Environmetrics

WILEY

RESEARCH ARTICLE OPEN ACCESS

Stacking Weights and Model Space Selection in Frequentist Model Averaging for Benchmark Dose Estimation

Jens Riis Baalkilde[1] | Niels Richard Hansen[2] | Signe Marie Jensen[1]

[1]Department of Plant and Environmental Science, University of Copenhagen, Copenhagen, Denmark | [2]Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark

Correspondence: Jens Riis Baalkilde (jba@plen.ku.dk)

Received: 26 January 2024 | **Revised:** 23 January 2025 | **Accepted:** 29 January 2025

Funding: This work was supported by NOVO Nordisk Fonden (Grant number NNF21OC0068954).

Keywords: AIC | dose-response analysis | frequentist model averaging | multi-model inference | stacked regression

ABSTRACT

In dose-response modeling, several models can often yield satisfactory fits to the observed data. The current practice in risk assessment is to use model averaging, which is a way to combine multiple models in a weighted average. A key parameter in risk assessment is the benchmark dose, the dose resulting in a predefined abnormal change in response. Current practice when applying frequentist model averaging is to use weights based on the Akaike Information Criterion (AIC). This paper introduces stacking weights as an alternative for dose-response modeling and generalizes a Diversity Index from dichotomous to continuous responses for model space selection. Three simulation studies were conducted to evaluate the new methods. They showed that, in three realistic scenarios, recommended strategies generally performed well, with stacking weights outperforming AIC weights in several cases. Strategies involving model selection were less effective. However, in a challenging scenario, none of the methods performed well. Due to the promising results of stacking weights, they have been added to the R package "bmd."

1 | Introduction

In toxicological and ecotoxicological risk assessment, toxic agents are evaluated in order to determine safe exposure levels. One important tool in risk assessment is the so-called benchmark dose (BMD), which is a dose level resulting in a predefined and adverse deviation from the background response (Haber et al. 2018; Crump 1984). The lower limit of the associated one-sided 95% confidence interval, denoted the benchmark dose lower limit (BMDL), is often used as the starting point for defining limit values of various toxic compounds. The BMD can be defined for continuous, binomial, and count response data as well as time-to-event data (Jensen et al. 2021).

The literature concerning the BMD methodology has developed rapidly during the last few decades (Jensen et al. 2019). Several efforts have focused on how to deal with model misspecification (West et al. 2012). For example, Piegorsch et al. (2014) and Piegorsch et al. (2012) deal with this issue through a nonparametric approach. However, the current practice in BMD estimation is to estimate the BMD by combining several models using a technique called model averaging (MA) (Kang et al. 2000; Namata et al. 2008; Wheeler and Bailer 2009; Ritz et al. 2013; Aerts et al. 2020). MA can be applied in a Bayesian or a frequentist context. This manuscript will focus on frequentist model averaging. Bayesian MA for dose-response modeling is investigated in Wheeler and Bailer (2007), Wheeler et al. (2020), and

Abbreviations: a.i., active ingredients; AIC, Akaike information criterion; BMD, benchmark dose; MA, model averaging.

Wheeler et al. (2022), and the software package ToxicR (Wheeler et al. 2023) is available for utilizing Bayesian MA in dose-response modeling.

MA works by constructing an estimator of the parameter of interest, based on a weighted average of the individual models. This can either be done as a weighted average of the estimates of the parameter of interest within the individual models or by combining the individual models into one MA dose-response model, from which the parameter of interest can be estimated. Both types of MA (referred to as post and curve MA in this paper) are recommended in the literature; however, a comparison of the two types is missing.

One important component in the definition of an MA estimator is the weights. The currently used weights for frequentist MA in BMD estimation are based on the AIC (Akaike Information Criterion Akaike 1973.) values for each of the included models (Ritz et al. 2013; Wheeler and Bailer 2009). However, there are other ways to construct weights. Stacked regression (Breiman 1996) is an MA method that finds the weighted combination of the fitted models with minimal squared prediction error. In contrast to weights based on the AIC values, this approach considers the final fit of the weighted model. The corresponding weights will be denoted stacking weights. Van der Laan et al. (2007) refer to this type of weights as SuperLearner weights. They have shown that the predictive performance of a model based on these weights is at least as good as the best individual model. While the idea is promising, this type of weight has not yet been utilized in frequentist MA for dose-response analysis and BMD estimation.

Another important component in MA is the set of candidate models to include in the weighted average. Current practice in this regard is to use a set of diverse models, often including the classical dose-response models such as the Log-Logistic, the Log-Normal, and the two types of Weibull models. Aerts et al. (2020) extended the set of dose-response models by introducing additional models based on various distribution functions, including the Gamma and Lomax distribution functions. Alternatively, augmenting the model space with models based on fractional polynomials has been suggested (Faes et al. 2003; Namata et al. 2008; Ritz et al. 2013).

Since the AIC weights do not take the fit of the MA model into account, it is suspected that MA using the AIC weights is particularly affected by the choice of model space and the possible inclusion of poorly fitting models in the model space. Wheeler and Bailer (2007) and Das (2018) found that when the true model lies on the boundary of the model space, MA using the AIC weights performed poorly.

Wheeler and Bailer (2009) concluded that in cases where the models in the model space in general did not provide an acceptable fit to the observations, the resulting MA BMD estimates were highly biased, and coverage of the corresponding confidence intervals was poor. In one of the cases, where the MA procedure performed poorly, they tried expanding the model space with so-called "supra-linear" dose-response models, and prior to the MA step, they included a model space selection procedure, where models were only included if the Pearson χ^2 goodness-of-fit statistic was non-significant at a 10% level. This initial screening of the models yielded a huge improvement in the performance of the MA estimate of the BMD.

Kim et al. (2014) suggested that a good model space is characterized by the goodness of fit of the individual models, as well as some degree of diversity around the parameter of interest. Based on this, they suggested a Diversity Index for model space selection, which takes exactly these two properties into account. Their proposed Diversity Index was, however, only defined for a binomial response and increasing dose-response curves, and no extensive study of its properties for MA has been conducted.

This paper takes a critical look at current practice for frequentist MA approaches to BMD estimation by comparing the standard approaches to different alternatives. These alternatives include applying the stacking weights and using alternative model spaces, including a model space selection procedure based on a modified version of the Diversity Index defined by Kim et al. (2014). The methods are illustrated in two small data examples from the literature, and their performance is examined in an extensive simulation study. Furthermore, convergence of the AIC and stacking weights is considered in a separate simulation study.

2 | Theory and Methods

2.1 | Dose-Response Analysis

A dose-response model is a model of the expected response level Y_m given a dose level x_m for a set of observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ (Ritz et al. 2020). Valid dose levels are positive numbers (including 0), and the responses can be continuous or discrete depending on the type of variable. When the response is continuous, a dose-response model is typically of the form

$$Y_m = g(x_m) + \varepsilon_m \tag{1}$$

where $\mathrm{E}(\varepsilon_m) = 0$ and $g : [0, \infty) \to \mathbb{R}$ is the true dose-response curve.

Model averaging using AIC weights, defined in Section 2.2.1, requires specific distributional assumptions about the additive errors ε_m, while stacking weights do not. Stacking weights are, however, most appropriate if the errors have approximately the same variance, and we have throughout the paper taken ε_m to be normally distributed with mean 0 and variance σ^2—both in the computations of AIC weights and in the simulation studies.

In applications, we model the dose-response curve by a (non-linear) parameterized function, g_θ, where the parameter vector, θ, determines the shape of the curve, and these parameters are estimated based on the observations. Common choices of dose-response models are listed in Table 1. They all include lower and upper limits, captured by the parameters c and d, respectively. The fixed parameters p_1 and p_2 in the fractional polynomial models are chosen as variations of $p_1 \in \{-2, -1, -0.5\}$ and $p_2 \in \{0.5, 1, 2\}$ throughout this paper, inspired by Ritz et al. (2013).

TABLE 1 | Popular choices for dose-response models, where Φ in the Log-Normal model is the cumulative distribution function of the standard normal distribution. The parameters p_1 and p_2 in the Fractional Polynomial are fixed parameters that must be set prior to fitting the model. In all models, it is required that $c < d$.

Model name	Abbreviation	Function $g_{b,c,d,e}(x)$
Log-Logistic (Hill)	LL	$c + \frac{d-c}{1+\exp[b(\log x - \log e)]}$
Log-Normal	LN	$c + (d-c)\Phi\left[b(\log x - \log e)\right]$
Weibull 1	W1	$c + (d-c)\exp\left[-\exp(b(\log x - \log e))\right]$
Weibull 2	W2	$c + (d-c)\left[1 - \exp(-\exp(b(\log x - \log e)))\right]$
Fractional Polynomial (logistic link)	FPL(p_1, p_2)	$c + \frac{d-c}{1+\exp[b\log(x+1)^{p_1} + e\log(x+1)^{p_2}]}$

2.1.1 | Benchmark Dose Methodology

The BMD can be defined in several ways. The definition used depends on the type of response variable considered (binomial, count, or continuous) and the aim of the analysis. This paper focuses on the case with a continuous response variable. There are several BMD definitions for continuous data (Crump 1995; Jensen et al. 2019) with the relative risk and the hybrid approach being the two most commonly used.

Let BMR denote the pre-specified Benchmark Response (typically in the range 1%–10%). Let g denote the dose-response function. The Benchmark dose (BMD) by the "relative risk" definition (EFSA Scientific Committee et al. 2022) is the dose level satisfying

$$\text{BMR} = \frac{g(\text{BMD}) - g(0)}{g(0)} \tag{2}$$

for an increasing dose-response curve, and

$$\text{BMR} = \frac{g(0) - g(\text{BMD})}{g(0)} \tag{3}$$

for a decreasing dose-response curve. The hybrid approach (Crump 1995) relies on a background level defined as

$$p_0 = 1 - \Phi\left(\frac{x_0 - g(0)}{\sigma}\right) \tag{4}$$

where x_0 is a prespecified cutoff (for example, a quantile of the response in the unexposed population), σ is the residual standard deviation, and Φ is the cumulative distribution function of the standard normal distribution. The BMD is then defined either by the extra risk definition as the solution to

$$\text{BMR} = \frac{1 - \Phi\left(\frac{x_0 - g(\text{BMD})}{\sigma}\right) - p_0}{1 - p_0} \tag{5}$$

and by the added risk definition as the solution to

$$\text{BMR} = 1 - \Phi\left(\frac{x_0 - g(\text{BMD})}{\sigma}\right) - p_0 \tag{6}$$

The European Food Safety Authority (EFSA) and the United States Environmental Protection Agency (US EPA) recommend the relative risk definition with BMR selected based on biological considerations of an adverse response, or, if no information on a biologically adverse response level is available, a BMR level corresponding to one control standard deviation from the control group (EFSA Scientific Committee et al. 2022; Davis et al. 2011). US EPA et al. (2012) also notes the potential for using the hybrid approach.

For the remaining of this work, we will consider the relative risk definition. However, the methods introduced and evaluated are equally applicable for the hybrid approach as well as other BMD definitions.

To account for uncertainty of data when estimating the BMD value, the lower limit of a (95%) one-sided confidence interval for the BMD estimate, denoted by BMDL, is often used in risk assessment.

2.2 | Frequentist Model Averaging

Let $\mathcal{M} = \{M_1, \ldots, M_k\}$ denote a set of k dose-response models. Let μ be a parameter of interest (for instance, the BMD). Two types of estimators of μ can be constructed by combining the models in $\mathcal{M}$ in a weighted average. Let $w = (w_1, \ldots, w_k)$ denote a set of weights such that $\sum_{i=1}^{k} w_i = 1$ and $0 \leq w_i \leq 1, i = 1, \ldots, k$. The weights can be combined with the individual estimates of the parameter $\hat{\mu}_1, \ldots, \hat{\mu}_k$ in the following way

$$\hat{\mu}_{\text{MA},w} := \sum_{i=1}^{k} w_i \hat{\mu}_i \tag{7}$$

This kind of MA estimator, where the MA step is applied *after* estimation of the parameter of interest, will be referred to as post-MA estimation. Another approach is to combine the individual fitted curves $\hat{g}_1, \ldots, \hat{g}_k$ in an MA curve

$$\hat{g}_{\text{MA},w}(x) := \sum_{i=1}^{k} w_i \hat{g}_i(x) \tag{8}$$

Subsequently, the parameter of interest can be estimated based on this curve. This technique will be referred to as curve MA estimation.

2.2.1 | Weight Choice

Buckland et al. (1997) suggested a set of weights based on AIC (Akaike 1973) values for the fitted models, see also Burnham and Anderson (2003) and Claeskens and Hjort (2008). These weights have been widely adopted in frequentist MA in

dose-response analysis, particularly for BMD estimation. The weights are defined as follows:

$$\hat{w}_i^{\text{AIC}} = \frac{\exp\left(-\frac{1}{2}\text{AIC}_i\right)}{\sum_{j=1}^{k}\exp\left(-\frac{1}{2}\text{AIC}_j\right)} \tag{9}$$

where AIC_i denotes the Akaike Information Criterion for model M_i. In general, $\text{AIC}_i = -2l_i + 2p_i$, where p_i is the number of parameters in model M_i and l_i is the log-likelihood of model M_i evaluated in the maximum-likelihood estimate of the parameters. For a dose-response model of the form (1), a common model of ε_m is a normal distribution with mean 0 and variance σ^2, in which case the maximum-likelihood estimate, $\hat{g}_i$, of the dose-response curve for model M_i is found by least squares, and the corresponding maximum-likelihood variance estimate is

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{m=1}^{n}\left(Y_m - \hat{g}_i(x_m)\right)^2$$

Using the normal model, $-2l_i = n\log(\hat{\sigma}_i^2) + n$ and the AIC weights are in this case given as

$$\hat{w}_i^{\text{AIC}} = \left(1 + \sum_{j\neq i}\left(\frac{\hat{\sigma}_i}{\hat{\sigma}_j}\right)^n e^{p_i - p_j}\right)^{-1}$$

These are the most widely used AIC weights, and they will also be used throughout this paper.

For curve MA, the resulting BMD estimate is a plug-in estimate of the MA curve. In this case, finding the weighted curve that minimizes the mean squared prediction error (MSPE) can be seen as a way to obtain the best-fitting curve. As shown in Lemma S1 in the Supporting Information, the AIC weights generally do not lead to the minimal squared prediction error of the true dose-response curve. The stacking weights, as introduced by Breiman (1996), explicitly do so.

With the fitted curves, $\hat{g}_1, \ldots, \hat{g}_k$, the optimal weights, w^{Stack}, are defined as

$$w^{\text{Stack}} = \underset{\substack{\sum_{i=1}^{k} w_i = 1 \\ 0\leq w_i \leq 1}}{\arg\min} \sum_{m=1}^{n}\left(\sum_{j=1}^{k} w_j \hat{g}_j(x_m) - g(x_m)\right)^2$$

These weights are estimated by V-fold cross-validation as described in detail in Algorithm 1.

The estimated weights can then be applied on the estimated curves, $\hat{g}_1, \ldots, \hat{g}_k$, obtained by fitting all models on the entire data set, resulting in a stacked MA dose-response curve

$$\hat{g}_{\text{MA},\tilde{w}^{\text{Stack}}}(x) = \sum_{j=1}^{k}\tilde{w}_j^{\text{Stack}}\hat{g}_j(x)$$

Alternatively, although less obvious, the stacking weights can be used in post-MA as well.

ALGORITHM 1 | Stacking weights.

1. Randomly split the data into V distinct data sets, denoted $(x^1, Y^1), (x^2, Y^2), \ldots, (x^V, Y^V)$.
2. For $v \in \{1, \ldots, V\}$:
   Fit all models, $\mathcal{M}$, on the data set consisting of all observations excluding (x^v, Y^v). Denote the fitted curves on this data set by $(\hat{g}_1^v, \hat{g}_2^v, \ldots, \hat{g}_k^v)$.
   Then, find the convex combination of the fitted curves that minimizes the squared prediction error on the data set (x^v, Y^v):
   $$\tilde{w}^{\text{Stack},v} := \underset{\substack{\sum_{i=1}^{k} w_j = 1 \\ 0\leq w_j \leq 1}}{\text{argmin}} \sum_{(x_i,Y_i)\in(x^v,Y^v)}\left(\sum_{j=1}^{k} w_j \hat{g}_j^v(x_i) - Y_i\right)^2.$$
3. Return the final weights
   $$\tilde{w}^{\text{Stack}} := \frac{1}{V}\sum_{v=1}^{V}\tilde{w}^{\text{Stack},v}.$$

TABLE 2 | Comparison of AIC weights and stacking weights. The optimal MA curve is defined as the MA curve minimizing the squared prediction error. *See Lemma S1.

	AIC weights	Stacking weights
Currently used in dose-response analysis	Yes	No
Motivation	Information theory	Minimize mean squared prediction error of MA curve
Random for a fixed data set	No	Yes, if V is less than the number of observations, the stacking weights depend on the random split of data
Convergence to optimal MA curve	No*	Yes

Since the stacking weights depend on a random split of the dataset, they are not fixed for a given observation. In contrast, AIC weights are fixed because they depend solely on the density function evaluated at the maximum likelihood estimate. Ideally, the data should be split within the dose levels, such that the design is fixed within each data split. However, if V is larger than the number of repetitions for each dose, this is not possible. Various values of V are investigated in Section 5.

General characteristics of the AIC weights and the stacking weights are summarized in Table 2.

2.3 | Model Space Selection

In MA, the model space plays an important role. Several published simulation studies suggest that it is sufficient to have a

model space consisting of a range of models providing somewhat satisfactory fits to the observations (Wheeler and Bailer 2009; Ritz et al. 2013; Namata et al. 2008; Aerts et al. 2020). However, various schemes for selecting a suitable (sub)set of models are suggested as well (Wheeler and Bailer 2009; EFSA 2011; Kim et al. 2014).

2.3.1 | F-Test

In order to avoid basing the MA estimate on poorly fitting models, a model space selection strategy based on conducting an F-test (inspired by Wheeler and Bailer (2009) where a model space selection procedure based on a χ^2 goodness-of-fit test is used) is investigated. In the following, let $\mathcal{M} = \{M_1, \ldots, M_k\}$ denote the set of fitted models. The procedure works as follows:

For each model $M_i \in \mathcal{M}$,

1. Fit two linear normal models on the residuals from the model. The first model is the normal null model with mean 0 and constant variance. The second model is a one-way ANOVA model with a separate mean value for each dose level and constant variance.
2. Carry out an F-test comparing the two linear normal models. If the F-test yields a p-value above 0.05, model M_i is used in the MA step. If not, the model is discarded.

This procedure is an automated procedure that determines if each model yields an acceptable fit to the observations. The intuition behind the procedure is that if model M_i is the correct model, the normal null model is the correct model for the residuals, and if model M_i is not the correct model, the one-way ANOVA model for the residuals is correct, under the assumption of normality, independence, and constant variance. Consequently, if M_i fits the observations well, the model should be accepted for use in the MA step, and if not, M_i should be discarded.

2.3.2 | A Diversity Index for Model Space Selection

Kim et al. (2014) suggested a Diversity Index (DI) for model space selection in the case of binomial response data and strictly increasing dose-response curves. In the following, their method is extended to a DI for model space selection in the case of continuous response data and either strictly increasing or strictly decreasing dose-response curves.

Definition 1. (Diversity Index). Let $\mathcal{M} = \{M_1, \ldots, M_k\}$ denote the full model space of size k. Let $\mathcal{M}_0 \subseteq \mathcal{M}$ be a subset of the model space with $k_0 = |\mathcal{M}_0| \geq 2$. Let $w = \{w_1, \ldots, w_k\}$ denote a set of weights for each of the k models, and let $\hat{g}_{\mathrm{MA},w}(x) = \sum_{i=1}^{k} w_i \hat{g}_i(x)$ denote the MA curve based on all models. The DI for the subspace $\mathcal{M}_0$ is given by

$$\mathrm{DI}_{w,\lambda,\gamma}(\mathcal{M}_0) := h_\lambda(w)^{1/\gamma} \frac{k}{k_0} \sum_{i:\, M_i \in \mathcal{M}_0} \mathcal{K}^*(\hat{g}_i, \hat{g}_{\mathrm{MA},w}) \qquad (10)$$

where $\mathcal{K}^*$ is the local pseudo Kullback-Leibler absolute divergence, which is given by

$$\mathcal{K}^*(g_0, g_1) := \int_L^U \left| \log\left(\frac{|g_0'(x)|}{|g_1'(x)|} \right) g_0'(x) \right| \mathrm{d}x \qquad (11)$$

for pre-specified lower and upper limits, L and U, where the derivatives of the dose-response curves are taken with respect to the dose x, and

$$h_\lambda(w) = \sum_{i:\, M_i \in \mathcal{M}_0} w_i^\lambda$$

The model space selected by the DI is then

$$\mathcal{M}_0^*(w, \lambda, \gamma) := \underset{\mathcal{M}_0 \subseteq \mathcal{M}, |\mathcal{M}_0| \geq 2}{\arg\max}\ \mathrm{DI}_{w,\lambda,\gamma}(\mathcal{M}_0)$$

The DI includes four tuning parameters: λ, γ, and the upper and lower limits in the local pseudo Kullback-Leibler absolute divergence. Kim et al. (2014) suggested that the latter two are chosen such that they contain the dose range in which diversity among the included models is desired.

