J Pharmacokinet Pharmacodyn (2017) 44:581–597
DOI 10.1007/s10928-017-9550-0

ORIGINAL PAPER

Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection

Yasunori Aoki[1,2] · Daniel Röshammar[3,4] · Bengt Hamrén[3] · Andrew C. Hooker[1]

Received: 23 June 2017 / Accepted: 14 October 2017 / Published online: 4 November 2017

Abstract Population model-based (pharmacometric) approaches are widely used for the analyses of phase IIb clinical trial data to increase the accuracy of the dose selection for phase III clinical trials. On the other hand, if the analysis is based on one selected model, model selection bias can potentially spoil the accuracy of the dose selection process. In this paper, four methods that assume a number of pre-defined model structure candidates, for example a set of dose–response shape functions, and then combine or select those candidate models are introduced. The key hypothesis is that by combining both model structure uncertainty and model parameter uncertainty using these methodologies, we can make a more robust model based dose selection decision at the end of a phase IIb clinical trial. These methods are investigated using realistic simulation studies based on the study protocol of an actual phase IIb trial for an oral asthma drug candidate (AZD1981). Based on the simulation study, it is demonstrated that a bootstrap model selection method properly avoids model selection bias and in most cases increases the accuracy of the end of phase IIb decision. Thus, we recommend using this bootstrap model selection method when conducting population model-based decision-making at the end of phase IIb clinical trials.

Keywords Model averaging · Model selection · Pharmacometrics · Phase IIb clinical trial · Dose finding study · Mathematical modelling · Dose–effect relationship

Introduction and background

Quantifying the probability of achieving the targeted efficacy and safety response is crucial for go/no-go investment decision-making in a drug development program. This is particularly crucial when analyzing phase IIb (PhIIb) dose-finding studies to select the phase III dose(s) given the costs of phase III studies.

It has previously been shown that population model-based (pharmacometric) approaches can drastically increase the power to identify drug effects in clinical trial data analysis compared to conventional statistical analysis (e.g., [1]). On the other hand, the model-based approach can be hindered by model selection bias if a single model structure is assumed and used for the analysis (e.g., [2, 3]). There have been several attempts through model averaging and model selection to weaken the model structure assumptions by considering multiple possible model candidates in the analysis [4–9].

In this paper, we introduce four methods that assume a number of pre-defined model candidates and then combine or select those candidate models in different ways to make predictions and to account for uncertainty in those predictions. The first method is "simple" model selection where a set of model structures are pre-specified and a model is

Electronic supplementary material The online version of this article (doi:10.1007/s10928-017-9550-0) contains supplementary material, which is available to authorized users.

✉ Yasunori Aoki
yaoki@uwaterloo.ca

[1] Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

[2] Present Address: National Institute of Informatics, Tokyo, Japan

[3] Quantitative Clinical Pharmacology, Innovative Medicines and Early Development, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden

[4] Present Address: SGS Exprimo, Mechelen, Belgium

chosen according to a statistical criterion. Uncertainty in model prediction is then derived from parameter uncertainty, based on a bootstrap procedure using the selected model. The second method is a bootstrapped model selection procedure, where, for each bootstrap dataset, the best-fit of the candidate models is chosen according to a statistical criterion. Simulation from each bootstrap selected model with its optimal parameter will then generate a distribution of the quantities of interest, accounting for both model and parameter uncertainty (similar methods can be found in the literature, e.g., [11, 12]). The third method is a conventional model averaging procedure where each candidate model is fit to the data and uncertainty is quantified via bootstrap. Simulations (including parameter uncertainty) from each candidate model of the distributions of the quantities of interest are then combined as a weighted average depending on model fit to the original data. The fourth method is a bootstrapped model averaging procedure, where the weighting for the weighted average calculations are based on model fit to each bootstrapped dataset (as opposed to the model-fit to the original data).

Comparison of these methods and a standard statistical method (pair-wise ANOVA and the groupwise estimate of an average change from baseline) are done using clinical trial simulations of dose-finding studies. To make the simulations as realistic as possible, we have based them on the protocol of an actual PhIIb trial for an oral asthma drug candidate (AZD1981) as well as the data from the placebo arm of that trial. Drug effects using various model structures were simulated for five different dose arms (placebo plus four active arms). The different analysis methods were then used to calculate the probability of achieving target endpoint and then choose the minimum effective dose (MED).

Methods

Phase IIb dose-finding case study

Part of the PhIIb clinical trial data and the study protocol of the asthma drug candidate AZD1981 (ClinicalTrials.gov/NCT01197794) was utilized in this work. One endpoint goal of the study was to demonstrate that the drug improved the forced expiratory volume in 1 s (FEV1) of asthma patients by, on average, at least 0.1 L (placebo and baseline adjusted). This clinical trial was chosen as a case study since FEV1 is a highly variable endpoint (standard deviation of 0.3 L in the placebo effect) relative to the expected effect magnitude; hence it is hard to characterize the dose–effect relationship from PhIIb clinical trials.

This study was conducted for 12 weeks and FEV1 was measured every 2 weeks (for a total of 7 measurements, or visits). The first measurement was a screening visit and the second measurement was used as a baseline measurement after which either placebo, AZD1981 10, 20, 100 or 400 mg was administered twice daily (bid).

The data from the placebo group and the lowest dose group of the PhIIb clinical trial for AZD1981 was provided for this analysis. Dosing information was not provided; however, as there were no statistically significant differences between the placebo group and the lowest dose group as described in [13], in this paper we refer this dataset as a "placebo" dataset. This dataset comprises 324 patients with a total of 1803 FEV1 measurements.

