Comparison of Model Averaging and Model Selection in Dose Finding Trials Analyzed by Nonlinear Mixed Effect Models

Simon Buatois, Sebastian Ueckert, Nicolas Frey, Sylvie Retout, France Mentré

Comparison of model averaging & model selection in dose finding trials analyzed by nonlinear mixed effect models.

Simon Buatois[1,2,3*], Sebastian Ueckert[4], Nicolas Frey[1], Sylvie Retout[1,2], France Mentré[3]

(1) Roche Pharma Research and Early Development, Clinical Pharmacology, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland

(2) INSTITUT ROCHE, 30, cours de l'île Seguin, 92650 Boulogne-Billancourt, France

(3) IAME, UMR 1137, INSERM, University Paris Diderot, Sorbonne Paris Cité, Paris, France

(4) Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

*Corresponding author: 16 rue henri huchard 75018 paris, France ;

simon.buatois@inserm.fr ; tel : +3378489218

Abstract

In drug development, pharmacometric approaches consist in identifying via a model selection (MS) process the model structure that best describes the data. However, making predictions using a selected model ignores model structure uncertainty, which could impair predictive performance. To overcome this drawback, model averaging (MA)takes into account the uncertainty across a set of candidate models by weighting them as a function of an information criterion. Our primary objective was to use clinical trial simulations (CTSs) to compare model selection (MS) with model averaging (MA) in dose-finding clinical trials, based on the AIC information criterion. A secondary aim of this analysis was to challenge the use of AIC by comparing MA and MS using 5 different information criteria. CTSs were based on a nonlinear mixed effects model characterizing the time course of visual acuity in wet age-related macular degeneration patients. Predictive performances of the modeling approaches were evaluated using 3 performance criteria focused on the main objectives of a phase II clinical trial. In this framework, MA adequately described the data and showed better predictive performance than MS, increasing the likelihood of accurately characterizing the

dose-response relationship and defining the minimum effective dose. Moreover, regardless of the modeling approach, AIC was associated with the best predictive performances.

Introduction

Finding the right dose remains a critical step in clinical drug development (1). Selecting too high a dose increases the risk of toxicity, while too low a dose may reduce the treatment efficacy. Uncertainty concerning the selected dose can lead to unsuccessful trials and delays in regulatory approval. Between 2000 and 2012, one of the greatest causes of failure of phase 3 submissions was uncertainty related to dose selection (2).

To tackle this challenge, 150 delegates from industry, academia and regulatory bodies representing different scientific disciplines attended a dose finding workshop under the leadership of the European Medicines Agency (EMA) (3).

Among the different discussions, the workshop reiterated the following statement of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E4 guidance (4): “dose finding should rely on model-based estimation rather than hypothesis testing via pairwise comparisons”. Hence, there is , an increased interest in innovative approaches to accurately characterization of the dose-response relationship (5). Approaches based on models, such as nonlinear models, provide a functional relationship between dose and response. Compared to a pairwise analysis, nonlinear models allow analysis of all the data simultaneously and interpolation between doses (6).

Based on recommendations from health authorities, the model should be specified prior to data analysis.. However, before phase 2, little is known regarding the dose response relationship. The Multiple Comparison Procedure – Modeling (MCP-MOD) method (7,8) addressed this issue by using a predefined set of candidate models for the description of the dose response relationship. Once the evidence of a drug effect is established at the MCP step

using multiple contrast tests, a MOD step is used to estimate the dose to be brought into the confirmatory phase.

Traditionally, model based approaches require selection of the model that best describes the data according to an information criterion. The model is then used to predict the dose response relationship. Nonetheless, making predictions with a selected model (MS) ignores uncertainty that could impair predictive performance, as recognized in the literature (9).

More recently, the method of model averaging (MA) has been proposed to take model uncertainty into account (10,11). MA associates a weight with each of the candidate models based on a chosen information criterion. Of information criteria proposed (12–14), most depend on the log-likelihood (*LL*) as well as a penalty term (*pen)* which varies depending on the selected criterion. MA predictions are then obtained from a weighted mixture of the candidate model predictions. This method could be applied alone or in combination with MCP-MOD. MA has recently been shown to provide consistently better predictive performances than MS in the context of nonlinear models for dose finding studies (15).

Compared to nonlinear models (NLMs), nonlinear mixed effects models (NLMEMs) allow longitudinal analysis of data and so leverage all the information provided by clinical trials (16). This feature is particularly meaningful in clinical trials with long-term end points where the disease status of a given patient evolves significantly over time. Repeated measures are used to identify the natural history of the disease, using a disease progression model, as well as the potential impact of a drug, distinguishing disease modification from symptomatic treatment effects (17,18).

MA was recently applied to NLMEMs by Aoki *et al* for selection of the minimum effective dose in the context of an asthma drug. This work shows that MA and bootstrap MS have better predictive properties than MS (19), (19),

The present article extends this work both in terms of results and methodology, by using a different simulation case study. As part of these extensions, we investigated different study designs and performance criteria. Moreover, the candidate models include a disease progression as well as inter-individual variability in the dose response relationships.

The focus of our study was comparison, through clinical trial simulations (CTS), of the predictive performances of MA and MS on a predefined set of NLMEMs with the same disease progression model and different dose-effect relationships. Knowing that there is no real guidance on which information criterion is preferable, our secondary objective was to evaluate 5 criteria and identify the best one for MA and MS. First, the predictive performances of the analysis methods were compared in 3 simulation scenarios using the Akaike information criterion (AIC). Predictive performances were evaluated using 3 performance criteria focusing on the main objectives of a phase 2 clinical trial, i.e. the ability to correctly identify (i) a clinically relevant effect, (ii) the target dose and (iii) the dose-response relationship. Second, using the last performance criterion, the predictive performances of MS and MA were evaluated using 5 different information criteria.

Materials and methods

Model-based data analysis

Nonlinear mixed effects models

Let $y_{i,j}$ be the observation of subject $i = 1, \dots, N$ at time $j = 1, \dots, n$ and Y the vector of all observations of size n_{tot}.

This work considered a set of candidate NLMEMs called $m = 1, \dots, M$ of the form:

$$y_{m,i,j} = f_m\left(d_i, t_j, \Phi_{m,i}\right) + \varepsilon_{m,i,j}$$

Where, d_i is the dose administered to the subject i, $\Phi_{m,i}$ the vector of individual parameters and $\varepsilon_{m,i,j}$ the residual error. The vector of individual parameters depends on fixed effects μ_m and random effects $\eta_{m,i} \sim \mathcal{N}(0, \Omega_m)$. For the sake of simplicity, a diagonal variance covariance

matrix Ω_m was assumed. Parameters that include a random effect were considered to have either a log-normal $\Phi_{m,i} = \mu_m \cdot e^{\eta_{m,i}}$ or a normal $\Phi_{m,i} = \mu_m + \eta_{m,i}$ distribution. Residual errors were assumed to be independent and normally distributed $\varepsilon_{m,i,j} \sim \mathcal{N}(0, \sigma^2{}_m)$.