The parameters λ and γ can be tuned to control the size of the suggested subspace of models. For a fixed $\gamma > 0$, increasing λ will reduce the selected subspace in such a way that any excluded model does not reappear for a larger value of λ. For a fixed $\lambda > 0$, the selected subspace will converge to the subspace consisting of the two most divergent models in the limit $\gamma \to \infty$ (Kim et al. 2014). Values of λ and γ in the range of 1 to 5 were investigated in this paper. These values were chosen to explore the effectiveness of the Diversity Index (DI) across a spectrum, ranging from cases where few or no models (0–2) were removed from the model space to scenarios where several models (5+) were excluded.

2.4 | Implementation

The methods are implemented in **R** (R Core Team 2024). BMD estimation by MA was implemented in the `bmd` package (Jensen et al. 2020a), and recently the stacking weights were added to the package for all available BMD definitions by the authors of this paper. The `bmd` package is available at GitHub in the www.github.com/doseResponse repository. The convex optimization in the estimation of the stacking weights was implemented using the `CVXR` package (Fu et al. 2017). An implementation of the DI is included in the Supporting Information.

3 | Data Example I: Lemna Minor Treated With Aciflourfen

The methods described in the previous section are demonstrated on a data example with observations of relative growth rate of *Lemna minor* treated with one of eight doses of the herbicide Aciflourfen and untreated control observations. Three replicates were included for each of the dose levels, and 12 replicates were included for the control level. The plants grew for seven days after the application of the treatment. The plants were then photographed next to a 1 cm^2 white square with a digital camera, and the frond area was determined by pixel counts. The experiment was described in detail in Cedergreen and Streibig (2005). The full data set is available in the `R` package `drcData`.

3.1 | Analysis

In the analysis of the Aciflourfen data, the model space $\mathcal{M}$ consisted of eight models, including the Log-Logistic, Log-Normal,

two types of Weibull models, and four Fractional Polynomial models with all combinations of fixed parameters $p_1 \in \{-2, -1\}, p_2 \in \{0.5, 1\}$ (see Table 1). All models were fitted in the four-parameter version where the lower and upper limits of the dose-response curve were estimated. The corresponding fitted dose-response curves and all observations can be seen in Figure 1.

The AIC weights and the stacking weights for the fitted models are listed in Table 3. Three-fold cross-validation was chosen since there were three replicates per treatment, ensuring a fixed design for each data split. In this case, the AIC weights favored the Log-Logistic, the Log-Normal, and the Weibull 1 models, while the stacking weights favored the Log-logistic, Weibull 1, and FPL$(-2, 1)$ models. Post as well as curve MA were applied for BMD estimation.

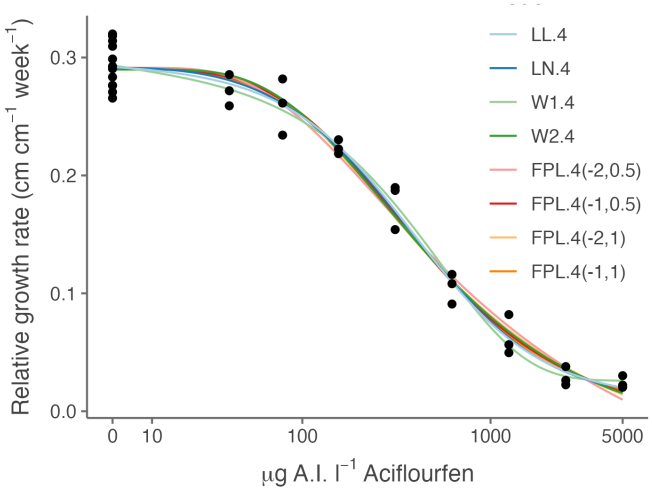


FIGURE 1 | All models fitted to the Aciflourfen data. The model abbreviations are explained in Table 1. All models were fitted in their four-parameter version, where the upper as well as the lower limit of the curve were estimated.

Model space selection by the DI was applied for $\lambda \in \{1, 3, 5\}$ and $\gamma = 1$ with the AIC weights. The lower and upper limits used in the pseudo Kullback-Leibler divergence were $L = 10$ and $U = 1000$, respectively.

The parameter of interest was chosen to be the BMD from the "relative risk" definition (3) with BMR = 10%. Table 4 shows the resulting BMD estimates and BMDL values from applying the different MA methods for BMD estimation. The resulting BMDL values were estimated by non-parametric bootstrap using the percentile method (Tibshirani and Efron 1993) based on 500 resampled data sets. The model weights were considered as a part of the model and were recomputed for each resampled data set.

The MA methods yielded quite similar BMD estimates ranging from 64.69 (curve MA, AIC weights on DI reduced subspace with $\lambda = 5$, Table 4) to 67.67 (curve MA, stacking weights), and similar BMDL values ranging from 44.04 (curve MA, AIC weights on DI reduced subspace with $\lambda = 5$) to 48.33 (curve MA, stacking weights). In contrast, the single model selected by the minimum AIC value yielded a BMD estimate of 58.16 and a BMDL value of 43.29. The DI model space selection removed the FPL$(-2, 0.5)$ model for $\lambda = 1, \gamma = 1$. For $\lambda = 3$, the FPL$(-2, 0.5)$ and FPL$(-1, 1)$ models were removed, and for $\lambda = 5$, the FPL$(-2, 0.5)$, FPL$(-1, 1)$, and Log-Normal models were removed from the model space, which resulted in slightly different BMD and BMDL estimates compared to the curve MA estimate using the AIC weights (Table 4).

TABLE 3 | AIC and stacking weights on the Aciflourfen data example. The model abbreviations are explained in Table 1. All models were fitted in their four-parameter version, where the upper as well as the lower limit of the curve were estimated.

	LL	LN	W1	W2	FPL(−2,0.5)	FPL(−1,0.5)	FPL(−2,1)	FPL(−1,1)
$\hat{w}^{\text{AIC}}$	0.31	0.18	0.34	0.02	0.01	0.05	0.01	0.08
$\tilde{w}^{\text{Stack}}$	0.31	0.00	0.51	0.00	0.02	0.00	0.16	0.00

TABLE 4 | Benchmark dose (BMD) and benchmark dose lower limit (BMDL) estimates on Aciflourfen data for the selection of model averaging (MA) methods included in the analysis. The BMD based on the "relative risk" definition (3) with BMR = 10% was estimated. Single model refers to model selection based on the lowest AIC value. MA estimates were based on all eight models included in the analysis. The Diversity Index (DI) was used for model space selection for different values of (λ, γ). Two different Fractional Polynomial models (FPL) were removed from the model space by application of the DI. The stacking weights were estimated based on three-fold cross-validation.

MA type	w	Model space	λ	γ	$\widehat{\text{BMD}}$	$\widehat{\text{BMDL}}$
Single model	—	min(AIC)	—	—	58.16	43.29
Post	$\hat{w}^{\text{AIC}}$	$\mathcal{M}$	—	—	66.49	45.16
Curve	$\hat{w}^{\text{AIC}}$	$\mathcal{M}$	—	—	66.82	45.19
Post	$\tilde{w}^{\text{Stack}}$	$\mathcal{M}$	—	—	66.76	47.50
Curve	$\tilde{w}^{\text{Stack}}$	$\mathcal{M}$	—	—	67.67	48.33
Curve (DI)	$\hat{w}^{\text{AIC}}$	$\mathcal{M} \setminus \{\text{FPL}(-2, 0.5)\}$	1	1	66.82	44.09
Curve (DI)	$\hat{w}^{\text{AIC}}$	$\mathcal{M} \setminus \{\text{FPL}(-2, 0.5); \text{FPL}(-1, 1)\}$	3	1	66.39	45.08
Curve (DI)	$\hat{w}^{\text{AIC}}$	$\mathcal{M} \setminus \{LN; \text{FPL}(-2, 0.5); \text{FPL}(-1, 1)\}$	5	1	64.70	44.04

4 | Data Example II: Mixture Experiment With Lemna Minor Treated With Mixtures of Aciflourfen and Diquat

The data used for the analysis in Section 3 was a subset of the data from a herbicide mixture experiment, including herbicides with Aciflourfen and Diquat as their active ingredients. In the full mixture experiment, seven mixtures of the two herbicides were included in eight dilutions. Since different herbicides require different concentrations to achieve the same effect on the considered endpoint, the original unit (μg a.i. l^{-1}) for each herbicide is often transformed to a common scale for both herbicides, based on an exchange ratio, which is usually derived from previously estimated ED50 (dose resulting in 50% effect) levels for each herbicide (Sørensen et al. 2007). In this data set, the Diquat concentrations were multiplied by an exchange ratio of 10, resulting in a dose given in μg a.i. l^{-1} Aciflourfen + 10 μg a.i. l^{-1} Diquat. The mixtures were given as the percentage of the mixture composed of Aciflourfen, meaning that for a dose of 10 with a 50% mixture of Aciflourfen, it consisted of 5 μg a.i. l^{-1} Aciflourfen and 50 μg a.i. l^{-1} Diquat.

4.1 | Analysis

The same set of models that was used in Section 3 was used for analyzing the full data set, meaning the following set of models was fitted

$$\mathcal{M} = \{\text{LL}, LN, \text{W1}, \text{W2}, \text{FPL}(-2, 0.5), \text{FPL}(-1, 0.5), \text{FPL}(-2, 1), \text{FPL}(-1, 1)\}$$

For each of the seven mixtures, a joint model was fitted with a separate dose-response curve for each mixture. All models were fitted with different b, c, e parameters for each mixture and a common d (upper asymptote) parameter for all mixtures. A common upper asymptote was chosen since the upper asymptote corresponds to the response at dose 0, where the mixture is irrelevant. The fitted curves for all models in the model space can be seen in Figure 2.

The AIC weights and the stacking weights based on three-fold cross-validation are listed in Table 5. Once again, three-fold cross-validation was chosen since there were three replicates per treatment, so this choice ensures a fixed design for each data split. For the models on the mixture data, the AIC weights were close to 1 for the Weibull 2 model and near 0 for the remaining models, while the stacking weights favored the Weibull 1 and 2 models and the fractional polynomial model with fixed parameters $p_1 = -2, p_2 = 1$.

The resulting BMD and BMDL estimates are listed in Table 6. For each mixture, the BMD and BMDL estimates were relatively similar for all applied estimation procedures. The BMD estimates based on the AIC weights were almost identical to the BMD estimates taken from the model with the lowest AIC value, since the weight for one single model was close to 1 (Table 5). However, the corresponding BMDL estimates differed. The BMD estimates based on MA with the stacking weights were lower than the MA BMD estimates using the AIC weights. The BMD and BMDL estimates were quite different between the different mixtures, with the largest values seen for the mixture percentages 33 and 50, suggesting an antagonistic effect of the two herbicides (Table 6).

5 | Simulation Study I

A small simulation study was conducted to see the effect of varying the number of data splits used in the cross-validation part of the stacking weights.

The true curve was chosen as the four-parameter Log-Logistic curve fitted to the Aciflourfen data in Section 3. The observations were simulated with independent normally distributed residuals with a standard deviation of 0.0015, which is similar to the observed standard deviation in the Aciflourfen data in Section 3. Two dosing scenarios were considered. In the first scenario, $n_{\text{rep}} = 5$ repetitions per dose level were simulated, with the dose levels being

$$\mathbf{x}_5 = (0, 78.125, 312.5, 1250, 5000)$$

In the second scenario, $n_{\text{rep}} = 10$ repetitions per dose level were simulated, with the following dose levels

$$\mathbf{x}_{10} = (0, 0, 39.0625, 78.125, 156.25, 312.5, 625, 1250, 2500, 5000)$$

The dose levels included in $\mathbf{x}_{10}$ were the same levels as in the Aciflourfen data analyzed in Section 3.

The BMD value based on the "relative risk" definition (3) with BMR = 10% was estimated based on a total of 17 models, including three- and four- parameter versions of the Log-Logistic, Log-Normal, Weibull 1, and Weibull 2 models, as well as Fractional polynomial models with all combinations of $p_1 \in \{-2, -1, -0.5\}$ and $p_2 \in \{0.5, 1, 2\}$. The true BMD value was 71.98 in this simulation study.

MA estimates of the BMD value were considered based on the stacking weights using two-fold, five-fold, ten-fold, and leave-one-out (LOO) cross validation. Post-MA estimation as well as curve MA estimation were applied.

For each configuration, $R = 500$ data sets were simulated. The confidence intervals were computed by percentile non-parametric bootstrapping, where the data set was resampled $B = 500$ times, and for each resampled data set, all strategies were employed to estimate the BMD. Then, the BMDL value was computed as the 5% quantile in the estimated BMD values on the resampled data sets. The weights were considered as a part of the model, and were recomputed for each resampled data set. All resampled data sets were found by resampling within the dose levels to obtain new data sets within the fixed design framework. The performance of the various strategies was assessed by considering the mean bias and the root mean squared error (RMSE) of the BMD estimates, as well as the coverage of the one-sided confidence intervals defining BMDL.

5.1 | Results

Overall, the stacking weights outperformed the AIC weights in terms of bias, RMSE, and coverage of BMDL (Figure 3). The

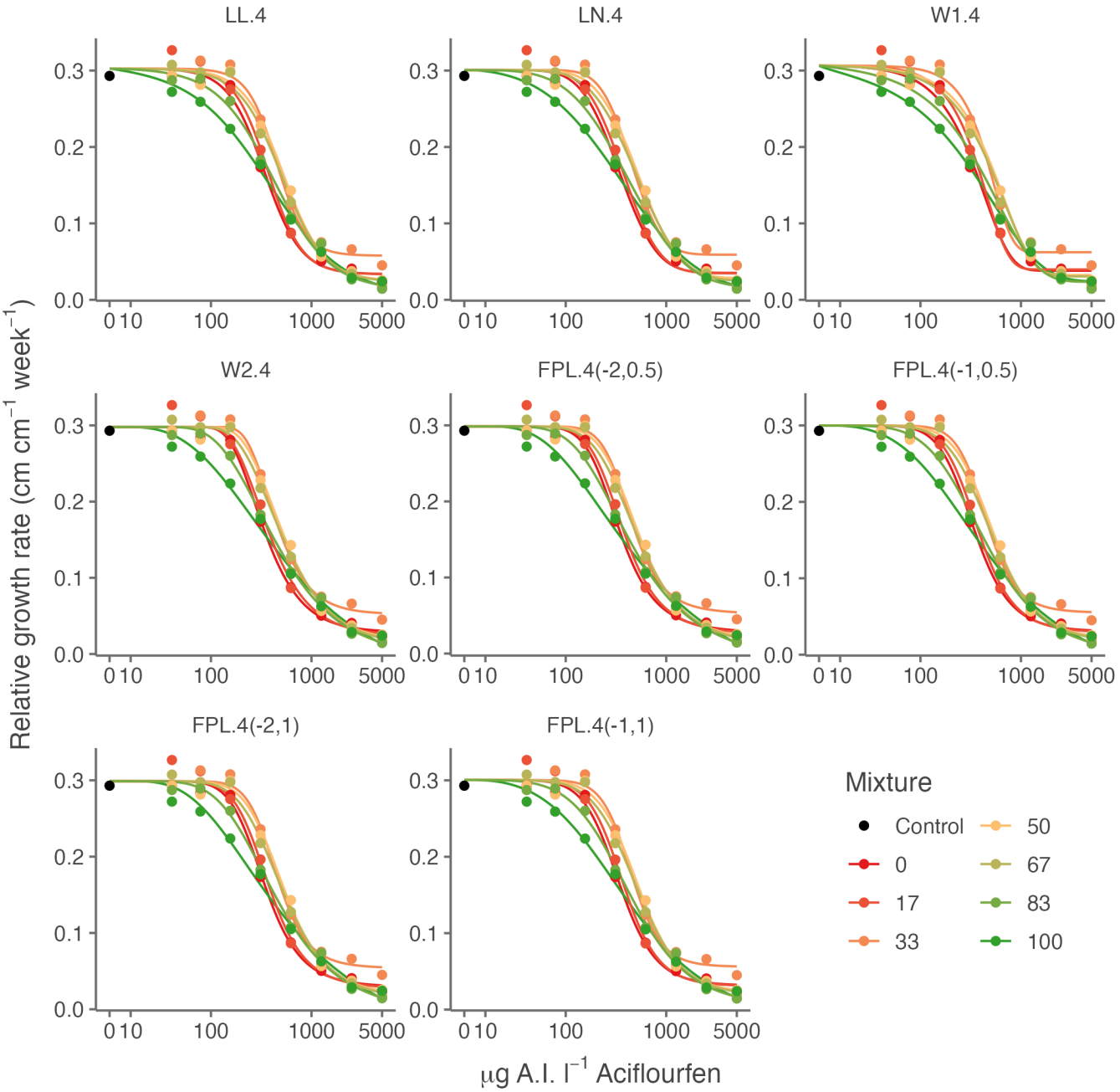


FIGURE 2 | Curves fitted to herbicide mixture data. The plotted points are the mean values for each dose and mixture. For all models, the curves are fitted with a common upper asymptote, while the remaining parameters in the model depend on the mixture.

TABLE 5 | AIC and stacking weights on the full mixture data example. The model abbreviations are explained in Table 1.

	LL	LN	W1	W2	FPL(−2,0.5)	FPL(−1,0.5)	FPL(−2,1)	FPL(−1,1)
$\hat{w}^{\text{AIC}}$	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
$\tilde{w}^{\text{Stack}}$	0.00	0.00	0.13	0.82	0.00	0.00	0.05	0.00

only exception being when the stacking weights were based on two-fold cross-validation, in which case, the resulting RMSE of the BMD estimates was slightly larger than the BMD estimates based on the AIC weights. The overall best performance was observed for the stacking weights based on LOO cross validation.

As expected, lower bias and RMSE were observed for all methods in the scenario with the highest number of observations. The coverage of BMDL was below the nominal level in both scenarios. For several estimation methods, BMDL coverage was higher for the scenario with five repetitions and five dose levels than for the scenario with ten repetitions and ten dose levels.

TABLE 6 | Benchmark dose (BMD) and benchmark dose lower limit (BMDL) estimates on mixture data for the different model averaging (MA) methods included in the analysis. The BMD based on the "relative risk" definition (3) with BMR = 10% was estimated. BMD and BMDL estimates are listed for each mixture denoted as the percentage of the mixture composed of Aciflourfen. Single model refers to model selection based on the lowest AIC value. MA estimates were based on all eight models included in the analysis. The stacking weights were estimated using three-fold cross-validation.