Models

Placebo model

The following placebo model was developed using the placebo dataset from the PhIIb clinical trial for AZD1981:

$$
\begin{aligned}
\mathrm{FEV1} &= \left(\mathrm{FEV1}_{\mathrm{Baseline}} + \begin{cases} 0 & \text{if visit} = 1,2 \\ \mathrm{FEV1}_{\mathrm{Placebo}} & \text{if visit} = 3,4,5,6,7 \end{cases}\right) \\
&\quad \cdot (1+\epsilon_1) + \epsilon_2 \\
\mathrm{FEV1}_{\mathrm{Placebo}} &= \theta_1 + \eta_1 \\
\mathrm{FEV1}_{\mathrm{Baseline}} &= \theta_2 e^{\eta_2} \\
&\quad \times (1 + \theta_3(\mathrm{FEV1}_{\%\,\mathrm{of\,normal}} - \overline{\mathrm{FEV1}}_{\%\,\mathrm{of\,normal}})) \\
&\quad \times (1 + \theta_4(\mathrm{Age} - \overline{\mathrm{Age}})) \\
&\quad \times \begin{cases} 1 & \text{if Male} \\ \theta_5 & \text{if Female} \end{cases} \\
\eta_1 &\sim \mathcal{N}(0, \omega_1^2) \\
\eta_2 &\sim \mathcal{N}(0, \omega_2^2) \\
\epsilon_1 &\sim \mathcal{N}(0, \sigma_1^2) \\
\epsilon_2 &\sim \mathcal{N}(0, \sigma_2^2)
\end{aligned}
$$

where $\mathrm{FEV1}_{\%\,\mathrm{of\,normal}}$ is the percentage of FEV1 at visit 2 compared to the predicted normal and $\overline{\mathrm{FEV1}}_{\%\,\mathrm{of\,normal}}$ is its population mean, $\overline{\mathrm{Age}}$ is the mean of the age of the patients. All the estimated model parameters can be found in Table 1.

Previously Wang et al. [14] have modelled a placebo effect on the FEV1 measurement. The model presented here differs slightly from Wang et al. in that this model employs a step function for the placebo effect model with respect to visit, while Wang et al. have used exponential models with time as the independent variable. Wang et al. state that the placebo effect plateaus at 0.806 week^{-1} while the current dataset contains FEV1 measurements only every 2 weeks; hence the rate constant of the exponential model was not estimable from this dataset.

Drug effect models

In this work, we simulate and estimate using a number of different dose–effect relationships $\mathrm{DE_j}$:

Table 1 Estimated Parameters of the placebo model of FEV1 of Asthma patients based on the placebo and lowest dose group of the PhIIb clinical trial for AZD 1981

Model parameter	Description	Estimated value (RSE%)
θ_1	Placebo effect	0.169 L (11.9)
θ_2	Baseline	2.51 L (0.828)
θ_3	Covariate effect of $\text{FEV1}_{\text{\% of normal}}$	$0.0129\ \text{L}^{-1}$(3.87)
θ_4	Covariate effect of age	$-0.0105\ \text{year}^{-1}$ (5.01)
θ_5	Covariate effect of sex	0.719 (1.43)
ω_1	SD of IIV of the placebo effect	0.303 L (8.88)
ω_2	SD of IIV of the baseline	0.105 (5.97)
σ_1	SD of proportional RUV	0.0832 (7.60)
σ_2	SD of additive RUV	0.102 L (18.8)

IIV inter-individual variability, *RUV* residual unexplained variability, *SD* standard deviation, *RSE* the relative standard error was approximated using a variance–covariance matrix, and the computational result was verified using preconditioning [19]

$$\text{DE}_0 = 0 \text{ (no treatment effect)}$$
$$\text{DE}_1(\text{dose}; p_1) = p_1 \cdot \text{dose (linear model)}$$
$$\text{DE}_2(\text{dose}; p_1, p_2) = p_1 \cdot \log(1 + p_2\text{dose}) \text{ (log-linear model)}$$
$$\text{DE}_3(\text{dose}; \text{EMAX}, \text{EC50}) = \text{EMAX} \cdot \frac{\text{dose}}{\text{EC50} + \text{dose}} \text{ (Emax model)}$$
$$\text{DE}_4(\text{dose}; \text{EMAX}, \text{EC50}, \gamma) = \text{EMAX} \cdot \frac{\text{dose}^{\gamma}}{\text{EC50}^{\gamma} + \text{dose}^{\gamma}} \text{ (sigmoid Emax model)}$$

To create simulated datasets, we add different simulated drug effects, with different parameters, using the above models, to the FEV1 measurements of the placebo data (more detail below). For estimation using the model-based analysis methods described below, we embed these dose–effect relationships into the placebo model as follows:

$$\text{FEV1}_j = \left(\text{Baseline} + \text{PlaceboEffect} + \text{DE}_j\right) \cdot (1 + \epsilon_1) + \epsilon_2$$

Analysis methods

Statistical analysis used for the PhIIb clinical trial for AZD1981

The primary statistical analysis of the data from the PhIIb clinical trial for AZD1981 to determine the MED was performed using a pair-wise ANOVA and a group wise estimate of treatment effect. Briefly, the treatment effect was measured as the change from baseline (average of all available data from visits 3–7 minus baseline) per dose group. The MED was identified via a two-stage step-down procedure to select either 400, 100, 40, 10 mg or "stop" (do not proceed to phase III). The procedure was as follows: (1) starting with the highest (400 mg) dose-arm conduct a one-sided ANOVA comparison with the placebo-arm. (2) If the difference is significant (significance level of 0.05) check that the average treatment effect in the arm is greater than the primary efficacy variable (0.1 L). (3) If both steps 1 and 2 are satisfied then proceed to the next dose dose-arm (100 mg) and repeat, otherwise move to step 4. (4) Choose the lowest dose arm where both steps 1 and 2 are satisfied (Note that if 100 mg satisfies steps 1 and 2 but 40 mg does not then the MED will be 100 mg in this process, even if 10 mg might also satisfy steps 1 and 2).

Model selection and averaging analysis methods

Below is an overview of four methods that assume a number of pre-defined model candidates and then combine or select those candidate models in different ways to make predictions and to account for uncertainty in those predictions. For the given example, the methods were meant to compare with the standard determination of the MED, identified in the original study using the two-stage step-down procedure described earlier in this section. Thus, in the following methods, there should be a test for drug effect as well as a determination if that effect is greater than a given minimum effect size. In all methods, the test for drug effect is done using a likelihood ratio test (LRT) against the placebo model (5% significance level) [8]. Determination of effect sizes at specific doses is done by first computing the change from baseline average (population mean) effect size, and uncertainty around that effect size, predicted by the different methods described below, for a given dose. MED is then chosen as the lowest studied dose with a predefined probability to be above a target effect. For a full technical description of the methods, we refer the readers to the Appendix. The methods below are previously presented by the authors as a conference contribution [9], and Method 3 was presented at an earlier conference [10].

