Overcoming model uncertainty – how equivalence tests can benefit from model averaging

Niklas Hagemann[1] and Kathrin Möllenhoff[1]
[1] Institute of Medical Statistics and Computational Biology (IMSB),
Faculty of Medicine, University of Cologne, Germany

May 3, 2024

Abstract

A common problem in numerous research areas, particularly in clinical trials, is to test whether the effect of an explanatory variable on an outcome variable is equivalent across different groups. In practice, these tests are frequently used to compare the effect between patient groups, e.g. based on gender, age or treatments. Equivalence is usually assessed by testing whether the difference between the groups does not exceed a pre-specified equivalence threshold. Classical approaches are based on testing the equivalence of single quantities, e.g. the mean, the area under the curve (AUC) or other values of interest. However, when differences depending on a particular covariate are observed, these approaches can turn out to be not very accurate. Instead, whole regression curves over the entire covariate range, describing for instance the time window or a dose range, are considered and tests are based on a suitable distance measure of two such curves, as, for example, the maximum absolute distance between them. In this regard, a key assumption is that the true underlying regression models are known, which is rarely the case in practice. However, misspecification can lead to severe problems as inflated type I errors or, on the other hand, conservative test procedures.

In this paper, we propose a solution to this problem by introducing a flexible extension of such an equivalence test using model averaging in order to overcome this assumption and making the test applicable under model uncertainty. Precisely, we introduce model averaging based on smooth AIC weights and we propose a testing procedure which makes use of the duality between confidence intervals and hypothesis testing. We demonstrate the validity of our approach by means of a simulation study and demonstrate the practical relevance of the approach considering a time-response case study with toxicological gene expression data.

1 Introduction

In numerous research areas, particularly in clinical trials, a common problem is to test whether the effect of an explanatory variable on an outcome variable is equivalent across different groups (see, e.g., Otto et al., 2008; Jhee et al., 2004). Equivalence is usually assessed by testing whether the difference between the groups does not exceed a pre-specified equivalence threshold. The choice of this threshold is crucial as it resembles the maximal amount of deviation for which equivalence can still be concluded. One usually chooses the threshold based on prior knowledge, as a percentile of the range of the outcome variable or resulting from regulatory guidelines. Equivalence tests provide a flexible tool for plenty of research questions. For instance, they can be used to test for equivalence across patient groups, e.g. based on gender or age, or between treatments. Moreover, they are a key ingredient of bioequivalence studies, investigating whether two formulations of a drug have nearly the

same effect and are hence considered to be interchangeable (e.g. Möllenhoff et al., 2022; Hauschke et al., 2007).

Classical approaches are based on testing the equivalence of single quantities, e.g. the mean, the area under the curve (AUC) or other values of interest (Schuirmann, 1987; Lakens, 2017). However, when differences depending on a particular covariate are observed, these approaches can turn out to be not very accurate. Instead, considering the entire covariate range, describing for instance the time window or a dose range, has recently been proposed by testing equivalence of whole regression curves. Such test are typically based on the principle of confidence interval inclusion (Liu et al., 2009; Gsteiger et al., 2011; Bretz et al., 2018). However, a more direct approach applying various distance measures has been introduced by Dette et al. (2018), which turned out to be particularly more powerful. Based on this, many further developments, e.g. for different outcome distributions, specific model structures and/or responses of higher dimensions, have been introduced (see, e.g., Möllenhoff et al., 2020, 2021, 2024; Hagemann et al., 2024).

All these approaches have one thing in common: they base on the assumption that the true underlying regression model is known. In practice this usually implies that the models need to be chosen manually, either based on prior knowledge or visually. Hence, these approaches might not be robust with regard to model misspecification and, consequently, suffer from problems like inflated type I errors or reduced power (Guhl et al., 2022; Dennis et al., 2019). One idea to tackle this problem is implementing a testing procedure which explicitly incorporates the model uncertainty. This can be based on a formal model selection procedure, see, e.g., Möllenhoff et al. (2018), who propose conducting a classical model choice procedure prior to preforming the equivalence test.

An alternative to this is the incorporation of a model averaging approach into the test procedure. As outlined by Bornkamp (2015) model selection has some disadvantages compared to model averaging. Particularly, model selection is not stable in the sense that minor changes in the data can lead to major changes in the results (Breiman, 1996). This also implies that model selection is non-robust with regard to outliers. In addition, the estimation of the distribution of post model selection parameter estimators is usually biased (Leeb and Pötscher, 2005, 2008). Model averaging is omnipresent whenever model uncertainty is present, which is, besides other applications, often the case in parametric dose response analysis. Besides practical applications, there are also several methodological studies regarding model averaging in dose-response studies (see, e.g., Schorning et al., 2016; Aoki et al., 2017; Buatois et al., 2018) and Bornkamp et al. (2009), who incorporated model averaging as an alternative to model selection in their widely used dose-finding method MCPMod.

Therefore, in this paper, we propose an approach utilising model averaging rather than model selection. There are frequentist as well as Bayesian model averaging approaches. The former almost always use the smooth weights structure introduced by Buckland et al. (1997). These weights depend on the values of an information criterion of the fitted models. Predominantly, the Akaike information criterion (AIC; Akaike, 1974) is used but other information criteria can be used as well. While only few of the Bayesian approaches perform fully Bayesian inference (see, e.g., Ley and Steel, 2009), the majority makes use of the fact that the posterior model probabilities can be approximated by weights based on the Bayesian information criterion (BIC; Schwarz, 1978) that have the same smooth weights structure as the frequentist weights (Wasserman, 2000).

In this paper, we propose an equivalence test incorporating model-averaging and hence overcoming the problems caused by model uncertainty. Precisely, we first make use of the duality between confidence intervals and hypothesis testing and propose a test based on the derivation of a confidence interval. By doing so, we both guarantee numerical stability of the procedure and provide confidence intervals for the measure of interest. We demonstrate the usefulness of our method with the example of toxicological gene expression data. In this application, using model averaging enables us to analyse

the equivalence of time-response curves between two groups for 1000 genes of interest without the necessity of specifying all 2000 correct models separately, thus avoiding both a time-consuming model selection step and potential model misspecifications.
The paper is structured as follows: In Section 2, dose-response models and the concept of model averaging are succinctly discussed. In Section 3, the testing approach is introduced, proposing three different variations. Finite sample properties in terms of Type I and II error rates are studied in Section 4. Section 5 illustrates the method using the toxicological gene expression example before Section 6 closes with a discussion.

2 Model averaging for dose-response models

2.1 Dose-response models

We consider two different groups, indicated by an index $l = 1, 2$, with corresponding response variables y_{lij} with $\mathcal{Y} \subseteq \mathbb{R}$ denoting the set of all possible outcomes. There are $i = 1, ..., I_l$ dose levels and $j = 1, ..., n_{li}$ denotes the observation index within each dose level. For each group the total number of observations is n_l and n is the overall number of observations, i.e. $n_l = \sum_{i=1}^{I_l} n_{li}$ and $n = n_1 + n_2$. For each group we introduce a flexible dose-response model

$$y_{lij} = m_l(x_{li}, \boldsymbol{\theta}_l) + e_{lij}, \quad j = 1, ..., n_{li}, \quad i = 1, ..., I_l, \quad l = 1, 2,$$

where $x_{li} \in \mathcal{X} \subseteq \mathbb{R}$ is the dose level, i.e. the deterministic explanatory variable. We assume the residuals e_{lij} to be independent, have expectation zero and finite variance σ_l^2. The function $m_l(\cdot)$ models the effect of x_{li} on y_{lij} via a regression curve with $\boldsymbol{\theta}_l \in \mathbb{R}^{\dim(\boldsymbol{\theta}_l)}$ being its parameter vector. We assume $m_l(\cdot)$ to be twice continuously differentiable. In dose-response studies, as well as in time-response studies, often either a linear model

