots, \hat{\Delta}_{n_{boot}}^{E*}$ for the comparison of the efficacy curves and $\hat{\Delta}_1^{T*}, \ldots, \hat{\Delta}_{n_{boot}}^{T*}$ for the comparison of the toxicity curves. Let $\hat{\Delta}_{(1)}^{E*} \leq \ldots \leq \hat{\Delta}_{(n_{boot})}^{E*}$ and $\hat{\Delta}_{(1)}^{T*} \leq \ldots \leq \hat{\Delta}_{(n_{boot})}^{T*}$ denote the corresponding order statistics and let $\hat{\Delta}_{(\lfloor n_{boot}\alpha \rfloor)}^{E*}$ and $\hat{\Delta}_{(\lfloor n_{boot}\alpha \rfloor)}^{T*}$ denote the corresponding empirical level α quantiles, respectively.

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

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

We do not need to adjust the level of the two individual tests and can thus use the α-quantile according to the intersection union principle. 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. Several approaches have been proposed in the literature, such as the inversion method (see Devroye, 1986), which is rather simple but comes along with computational disadvantages (for details see e.g., Leisch *and others*, 1998). Here, we use the algorithm of Emrich and Piedmonte (1991), as implemented with the function generate.binary in the R package MultiOrd (Amatya and Demirtas, 2015). For this purpose, we calculate the correlation (3.4) 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

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

and

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

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 $\hat{\tilde{\theta}}_1$ and $\hat{\tilde{\theta}}_2$ generate a distribution and bootstrap data can be obtained.

3.4. *Shared parameters*

As pointed out in the introduction there exist also situations, where it is reasonable to assume that certain model parameters are the same for both groups. In such cases, the total number of parameters to be estimated is reduced, which yields to more efficient inference if the assumption is correct. For example, Möllenhoff *and others* (2020) describe a trial assessing similarity of Japanese and Caucasian patients, where a similar response to placebo and a common maximum treatment effect is assumed.

The new methodology can be further developed to address this situation by considering a joint likelihood function instead of fitting two separate models. For this purpose, we adopt Algorithm 3.1 to that situation as follows. Let $\theta = (\theta_0, \tilde{\theta}_1, \tilde{\theta}_2)$, where θ_0 denotes the vector of common parameters and $\tilde{\theta}_1, \tilde{\theta}_2$ denote the remaining parameters of the Gumbel models, such that $\theta_\ell = (\theta_0, \tilde{\theta}_\ell)$, $\ell = 1, 2$. We then estimate an MLE $\hat{\theta}$ by using the combined sample and maximizing $l_1(\theta_0, \tilde{\theta}_1) + l_2(\theta_0, \tilde{\theta}_2)$, where l_1 and l_2 are the log-likelihood functions given in (3.6). The calculation of the test statistic, the constrained optimization, and the generation of bootstrap data described in step (2) of Algorithm 3.1 are performed similarly, now using joint likelihood functions instead of fitting two separate models throughout. The details are omitted for the sake of brevity.

4. Finite sample properties

We now investigate the finite sample properties of the two tests based on Algorithms 2.1 and 3.1. We consider the dose range $\mathcal{D} = [0, 2]$ with the seven dose levels $0, 0.1, 0.2, 0.5, 1, 1.5$, and 2. We assume $n_{\ell,i} = 7,\ 14,\ 21,\ 28,\ 50$ patients per dose level, $i = 1, \ldots, 7$ and group $\ell = 1, 2$, resulting in $n_\ell = 49, 98, 147, 196, 350$, $\ell = 1, 2$. The significance level is $\alpha = 0.05$ throughout and we consider three different margins in (3.7) and (3.8), that is 0.1, 0.15, and 0.2. All simulations are performed using 1000 simulation runs and $n_{boot} = 400$ bootstrap replications. Binary data are generated as described in Section 3.3. We set $\nu = 0$ for the univariate (efficacy) case. For the sake of brevity, we present in this section only the results for the bivariate case and a short summary of the findings for the univariate case. We refer to Supplementary Section S1 available at *Biostatistics* online for the complete simulation results and a more detailed discussion.