Method 1: model selection

1. Fit each candidate model structure to the original data and estimate the model parameters and maximum likelihood (see Fig. 1).
2. For each candidate model, perform a numerical identifiability test (see appendix for detail) and LRT against the placebo model and reject any model structure that fails either of these tests.
3. Select one model structure among the remaining model candidates using a statistical criterion based on the maximum likelihood (e.g., the Akaike information criterion, AIC, the Bayesian information criterion, BIC, etc.).
4. Quantify parameter uncertainty using case sampling bootstrap and the selected model structure.
5. Simulate the quantities of interest (with uncertainty); in this case, the dose-endpoint (change from baseline population mean effect size) relationships using the selected model structure and model parameters obtained from the bootstrap procedure.
6. Make a decision; in this case, choose the lowest dose (given allowed dose levels) that has a probability of achieving target endpoint greater than a predefined limit.

Method 2: bootstrap model selection

1. Create bootstrap datasets based on the original data using a case sampling bootstrap procedure (see Fig. 2).
2. For each bootstrap dataset estimate parameters and the maximum likelihood for each candidate model structure.
3. For each bootstrap dataset, and for each candidate model, perform a numerical identifiability test and LRT against the placebo model and reject any model structure that fails either of these tests.
4. For each bootstrap dataset, select one model structure among the remaining model candidates using a statistical criterion based on the maximum likelihood (e.g., AIC, BIC, etc.).
5. For each bootstrap dataset, simulate the quantities of interest; in this case, the dose-endpoint (change from baseline population mean effect size) relationships using the selected model structure and model parameters obtained from that bootstrap dataset.
6. Summarize the simulations; in this case, compute the probability of achieving the target endpoint at each dose of interest using the simulated dose-endpoint relationships.
7. Make a decision; in this case, choose the lowest dose (given allowed dose levels) that has a probability of achieving the target endpoint greater than a predefined limit.

Method 3: model averaging

1. Fit each candidate model structure to the original data and estimate model parameters and maximum likelihood (see Fig. 3).
2. For each candidate model, perform a numerical identifiability test and LRT against the placebo model and reject any model structure that fails either of these tests.

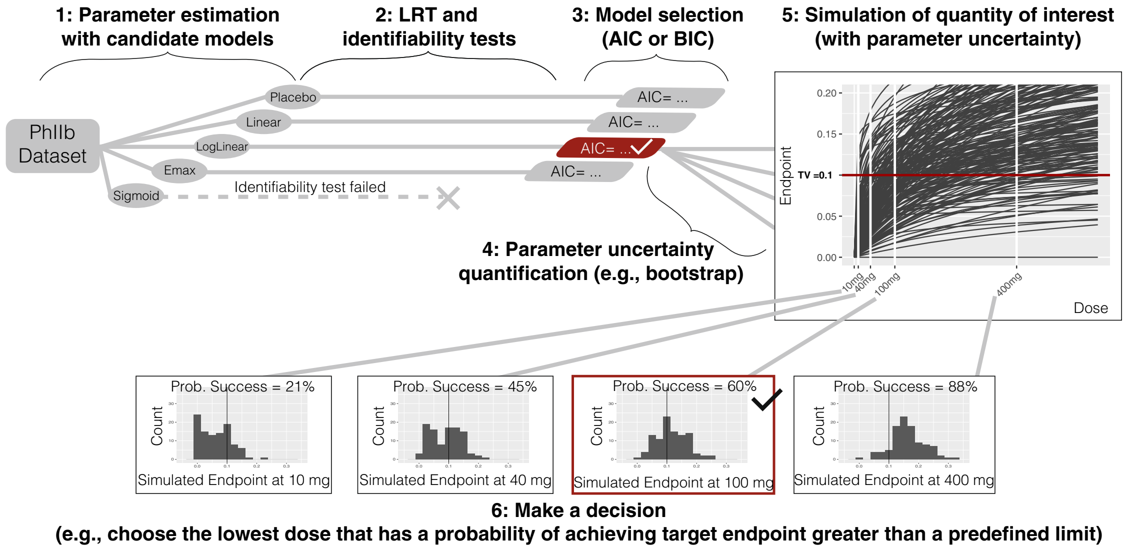
1: Parameter estimation with candidate models
2: LRT and identifiability tests
3: Model selection (AIC or BIC)
5: Simulation of quantity of interest (with parameter uncertainty)
PhIIb Dataset
Placebo
Linear
LogLinear
Emax
Sigmoid
Identifiability test failed
AIC= ...
AIC= ...
AIC= ...
AIC= ...
4: Parameter uncertainty quantification (e.g., bootstrap)
Endpoint
0.20
0.15
TV =0.1
0.05
0.00
10mg
40mg
100mg
400mg
Dose
Prob. Success = 21%
Count
Simulated Endpoint at 10 mg
Prob. Success = 45%
Count
Simulated Endpoint at 40 mg
Prob. Success = 60%
Count
Simulated Endpoint at 100 mg
Prob. Success = 88%
Count
Simulated Endpoint at 400 mg
6: Make a decision
(e.g., choose the lowest dose that has a probability of achieving target endpoint greater than a predefined limit)