Finally, for each model m, Ψ_m is defined as the vector of the population parameters of size p_m, i.e. μ_m, Ω_m and $\sigma^2{}_m$.

Modeling approaches

Starting from a set of candidate models, different modeling approaches can be used to estimate the dose-response relationship. One can: (a) use a given candidate model regardless of its properties to describe the data; (b) select the candidate model that describe the data the best using MS; or (c) compute a weighted mixture of the candidate models using MA. When focusing on MS and MA, both approaches rely on an information criterion I.

Information criteria

Different information criteria can be used in NLMEM (13,14). Here, the list of investigated information criteria is derived from the work of Bertrand *et al* (12) and consists of the Akaike information criterion (AIC), the consistent Akaike information criterion (CAIC) and the Bayesian information criterion (BIC). All investigated criteria balance the log-likelihood (LL) with a penalty term (pen) which varies depending on the selected criterion I:

$$I_m = -2LL\left(Y, \widehat{\Psi}_m\right) + 2pen_{I,m}$$

The penalty terms are based on the parsimony principle, for similar information, the simplest of competing models is to be preferred. Penalty terms depend upon the number of estimated parameters as well as the sample size for BIC and CAIC. The definition of sample size, however, is not straightforward in mixed effect models and refers either to the number of subjects or to the total number of observations (20) leading to 5 information criteria: AIC, $BIC_{N,}$ BIC_{ntot}, $CAIC_N$ and $CAIC_{ntot.}$The penalty terms are defined as:

$$\text{pen}_{AIC,m} = p_m$$

$$\text{pen}_{BIC_{N,m}} = 0.5 \cdot p_m \cdot \log(\text{N})$$

$$\text{pen}_{BIC_{ntot,m}} = 0.5 \cdot p_m \cdot \log(n_{tot})$$

$$\text{pen}_{CAIC_{N,m}} = 0.5 \cdot p_m \cdot (\log(\text{N}) + 1)$$

$$\text{pen}_{CAIC_{ntot,m}} = 0.5 \cdot p_m \cdot (\log(n_{tot}) + 1)$$

Model selection:

The best model (m_{min}) was defined as the one with the lowest I_m value among the M candidate models.

Model averaging estimator:

MA corresponds to a mixture of the M candidate models weighted by their I_m values. Weights (w_m) were calculated as (9,21):

$$w_m = \frac{e^{\frac{-\Delta I_m}{2}}}{\sum_{m'=1}^{M} e^{\frac{-\Delta I_{m'}}{2}}}$$

Where,

$$\Delta I_m = I_m - I_{m_{min}}$$

ΔI_m is used instead of I_m to avoid numerical problems when calculating the exponential of I_m.

Simulation & Estimation

Clinical study:

The design is inspired by a study of a monoclonal antibody indicated in the treatment of wet age related macular degeneration (wet-AMD) (22). Wet-AMD, is a chronic eye disease associated with abnormal blood vessels that grow underneath the retina, damage the macula and may result in blurred vision or loss of vision in the center of the visual field. The treatment reduces vessel leakiness and improves visual acuity (VA) by neutralizing the vascular endothelial growth factor (VEGF) in the retina.

CTS were performed in the scenario of a hypothetical phase 2 trial inspired by the clinical development of a monoclonal antibody indicated in the treatment of wet-AMD. The trial duration was set to 24 months with a total of 26 VA measurements per patient (at baseline, day 7 and every 28 days). The study size was set to 300 patients who were equally randomized to receive either the placebo or 1 out of 3 doses.

Model:

In wet-AMD, disease progression and treatment effect are assessed using a VA test based on the early treatment diabetic retinopathy study (ETDRS) chart (22) which contains 14 lines (70 letters). The ETDRS chart is used to measure the number of letters successfully read by the patient at a given visit. The simulation model describes the time course of VA in wet-AMD patients, with or without anti-VEGF treatment (Diack C. *et al*, PAGE 2015, https://www.page-meeting.org/?abstract=3569).

$$f(d_i, t_j, \Phi_i) = VA_{0,i} + \left(1 - e^{-k_{pr,i} \cdot t_j}\right) \cdot \left(\frac{emax_i \cdot d_i}{ED_{50} + d_i} - \beta_i \cdot VA_{0,i}\right)$$

VA_0 is the visual acuity at baseline and in the absence of treatment. VA exponentially decreases over time t_j in untreated patients, reaching an asymptote $VA_{0,i} \cdot (1 - \beta_i)$ at a rate k_{pr}. Based on k_{pr}, one can derive the average time to steady state $(5 \cdot \log(2)/k_{pr})$ which equals approximately 24 months and corresponds to the end of trial. In line with the literature, the model mimic a mixture of both a disease modifying and symptomatic effect. The dose response relationship was assumed to be an Emax function, where, $emax_i$ represents the maximum number of letters that an individual can gain and ED_{50} the dose at which 50% of the maximal effect is achieved. For a very high dose and at steady state, the predicted visual acuity is $VA_{0,i} \cdot (1 - \beta_i) + emax_i$.

Interindividual variability was assumed for the parameters VA_0, k_{pr}, β and *emax*. The *emax* parameter follows a normal distribution and a log normal distribution was assumed for the

remaining ones. The parameter values (Ψ^*) used to simulate the datasets are reported in Table I.

Simulation scenarios

Clinical trial simulations were used to compare the predictive performances of MA and MS in different scenarios. In total, three simulation scenarios were investigated assuming (I) an informative design with doses around the ED50*, (II) only small doses that lay in the linear part of the dose-response curve, or (III) no drug effect (Figure 1):

- In scenario I, MA and MS are compared using a hypothetical informative design where emax equals 30 letters and doses are equal to 0, 150, 300 and 500 µg.
- In scenario II, the investigated doses are in the linear part of the dose response curve i.e. 0, 25, 50 and 100 µg; therefore little is known regarding the maximal effect of the drug (30 letters).
- In scenario III, the investigated doses are 0, 25, 50 and 100 µg with a flat dose response relationship (i.e. no drug effect).

Estimation:

For each scenario, S datasets were simulated and for each simulated dataset *s*, parameters ($\widehat{\Psi}_{m,s}$) of the M candidate models were estimated by maximizing the log likelihood using the Monte Carlo importance sampling expectation maximization method (IMP).