$$m_l(x, \theta_l) = \beta_{l0} + \beta_{l1} x, \tag{1}$$

a quadratic model

$$m_l(x, \theta_l) = \beta_{l0} + \beta_{l1} x + \beta_{l2} x^2, \tag{2}$$

an emax model

$$m_l(x, \theta_l) = \beta_{l0} + \beta_{l1} \frac{x}{\beta_{l2} + x}, \tag{3}$$

an exponential (exp) model

$$m_l(x, \theta_l) = \beta_{l0} + \beta_{l1} \left(\exp\left(\frac{x}{\beta_{l2}} \right) - 1 \right), \tag{4}$$

a sigmoid emax (sigEmax) model

$$m_l(x, \theta_l) = \beta_{l0} + \beta_{l1} \frac{x^{\beta_{l3}}}{(\beta_{l2})^{\beta_{l3}} + x^{\beta_{l3}}}, \tag{5}$$

also known as Hill model or 4pLL-model, or a beta model

$$m_l(x, \theta_l) = \beta_{l0} + \beta_{l1} \left(\frac{(\beta_{l2} + \beta_{l3})^{\beta_{l2}+\beta_{l3}}}{(\beta_{l2})^{\beta_{l2}} + (\beta_{l3})^{\beta_{l3}}} \right) \left(\frac{x}{s} \right)^{\beta_{l2}} \left(1 - \frac{x}{s} \right)^{\beta_{l3}}, \tag{6}$$

where s is a fixed scaling parameter, is deployed (Bretz et al., 2005; Pinheiro et al., 2006, 2014; Duda et al., 2022). These models strongly vary in the assumed underlying dose-response relation, e.g. in terms of monotony, and consequently in the shape of their curves. Therefore, choosing a suitable dose-response model is crucial for all subsequent analyses.
However, in practical applications the true underlying model shape is in general unknown. Thus, it might not always be clear which functional form of 1-6 should be imployed. A possible answer to this is implementing model averaging which, as outlined in Section 1, has several advantages over the simpler alternative of model selection.

2.2 Model averaging

As outlined before, frequentist as well as Bayesian model averaging approaches usually both use the same smooth weights structure introduced by Buckland et al. (1997) and Wasserman (2000), respectively. Accordingly, by leaving out the group index $l = 1, 2$ for better readability the averaged model is given by

$$m(x, \hat{\boldsymbol{\theta}}) := \sum_{k=1}^{K} w_k \, m_k(x, \hat{\boldsymbol{\theta}}_k), \tag{7}$$

where the $m_k(x, \hat{\boldsymbol{\theta}}_k), k = 1, .., K$, correspond to the K candidate models,

$$w_k = \frac{\exp(0.5\, I(m_k(x, \hat{\boldsymbol{\theta}}_k))}{\sum_{\tilde{k}=1}^{K} \exp(0.5\, I(m_{\tilde{k}}(x, \hat{\boldsymbol{\theta}}_{\tilde{k}})))} \tag{8}$$

are the corresponding weights and $I(\cdot)$ is an information criterion. Usually the AIC is used for frequentist model averaging, while the BIC is usually deployed for Bayesian model averaging (Schorning et al., 2016).
Despite the prevalence of the AIC and BIC, other information criteria are sometimes used as well, e.g. the deviance information criterion (see, e.g., Price et al., 2011). Occasionally, model averaging also bases on cross-validation or machine learning methods. For a more general introduction to model averaging techniques the reader is refereed to e.g. Fletcher (2018) or Claeskens and Hjort (2008) and an overview specifically focusing on dose-response models is given by Schorning et al. (2016).

2.3 Inference

As the parameter estimation is conducted for each of the candidate models separately, it is not influenced by the subsequent model averaging. Therefore, here the index k is left out. Inference can be based on an ordinary least squares (OLS) estimator, i.e. minimising

$$\sum_{i=1}^{I_l} \sum_{j=1}^{n_{li}} (y_{lij} - m_l(x_{li}, \boldsymbol{\theta}_l))^2, \quad l = 1, 2.$$

Alternatively, under the assumption of normality of the residuals, i.e.

$$e_{lij} \overset{iid}{\sim} N(0, \sigma_l^2), \quad l = 1, 2,$$

a maximum likelihood estimator (MLE) can also be deployed where the log-likelihood is given by

$$\ell(\boldsymbol{\theta}_l, \sigma_l^2) = -\frac{n_l}{2} \ln(2\pi\sigma_l^2) - \frac{1}{2\sigma_l^2} \sum_{i=1}^{I_l} \sum_{j=1}^{n_{li}} (y_{lij} - m_l(x_{li}, \boldsymbol{\theta}_l))^2, \quad l = 1, 2. \tag{9}$$

Under normality both approaches are identical and, hence, lead to the same parameter estimates $\hat{\boldsymbol{\theta}}_l$. From (9) an MLE for the variance

$$\hat{\sigma}_l^2 = \frac{1}{n_l} \sum_{i=1}^{I_l} \sum_{j=1}^{n_{li}} (y_{lij} - m_l(x_{li}, \boldsymbol{\theta}_l))^2, \quad l = 1, 2 \tag{10}$$

can be derived as well. In `R` inference is performed with the function `fitMod` from the package `Dosefinding` (Bornkamp et al., 2009; Pinheiro et al., 2014) which performs OLS estimation. The value of the log-likelihood needed for the AIC or BIC is then calculated by plugging the OLS estimator into the log-likelihood (9).

3 Model-based equivalence tests under model uncertainty

3.1 Equivalence testing based on confidence intervals

Model-based equivalence tests have been introduced in terms of the L^2-distance, the L^1-distance or the maximal absolute deviation (also called L^∞-distance) of the model curves (Dette et al., 2018; Bastian et al., 2024). Although all of these approaches have their specific advantages and disadvantages as well as specific applications, subsequent research (see, e.g., Möllenhoff et al., 2018, 2020, 2021; Hagemann et al., 2024) is predominately based on the maximal absolute deviation due to its easy interpretability. Accordingly, we state the hypotheses

$$H_0 : d \geq \varepsilon \text{ vs. } H_1 : d < \varepsilon \tag{11}$$

of equivalence of regression curves with respect to the maximal absolute deviation, that is

$$d = \max_{x \in \mathcal{X}} |m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2)|$$

is the maximal absolute deviation of the curves and ε is the pre-specified equivalence threshold, i.e. that a difference of ε is believed not to be clinically relevant. The test statistic is given as the estimated maximal deviation between the curves

$$\hat{d} = \max_{x \in \mathcal{X}} |m_1(x, \hat{\boldsymbol{\theta}}_1) - m_2(x, \hat{\boldsymbol{\theta}}_2)|. \tag{12}$$

As the distribution of $\hat{d}$ under the null hypothesis is in general unknown, it is usually either approximated based on a parametric bootstrap procedure or by asymptotic theory. In Dette et al. (2018) the asymptotic validity of both approaches is proven, but the corresponding simulation study shows that the bootstrap test outperforms the asymptotic test in finite samples. For the bootstrap test several studies (see, e.g., Dette et al., 2018; Möllenhoff et al., 2018, 2020) show reasonable results for finite samples across applications.

However, in light of practical application, this approach can have two disadvantages: First, it does not directly provide confidence intervals (CI) which provide useful information about the precision of the test statistic. Further, they would have an important interpretation analogously to their interpretation in classical equivalence testing known as TOST (two one-sided tests; Schuirmann, 1987), where the bounds of the confidence interval are typically compared to the confidence region of $[-\varepsilon, \varepsilon]$.

Second, it requires the estimation of the models under the constraint of being on the edge of the null hypothesis, i.e. the maximal absolute deviation being equal to ε (see Algorithm 1 in Dette et al., 2018). Technically, this is usually conducted using augmented Lagrangian optimisation. However, with

increasing model complexity, this becomes numerically challenging. In the context of model averaging, these numerical issues are particularly relevant since all models would need to be estimated jointly as they need to jointly fulfil the constraint. This leads to a potentially high dimensional optimisation problem with a large number of parameters. In addition, for model averaging the side constraint has a highly complex structure because with every parameter update not only the model curves change but also the model weights do.