4.1. *Univariate efficacy outcomes*

Supplementary Table S1 and Table S2 available at *Biostatistics* online display the simulated Type I error rates and the power of the bootstrap test (2.6), respectively, for margins $\epsilon^E = 0.1, 0.15, 0.2$. 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, $\Delta^E = \epsilon^E$) and this accuracy increases with increasing sample sizes. Moreover, in the interior of the null hypothesis (that is, $\Delta^E > \epsilon^E$) the number of rejections is close to 0 in all scenarios, indicating that the Type I error rate is very small in these cases. We further conclude that the procedure has reasonable power for sufficiently large sample sizes. For example, the test achieves more than 80% power for sample sizes of 28 or 50 patients per dose level, depending on the margin under consideration. We also observe that the power increases with increasing sample sizes.

4.2. *Bivariate efficacy–toxicity outcomes*

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

$$\theta_1 = (\beta_{1,1}, \gamma_{1,1}, \beta_{1,2}, \gamma_{1,2}, \nu_1) = (-1, 2, -3, 3, \nu_1) \tag{4.1}$$

and we assume two different levels of dependence. The first setting represents a moderate correlation between the efficacy and toxicity outcomes ($\nu_1 = 1$). In the second setting, we fix ν_1 to the maximum value such that all cell probabilities in (3.2) with regard to the model parameter θ_1 are still between 0 and 1, that is $\nu_1 = 2.4$. According to (3.4), the correlation of Y_1^E and Y_1^T at dose $d \in \mathcal{D}$ is given by

$$corr(Y_1^E, Y_1^T | \, d) = \frac{\nu_1}{(\mathrm{e}^{-0.5+d} + \mathrm{e}^{0.5-d})(\mathrm{e}^{-1.5+1.5d} + \mathrm{e}^{1.5-1.5d})}, \tag{4.2}$$

which ranges from 0.09 to 0.23 for $\nu_1 = 1$ and from 0.23 to 0.55 for $\nu_1 = 2.4$. Figure 1(a) displays the probability of efficacy without toxicity response, $\mathbb{P}(Y^E = 1, Y^T = 0 | \, d) = p_{10}(d)$. Figure 1(b) displays the correlation for three different choices of ν in dependence of the dose. In order to investigate the performance 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 1. 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 1(c).

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 2 and 3, 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 margins $\epsilon = (\epsilon^E, \epsilon^T) = (0.15, 0.15)$ and $(0.2, 0.2)$. 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 $n_{\ell,i} = 14$, $\nu_1 = \nu_2 = 1$ and $\Delta = (\Delta^E, \Delta^T) = \epsilon = (0.2, 0.2)$ the individual proportions of rejection are 0.046 for efficacy and 0.058 for toxicity, whereas the Type I error rate for the global test is 0.001, which is well below the nominal level. This is a common feature of the intersection union principle for the problem of testing equivalence in multivariate responses (Berger and Hsu, 1996). A similar conclusion holds for the high level of dependence, that is $\nu_1 = \nu_2 = 2.4$. Considering the same configuration as above, that is $n_{\ell,i} = 14$ and $\Delta = \epsilon = (0.2, 0.2)$, the individual proportions of rejection are 0.088 for efficacy and 0.089 for toxicity, whereas the Type I error rate for the global test is 0.002.

In general, we conclude that for a low level of dependence the individual tests on efficacy and toxicity yield rejection probabilities that are close to 0.05 when simulating on the margin of the global null hypothesis (that is $\Delta = \epsilon$) and hence the global Type I error rates are well below α in these cases. However, for a high level of correlation, that is $\nu_1 = \nu_2 = 2.4$, there are a few scenarios where the Type I error rate is too large. For instance, we observe the largest proportion of rejections of the global null hypothesis given by 0.127 for $n_{\ell,i} = 50$, $\epsilon = (0.2, 0.2)$ and $\Delta = (0, 0.2)$. Considering the same configurations but $\epsilon = (0.15, 0.15)$, yields a proportion of 0.089, which is lower but still above the desired value of 0.05. For all other scenarios, the Type I error of the global test is well below the nominal level. The size of the parameter ν_ℓ affects the precision of the estimates for the parameter θ_ℓ of the Gumbel model, which explains the different results for the rather low correlations corresponding to $\nu_\ell = 1$ and the high correlations obtained for $\nu_\ell = 2.4$, $\ell = 1, 2$. In other words, a high correlation makes the estimation of the curves more difficult, even for large sample sizes. A more detailed discussion, including a table

