BIOMETRIC METHODOLOGY

Equivalence of regression curves sharing common parameters

Kathrin Möllenhoff1 | Frank Bretz2,3 | Holger Dette1

1Department of Mathematics, Ruhr-Universität Bochum, Bochum, Germany
2Novartis Pharma AG, Basel, Switzerland
3Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Correspondence
Kathrin Möllenhoff, Department of Mathematics, Ruhr-Universität Bochum, 44801 Bochum, Germany.
Email: kathrin.moellenhoff@rub.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: Sonderforschungsbereich 823, Teilprojekt T1

Abstract
In clinical trials, the comparison of two different populations is a common problem. Nonlinear (parametric) regression models are commonly used to describe the relationship between covariates, such as concentration or dose, and a response variable in the two groups. In some situations, it is reasonable to assume some model parameters to be the same, for instance, the placebo effect or the maximum treatment effect. In this paper, we develop a (parametric) bootstrap test to establish the similarity of two regression curves sharing some common parameters. We show by theoretical arguments and by means of a simulation study that the new test controls its significance level and achieves a reasonable power. Moreover, it is demonstrated that under the assumption of common parameters, a considerably more powerful test can be constructed compared with the test that does not use this assumption. Finally, we illustrate the potential applications of the new methodology by a clinical trial example.

KEYWORDS
dose-finding studies, equivalence testing, nonlinear regression, parametric bootstrap, similarity of regression curves

1 | INTRODUCTION

Regression models are commonly used to describe the relationship between multiple covariates and a response variable. In certain applications, more than one regression models are available, such as when assessing the relationship between the covariates and the response variables in more than one population (eg, in males and females). It is then often of interest to demonstrate the equivalence of the regression curves: If equivalence can be claimed, conclusions can be drawn from the pooled sample and a single regression model is sufficient to describe the data. This can be achieved by testing a suitable null hypothesis that the distance between the regression curves (measured in an appropriate sense) is smaller than a prespecified equivalence margin at a controlled type I error rate. Note that the problem of equivalence testing, as considered in this paper, is conceptually different from the more frequent problem of testing for the difference of curves and is much less studied in the literature due to methodological difficulties.

The problem of testing for differences between regression models has been intensively discussed in the nonparametric context, and we refer to the recent work of Feng et al. (2015), which contains a rather comprehensive list of references. In applied regression analysis, however, parametric models are usually preferred to a purely nonparametric approach as they admit a direct interpretation of the observed effects in terms of the model parameters. In addition, the available information of the observations is increased by applying
more efficient estimation or test procedures, provided that the assumed model is valid. Despite its importance, the problem of establishing the equivalence of two parametric regression models while controlling the type I error rate has only recently found attention in the literature. Using the intersection-union test device from Berger (1982), Liu et al. (2009) investigated the assessment of nonsuperiority, noninferiority, and equivalence when comparing two regression models over a restricted covariate region. Building upon this work, Gsteiger et al. (2011) derived equivalence tests based on simultaneous confidence bands for nonlinear regression models, with application to population pharmacokinetic analyses. Likewise, Bretz et al. (2018) assessed the similarity of dose-response curves in two nonoverlapping subgroups of patients. Alternatively, Dette et al. (2018) suggested directly estimating the distance between the regression curves and using a parametric bootstrap test to decide for equivalence of the two curves if the estimate is less than a certain threshold, where the bootstrap replicates are generated under a constraint of the model parameters. Expanding this approach, Möllenhoff et al. (2018) assessed the comparability of drug dissolution profiles via maximum deviation, whereas Hoffelder (2018) demonstrated the equivalence of dissolution profiles using the Mahalanobis distance; see also Collignon et al. (2019).

In these papers, the authors assumed that the regression models have different parameters and can, therefore, be evaluated separately. In some applications, however, this assumption cannot be justified, and it is more reasonable to assume that the regression models may have some common parameters. The total number of parameters to estimate is then reduced to the common and remaining parameters of each model, affecting the asymptotic behavior of the estimators. Consider, for example, the phase II dose-finding trial for a weight loss drug described in Bretz et al. (2018). This trial aimed at comparing the dose-response relationship for two regimens administered to patients suffering from overweight or obesity: Three doses each for once daily (o.d.) and twice daily (b.i.d.) use of the medication, and placebo. It is reasonable to assume that the placebo response is the same under both the o.d. and the b.i.d. regimens. Since the regression models typically used for dose-response modeling contain a parameter for the placebo response (Dette et al., 2010), they will share this common parameter for both the o.d. and the b.i.d. regimens. In some instances, it might even be reasonable to assume that the maximum efficacy for high doses is similar in both groups. Moreover, clinical trial sponsors may even decide to use the same placebo group for logistical reasons. The response of each patient on placebo is then used twice in the estimation of the o.d. and b.i.d. dose-response models, further complicating the statistical problem.

In this paper, we investigate the equivalence of two parametric regression curves that share common parameters. In Section 2, we first introduce the regression models to be estimated under the assumption of common parameters. We then develop a constrained parametric bootstrap test, which performs the resampling under the constraints of the null hypotheses implied by the equivalence test problem. The new tests improve the procedure proposed in Dette et al. (2018) using the additional information of common parameter in both groups. We also discuss testing the equivalence of model parameters to assess whether the assumption of common parameters is plausible. In Section 3, we investigate the finite sample properties of the proposed bootstrap test proposed in terms of power and size. In Section 4, we illustrate the methods using a multiregional clinical trial example where it is conceivable that the placebo and maximum treatment responses are the same across geographic regions but the onset of treatment differs due to intrinsic and extrinsic factors (Malinowski et al., 2008; ICH, 2017). Technical details and proofs are deferred to Section 1 of Supporting Information.

2 | METHODOLOGY

2.1 | Models with common parameters

Let

\[ Y_{\ell,i,j} = m_\ell(d_{\ell,i}, \beta_\ell) + \eta_{\ell,i,j}, \quad j = 1, \ldots, n_{\ell,j}, \]

\[ i = 1, \ldots, k_\ell, \]

(1)
denote the observed response of the \(j\)th subject at the \(i\)th dose level \(d_{\ell,i}\) under the \(\ell\)th dose-response model \(m_\ell\), where \(\ell = 1, 2\) denotes the index of the two groups under consideration and \(k_\ell\) the number of dose levels in group \(\ell\). We assume that the (nonlinear) regression model \(m_\ell\) is parametrized through a \(p_\ell\)-dimensional vector \(\beta_\ell\), \(\ell = 1, 2\). Note that the regression models \(m_1\) and \(m_2\) may be different. Likewise, the parameters \(\beta_1\) and \(\beta_2\) may be different even if \(m_1 = m_2\). We further assume that the error terms \(\eta_{\ell,i,j}\) are independent and identically distributed with expectation 0 and variance \(\sigma_\ell^2\). The dose levels \(d_{\ell,i}\) may be different in both groups but they are attained on the same (restricted) covariate region \(D\). In this paper, \(D\) is assumed to be the dose range, although the results can be generalized to include other covariates. Further, \(n_\ell = \sum_{i=1}^{k_\ell} n_{\ell,i}\) denotes the sample size in group \(\ell\) where we assume \(n_{\ell,i}\) observations in the \(i\)th dose level \((i = 1, \ldots, k_\ell, \ell = 1, 2)\). The sample sizes \(n_\ell\) can be unequal and the total number of observations is denoted by \(n = n_1 + n_2\).