As an alternative to approximating the distribution under the null hypothesis, we propose to test hypotheses (11) based on the well-known duality between confidence intervals and hypothesis testing (Aitchison, 1964). This testing approach is similar to what Bastian et al. (2024, Algorithm 1) introduced for the L^1 distance of regression models. Therefore, let $(-\infty, u]$ be a one-sided lower $(1-\alpha)$-CI for d which we can rewrite as $[0, u]$ due to the non-negativity of d, i.e.

$$\mathbb{P}(d \leq u) = \mathbb{P}(d \in (-\infty, u]) = \mathbb{P}(d \in [0, u]) \geq 1 - \alpha.$$

According to the duality between CI and hypothesis testing, we reject the null hypothesis and conclude equivalence if

$$\varepsilon > u. \tag{13}$$

This testing procedure is an α-level test as

$$\begin{aligned} &\mathbb{P}_{H_0}(\varepsilon > u) \\ \leq &\mathbb{P}(d > u) \\ = &1 - \mathbb{P}(d \leq u) \\ \leq &1 - (1 - \alpha) = \alpha. \end{aligned}$$

However, as the distribution of $\hat{d}$ is in general unknown, obtaining u is again a challenging problem. It is obvious from (13) that the quality of the testing procedure crucially depends on the quality of the estimator for u. If the CI is too wide the test procedure is conservative and lacks power. In contrast, a too narrow CI can lead to type I error inflation due to not reaching the desired coverage probability $1 - \alpha$. We propose three different possibilities to calculate the CI, namely

1. CI based on a parametric percentile bootstrap,
2. asymptotic CI based on the asymptotic distribution of $\hat{d}$ derived by Dette et al. (2018), and
3. a hybrid approach using the asymptotic normality of $\hat{d}$ but estimating its standard error based on a parametric bootstrap.

One-sided CIs based on a parametric percentile bootstrap can be constructed in the same way Möllenhoff et al. (2018) proposed for two-sided CIs. In order to do so, they obtain parameter estimates (either via OLS or maximum likelihood optimisation), generate bootstrap data from these estimates and calculate the percentiles from the ordered bootstrap sample. The resulting test is similar to what Bastian et al. (2024, Algorithm 1) derived for the L^1 distance of regression models. That is

$$\left[0, \hat{q}^*(1-\alpha)\right],$$

where $\hat{q}^*(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the ordered bootstrap sample.

Asymptotic CIs can be derived directly from test (5.4) of Dette et al. (2018) and are given by

$$\left[0, \hat{d} + \sqrt{\frac{\widehat{\mathrm{Var}}(d)}{n}} z\right] \tag{14}$$

where z is the $(1-\alpha)$-quantile of the standard normal distribution and $\widehat{\text{Var}}(d)$ is the closed-form estimator for the variance of d given by equation (4.7) of Dette et al. (2018). However, the asymptotic validity of this variance estimator is only given under the assumption that within $\mathcal{X}$ there is only one unique value x_0 where the absolute difference curve attains its maximum, i.e. $x_0 = \text{argmax}_{x\in\mathcal{X}} |m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2)|$ and, moreover, that this value x_0 is known. This does not hold in general as Dette et al. (2018) give two explicit counterexamples in terms of two shifted emax or exponential models. In addition, in practical applications x_0 is in general not known and needs to be estimated. If the absolute deviation along x is small, the estimation of x_0 can become unstable leading to an unstable variance estimator. Moreover, as mentioned before, the simulation study of Dette et al. (2018) shows that for finite samples the bootstrap test is superior to the asymptotic test. Given the disadvantages of the asymptotic CI and especially of the underlying variance estimator, we introduce a hybrid approach which is a combination of both approaches. It is based on the asymptotic normality of $\hat{d}$ but estimates the standard error of $\hat{d}$ based on a parametric bootstrap leading to

$$\left[0, \hat{d} + \widehat{\text{se}}(\hat{d})z\right],$$

where the estimator $\widehat{\text{se}}(\hat{d})$ of the standard error of $\hat{d}$ is the empirical standard deviation of the bootstrap sample.
Under the assumptions introduced by Dette et al. (2018), all three approaches are asymptotically valid. For the test based on the asymptotic CI, this follows directly from Dette et al. (2018). This also applies to the hybrid CI-based test due to $\widehat{\text{se}}(\hat{d})$ being an asymptotically unbiased estimator for the standard error of $\hat{d}$ as outlined by Efron and Tibshirani (1994). The asymptotic validity of the percentile approach follows from Dette et al. (2018, Appendix: proof of Theorem 4) analogously to how Bastian et al. (2024, Appendix A.3) derived the validity of their approach from there. The finite sample properties of the three methods are compared in Section 4.1.

3.2 Model-based equivalence tests incorporating model averaging

We now combine the model averaging approach presented in Section 2.2 with the CI-based test introduced in Section 3.1. For the asymptotic test that is estimating $m_l(x, \hat{\boldsymbol{\theta}}_l)$, $l = 1, 2$ using (7) with model weights (8) and then calculating the test statistic (12). Subsequently, the asymptotic CI (14) can be determined using the closed form variance estimator given by Dette et al. (2018). Using this CI, the test decision is based on decision rule (13).
The testing procedure of the percentile and hybrid approach is shown in Algorithm 1, where the first two steps are essentially the same as for the asymptotic test. The percentile test is conducted by performing Algorithm step 4a, while conducting step 4b instead leads to the hybrid test. In the following we will refer to this as Algorithm 1a and Algorithm 1b, respectively.

Algorithm 1

1. Obtain parameter estimates $\hat{\boldsymbol{\theta}}_{lk}$, $k = 1, ..., K_l$, $l = 1, 2$ for the candidate models, either via OLS or maximum likelihood optimisation (see Section 2.3). Determine the averaged models from the candidate models using (7), i.e. by calculating

$$m_l(x, \hat{\boldsymbol{\theta}}_l) = \sum_{k=1}^{K_l} w_{lk}\, m_{lk}(x, \hat{\boldsymbol{\theta}}_{lk}), \quad l = 1, 2,$$

with weights (8) as well as the variance estimator $\hat{\sigma}_l^2$, $l = 1, 2$ from (10).

2. Calculate the test statistic (12).

3. Execute the following steps:

   3.1 Obtain bootstrap samples by generating data according to the model parameters $\hat{\boldsymbol{\theta}}_l$, $l = 1, 2$. Under the assumption of normality that is

   $$y^*_{lij} \sim N(m_l(x_{li}, \hat{\boldsymbol{\theta}}_l), \hat{\sigma}^2_l), \quad i = 1, ..., n_l,\ l = 1, 2.$$

   3.2 From the bootstrap samples, estimate the models $m_l(x_{li}, \hat{\boldsymbol{\theta}}^*_l)$, $l = 1, 2$ as in step (1) and the test statistic

   $$\hat{d}^* = \max_{x \in \mathcal{X}} |m_1(x, \hat{\boldsymbol{\theta}}^*_1) - m_2(x, \hat{\boldsymbol{\theta}}^*_2)|. \tag{15}$$

   3.3 Repeat steps (3.1) and (3.2) n_{boot} times to generate replicates $\hat{d}^*_1, \ldots, \hat{d}^*_{n_{boot}}$ of $\hat{d}^*$. Let $\hat{d}^*_{(1)} \leq \ldots \leq \hat{d}^*_{(n_{boot})}$ denote the corresponding order statistic.