Candidate models:

The candidate models resulted from the combination of the disease progression model with one of four dose effect relationships (8,15,19,23):

- Emax, $$E(d) = \frac{emax_i \cdot d}{ED_{50} + d} \quad (1)$$
- Linear, $$E(d) = \alpha_i \cdot d \quad (2)$$
- Log-linear, $$E(d) = \alpha_i \cdot \log(d + 1) \quad (3)$$

- Sigmoid, $$E(d) = \frac{emax_i \cdot d^{\gamma}}{{ED_{50}}^{\gamma} + d^{\gamma}} \quad (4)$$

Where, α_i represents the slope of the Linear and Log-linear equations and γ is the Hill coefficient of the sigmoidal equation. The hypothesis of a flat dose response relationship was included in each candidate model by assuming a normal distribution for the $emax_i$ and α_i parameters.

Model predictions

In wet-AMD, the primary endpoint is the individual visual acuity change from baseline (ΔVA) to the end of trial (EOT). Therefore, Monte Carlo simulations were used to compute the distribution of the individuals ΔVA at 24 months for each dose.

These distributions depend upon the scenario, the dose, the trial replicate, the modeling approach and the information criterion. For the sake of clarity, the scenario, trial replicate and information criterion subscripts were ignored in the following equations.

The true probability density of the change from baseline at a given dose d, $p^*(d, \Delta VA)$, is calculated as follows:

$$p^*(d, \Delta VA) = p\big(f(d\,, t_{EOT}, \Phi) - f(d\,, 0, \Phi)\big), \Phi \sim p(\Psi^*)$$

In total $a = 1, \ldots, 6$ modeling approaches were compared. From $a = 1$ to 4, the distribution corresponds to the probability density function of the corresponding candidate model m. Approach 5 corresponds to model selection and approach 6 to model averaging.

The estimated probability density changes from baseline at a given dose d for each of the candidate models, $p_m(d, \Delta VA)$, were calculated as follows:

$$p_m(d, \Delta VA) = p\big(f_m(d\,, t_{EOT}, \Phi) - f_m(d, 0, \Phi)\big), \Phi \sim p(\widehat{\Psi}_m)$$

If $a = 5$, $p_a(d, \Delta VA)$ corresponds to the estimated probability density of the best candidate model $p_{m_{min}}(d, \Delta VA)$; and if $a = 6$, $p_a(d, \Delta VA)$ corresponds to the estimated probability density of the mixture of the M candidate models weighted by w_m:

$$p_a(d, \Delta_{VA}) = \sum_m w_m \cdot p_m(d, \Delta VA)$$

Evaluation of predictive performance

Using AIC as the information criterion, predictive performances of the $a = 1, \ldots, 6$ modeling approaches were evaluated using three criteria based on the clinically relevant drug effect (CRE), the minimum effective dose (MED) and the Kullback–Leibler divergence (D_{KL}). Secondly, MS and MA D_{KL} values were compared using 5 different information criteria. In line with the model prediction section, the scenario, trial replicate and information criterion subscripts were ignored in the following equations.

Clinically relevant drug effect

The **clinically relevant drug effect** (CRE) was defined as an increase of the median ΔVA at EOT of at least 15 letters at dose 500 µg, compared to patients treated by placebo, such as:

$$1_{CRE_a}(\Delta VA) \coloneqq \begin{cases} 1 \; if \; \{median(p_a(500, \Delta VA)) - median(p_a(0, \Delta VA) \geq 15\} \\ 0 \; if \; \{median(p_a(500, \Delta VA)) - median(p_a(0, \Delta VA) < 15\} \end{cases}$$

Replicates were then used to calculate the percentage of trials indicating a CRE for each simulation scenario and modeling approach.

Minimum effective dose

The MED corresponds to the minimum dose at which a CRE is achieved. Derived from the probability density, one can predict the true and estimated MED, respectively:

$$MED^* = argmin_d\{median(p^*(d, \Delta VA)) - median(p^*(0, \Delta VA)) \geq 15\}$$

$$MED_a = argmin_{d,a}\{median(p_a(d, \Delta VA)) - median(p_a(0, \Delta VA)) \geq 15\}$$

Where, d range from 0 to 500 μg in steps of 50 μg. For a given trial replicate, if none of the simulated doses led to a CRE, then MED_a was capped at 500 μg.

In the first and second simulation scenarios, the relative root mean squared error (RRMSE) and the relative bias between the true and estimated MED were used to compare the precision and accuracy of the different approaches.

Kullback–Leibler divergence

The Kullback–Leibler divergence (D_{KL}) or relative entropy represents the divergence between two probability distributions (24). The relative entropy satisfies Gibbs' inequality, $D_{KL}(p^*|p) \geq 0$. Thus, a Kullback–Leibler divergence between two identical distributions equals zero.

In this study, the Kullback–Leibler divergence between the true and the estimated probability density was calculated according to (25):

$$D_{KL_a}(d, p^*, p) = \int p^*(d, \Delta VA) \cdot \ log \frac{p^*(d, \Delta VA)}{p_a(d, \Delta VA)}$$

The $Total\, D_{KL}$ represents here the divergence over the doses d^k, $k = 1, \ldots, 4$, in $\{0, 150, 300, 500\}\ \mu g$:

$$Total\, D_{KL_a}(p^*, p) = \sum_{k=1}^{4} D_{KL_a}\,(d^k)$$

The distribution of the $Total\, D_{KL}$ values was then used to compare the predictive performances of the modeling approaches.

Technical implementation

In each scenario, $S = 500$ trial replicates were simulated. For each simulated dataset and each candidate model, $\hat{\Psi}_{s,m}$ were estimated in NONMEM7.3 (26) using IMP with the option AUTO=1, which allows the best settings to be determined. The CTYPE option was, however, over-ridden and set to 0 in order to let the process go through the full set of iterations (NITER=500). Then 10000 Monte Carlo simulations were used to compute the distribution of the true and the predicted visual acuity changed from baseline at 24 months.

Results

Model selection and model averaging

Figure 2 represents the selected proportion and the distribution of estimated weights per candidate model as a function of the simulation scenario using AIC as the information criterion. In scenario I, where doses are around the ED_{50} and *emax* is set to 30 letters,the proportion and weights are notably higher for the true (Emax) candidate model. However, from scenario II, by investigating doses only below ED_{50}, both the Emax and Linear models are likely to be selected in 50.8% and 39.6% of cases, respectively. Using MA, the higher weights are for the Emax and Linear models with a median of 0.41 for the former and 0.29 for the latter. When little is known regarding the maximal effect of the drug, the Emax and Linear models are almost equally likely.

The last simulation scenario explores the case of a flat dose-response relationship. With these settings, the proportion and weights are higher for the candidate models with the lowest number of estimated parameters i.e. Linear and Log-linear.

Evaluation

Clinically relevant drug effect:

Table II reports the percentage of trials indicating a CRE as a function of the scenario and the modeling approaches. Based on the true model and the population parameters Ψ^*, the percentage of trials indicating a CRE is 100% for scenario I and II and 0% for scenario III.