In this paper, we consider the situation, where the regression models have some common parameters. More
We assume without loss of generality that these parameters are given by the first $p'$ model parameters of the parameter $\beta_\ell$ in model (1) that is $\beta_\ell = (\beta_{0,\ell}, \beta_{1,\ell}) \in \mathbb{R}^p$, $\ell = 1, 2$, where $\beta_0 \in \mathbb{R}^{p'}$ denotes the vector of common parameters in both regression models and $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the remaining parameters in the models $m_1$ and $m_2$, respectively, which do not necessarily coincide. The case where the models $m_1$ and $m_2$ do not share any common parameters is included and corresponds to $\beta_\ell = \hat{\beta}_\ell$ for $\ell = 1, 2$ (that is $p' = 0$). As a consequence, the $p_1 + p_2 - p'$-dimensional vector of all parameters of the regression functions in model (1) under the assumption above is given by $\beta = (\beta_0, \hat{\beta}_1, \hat{\beta}_2)$. Throughout this paper, we assume that $\beta \in B$ where $B \subset \mathbb{R}^{p_1 + p_2 - p'}$ is a compact set. In many important applications, the interpretation of common parameters for the two models is easy. For example, frequently the same models are used for both samples, that is $m_1 = m_2$.

A further important example is location-scale models, that is, $m_\ell(d, \beta_\ell) = \beta_{1,\ell} f_\ell(d, \beta_{0,\ell}, \beta_{2,\ell})$, $\ell = 1, 2$, where $\beta_\ell = (\beta_{0,\ell}, \beta_{1,\ell}^T, \beta_{2,\ell}^T)$ ($\ell = 1, 2$), $f_1$ and $f_2$ are functions with the same range and it is assumed that $\beta_{1,1} = \beta_{1,2}$ and/or $\beta_{1,2} = \beta_{1,2}$. Further issues of interpretability are also briefly discussed in Section 5.

The parameter $\beta$ is estimated by least squares using the combined sample $\{Y_{\ell, i,j} : j = 1, ..., n_{\ell, i}, i = 1, ..., k_\ell, \ell = 1, 2\}$, that is

$$\hat{\beta} = \arg\min_{(b_0, \hat{\beta}_1, \hat{\beta}_2) \in B} \sum_{j=1}^{2} \sum_{i=1}^{k_\ell} \sum_{j=1}^{n_{\ell, i}} (Y_{\ell, i,j} - m_\ell(d_{\ell, i,j}, (b_0, \hat{\beta}_1)))^2.$$  

(2)

### 2.2 Testing equivalence of regression curves

Following Liu et al. (2009) and Gsteiger et al. (2011) we consider the regression curves $m_1$ and $m_2$ to be equivalent if the maximum distance between the two curves is smaller than a given pre-specified constant, say $\varepsilon > 0$, that is,

$$d_\infty(\beta_1, \beta_2) = \max_{d \in \mathcal{D}} |m_1(d, \beta_1) - m_2(d, \beta_2)| < \varepsilon.$$  

In clinical trial practice, $\varepsilon$ is often referred to as a relevance threshold in the sense that if $d_\infty(\beta_1, \beta_2) < \varepsilon$ the difference between the two curves is believed not to be clinically relevant. In order to establish equivalence of the two curves $m_1$ and $m_2$ at a controlled type I error rate, we develop a test for the hypotheses

$$H_0 : d_\infty(\beta_1, \beta_2) \geq \varepsilon \quad \text{vs} \quad H_1 : d_\infty(\beta_1, \beta_2) < \varepsilon.$$  

(3)

In the following, we extend the bootstrap approach from Dette et al. (2018) to test the hypotheses (3) in the situation of common parameters. Note that the test procedure proposed below could also be applied to alternative measures of equivalence, such as the integrated deviation $\int_{\mathcal{D}} |m_1(t, \beta_1) - m_2(t, \beta_2)| dt$. We illustrate this fact in Section 2.5 of the Supporting Information by means of a small simulation study.

**Algorithm 1.** Parametric bootstrap for testing equivalence under the assumption of common parameters

1. Calculate the ordinary least-square (OLS) parameter estimate (2) assuming a common parameter $\beta_0$. The corresponding variance estimates are given by

$$\hat{\sigma}_\ell^2 = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} (Y_{\ell, i,j} - m_\ell(d_{\ell, i,j}, \hat{\beta}))^2, \quad \ell = 1, 2.$$  

(4)

where $\hat{\beta}_\ell = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, $\ell = 1, 2$. Calculate the estimate

$$\hat{d}_\infty = d_\infty(\hat{\beta}_1, \hat{\beta}_2) = \max_{d \in \mathcal{D}} |m_1(d, \beta_1) - m_2(d, \beta_2)|$$

for the maximal deviation between the two regression curves.

2. Define the constrained estimates

$$\hat{\beta}_\ell = \begin{cases} \hat{\beta}_\ell & \text{if } \hat{d}_\infty \geq \varepsilon, \\ \overline{\beta}_\ell & \text{if } \hat{d}_\infty < \varepsilon, \end{cases} \quad \ell = 1, 2,$$  

(5)

where $\overline{\beta}_1, \overline{\beta}_2$ minimize the objective function in (2) under the additional restriction

$$d_\infty(\beta_1, \beta_2) = \max_{d \in \mathcal{D}} |m_1(d, \beta_1) - m_2(d, \beta_2)| = \varepsilon.$$  

(6)

Define $\hat{d}_\infty = d_\infty(\hat{\beta}_1, \hat{\beta}_2)$ and note that $\hat{d}_\infty \geq \varepsilon$. This constrained optimization of the function (2) can be performed using, for example, the auglag function from the alabama package (Varadhan, 2014). This function is based on the augmented Lagrangian minimization algorithm, which is an optimization technique for solving constrained optimization problems. The next two steps describe the (parametric) bootstrap procedure.

3. Recall Equation (5) with notation $\hat{\beta}_\ell = (\hat{\beta}_0, \hat{\beta}_\ell)$ and generate data

$$Y_{\ell, i,j}^* = Y_{\ell, i,j} + \tilde{\varepsilon}_{\ell, i,j}, \quad i = 1, ..., n_{\ell, i}, \quad \ell = 1, 2.$$  

(7)
with independent and normally distributed errors 
\[ \eta_{i,j} \sim N(0, \sigma_i^2). \]

(4) Calculate the OLS estimate \( \hat{\beta}^* \) as in (2) and the test statistic as in Step (1), that is
\[ \hat{d}^*_\infty = \max_{d \in \mathcal{D}} |m_1(d, \hat{\beta}_1^*) - m_2(d, \hat{\beta}_2^*)|, \]
where \( \beta_\ell^* = (\beta_{1,\ell}^*, \beta_{2,\ell}^*) \), \( \ell = 1, 2 \). The \( \alpha \)-quantile of the distribution of the statistic \( \hat{d}^*_\infty \) is denoted by \( q_\alpha^* \) and the null hypotheses in (3) is rejected at the targeted significance level \( \alpha \) whenever \( \hat{d}^*_\infty < q_\alpha^* \).