4. Calculate the CI using one of the following approaches:

   a Percentile CI: Obtain the estimated right bound of the percentile bootstrap CI as the $(1 - \alpha)$-quantile of the bootstrap sample

   $$\hat{u} = \hat{q}^*(1 - \alpha) = \hat{d}^*_{(\lfloor n_{boot}(1-\alpha) \rfloor)}.$$

   b Hybrid CI: Obtain the estimator for the standard error of $\hat{d}$ as $\widehat{\text{se}}(\hat{d}) = \sqrt{\widehat{\text{Var}}(\hat{d}^*_1, ..., \hat{d}^*_{n_{boot}})}$ and the estimated right bound of the hybrid CI as

   $$\hat{u} = \hat{d} + \widehat{\text{se}}(\hat{d}) z.$$

5. Reject the null hypothesis in (11) and assess equivalence if

   $$\varepsilon > \hat{u}.$$

---

4 Finite sample properties

In the following we investigate the finite sample properties of the proposed tests by a simulation study. In order to ensure comparability, we reanalyse the simulation scenarios given by Dette et al. (2018). The dose range is given by $\mathcal{X} = [0, 4]$ and data is observed for dose levels $x = 0, 1, 2, 3$ and 4 with equal number of observations $n_{li} = \frac{n_l}{5}$ for each dose level. All three simulation scenarios use the same three variance configurations $(\sigma^2_1, \sigma^2_2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$ as well as the same four different sample sizes $(n_1, n_2) \in \{(10, 10), (10, 20), (20, 20), (50, 50)\}$ and the same significance level of $\alpha = 0.05$. In the first simulation scenario the equivalence of an emax model and an exponential model is investigated. The other two simulation scenarios consist of testing for equivalence of two shifted models, either both being emax models or both being exponential models. In contrast to the first scenario where the absolute deviation of the models is observed at one unique x_0, this is not the case for the latter two scenarios. Here, the deviation of both models is constant across the whole dose range $\mathcal{X} = [0, 4]$, i.e.

$$|m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2)| = d \ \forall\ x \in \mathcal{X}$$

as m_1, m_2 are just shifted.
Hence, for these two scenarios the asymptotic test is not applicable as its close form variance estimator bases on the uniqueness of x_0. Therefore, only simulation scenario 1 is used to compare the three CI-based tests to each other as well as to the results observed by Dette et al. (2018). Subsequently, all three scenarios are used to compare the performance of the test using model averaging to the one based on the correct specification of the models as well as under model misspecification.

4.1 Finite sample properties of confidence interval-based equivalence testing

Prior to the investigation of the effect of model averaging onto the finite sample properties, we first inspect the performance of the CI-based testing approach, i.e. test (13), itself. For the asymptotic test, the CI is defined by (14). For the percentile as well as the hybrid approach, the tests are conducted as explained in Section 3.1 which is formally defined by setting $K_1 = K_2 = 1$ in Algorithm 1.
As outlined before, simulation scenario 1 of Dette et al. (2018) is given by tesing for the equivalence of an emax model (3) with $\theta_1 = (\beta_{10}, \beta_{11}, \beta_{12}) = (1, 2, 1)$ and an exponential model (4) with $\theta_2 = (\beta_{20}, \beta_{21}, \beta_{22}) = (\beta_{20}, 2.2, 8)$. It consists of 60 sub-scenarios resulting from the three different variance configurations each being combined with the four different sample sizes and five different choices of $\beta_{20} \in \{0.25, 0.5, 0.75, 1, 1.5\}$, leading to the corresponding deviations of the regression curves being $d \in \{1.5, 1.25, 1, 0.75, 0.5\}$. The test is conducted for $\varepsilon = 1$ such that the first three deviations are under the null hypothesis and, therefore, used to investigate the type I error rates. The latter two deviations correspond to the alternative and are used to estimate the power of the tests.
As the type I error rates are always smaller than the nominal level of $\alpha = 0.05$ for all three approaches (see table S1 of the supplementary material for exact values), i.e. all testing approaches always hold the nominal level, the following analysis focuses on the power of the tests. Figure 1 shows the power for all three tests for all sub-scenarios under the alternative as well as the corresponding power of the tests of Dette et al. (2018). In each sub-scenario we observe that the hybrid test has superior power compared to the other two CI-based tests. In addition, one can observe that the power achieved by the hybrid test is quite similar to the one Dette et al. (2018) observed for their bootstrap test and, therefore, is also superior to the power of their asymptotic test. The power of the test based on the percentile CI is considerably smaller which indicates that the test might be overly conservative in finite samples. The test based on the asymptotic CI leads to nearly the same results as the asymptotic test of Dette et al. (2018) which is not surprising as it is directly derived from it. Consequentially, the lack of power that Dette et al. (2018) observed for their asymptotic test in comparison to their bootstrap test, is also present for the test based on the asymptotic CI.
In conclusion, the hybrid approach which provides numerically advantages compared to the bootstrap test of Dette et al. (2018) and also leads to additional interpretability due to providing CIs, achieves nearly the same power as the bootstrap test while holding the nominal level.

4.2 Finite sample properties under model uncertainty

We now investigate the finite sample properties under model uncertainty. Due to the clear superiority of the hybrid approach observed in Section 4.1, only the hybrid test is used for this analysis. We compare the performance of the test using model averaging to the one based on the correct specification of the models as well as under model misspecification. We use frequentist model averaging with AIC-based weights as the inference is frequentist as well. The corresponding equivalence tests are conducted using Algorithm 1b.

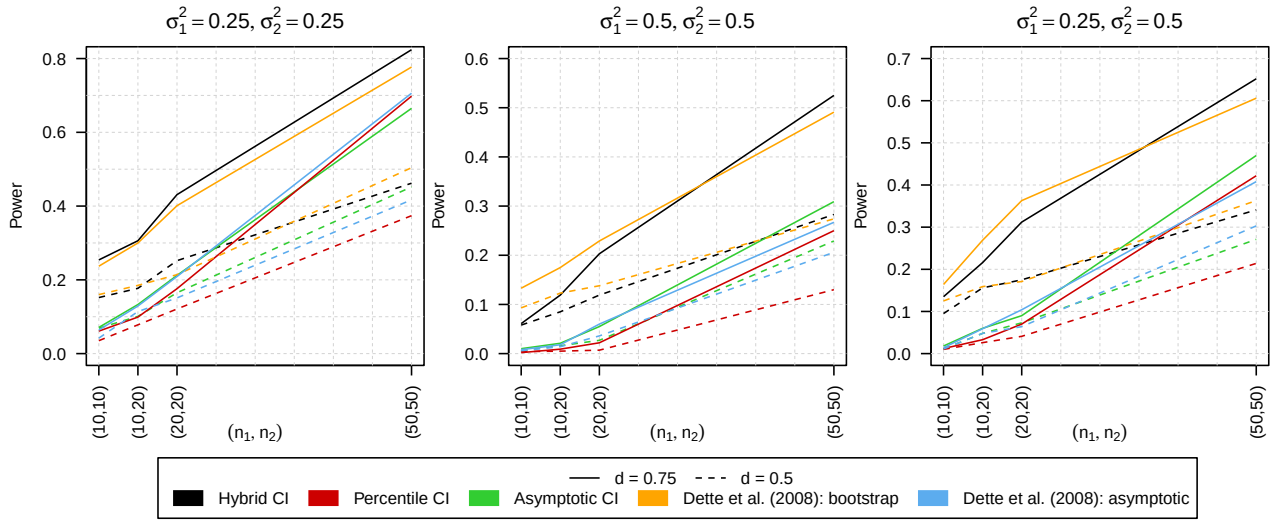

Figure 1: Comparison of the power of the CI-based testing approaches to the testing approaches proposed by Dette et al. (2018) with $\varepsilon = 1$. The results are shown for two distances of the regression curves $d \in \{0.5, 0.75\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.5, 0.5), (0.25, 0.5)\}$.

4.2.1 Comparison of an emax with an exponential model

First, we again investigate the first simulation scenario introduced in Section 4.1 but now under model uncertainty where it is unclear if an emax or an exponential model applies for each of the groups implying $K_1 = K_2 = 2$ and leading to one correct specification, as well as three misspecifications. Figure 2 shows the corresponding type I error rates for all sub-scenarios under the null hypothesis, i.e. for $d \in \{1, 1.25, 1.5\}$ and $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$ for the correct specification, the three misspecifications as well as under model averaging.

One can observe that falsely specifying the same model for both responses leads to highly inflated type I error rates which, in addition, even increase for increasing sample size. The highest type I errors are present if an exponential model is specified for both responses leading to type I error rates being as large as 0.410 which is observed for $\sigma_1^2 = \sigma_2^2 = 0.25$ and $n_1 = n_2 = 50$. If an emax model is specified for both responses, the type I error inflation is smaller but still present and reaches up to 0.187 which is observed for $\sigma_1^2 = \sigma_2^2 = 0.25$ and $n_1 = n_2 = 50$. The third misspecification under investigation is that specifying the models the wrong way round, i.e. an exponential model for the first group and an emax model for the second one. In comparison to the other two misspecifications, this leads to less extreme results but type I error inflation is still observable. This results from the fact that using one convex (exponential) and one concave (emax) model usually leads to a larger maximal absolute deviation than using two convex or two concave models and, therefore, in general to fewer rejections of the null hypothesis.