Fig. 1 Method 1 model selection

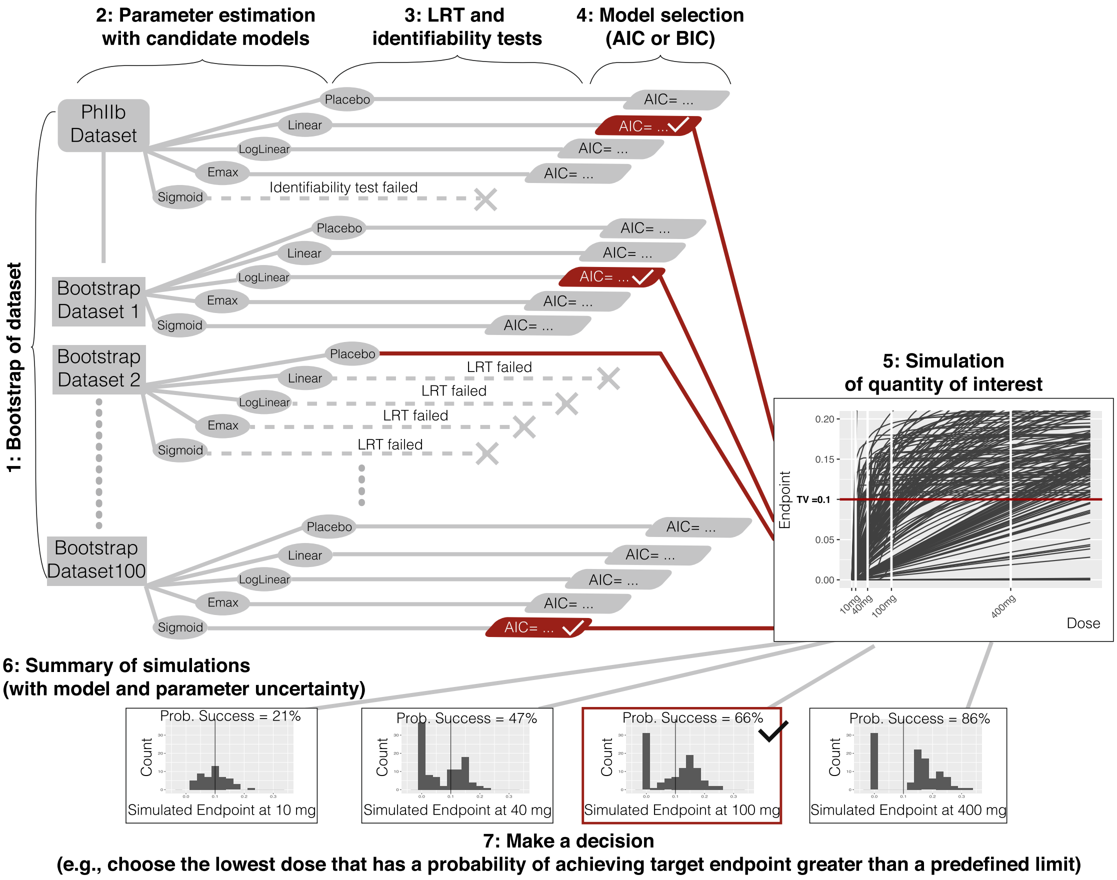
2: Parameter estimation with candidate models
3: LRT and identifiability tests
4: Model selection (AIC or BIC)
1: Bootstrap of dataset
PhIIb Dataset
Placebo
Linear
LogLinear
Emax
Sigmoid
Identifiability test failed
AIC= ...
Bootstrap Dataset 1
Bootstrap Dataset 2
LRT failed
Bootstrap Dataset100
5: Simulation of quantity of interest
Endpoint
TV =0.1
0.20
0.15
0.05
0.00
10mg
40mg
100mg
400mg
Dose
6: Summary of simulations (with model and parameter uncertainty)
Prob. Success = 21%
Prob. Success = 47%
Prob. Success = 66%
Prob. Success = 86%
Count
Simulated Endpoint at 10 mg
Simulated Endpoint at 40 mg
Simulated Endpoint at 100 mg
Simulated Endpoint at 400 mg
7: Make a decision
(e.g., choose the lowest dose that has a probability of achieving target endpoint greater than a predefined limit)

Fig. 2 Method 2 bootstrap model selection

3. For each model structure, quantify parameter uncertainty using case sampling bootstrap methodology to obtain the distribution of the model parameters.
4. For each model structure, simulate the quantities of interest; in this case, the dose-endpoint relationships using the model parameters obtained from the bootstrap method.
5. For each model structure, summarize the simulations; in this case, compute the probability of achieving the target endpoint at each dose of interest using the simulated dose-endpoint relationships.
6. Compute the weighted average of the summary variables obtained in step 5; in this case, the probability of achieving the target endpoint at each dose over the model structures, where the weights are derived from the maximum likelihood obtained in step 1 (e.g., AIC, BIC, etc.).
7. Make a decision; in this case, choose the lowest dose (given allowed dose levels) that has a probability of achieving target endpoint greater than a predefined limit.

Method 4: bootstrap model averaging

1. Create bootstrap datasets based on the original data using a case sampling bootstrap procedure (see Fig. 4).
2. For each bootstrap dataset estimate parameters and the maximum likelihood for each candidate model structure.
3. For each bootstrap dataset, and for each candidate model, perform a numerical identifiability test and LRT against the placebo model and reject any model structure that fails either of these tests.
4. For each bootstrap dataset and each model structure, simulate the quantities of interest; in this case, the dose-endpoint (change from baseline population mean effect size) relationships using the selected model structure and model parameters obtained from that bootstrap dataset.
5. Summarize the simulations; in this case, compute the weighted average of the probability of achieving the target endpoint at each dose using the dose-endpoint relationships for all the model structures and all the bootstrap datasets (except the ones that failed the LRT or identifiability test). The weights are derived from