In practice, the quantile \( q_\alpha^* \) is not known, but can be estimated with arbitrary precision repeating steps (3) and (4), say \( n_{\text{boot}} \) times, in order to obtain replicates \( \hat{d}^*_{\infty,1}, \ldots, \hat{d}^*_{\infty,n_{\text{boot}}} \) of \( \hat{d}^*_\infty \). An estimate \( \hat{q}^*_\alpha \) of \( q_\alpha^* \) is then defined by \( \hat{q}^*_\alpha := \hat{d}^*_\infty \left( \alpha | \{n_{\text{boot}} = 1\} \right) \), where \( \hat{d}^*_\infty(1) \leq \cdots \leq \hat{d}^*_\infty(n_{\text{boot}}) \) denotes the corresponding order statistic. The resulting rejecting rule is finally given by
\[ \hat{d}^*_\infty < \hat{q}^*_\alpha. \] (8)

Note that a resampling procedure has to mimic the distribution of the test statistic under the null hypothesis, independently of the truth. In other words, if the true distance between the two regression curves is below the threshold (under the alternative) the bootstrap sample should still mimic the distribution under the null hypothesis (distance of the two regression curves is larger than the threshold). This property is achieved by the parametric bootstrap described in Algorithm 1. By generating data under the constraint (6) we try to maximize the power of the bootstrap test. We also note that the bootstrap procedure is robust with respect to the assumption of normal distributed errors in model (1) if the samples sizes \( n_{\ell} \) are not too small. This is a consequence of the fact that all estimates depend only on the means \( \mathcal{Y}_{\ell,i,\ell'} = 1/n_{\ell,i} \sum_{j=1}^{n_{\ell,i}} Y_{\ell,i,j} \), which are—by the central limit theorem—asymptotically normally distributed (see section 2.4 of the Supporting Information for some simulation results for an alternative error distribution). The following theorem states that Algorithm 1 yields a valid test procedure. The proof is left to the Supporting Information.

**Theorem 1.** The test defined by (8) is a consistent, asymptotic \( \alpha \)-level test. That is
\[ \lim_{n_1, n_2 \to \infty} P(\hat{d}^*_\infty < \hat{q}^*_\alpha) = 1, \] (9) whenever \( d_\infty < \varepsilon \), and, if \( d_\infty \geq \varepsilon \)
\[ \limsup_{n_1, n_2 \to \infty} P(\hat{d}^*_\infty < \hat{q}^*_\alpha) \leq \alpha. \] (10)

**Remark 1.** The results presented in this section remain correct in trials with a common placebo group, where \( n_0 \) observations are taken at dose level \( d_1 = 0 \) (corresponding to placebo), which are modelled by the random variables \( Y_{0,1}, \ldots, Y_{0,n_0} \). For the sake of a simple presentation, we consider location-scale type models, such that the common effect at the placebo can easily be modelled. More general models can be considered as well by introducing additional constraints for the parameter. Assume that the models in (1) are given by
\[ m_{\ell}(d, \hat{\beta}_\ell) = \beta_{0,\ell} + \hat{\beta}_{\ell,1} \cdot m_{\ell}(d, \hat{\beta}_0^0), \quad \ell = 1, 2, \]
\[ i = 1, \ldots, k_{\ell}, \] (11)
where \( m_{\ell}(0, \hat{\beta}_0^0) = 0 \) (\( \ell = 1, 2 \)), such that the condition \( m_1(0, \hat{\beta}_1) = m_2(0, \hat{\beta}_2) = \beta_{0,1} \) reflects the fact that there is only one placebo group (and as a consequence a common placebo parameter). Models of this type cover the most frequently used functional forms used in drug development; see, for example, Bretz et al. (2005) and LaVange (2019). Besides the location parameter \( \beta_{0,1} \) there may be also other shared parameters, which we do not reflect in our notations for better readability. Consequently, \( \beta_{0,1} \) is a part of \( \hat{\beta}_0 \). The \( \ell \)-th model is completely characterized by its parameter \( \beta_{\ell} = (\beta_{0,\ell}, \hat{\beta}_{\ell,1}) = (\beta_{0,\ell}, \hat{\beta}_{0,\ell,1}, \hat{\beta}_{0,\ell,1}^0), \ell = 1, 2, \) and we obtain estimates of the model parameters by minimizing the sum of squares
\[ \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_{1,1}, \hat{\beta}_{2,1}) = \min_{b \in \mathbb{R}} \sum_{j=1}^{n_0} (Y_{0,j} - b_{0,1})^2 + \sum_{\ell=1}^{2} k_{\ell} \sum_{i=1}^{n_{\ell,i}} \left( Y_{\ell,i,j} - \left( b_{0,1} + \hat{\beta}_{\ell,1} \cdot m_{\ell}(d_{\ell,i}, \hat{\beta}_0^0) \right) \right)^2. \] (12)

Theorem 1 still holds in this case and a proof can be found in the Supporting Information.

**Remark 2.** The methodology can easily be extended to test relevant hypotheses of the form \( H_0: d_\infty(\hat{\beta}_1, \hat{\beta}_2) \leq \varepsilon \) vs \( H_1: d_\infty(\hat{\beta}_1, \hat{\beta}_2) > \varepsilon \). In this case, the null hypothesis is rejected, whenever \( \hat{d}^*_\infty > \hat{q}^*_\alpha \), where the quantile is defined in Algorithm 1.
2.3 Testing equivalence of model parameters

So far, we assumed that the two regression models $m_1$ and $m_2$ share the common parameter $\beta_0$. In practice, it may be necessary to assess whether this assumption is plausible using an appropriate equivalence test for the shared model parameters. To be more precise, we recall the definition the parameters $\beta_\ell$ in model (1), that is, $\beta_\ell = (\beta_{1,\ell}, ..., \beta_{p',\ell}, ..., \beta_{p,\ell})$, $\ell = 1, 2$, and note that the assumption of $p'$ common parameters in the models $m_1$ and $m_2$ can be represented as $(\beta_{1,1}, ..., \beta_{1,2}) = (\beta_{2,1}, ..., \beta_{2,2})$ for $\ell = 1, 2$. In order to investigate if this assumption holds, at least approximately, we construct a test for the hypotheses

$$K_0 : \max_{i=1, ..., p'} |\beta_{i,1} - \beta_{i,2}| \geq \delta \quad \text{vs} \quad K_1 : \max_{i=1, ..., p'} |\beta_{i,1} - \beta_{i,2}| < \delta,$$

(13)

where $\delta$ denotes the equivalence margin. To be precise, let $\hat{\beta}_\ell^{(\ell)}$ denote the least squares estimates in model $m_\ell$ for the sample $\{y_{\ell,i,j} : j = 1, ..., n_\ell,i, \ i = 1, ..., k_\ell\}$ ($\ell = 1, 2$), and assume that the sample sizes $n_\ell$ and $n_{\ell,i}$ converge to infinity such that

$$\lim_{n_\ell \to \infty} \frac{n_{\ell,i}}{n_\ell} = \zeta_{\ell,i} > 0, \quad i = 1, ..., k_\ell, \ \ell = 1, 2, \quad \text{and}$$

$$\lim_{n, n_\ell \to \infty} \frac{n}{n_\ell} = \lambda \in (1, \infty).$$

(14)