Compared to these results, the type I errors resulting from model averaging are closer to the nominal level of the test. However, for two out of the 36 investigated sub-scenarios ($\sigma_1^2 = \sigma_2^2 = 0.25$ and $n_1 = n_2 = 20$ as well as $n_1 = n_2 = 50$) the type I errors still exceeds the nominal level but to a much lesser extent compared to model misspecification, reaching a maximum of 0.061. As expected, when using the true underlying model, the test holds the nominal level of $\alpha = 0.05$. Comparison of the power of the tests is not meaningful as some of them are not holding the nominal level. However, the

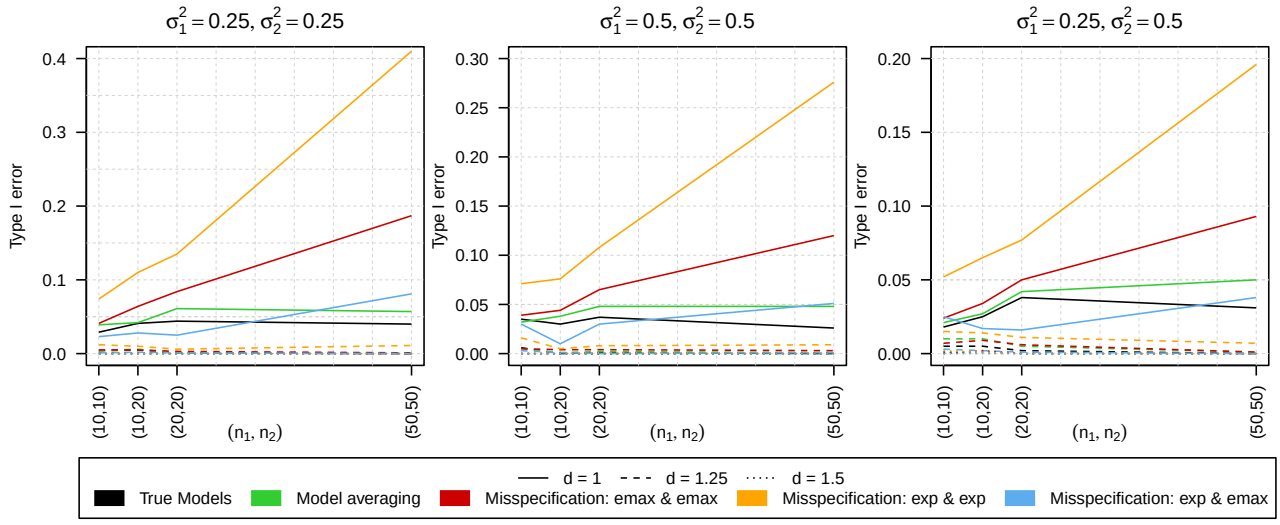

Figure 2: Comparison of the type I error rates of the test using the true model, the model averaging-based test and the tests under model misspecification in scenario 1. The results are shown for $\varepsilon = 1$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

estimated power is shown in table S2 of the supplementary material.

4.2.2 Comparison of two shifted emax models

We continue by investigating the fine sample properties for the case of two shifted emax models, i.e. model 3 now applies for both groups, where $\theta_1 = (\beta_{10}, \beta_{11}, \beta_{12}) = (\beta_{10}, 5, 1)$ and $\theta_2 = (\beta_{20}, \beta_{21}, \beta_{22}) = (0, 5, 1)$, which implies $d = \beta_{10}$. The levels of d under investigation are $1, 0.75, 0.5, 0.25, 0.1$ and 0. The test is conducted for $\varepsilon = 0.5$ such that the first three deviations are under the null hypothesis and, therefore, used to investigate the type I error rates. The latter three deviations are under the alternative and used to estimate the power of the tests.
We only observe few type I error rates which are non-zero and these are still much smaller than the nominal level of $\alpha = 0.05$, reaching a maximum of only 0.003 (all values can be found in table S3 of the supplementary material). Hence, the analysis focuses on the power of the tests which is shown in Figure 3. One can observe falsely specifying one of the models to be an exponential model leads to the power being constantly equal to zero even for sub-scenarios which are quite far under the alternative. For misspecification in terms of using an exponential model for both responses, the power loss is not that extensive but still occurs for smaller sample sizes, which is especially visible for $\sigma_1^2 = \sigma_2^2 = 0.25$ due to the estimation uncertainty being the smallest. In contrast, model averaging results in nearly the same power as using the true model. This also leads to the fact that in some cases the black line is even hardly visual as it is nearly perfectly overlapped by the green one.

4.2.3 Comparison of two shifted exponential models

The third simulation scenario is given by two shifted exponential models, i.e. model 4 now applies for both groups, where $\theta_1 = (\beta_{10}, \beta_{11}, \beta_{12}) = (\beta_{10}, 2.2, 8)$ and $\theta_2 = (\beta_{20}, \beta_{21}, \beta_{22}) = (0, 2.2, 8)$ which implies $d = \beta_{10}$, resulting in the same values for d as in Section 4.2.2. The test is conducted for $\varepsilon = 0.5$

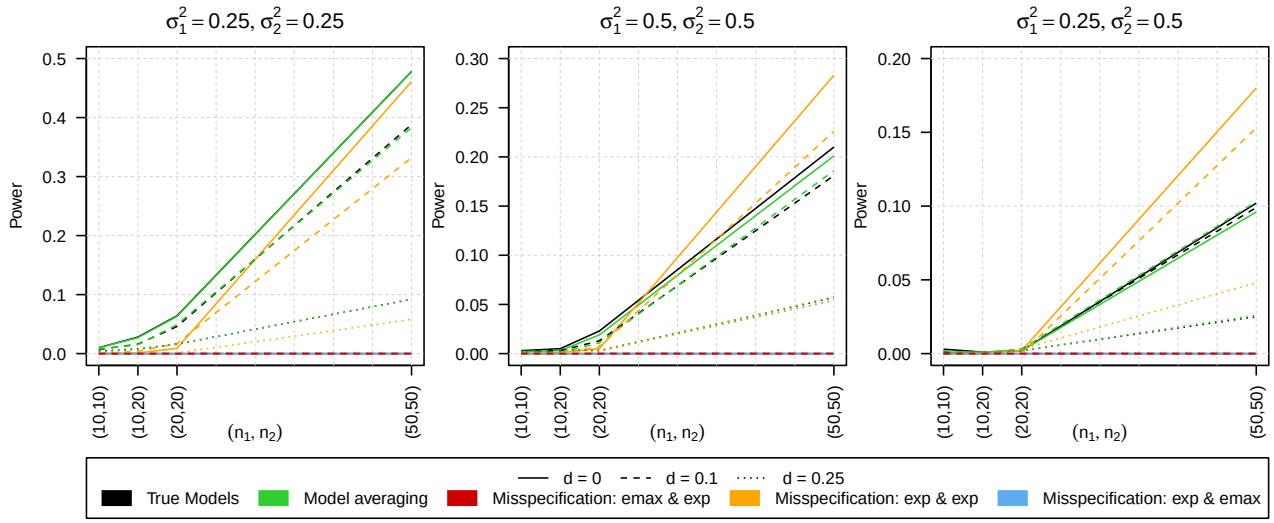

Figure 3: Comparison of the power of the test using the true model, the model averaging-based test and the tests under model misspecification in scenario 2. The results are shown for $\varepsilon = 0.5$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

such that the first three deviations are under the null hypothesis and, therefore, used to investigate the type I error rates. The latter three deviations are under the alternative and used to estimate the power of the tests.
As previously observed in Section 4.2.2 only few type I error rates are non-zero and these exceptions are still much smaller than the nominal level of $\alpha = 0.05$, reaching a maximum of only 0.009 (all values can be found in table S4 of the supplementary material). Hence, the analysis focuses on the power of the tests which is shown in Figure 4. The loss of power resulting from model misspecification is not as large as in Section 4.2.2 but still present. Especially if one of the models is falsely specified to be an emax model but the other one specified correctly, we observe a notable loss of power not only compared to using the true model but also compared to using model averaging. Moreover, this effect is increasing with increasing sample size. In contrast to Section 4.2.2, the power resulting from using model averaging is notably smaller than the one observed when using the true model. However, compared to two out of the three misspecifications, the loss in power is extensively reduced. In conclusion, if the models are misspecified we observe either type I error inflation or a lack of power, both often of substantial extend, in all three scenarios. Using model averaging considerably reduces these problems, often leading to similar results as knowing and using the true underlying model.