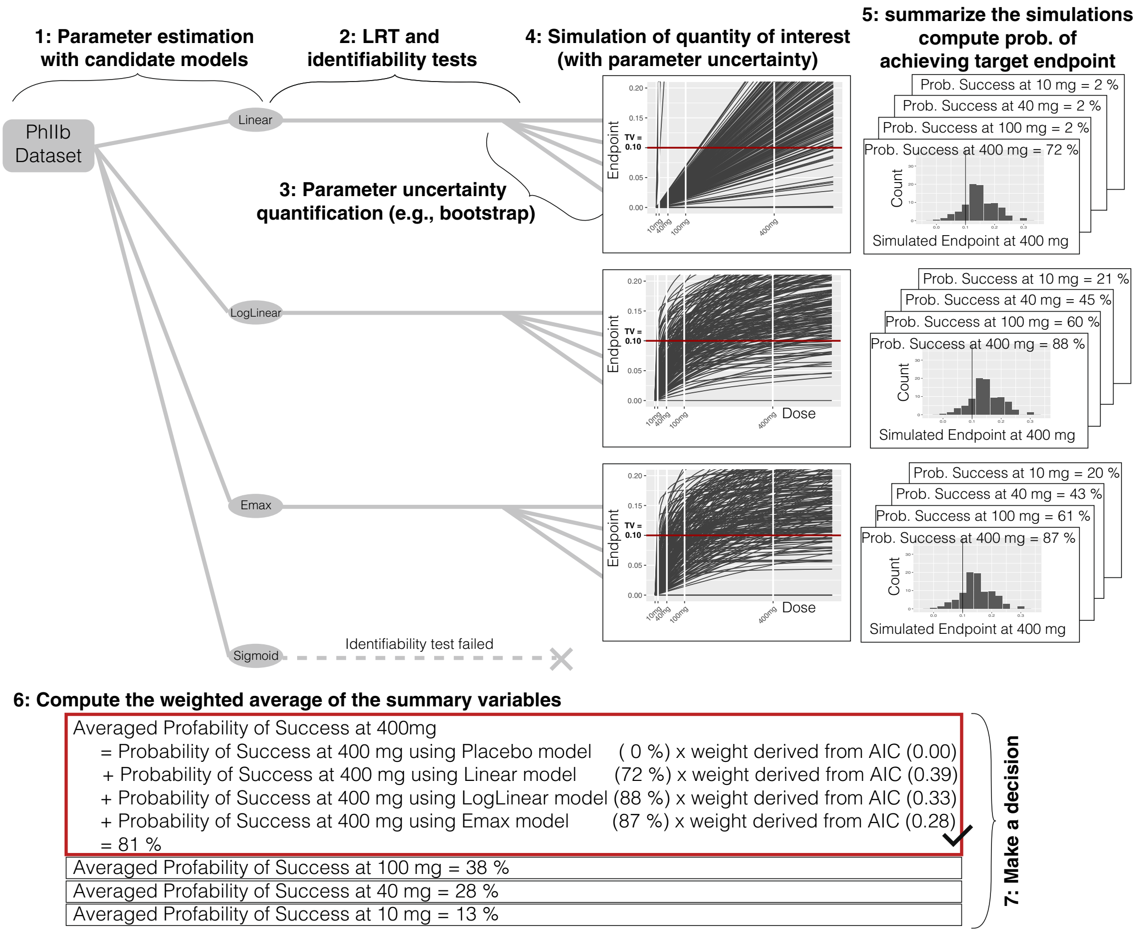
1: Parameter estimation with candidate models
2: LRT and identifiability tests
3: Parameter uncertainty quantification (e.g., bootstrap)
4: Simulation of quantity of interest (with parameter uncertainty)
5: summarize the simulations compute prob. of achieving target endpoint
PhIIb Dataset
Linear
LogLinear
Emax
Sigmoid
Identifiability test failed
Endpoint
TV = 0.10
Dose
10mg
40mg
100mg
400mg
Prob. Success at 10 mg = 2 %
Prob. Success at 40 mg = 2 %
Prob. Success at 100 mg = 2 %
Prob. Success at 400 mg = 72 %
Count
Simulated Endpoint at 400 mg
Prob. Success at 10 mg = 21 %
Prob. Success at 40 mg = 45 %
Prob. Success at 100 mg = 60 %
Prob. Success at 400 mg = 88 %
Prob. Success at 10 mg = 20 %
Prob. Success at 40 mg = 43 %
Prob. Success at 100 mg = 61 %
Prob. Success at 400 mg = 87 %
6: Compute the weighted average of the summary variables
Averaged Profability of Success at 400mg
= Probability of Success at 400 mg using Placebo model (0 %) x weight derived from AIC (0.00)
+ Probability of Success at 400 mg using Linear model (72 %) x weight derived from AIC (0.39)
+ Probability of Success at 400 mg using LogLinear model (88 %) x weight derived from AIC (0.33)
+ Probability of Success at 400 mg using Emax model (87 %) x weight derived from AIC (0.28)
= 81 %
Averaged Profability of Success at 100 mg = 38 %
Averaged Profability of Success at 40 mg = 28 %
Averaged Profability of Success at 10 mg = 13 %
7: Make a decision

Fig. 3 Method 3 model averaging

the maximum likelihood obtained in step 2 (using AIC, BIC, etc.).

6. Make a decision; in this case, choose the lowest dose (given allowed dose levels) that has a probability of achieving the target endpoint greater than a predefined limit.

Single model based approach

To compare the proposed methods against the idealized situation where the underlining true model structure is known before the analysis, we compare with a single model based approach where the model used to analyze the dataset is the same as the model used to simulate that dataset. Note that this single model based analysis using the simulation model is an idealistic scenario. In a real PhIIb dataset analysis (i.e., when analyzing data that was not simulated) it is not realistic to assume the exact underlying model structure is known a priori. The method has the following steps:

1. Perform LRT between the model with and without drug effect. If the model does not pass the LRT, make a “stop” decision.
2. If the model with drug effect passes the LRT, estimate the parameter uncertainty using a case sampling bootstrap.
3. Simulate the quantities of interest (with uncertainty); in this case, the dose-endpoint (change from baseline population mean effect size) relationships using the model parameter distribution obtained from the bootstrap procedure.
4. Make a decision; in this case, select the dose based on the required probability of achieving the target endpoint.

Numerical experiments

To test the proposed model averaging and selection methodologies, we have simulated dose-finding studies under various designs and experimental scenarios. All numerical computations were done using NONMEM [15]