		0		17		33		50		67		83		100	
MA type	w	**BMD**	**BMDL**	**BMD**	**BMDL**	**BMD**	**BMDL**	**BMD**	**BMDL**	**BMD**	**BMDL**	**BMD**	**BMDL**	**BMD**	**BMDL**
Single model	—	171.52	143.95	180.19	164.68	252.35	231.57	235.62	214.76	208.71	191.37	132.59	110.35	69.42	50.52
Post	$\hat{w}^{\text{AIC}}$	171.49	124.51	180.18	160.94	252.33	222.67	235.60	195.92	208.69	178.16	132.58	104.79	69.43	49.49
Curve	$\hat{w}^{\text{AIC}}$	171.50	124.52	180.19	161.03	252.33	222.98	235.60	196.01	208.69	178.55	132.58	104.79	69.43	49.49
Post	$\tilde{w}^{\text{Stack}}$	164.09	87.13	176.44	161.96	247.98	198.23	226.15	203.70	200.19	173.78	125.25	117.71	66.14	42.38
Curve	$\tilde{w}^{\text{Stack}}$	166.25	88.07	177.54	162.82	249.32	198.49	228.80	206.84	202.48	176.09	126.84	118.97	66.71	42.93

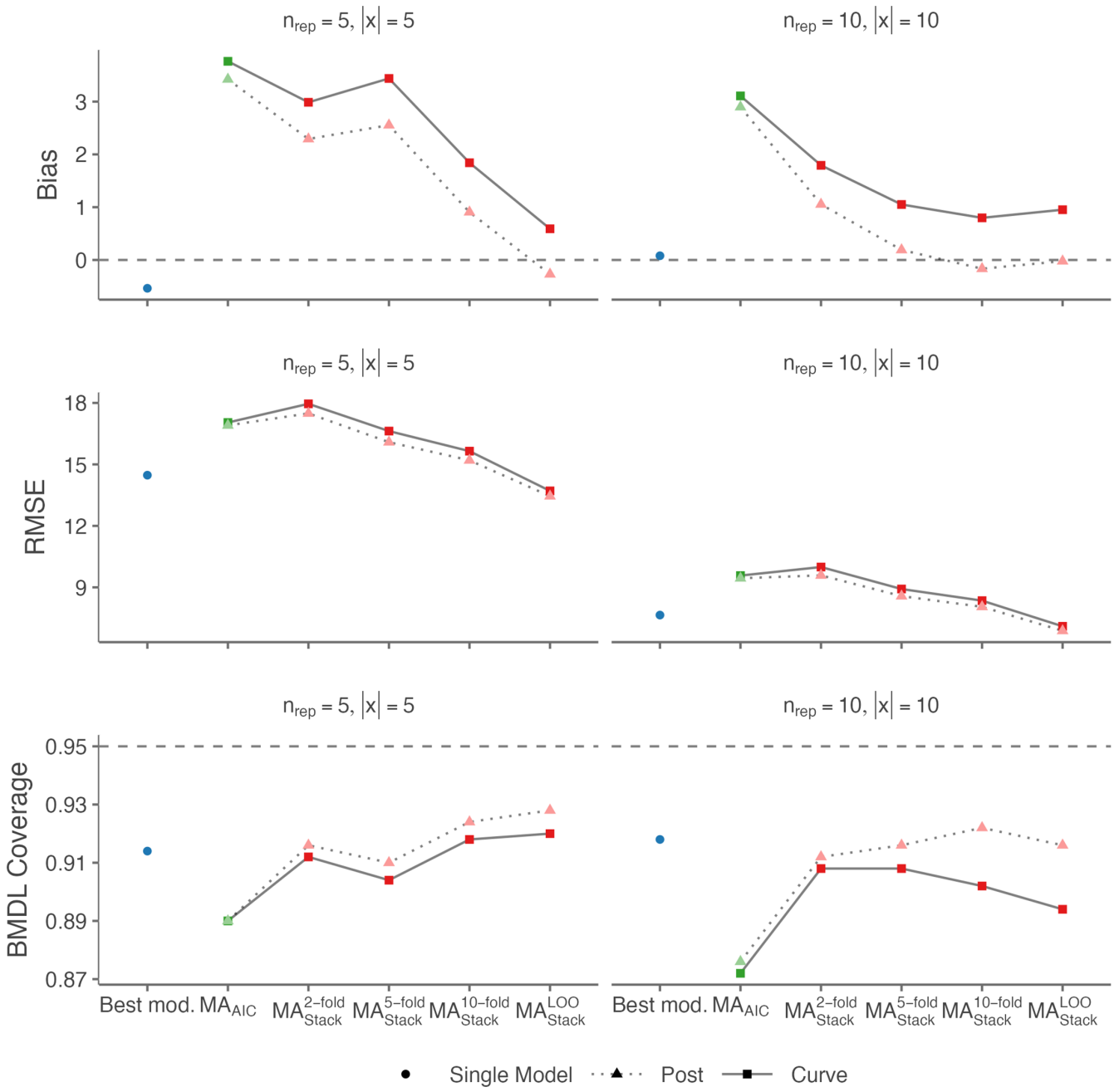


FIGURE 3 | Observed bias, RMSE, and BMDL coverage in simulation study I. Blue points are estimators based on a single model, while green points are MA estimators based on AIC weights, and red points are MA estimators based on the stacking weights.

Generally, a larger number of folds for the stacking weights led to smaller bias and RMSE. However, this came at the expense of longer computation times, since more splits of the data set meant more models needed to be fitted to larger data sets. This was particularly intense for LOO cross-validation.

6 | Simulation Study II

To assess the performance of the various strategies for estimating the BMD, a simulation study was conducted. The strategies included different MA strategies, such as different choices of weights, post vs. curve MA, as well as different choices of model spaces. The parameter of interest was the BMD from the "relative risk" definition (3) with BMR = 10%.

6.1 | Simulation Setup

Four different data-generating models were considered in this simulation study. In all four setups, data were simulated with five and ten repetitions per dose level from the two dose vectors $\mathbf{x}_5$ and $\mathbf{x}_{10}$ used in simulation study I.

The four setups were:

- **Setup A**: The true curve was the four-parameter Log-Logistic model fitted to the Aciflourfen data in Section 3. Residuals were independent and normally distributed with standard deviation $\sigma = 0.015$. The true BMD value in this setup was 71.98.
- **Setup B**: The true curve was the MA curve based on the AIC weights and all models fitted to the Aciflourfen data in Section 3. Residuals were independent and normally distributed with standard deviation $\sigma = 0.015$. The true BMD value in this setup was 70.25.
- **Setup C**: The true curve was a monotonic spline function based on an I-spline basis (Wang and Yan 2021) with coefficients chosen based on the Aciflourfen data. Residuals were independent and normally distributed with standard deviation $\sigma = 0.015$. The true BMD value in this setup was 52.62.
- **Setup D**: The true curve was a monotonic spline function based on an I-spline basis with coefficients chosen to obtain a steeper and more irregular dose-response curve than in setup C. Residuals were independent and normally distributed with standard deviation $\sigma = 0.03$. The true BMD value in this setup was 71.98.

The models fitted to the simulated data sets included three- and four-parameter versions of the Log-Logistic, Log-Normal, and the two types of Weibull models, as well as a total of nine Fractional Polynomial models with all combinations of $p_1 \in \{-2, -1, -0.5\}$ and $p_2 \in \{0.5, 1, 2\}$. The true dose-response curves and all models fitted to data with no residual variance are shown in Figure 4. The bias of the BMD estimate from each model fitted to data from each of the four scenarios with no residual variance can be seen in Table S1 in the Supporting Information.

6.1.1 | Model Averaging Strategies

A range of strategies were examined in this simulation study, including strategies currently used in available software packages, strategies recommended in the literature, as well as strategies based on the stacking weights and the DI (which has not been used before in BMD estimation with a continuous response). The full list of investigated strategies is described below.

Single Model

- min(AIC): The best-fitting model was selected based on the lowest AIC value.
- Best Model: The model for which the BMD in the optimal parameters was closest to the true BMD (see Table S1) was identified as the best model. In setup A, this model was also the true model.
- Worst Model: The model for which the BMD in the optimal parameters was furthest away from the true BMD (see Table S1) was identified as the worst model.

Weight Choice on Full Model Space

- MA_{AIC}(post): The BMD in all fitted models was computed, and the AIC weights were used to construct the MA estimate.
- MA_{AIC}(curve): The MA curve based on all models weighted with the AIC weights was constructed, and the BMD was estimated from this curve.
- MA_{Stack}(post): The BMD in all fitted models was computed, and the stacking weights were used to construct the MA estimate.
- MA_{Stack}(curve): The MA curve based on all models weighted with the stacking weights was constructed, and the BMD was estimated from this curve.

Model Space Selection

- Classical Models: The BMD was estimated by post and curve MA on the model space $\mathcal{M}_{\text{ClM}}$ consisting of three- and four-parameter versions of the Log-Logistic, Log-Normal, and Weibull 1 and 2 models, using the AIC weights as well as the stacking weights.
- Fractional Polynomials: The BMD was estimated by post and curve MA on the model space $\mathcal{M}_{\text{FPL}}$ consisting of four-parameter Fractional Polynomial models with a logistic link function and fixed parameters $p_1 \in \{-2, -1, -0.5\}$ and $p_2 \in \{0.5, 1, 2\}$ using the AIC weights as well as the stacking weights.
- F-Test: An F-test was conducted on the residuals from each model as described in Section 2.3.1. All models where the p-value from the F-test was above 0.05 were used in the following curve MA step using the AIC weights.
- $\text{MA}_{\text{AIC}}\text{DI}(\lambda, \gamma)$: For all combinations of $\lambda \in \{1, 3, 5\}$ and $\gamma \in \{1, 5\}$, the subset of models with the highest DI based on the AIC weights was found. The lower and upper limits used in the DI were 10 and 1000, respectively. The BMD was estimated from the MA curve based on this subset of models using the AIC weights.

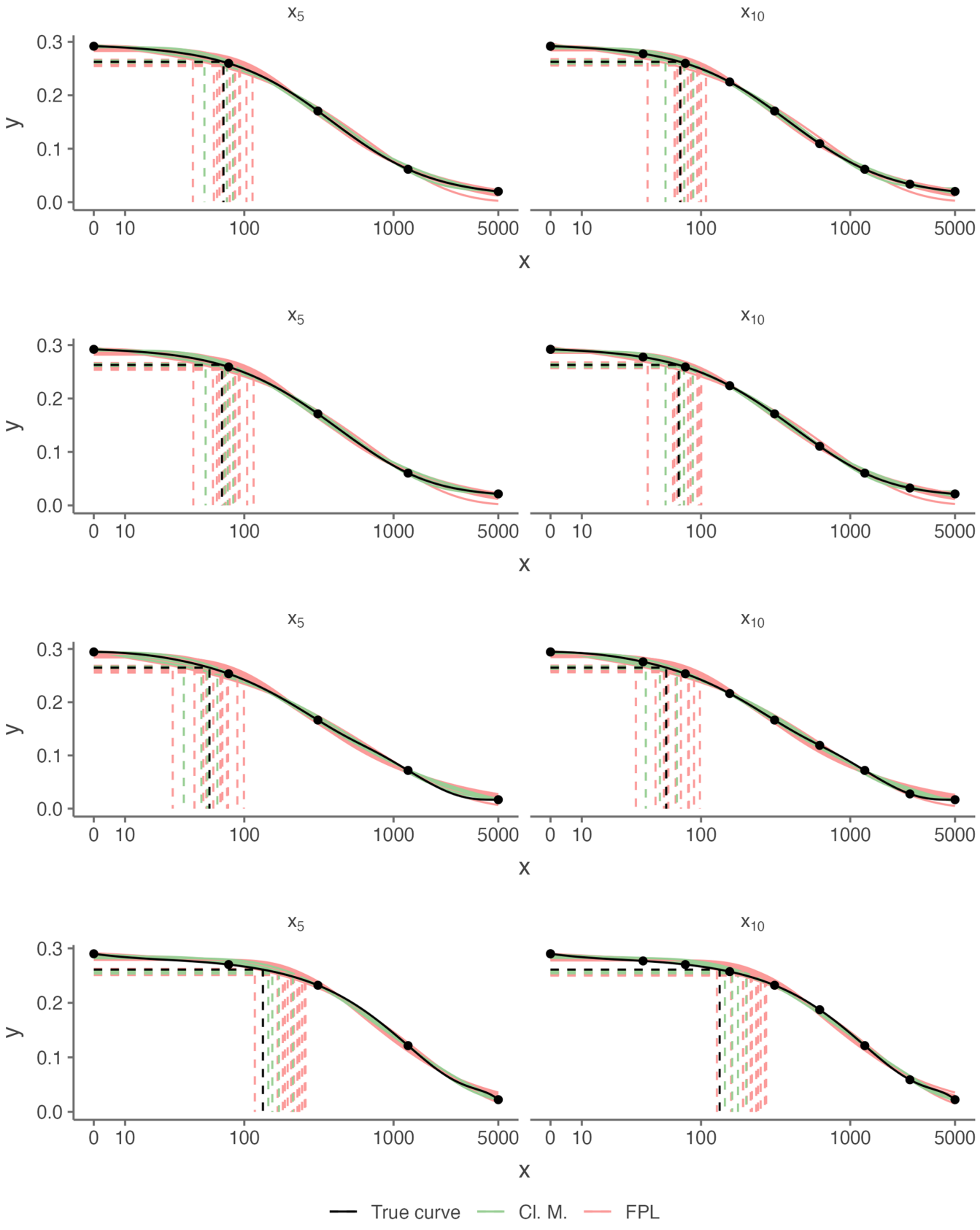


FIGURE 4 | True underlying dose-response curve, dose levels, and optimal fitted curves to data with no residual variance for all types of models included in simulation study II. The colors on the optimally fitted curves refer to whether each model is in the set of classical dose-response models (Cl. M., including three- and four-parameter Log-Logistic, Log-Normal, and Weibull 1 and 2 curves) or in the set of included Fractional Polynomial models (FPL, models with fixed parameters $p_1 \in \{-2, -1, -0.5\}$ and $p_2 \in \{0.5, 1, 2\}$ are included).

- $MA_{Stack}DI(\lambda, \gamma)$: For all combinations of $\lambda \in \{1, 3, 5\}$ and $\gamma \in \{1, 5\}$, the subset of models with the highest DI based on the stacking weights was found. The lower and upper limits used in the DI were 10 and 1000, respectively. The BMD wass estimated from the MA curve based on this subset of models using the stacking weights.

For an overview of the strategies, see Table S2 in the Supporting Information. Note that identifying the single best and worst model requires oracle knowledge about the true data-generating mechanism and is accordingly not a useful strategy in practice.

All stacking weights were estimated by ten-fold cross-validation (chosen based on the results of Section 5). For each configuration, $R = 500$ data sets were simulated. In all configurations, the confidence intervals were computed by percentile non-parametric bootstrapping, where the data set was resampled $B = 500$ times, and for each resampled data set, all strategies were employed to estimate the BMD. Then, the BMDL was computed as the 5% quantile in the estimated BMD values on the resampled data sets. Both types of weights were considered as a part of the model and were recomputed for each resampled data set. All resampled data sets were resampled within the dose levels to obtain new data sets within the fixed design framework. The performance of the various strategies was assessed by considering the mean bias and the RMSE of the BMD estimates, as well as the coverage of the one-sided confidence intervals defining BMDL.

6.2 | Results

6.2.1 | AIC Weights Vs. Stacking Weights on Full Model Space

Mean bias and RMSE for the methods in simulation study II can be seen in Table S3 in the Supporting Information, with a selection of the methods illustrated in Figure 5. Across nearly all setups and combinations of n_{rep} and $\mathbf{x}_5$ or $\mathbf{x}_{10}$, the MA estimators based on the full model space using the stacking weights had bias and RMSE smaller than or comparable to the corresponding MA estimators using the AIC weights.

In setup A, the bias for the best model was smaller than that of the MA estimators in all cases except for the post-MA estimate using the stacking weights in the case with most observations. In setup B, the bias of the MA estimators and the best model were comparable. The lowest bias was obtained by MA with the stacking weights. In Setup C, the bias and RMSE for the best model were smaller than the MA estimators in the two scenarios with five dose levels, but similar in the scenarios with ten dose levels. In setup D, the best model had the lowest mean bias and RMSE in all cases.

In all cases, choosing the model with the lowest AIC value for each data set resulted in an estimator with RMSE similar to the MA estimators based on the full model space. The mean bias for the BMD estimate from the model with the minimum AIC value was smaller or similar to that of the MA estimators.

Coverage of BMDL (see Figure 6 and Table S4) was observed to be below the nominal level of 0.95 for nearly all strategies in all scenarios. In setups A and B, and C (excluding setup C with $n_{rep} = 5$ and the dose vector $\mathbf{x}_{10}$), the coverage of the MA estimators using the stacking weights was higher and closer to the nominal level than the MA estimators using the AIC weights. In setup A, the observed coverage of BMDL based on the best (true) model was closer to the nominal level than the MA estimators in all cases. However, in setup B, the observed coverage of BMDL was better for all the MA estimators using the full model space. In setup C, the best model had coverage above the nominal level for the cases with five dose levels (Figure 6).

Excluding setup D, the MA estimators using the full model space had an observed coverage of BMDL closer to the nominal level than the estimator based on selecting the model with the lowest AIC value.

6.2.2 | Effect of Model Space

The mean bias and RMSE of the MA curve estimators based on the different included model spaces are visualized in Figure 7.

The best overall performance in terms of mean bias and RMSE was seen on the MA estimators using the full model space or exclusively the classical models (Log-Logistic, Log-Normal, and Weibull 1 and 2 models in three- and four-parameter versions). The included schemes for selecting a subset of the full model space to use in the MA did not improve the performance of the MA estimators. The DI affected the performance slightly for low values of the tuning parameters λ and γ, resulting in higher observed mean bias and RMSE. For increasing values of the tuning parameters, the performance of the MA estimators was severely affected (Table S2).

In setups A, B, and C, the MA estimators based on the fractional polynomial models had a larger mean bias and RMSE than the corresponding MA estimators based on the full model space and the space of classical dose-response models (Figure 7). Additionally, the classical model space compared to the full model space generally resulted in lower mean bas and RMSE. In setup D, all model spaces resulted in highly biased results.

For the MA estimators using the AIC weights, applying the DI model space selection resulted in larger RMSE, but comparable, and in some cases even smaller mean bias. Conversely, for the MA estimators using the stacking weights, applying the DI model space selection resulted in larger mean bias and RMSE in most cases, particularly for larger values of the tuning parameters λ and γ (Table S3 and Figure 7).

In almost all scenarios in setups A, B, and C, the MA estimators based exclusively on the classical models (Log-Logistic, Log-Normal, and Weibull 1 and 2 models) had an observed coverage of BMDL closer to the nominal level compared to the MA estimators based on the full model space, and the MA estimators based on exclusively the fractional polynomial models performed worse than the full model space (Figure 8).

In all scenarios in setup A and B, the observed coverage of BMDL was lower for the estimators with the DI applied than for the corresponding estimators based on the full model space. In setup

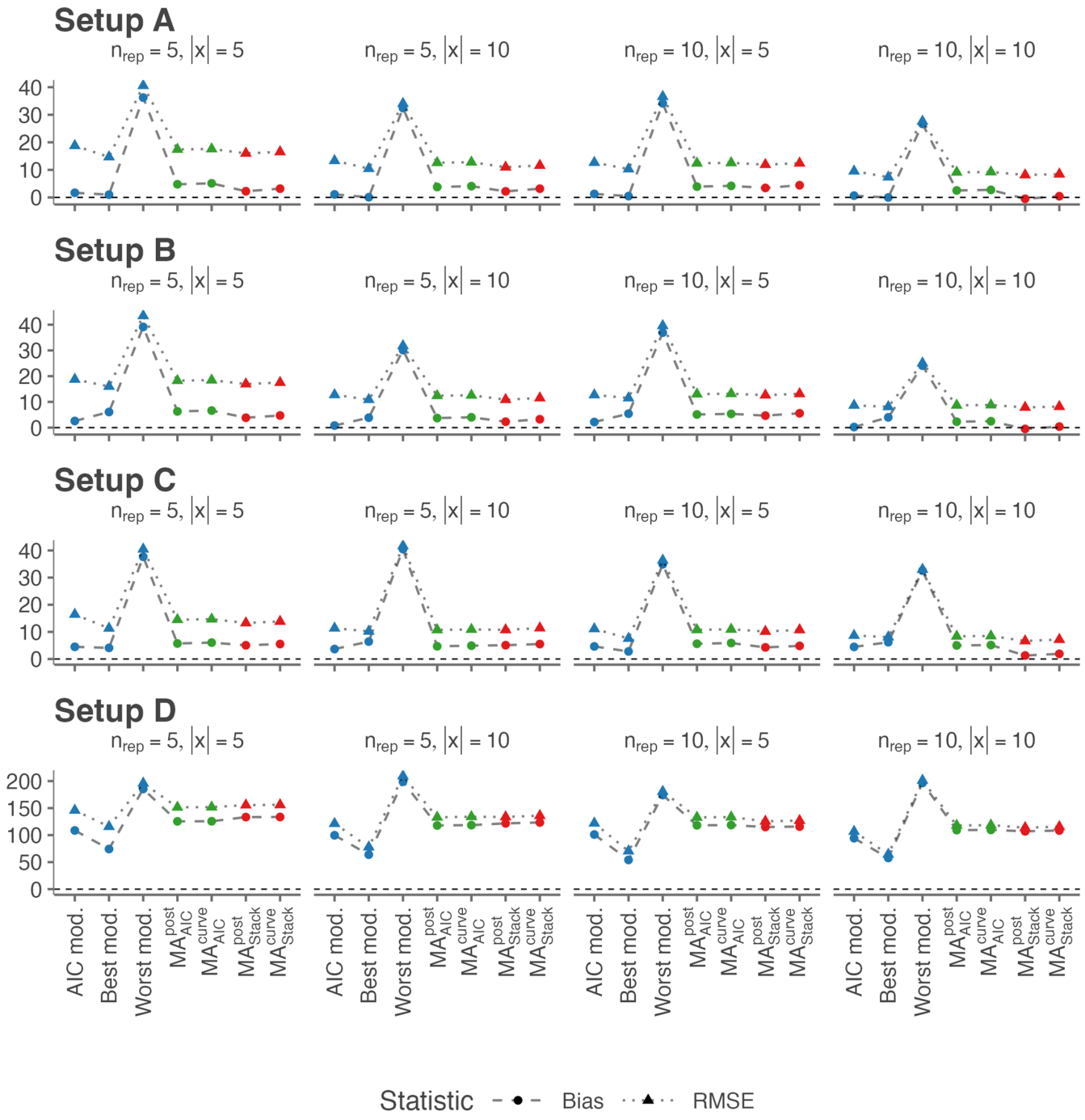


FIGURE 5 | Observed mean bias and root mean square error (RMSE) of the benchmark dose (BMD) estimates of the single model estimating strategies and the model averaging (MA) estimating strategies using the full model space in Simulation Study II. Blue points are estimators based on a single model, while green points are MA estimators based on AIC weights, and red points are MA estimators based on the stacking weights.