5 Case study

We illustrate the proposed methodology through a case study analysing the equivalence of time-response curves (also known as exposure duration-response curves) using data which was originally published by Ghallab et al. (2021). The study aims to investigate dietary effects onto the gene expression. The dataset consists of two groups of mice which were fed with two different diets and then sacrificed at different time points. The first one is a high-fat or "Western" diet (WD) while the other one is a standard diet (SD). As no data has been collected in the first 3 weeks, they are not

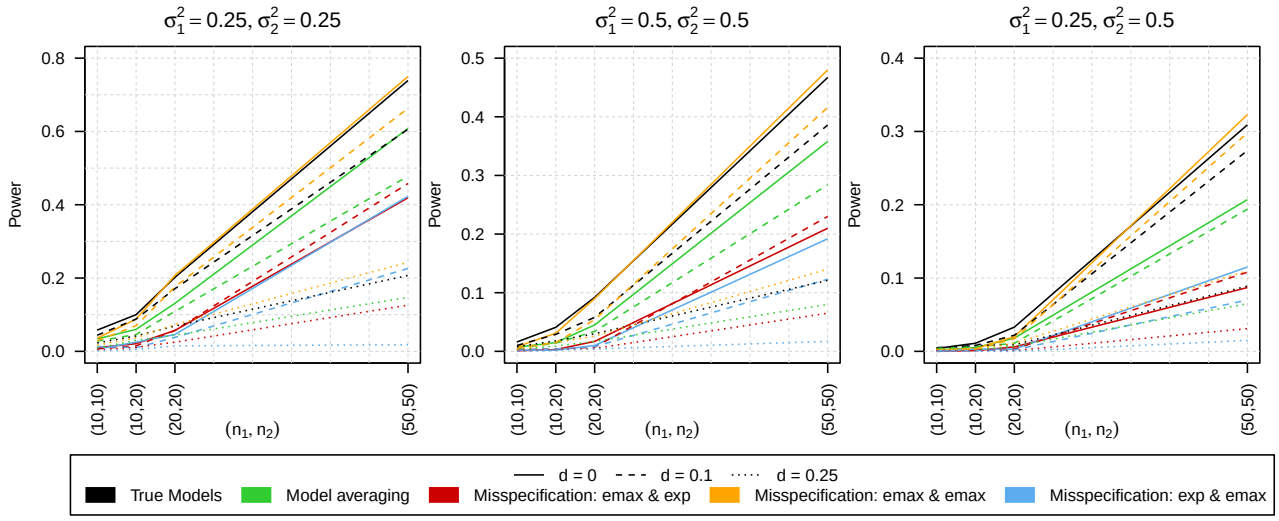

Figure 4: Comparison of the power of the test using the true model, the model averaging-based test and the tests under model misspecification in scenario 3. The results are shown for $\varepsilon = 0.5$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

included into our analysis. Consequentially, the beginning of the study ($t = 0$) resembles week 3 of the actual experiment. Data is then observed at $t = 0, 3, 9, 15, 21, 27, 33, 39$ and 45 for the Western diet and at $t = 0, 3, 27, 33, 39$ and 45 for the standard diet with sample sizes $5, 5, 5, 5, 5, 5, 5, 4, 8$ and $7, 5, 5, 7, 3, 5$, respectively. For each group the gene expression of 20733 genes is measured in terms of gene counts. For our analysis, we focus on the 1000 genes Ghallab et al. (2021) classified as especially interesting due to high activity. Although gene expression is measured as count data, it is treated as continuous due to the very high number of counts. The raw count data is preprocessed in terms of the gene count normalisation conducted by Ghallab et al. (2021) and subsequent $\log_2$-transformation of the normalised counts as suggested by Duda et al. (2022, 2023).

Using this data, we aim to investigate the equivalence of the time-gene expressions curves between the two diets at a 5 % significance level. From Ghallab et al. (2021) is is known that there are quite large differences between the diets, such that for the majority of the genes we expect not to conclude equivalence. However, precisely for this reason it is of interest for which genes equivalence can be concluded nevertheless. As we are interested in the results for each gene separately and are not aiming for a global conclusion, we do not adjust for multiple testing.

As time-response studies are relatively rare, no specific time-response models have been developed. Hence, dose-response models are deployed for time-response relations as well. Methodological review studies (e.g. Kappenberg et al., 2023) do also not distinguish between dose-response and time-response studies. In addition, it seems intuitive that the effects of the high-fat diet accumulate with increasing time of consumption in a similar manner as the effects in dose response-studies accumulate with increasing dose.

As outlined by Ghallab et al. (2021), the dose-response relations vary across genes such that there is no single model which fits to all of them and, hence, model uncertainty is present. In addition, the models cannot be chosen manually due to the high number of genes. We introduce frequentist model averaging using AIC-based weights and the equivalence tests are performed using hybrid CI, i.e. by conducting Algorithm 1b. We deploy the set of candidate models suggested by Duda et al. (2022), i.e.

a linear model (1), a quadratic model (2), an emax model (3), an exponential model (4), a sigmoid emax model (5) and a beta model (6). This set of candidate models can capture quite diverse effects, as it includes linear and nonlinear, increasing and decreasing, monotone and non-monotone as well as convex, concave and sigmoid curves.
The ranges of the response variables, i.e. the ranges of $\log_2$(normalised counts), are not comparable across different genes. Hence, different equivalence thresholds are needed for each of the genes. As such thresholds can not be chosen manually due to the high number of genes, we determine the thresholds as a percentile of the range of the response variable. For a gene $g \in \{1, ..., 1000\}$ that is

$$\varepsilon_g = \tilde{\varepsilon} \left(\max_{l,i}(\hat{y}_{gli}) - \min_{l,i}(\hat{y}_{gli}) \right),$$

where $\tilde{\varepsilon} \in (0, 1)$ is the corresponding percentile and $\tilde{\varepsilon} = 0.2$ or 0.25 would be typical choices. Alternatively, one can proceed the other way around, calculate

$$\tilde{u}_g = \frac{u_g}{\max_{l,i}(\hat{y}_{gli}) - \min_{l,i}(\hat{y}_{gli})}$$