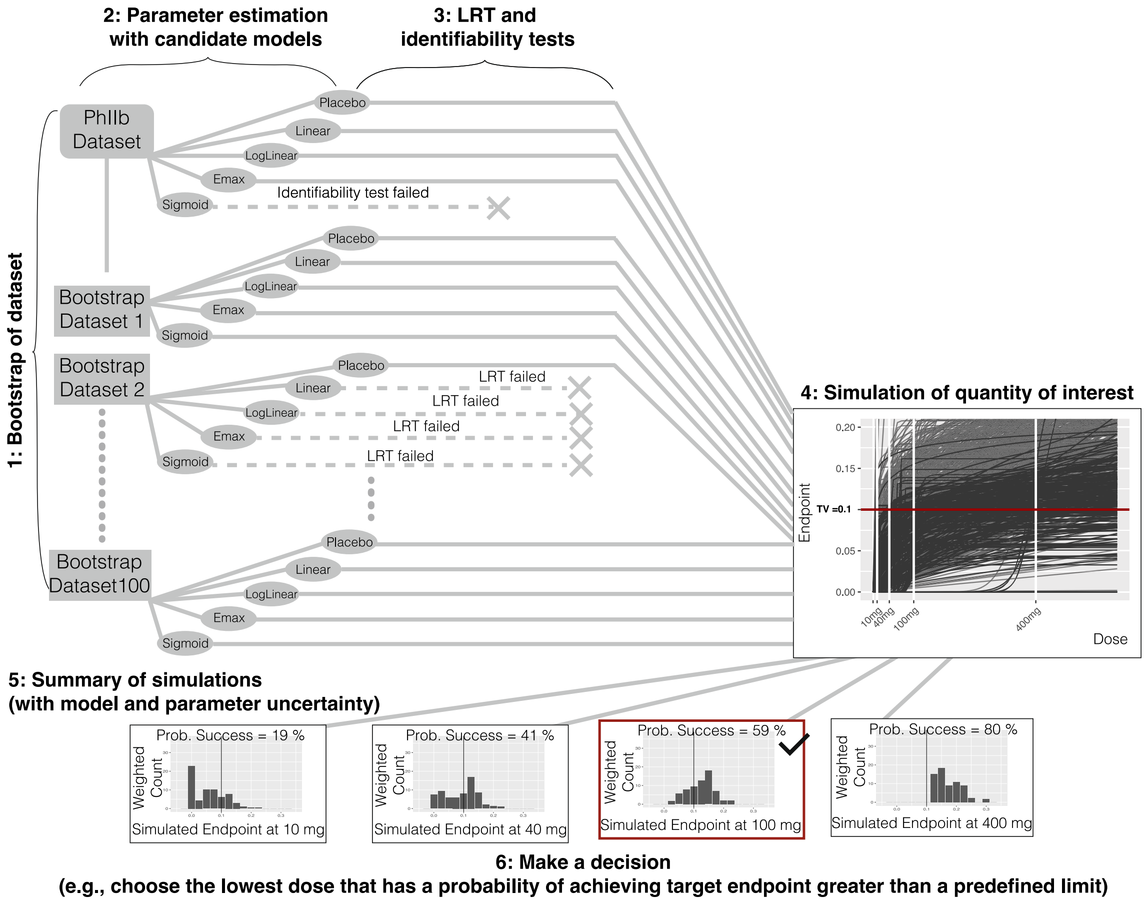
2: Parameter estimation with candidate models
3: LRT and identifiability tests
1: Bootstrap of dataset
PhIIb Dataset
Placebo
Linear
LogLinear
Emax
Sigmoid
Identifiability test failed
Bootstrap Dataset 1
Bootstrap Dataset 2
LRT failed
LRT failed
LRT failed
LRT failed
Bootstrap Dataset100
4: Simulation of quantity of interest
Endpoint
0.20
0.15
TV =0.1
0.05
0.00
10mg
40mg
100mg
400mg
Dose
5: Summary of simulations (with model and parameter uncertainty)
Prob. Success = 19 %
Prob. Success = 41 %
Prob. Success = 59 %
Prob. Success = 80 %
Weighted Count
Simulated Endpoint at 10 mg
Simulated Endpoint at 40 mg
Simulated Endpoint at 100 mg
Simulated Endpoint at 400 mg
6: Make a decision
(e.g., choose the lowest dose that has a probability of achieving target endpoint greater than a predefined limit)

Fig. 4 Method 4 bootstrap model averaging

version 7.3, PsN [16] version 4.6 on a Linux Cluster, with Intel Xeon E5645 2.4 GHz processors, 90 GB of memory, Scientific Linux release 6.5, GCC 4.4.7 and Perl version 5.10.1. To assure reproducibility of the numerical experiments we had a fixed random seed when the bootstrap method was performed using PsN. All computation outside of NONMEM and PsN was done using R version 3.2 [17] and all plots are made using ggplot2 [18].

Simulation studies based on placebo data

To create simulated datasets, we have simply added different simulated drug effects to the FEV1 measurements of the placebo data. We have randomly generated the artificial drug effect so that the theoretical minimum effective dose (tMED, i.e., the exact dose that achieves a drug effect of 0.1 L) is uniformly distributed in the ranges shown in Table 2.

For each Simulation Study 1–5, we have constructed 300 PhIIb clinical trial simulation datasets (1500 datasets in total). Simulation Studies 1–4 are constructed to test each analysis method for the accuracy of finding tMED, while Simulation Study 5 is constructed to test each method for the accuracy of Type-1 error control.

In each Simulation study, log-linear, emax and sigmoidal models (described above) were used to simulate the drug effects ($\mathrm{DE_j}$) (100 datasets each for each of three model structures, hence 300 total simulated datasets in one simulation study). For each data set, we first randomly choose tMED in the range shown in Table 2. Then the model parameters are chosen randomly as follows:

For the log-linear model, p_1 and p_2 are chosen so that

$$\mathrm{DE}_2(1000, p_1, p_2) \sim \mathrm{unif}(0.2, 0.3)$$
$$\mathrm{DE}_2(\mathrm{tMED}; p_1, p_2) = 0.1$$

For the emax model, EMAX and EC50 are chosen so that

$$\mathrm{EMAX} \sim \mathrm{unif}(0.2, 0.3)$$
$$\mathrm{DE}_3(\mathrm{tMED}; \mathrm{EMAX}, \mathrm{EC50}) = 0.1$$

For the sigmoidal model, EMAX, EC50 and γ are chosen so that

Table 2 Various scenarios of the simulation studies

	Theoretical minimum effective dose (tMED)	Correct dose finding/decision
Simulation Study 1	0–10 mg	10 mg
Simulation Study 2	10–40 mg	40 mg
Simulation Study 3	40–100 mg	100 mg
Simulation Study 4	100–400 mg	400 mg
Simulation Study 5	No drug effect	Stop

$$\mathrm{EMAX} \sim \mathrm{unif}(0.2, 0.3)$$
$$\gamma \sim \mathrm{unif}(0.5, 4)$$
$$\mathrm{DE}_4(\mathrm{tMED}; \mathrm{EMAX}, \mathrm{EC50}, N) = 0.1$$

Note that to determine the parameters p_1 and p_2 for the log-linear model, we need to solve a nonlinear equation numerically and we do so by using the *uniroot* function in R. As can be seen in Fig. 5, we can create diverse realistic simulated drug effects by the above choice of model parameters while the range of tMED is constrained.