C, BMDL coverage was slightly closer to the nominal level for the MA estimators using the stacking weights with the DI model space selection with the lowest values of the tuning parameters, compared to using the full model space. In general, the coverage of BMDL decreased with the size of the tuning parameters λ and γ. The MA estimators based on the stacking weights were particularly affected by the tuning parameters (Table S4).

Compared to the MA estimator using the AIC weights on the full model space, there was no observed effect of conducting the F-test on the models prior to MA on mean bias or RMSE, and the observed coverage of BMDL was unchanged or slightly lower.

6.2.3 | Post MA Vs. Curve MA

In nearly all scenarios, the mean bias of the BMD estimates was slightly larger, and the coverage of BMDL was slightly lower for the curve MA estimators compared to the post-MA estimators. Otherwise, the two types performed similarly within each

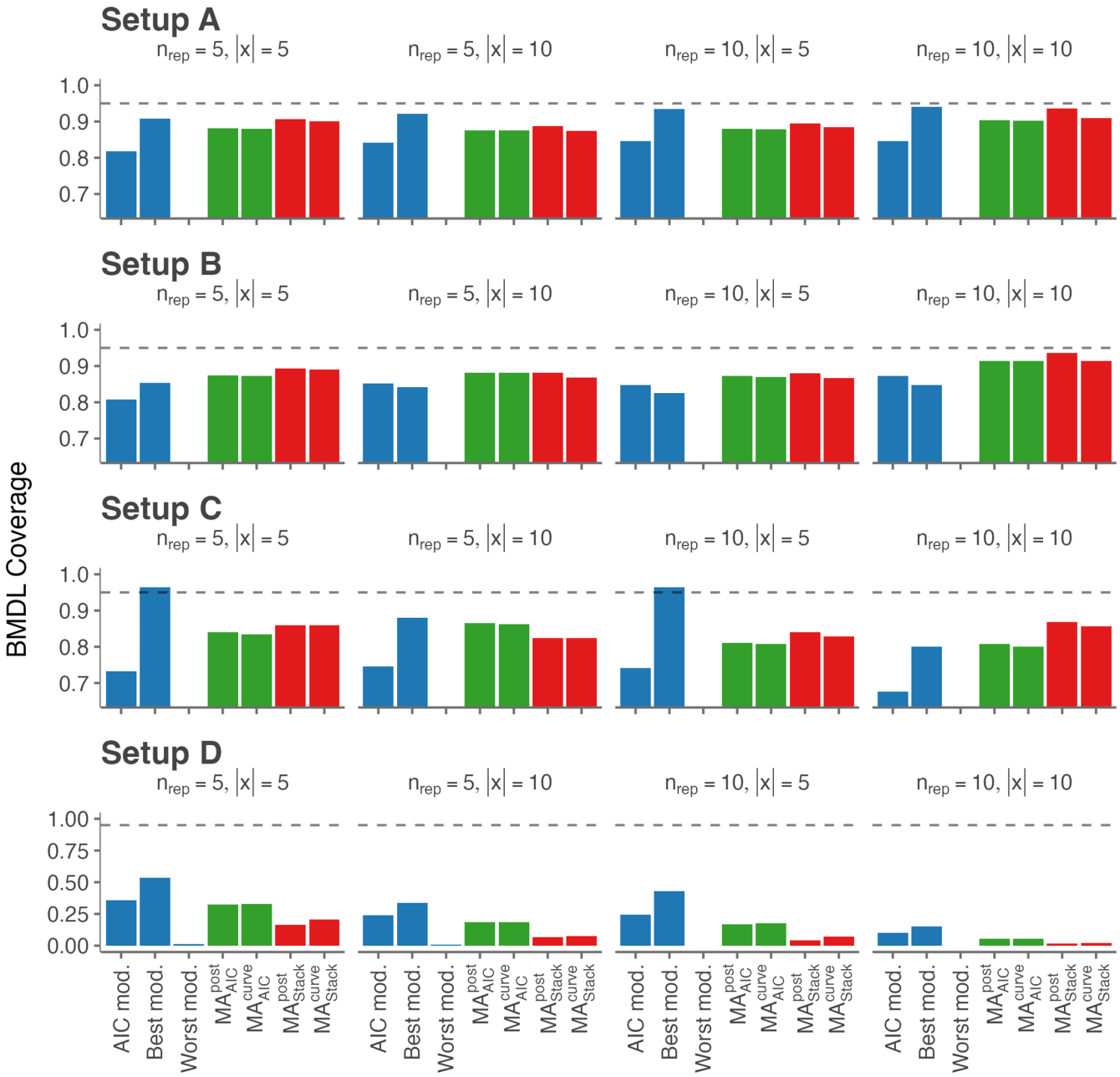


FIGURE 6 | Observed coverage of the benchmark dose lower limit (BMDL) of the single model estimating strategies and the model averaging (MA) estimating strategies using the full model space in Simulation Study II. Note the different range of the y-axis for Setup D. Coverages for the worst model are not visible on the figure, since they are near zero in most cases (Table S4 in Supporting Information). Blue bars denote estimators based on a single model, while green bars are MA estimators based on AIC weights, and red bars are MA estimators based on the stacking weights.

MA configuration of weights and model space (Figures 5–8 and Tables S3 + S4).

7 | Simulation Study III

For a dose-response model of the type (1) and a fixed number of observations, the AIC weights converge to a fixed set of weights in the limit $\sigma \to 0$, where σ denotes the residual standard deviation (Lemma S1 in the Supporting Information). However, the AIC weights in the limit are not necessarily the optimal weights to use in MA.

Based on this result, a third simulation study was conducted to assess the performance of the AIC weights compared to the stacking weights in a scenario with decreasing values of the residual standard deviation.

The true curve was constructed as a convex combination of Log-Logistic, Log-Normal, and Weibull 1 and 2 curves, such that the true curve was not in the set of fitted models. Points were simulated from the dose vector $\mathbf{x} = (0, 2.5, 5, 10, 20)$, and residuals were simulated from a normal distribution with decreasing standard deviation $\sigma \in \{5, 1.58, 0.5, 0.158, 0.05, 0.0158\}$. The number of repetitions per dose level in this simulation study was fixed at

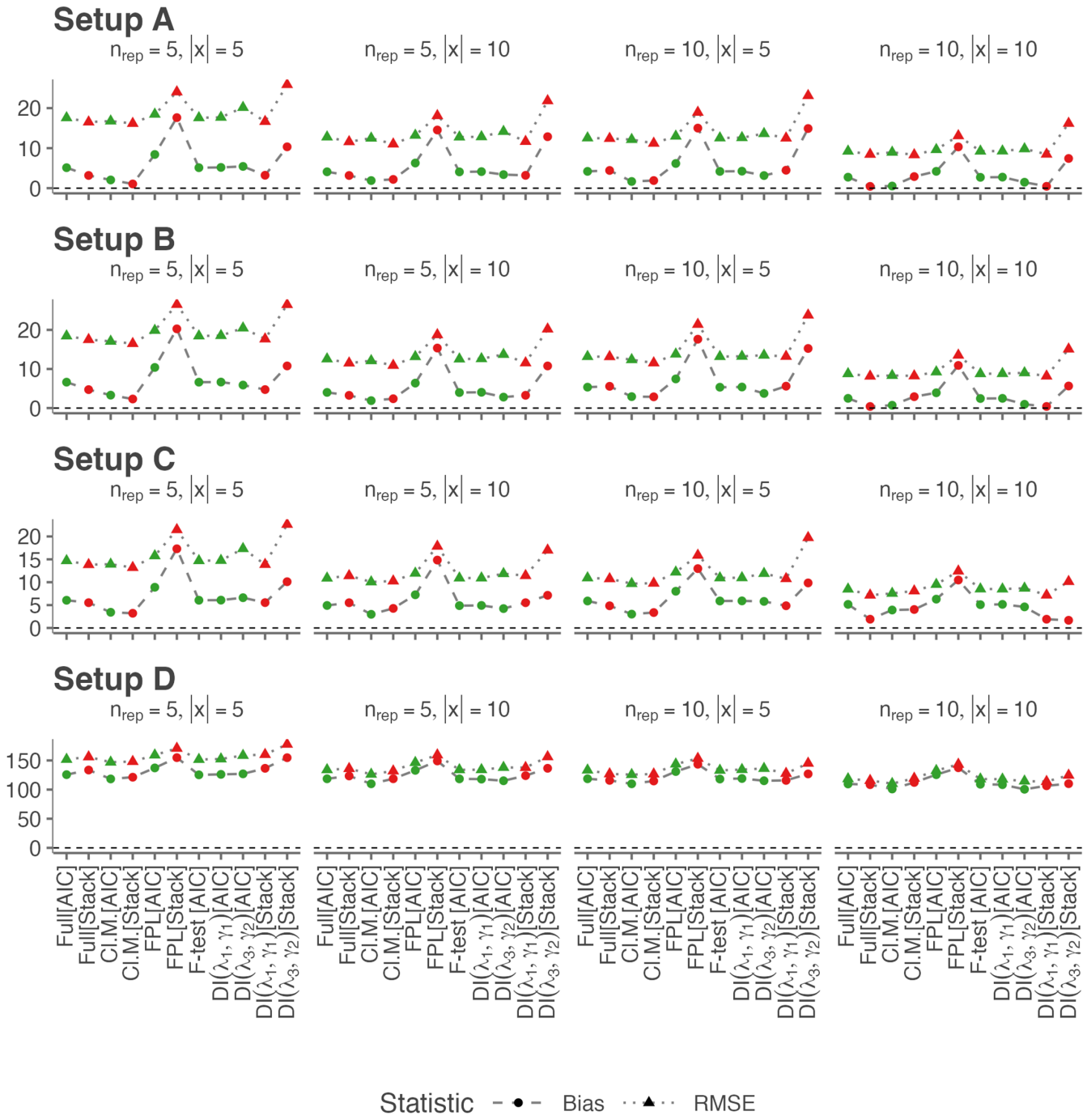


FIGURE 7 | Observed mean bias and root mean square error (RMSE) for a subset of the estimation strategies included in Simulation Study II. The subset was selected to compare the model spaces underlying model averaging (MA). The figure includes curve MA estimation using the full model space, the classical dose-response models, the fractional polynomial models, and model space selection based on the diversity index (DI) for $\lambda_1 = 1$, $\lambda_3 = 5$ and $\gamma_1 = 1, \gamma_2 = 5$. Green points are MA estimators based on AIC weights, and red points are MA estimators based on the stacking weights.

six for all values of σ. The true BMD value in this simulation study was 3.42.

The model space included the same selection of models as in Simulation Study II. The true curve and the included models fitted to a data set with no residual variance can be seen in Figure S1 in the Supporting Information.

This simulation study included the following five strategies for BMD estimation also included in simulation study II:

- Best model
- MA_{AIC}(post)
- MA_{AIC}(curve)
- MA_{Stack}(post)
- MA_{Stack}(curve)

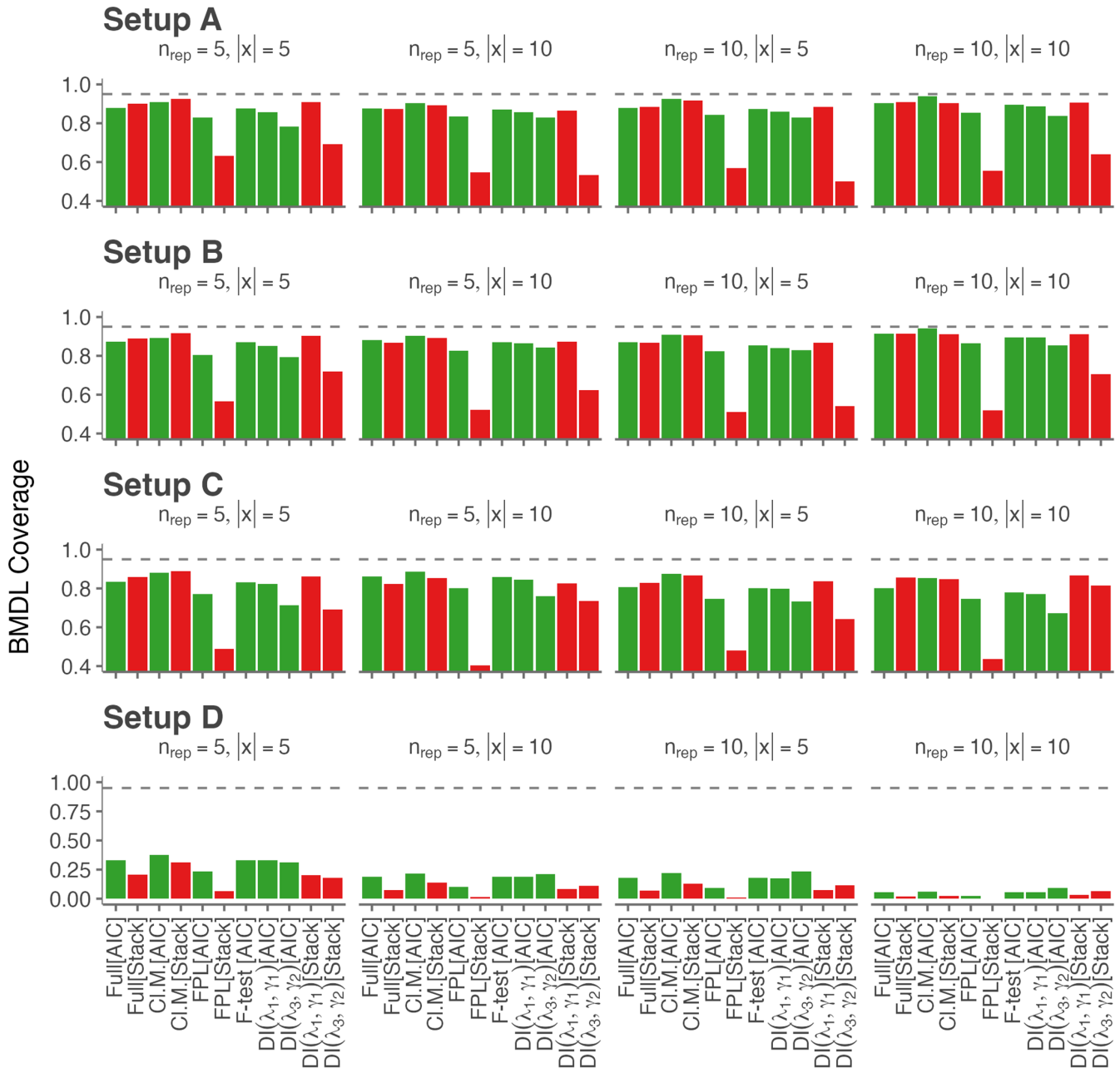


FIGURE 8 | Observed coverage of the benchmark dose lower limit (BMDL) for a subset of the estimation strategies included in Simulation Study II. The subset was selected to compare the model spaces underlying model averaging (MA). The figure includes curve MA estimation using the full model space, the classical dose-response models, the fractional polynomial models, and model space selection based on the diversity index (DI) for $\lambda_1 = 1, \lambda_3 = 5$ and $\gamma_1 = 1, \gamma_2 = 5$. Note the different range of the y-axis for Setup D. Coverages for the worst model are not visible on the figure, since they are near zero in most cases (Table S4 in Supporting Information). Green bars are MA estimators based on AIC weights, and red bars are MA estimators based on the stacking weights.

The best model in this setup was the Fractional Polynomial model with fixed parameters $p_1 = -0.5$ and $p_2 = 0.5$. All MA methods used the full model space. The stacking weights were estimated based on ten-fold cross-validation. For each value of σ, $R = 500$ data sets were simulated. The performance of the various strategies was assessed by considering the mean bias and the normalized RMSE (defined as RMSE divided by the residual standard deviation) of the BMD estimates.

7.1 | Results III

A visualization of summarized results from simulation study III can be seen in Figure 9. For small values of σ, the observed mean bias of the MA estimators based on the stacking weights was closer to zero than the mean bias of the MA estimators based on the AIC weights and the mean bias of the best model. The normalized RMSE was similar for large values of σ, and increasing

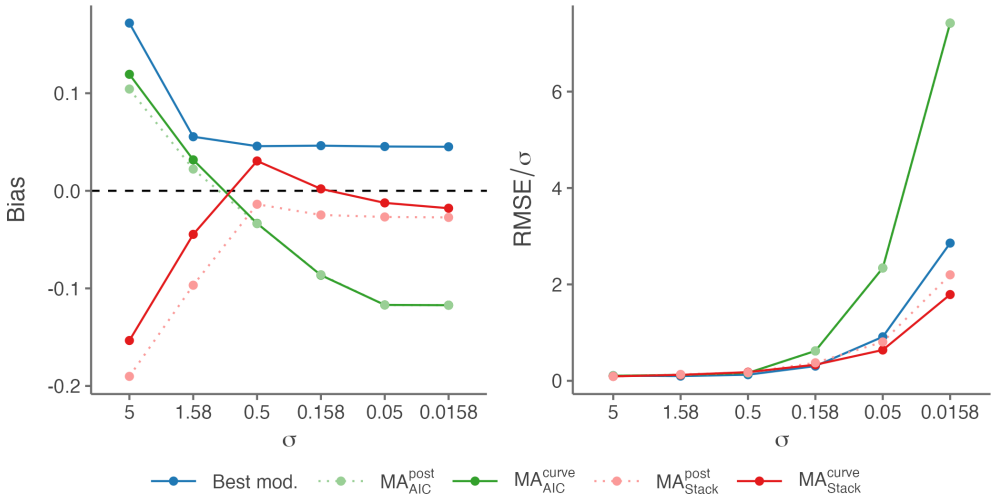


FIGURE 9 | Performance of the five strategies included in simulation study III in terms of empirical mean (left) and normalized root mean squared error (RMSE, right).

at different rates for the different MA strategies, with the largest normalized RMSE being observed for the MA estimators using the AIC weights. The increase in normalised RMSE was a consequence of the bias of each strategy in the limit $\sigma \to 0$.

8 | Discussion

Frequentist MA is a widely used technique in BMD estimation, and our results generally support the use of this technique, although some disadvantages of current practice have been highlighted in this paper. In frequentist MA, weights based on the AIC values for each individual fitted model are widely used (Jensen et al. 2019; Wheeler and Bailer 2009; Aerts et al. 2020). However, their asymptotic behavior, or, as considered in this paper, the limit when the residual variance tends to zero, is less than optimal. If the residual variance tends to zero, the AIC weights are not guaranteed to converge to the optimal combination of the models. A similar choice of weights, which are based on the Bayesian Information Criterion (BIC) values instead of the AIC values, can be defined. Hoeting et al. (1999) report that these BIC weights are O(1) estimators of the posterior model probabilities, which will not converge to the correct model probabilities. This is partly why the current Bayesian model averaging (Wheeler et al. 2020, 2022; EFSA Scientific Committee et al. 2022) uses Laplace or Bridge Sampling estimators for the posterior model probability as they converge to the correct posterior model probabilities. In contrast, the stacking weights (Breiman 1996) are defined such that they converge to the set of weights that result in the optimal MA curve in terms of MSPE. This was illustrated in simulation study III, which showed that the MA estimators using the stacking weights performed better than MA estimators using the AIC weights for low values of the residual standard deviation. A similar result regarding convergence of the AIC weights can be shown for the number of observations tending to infinity. However, the number of observations needed for this effect to set in is very large and certainly much larger than the number of observations usually seen in dose-response modeling. If the asymptotic behavior is considered, it can even be shown that if one model in the model space has lower Kullback-Leibler divergence to the true model in the Kullback-Leibler optimal parameters than the rest of the models, the AIC weight for this particular model will converge to one, while the AIC weights for the remaining models will converge to zero. This is not exactly the case when the convergence of the residual variance to zero is considered, as was done here, since the number of observations is assumed to stay the same. However, in the design stage of simulation study III, it was found that even for a small number of observations (≈ 25), the AIC weights in cases with zero residual variance were practically one for the best-fitting model and zero for the remaining models (results not shown).