In the first and third scenarios, in most of the trial replicates, the modeling approaches correctly concluded that there is a CRE for the former and no CRE for the latter.

Regarding the second scenario, the highest percentages of trials indicating a CRE are achieved by the Linear and MA approaches, with 100% and 89%, respectively. Comparatively, using the true (Emax) candidate model or MS, the percentage is below 82%.

Minimum effective dose:

The ability of the 6 modeling approaches to predict the correct target dose, MED^*, is assessed via the RRMSE, the relative bias (Table III) and a boxplot representation of the estimated MED (Figure 3). Derived from the true model and the population parameters $\Psi^* MED^*$ is equal to 250 μg in the first and second scenario and capped at 500 μg in the third scenario.

When focusing on the first scenario, apart from Linear and Log-linear candidate models, all modeling approaches provide a precise and accurate prediction of the target dose. In the second scenario, the lower and upper quartiles are closer to MED^* using MA leading to a lower RRMSE and a smaller bias compared to the other modeling approaches. When comparing MS and MA, the RRMSE drops from 53.4% to 45.0% and the relative bias from 12.2 to 6.4%. The Linear candidate model leads to a precise but biased prediction of the target dose.

Finally, in the absence of a drug effect (scenario III), the MED is capped at 500 μg for almost all clinical trials regardless of the modeling approach.

Kullback–Leibler divergence:

Comparison of the different modeling approaches

The Kullback-Leibler divergence is first used to compare predictive performances of the different modeling approaches over a set of 4 doses. Figure 4 is a boxplot representation of the total Kullback-Leibler divergence for the three different scenarios. Sigmoid and Emax models provide equivalent or better predictive performances than the other modeling approaches inon all scenarios. The mean total Kullback-Leibler divergence is reduced, up to 50%, when using MA compared to MS.

Comparison of the different information criteria

MS and MA predictive performances were compared using the 5 different information criteria (Figure 5). In scenario II, regardless of the modeling approach, the distribution of the total Kullback-Leibler divergence is closer to 0 when using AIC compared to the other investigated

criteria. Thus, AIC provide better predictive performances than the other information criteria. When comparing the total Kullback-Leibler divergence of MS and MA as a function of the information criteria, MA has a consistently better predictive performance than MS.

In simulation scenarios I and III, the 5 information criteria provide similar predictive performances.

Discussion

The primary objective of this study was to use clinical trial simulations (CTS),to compare the use of model selection (MS) versus model averaging (MA) in dose-finding clinical trials. CTS were based on a disease model characterizing the time course of visual acuity in wet age-related macular degeneration patients, and an Emax dose-response relationship. Different scenarios were investigated assuming either (I) an informative design with doses around the ED50*, (II) only doses in the linear part of the dose response curve, and (III) no drug effect. Then, for each simulated trial, parameters of four candidate models (Emax, Sigmoid, Log-linear and Linear) were estimated. Finally, using AIC as the information criterion, predictive performances of MS and MA modeling approaches were evaluated through three performance criteria focusing on the main objectives of a phase 2 clinical trial, i.e. the ability to correctly identify (i) a clinically relevant effect (CRE), (ii) the minimum effective dose (MED) and (iii) the dose-response relationship using the total Kullback-Leibler divergence ($total\ \mathrm{D_{KL}}$). Results highlight that, under the scenario I and III, MS and MA provided similar predictive performances and led to an accurate and unbiased prediction of the true dose-response relationship. In the case where the investigated doses are below the ED50 (scenario II) and therefore when little is known regarding the maximal effect of the drug, MA was leading to better predictive performances than MS. MA was associated with (i) a higher percentage of trials indicating a CRE, (ii) a lower RRMSE with a smaller bias, and (iii) lower $total\ \mathrm{D_{KL}}$ values compared to MS. A common misconception in MS and MA is to think that the goal is

to identify the true structural model. In scenario II, where the true (Emax) model cannot be supported by the trial data, bias with Emax is similar to bias with Linear, and the RRMSE is even lower for Linear. This explains why MA performs better than MS even if the median weight attributed to emax (0.41) is lower than the proportion of emax model (50.8%).A secondary aim of this analysis was to challenge the use of AIC by comparing MA and MS $total$ D_{KL} values using 5 different information criteria. In this framework, regardless of the modeling approach, AIC was associated with lower $total$ D_{KL}than $BIC_{N,}$ BIC_{ntot}, $CAIC_N$ and $CAIC_{ntot}$. Moreover, regardless of the information criterion, MA was consistently associated with lower $total$ D_{KL}values than MS.

Our results based on NLMEMs are in accordance with the literature. In the field of NLM, Schorning *et al* (15) have shown that MA outperforms MS in dose-finding trials. In NLMEMs, Aoki *et al* (19) highlighted that the MA method significantly decreases the effect of MS bias. Both studies show that AIC outperforms other information criteria. In addition to these two studies, we the present article explored the impact of the set of investigated doses on the predictive performances of MS and MA. Moreover, the predefined set of candidate models includes a disease progression model and an inter-individual variability within the dose response relationship. Finally, we extended the information criterion comparison from 2 to 5. This comparison is based on the $total$ D_{KL} which uses the entire probability distribution and is therefore more informative than other performance criteria.

In our analysis, uncertainties around the estimated parameters were ignored, thus excluding the possibility of a clinically significant effect. In the context of NLMEM it can be extremely challenging to include parameter uncertainty. The normality assumption on parameter uncertainty distribution might not hold and, due to the study design and number of trial replicates, bootstrap/SIR become prohibitively costly in computation. However, one could argue that the present results are relevant in pharmacometry where simulations are often

performed without uncertainty. Predictions were used to compare the percentages of trials indicating a clinically relevant effect at the highest simulated doses. All candidate models may conclude to a flat dose response relationship, although another possibility could have been to include a model without a dose-response relationship as one of the candidate models.

For all scenarios, the Emax and Sigmoid model were associated with the lowest $total\ \mathrm{D_{KL}}$ values emphasizing the notion that they can be used for all practical purposes (27,28). However, one should note that the true candidate model corresponds to the former and is nested to the latter. In addition, knowing that uncertainties around the estimated parameters were ignored in this framework, about it may be asked whether the Emax and Sigmoid models over-fit the data.

In contrast to existing work, we explored the impact of the set of doses on the predictive performances of MA and MS. However, we did not quantify the impact of the number of patients, the study duration and the set of candidate models. The set of probable dose-response relationships was selected from the literature (8,15,19), but other dose-response relationships, such as the umbrella or exponential one, could have been used. Additional scenarios are reported in the online supplementary information, to investigate the case where the true model (umbrella) is not part of the set of candidate models. The results highlights that, when the true model cannot be approximated by the set of candidate models, MS and MA have similar predictive performances.