(15)

Under standard assumptions listed in section 1.1 of the Supporting Information, it can be shown that the least-squares estimate $\hat{\beta}_\ell^{(\ell)}$ of the parameter $\beta_\ell$ in model $m_\ell$ is approximately normal distributed, that is

$$\sqrt{n_\ell} (\hat{\beta}_\ell^{(\ell)} - \beta_\ell) \overset{D}{\to} N(0, \Sigma_\ell^{-1}), \quad \ell = 1, 2,$$

(16)

where the symbol $\overset{D}{\to}$ means convergence in distribution and the matrix $\Sigma_\ell$ is defined by

$$\Sigma_\ell = \frac{1}{\sigma_\ell^2} \sum_{i=1}^{k_\ell} \zeta_{\ell,i} \frac{\partial}{\partial b_\ell} m_\ell(d_{\ell,i}, b_\ell) \bigg|_{b_\ell = \beta_\ell} \left( \frac{\partial}{\partial b_\ell} m_\ell(d_{\ell,i}, b_\ell) \bigg|_{b_\ell = \beta_\ell} \right)^T, \quad \ell = 1, 2.$$  

(17)

Here we assume that the matrices $\Sigma_1$ and $\Sigma_2$ are non-singular. Consequently, the difference $\sqrt{n_\ell} (\hat{\beta}_1^{(1)} - \hat{\beta}_2^{(2)})$ is also asymptotically normally distributed, and in particular it follows for the first $p'$ components of the difference that

$$\sqrt{n_\ell} (\hat{\beta}_{1,1}^{(1)}, ..., \hat{\beta}_{p',1}^{(1)} - (\hat{\beta}_{1,1}^{(2)}, ..., \hat{\beta}_{p',2}^{(2)})) \overset{D}{\to} N(0, \Omega),$$

(18)

where the matrix $\Omega$ is defined by

$$\Omega := \lambda \Lambda_1^{-1} + \frac{\lambda}{\lambda - 1} \Lambda_2^{-1},$$

(19)

$\Lambda^{-1}_\ell = ((\Sigma^{-1}_\ell)_{ij})_{i,j=1}^{p'}$ denotes the upper-left $p' \times p'$-block of the matrix $\Sigma^{-1}_\ell$ ($\ell = 1, 2$) and $\lambda$ is defined in Equation (15). Therefore we obtain the approximation

$$(\hat{\beta}_{1,1}^{(1)}, ..., \hat{\beta}_{p',1}^{(1)} - (\hat{\beta}_{2,1}^{(2)}, ..., \hat{\beta}_{2,p'}^{(2)}) \overset{D}{\to} N \left( (\beta_{1,1}^{(1)}, ..., \beta_{1,p'}^{(1)}) - (\beta_{2,1}^{(2)}, ..., \beta_{2,p'}^{(2)}), \frac{1}{n} \Omega \right),$$

where $\Omega$ is defined in Equation (19). We can now apply the test (2.2) proposed in Wang et al. (1999) by rejecting the null hypothesis $K_0$ in Equation (13), whenever

$$|\hat{\beta}_{i,1}^{(1)} - \hat{\beta}_{i,2}^{(2)}| < \delta - t_{1-a, n-2} \left( \frac{\hat{\Omega}_{ii}}{n(n-2)} \right)^{1/2}$$

for all $i = 1, ..., p'$,

(20)

where $t_{1-a, n-2}$ denotes the $1 - \alpha$ quantile of the $t$-distribution with $n-2$ degrees of freedom and $\hat{\Omega}_{ii}$ the $i$th diagonal element of the matrix $\hat{\Omega}$ which is an estimate for the (unknown) covariance matrix $\Omega$. We obtain $\hat{\Omega}$ by replacing the unknown parameters $\beta_\ell$, $\sigma_\ell^2$ and weights $\zeta_{\ell,i}$ in Equation (17) by their corresponding estimates and $n_{\ell,i}/n_\ell$, respectively.

Remark 3. One reviewer suggested a two-stage procedure that formally combines the approaches in Section 2.2 and in this section: If the preliminary test for equivalence of model parameters is significant, one proceeds with the assumption of $p'$ common parameters in the models $m_1$ and $m_2$; otherwise, one proceeds in establishing equivalence of the two models $m_1$ and $m_2$ without shared parameters. Several authors have identified problems in such two-stage procedures based on preliminary tests in other applications; see, for example, Bancroft (1944) and Paull (1950) for early references on this topic. It would be interesting to investigate the performance (type I error rates and power) of such an adaptive test compared with the one.
that fits two models separately and we leave this topic for future research.

3 FINE SAMPLE PROPERTIES

We now investigate the finite sample properties of the bootstrap test proposed in Section 2.2 in terms of power and size using numerical simulations. The data are generated as follows:

1. We choose the functional form of the models \( m_1, m_2 \) and specify their parameters \( \beta_1, \beta_2 \) (including a common parameter \( \beta_0 \)), which determine the true underlying models. Further, we choose variances \( \sigma^2 \) and the actual dose levels \( d_{\epsilon, i}, \epsilon = 1, 2 \).

2. For each dose \( d_{\epsilon, i} \), we calculate \( n_{\epsilon, i} \) values for the response given by \( m_\epsilon(d_{\epsilon, i}, (\beta_0, \hat{\beta}_\epsilon)) \). By generating residual errors \( \eta_{\epsilon, i,j} \sim N(0, \sigma^2) \), we obtain the final response data

\[
Y_{\epsilon, i,j} = m_\epsilon(d_{\epsilon, i}, (\beta_0, \hat{\beta}_\epsilon)) + \eta_{\epsilon, i,j}, \quad j = 1, \ldots, n_{\epsilon, i},
\]

\[
i = 1, \ldots, k_\epsilon, \quad \epsilon = 1, 2. \tag{21}
\]

The simulation results below were obtained using 1000 simulation runs, where \( n_{\text{boot}} = 500 \) bootstrap replications were used to calculate quantiles of the bootstrap test. In the following, we report the simulations results for power and size under three different scenarios. We consider the four-parameter sigmoid \( E_{\text{max}} \) model

\[
m(d, \beta) = \hat{\beta}_1 + \frac{\beta_2 d^{\hat{\beta}_3}}{\hat{\beta}_2^3 + d^{\hat{\beta}_3}}, \tag{22}
\]

which is frequently used in practice when modeling dose-response relationships (see, eg, Gabrielsson and Weiner, 2007) or Thomas et al. (2014). In model (22) the parameter \( \beta = (\hat{\beta}_1, \beta_2, \hat{\beta}_3, \beta_0) \) corresponds (in this order) to the placebo effect \( E_0 \), the maximum effect \( E_{\text{max}} \), the Hill parameter \( h \) determining the steepness of the dose-response curve and the dose \( ED_{50} \) producing half of the maximum effect (MacDougall, 2006). In what follows, we add an index \( \ell = 0 \) for a shared parameter or \( \ell = 1, 2 \) for the group under consideration.