We have focused on assessing and comparing frequentist MA estimators of BMD for continuous response variables and in settings where residual variances are constant. This is the most appropriate setting for applying stacking weights. If the residual variances are not constant on the original scale of the responses, it is often possible in practice to transform the responses to a scale where the residual variances are approximately constant, for example, by a log-transformation. In such cases, we recommend applying MA using stacking weights on the transformed scale. A different approach, which we have not investigated, is a full model specification of multiple error distributions, which can be leveraged by frequentist AIC weights as well as in Bayesian MA. For a detailed investigation of Bayesian MA with multiple error distributions, we refer to Wheeler et al. (2022).

Nair et al. (2022) and Gomes et al. (2012) applied stacking to binary classification, using a modified approach where the weights are optimized through cross-validated logistic regression models rather than the cross-validated ordinary least squares method used for stacking weights in this paper. With this modification, stacking weights could potentially be applied to BMD estimation in cases involving binary response variables. However, this approach has not yet been implemented, and no comparisons between AIC weights and stacking weights have been conducted.

In simulation study II, several frequentist MA estimators of the BMD were evaluated across a range of scenarios, from realistic setups inspired by actual experiments to more challenging conditions. In setups A, B, and C, where the true dose-response curve was based on real experimental data, various MA strategies

demonstrated performance comparable to the best or true model, and in some cases even exceeded it in terms of bias, RMSE, and coverage of BMDL. However, in a fourth setup, D, where the true dose-response curve was generated using an I-spline basis designed to differ from any of the models in the model space, all methods, including the best individual model, showed poor performance. The satisfactory performance of MA in the realistic scenarios aligns with previous studies and supports its use in risk assessment (EFSA 2011; Jensen et al. 2019; Aerts et al. 2020; Ritz et al. 2013; Wheeler and Bailer 2009). Nevertheless, the poor performance in the adverse scenarios highlights the importance of evaluating the individual model fits, even when using MA (OECD 2006).

It was expected that for a fixed number of observations, more dose levels would be preferred over more repetitions per dose level, since more dose levels would entail a more accurate description of the dose-response curve (Ringblom et al. 2018; Shao and Small 2012). This was confirmed for the MA estimators in setups A, B, and C, where scenarios with five repetitions and ten dose levels resulted in slightly lower bias and RMSE compared to the scenarios with ten repetitions and five dose levels. This difference highlights the need for careful consideration of the experiment design in order to obtain a reliable BMD estimate.

The included MA strategies were based on 17 different dose-response models. Overall, the best MA performance was observed for strategies using the full model space or strategies only including the Log-Logistic, Log-Normal, and Weibull models. MA strategies based exclusively on the fractional polynomial models did not perform satisfactorily in this simulation study.

In setups A, B, and C, the stacking weights showed a promising performance, which, in terms of the mean bias of the MA estimator and coverage of BMDL, was observed to be slightly better than the AIC weights in several scenarios. While the stacking weights outperformed the AIC weights in several scenarios in the simulation studies conducted in this paper, it is difficult to determine a set of criteria where one set of weights is guaranteed to outperform the other. The performance will depend on various factors, such as the endpoint under consideration and factors related to the data-generating model, including the variance structure, the true dose-response curve, and the specific dose values considered.

Two-, five-, and ten-fold and LOO cross-validation were applied for the estimation of the stacking weights in simulation study I. By Arlot and Celisse (2010) and Zhang and Yang (2015), the lowest variance and bias can be expected for LOO cross validation compared to V-fold cross validation, if the optimisation procedure satisfies a certain stability condition. While it is out of the scope of this paper to show that stacking satisfies this stability condition, our results are consistent with this result, with the lowest bias and normalized RMSE observed when LOO cross-validation was applied. However, due to the excessive computation time of fitting models a large number of times, it can be impractical to use LOO cross-validation when the stacking weights are combined with confidence intervals obtained by bootstrap. In order to do this, all models need to be fitted, and weights need to be estimated by convex optimization a total of $n \cdot (R_{\text{boot}} + 1)$ times (here n is the number of observations, and R_{boot} is the number of resampled data sets used for bootstrap).

Relying on cross-validation when estimating the stacking weights means that a full data set of the individual experimental units is required. As a result, they cannot be estimated for dose-response analyses based on sufficient statistics. Bayesian stacking, on the other hand, does not have this issue (Yao et al. 2018). However, Bayesian stacking has so far not been applied to dose-response analysis.

Van der Laan et al. (2007) showed that MA using the stacking weights results in a model that works at least as well asymptotically for prediction (in terms of mean square prediction error) as the single best model in the set of initial models. This does, however, not imply that the BMD estimate from the stacked model results in an estimate, which is at least as good asymptotically as estimating the BMD based on the best model for this purpose. Nonetheless, the results from the conducted simulation study indicate that this is the case.

In simulation studies I and II, the coverage of BMDL was below the nominal level in most cases. It is suspected that this was caused primarily by the bias of the estimators. The cases where the coverage of BMDL was below the nominal level coincide with the cases where the MA estimators of the BMD were positively biased. It is also suspected that this was caused by the nature of percentile bootstrap confidence intervals, which are known to assert this kind of behavior, in particular when the distribution of the estimator is skewed (Dicicio and Romano 1988). As an alternative, bias-corrected bootstrap intervals can be considered, although no improvement in performance is guaranteed (Jensen et al. 2020b).

MA, in particular when constructing an MA curve, can be thought of as a way to combine the initial models into one MA model, which (hopefully) inherits the best traits of the individual models. This is the general concept behind the stacking weights (Breiman 1996), which are defined such that they actively seek to combine the initial models in an optimal way in terms of the MSPE. Unlike the stacking weights, the AIC weights are not designed to take the fit of the resulting MA curve into account. This might result in too much weight being assigned to models that do not improve the fit of the MA curve. These considerations were the motivation behind investigating the influence of the model space. A quite large effect of the choice of model space was observed. The best overall performance across all scenarios was seen for the MA estimators using either the full model space, or only the classical models.

The AIC weights in particular were suspected to be affected by the presence of poor-fitting models in the model space. For this reason, an adaptation of a model space selection scheme used in Wheeler and Bailer (2009) was considered. However, no improvement in performance was observed when applying this model selection scheme prior to applying MA compared to using the full model space. Also not in scenario D, where this model space reduction was expected to have the largest effect. This model selection scheme also has the built-in disadvantage that if all the models are wrong, asymptotically they would all be rejected.

Finally, model space selection based on a DI was examined. The DI was proposed by Kim et al. (2014) for dose-response models on quantal data to aid in selecting a model space with diverse

model fits. In this paper, their proposed DI was modified to work with continuous response data as well. However, the results in the simulation study were less than promising. At best, the MA strategies involving model space selection by the DI performed similarly to the corresponding strategies including the full model space, and in several cases, they resulted in higher bias of the BMD estimate and lower coverage of BMDL. It is suspected that this is caused by the fact that model space selection based on the DI by design favors the "extreme" models in the model space and discards similar models. Consequently, a subset of models providing good (and therefore also similar) fits to the observations might be discarded, resulting in a model space that only includes models that do not fit the data well.

8.1 | Conclusion

The stacking weights are a new addition to BMD estimation by frequentist MA, and the results in this paper warrant more research in the application of the stacked regression approach in frequentist dose-response MA. The MA estimators using the stacking weights had a lower mean bias and a coverage of BMDL slightly closer to the nominal level compared to the estimators using the AIC weights. For small values of the residual variance, the stacking weights outperformed the AIC weights. Due to the promising performance of the stacking weights, they have been added as an option in the `bmd` R package.

Acknowledgments

This work was supported by NOVO Nordisk Fonden (Grant number NNF21OC0068954).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data analyzed in Section 3 is available through the R package `drcData`, and the code for the data analysis is available in the supplementary information. The code for the simulation study is available from the corresponding author on request.

References

Aerts, M., M. W. Wheeler, and J. C. Abrahantes. 2020. "An Extended and Unified Modeling Framework for Benchmark Dose Estimation for Both Continuous and Binary Data." *Environmetrics* 31, no. 7: e2630.

Akaike, H. 1973. *Information Theory and an Extension of the Maximum Likelihood Principle, Proceedings of the 2nd International Symposium on Information*. Czaki.

Arlot, S., and A. Celisse. 2010. "A Survey of Cross-Validation Procedures for Model Selection." *Statistics Surveys* 4: 4079.

Breiman, L. 1996. "Stacked Regressions." *Machine Learning* 24: 4964.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. "Model Selection: An Integral Part of Inference." *Biometrics* 53: 603618.

Burnham, K., and D. Anderson. 2003. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer. https://books.google.dk/books?id=BQYR6js0CC8C.

Cedergreen, N., and J. C. Streibig. 2005. "Can the Choice of Endpoint Lead to Contradictory Results of Mixture-Toxicity Experiments?" *Environmental Toxicology and Chemistry: An International Journal* 24, no. 7: 16761683.

Claeskens, G., and N. L. Hjort. 2008. *Model Selection and Model Averaging*. Cambridge Books.

Crump, K. S. 1984. "A New Method for Determining Allowable Daily Intakes." *Toxicological Sciences* 4, no. 5: 854871.

Crump, K. S. 1995. "Calculation of Benchmark Doses From Continuous Data." *Risk Analysis* 15, no. 1: 7989.

Das, I. 2018. "Robust Benchmark Dose Estimation Using an Infinite Family of Response Functions." *Human and Ecological Risk Assessment: An International Journal* 24, no. 8: 20542069.

Davis, J. A., J. S. Gift, and Q. J. Zhao. 2011. "Introduction to Benchmark Dose Methods and U.S. EPAs Benchmark Dose Software (BMDS) Version 2.1.1." *Toxicology and Applied Pharmacology* 254, no. 2: 181–191. https://www.sciencedirect.com/science/article/pii/S0041008X10004096TRAC 2008/2009 meeting. https://doi.org/10.1016/j.taap.2010.10.016.

Diciccio, T. J., and J. P. Romano. 1988. "A Review of Bootstrap Confidence Intervals." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 50, no. 3: 338354.

EFSA. 2011. "Use of BMDS and PROAST Software Packages by EFSA Scientific Panels and Units for Applying the Benchmark Dose (BMD) Approach in Risk Assessment." *EFSA Supporting Publications* 8, no. 2: 113E.

EFSA Scientific Committee, More, S. J., Bampidis, V., et al. 2022. "Guidance on the Use of the Benchmark Dose Approach in Risk Assessment." *EFSA Journal* 20, no. 10: e07584.

Faes, C., H. Geys, M. Aerts, and G. Molenberghs. 2003. "Use of Fractional Polynomials for Dose-Response Modelling and Quantitative Risk Assessment in Developmental Toxicity Studies." *Statistical Modelling* 3, no. 2: 109125.

Fu, A., B. Narasimhan, and S. Boyd. 2017. CVXR: An R Package for Disciplined Convex Optimization. *arXiv Preprint arXiv:1711.07582*.

Gomes, C., H. Nocairi, M. Thomas, F. Ibanez, J.-F. Collin, and G. Saporta. 2012. "Stacking Prediction for a Binary Outcome." In *Compstat 2012*. HAL science ouverte, 271282.

Haber, L. T., M. L. Dourson, B. C. Allen, et al. 2018. "Benchmark Dose (BMD) Modeling: Current Practice, Issues, and Challenges." *Critical Reviews in Toxicology* 48, no. 5: 387415.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial (With Comments by M. Clyde, David Draper and EI George, and a Rejoinder by the Authors)." *Statistical Science* 14, no. 4: 382417.

Jensen, S. M., N. Cedergreen, F. M. Kluxen, and C. Ritz. 2021. "A Nonmechanistic Parametric Modeling Approach for Benchmark Dose Estimation of Event-Time Data." *Risk Analysis* 41, no. 11: 20812093.

Jensen, S. M., F. M. Kluxen, and C. Ritz. 2019. "A Review of Recent Advances in Benchmark Dose Methodology." *Risk Analysis* 39, no. 10: 22952315.

Jensen, S. M., F. M. Kluxen, J. C. Streibig, N. Cedergreen, and C. Ritz. 2020a. "Bmd: An R Package for Benchmark Dose Estimation." *PeerJ* 8: e10557.

Jensen, S. M., J. C. Streibig, N. Cedergreen, and C. Ritz, eds. 2020b. *Benchmark Dose: Bedre Udnyttelse AF Data I Pesticid Registrering OG Regulering*. Miljstyrelsen.

Kang, S.-H., R. L. Kodell, and J. J. Chen. 2000. "Incorporating Model Uncertainties Along With Data Uncertainties in Microbial Risk Assessment." *Regulatory Toxicology and Pharmacology* 32, no. 1: 68–72. https://www.sciencedirect.com/science/article/pii/S0273230000914041. https://doi.org/10.1006/rtph.2000.1404.

Kim, S. B., R. L. Kodell, and H. Moon. 2014. "A Diversity Index for Model Space Selection in the Estimation of Benchmark and Infectious Doses via Model Averaging." *Risk Analysis* 34, no. 3: 453464.

Nair, S., A. Gupta, R. Joshi, and V. Chitre. 2022. Combining Varied Learners for Binary Classification Using Stacked Generalization. arxiv Preprint *arXiv:2202.08910*.

Namata, H., M. Aerts, C. Faes, and P. Teunis. 2008. "Model Averaging in Microbial Risk Assessment Using Fractional Polynomials." *Risk Analysis: An International Journal* 28, no. 4: 891905.

OECD. 2006. "Current Approaches in the Statistical Analysis of Ecotoxicity Data." https://doi.org/10.1787/9789264085275-en. https://www.oecd-ilibrary.org/content/publication/9789264085275-en.

Piegorsch, W. W., H. Xiong, R. N. Bhattacharya, and L. Lin. 2012. "Nonparametric Estimation of Benchmark Doses in Environmental Risk Assessment." *Environmetrics* 23, no. 8: 717728.

Piegorsch, W. W., H. Xiong, R. N. Bhattacharya, and L. Lin. 2014. "Benchmark Dose Analysis via Nonparametric Regression Modeling." *Risk Analysis* 34, no. 1: 135151.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing [Computer Software Manual]*. Austria. https://www.R-project.org/.

Ringblom, J., F. Kalantari, G. Johanson, and M. Öberg, 2018. "Influence of Distribution of Animals Between Dose Groups on Estimated Benchmark Dose and Animal Welfare for Continuous Effects." *Risk Analysis* 38, no. 6: 11431153.

Ritz, C., D. Gerhard, and L. Hothorn. 2013. "A Unified Framework for Benchmark Dose Estimation Applied to Mixed Models and Model Averaging." *Statistics in Biopharmaceutical Research* 1: 79–90.

Ritz, C., S. M. Jensen, D. Gerhard, and J. C. Streibig. 2020. *Dose-Response Analysis Using R*. Vol. 2. 1st ed. Chapman & Hall.

Shao, K., and M. J. Small. 2012. "Statistical Evaluation of Toxicological Experimental Design for Bayesian Model Averaged Benchmark Dose Estimation With Dichotomous Data." *Human and Ecological Risk Assessment: An International Journal* 18, no. 5: 10961119.

Sørensen, H., N. Cedergreen, I. M. Skovgaard, and J. C. Streibig. 2007. "An Isobole-Based Statistical Model and Test for Synergism/Antagonism in Binary Mixture Toxicity Experiments." *Environmental and Ecological Statistics* 14: 383397.

Tibshirani, R. J., and B. Efron. 1993. "An Introduction to the Bootstrap." *Monographs on Statistics and Applied Probability* 57, no. 1: 1–436.

US EPA, More, S. J., Bampidis, V., et al. 2012. "Benchmark Dose Technical Guidance." *Risk Assessment Forum*: 42e.

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6, no. 1: 1–25.

Wang, W., and J. Yan. 2021. "Shape-Restricted Regression Splines With R Package Splines2." *Journal of Data Science* 19, no. 3: 498517. https://doi.org/10.6339/21-JDS1020.

West, R. W., W. W. Piegorsch, E. A. Pea, et al. 2012. "The Impact of Model Uncertainty on Benchmark Dose Estimation." *Environmetrics* 23, no. 8: 706716.

Wheeler, M. W., and A. J. Bailer. 2007. "Properties of Model-Averaged BMDLs: A Study of Model Averaging in Dichotomous Response Risk Estimation." *Risk Analysis: An International Journal* 27, no. 3: 659670.

Wheeler, M. W., and A. J. Bailer. 2009. "Comparing Model Averaging With Other Model Selection Strategies for Benchmark Dose Estimation." *Environmental and Ecological Statistics* 16, no. 1: 37.

Wheeler, M. W., T. Blessinger, K. Shao, et al. 2020. "Quantitative Risk Assessment: Developing a Bayesian Approach to Dichotomous Doseresponse Uncertainty." *Risk Analysis* 40, no. 9: 17061722.

Wheeler, M. W., J. Cortias Abrahantes, M. Aerts, J. S. Gift, and J. Allen Davis. 2022. "Continuous Model Averaging for Benchmark Dose Analysis: Averaging Over Distributional Forms." *Environmetrics* 33, no. 5: e2728.

Wheeler, M. W., S. Lim, J. S. House, K. R. Shockley, A. J. Bailer, and J. Fostel. 2023. "ToxicR: A Computational Platform in R for Computational Toxicology and Doseresponse Analyses." *Computational Toxicology* 25: 100259.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions (With Discussion)." *Bayesian Analysis* 13, no. 3: 9171003.

Zhang, Y., and Y. Yang. 2015. "Cross-Validation for Selecting a Model Selection Procedure." *Journal of Econometrics* 187, no. 1: 95112.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

An extension to current frequentist model averaging methods for benchmark dose estimation - Supplementary material

Jens Riis Baalkilde*, Niels Richard Hansen, Signe Marie Jensen

*Corresponding author: jba@plen.ku.dk

Convergence of AIC weights

In the following, it is shown that if the residual variance in the true model tends to zero, the Gaussian AIC weights converge to a non-random set of weights. We identify these limit weights and note that they are not generally the set of weights resulting in the optimal MA curve.

The setup is as follows: we let $x_1, \ldots, x_n$ denote a fixed set of dose values, and we assume that the true model is given as

$$Y_m = g(x_m) + \sigma\varepsilon_m \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with mean 0 and variance 1. We let $\mathcal{M} = \{M_1, \ldots, M_k\}$ be the set of models, and for a curve $g_i \in M_i$ we introduce

$$R(g_i) = \frac{1}{n}\sum_{m=1}^{n}(g_i(x_m) - Y_m)^2$$
$$R^*(g_i) = \frac{1}{n}\sum_{m=1}^{n}(g_i(x_m) - g(x_m))^2.$$

Now $\hat{g}_i$ denotes the least squares estimate in M_i, which is a minimizer of $R(g_i)$ within model M_i. Likewise, g_i^* denotes a minimizer of $R^*(g_i)$ within model M_i. With these definitions we note that

$$\hat{\sigma}_i^2 = R(\hat{g}_i)$$

is the Gaussian maximum-likelihood estimate of the variance within model M_i. We suppose that $\hat{\sigma}_i^2 > 0$ for $i = 1, \ldots, k$ to make the AIC weights well defined.

Lemma S1. *Within the setup above, suppose that*

$$\sup_{g_i \in M_i} |R(g_i) - R^*(g_i)| \to 0 \tag{2}$$

for $\sigma \to 0$ and $i = 1, \ldots, k$, and that there is at most one model with $R^*(g_i^*) = 0$, then

$$\hat{w}_i^{\mathrm{AIC}} \to \left(1 + \sum_{j \neq i}^{k}\left(\frac{R^*(g_i^*)}{R^*(g_j^*)}\right)^{\frac{n}{2}} e^{p_i - p_j}\right)^{-1} \tag{3}$$

for $\sigma \to 0$. A sufficient condition for (2) to hold is that there is a constant C such that $|g_i(x_m)| \leq C$ for $m = 1, \ldots, n$ and $g_i \in M_i$.