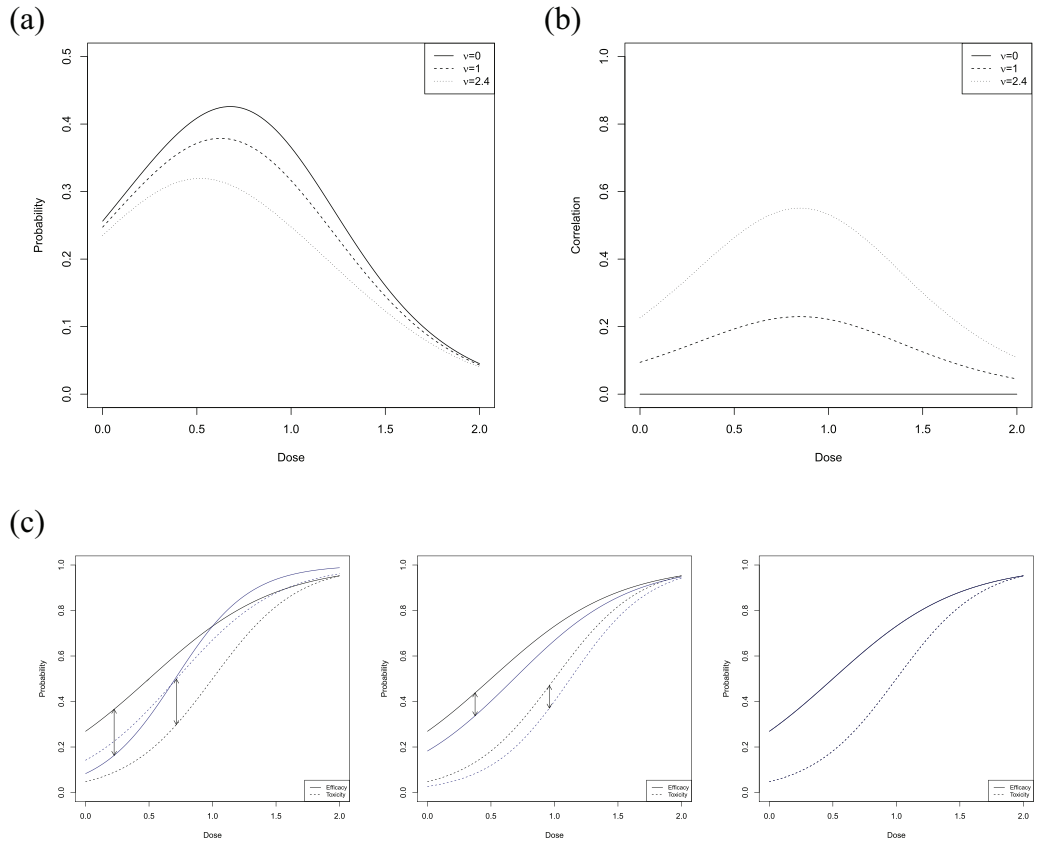

Fig. 1. (a) Probability $\mathbb{P}(Y^E = 1, Y^T = 0) = p_{10}(d)$ in dependence of the dose for the reference model (4.1) for different choices of the correlation parameter ν; (b) Correlation of efficacy and toxicity response for different choices of ν in dependence of the dose; (c) Efficacy curves (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 $\Delta^E = \Delta^T = 0.2,\ 0.1$ and 0 (from left to right).

Table 1. *Different scenarios corresponding to the null hypothesis* (3.7) *and the alternative* (3.8)

	θ_1	θ_2	$\Delta = (\Delta^E, \Delta^T)$
Alternative	$(-1, 2, -3, 3, \nu_1)$	$(-1, 2, -3, 3, \nu_2)$	$(0, 0)$
	$(-1, 2, -3, 3, \nu_1)$	$(-1.2, 2, -3.3, 3.1, \nu_2)$	$(0.05, 0.05)$
	$(-1, 2, -3, 3, \nu_1)$	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	$(0.1, 0.1)$
Null hypothesis	$(-1, 2, -3, 3, \nu_1)$	$(-2, 3.4, -2, 2.5, \nu_2)$	$(0.15, 0.15)$
	$(-1, 2, -3, 3, \nu_1)$	$(-1, 2, -2, 2.5, \nu_2)$	$(0, 0.15)$
	$(-1, 2, -3, 3, \nu_1)$	$(-2.4, 3.4, -1.8, 2.5, \nu_2)$	$(0.2, 0.2)$
	$(-1, 2, -3, 3, \nu_1)$	$(-1, 2, -1.8, 2.5, \nu_2)$	$(0, 0.2)$