Regarding the information criteria comparison, results should be interpreted with caution as the comparison favors the AIC. In fact, all information criteria but one, the AIC, depend on the total number of patients, leading to over-penalization of complex models in the case, like here, of large study size. Moreover, the AIC is derived from the Kullback-Leibler divergence.

It would be beneficial to extend this work by including parameter uncertainty as well as investigating other dose response relationships. It would also be interesting to vary the

number of patients and study duration. Finally, we believe that inclusion of different disease progression structures in the set of candidate models should be studied in more depth in the future in order to capture the model uncertainty around both the dose response relationship and the disease progression structure.

We present here the interesting properties of MA compared to MS based on a simulation case study. Our results highlight that, in an informative design, MA and MS provided similar predictive performances and led to accurate prediction of the target dose. However, with less informative designs, by estimating weights for a predefined set of NLMEMs, MA showed better overall predictive performances than MS increasing the likelihood of accurately characterizing the dose-response relationship.

Conclusion

Dose-finding clinical trials data can be more efficiently analyzed by combining NLME models and model averaging. This leverages the information provided by clinical trials through a longitudinal analysis of the data while taking into account model uncertainty over a predefined set of candidate models.

Acknowledgements

This work was financed by a CIFRE agreement (Conventions Industrielles de Formation par la Recherche) and was conducted under the supervision of the ANRT (Association Nationale Recherche Technologie). The CIFRE agreement is a partnership between a public laboratory and a company, here the UMR 1137 and INSTITUT ROCHE, respectively.

References

1. Cross J, Lee H, Westelinck A, Nelson J, Grudzinskas C, Peck C. Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980–1999. Pharmacoepidemiol Drug Saf. 2002 Sep 1;11(6):439–46.

2. Sacks LV, Shamsuddin HH, Yasinskaya YI, Bouri K, Lanthier ML, Sherman RE. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000-2012. JAMA. 2014 Jan 22;311(4):378–84.

3. Musuamba FT, Manolis E, Holford N, Cheung S, Friberg LE, Ogungbenro K, et al. Advanced methods for dose and regimen finding during drug development: summary of the EMA/EFPIA workshop on dose finding (London 4-5 December 2014). CPT Pharmacomet Syst Pharmacol. 2017 Jul;6(7):418–29.

4. Dose-response information to support drug registration (ICH Harmonized Tripartite Guideline), Page 2. 1994 [cited 2017 Jun 14]. Available from: http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/dose-response-information-to-support-drug-registration.html

5. Bornkamp B, Bretz F, Dmitrienko A, Enas G, Gaydos B, Hsu C-H, et al. Innovative approaches for designing and analyzing adaptive dose-ranging trials. J Biopharm Stat. 2007 Nov 8;17(6):965–95.

6. Karlsson KE, Vong C, Bergstrand M, Jonsson EN, Karlsson MO. Comparisons of analysis methods for proof-of-concept trials. CPT Pharmacomet Syst Pharmacol. 2013 Jan;2(1):e23.

7. Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics. 2005 Sep;61(3):738–48.

8. Pinheiro J, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. Stat Med. 2014 May 10;33(10):1646–61.

9. Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. Biometrics. 1997;53(2):603–18.

10. Sébastien B, Hoffman D, Rigaux C, Pellissier F, Msihid J. Model averaging inconcentration–QT analyses. Pharm Stat. 2016 Nov 1;15(6):450–8.

11. Dosne AG, Bergstrand M, Karlsson MO, Renard D, Heimann G. Model averaging for robust assessment of QT prolongation by concentration-response analysis. Stat Med. 2017 Oct 30;36(24):3844–57.

12. Bertrand J, Comets E, Mentre F. Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters. J Biopharm Stat. 2008;18(6):1084–102.

13. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987 Sep 1;52(3):345–70.

14. Anderson DR, Burnham KP. Understanding information criteria for selection among capture-recapture or ring recovery models. Bird Study. 1999 Jan 1;46(sup1):S14–21.

15. Schorning K, Bornkamp B, Bretz F, Dette H. Model selection versus model averaging in dose finding studies. Stat Med. 2016 Sep 30;35(22):4021–40.

16. Bates DM. Nonlinear mixed effects models for longitudinal data. In: Wiley StatsRef: Statistics Reference Online [Internet]. John Wiley & Sons, Ltd; 2014. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat05806/abstract

17. Holford N. Clinical pharmacology = disease progression + drug action. Br J Clin Pharmacol. 2015 Jan 1;79(1):18–27.

18. Buatois S, Retout S, Frey N, Ueckert S. Item response theory as an efficient tool to describe a heterogeneous clinical rating scale in de novo idiopathic parkinson's disease patients. Pharm Res. 2017 Oct 1;34(10):2109–18.

19. Aoki Y, Röshammar D, Hamrén B, Hooker AC. Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection. J Pharmacokinet Pharmacodyn. 2017 Nov 4;1–17.

20. Delattre M, Lavielle M, Poursat M-A. A note on BIC in mixed-effects models. Electron J Stat. 2014;8:456–475.

21. Claeskens G, Hjort NL. Model Selection and Model Averaging. 1 edition. Cambridge ; New York: Cambridge University Press; 2008. 332 p.

22. Rosenfeld PJ, Brown DM, Heier JS, Boyer DS, Kaiser PK, Chung CY, et al. Ranibizumab for neovascular age-related macular degeneration. N Engl J Med. 2006 Oct 5;355(14):1419–31.

23. Thomas N, Sweeney K, Somayaji V. Meta-analysis of clinical dose–response in a large drug development portfolio. Stat Biopharm Res. 2014 Oct 2;6(4):302–17.

24. Kullback S. Information theory and statistics. New edition edition. Mineola, N.Y: Dover Publications; 1997. 432 p.

25. MacKay DJC. Information theory, inference and learning algorithms. 1 edition. Cambridge, UK ; New York: Cambridge University Press; 2003. 640 p.

26. Beal S, Sheiner LB, Boeckmann A, Bauer RJ. NONMEM User's Guides. (1989–2009). Icon Development Solutions, Ellicott City, MD USA; 2009.

27. Thomas N. Hypothesis testing and bayesian estimation using a sigmoid Emax model applied to sparse dose-response designs. J Biopharm Stat. 2006;16(5):657–77.

28. Dragalin V, Hsuan F, Padmanabhan SK. Adaptive designs for dose-finding studies based on sigmoid Emax model. J Biopharm Stat. 2007;17(6):1051–70.