Proof. Since g_i^* and $\hat{g}_i$ are minimizers in M_i of R^* and R, respectively, we have

$$\begin{aligned} 0 &\leq R^*(\hat{g}_i) - R^*(g_i^*) \\ &= R^*(\hat{g}_i) - R(\hat{g}_i) + R(\hat{g}_i) - R^*(g_i^*) \\ &\leq R^*(\hat{g}_i) - R(\hat{g}_i) + R(g_i^*) - R^*(g_i^*) \\ &\leq 2 \sup_{g_i \in M_i} |R(g_i) - R^*(g_i)|. \end{aligned}$$

It follows by assumption (2) that $|R^*(\hat{g}_i) - R^*(g_i^*)| \to 0$ for $\sigma \to 0$. We then get that

$$\begin{aligned} |\hat{\sigma}_i^2 - R^*(g_i^*)| &= |R(\hat{g}_i) - R^*(g_i^*)| \\ &\leq |R(\hat{g}_i) - R^*(\hat{g}_i)| + |R^*(\hat{g}_i) - R^*(g_i^*)| \\ &\leq \sup_{g_i \in M_i} |R(g_i) - R^*(g_i)| + |R^*(\hat{g}_i) - R^*(g_i^*)| \to 0. \end{aligned}$$

We generally have that

$$\left(\hat{w}_i^{\mathrm{AIC}}\right)^{-1} = 1 + \sum_{j \neq i} \left(\frac{\hat{\sigma}_i}{\hat{\sigma}_j}\right)^n e^{p_i - p_j}, \tag{4}$$

and if $R^*(g_i^*) > 0$ for all models, then (3) follows by continuity. If $R^*(g_{i_0}^*) = 0$ and $R^*(g_j^*) > 0$ for $j \neq i_0$, the sum in (4) converges to 0 for $i = i_0$ and $\hat{w}_{i_0}^{\mathrm{AIC}} \to 1$. If $i \neq i_0$, then $\hat{\sigma}_i \to \sqrt{R^*(g_i^*)} > 0$ while one of the denominators in the sum in (4) tends to 0, and the entire sum tends to ∞. Consequently, $\hat{w}_i^{\mathrm{AIC}} \to 0$ for $i \neq i_0$. This is all in accordance with (3) with the convention that $1/\infty = 0$.

To establish (2) under the boundedness assumption, we compute

$$\begin{aligned} R(g_i) &= \frac{1}{n} \sum_{m=1}^n (g_i(x_m) - g(x_m) - \sigma\varepsilon_m)^2 \\ &= \frac{1}{n} \sum_{m=1}^n (g_i(x_m) - g(x_m))^2 - \frac{2\sigma}{n} \sum_{m=1}^n (g_i(x_m) - g(x_m))\varepsilon_m + \frac{\sigma^2}{n} \sum_{m=1}^n \varepsilon_m^2 \\ &= R^*(g_i) - \frac{2\sigma}{n} \sum_{m=1}^n g_i(x_m)\varepsilon_m + \frac{2\sigma}{n} \sum_{m=1}^n g(x_m)\varepsilon_m + \frac{\sigma^2}{n} \sum_{m=1}^n \varepsilon_m^2. \end{aligned}$$

This gives

$$\begin{aligned} |R(g_i) - R^*(g_i)| &\leq \sigma \left(\frac{2}{n} \sum_{m=1}^n |g_i(x_m)||\varepsilon_m| + \frac{2}{n} \sum_{m=1}^n |g(x_m)||\varepsilon_m| + \frac{\sigma}{n} \sum_{m=1}^n \varepsilon_m^2 \right) \\ &\leq \sigma \left(\frac{2C}{n} \sum_{m=1}^n |\varepsilon_m| + \frac{2}{n} \sum_{m=1}^n |g(x_m)||\varepsilon_m| + \frac{\sigma}{n} \sum_{m=1}^n \varepsilon_m^2 \right). \end{aligned}$$

The last expression above is independent of g_i and tends to 0 for $\sigma \to 0$, and this shows (2). □

Note that $R^*(g_i^*) = 0$ is equivalent to the model M_i containing a function identical to the true dose-response curve in the chosen dose values. If this holds for more than one model it becomes more subtle to determine if the AIC weights converge, as this will require detailed knowledge of the rates of convergence to 0 of each of the corresponding $\hat{\sigma}_i$-s.

Note also that the first part of the proof above is a standard argument for establishing convergence of empirical risk, which is usually studied for $n \to \infty$ instead of $\sigma \to 0$. By replacing $R^*(g_i)$ with $R^*(g_i) + \sigma^2$, the above analysis could also be modified to obtain convergence of the AIC weights for $n \to \infty$. The conditions for (2) to hold would be more subtle, the convergence would be "in probability", and the limit weights would be degenerate at the model with the smallest $R^*(g_i^*)$-value. Since n is moderate in typical dose-response experiments, we found the "small noise asymptotics" more interesting, and even if the limit weights (3) are more complicated, it is clear from the formula that the weights will concentrate sharply on the models with the smallest values of $R^*(g_i^*)$ due to the n-th power.

Tables

Model	Setup A		Setup B		Setup C		Setup D	
	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$
LL.3	−7.00	−6.18	−5.48	−4.73	−5.47	−2.06	56.38	49.92
LL.4	†**0.00**	†**0.00**	3.04	1.36	−7.15	−5.56	12.00	17.43
LN.3	−4.42	−3.21	−3.08	−1.82	−2.70	0.95	64.18	62.37
LN.4	4.04	4.76	6.78	6.22	†**−2.00**	†**0.09**	32.17	35.89
W1.3	−28.29	−29.79	−26.30	−28.00	−26.78	−22.99	−23.32	−16.63
W1.4	−19.09	−15.18	−16.30	−13.40	−20.08	−16.50	†**4.21**	†**−0.82**
W2.3	21.54	24.86	22.66	25.91	18.82	24.33	120.02	‡*128.90*
W2.4	11.77	16.55	13.79	17.53	7.80	10.85	69.62	58.64
FPL.4($p_1 = -2, p_2 = 0.5$)	‡*41.75*	‡*36.10*	‡*45.38*	27.14	‡*42.01*	‡*40.87*	114.98	115.22
FPL.4($p_1 = -1, p_2 = 0.5$)	20.31	22.27	21.26	24.27	20.00	25.23	102.50	94.53
FPL.4($p_1 = -0.5, p_2 = 0.5$)	7.62	9.24	9.09	11.29	11.01	9.48	69.38	52.92
FPL.4($p_1 = -2, p_2 = 1$)	31.51	28.14	34.26	‡*30.06*	32.24	32.20	‡*121.30*	89.68
FPL.4($p_1 = -1, p_2 = 1$)	12.79	13.39	15.19	14.55	11.58	15.41	88.15	87.05
FPL.4($p_1 = -0.5, p_2 = 1$)	4.16	4.94	4.13	6.24	3.72	2.04	75.40	62.02
FPL.4($p_1 = -2, p_2 = 2$)	14.51	15.42	16.42	17.60	13.27	15.37	95.60	79.41
FPL.4($p_1 = -1, p_2 = 2$)	−1.07	1.24	†**1.62**	†**1.26**	2.89	0.31	50.06	57.27
FPL.4($p_1 = -0.5, p_2 = 2$)	−10.25	−5.90	−9.50	−6.14	−12.61	−9.21	34.29	22.06

Table S1: Asymptotic bias of all models included in the simulation study. The number marked with † and written in bold, was the asymptotic bias of the model with the lowest absolute bias for each scenario (referred to as the best model). The number marked with ‡ and written in italics, was the asymptotic bias of the model with the *largest* absolute bias for each scenario (referred to as the worst model in the set). ".3" and ".4" refer to 3- or 4-parameter versions of each model.

Strategy	$\mathcal{M}$	w
min(AIC)	Model with smallest AIC value	-
Best model	Single best model	-
Worst model	Single worst model	-
MA_{AIC} (post)	Full	AIC
MA_{AIC} (curve)	Full	AIC
MA_{Stack} (post)	Full	Stacking
MA_{Stack} (curve)	Full	Stacking
Cl.M_{AIC} (post)	$\mathcal{M}_{\text{ClM}}$	AIC
Cl.M_{AIC} (curve)	$\mathcal{M}_{\text{ClM}}$	AIC
$\text{Cl.M}_{\text{Stack}}$ (post)	$\mathcal{M}_{\text{ClM}}$	Stacking
$\text{Cl.M}_{\text{Stack}}$ (curve)	$\mathcal{M}_{\text{ClM}}$	Stacking
FPL_{AIC} (post)	$\mathcal{M}_{\text{FPL}}$	AIC
FPL_{AIC} (curve)	$\mathcal{M}_{\text{FPL}}$	AIC
$\text{FPL}_{\text{Stack}}$ (post)	$\mathcal{M}_{\text{FPL}}$	Stacking
$\text{FPL}_{\text{Stack}}$ (curve)	$\mathcal{M}_{\text{FPL}}$	Stacking
F-Test	F-Test selection	AIC
$\text{MA}_{\text{AIC}}\,\text{DI}_{(}\lambda = 1, \gamma = 1)$	$\mathcal{M}_0^*(\hat{w}^{\text{AIC}}, 1, 1)$	AIC
$\text{MA}_{\text{AIC}}\,\text{DI}(\lambda = 3, \gamma = 1)$	$\mathcal{M}_0^*(\hat{w}^{\text{AIC}}, 3, 1)$	AIC
$\text{MA}_{\text{AIC}}\,\text{DI}(\lambda = 5, \gamma = 1)$	$\mathcal{M}_0^*(\hat{w}^{\text{AIC}}, 5, 1)$	AIC
$\text{MA}_{\text{AIC}}\,\text{DI}(\lambda = 1, \gamma = 5)$	$\mathcal{M}_0^*(\hat{w}^{\text{AIC}}, 1, 5)$	AIC
$\text{MA}_{\text{AIC}}\,\text{DI}(\lambda = 3, \gamma = 5)$	$\mathcal{M}_0^*(\hat{w}^{\text{AIC}}, 3, 5)$	AIC
$\text{MA}_{\text{AIC}}\,\text{DI}(\lambda = 5, \gamma = 5)$	$\mathcal{M}_0^*(\hat{w}^{\text{AIC}}, 5, 5)$	AIC
$\text{MA}_{\text{Stack}}\,\text{DI}(\lambda = 1, \gamma = 1)$	$\mathcal{M}_0^*(\tilde{w}^{\text{Stack}}, 1, 1)$	Stacking
$\text{MA}_{\text{Stack}}\,\text{DI}(\lambda = 3, \gamma = 1)$	$\mathcal{M}_0^*(\tilde{w}^{\text{Stack}}, 3, 1)$	Stacking
$\text{MA}_{\text{Stack}}\,\text{DI}(\lambda = 5, \gamma = 1)$	$\mathcal{M}_0^*(\tilde{w}^{\text{Stack}}, 5, 1)$	Stacking
$\text{MA}_{\text{Stack}}\,\text{DI}I(\lambda = 1, \gamma = 5)$	$\mathcal{M}_0^*(\tilde{w}^{\text{Stack}}, 1, 5)$	Stacking
$\text{MA}_{\text{Stack}}\,\text{DI}(\lambda = 3, \gamma = 5)$	$\mathcal{M}_0^*(\tilde{w}^{\text{Stack}}, 3, 5)$	Stacking
$\text{MA}_{\text{Stack}}\,\text{DI}(\lambda = 5, \gamma = 5)$	$\mathcal{M}_0^*(\tilde{w}^{\text{Stack}}, 5, 5)$	Stacking

Table S2: Overview of model averaging strategies with model space ($\mathcal{M}$) and weight type (w) included in simulation study I.

	Setup A				Setup B				Setup C				Setup D			
	$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$		$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$		$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$		$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$	
Method	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$
AIC mod.	1.69 (18.77)	1.10 (13.33)	1.28 (12.66)	0.67 (9.54)	2.58 (18.75)	0.81 (12.76)	2.18 (12.71)	0.22 (8.64)	4.50 (16.43)	3.67 (11.36)	4.64 (11.13)	4.47 (8.67)	108.60 (145.90)	99.70 (121.29)	101.09 (122.05)	94.28 (106.76)
Best mod.	1.01 (14.68)	0.07 (10.45)	0.48 (10.31)	-0.02 (7.37)	6.07 (15.99)	3.86 (10.87)	5.42 (11.50)	3.99 (8.09)	4.11 (11.31)	6.44 (10.19)	2.80 (7.57)	6.19 (8.08)	74.27 (115.73)	63.85 (77.86)	54.08 (70.52)	57.67 (64.39)
Worst mod.	36.23 (40.56)	32.43 (34.07)	34.16 (36.60)	26.69 (27.61)	39.14 (43.46)	30.13 (31.80)	36.91 (39.59)	24.07 (25.07)	37.68 (40.47)	40.49 (41.67)	34.91 (36.38)	32.57 (32.99)	185.45 (196.01)	198.85 (209.39)	174.16 (180.70)	196.55 (201.30)
$\text{MA}^{\text{post}}_{\text{AIC}}$	4.77 (17.42)	3.79 (12.60)	3.93 (12.41)	2.54 (9.12)	6.30 (18.26)	3.72 (12.39)	5.11 (13.03)	2.30 (8.67)	5.73 (14.49)	4.70 (10.74)	5.63 (10.77)	5.00 (8.38)	125.47 (151.46)	117.85 (133.06)	118.13 (132.78)	109.23 (117.91)
$\text{MA}^{\text{curve}}_{\text{AIC}}$	5.12 (17.58)	4.09 (12.78)	4.20 (12.55)	2.75 (9.24)	6.63 (18.44)	4.01 (12.56)	5.34 (13.17)	2.48 (8.78)	6.05 (14.70)	4.91 (10.88)	5.90 (10.94)	5.15 (8.49)	125.63 (151.77)	118.53 (133.99)	118.56 (133.39)	109.63 (118.46)
$\text{MA}^{\text{post}}_{\text{Stack}}$	2.24 (15.99)	2.20 (10.97)	3.45 (11.91)	-0.50 (8.15)	3.84 (16.95)	2.30 (10.87)	4.66 (12.62)	-0.51 (7.86)	5.06 (13.28)	5.09 (10.75)	4.29 (10.15)	1.29 (6.66)	133.36 (155.55)	121.74 (133.65)	114.99 (125.56)	107.03 (113.70)
$\text{MA}^{\text{curve}}_{\text{Stack}}$	3.18 (16.52)	3.17 (11.60)	4.43 (12.41)	0.45 (8.45)	4.72 (17.51)	3.25 (11.51)	5.56 (13.14)	0.40 (8.16)	5.53 (13.84)	5.52 (11.42)	4.85 (10.75)	1.91 (7.17)	133.56 (156.04)	123.42 (135.97)	115.77 (126.87)	108.47 (115.52)
$\text{Cl.M.}^{\text{post}}_{\text{AIC}}$	1.63 (16.56)	1.57 (12.27)	1.39 (11.97)	0.36 (8.82)	2.91 (16.86)	1.61 (11.91)	2.69 (12.20)	0.59 (8.22)	3.11 (13.68)	2.83 (9.92)	2.82 (9.54)	3.84 (7.49)	117.82 (146.24)	109.09 (124.96)	109.42 (124.79)	100.72 (109.74)
$\text{Cl.M.}^{\text{curve}}_{\text{AIC}}$	2.05 (16.76)	1.90 (12.49)	1.67 (12.15)	0.53 (8.96)	3.30 (17.08)	1.90 (12.10)	2.93 (12.37)	0.72 (8.33)	3.43 (13.93)	2.98 (10.06)	3.02 (9.69)	3.92 (7.56)	118.29 (146.98)	109.91 (126.23)	109.90 (125.68)	101.01 (110.28)
$\text{Cl.M.}^{\text{post}}_{\text{Stack}}$	-0.35 (15.86)	0.92 (10.55)	0.57 (10.87)	1.73 (7.79)	1.01 (16.12)	1.14 (10.41)	1.63 (11.14)	1.79 (7.70)	2.17 (12.60)	3.35 (9.57)	2.37 (9.16)	3.20 (7.43)	120.13 (146.60)	116.09 (128.69)	112.77 (124.05)	109.98 (116.57)
$\text{Cl.M.}^{\text{curve}}_{\text{Stack}}$	1.06 (16.17)	2.19 (11.02)	1.91 (11.21)	2.90 (8.34)	2.32 (16.49)	2.38 (10.91)	2.88 (11.55)	2.91 (8.23)	3.21 (13.20)	4.28 (10.26)	3.37 (9.77)	4.04 (8.09)	121.22 (148.11)	118.58 (131.84)	114.77 (126.66)	112.26 (119.32)
$\text{FPL}^{\text{post}}_{\text{AIC}}$	8.10 (18.32)	5.99 (13.04)	5.89 (12.85)	3.98 (9.50)	10.07 (19.69)	6.09 (12.98)	7.19 (13.65)	3.70 (9.12)	8.60 (15.60)	7.00 (11.78)	7.77 (12.04)	6.10 (9.37)	137.29 (159.22)	132.52 (145.81)	130.53 (143.26)	124.95 (131.93)
$\text{FPL}^{\text{curve}}_{\text{AIC}}$	8.42 (18.47)	6.28 (13.21)	6.17 (12.99)	4.21 (9.63)	10.39 (19.86)	6.39 (13.15)	7.46 (13.80)	3.91 (9.24)	8.90 (15.77)	7.25 (11.95)	8.03 (12.20)	6.30 (9.51)	137.19 (159.25)	133.01 (146.35)	130.99 (143.75)	125.57 (132.57)
$\text{FPL}^{post}_{\text{Stack}}$	17.39 (23.86)	14.18 (17.79)	14.72 (18.71)	9.94 (12.77)	20.03 (26.31)	15.02 (18.44)	17.34 (21.18)	10.55 (13.21)	17.14 (21.33)	14.74 (17.70)	12.75 (15.66)	10.28 (12.21)	155.32 (171.00)	148.86 (159.09)	143.32 (153.15)	136.96 (142.81)
$\text{FPL}^{\text{curve}}_{\text{Stack}}$	17.62 (24.05)	14.53 (18.11)	15.01 (18.95)	10.32 (13.09)	20.26 (26.49)	15.35 (18.75)	17.62 (21.40)	10.91 (13.52)	17.30 (21.47)	14.84 (17.85)	12.97 (15.88)	10.48 (12.42)	154.82 (170.73)	149.13 (159.45)	143.49 (153.41)	137.48 (143.35)
F-test	5.11 (17.59)	4.07 (12.77)	4.19 (12.55)	2.71 (9.25)	6.62 (18.44)	3.97 (12.54)	5.33 (13.17)	2.44 (8.79)	6.06 (14.70)	4.88 (10.91)	5.90 (10.94)	5.09 (8.52)	125.31 (151.63)	118.41 (133.94)	118.12 (133.02)	109.14 (118.08)
$\text{DI}(\lambda_1, \gamma_1)_{\text{AIC}}$	5.18 (17.70)	4.14 (12.84)	4.25 (12.61)	2.76 (9.28)	6.65 (18.54)	4.04 (12.61)	5.38 (13.24)	2.48 (8.81)	6.09 (14.75)	4.92 (10.91)	5.92 (10.97)	5.16 (8.50)	126.11 (152.53)	117.97 (134.09)	119.02 (134.31)	108.43 (117.85)
$\text{DI}(\lambda_2, \gamma_1)_{\text{AIC}}$	5.23 (18.58)	3.77 (13.36)	3.96 (12.99)	2.27 (9.55)	6.33 (19.12)	3.40 (13.00)	4.84 (13.45)	1.81 (8.92)	6.27 (15.52)	4.48 (11.25)	6.01 (11.39)	4.82 (8.55)	122.71 (153.04)	111.85 (132.19)	113.64 (132.83)	99.29 (111.61)
$\text{DI}(\lambda_3, \gamma_1)_{\text{AIC}}$	4.99 (19.11)	3.34 (13.58)	3.49 (13.30)	1.80 (9.60)	5.88 (19.65)	2.94 (13.23)	4.23 (13.53)	1.31 (9.00)	6.36 (16.38)	4.15 (11.52)	5.88 (11.69)	4.67 (8.64)	122.56 (153.81)	111.89 (133.03)	112.76 (132.72)	98.99 (111.89)
$\text{DI}(\lambda_1, \gamma_2)_{\text{AIC}}$	5.96 (19.51)	4.32 (13.53)	4.43 (13.15)	2.73 (9.57)	7.27 (20.32)	4.02 (13.12)	5.30 (13.54)	2.30 (8.96)	6.31 (15.28)	4.87 (11.16)	6.04 (11.23)	5.11 (8.57)	142.67 (175.18)	127.80 (155.31)	126.81 (150.36)	104.90 (121.00)
$\text{DI}(\lambda_2, \gamma_2)_{\text{AIC}}$	5.28 (19.61)	3.47 (13.93)	3.43 (13.44)	1.75 (9.72)	6.08 (20.22)	2.91 (13.49)	4.10 (13.56)	1.27 (9.06)	6.43 (16.64)	4.15 (11.72)	5.95 (11.77)	4.68 (8.68)	129.02 (161.07)	115.72 (139.03)	115.66 (137.06)	100.76 (114.79)
$\text{DI}(\lambda_3, \gamma_2)_{\text{AIC}}$	5.44 (20.16)	3.39 (14.16)	3.15 (13.63)	1.48 (9.83)	5.89 (20.47)	2.80 (13.71)	3.74 (13.51)	0.96 (9.01)	6.61 (17.33)	4.23 (11.87)	5.80 (11.90)	4.61 (8.69)	127.00 (158.53)	115.13 (137.94)	115.14 (136.23)	100.54 (114.49)
$\text{DI}(\lambda_1, \gamma_1)_{\text{Stack}}$	3.23 (16.65)	3.21 (11.66)	4.48 (12.51)	0.46 (8.48)	4.73 (17.67)	3.26 (11.55)	5.60 (13.22)	0.41 (8.18)	5.54 (13.87)	5.52 (11.44)	4.85 (10.77)	1.91 (7.18)	136.51 (160.15)	123.94 (137.76)	115.75 (127.69)	106.51 (113.77)
$\text{DI}(\lambda_2, \gamma_1)_{\text{Stack}}$	3.78 (19.00)	4.64 (14.43)	6.83 (15.91)	1.15 (10.17)	5.03 (19.95)	4.01 (13.29)	7.42 (16.23)	0.57 (9.41)	5.78 (15.77)	5.14 (12.49)	4.99 (12.09)	1.00 (7.66)	139.57 (164.45)	122.96 (138.96)	114.07 (127.98)	100.02 (109.60)
$\text{DI}(\lambda_3, \gamma_1)_{\text{Stack}}$	5.98 (22.07)	6.97 (17.00)	9.63 (19.07)	2.89 (12.41)	6.48 (22.11)	6.07 (16.20)	10.42 (19.62)	2.09 (11.47)	6.54 (17.96)	5.13 (14.06)	6.02 (14.71)	0.64 (8.41)	141.43 (166.35)	124.30 (141.23)	113.15 (128.38)	98.94 (109.73)
$\text{DI}(\lambda_1, \gamma_2)_{\text{Stack}}$	9.17 (27.53)	7.31 (18.41)	9.59 (20.31)	2.37 (12.02)	10.64 (29.20)	6.17 (16.92)	10.47 (21.17)	1.72 (11.04)	6.80 (17.48)	5.90 (13.34)	5.19 (11.94)	1.56 (7.32)	183.61 (200.36)	177.12 (194.79)	168.37 (181.66)	158.45 (174.76)
$\text{DI}(\lambda_2, \gamma_2)_{\text{Stack}}$	9.73 (26.37)	11.08 (20.94)	13.30 (22.21)	5.93 (15.23)	10.66 (27.53)	9.17 (19.40)	13.81 (22.94)	4.55 (14.33)	8.95 (21.23)	6.33 (15.94)	7.79 (17.58)	1.01 (9.07)	159.82 (182.39)	145.05 (164.96)	135.13 (153.29)	115.71 (131.34)
$\text{DI}(\lambda_3, \gamma_2)_{\text{Stack}}$	10.33 (25.88)	12.86 (21.83)	14.90 (23.11)	7.42 (16.18)	10.78 (26.46)	10.77 (20.21)	15.25 (23.77)	5.64 (15.08)	10.11 (22.64)	7.14 (17.01)	9.84 (19.74)	1.68 (10.14)	154.72 (177.74)	136.57 (156.24)	126.91 (145.09)	110.12 (124.52)