Table 2. *Simulated Type I error rates of the global bootstrap test* (3.11) *for two different choices of* ν_ℓ, $\ell = 1, 2$

$\epsilon = (\epsilon^E, \epsilon^T)$	$n_{\ell,i}$	θ_2	$\Delta = (\Delta^E, \Delta^T)$	$\nu_\ell = 1$	$\nu_\ell = 2.4$
$(0.15, 0.15)$	7	$(-2, 3.4, -2, 2.5, \nu_2)$	$(0.15, 0.15)$	0.001 (0.063/0.074)	0.006 (0.078/0.064)
		$(-1, 2, -2, 2.5, \nu_2)$	$(0, 0.15)$	0.008 (0.122/0.060)	0.012 (0.112/0.075)
	14	$(-2, 3.4, -2, 2.5, \nu_2)$	$(0.15, 0.15)$	0.003 (0.040/0.047)	0.003 (0.082/0.065)
		$(-1, 2, -2, 2.5, \nu_2)$	$(0, 0.15)$	0.001 (0.207/0.052)	0.020 (0.230/0.068)
	21	$(-2, 3.4, -2, 2.5, \nu_2)$	$(0.15, 0.15)$	0.000 (0.026/0.046)	0.002 (0.051/0.057)
		$(-1, 2, -2, 2.5, \nu_2)$	$(0, 0.15)$	0.016 (0.325/0.041)	0.029 (0.326/0.084)
	28	$(-2, 3.4, -2, 2.5, \nu_2)$	$(0.15, 0.15)$	0.000 (0.049/0.053)	0.004 (0.125/0.090)
		$(-1, 2, -2, 2.5, \nu_2)$	$(0, 0.15)$	0.032 (0.476/0.058)	0.034 (0.455/0.076)
	50	$(-2, 3.4, -2, 2.5, \nu_2)$	$(0.15, 0.15)$	0.000 (0.035/0.078)	0.012 (0.210/0.084)
		$(-1, 2, -2, 2.5, \nu_2)$	$(0, 0.15)$	0.074 (0.827/0.085)	0.089 (0.815/0.111)
$(0.2, 0.2)$	7	$(-2.4, 3.4, -1.8, 2.5, \nu_2)$	$(0.2, 0.2)$	0.004 (0.061/0.063)	0.006 (0.091/0.101)
		$(-1, 2, -1.8, 2.5, \nu_2)$	$(0, 0.2)$	0.012 (0.218/0.055)	0.019 (0.233/0.084)
	14	$(-2.4, 3.4, -1.8, 2.5, \nu_2)$	$(0.2, 0.2)$	0.001 (0.046/0.058)	0.002 (0.088/0.089)
		$(-1, 2, -1.8, 2.5, \nu_2)$	$(0, 0.2)$	0.024 (0.396/0.067)	0.027 (0.442/0.065)
	21	$(-2.4, 3.4, -1.8, 2.5, \nu_2)$	$(0.2, 0.2)$	0.003 (0.048/0.070)	0.003 (0.090/0.087)
		$(-1, 2, -1.8, 2.5, \nu_2)$	$(0, 0.2)$	0.033 (0.672/0.051)	0.040 (0.648/0.070)
	28	$(-2.4, 3.4, -1.8, 2.5, \nu_2)$	$(0.2, 0.2)$	0.003 (0.069/0.072)	0.004 (0.124/0.077)
		$(-1, 2, -1.8, 2.5, \nu_2)$	$(0, 0.2)$	0.050 (0.813/0.065)	0.068 (0.870/0.078)
	50	$(-2.4, 3.4, -1.8, 2.5, \nu_2)$	$(0.2, 0.2)$	0.004 (0.054/0.076)	0.003 (0.145/0.103)
		$(-1, 2, -1.8, 2.5, \nu_2)$	$(0, 0.2)$	0.060 (0.982/0.061)	0.127 (0.986/0.132)

The numbers in brackets show the proportion of rejections for the individual tests according to the hypotheses (3.9) and (3.10).

presenting the bias of the estimates for some configurations, can be found in Supplementary Section S3 available at *Biostatistics* online.