**Scenario 1:** We assume the dose range \( D = [0, 4] \) with identical dose levels \( d_{\epsilon, i} = i - 1, i = 1, 2, 3, 4, 5 \) for both regression models \( \ell = 1, 2 \). For each configuration of \( \sigma^2 = 1, 2, 3, 4, 5 \), we use (21) to simulate \( n_{\epsilon, i} = 6, 18, 30, 40, 100 \) observations at each dose level \( d_{\epsilon, i} \), resulting in total sample sizes of \( n_\epsilon = 30, 90, 150, 200, 500 \). We first compare the two sigmoid \( E_{\text{max}} \) models

\[
m_1(d, \beta_1) = \beta_{0,1} + \frac{\beta_{0,3} d^{\beta_{1,3}}}{\beta_{1,4}^3 + d^{\beta_{1,3}}} \quad \text{and}
\]

\[
m_2(d, \beta_2) = \beta_{0,1} + \frac{\beta_{0,3} d^{\beta_{2,3}}}{\beta_{2,4}^3 + d^{\beta_{2,3}}}. \tag{23}
\]

assuming the shared parameters \( (\beta_{0,1}, \beta_{0,2}, \beta_{0,3}) \). The only difference between the two models is in the \( ED_{50} \) parameters \( \hat{\beta}_{1,4} \) and \( \hat{\beta}_{2,4} \), which results in estimating five parameters in total. We consider the reference sigmoid \( E_{\text{max}} \) model \( m_1 \) with the parameters \( (\beta_{0,1}, \beta_{0,2}, \beta_{0,3}) = (1, 5, 4) \) and \( \beta_{1,4} = 1 \). This reference model is compared to various specifications of the second model \( m_2 \) determined by \( \hat{\beta}_{2,4} = 1.99, 1.77, 1.59, 1.43, 1.37, 1 \) and common shared parameters \( (\beta_{0,1}, \beta_{0,2}, \beta_{0,3}) \). The values for \( \beta_{2,4} \) were chosen such that the maximum absolute distances \( d_{\infty} = \max_{d \in D} |m_1(d, \beta_1) - m_2(d, \beta_2)| \) are given by 2, 1.5, 1, 0.5, 0.25, 0, respectively. For \( d_{\infty} > 0 \) these are attained at the dose levels 1.61, 1.52, 1.44, 1.37, and 1.33; see Figure 1A. For \( d_{\infty} = 0 \), that is \( \hat{\beta}_{1,4} = \hat{\beta}_{2,4} \), the maximum distance is attained at every point in \( D \).

In Table 1 we display the simulated rejection probabilities of the bootstrap test (8) under the null hypothesis (3) with \( d_{\infty} = 2, 1.5, 1 \) and \( \epsilon = 1 \). We conclude that the test controls its level in all cases under consideration. At the margin of the null hypothesis (ie, \( d_{\infty} = 1 \)) the approximation of the level is very precise, even for sample sizes as small as \( n_{\epsilon, i} = 6 \).

We also repeated the type I error rate simulations when fixing the Hill parameter at \( \beta_{0,3} = 4 \). The results are reported in Table 1 (numbers in brackets) and we conclude that the size is well controlled within the simulation error. In order to further investigate the role of the Hill parameter, we observe the relative residual mean squared errors (RRMSE) of the parameters estimates for this scenario in both cases, that is estimating the Hill parameter and fixing it. For the sake of brevity, the table presenting the results and a detailed interpretation is deferred to Section 2.1 of the Supporting Information (see Table 1). It turns out that the RRMSE for estimating the Hill parameter \( \beta_{0,3} \) is (by far) the largest. Further, we observe that all estimation errors decrease with larger sample sizes and smaller variances and are in general smaller when fixing the Hill parameter.

In Table 2, we summarize the power of the bootstrap when generating the data under the alternative \( d_{\infty} = 0.5, 0.25, 0 \) and \( \epsilon = 1 \). As expected, the power increases with larger sample sizes and smaller variances and is reasonably high across all configurations. Fixing the Hill parameter significantly improves the power, which can be explained by the difficulty of estimating this
parameter precisely, as discussed above. Considering the two largest sample sizes, that is \( n_{\ell} = 200, 500 \), underlines the asymptotic theory derived in Section 2, as the power clearly tends to one.

We investigated further scenarios and all results are deferred to the Supporting Information. First, we considered two sigmoid \( E_{\text{max}} \) models assuming two shared parameters, namely that the placebo effect and the maximum treatment effect are the same (see Figure 1B). We conclude again that the bootstrap test controls its level in all cases under consideration and we observe a reasonable power; see section 2.2 of the Supporting Information. Second, we compared the operating characteristics of the bootstrap test assuming three, two, one, and no shared parameters. The test assuming three shared parameters has the highest power among all four tests, followed by the test assuming two shared parameters; see section 2.3 of the Supporting Information. Finally, we investigated another type of error distribution and an alternative measure of equivalence for scenario 1.

![Figure 1](image)

**Figure 1** Graphical illustration of scenarios 1 and 2. Open dots indicate the doses and corresponding responses where the maximum distance to the reference curve \( m_{1} \) (dashed line) is attained.

| \( n_{\ell} \) | \( d_{\alpha} \) | \( \alpha = 0.05 \) | \( \alpha = 0.1 \) |
|---|---|---|---|
| \( \sigma^2 = 1 \) | \( \sigma^2 = 2 \) | \( \sigma^2 = 3 \) | \( \sigma^2 = 1 \) | \( \sigma^2 = 2 \) | \( \sigma^2 = 3 \) |
| 30 | 2 | 0.002 (0.000) | 0.010 (0.005) | 0.012 (0.009) | 0.002 (0.004) | 0.018 (0.013) | 0.031 (0.019) |
| 30 | 1.5 | 0.014 (0.013) | 0.035 (0.028) | 0.027 (0.043) | 0.026 (0.020) | 0.055 (0.049) | 0.052 (0.070) |
| 30 | 1 | 0.068 (0.058) | 0.054 (0.065) | 0.051 (0.060) | 0.106 (0.099) | 0.109 (0.114) | 0.115 (0.121) |
| 90 | 2 | 0.000 (0.000) | 0.000 (0.000) | 0.001 (0.002) | 0.000 (0.000) | 0.000 (0.000) | 0.002 (0.002) |
| 90 | 1.5 | 0.004 (0.001) | 0.010 (0.006) | 0.013 (0.015) | 0.005 (0.005) | 0.021 (0.014) | 0.026 (0.020) |
| 90 | 1 | 0.048 (0.071) | 0.053 (0.043) | 0.062 (0.059) | 0.104 (0.122) | 0.101 (0.097) | 0.129 (0.117) |
| 150 | 2 | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) |
| 150 | 1.5 | 0.000 (0.000) | 0.003 (0.000) | 0.002 (0.000) | 0.002 (0.000) | 0.011 (0.002) | 0.012 (0.009) |
| 150 | 1 | 0.056 (0.061) | 0.040 (0.061) | 0.042 (0.063) | 0.103 (0.102) | 0.090 (0.102) | 0.096 (0.109) |
| 200 | 2 | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) |
| 200 | 1.5 | 0.000 (0.000) | 0.002 (0.000) | 0.004 (0.003) | 0.000 (0.001) | 0.002 (0.001) | 0.008 (0.006) |
| 200 | 1 | 0.048 (0.049) | 0.053 (0.057) | 0.045 (0.048) | 0.097 (0.087) | 0.110 (0.109) | 0.112 (0.081) |
| 500 | 2 | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) |
| 500 | 1.5 | 0.000 (0.000) | 0.001 (0.000) | 0.000 (0.000) | 0.000 (0.000) | 0.001 (0.002) | 0.001 (0.000) |
| 500 | 1 | 0.056 (0.060) | 0.055 (0.056) | 0.058 (0.054) | 0.105 (0.110) | 0.111 (0.103) | 0.112 (0.103) |