Numerical experiment 1: dose finding accuracy

The simulated data from Simulation Studies 1–4 (when a drug effect is present) was analyzed using the methods presented above to determine the dose finding accuracy of the methods. Each method was used to find the MED for each trial simulation dataset and the probability of finding the correct dose was calculated (see Table 2).

For the model-based approaches, the MED dose was chosen as the minimum dose arm (of the investigated doses) with more than a 50% probability of achieving the target endpoint. 50% was chosen to match the statistical analysis used for the PhIIb clinical trial for AZD198, which evaluates if the average treatment effect in a dose arm is greater than the primary efficacy variable.

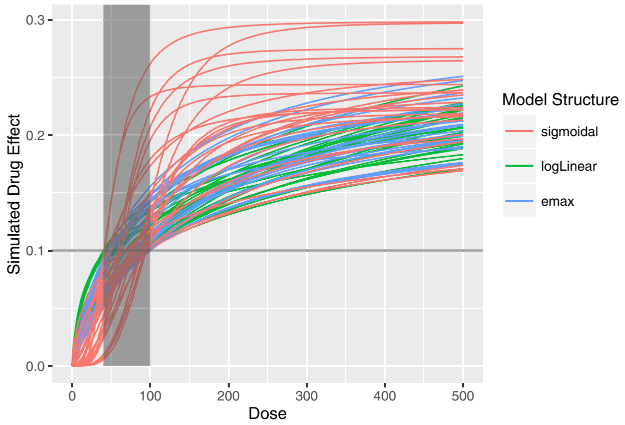

Fig. 5 Plot of (some of) the simulated drug effect for Simulation Study 3. The theoretical minimum effective dose (the exact dose that achieves the target endpoint of 0.1 L) ranges between 40 and 100 mg hence the 100 mg dose is the correct dose selection for this simulation study

Numerical experiment 2: type-1 error control accuracy

All methods presented above were used to determine the MED based on the data from Simulation Study 5 (the simulation study without simulated drug effect) to test the type-I error rate of the proposed methods. That is, the probability of choosing the MED to be either 10, 40, 100, or 400 mg while there is no simulated drug effect. The MED selection using the model-based approaches were determined at a 50% confidence level to fairly compare the method with the pairwise ANOVA method.

Numerical experiment 3: decision-making accuracy

In the previous two numerical experiments the MED using the model-based approaches are determined at a 50% confidence level to fairly compare the method with the pairwise ANOVA method. However, in reality, more than 50% certainty may be desired when making a decision about which dose to use in a phase III trial [20]. For example, from an investment perspective, it may be more crucial to reduce the risk of proceeding to a phase III trial with insufficient effect than to determine the exact MED of a drug.

For this experiment, we define the "correct" decision to be when any dose higher than the theoretical MED is selected. For example, for Simulation Study 3 ($40\ \mathrm{mg} < \mathrm{tMED} \leq 100\ \mathrm{mg}$), if either 100 or 400 mg is chosen then the correct decision was made; while if dose 10 or 40 mg is chosen, or a "stop" decision is made, then the incorrect decision was made. Each method was then used to find the MED (70% confidence level for the model-based approaches) for each simulated dataset from Simulation Studies 1–4 (when a drug effect is present). The probability of each method making the correct decision was then calculated.

Numerical experiment 4: probability of achieving target endpoint estimation accuracy

In the model averaging and selection methods investigated here, the dose selection is based on the probability of achieving the target endpoint, hence, accurate estimation of this probability is crucial. In this experiment, we investigate this probability estimation for each simulated dataset from Simulation Studies 1–4 (when a drug effect is present) in the following manner:

1. Select a predefined limit, p, for the probability of achieving the target effect.
2. Allow any dose (any positive real number) to be selected (not just the investigated dose levels) and choose the dose that is estimated to achieve the target endpoint with probability p using the proposed model-based methods.
3. Repeat steps 1 and 2 for all 1200 simulated phase IIb datasets and count the number of times a dose above the theoretical minimum effective dose (tMED) is selected, from which the empirical probability of achieving the target effect is calculated.
4. Repeat steps 1–3 for $p = 0.01, 0.02, \ldots, 0.99$.

Note that if a method estimates the probability of achieving the target endpoint without bias, then the selected doses should be above tMED with probability p.

Results

To concisely present the results for each of Numerical Experiments 1, 3, and 4, we has combined the results of Simulation Studies 1–4. Hence, for those experiments, the results are based on 1200 PhIIb clinical trial simulations. We refer the readers to the Appendix for a detailed discussion of the result for each simulation study. Further, the uncertainty of the numerical experiments has been quantified by randomly sampling trial simulations with replacement (1200 trial simulations for Numerical Experiments 1, 3, and 4, and 300 trial simulations for Numerical Experiment 2) and repeated the numerical experiments. For example, for Numerical Experiments 1, 3, and 4, 1200 trial simulations were sampled with replacement 100 times to produce 100 sets of the 1200 trial simulations. For each set of trial simulations, the numerical experiments were performed.

Numerical experiment 1: dose finding accuracy

The dose finding accuracy of the various investigated methods is presented in Fig. 6. As can be seen, all the model based methods could find the correct dose more often than the statistical method used in the PhIIb AZD1981 study protocol. In addition, we can see that Methods 2 and 4 outperform Methods 1 and 3 and the Single Model Based approach (using the simulation model).