Legend to Figures

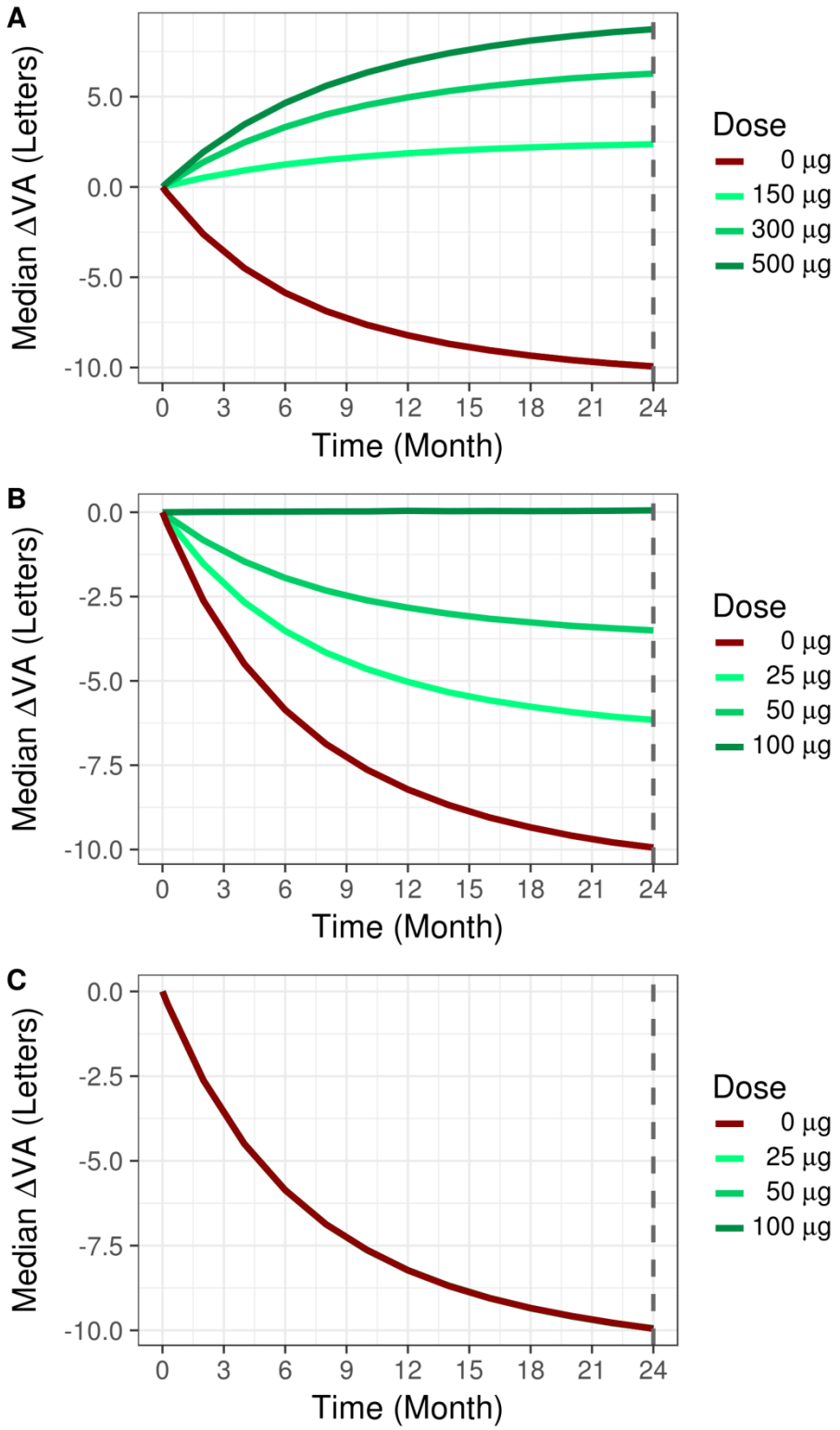

Figure 1: Representation of the simulated median visual acuity change from baseline in function of the time and per dose group. The dashed horizontal line represents the end of trial. Panels A, B and C correspond to the simulation scenarios I, II and III, respectively.

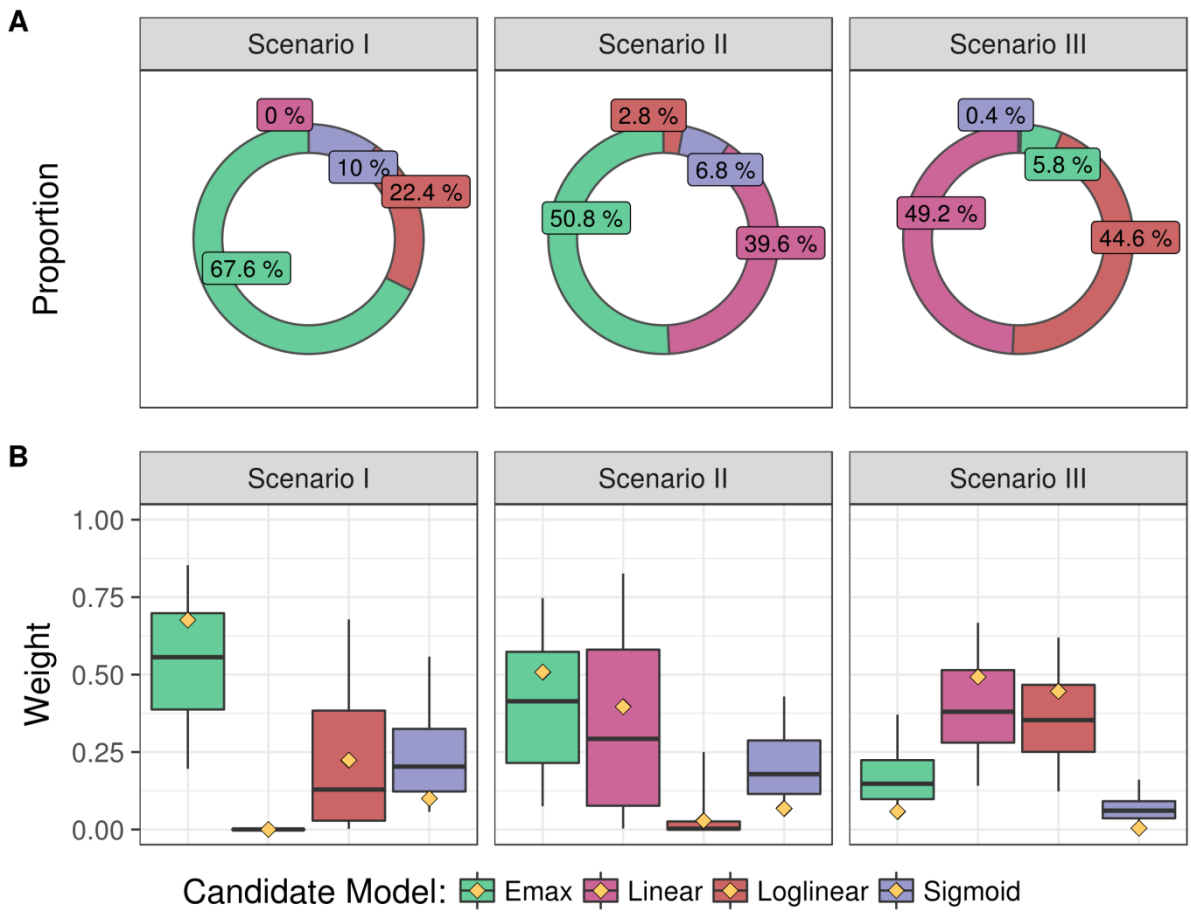

Figure 2: Representation of the selected proportions, MS (panel A), and distribution of the weights, MA (panel B), per candidate model and for each simulation scenario using AIC as the information criterion. Yellow diamonds represent the selected proportions using MS.