*Note:* The numbers in brackets show the simulated type I error when fixing the Hill parameter at \( \hat{\beta}_{4,3} = 4 \).
More precisely, we generated error terms following a $t$-distribution with five degrees of freedom; see section 2.4 of the Supporting Information. We then investigated the integrated deviation $d_{t} := \int_{0}^{\infty} |m_{1}(t, \hat{\beta}_{1}) - m_{2}(t, \hat{\beta}_{2})| dt$ as a measure for similarity. For both additional simulation studies, we considered again two situations, that is fixed and estimating the Hill Parameter, respectively. It turns out that irrespective of the distribution under consideration, the results are very similar. The test controls its level for all specifications of $\sigma$ and $n_{e}$ and achieves a power, which is qualitatively the same as that when considering normally distributed errors. For the distance $d_{t}$ we recognize some minor differences, that is a slightly higher power in all situations, but as we are using two different measures of similarity, a comparison to the results concerning $d_{\infty}$ is hardly possible. However, we can conclude that for this alternative measure of equivalence also, we obtain a valid test procedure controlling its level and yielding a reasonably high power. For a more detailed interpretation, see sections 2.4 and 2.5 of the Supporting Information.

### Table 2: Simulated power of the bootstrap test (8) for the equivalence of two sigmoid $E_{\text{max}}$ models defined in scenario 1 with $\epsilon = 1$

| $n_{e}$ | $d_{w}$ | $\sigma^{2} = 1$ | $\sigma^{2} = 2$ | $\sigma^{2} = 3$ | $\sigma^{2} = 1$ | $\sigma^{2} = 2$ | $\sigma^{2} = 3$ |
|--------|--------|----------------|----------------|----------------|----------------|----------------|----------------|
| 30     | 0.5    | 0.137 (0.154) | 0.075 (0.078) | 0.070 (0.073) | 0.238 (0.266) | 0.172 (0.155) | 0.137 (0.145) |
| 50     | 0.25   | 0.208 (0.190) | 0.102 (0.101) | 0.081 (0.088) | 0.344 (0.349) | 0.196 (0.188) | 0.152 (0.170) |
| 30     | 0      | 0.181 (0.203) | 0.105 (0.105) | 0.086 (0.092) | 0.333 (0.361) | 0.196 (0.213) | 0.154 (0.154) |
| 90     | 0.5    | 0.341 (0.424) | 0.180 (0.230) | 0.132 (0.153) | 0.505 (0.581) | 0.311 (0.357) | 0.246 (0.279) |
| 90     | 0.25   | 0.550 (0.675) | 0.249 (0.315) | 0.166 (0.190) | 0.733 (0.802) | 0.428 (0.484) | 0.305 (0.348) |
| 90     | 0      | 0.664 (0.783) | 0.286 (0.353) | 0.191 (0.188) | 0.822 (0.884) | 0.463 (0.562) | 0.338 (0.367) |
| 150    | 0.5    | 0.481 (0.593) | 0.297 (0.359) | 0.207 (0.273) | 0.635 (0.729) | 0.460 (0.502) | 0.355 (0.406) |
| 150    | 0.25   | 0.826 (0.868) | 0.448 (0.569) | 0.280 (0.357) | 0.902 (0.933) | 0.635 (0.719) | 0.477 (0.545) |
| 150    | 0      | 0.917 (0.961) | 0.559 (0.665) | 0.342 (0.415) | 0.966 (0.989) | 0.740 (0.812) | 0.520 (0.596) |
| 200    | 0.5    | 0.616 (0.665) | 0.391 (0.439) | 0.248 (0.348) | 0.743 (0.783) | 0.545 (0.583) | 0.420 (0.505) |
| 200    | 0.25   | 0.880 (0.951) | 0.618 (0.718) | 0.417 (0.533) | 0.939 (0.974) | 0.747 (0.833) | 0.573 (0.697) |
| 200    | 0      | 0.972 (0.980) | 0.738 (0.839) | 0.518 (0.584) | 0.987 (0.991) | 0.875 (0.911) | 0.684 (0.767) |
| 500    | 0.5    | 0.921 (0.959) | 0.697 (0.790) | 0.523 (0.637) | 0.962 (0.985) | 0.814 (0.879) | 0.678 (0.754) |
| 500    | 0.25   | 0.999 (1.000) | 0.950 (0.975) | 0.841 (0.898) | 1.000 (1.000) | 0.979 (0.988) | 0.916 (0.951) |
| 500    | 0      | 1.000 (1.000) | 0.991 (1.000) | 0.957 (0.981) | 1.000 (1.000) | 0.998 (1.000) | 0.985 (0.996) |

Note: The numbers in brackets show the simulated power when fixing the Hill parameter at $\tilde{\beta}_{23} = 4$.

4 | CLINICAL TRIAL EXAMPLE

We now illustrate the proposed method with a multi-regional clinical trial example. The objective of this trial is to evaluate the dose-response relationships in Caucasian and Japanese patients and assess their similarity. Based on data from previous clinical trials investigating a drug with a similar mode of action, it is reasonable to assume a similar response to placebo and a common maximum treatment effect in both populations, with the main difference expected to be in a different onset of treatment effect. Using the sigmoid $E_{\text{max}}$ model (22), these considerations thus lead to different $ED_{50}$ and Hill parameters for the two dose-response curves. Because the trial is still in its design stage, we simulate data based on the trial assumptions. To maintain confidentiality, we scale the actual doses to lie within the [0, 15] interval. These limitations do not change the utility of the calculations below.

We assume 60 Japanese and 240 Caucasian patients, resulting in 300 patients overall. Patients from both populations are randomized to receive either placebo (dose level 0) or one of three active dose levels, namely 1, 3, and 15 for the Japanese and 0.5, 9, and 15 for the Caucasian patients. Assuming equal allocation of patients within each population, we thus have 75, 60, 15, 15, 60, and 75 patients randomized to the dose levels 0, 0.5, 1, 3, 9, and 15, respectively. The response variable is assumed to be normally distributed and larger values indicate a better outcome. Pharmacological and clinical considerations suggest the use of the (three-parameter) $E_{\text{max}}$ model with the Hill parameter fixed at 1. Later on, we relax this assumption as part of a sensitivity analysis.