The simulated power is shown in Table 3. It turns out that the global test achieves reasonable power for sufficiently large sample sizes. For example, a maximum power (always attained at $\Delta = (0, 0)$) of 0.933 is achieved for the global test for a choice of $n_{\ell,i} = 50$, $\nu_1 = \nu_2 = 2.4$, and $\epsilon = (0.2, 0.2)$. For a lower margin, that is, $\epsilon = (0.15, 0.15)$, the maximum power is smaller, but still increasing with growing sample sizes, reaching for instance 0.581 for $n_{\ell,i} = 50$ and $\nu_1 = \nu_2 = 2.4$. The same statement holds for a lower correlation of $\nu_\ell = 1$, $\ell = 1, 2$. For example, considering $n_{\ell,i} = 28$, $\nu_1 = \nu_2 = 1$ and $\epsilon = (0.2, 0.2)$, we observe a maximum power of 0.541.

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 outcome being measured within a few hours. It is 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, 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

Table 3. *Simulated power of the global bootstrap test* (3.11) *for two different choices of* ν_ℓ, $\ell = 1, 2$

$\epsilon = (\epsilon^E, \epsilon^T)$	$n_{\ell,i}$	θ_2	$\Delta = (\Delta^E, \Delta^T)$	$\nu_\ell = 1$	$\nu_\ell = 2.4$
(0.15, 0.15)	7	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.009 (0.092/0.125)	0.007 (0.089/0.125)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05,0.05)	0.009 (0.129/0.108)	0.010 (0.114/0.116)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.002 (0.128/0.133)	0.018 (0.153/0.121)
	14	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.008 (0.105/0.102)	0.014 (0.119/0.104)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.031 (0.176/0.146)	0.042 (0.183/0.172)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.035 (0.196/0.162)	0.045 (0.209/0.214)
	21	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.020 (0.145/0.150)	0.025 (0.145/0.155)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.051 (0.288/0.201)	0.075 (0.242/0.254)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.085 (0.345/0.265)	0.077 (0.309/0.269)
	28	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.038 (0.185/0.166)	0.057 (0.137/0.189)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.098 (0.387/0.266)	0.121 (0.356/0.313)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.201 (0.484/0.385)	0.202 (0.453/0.403)
	50	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.066 (0.295/0.263)	0.106 (0.239/0.234)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.318 (0.624/0.484)	0.326 (0.565/0.527)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.566 (0.842/0.656)	0.581 (0.827/0.686)
(0.2, 0.2)	7	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.018 (0.133/0.140)	0.029 (0.159/0.129)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05,0.05)	0.027 (0.159/0.151)	0.032 (0.213/0.155)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.026 (0.183/0.189)	0.049 (0.221/0.191)
	14	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.063 (0.277/0.210)	0.076 (0.278/0.230)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.112 (0.352/0.299)	0.099 (0.335/0.282)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.124 (0.409/0.300)	0.171 (0.451/0.356)
	21	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.119 (0.369/0.310)	0.142 (0.343/0.321)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.243 (0.585/0.388)	0.254 (0.527/0.416)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.328 (0.658/0.505)	0.322 (0.593/0.536)
	28	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.177 (0.468/0.348)	0.212 (0.429/0.418)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.445 (0.716/0.608)	0.472 (0.688/0.622)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.541 (0.816/0.660)	0.581 (0.822/0.705)
	50	$(-1.5, 2.2, -3.6, 3.2, \nu_2)$	(0.1, 0.1)	0.404 (0.717/0.543)	0.437 (0.653/0.602)
		$(-1.2, 2, -3.3, 3.1, \nu_2)$	(0.05, 0.05)	0.740 (0.933/0.783)	0.765 (0.897/0.836)
		$(-1, 2, -3, 3, \nu_2)$	(0, 0)	0.900 (0.987/0.914)	0.933 (0.985/0.945)

The numbers in brackets show the proportion of rejections for the individual tests according to the hypotheses (3.9) and (3.10).

and a known dose-response relationship for tolerability are available, the objective of the study at hand was to demonstrate similarity with a marketed product for the bivariate efficacy–toxicity outcome in a proof of concept setting.

This was a randomized double-blind multi-regional parallel group clinical 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 investigational drug) 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.