and directly compare $\tilde{u}_g$ to $\tilde{\varepsilon}$, i.e. the decision rules $\varepsilon_g > u_g$ and $\tilde{\varepsilon} > \tilde{u}_g$ are equivalent.
Subfigures (a) and (b) of Figure 5 show boxplots of the model weights for both diets. It can be observed that less complex models, i.e. the linear and quadratic model, have higher weights for the standard diet compared to the Western diet. In contrast, for the two most complex models, i.e. the beta and sigEmax model, the opposite can be observed: they have higher weights for the Western diet compared to the standard diet. In addition, Subfigures (c) and (d) of Figure 5 show histograms of the highest model weight per gene. It can be observed that for the Western diet the model weights tend to be larger compared to the standard diet. In addition, for the Western diet there are notably many genes for which the highest model weight is very close to 1, i.e. the averaged model consists nearly fully of only one of the candidate models.
Subfigure (e) of Figure 5 shows the number of genes for which equivalence of the time-gene expression curves of both diets can be concluded, i.e. the number of genes for which H_0 can be rejected depending on the choice of $\tilde{\varepsilon}$. For very small choices of $\tilde{\varepsilon}$ (e.g. 0.05 or 0.075) equivalence cannot be concluded for any gene and for $\tilde{\varepsilon} = 0.1$ only three genes would be assessed as equivalent. For more typical choices of $\tilde{\varepsilon}$ being 0.2, 0.25 or 0.3, equivalence could be concluded for 37, 68 and 114 genes, respectively. With further increasing $\tilde{\varepsilon}$ the number of rejections also further increases and approaches 1000. However, this is not shown for $\tilde{\varepsilon} > 0.3$ as performing an equivalence test with a threshold larger than 30 % of the range of the response variable might not have practical relevance.
Figure 6 shows the results for three exemplary genes. For ENSMUSG00000095335 it can be observed that both time-response curves are extremely close to each other and that the maximum absolute deviation of the curves is quite small. This leads to $\tilde{u} = 0.098$. Regarding the model weights it can be observed that both time-response curves consist essentially of the same models. For ENSMUSG00000024589 we observe that both time-response curves have a similar shape both being emax-like, although their model weights are not as similar as before. However, their distance is larger than for ENSMUSG00000095335 which leads to $\tilde{u} \approx 0.275$. Hence, the curves are not equivalent for typical choices of $\tilde{\varepsilon}$ being e.g. 0.2 or 0.25 but only for a more liberal choice of $\tilde{\varepsilon} = 0.3$. For the last example ENSMUSG00000029816 we observe that the two curves are completely different with regard to both, shape and location. For the standard diet an almost constant curve is present while for the Western diet a typical emax shape is observable. This is also reflected by the model weights where models which have high weights for one curve, have small ones for the other one and vise versa, the only exempt to this is the emax model which has a medium large weight for both of the groups. Due

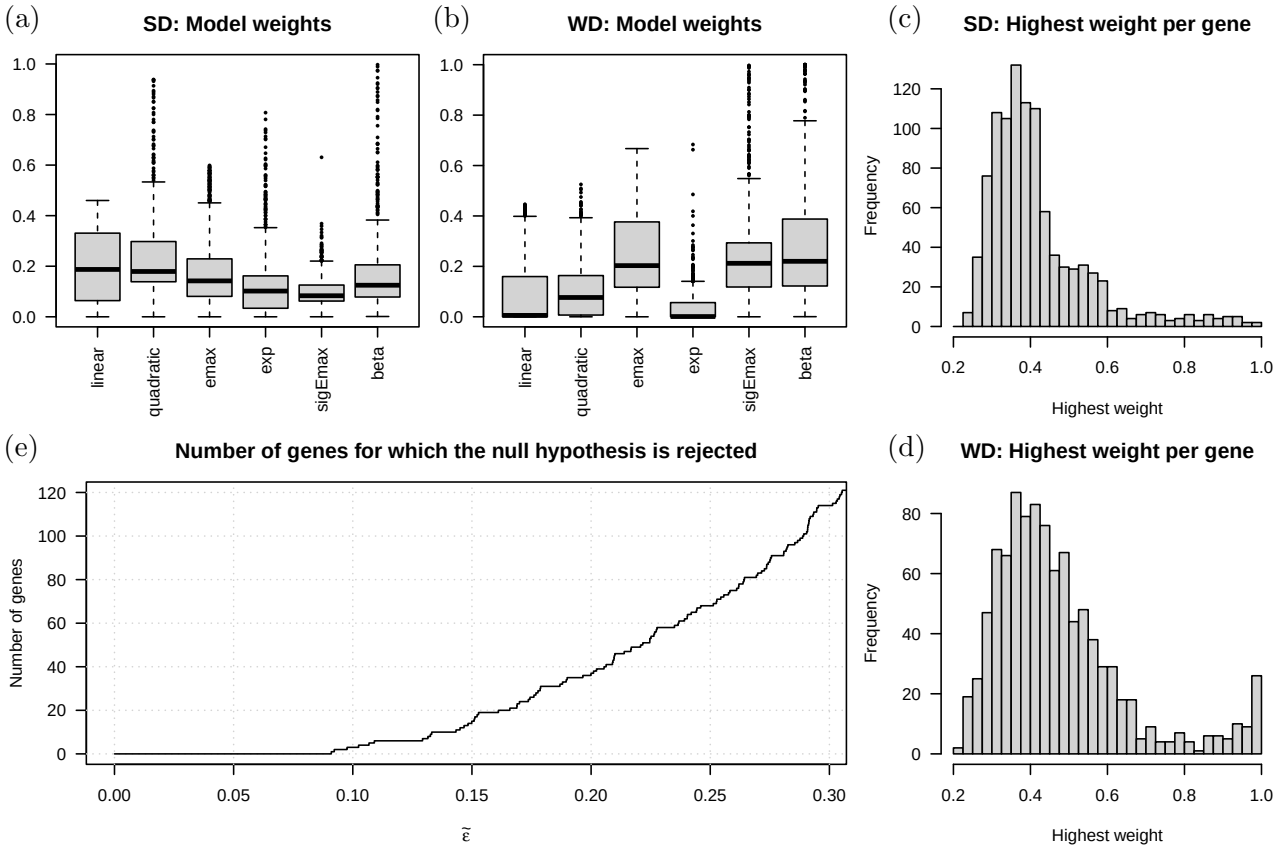

Figure 5: Subfigures (a) and (b) show boxplots of the model weights for each of the two diets. Subfigures (c) and (d) show histograms of the highest model weight per gene for both genes. Subfigure (e) shows the number of genes for which equivalence between the time-gene expression curves of the two diets can be concluded in dependence of the equivalence threshold $\tilde{\varepsilon}$.

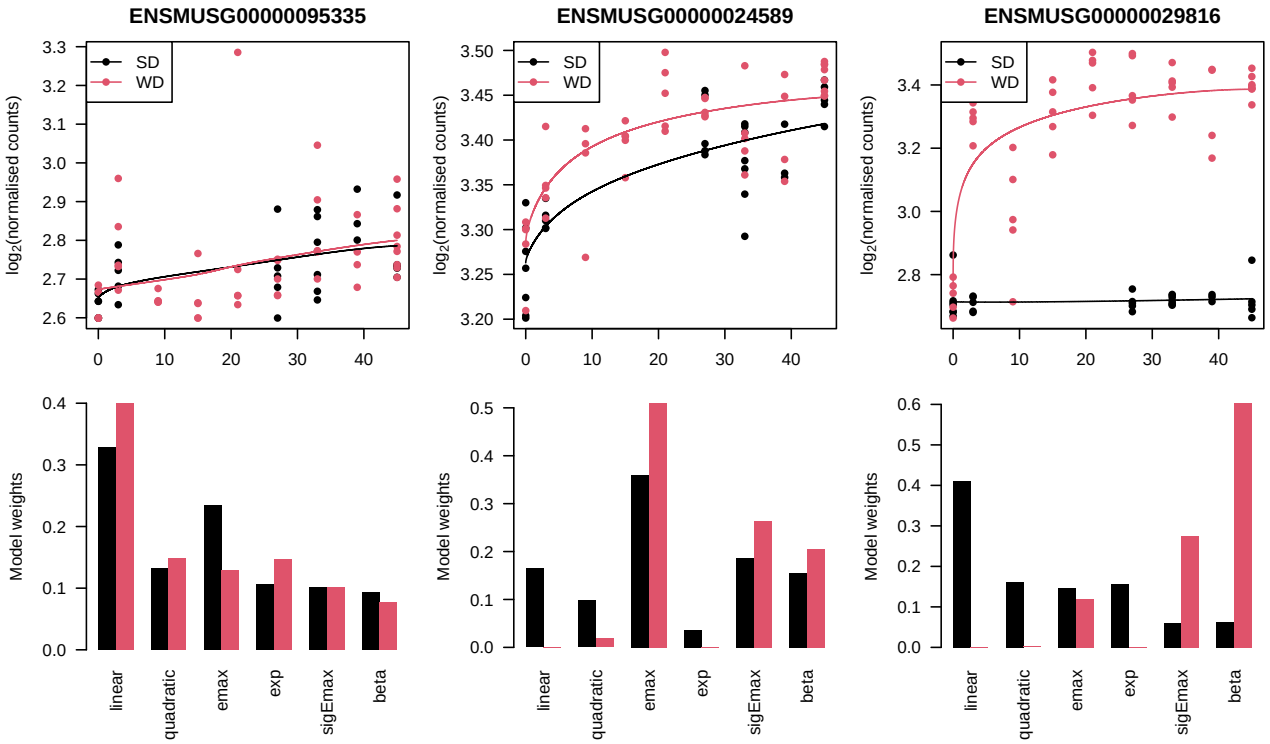

Figure 6: Results for three exemplary genes. The first row of figures shows the data for both diets as well as the fitted models. The second row of figures shows the corresponding model weights.

to the large maximum absolute deviation between the curves given by $\hat{d} = 0.847$, similarity cannot be concluded for any reasonable equivalence threshold.