Table S3: Mean bias and root mean squared error (values in parenthesis) observed in simulation study I.

	Setup A				Setup B				Setup C				Setup D			
	$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$		$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$		$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$		$n_{\text{rep}} = 5$		$n_{\text{rep}} = 10$	
Method	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$	$\mathbf{x}_5$	$\mathbf{x}_{10}$
AIC mod.	0.82	0.84	0.85	0.85	0.81	0.85	0.85	0.87	0.73	0.75	0.74	0.68	0.36	0.24	0.25	0.10
Best mod.	0.91	0.92	0.93	0.94	0.85	0.84	0.83	0.85	0.96	0.88	0.96	0.80	0.54	0.34	0.43	0.15
Worst mod.	0.17	0.00	0.05	0.00	0.13	0.01	0.03	0.00	0.15	0.00	0.00	0.00	0.01	0.01	0.00	0.00
$\text{MA}_{\text{AIC}}^{\text{post}}$	0.88	0.88	0.88	0.90	0.87	0.88	0.87	0.91	0.84	0.87	0.81	0.81	0.32	0.19	0.17	0.06
$\text{MA}_{\text{AIC}}^{\text{curve}}$	0.88	0.88	0.88	0.90	0.87	0.88	0.87	0.91	0.83	0.86	0.81	0.80	0.33	0.19	0.18	0.06
$\text{MA}_{\text{Stack}}^{\text{post}}$	0.91	0.89	0.89	0.94	0.89	0.88	0.88	0.94	0.86	0.82	0.84	0.87	0.16	0.07	0.04	0.02
$\text{MA}_{\text{Stack}}^{\text{curve}}$	0.90	0.87	0.88	0.91	0.89	0.87	0.87	0.91	0.86	0.82	0.83	0.86	0.21	0.08	0.07	0.02
$\text{Cl.M.}_{\text{AIC}}^{\text{post}}$	0.91	0.91	0.93	0.94	0.90	0.90	0.91	0.94	0.89	0.89	0.88	0.85	0.37	0.21	0.22	0.06
$\text{Cl.M.}_{\text{AIC}}^{\text{curve}}$	0.91	0.90	0.92	0.94	0.89	0.90	0.91	0.94	0.88	0.89	0.88	0.85	0.37	0.21	0.22	0.06
$\text{Cl.M.}_{\text{Stack}}^{\text{post}}$	0.93	0.91	0.93	0.92	0.92	0.91	0.92	0.92	0.91	0.87	0.88	0.88	0.30	0.12	0.12	0.02
$\text{Cl.M.}_{\text{Stack}}^{\text{curve}}$	0.93	0.89	0.92	0.90	0.92	0.89	0.91	0.91	0.89	0.85	0.87	0.85	0.31	0.14	0.13	0.02
$\text{FPL}_{\text{AIC}}^{\text{post}}$	0.84	0.83	0.84	0.86	0.81	0.83	0.83	0.87	0.78	0.81	0.76	0.76	0.22	0.10	0.09	0.02
$\text{FPL}_{\text{AIC}}^{\text{curve}}$	0.83	0.83	0.84	0.85	0.80	0.83	0.82	0.86	0.77	0.80	0.75	0.75	0.23	0.10	0.09	0.02
$\text{FPL}_{\text{Stack}}^{post}$	0.64	0.56	0.58	0.58	0.57	0.54	0.52	0.53	0.48	0.40	0.48	0.44	0.06	0.01	0.00	0.00
$\text{FPL}_{\text{Stack}}^{\text{curve}}$	0.63	0.55	0.57	0.55	0.57	0.52	0.51	0.52	0.49	0.40	0.48	0.44	0.07	0.02	0.01	0.00
F-test	0.88	0.87	0.87	0.89	0.87	0.87	0.85	0.90	0.83	0.86	0.80	0.78	0.33	0.19	0.18	0.06
$\text{DI}(\lambda_1, \gamma_1)_{\text{AIC}}$	0.86	0.86	0.86	0.89	0.85	0.86	0.84	0.90	0.82	0.84	0.80	0.77	0.33	0.19	0.18	0.06
$\text{DI}(\lambda_2, \gamma_1)_{\text{AIC}}$	0.84	0.85	0.84	0.86	0.82	0.86	0.83	0.88	0.78	0.81	0.77	0.72	0.33	0.21	0.24	0.08
$\text{DI}(\lambda_3, \gamma_1)_{\text{AIC}}$	0.82	0.84	0.84	0.85	0.80	0.86	0.82	0.87	0.74	0.78	0.75	0.70	0.32	0.22	0.24	0.09
$\text{DI}(\lambda_1, \gamma_2)_{\text{AIC}}$	0.83	0.85	0.84	0.86	0.81	0.86	0.82	0.88	0.79	0.82	0.78	0.75	0.28	0.21	0.22	0.09
$\text{DI}(\lambda_2, \gamma_2)_{\text{AIC}}$	0.81	0.84	0.84	0.84	0.79	0.85	0.83	0.87	0.74	0.77	0.74	0.69	0.31	0.22	0.23	0.09
$\text{DI}(\lambda_3, \gamma_2)_{\text{AIC}}$	0.78	0.83	0.83	0.84	0.79	0.84	0.83	0.85	0.71	0.76	0.73	0.67	0.31	0.21	0.23	0.09
$\text{DI}(\lambda_1, \gamma_1)_{\text{Stack}}$	0.91	0.87	0.88	0.91	0.90	0.87	0.87	0.91	0.86	0.83	0.84	0.87	0.20	0.08	0.08	0.03
$\text{DI}(\lambda_2, \gamma_1)_{\text{Stack}}$	0.88	0.82	0.79	0.85	0.88	0.83	0.81	0.88	0.86	0.82	0.82	0.88	0.21	0.13	0.09	0.05
$\text{DI}(\lambda_3, \gamma_1)_{\text{Stack}}$	0.81	0.73	0.68	0.77	0.83	0.76	0.69	0.82	0.81	0.81	0.78	0.86	0.21	0.13	0.10	0.06
$\text{DI}(\lambda_1, \gamma_2)_{\text{Stack}}$	0.76	0.75	0.70	0.81	0.76	0.77	0.72	0.85	0.85	0.82	0.83	0.88	0.09	0.05	0.03	0.03
$\text{DI}(\lambda_2, \gamma_2)_{\text{Stack}}$	0.71	0.60	0.56	0.68	0.73	0.67	0.59	0.75	0.74	0.77	0.72	0.85	0.16	0.11	0.10	0.06
$\text{DI}(\lambda_3, \gamma_2)_{\text{Stack}}$	0.69	0.53	0.50	0.64	0.72	0.62	0.54	0.71	0.69	0.74	0.64	0.81	0.18	0.11	0.12	0.06

Table S4: Coverage of benchmark dose lower limit (BMDL) observed in simulation study II.

Figures

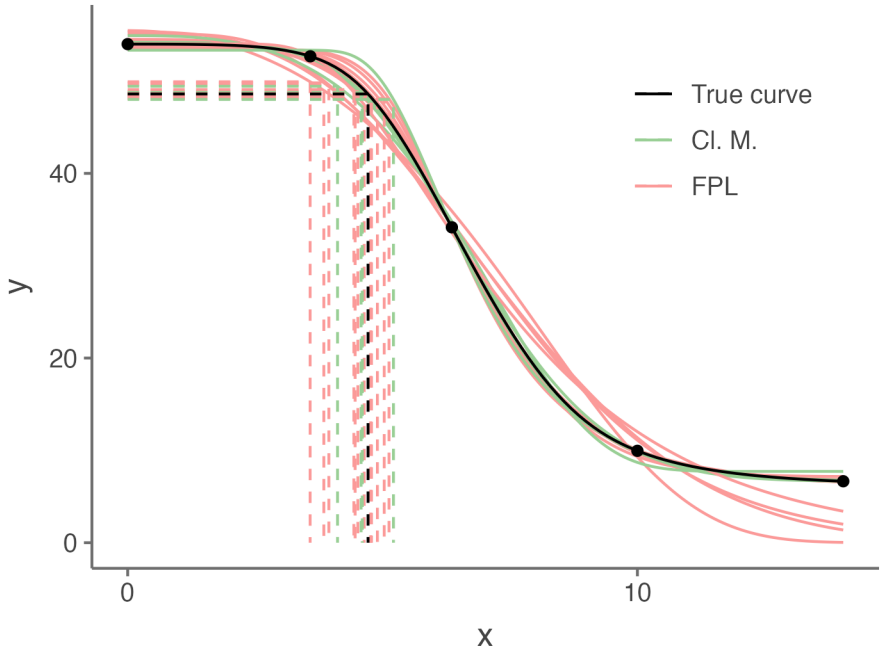


Figure S1: True underlying dose-response curve, dose levels and optimal fitted curves to data with no residual variance for all types of models included in simulation study III. The colors on the optimal fitted curves refer to whether each model is in the set of classical dose-response models (Cl. Mods, including three- and four-parameter Log-Logistic, Log-Normal, and Weibull 1 and 2 curves) or in the set of included Fractional Polynomial models (FPL, models with fixed parameters $p_1 \in \{-2, -1, -0.5\}$ and $p_2 \in \{0.5, 1, 2\}$ are included).

Aciflourfen data example

Required packages

```
# install.packages("tidyverse")
library(tidyverse)
# install.packages("latex2exp")
library(latex2exp) # for nice axes on plots
# install.packages("CVXR",repos = "https://cloud.r-project.org/")
library(CVXR)
# install.packages("devtools",repos = "https://cloud.r-project.org/")
library(devtools)
# install_github("doseResponse/drcData")
# install_github("doseResponse/drc")
# install_github("doseResponse/bmd")
library(drcData)
library(drc)
library(bmd)
```

Functions for DI model space selection

```
# Diversity Index ------------------------------------------------------------
DI_single_subset <- function(subset, KL_div, w, lambda = 1, gamma = 1, data){
  k0 <- length(subset)
  k <- length(KL_div)

  g <- sum(w[subset]^lambda)
  value <- g^(1/gamma) * k/k0 * sum(KL_div[subset])
  value
}

KL_div_est <- function(f0, f1, lower = 1, upper = 5){
  int <- try(integrate(f = function(d){
    abs(log(abs(f0(d))) - log(abs(f1(d)))) * abs(f0(d))},
    lower, upper), silent = TRUE)
  if(inherits(int, "try-error")){return(0)}
  else{return(int$value)}
}

get_all_KL_div <- function(model_fit_list, fct_derivx_list, w, lower, upper){
  n_models <- length(model_fit_list)
  derivx_MA <- function(x) sum((sapply(fct_derivx_list,
                                       function(fct) fct(x))) * w)
  vals <- sapply(fct_derivx_list, function(df){
    KL_div_est(f0 = derivx_MA,
               f1 = df,
               lower, upper)
  })

  vals
```

```
}

DI_model_space <- function(modelList, w, lambda = 1, gamma = 1,
                           lower = 2, upper = 4, data){
  k <- length(w)
  fct_derivx_list <- lapply(modelList, curveDerivx)
  KL_div <- get_all_KL_div(modelList, fct_derivx_list, w, lower, upper)

  subsets <- lapply(2:k, function(setsize){
    combn(1:k, setsize, simplify = FALSE)}) |>
    unlist(recursive = FALSE)
  # combn returns combinations of size setsize of 1:J
  # previous output is a list of lists, this makes sure we just have a list of the sets.

  DI_vals <- sapply(subsets, function(subset){
    DI_single_subset(subset, KL_div, w, lambda, gamma, data)})
  return(subsets[[which.max(DI_vals)]])
}



# Curve functions derived in x ---------------------------------------------
curveDerivx <- function(model){
  fctName <- model$fct$name

  if(fctName == "LL.3") LL3_derivx(model$fit$par)
  else if(fctName == "LL.4") LL4_derivx(model$fit$par)
  else if(fctName == "LN.3") LN3_derivx(model$fit$par)
  else if(fctName == "LN.4") LN4_derivx(model$fit$par)
  else if(fctName == "W1.3") W13_derivx(model$fit$par)
  else if(fctName == "W1.4") W14_derivx(model$fit$par)
  else if(fctName == "W2.3") W23_derivx(model$fit$par)
  else if(fctName == "W2.4") W24_derivx(model$fit$par)
  else if(substr(fctName, 1,5) == "FPL.4") get_FP_derivx(
    eval(parse(text=paste("c", substring(fctName, 6), sep = "")))) # extract c(p1,p2)
  )(model$fit$par)
}


# LL4
LL4_derivx <- function(par){
  function(x){
    b <- par[1]
    c <- par[2]
    d <- par[3]
    e <- par[4]

    dmc <- d-c
    exp_bxme <- exp(b*(log(x)-log(e)))
    -dmc * b * exp_bxme / (1 + exp_bxme)^2 / x
  }
}
```

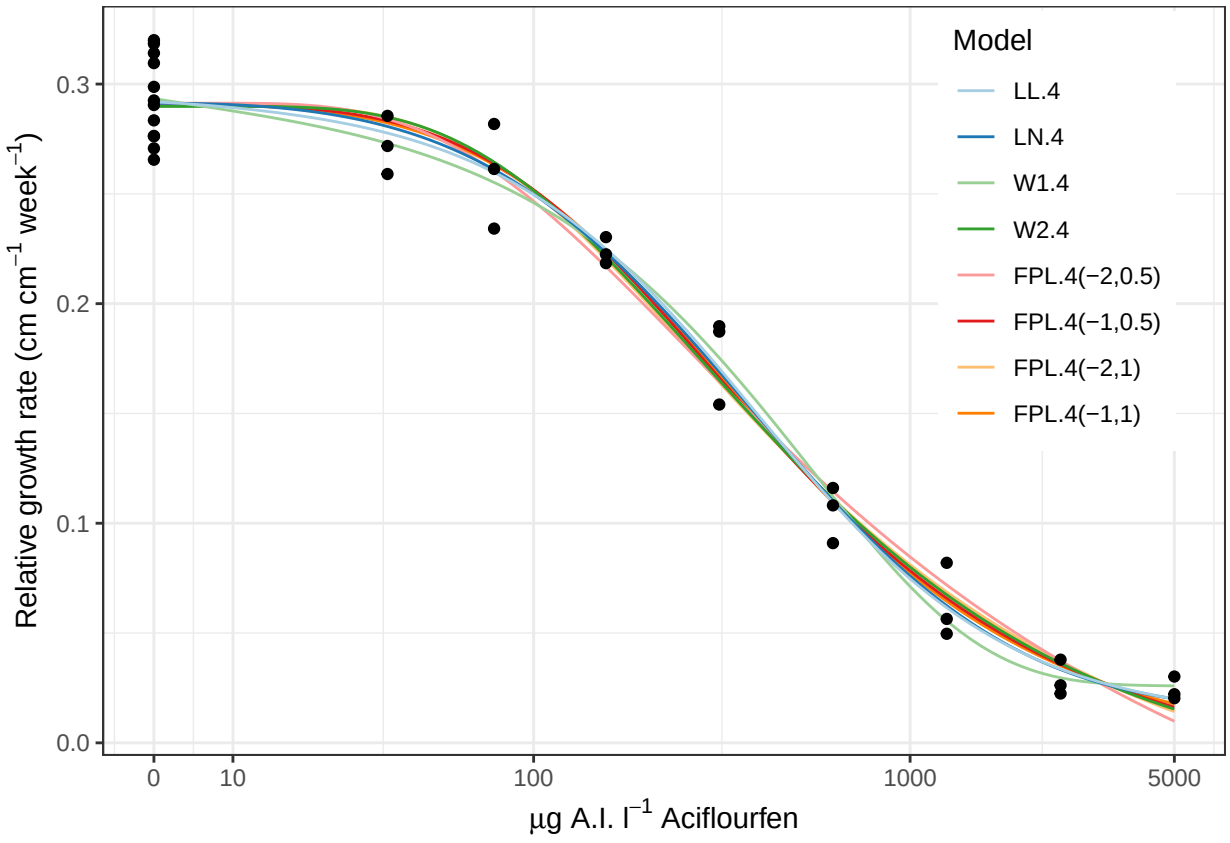
```r
LL3_derivx <- function(par){
  LL4_derivx(par = c(par[1], 0, par[2], par[3]))
}

# LN4
LN4_derivx <- function(par){
  function(x){
    b <- par[1]
    c <- par[2]
    d <- par[3]
    e <- par[4]

    dmc <- d-c
    dmc * dnorm(b*(log(x)-log(e))) * b/x
  }
}

LN3_derivx <- function(par){LN4_derivx(c(par[1], 0, par[2], par[3]))}

# W14
W14_derivx <- function(par){
  function(x){
    b <- par[1]
    c <- par[2]
    d <- par[3]
    e <- par[4]

    dmc <- d - c
    log_xme <- log(x) - log(e)
    exp_bxme <- exp(b*(log_xme))
    exp_mexp_bxme <- exp(-exp_bxme)

    - dmc * b * exp_bxme * exp_mexp_bxme / x
  }
}

W13_derivx <- function(par){W14_derivx(c(par[1], 0, par[2], par[3]))}

# W24
W24_derivx <- function(par){
  function(x){
    b <- par[1]
    c <- par[2]
    d <- par[3]
    e <- par[4]

    dmc <- d - c
    log_xme <- log(x) - log(e)
    exp_bxme <- exp(b*(log_xme))
    exp_mexp_bxme <- exp(-exp_bxme)

    dmc * b * exp_bxme * exp_mexp_bxme / x
  }
```

```
}

W23_derivx <- function(par){W24_derivx(c(par[1], 0, par[2], par[3]))}

# FP4
get_FP_derivx <- function(p){
  FP_derivx <- function(par){
    function(x){
      b <- par[1]
      c <- par[2]
      d <- par[3]
      e <- par[4]

      dmc <- d - c
      xp1 <- x + 1
      log_xp1 <- log(x + 1)
      exp_tmp <- exp(b * log_xp1^p[1] + e*log_xp1^p[2])
      frac1 <- b * log_xp1^p[1]*p[1] / (xp1*log_xp1)
      frac2 <- e * log_xp1^p[2]*p[2] / (xp1*log_xp1)

      - dmc * (frac1 + frac2) * exp_tmp / (1 + exp_tmp)^2
    }
  }
  FP_derivx
}
```