This trial included half of the patients each from Western and Eastern Europe. Prior investigations suggested that the differences across both geographic regions are negligible. We thus compare the efficacy and toxicity data of the 150 patients randomized to the marketed drug across both regions. For this purpose, we fit two Gumbel models as defined in Section 3.1 to the data, one for the 75 patients from

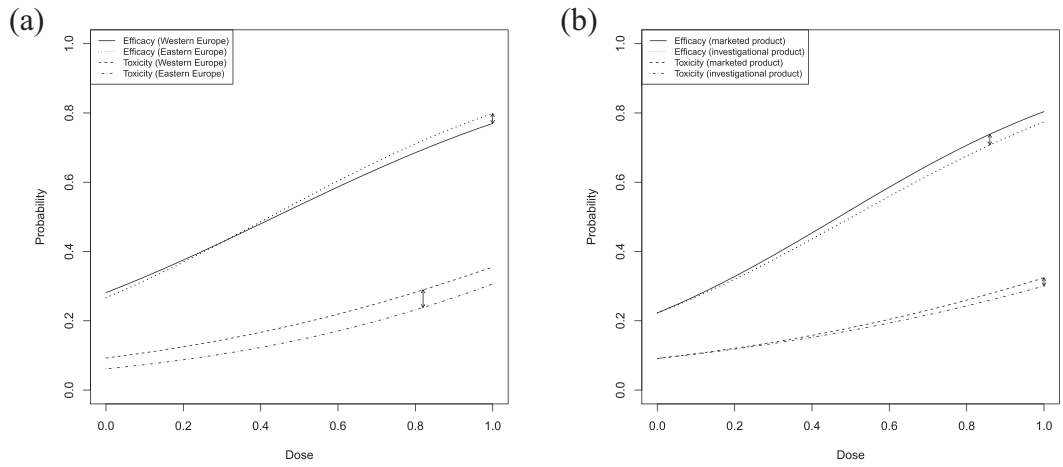

Fig. 2. (a) Efficacy and toxicity curves corresponding to the fitted Gumbel models (5.1) for the hypothetical data described in Section 5. The solid (efficacy) and the dashed line (toxicity) correspond to the patients from Western Europe, the dotted (efficacy) and the dotted-dashed (toxicity) to those from Eastern Europe. (b) Efficacy and toxicity curves under the assumption of shared placebo parameters. Here the solid (efficacy) and the dashed line (toxicity) correspond to the marketed product, the dotted (efficacy) and the dotted-dashed (toxicity) to the investigational drug, respectively. The arrows indicate the maximum absolute distances.

Western Europe ($\ell = 1$) and one for the 75 patients from Eastern Europe ($\ell = 2$). The estimated model parameters are

$$\hat{\theta}_1 = (-0.938, 2.145, -2.284, 1.689, 0.498),\ \hat{\theta}_2 = (-1.012, 2.388, -2.728, 1.910, -0.475), \quad (5.1)$$

see Figure 2(a) for the corresponding efficacy and toxicity curves. The maximum distances are $\hat{\Delta}^E = 0.029$ and $\hat{\Delta}^T = 0.051$, attained at the maximum dose 1 and the dose 0.82, respectively. We perform a test (3.11) for similarity in Algorithm 3.1 (significance level of $\alpha = 0.05$) choosing the margin $\epsilon = (0.2, 0.2)$, which means that we allow the responses between populations to differ about 20%.

Using $n_{boot} = 1000$ bootstrap replications, we obtain critical values $q^E_{0.05} = 0.061$ and $q^T_{0.05} = 0.056$ and test the global null hypothesis (3.7) against the alternative (3.8). Since $\hat{\Delta}^E = 0.029 < 0.061 = \hat{q}^E_{0.05}$ and $\hat{\Delta}^T = 0.051 < 0.056 = \hat{q}^T_{0.05}$, we can reject (3.7) at level $\alpha = 0.05$. This conclusion can also be drawn by directly considering the p-values 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. Since the individual p-values are given by $\hat{F}^E_{n_{boot}}(\hat{\Delta}^E) = 0.015$ and $\hat{F}^T_{n_{boot}}(\hat{\Delta}^T) = 0.042$, we have $\max(0.015, 0.042) = 0.042 < 0.05 = \alpha$ and can reject the null hypothesis (3.7), thus concluding similarity of efficacy and toxicity across the two geographic regions.