In Figure 2, we display the fitted dose-response models $m_{1}(d, \hat{\beta}_{1})$ and $m_{2}(d, \hat{\beta}_{1})$ for the Japanese and Caucasian patients, respectively, together with the
individual observations, where \( d \in [0, 15] \) and the \( y \)-axis is truncated to \([-1,6]\) for better readability. The parameter estimates from the two separate model fits are given by \( \hat{\beta}_1 = (-0.195, 4.751, 11.991) \) and \( \hat{\beta}_2 = (-0.002, 5.676, 33.887) \). The observed differences for the placebo response and the maximum treatment effect are given by \( |\hat{\beta}_{1,1} - \hat{\beta}_{2,1}| = 0.193 \) and \( |\hat{\beta}_{1,2} - \hat{\beta}_{2,2}| = 0.925 \), respectively, and are thus relatively small, as also transpires from the plots in Figure 2. To corroborate this empirical observation, we formally test whether the assumption of shared parameters is plausible by applying the equivalence test described in Section 2.3 on the data set under consideration. We choose the threshold \( \delta = 1.5 \) and therefore test the null hypothesis \( K_0: \max_{i=1,2} |\hat{\beta}_{1,i} - \hat{\beta}_{2,i}| \geq 1.5 \) against the alternative \( K_1: \max_{i=1,2} |\hat{\beta}_{1,i} - \hat{\beta}_{2,i}| < 1.5 \). Applying the test (20) for \( \alpha = 0.05 \), we obtain \( \hat{\Omega}_{11} = 3127.91 \) and \( \hat{\Omega}_{22} = 10748.27 \) and, therefore, \( \delta - t_{1-\alpha,n-2}(\hat{\Omega}_{11}/n(n-2))^{1/2} = 1.191 \) and \( \delta - t_{1-\alpha,n-2}(\hat{\Omega}_{22}/n(n-2))^{1/2} = 0.928 \), respectively. We can thus reject \( K_0 \) at the relatively stringent 5% level and conclude equivalence of the two parameters, which justifies using the bootstrap test (1) with shared parameters.

We now evaluate the similarity of the dose-response curves for the Japanese and Caucasian patients, assuming the same placebo and maximum treatment effect. In order to compute the nonlinear least squares estimates in model (1) with common parameters \( \hat{\beta}_{0,1} \) and \( \hat{\beta}_{0,2} \) we formulate the objective function of the minimization step as

\[
\sum_{i=1}^{4} \sum_{j=1}^{60} \left( Y_{1,i,j} - \left( \hat{\beta}_{0,1} + \hat{\beta}_{0,2} d_{1,i,j} \right) / \hat{\beta}_{1,1} + d_{1,i,j} \right)^2 + \sum_{i=1}^{4} \sum_{j=1}^{240} \left( Y_{2,i,j} - \left( \hat{\beta}_{0,1} + \hat{\beta}_{0,2} d_{2,i,j} \right) / \hat{\beta}_{2,3} + d_{2,i,j} \right)^2.
\]

Here, \( \hat{\beta}_{0,1} \) denotes the (shared) placebo effect, \( \hat{\beta}_{0,2} \) the (shared) maximum treatment effect \( E_{max} \), and \( \hat{\beta}_{1,3} \) and \( \hat{\beta}_{2,3} \) the \( ED_{50} \) parameters of the two models. Using the \texttt{auglag} function from the \texttt{alabama} package \texttt{Varadhan (2014)} to solve the above optimization problem, we obtain the parameter estimates

\[
\hat{\beta}_{0,1} = -0.064 (0.074), \quad \hat{\beta}_{0,2} = 5.366 (0.137), \\
\hat{\beta}_{1,3} = 19.400 (2.634) \text{ and } \hat{\beta}_{2,3} = 25.681 (3.256).
\]

In brackets, we report the associated standard errors, which have to be calculated manually. The estimates for the population variances are \( \hat{\sigma}_1^2 = 0.508 \) and \( \hat{\sigma}_2^2 = 0.455 \). The observed maximum difference between both curves over the investigated dose range \([0, 15]\) is \( \hat{d}_\infty = 0.376 \), attained at dose 2.23. We apply the bootstrap test (8) using \( n_{\text{boot}} = 1000 \) bootstrap replications. Setting \( \varepsilon = 0.7 \) for the equivalence margin in (3), we obtain the quantile \( q_{0.05} = 0.438 \) for \( \alpha = 0.05 \). Thus, we reject the null hypothesis (3) at the 5% significance level and conclude that the dose-response curves for the Japanese and Caucasian populations are similar, under the shared parameter assumption. Alternatively, we can calculate the \( P \) value \( 1/B \sum_{i=1}^{B} I(d_{\infty}^{(i)} \leq \hat{d}_\infty) = 0.023 \) for the bootstrap test and obtain the same test decision at level \( \alpha = 0.05 \). For illustration purposes, we also apply the bootstrap test (8) but without shared parameters (yet under the assumption of a fixed Hill parameter). Accordingly, we obtain a considerably larger \( p = 0.458 \), which supports our findings from scenario 3 in Section 3 about the loss in power when no shared parameters are assumed. In this case, the observed maximum distance is \( \hat{d}_\infty = 0.706 \), attained at dose 1.42, and the quantile of the bootstrap distribution is \( q_{0.05} = 0.449 \).

Finally, we perform a sensitivity analysis to investigate the assumption of the Hill parameter being equal to 1. As part of this analysis, we repeat the model fit and the bootstrap test using the sigmoid model (22) where the Hill parameter is now part of the estimation. The parameter estimates (standard errors in brackets) are

\[
\hat{\beta}_{0,1} = 0.037 (0.082), \quad \hat{\beta}_{0,2} = 4.544 (0.218), \\
\hat{\beta}_{1,3} = 1.05 (0.229), \quad \hat{\beta}_{1,4} = 13.542 (2.095), \\
\hat{\beta}_{2,3} = 1.650 (0.331) \text{ and } \hat{\beta}_{2,4} = 16.558 (4.521).
\]

Now, the maximum distance between the curves is \( \hat{d}_\infty = 0.640 \), attained at dose 0.6. It turns out that the
standard errors of the estimates are slightly higher which is in line with the results shown in the simulation studies in Section 3. Performing again the bootstrap test with two shared parameters results in the quantile \( q_{0.05} = 0.429 \) and the \( P = .285 \). Consequently, we cannot reject the null hypothesis in this case. In conclusion, fixing the Hill parameter to 1 and assuming both the placebo effect and the maximum treatment effect to be the same in both populations clearly results in the most powerful procedure. We can demonstrate equivalence at the significance level of \( \alpha = .05 \), whereas in case of estimating both models separately (ie, no shared parameters) or including the Hill parameter in the estimation we obtain considerably larger \( P \) values.

One practical issue about working with different regression models may emerge after equivalence is claimed: one might then want to use a single regression model to fit both datasets for further inference and the question arises which one of the two models shall be chosen. Formally, one can apply model selection criteria, such as the information criterion from Akaike (1974) which tends to select models with fewer parameters. In practice, however, if we succeed in establishing the equivalence of two regression curves, evidence is provided that the difference in response is negligible over the entire covariate region \( D \). Such a result may provide sufficient evidence that any of the two models could be used for further inference. For example, in the context of multiregional clinical trials above, such a result may sufficiently prove that the same dose can be administered in both Japanese and Caucasian patients.

5 | DISCUSSION

In this paper, we developed a new test for the equivalence of two regression curves when it is reasonable to assume that some model parameters are the same. Our approach is based on an estimate of the maximum deviation between the two curves, where critical values are obtained by a novel constraint bootstrap procedure. We demonstrated that the new test controls its level properly and is consistent. The choice of the equivalence margins \( \varepsilon \) in (3) and \( \delta \) for the test described in (13) remains a practical problem. This choice depends on the particular application and has to be made by clinical experts, possibly with input from statisticians and other quantitative scientists. Regulatory guidance documents are available in specific settings, such as for the problem of demonstrating bioequivalence (CHMP, 2010).