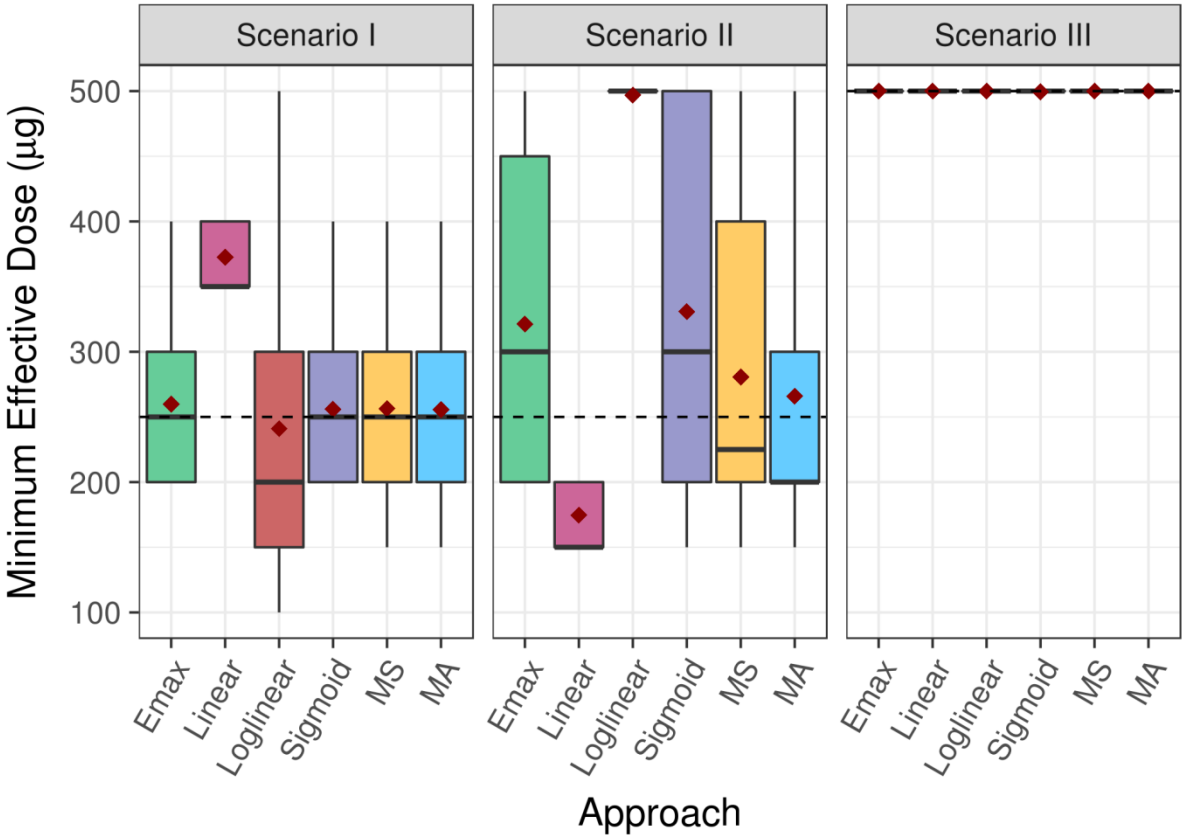

Figure 3: Representation of the distribution of the predicted minimum effective dose for each modeling approach and each simulation scenario using AIC as the information criterion. Red diamonds represent the mean values and the dashed line represents the predicted MED using the true model and the true population parameters.

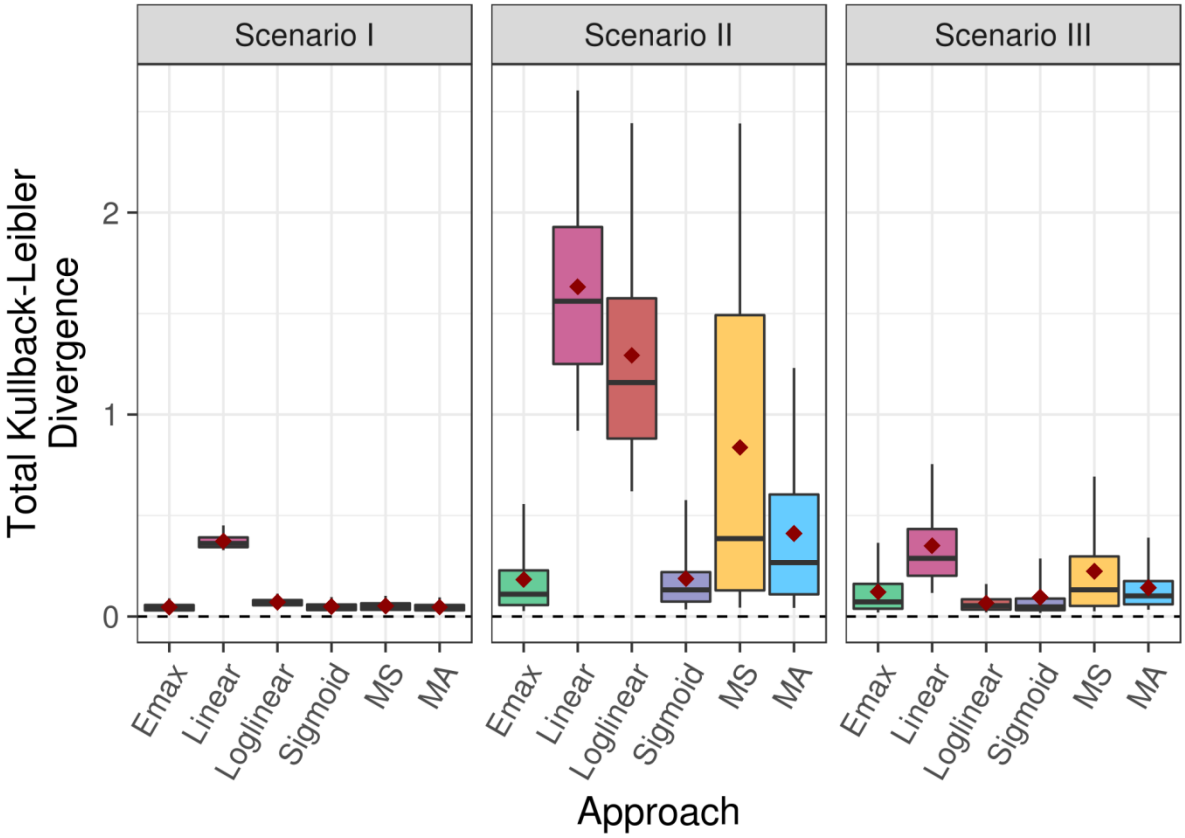

Figure 4: Representation of the distribution of the total Kullback-Leibler divergence for each modeling approach and each simulation scenario using AIC as the information criterion. The dashed line represents the total Kullback-Leibler divergence calculated using the true model and the true population parameters. Red diamonds represent the mean values.