Fit models

```
# Load data, choose models ------------------------------------------------

data("acidiq")
exData <- subset(acidiq, pct %in% c(999,100))

p1s <- c(-2, -1)
p2s <- c(0.5, 1)
FPs <- outer(X = p1s, Y = p2s, FUN = function(p1,p2){
  paste("FPL.4(p1=", p1, ",", " p2=", p2,")", sep = "")}) |> as.character()
modelFcts <- c("LL.4()", "LN.4()", "W1.4()", "W2.4()", FPs)


# Set BMD definition and BMR ----------------------------------------------
def <- "relative"
bmr <- 0.1


# Parameters for confidence intervals -------------------------------------
bootR <- 500
level <- 0.95


# Fit models --------------------------------------------------------------
modelList <- lapply(modelFcts,
```

```
                    function(modelString) drm(rgr ~ dose, data = exData,
                                              fct = eval(parse(text = modelString)),
                                              type = "continuous"))
# Plot
colours <- c("black", "#A6CEE3", "#1F78B4","#99cf95", "#33A02C", "#FB9A99", "#E31A1C",
             "#FDBF6F", "#FF7F00", "#CAB2D6", "#6A3D9A", "#FFFF99", "#B15928")
app1_plot <- ggplot(exData) +
  lapply(modelList[8:1], function(x){
    geom_function(aes(col = x$fct$name), fun = x$curve[[1]])
  }) +
  geom_point(aes(x = dose, y = rgr), col = colours[1]) +
  scale_color_manual(breaks = sapply(modelList, function(x) x$fct$name),
                     values = colours[2:9]) +
  scale_x_continuous(transform = scales::pseudo_log_trans(sigma = 10),
                     breaks = c(0,10,100,1000,5000)) +
  labs(col = "Model", x = TeX("$\\mu g$ A.I. $l^{-1}$ Aciflourfen"),
       y = TeX("Relative growth rate (cm $cm^{-1}$ $week^{-1}$)")) +
  theme_bw() +
  theme(legend.position = c(0.85,0.7))
app1_plot
```



BMD estimation

```
# BMD Estimation ------------------------------------------------------------
# minimum AIC value
whichAICMin <- which.min(sapply(modelList, function(model) AIC(model)))

set.seed(2023)
bmdAICMin <- bmdBoot(modelList[[whichAICMin]], bmr = bmr, backgType = "modelBased",
                     def = def, bootType = "nonparametric",
                     bootInterval = "percentile", R = bootR, level = level)
```

```
##       BMD     BMDL
##  58.16225 43.28719
```

```
# Compute weights
AICWeights <- exp(-(sapply(modelList,AIC)-min(sapply(modelList,AIC))))/
  sum(exp(-(sapply(modelList,AIC)-min(sapply(modelList,AIC)))))
set.seed(2023)
StackWeights <- getStackingWeights(modelList, nSplits = 3)
print(StackWeights)
```

```
## [1] 3.134158e-01 1.815964e-21 5.066936e-01 7.902277e-20 2.215077e-02
## [6] 4.799909e-20 1.577399e-01 2.449790e-20
```

```
# bmdMA
set.seed(2023)
bmdMAAICPost <- bmdMA(modelList, modelWeights = "AIC", bmr = bmr,
                      backgType = "modelBased", def = def, type = "bootstrap",
                      bootstrapType = "nonparametric",
                      bootInterval = "percentile", R = bootR, level = level,
                      progressInfo = FALSE)
```

```
##    BMD_MA  BMDL_MA
##  66.48764 45.16161
```

```
set.seed(2023)
bmdMAAICCurve <- bmdMA(modelList, modelWeights = "AIC", bmr = bmr,
                       backgType = "modelBased", def = def, type = "curve",
                       bootstrapType = "nonparametric",
                       bootInterval = "percentile", R = bootR, level = level,
                       progressInfo = FALSE)
```

```
##   BMD_MA BMDL_MA
##  66.8208 45.1868
```

```
set.seed(2023)
bmdMAStackPost <- bmdMA(modelList, modelWeights = "Stack", bmr = bmr,
                        backgType = "modelBased", def = def, type = "bootstrap",
                        bootstrapType = "nonparametric",
                        bootInterval = "percentile", R = bootR, level = level,
                        stackingSplits = 3, progressInfo = FALSE)
```

```
##    BMD_MA BMDL_MA
##  66.76402 47.5012
```

```
set.seed(2023)
bmdMAStackCurve <- bmdMA(modelList, modelWeights = "Stack", bmr = bmr,
                         backgType = "modelBased", def = def, type = "curve",
                         bootstrapType = "nonparametric",
                         bootInterval = "percentile", R = bootR, level = level,
                         stackingSplits = 3, progressInfo = FALSE)
```

```
##    BMD_MA  BMDL_MA
##  67.67184 48.32985
```

Apply DI

```
# Apply DI
lambda <- c(1,3,5)
gamma <- 1

(AICDILambda1 <- DI_model_space(modelList, AICWeights, lambda = lambda[1],
                                gamma = gamma, lower = 10, upper = 100, data = exData))
```

```
## [1] 1 2 3 4 6 7 8
```

```
(AICDILambda2 <- DI_model_space(modelList, AICWeights, lambda = lambda[2],
                                gamma = gamma, lower = 10, upper = 100, data = exData))
```

```
## [1] 1 2 3 4 6 7
```

```
(AICDILambda3 <- DI_model_space(modelList, AICWeights, lambda = lambda[3],
                                gamma = gamma, lower = 10, upper = 100, data = exData))
```

```
## [1] 1 3 4 6 7
```

```
AICModelSubspaceList <- list(AICDILambda1,
                             AICDILambda2,
                             AICDILambda3)

set.seed(2023)
bmdMAAICCurveDILambda1 <- bmdMA(modelList[AICDILambda1], modelWeights = "AIC",
                                bmr = bmr, backgType = "modelBased", def = def,
                                type = "curve", bootstrapType = "nonparametric",
                                bootInterval = "percentile", R = bootR, level = level,
                                progressInfo = FALSE)
```

```
##   BMD_MA  BMDL_MA
##  66.8196 44.08511
```

```r
set.seed(2023)
bmdMAAICCurveDILambda2 <- bmdMA(modelList[AICDILambda2], modelWeights = "AIC",
                                bmr = bmr, backgType = "modelBased", def = def,
                                type = "curve", bootstrapType = "nonparametric",
                                bootInterval = "percentile", R = bootR, level = level,
                                progressInfo = FALSE)
```

```
##    BMD_MA  BMDL_MA
##  66.38595 44.08293
```

```r
set.seed(2023)
bmdMAAICCurveDILambda3 <- bmdMA(modelList[AICDILambda3], modelWeights = "AIC",
                                bmr = bmr, backgType = "modelBased", def = def,
                                type = "curve", bootstrapType = "nonparametric",
                                bootInterval = "percentile", R = bootR, level = level,
                                progressInfo = FALSE)
```

```
##    BMD_MA  BMDL_MA
##  64.69567 44.04452
```

```r
# Results ----------------------------------------------------------------
methodNames <- c("min(AIC)", "MA_AIC(post)",
                 "MA_AIC(curve)", "MA_Stack(post)", "MA_Stack(curve)",
                 "MA_AIC_DI_lambda1", "MA_AIC_DI_lambda2", "MA_AIC_DI_lambda3")
bmdList <- list(bmdAICMin, bmdMAAICPost, bmdMAAICCurve,
                bmdMAStackPost, bmdMAStackCurve,
                bmdMAAICCurveDILambda1,
                bmdMAAICCurveDILambda2,
                bmdMAAICCurveDILambda3)
MABmdEst <- sapply(bmdList, function(x) x$Results[1])
names(MABmdEst) <- methodNames
MABmdLower <- sapply(bmdList, function(x) x$Results[2])
names(MABmdLower) <- methodNames

appRes <- list(modelFcts = modelFcts,
               methodNames = methodNames,
               lambda = lambda,
               gamma = gamma,
               AICWeights = AICWeights,
               StackWeights = StackWeights,
               MABmdEst = MABmdEst,
               MABmdLower = MABmdLower,
               AICDILambda1 = AICDILambda1,
               AICDILambda2 = AICDILambda2,
               AICDILambda3 = AICDILambda3)
```

Full mixture data example

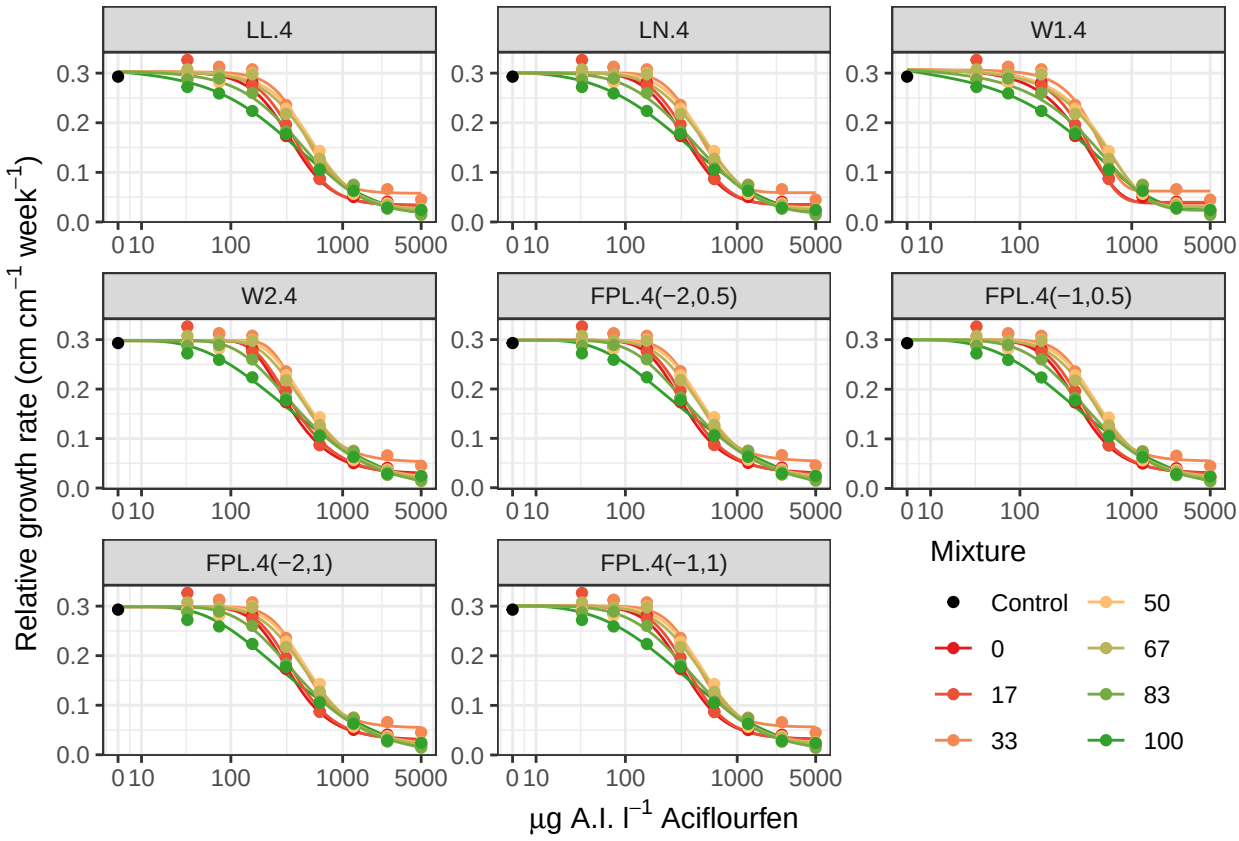
Fit models

```
# Analysis of full mixture experiment ------------------------------------
acidiq <- transform(acidiq, pct = factor(pct))

# Fit models -------------------------------------------------------------
modelList2 <- lapply(modelFcts,
                     function(modelString){
                       if(!substr(modelString, 1,3) == "FPL"){
                         drm(rgr ~ dose, data = acidiq, curveid = pct,
                             pmodels = list(~ pct, ~ pct, ~ 1, ~ pct),
                             fct = eval(parse(text = modelString)),type = "continuous")
                       } else {
                         drm(rgr ~ dose, data = acidiq, curveid = pct,
                             pmodels = list(~ pct, ~ pct, ~ 1, ~ pct),
                             fct = eval(parse(text = modelString)),
                             start = c(rep(-50,7), rep(0,7), 0.3, rep(3,7)),
                             type = "continuous")
                       }
                     })
```

```
## Control measurements detected for level: 999
## Control measurements detected for level: 999
## Control measurements detected for level: 999
## Control measurements detected for level: 999
## Control measurements detected for level: 999
## Control measurements detected for level: 999
## Control measurements detected for level: 999
## Control measurements detected for level: 999
```

```
xx <- c(0,exp(seq(log(0.8), log(5000), length.out = 300)))

app2_curves <- function(x,model){
  df <- as.data.frame(model$curve[[1]](x))
  colnames(df) <- c("Model",levels(acidiq$pct)[c(7:1)])
  df$Model <- model$fct$name
  df$x <- x
  df_long <- df %>% pivot_longer(cols = all_of(levels(acidiq$pct)[c(7:1)]),
                                 names_to = "pct")
  df_long
}

curve_values <- do.call(rbind, lapply(modelList2, function(mod) app2_curves(xx, mod))) %>%
  mutate(Model = factor(Model, levels = sapply(modelList2, function(x) x$fct$name)))

colorRamp <- colorRampPalette(colours[c(7,8,5)])(7) # colours used for mixture curves

app2_plot <- ggplot(acidiq) +
  geom_line(aes(x = x, y = value, col = pct), data = curve_values) +
  stat_summary(aes(x = dose, y = rgr, col = pct), fun = mean, geom = "point") +
```

```
  facet_wrap(~Model, scales = "free", nrow = 3) +
  scale_color_manual(breaks = levels(acidiq$pct)[c(8,1:7)],
                     values = c(colours[1], colorRamp),
                     labels = c("Control", levels(acidiq$pct)[1:7])) +
  scale_x_continuous(transform = scales::pseudo_log_trans(sigma = 10),
                     breaks = c(0,10,100,1000,5000)) +
  labs(x = TeX("$\\mu g$ A.I. $l^{-1}$ Aciflourfen"),
       y = TeX("Relative growth rate (cm $cm^{-1}$ $week^{-1}$)"),
       color = "Mixture") +
  theme_bw() +
  theme(legend.position = "bottom") +
  guides(col = guide_legend(nrow = 4)) + theme(legend.position = c(0.85,0.15))
app2_plot
```



BMD estimation

```
# BMD Estimation ------------------------------------------------------------
# minimum AIC value
whichAICMin2 <- which.min(sapply(modelList2, function(model) AIC(model)))

set.seed(2023)
bmdAICMin2 <- bmdBoot(modelList2[[whichAICMin2]], bmr = bmr, backgType = "modelBased",
                      def = def, bootType = "nonparametric",
                      bootInterval = "percentile", R = bootR, level = level)
```

```
##           BMD      BMDL
## 100  69.42293  50.51751
## 83  132.59369 110.35202
## 67  208.71374 191.37198
## 50  235.62305 214.75666
## 33  252.35091 231.57478
## 17  180.19215 164.68192
## 0   171.52023 143.94582
```

```
# Compute weights
AICWeights2 <- exp(-(sapply(modelList2,AIC)-min(sapply(modelList2,AIC))))/
  sum(exp(-(sapply(modelList2,AIC)-min(sapply(modelList2,AIC)))))
AICWeights2
```

```
## [1] 7.835780e-07 4.695745e-07 2.773048e-22 9.973555e-01 1.593816e-03
## [6] 5.871684e-04 3.646200e-04 9.761762e-05
```

```
set.seed(2023)
StackWeights2 <- getStackingWeights(modelList2, nSplits = 3)
StackWeights2
```

```
## [1] 3.757920e-23 0.000000e+00 1.317828e-01 8.215317e-01 5.078453e-22
## [6] 0.000000e+00 4.668556e-02 2.583136e-22
```

```
# bmdMA
set.seed(2023)
bmdMAAICPost2 <- bmdMA(modelList2, modelWeights = "AIC", bmr = bmr,
                       backgType = "modelBased", def = def, type = "bootstrap",
                       bootstrapType = "nonparametric",
                       bootInterval = "percentile", R = bootR, level = level,
                       progressInfo = FALSE)
```

```
##       BMD_MA   BMDL_MA
## 100  69.4318  49.49182
## 83  132.5781 104.78683
## 67  208.6855 178.16338
## 50  235.5963 195.91563
## 33  252.3281 222.67002
## 17  180.1846 160.93537
## 0   171.4948 124.51468
```

```
set.seed(2023)
bmdMAAICCurve2 <- bmdMA(modelList2, modelWeights = "AIC", bmr = bmr,
                        backgType = "modelBased", def = def, type = "curve",
                        bootstrapType = "nonparametric",
                        bootInterval = "percentile", R = bootR, level = level,
                        progressInfo = FALSE)
```

```
##        BMD_MA   BMDL_MA
## 100  69.43216  49.49229
## 83  132.57950 104.79378
## 67  208.68933 178.54598
## 50  235.60069 196.00994
## 33  252.33251 222.97545
## 17  180.18591 161.02768
## 0   171.49895 124.51844
```

```
set.seed(2023)
bmdMAStackPost2 <- bmdMA(modelList2, modelWeights = "Stack", bmr = bmr,
                         backgType = "modelBased", def = def, type = "bootstrap",
                         bootstrapType = "nonparametric",
                         bootInterval = "percentile", R = bootR, level = level,
                         stackingSplits = 3, progressInfo = FALSE)
```

```
## Error in optim(startVec, opfct, hessian = TRUE, method = optMethod, control = list(maxit = maxIt,  :
##   non-finite finite-difference value [22]
```

```
##        BMD_MA   BMDL_MA
## 100  66.13601  49.05942
## 83  125.25277 101.03697
## 67  200.19203 174.74233
## 50  226.14878 195.75679
## 33  247.97640 217.74103
## 17  176.43558 161.45547
## 0   164.09396 121.49850
```

```
set.seed(2023)
bmdMAStackCurve2 <- bmdMA(modelList2, modelWeights = "Stack", bmr = bmr,
                          backgType = "modelBased", def = def, type = "curve",
                          bootstrapType = "nonparametric",
                          bootInterval = "percentile", R = bootR, level = level,
                          stackingSplits = 3, progressInfo = FALSE)
```

```
## Error in optim(startVec, opfct, hessian = TRUE, method = optMethod, control = list(maxit = maxIt,  :
##   non-finite finite-difference value [22]
```

```
##        BMD_MA   BMDL_MA
## 100  66.71447  49.29788
## 83  126.83799 102.42069
## 67  202.48457 176.88400
## 50  228.79535 199.24606
## 33  249.32155 218.30074
## 17  177.53925 161.94172
## 0   166.24587 121.36773
```

Results

```r
# All results ------------------------------------------------------------------
methodNames2 <- c("min(AIC)", "MA_AIC(post)","MA_AIC(curve)",
                  "MA_Stack(post)", "MA_Stack(curve)")
bmdList2 <- list(bmdAICMin2, bmdMAAICPost2, bmdMAAICCurve2,
                 bmdMAStackPost2, bmdMAStackCurve2)
names(bmdList2) <- methodNames2


appRes2 <- list(modelFcts = modelFcts,
                methodNames2 = methodNames2,
                AICWeights2 = AICWeights2,
                StackWeights2 = StackWeights2,
                bmdList2 = bmdList2)
```