We investigated the finite sample properties of the proposed procedure using extensive simulations and observed that the type I error rate is controlled in all scenarios under consideration, even for sample sizes as small as six patients per dose level. Further, we concluded that the test reaches a reasonable power that increases with larger sample sizes. In particular, we demonstrated that the power of tests for the equivalence of curves can be improved substantially by using the additional information on common parameters in the two regression curves. This effect could also be observed in the clinical trial example, which showed the power advantage of the bootstrap test (8) if the underlying assumptions are well justified. Relaxing those assumptions may lead to more robust conclusions but only at the cost of a loss in power.

In this paper, we proposed to apply potentially different regression models \( m_1(d, \beta_1) \) and \( m_2(d, \beta_2) \) to two populations, and assumed that some components of \( \beta_1 \) and \( \beta_2 \) can be the same. When \( m_1 = m_2 \) (as motivated in Section 2.1 and used in all simulations), such a common-parameter assumption seems reasonable. However, one reviewer pointed out that its meaning becomes less clear when \( m_1(\cdot, \cdot) \neq m_2(\cdot, \cdot) \) and suggested the investigation of \( m_1(d, \cdot) = m_2(d, \cdot) \), for all \( d \in D_0 \), where \( D_0 \) corresponds to some subset of the covariate region \( D \). The set \( D_0 \) could contain, for example, placebo, reflecting the common placebo effect assumption made in (11). Under this setup, the goal is to test the equivalence assumption \( \sup_{d \in \partial D_0} \left| m_1(d, \beta_1) - m_2(d, \beta_2) \right| < \varepsilon \) under the constraint above. We leave the investigation of this problem for future research.

Another line of future investigation stems from the fact that often only limited information about the shape of the regression models \( m_1 \) and \( m_2 \) is available before the start of an experiment. While empirical evidence suggests that the three- or four-parameter \( \bar{P}_{\text{max}} \) model is commonly observed in, for example, clinical dose-finding trials (MacDougall, 2006), model uncertainty remains of practical concern and is often underestimated. Especially, in the context of subgroup analysis and multiregional clinical trials, the dose-response models could be very different between subgroups and regions, respectively. Selecting a single model discards model uncertainty and tailored model selection or averaging approaches have been investigated in the context of clinical dose-finding trials (Schorning et al., 2016).

An interesting extension of the proposed methodology arises from the need to include covariates in clinical trial practice. Covariates can be continuous (eg, age or body mass index), categorical (eg, disease status or race), or binary (eg, gender or smoking yes/no), possibly changing over time. These cases may have to be treated differently and we leave this problem for future research. Another area of research could be the assessment of similarity in two nested populations, thus relaxing the assumption...
of independence between the observations. In our multi-regional clinical trial example, we compared the Japanese with Caucasian patients. It will be interesting and relevant to clinical trials to explore the development of the proposed methods when comparing the Japanese with an overall population that includes Japanese and Caucasian patients. Again, we leave this topic for future research.

ACKNOWLEDGMENTS

Parts of this manuscript were written while Frank Bretz was on a sabbatical at the University of Canterbury in Christchurch, New Zealand. He would like to thank Dr Daniel Gerhard for his support. We are also grateful to three reviewers for their constructive comments on an earlier version of this paper. This work has been supported in part by the Collaborative Research Center “Statistical modeling of nonlinear dynamic processes” (SFB 823, Teilprojekt T1) of the German Research Foundation and by a grant from the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM107639. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID

Kathrin Möllenhoff http://orcid.org/0000-0001-7861-3892

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. In: Emanuel, P., Kunio, T. and Genshiro, K. (Eds.) Perspectives in Statistics: Selected Papers of Hirotugu Akaike. New York, NY: Springer, pp. 215–222.

Bancroft, T.A. (1944) On biases in estimation due to the use of preliminary tests of significance. The Annals of Mathematical Statistics, 15, 190–204.

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

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.

CHMP (2010) Committee for Medicinal Products for Human Use: Guideline on the Investigation of Bioequivalence. London: European Medicines Agency.

Collignon, O., Möllenhoff, K. and Dette, H. (2019) Equivalence analyses of dissolution profiles with the mahalanobis distance: a regulatory perspective and a comparison with a parametric maximum deviation-based approach. Biometrical Journal, 61, 779–782.

Dette, H., Kiss, C., Bevanda, M. and Bretz, F. (2010) Optimal designs for the Emax, log-linear and Bretz, F. (2018) Equivalence of regression curves. Journal of the American Statistical Association, 113, 711–729.

Feng, L., Zou, C., Wang, Z. and Zhu, L. (2015) Robust comparison of regression curves. Test, 24, 185–204.

Gabrielsson, J. and Weiner, D. (2007) Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications, 4th edition. Stockholm: Swedish Pharmaceutical Press.

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.

Hoffelder, T. (2018) Equivalence analyses of dissolution profiles with the mahalanobis distance. Biometrical Journal, 61(5), 1120–1137.

ICH (2017) E17-general principles for planning and design of multi-regional clinical trials. Available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17EWG_Step4_2017_1116.pdf

LaVange, L.M. (2019) Statistics at fda: reflections on the past six years. Statistics in Biopharmaceutical Research, 11, 1–12.

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, 1279–1287.

MacDougall, J. (2006) Analysis of dose-response studies-E max model. In: Ting, N. (Ed.) Dose Finding in Drug Development. New York, NY: Springer, pp. 127–145.

Malinowski, H.J., Westelinck, A., Sato, J. and Ong, T. (2008) Same drug, different dosing: differences in dosing for drugs approved in the United States, Europe, and Japan. The Journal of Clinical Pharmacology, 48, 900–908.

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.

Paull, A.E. (1950) On a preliminary test for pooling mean squares in the analysis of variance. The Annals of Mathematical Statistics, 21, 539–556.

Schorning, K., Bornkamp, B., Bretz, F. and Dette, H. (2016) Model selection versus model averaging in dose finding studies. Statistics in Medicine, 35, 4021–4040.

Thomas, N., Sweeney, K. and Somayaji, V. (2014) Meta-analysis of clinical dose-response in a large drug development portfolio. Statistics in Biopharmaceutical Research, 6, 302–317.

Varadhan, R. (2014) Constrained nonlinear optimization. R package version 2015.3-1. Available at http://cran.r-project.org/web/packages/alabama/index.html

Wang, W., GeneHwang, J. and Dasgupta, A. (1999) Statistical tests for multivariate bioequivalence. Biometrika, 86, 395–402.
SUPPORTING INFORMATION

Web Appendices and Tables referenced in Sections 2 and 3 are available with this paper at the Biometrics website on Wiley Online Library. R Code for reproducing the simulation results from Section 3 and the case study from Section 4 is also provided.

How to cite this article: Möllenhoff K, Bretz F, Dette H. Equivalence of regression curves sharing common parameters. Biometrics. 2020;76:518–529. https://doi.org/10.1111/biom.13149