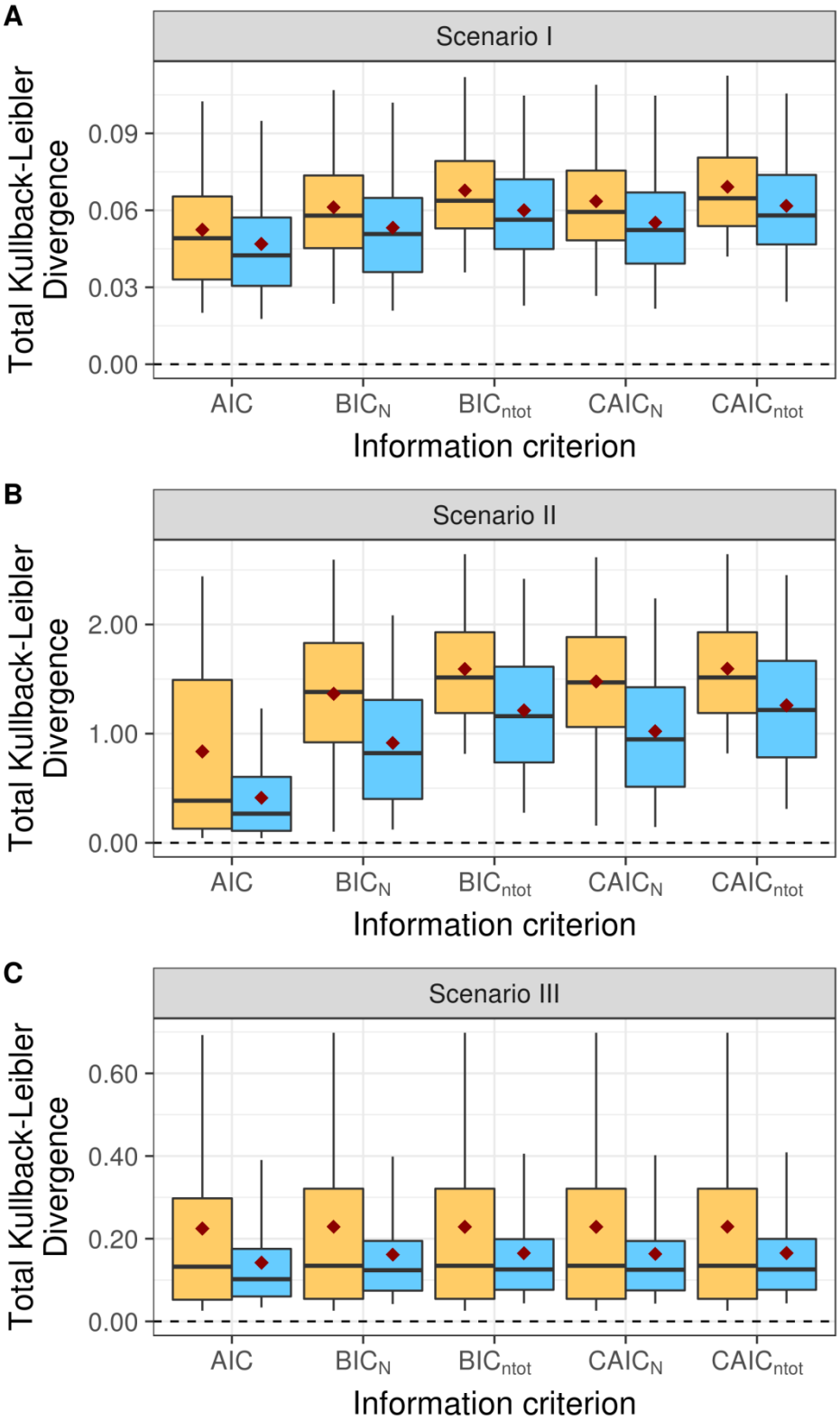

Figure 5: Representation of the distribution of the total Kullback-Leibler divergence using MS (in yellow) and MA (in blue) for each information criterion and in simulation scenario II. The dashed line represents the total Kullback-Leibler divergence calculated using the true model and the true population parameters. Panel A, B and C correspond to the simulation scenarios I, II and III, respectively. Red diamonds represent the mean values

Table I: Parameter values Ψ^* used to simulate the data assuming an Emax dose-response model.

Parameter	μ	ω^2	σ^2
$VA_0{}^*(let)$	55	0.07	-
$k_{pr}{}^*\ (Day^{-1})$	0.005	0.5	-
β^*	0.2	1.0	-
$emax^*\ (let)$	30	150	-
$ED_{50}{}^*\ (\mu g)$	150	-	-
$\sigma^{2*}(let)$	-	-	28

μ: Fixed effect
ω^2: Variance of the random effect
σ^2: Variance of the residual error
let: Letter

Table II: Percentage of trials indicating a clinically relevant effect at the dose of 500 μg for each modeling approach and each simulation scenario using AIC as the information criterion.

	(%) **of trials indicating a clinically relevant effect**		
	Scenario I	**Scenario II**	**Scenario III**
Simulation values	100	100	0
Emax	98.6	80.0	0.0
Linear	100.0	100.0	0.0
Log-linear	95.6	3.6	0.0
Sigmoid	98.8	76.0	0.4
Model selection	98.4	81.8	0.0
Model averaging	98.4	89.0	0.0

Table III: Relative bias and Relative root mean squared error (RRMSE) in the predicted minimal effective dose for each modeling approach and each simulation scenario using AIC as the information criterion.

Approach	**Relative bias (%)**		**RRMSE** (%)	
	Scenario I	**Scenario II**	**Scenario I**	**Scenario II**
Emax	4.0	28.5	27.3	54.6
Linear	49.0	-30.2	50.7	32.6
Log-linear	-3.6	98.8	42.7	99.1
Sigmoid	2.4	32.3	26.4	59.5
Model selection	2.6	12.2	30.0	53.4
Model averaging	2.3	6.4	28.8	45.0