Based on this result, it is therefore reasonable to proceed with a further analysis of this trial using the data pooled from both regions. We now compare the investigational drug with the marketed product across all dose levels for the bivariate efficacy–toxicity outcomes of all 300 patients randomized into the study. For this analysis, it is reasonable to assume that the placebo effect is the same for both products with regard to efficacy and toxicity and to investigate the question of similarity under the assumption of shared placebo parameters as described in Section 3.4. More precisely we assume that $\beta_{1,1} = \beta_{2,1}$ and $\beta_{1,2} = \beta_{2,2}$, which reduces the number of parameters to be estimated from 10 to 8. The parameter estimates are $\hat{\theta}_1 = (-1.250, 2.661, -2.299, 1.564, -0.066)$ and $\hat{\theta}_2 = (-1.250, 2.481, -2.299, 1.453, 0.941)$, see Figure 2(b)

for the corresponding efficacy and toxicity curves. The maximum distances are now given by $\hat{\Delta}^E = 0.031$ and $\hat{\Delta}^T = 0.024$, attained at the dose levels 0.86 and 1, respectively. We perform the test at a significance level of $\alpha = 0.05$ and generate bootstrap data under the assumption of common placebo parameters. The critical values are now given by $q^E_{0.05} = 0.060$ and $q^T_{0.05} = 0.035$ and hence we conclude that the null hypothesis (3.7) can be rejected as $\hat{\Delta}^E = 0.031 < \hat{q}^E_{0.05} = 0.06$ and $\hat{\Delta}^T = 0.024 < \hat{q}^T_{0.05} = 0.035$. The p-values are given by $\hat{F}^E_{n_{boot}}(\hat{\Delta}^E) = 0.021$ and $\hat{F}^T_{n_{boot}}(\hat{\Delta}^T) = 0.031$, respectively.

Finally, we note that fitting separate models as shown above also implies that the dependence parameter is allowed to differ between the two drugs. Such an approach seems sensible in practice as it would be hard to justify clinically that the dependence parameter is the same, unless the two products are from the same chemical class or have a common mode of action. If for a given problem at hand it can be argued in favor of a shared dependence parameter then the methods in this article can be extended following Möllenhoff *and others* (2020).

6. Conclusions and discussion

In the first part of this article, we investigated a single efficacy response given by a binary outcome and derived a test procedure for the similarity of the corresponding dose-response curves, which can be modeled, for instance, by a parametric logistic regression or a probit model. We developed a parametric bootstrap test and decide for similarity 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 article, we investigated the operating characteristics by means of an extensive simulation study. The choice of the margin ϵ measuring the degree of similarity 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. We demonstrated that the resulting procedures control their level in most of the configurations and achieve reasonable power. However, for a high level of dependence between the efficacy and the toxicity outcome we observed a slight inflation of the Type I error in some few scenarios. This can be explained by the difficulty in estimating the model parameters with sufficient precision for large correlations: Increasing correlations severely increases the bias of the estimates and hence affects the performance of the test. We provide a more detailed discussion in the Supplementary Material available at *Biostatistics* online.

In this article, we used a Gumbel-type copula to model the dependency of bivariate binary outcomes. In the Supplementary Material available at *Biostatistics* online, we demonstrate that the methodology is easily applicable to other copula models. Moreover, we also investigate the sensitivity of our approach with respect to the choice of the copula by means of a simulation study and demonstrate that the approach is remarkably robust. A heuristic explanation for this observation is that parametric copula models provide some flexibility for modeling different dependencies by choosing different parameters. Therefore, a given dependency structure can often be reasonably well modeled by different copula models choosing appropriate parameters. A similar observation was also made in Dette *and others* (2014) in the context of copula-based regression models.

The methods proposed in this article are broadly applicable whenever binary efficacy and toxicity responses are compared. These groups can be, for example, different populations or treatments. The methodology can also be extended to models with shared parameters, such as a common placebo effect. A standard application for the latter is the comparison of a new with an old formulation in the development of a generic product because doses are immediately comparable. Our approach is different to the standard bioequivalence assessment based on pharmacokinetic (PK) parameters, such as the area under the curve

or the maximum concentration. One reviewer argued that the PK is often linear in dose meaning that a factor on the "vertical" outcome scale can be translated to a factor on the "horizontal" dose scale and this implies that two dilutions of the same drug can only be bioequivalent if the concentrations are very close to each other. With the suggested approach in this article, equivalence, and therefore similarity, is based on small absolute differences on the "vertical scale" (recall (3.7)). This means that drugs are similar if the dose range only covers low doses or, as an alternative formulation, a low dose of a drug is similar to placebo in this metric. In clinical applications, however, the dose range should be chosen sufficient large (including high doses) such that a relevant difference to placebo can be detected.