6 Conclusion

In this paper, we introduced a new approach for model-based equivalence testing which can also be applied in the presence of model uncertainty – a problem which is usually faced in practical applications. Our approach is based on a flexible model averaging method which relies on information criteria and a testing procedure which makes use of the duality of tests and confidence intervals rather than simulating the distribution under the null hypothesis, providing a numerically stable procedure. Moreover, our approach leads to additional interpretability due to the provided confidence intervals while retaining the asymptotic validity and a similar performance in finite samples as the bootstrap based test proposed by Dette et al. (2018).

Precisely, we investigated the finite sample properties of the proposed method by reanalysing the simulation study of Dette et al. (2018) and observed similar results for the CI-based test compared to their test. In the presence of model uncertainty, model misspecification frequently led to either type I error inflation or a lack of power, both often of substantial extend. In contrast, our approach considerably reduced these problems and in many cases even achieved similar results as knowing and using the true underlying model. The presented case study outlines the practical usefulness of the proposed method based on a large data application where choosing the models manually would be time-consuming and could easily lead to many model misspecifications. Hence, introducing model

averaging here is essential to test for the equivalence of time-gene expression curves for such large numbers of genes, typically occurring in practice.
Future possible research includes extending the presented method for other model averaging techniques, e.g. cross validation-based model averaging. In addition, transferring this approach to other model classes (e.g. survival models) as well as to multidimensional responses, i.e. multiple endpoints, merits further exploration.

Software and data availability

Software in the form of R code available at https://github.com/Niklas191/equivalence_tests_with_model_averaging.git. The case study data set is publicly available at the SRA database with reference number PRJNA953810.

Funding

This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project Number 427806116).

Competing interests

The authors declare no competing interests.

References

Aitchison, J. (1964). Confidence-region tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(3):462–476.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723.

Aoki, Y., Röshammar, D., Hamrén, B., and Hooker, A. C. (2017). Model selection and averaging of nonlinear mixed-effect models for robust phase iii dose selection. *Journal of pharmacokinetics and pharmacodynamics*, 44:581–597.

Bastian, P., Dette, H., Koletzko, L., and Möllenhoff, K. (2024). Comparing regression curves – an l^1-point of view. *Annals of the Institute of Statistical Mathematics*, 76(1):159–183.

Bornkamp, B. (2015). Viewpoint: model selection uncertainty, pre-specification, and model averaging. *Pharmaceutical Statistics*, 14(2):79–81.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2009). Mcpmod: An r package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, 29(7):1–23.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383.

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

Bretz, F., Pinheiro, J. C., and Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3):738–748.

Buatois, S., Ueckert, S., Frey, N., Retout, S., and Mentré, F. (2018). Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. *The AAPS journal*, 20:1–9.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2):603–618.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Dennis, B., Ponciano, J. M., Taper, M. L., and Lele, S. R. (2019). Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and aic. *Frontiers in Ecology and Evolution*, 7.

Dette, H., Möllenhoff, K., Volgushev, S., and Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, 113(522):711–729.

Duda, J., Drenda, C., Kästel, H., Rahnenführer, J., and Kappenberg, F. (2023). Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons. *Scientific Reports*, 13:20804.

Duda, J. C., Kappenberg, F., and Rahnenführer, J. (2022). Model selection characteristics when using mcp-mod for dose-response gene expression data. *Biometrical Journal*, 64(5):883–897.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Monographs on Statistics & Applied Probability. Chapman and Hall, New York.

Fletcher, D. (2018). *Model Averaging*. Springer Briefs in Statistics. Springer.

Ghallab, A., Myllys, M., Friebel, A., Duda, J., Edlund, K., Halilbasic, E., Vucur, M., Hobloss, Z., Brackhagen, L., Begher-Tibbe, B., Hassan, R., Burke, M., Genc, E., Frohwein, L. J., Hofmann, U., Holland, C. H., González, D., Keller, M., Seddek, A.-l., Abbas, T., Mohammed, E. S. I., Teufel, A., Itzel, T., Metzler, S., Marchan, R., Cadenas, C., Watzl, C., Nitsche, M. A., Kappenberg, F., Luedde, T., Longerich, T., Rahnenführer, J., Hoehme, S., Trauner, M., and Hengstler, J. G. (2021). Spatio-temporal multiscale analysis of western diet-fed mice reveals a translationally relevant sequence of events during nafld progression. *Cells*, 10(10):1–29.

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

Guhl, M., Mercier, F., Hofmann, C., Sharan, S., Donnelly, M., Feng, K., Sun, W., Sun, G., Grosser, S., Zhao, L., Fang, L., Mentré, F., Comets, E., and Bertrand, J. (2022). Impact of model misspecification on model-based tests in pk studies with parallel design: real case and simulation studies. *Journal of Pharmacokinetics and Pharmacodynamics*, 49(5):557–577.

Hagemann, N., Marra, G., Bretz, F., and Möllenhoff, K. (2024). Testing for similarity of multivariate mixed outcomes using generalised joint regression models with application to efficacy-toxicity responses. *arXiv: 2401.05817 [stat.ME]*.

Hauschke, D., Steinijans, V., and Pigeot, I. (2007). *Bioequivalence Studies in Drug Development*. John Wiley & Sons, Ltd.

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(3):258–264.

Kappenberg, F., Duda, J., Schürmeyer, L., Gül, O., Brecklinghaus, T., Hengstler, J., Schorning, K., and Rahnenführer, J. (2023). Guidance for statistical design and analysis of toxicological dose-response experiments, based on a comprehensive literature review. *Archives of Toxicology*, 97(10):2741 – 2761.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674.

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

Möllenhoff, K., Binder, N., and Dette, H. (2024). Testing similarity of parametric competing risks models for identifying potentially similar pathways in healthcare. *arXiv: 2401.04490 [stat.ME]*.

Möllenhoff, K., Bretz, F., and Dette, H. (2020). Equivalence of regression curves sharing common parameters. *Biometrics*, 76(2):518–529.

Möllenhoff, K., Dette, H., and Bretz, F. (2021). Testing for similarity of binary efficacy-toxicity responses. *Biostatistics*, 23(3):949–966.

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

Möllenhoff, K., Loingeville, F., Bertrand, J., Nguyen, T. T., Sharan, S., Zhao, L., Fang, L., Sun, G., Grosser, S., Mentré, F., and Dette, H. (2022). Efficient model-based bioequivalence testing. *Biostatistics*, 23(1):314–327.

Otto, C., Fuchs, I., Altmann, H., Klewer, M., Walter, A., Prelle, K., Vonk, R., and Fritzemeier, K.-H. (2008). Comparative Analysis of the Uterine and Mammary Gland Effects of Drospirenone and Medroxyprogesterone Acetate. *Endocrinology*, 149(8):3952–3959.

Pinheiro, J., Bornkamp, B., Glimm, E., and Bretz, F. (2014). Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, 33(10):1646–1661.

Pinheiro, J. C., Bretz, F., and Branson, M. (2006). *Analysis of Dose–Response Studies–Modeling Approaches*. Dose Finding in Drug Development. Springer, New York.

Price, M. J., Welton, N. J., Briggs, A. H., and Ades, A. (2011). Model averaging in the presence of structural uncertainty about treatment effects: influence on treatment decision and expected value of information. *Value in Health*, 14(2):205–218.

Schorning, K., Bornkamp, B., Bretz, F., and Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, 35(22):4021–4040.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.