In some settings, the efficacy or toxicity responses are not modeled 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 article 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. Modeling these types of responses is still a challenging problem. Tao *and others* (2013) investigated this situation by modeling these multiple endpoints by a joint model constructed with archimedean copula. A test approach corresponding to 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.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors would like to thank two referees and the Associate Editor for their useful suggestions which greatly improved the manuscript. *Conflict of Interest*: None declared.

Funding

The authors gratefully acknowledge financial support by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt 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**, 1683–1691.

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

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

Bradley, R. and Gart, J. (1962). The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika* **49**, 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**, 722–738.

BRETZ, F., PINHEIRO, J. C. AND BRANSON, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61**, 738–748.

DELDOSSI, L., OSMETTI, S. A. AND TOMMASI, C. (2019). Optimal design to discriminate between rival copula models for a bivariate binary response. *TEST* **28**, 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.

DETTE, H., VAN HECKE, R. AND VOLGUSHEV, S. (2014). Some comments on copula-based regression. *Journal of the American Statistical Association* **109**, 1319–1324.

DEVROYE, L. (1986). Sample-based non-uniform random variate generation. In: *Proceedings of the 18th Conference on Winter simulation*. Washington, DC, USA, pp. 260–265.

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

EMRICH, L. AND PIEDMONTE, M. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45**, 302–304.

FEDOROV, V. AND WU, Y. (2007). Dose finding designs for continuous responses and binary utility. *Journal of Biopharmaceutical Statistics* **17**, 1085–1096.

GAYDOS, B., KRAMS, M., PEREVOZSKAYA, I., BRETZ, F., LIU, Q., GALLO, P., BERRY, D., CHUANG-STEIN, C., PINHEIRO, J. AND BEDDING, A. (2006). Adaptive dose-response studies. *Drug Information Journal* **40**, 451–461.

GLONEK, G. AND MCCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 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**, 708–725.

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

HEISE, M. AND MYERS, R. (1996). Optimal designs for bivariate logistic regression. *Biometrics*, 613–624.

JHEE, S. S., LYNESS, W. H., ROJAS, P. B., LEIBOWITZ, M. T., ZAROTSKY, V. AND JACOBSEN, L. V. (2004). Similarity of insulin detemir pharmacokinetics, safety, and tolerability profiles in healthy caucasian and Japanese American subjects. *The Journal of Clinical Pharmacology* **44**, 258–264.

LEISCH, F., WEINGESSEL, A. AND HORNIK, K. (1998). On the generation of correlated artificial binary data, *Working Papers SFB* "*Adaptive Information Systems and Modelling in Economics and Management Science*", **13**.

LIU, W., BRETZ, F., HAYTER, A. J. AND WYNN, H. P. (2009). Assessing non-superiority, non-inferiority or equivalence when comparing two regression models over a restricted covariate region. *Biometrics* **65**, 1279–1287.

MÖLLENHOFF, K., BRETZ, F. AND DETTE, H. (2020). Equivalence of regression curves sharing common parameters. *Biometrics* **76**, 518–529.

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**, 2968–2981.

MURTAUGH, P. AND FISHER, L. (1990). Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Communications in Statistics Theory and Methods* **19**, 2003–2020.

OTTO, C., FUCHS, I., ALTMANN, H., KLEWER, M., WALTER, A., PRELLE, K., VONK, R. AND FRITZEMEIER, K. (2008). Comparative analysis of the uterine and mammary gland effects of drospirenone and medroxyprogesterone acetate. *Endocrinology* **149**, 3952–3959.

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**, e78805.

VARADHAN, R. (2014). *Constrained Nonlinear Optimization*. R package version 2011.9-1. 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**, 777–787.

ZHANG, W., SARGENT, D. AND MANDREKAR, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine* **25**, 2365–2383.

[*Received October 18, 2019; revised December 7, 2020; accepted for publication December 8, 2020*]