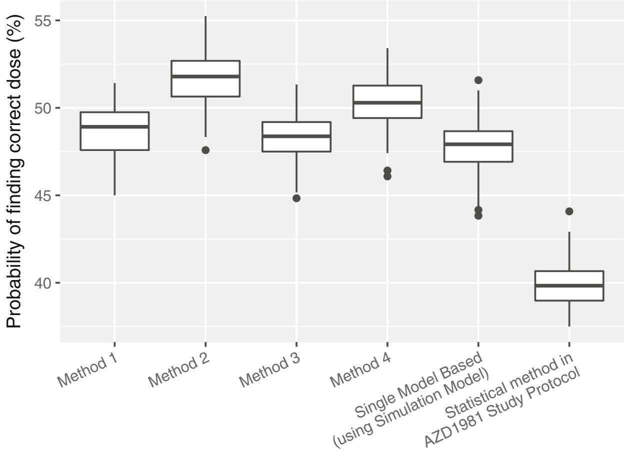

Fig. 6 Probability of finding the correct dose. The edges of the boxes are 75th and 25th percentiles. The line in the box is the median and the whiskers extend to the largest and the smallest value within 1.5*inter-quartile range. Dots are the outliers outside of the whiskers

Numerical experiment 2: type-1 error control accuracy

The Type-I error control of the various investigated methods is presented in Fig. 7. As can be seen, Methods 1–4 control the type-I error accurately. Furthermore, we can see that the LRT is necessary for Methods 1, 2, and 4 to properly control the Type-1 error. Lastly, we see that the

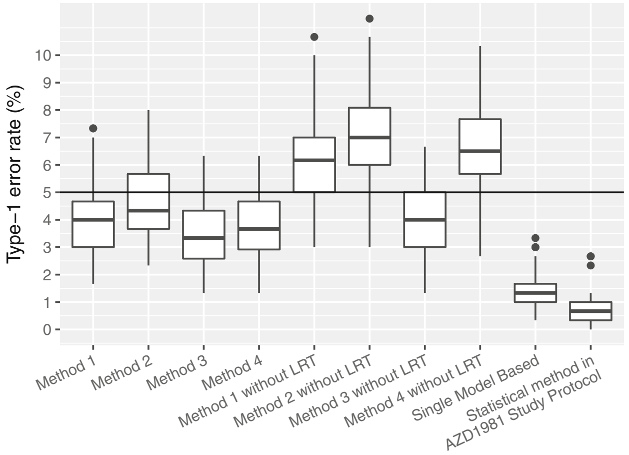

Fig. 7 Type-1 error rate, the probability of choosing either 10 mg, 40 mg, 100 mg, or 400 mg while there is no simulated drug effect. The significance level of all the methods was set to 0.05 hence if the Type-1 error is correctly controlled the Type-1 error rate should be at 5% (indicated by the horizontal line)

type-I error is lower than expected for the standard statistical test and Single Model Based method (using the Simulation Model).

Numerical experiment 3: decision-making accuracy

The decision-making accuracy of the various investigated methods is presented in Fig. 8. As can be seen, all model based method (Methods 1–4 and the Single Model Based method) makes the correct decision more often than the Statistical method employed in the AZD 1981 study protocol. Also, we can see that Method 4 performs relatively poorly compared to Methods 1–3.

Numerical experiment 4: probability of achieving target endpoint estimation accuracy

The Probability of achieving target endpoint estimation accuracy of the various investigated model-based methods is presented in Fig. 9. Note that if the investigated method estimates the probability of achieving the target endpoint without bias then the QQplot in Fig. 9 should follow the line of unity.

As can be seen in Fig. 9, Methods 2 and 4, using AIC as the statistical criteria in the methods, can calculate the probability of achieving target endpoint accurately. The bias on the calculated probability of achieving the target endpoint of the conventional model selection method (Method 1) is clearly observed. As discussed in literature (e.g., [2, 3]), if model selection is made based on one dataset the bias in the model selection procedure will be carried forward to subsequent analyses and any resulting quantity may be biased. Although the regular model averaging method (Method 3) should significantly decrease the effect of model selection bias, we still observe the presence of bias. Lastly, we observed that AIC is a more suitable statistical criterion than BIC for the proposed model averaging and selection methods.

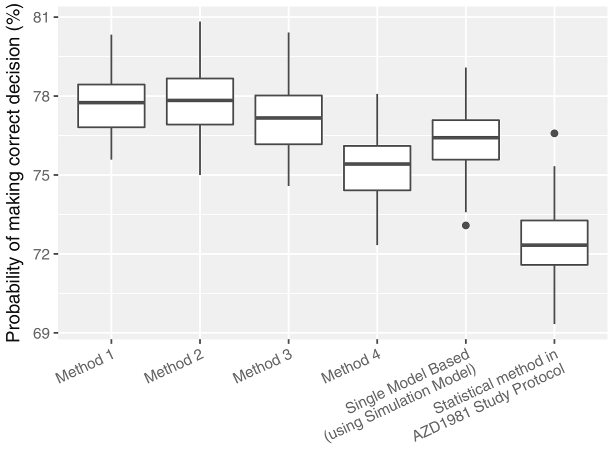

Fig. 8 Probability of making the correct decision (the correct decision is defined as choosing the dose that is above tMED)

Discussion

This work presents model averaging and selection methods that incorporate both model structure and parameter estimation. We have tested the proposed methods through realistic PhIIb dose finding and decision-making scenarios and demonstrated that the proposed methods could help increase the overall probability of making the correct decision at the end of PhIIb studies.

Through all the numerical experiments, the model based approaches (Methods 1–4 and Single Model Based method) outperformed the pairwise ANOVA based method used in the AZD1981 study protocol. Numerical Experiments 1 and 4 have shown that Methods 2 and 4 perform better than other methods for finding MED and estimating probability of achieving endpoint. Numerical Experiment 3 has shown that Method 2 can be used to make the investment decision more accurately than Method 4. Experiment 2 has shown that Type-1 error can be appropriately controlled using the LRT and the Type-1 error control of Method 2 is marginally better than the other methods (Method 1, 3 and 4). Thus, within the scope of our numerical experiment, Method 2 was the most accurate and precise compared to the other tested methods.

The numerical experiments indicated that AIC is a more suitable statistical criterion than BIC for the model averaging and selection methods we have tested. BIC takes the number of observations into account when weighing the penalty for the extra degrees of freedom. For nonlinear mixed effect models, the informativeness of the dataset not only depends on the number of observations but also a number of individuals. Hence, we conjecture that, by naively using the number of observations, BIC does not properly weigh the penalty term and some other way of quantifying the ‘informativeness of observations’ is necessary.

Although we have conducted a wide-range of numerical experiments within the scope of this project, we believe the accumulation of more experiences of these and other methods through applying them to more scenarios would be desirable. For example, it would be interesting to compare and/or integrate the methods presented here with the MCP-Mod approach [21, 22]. The MCP-Mod methodology allows model averaging and selection methods for the “Mod” portion of that framework, but entails an initial multiple comparison procedure (the “MCP” portion) that may be redundant with the LRT used here. To promote the application and further development of the proposed methodologies, we have made the methodologies

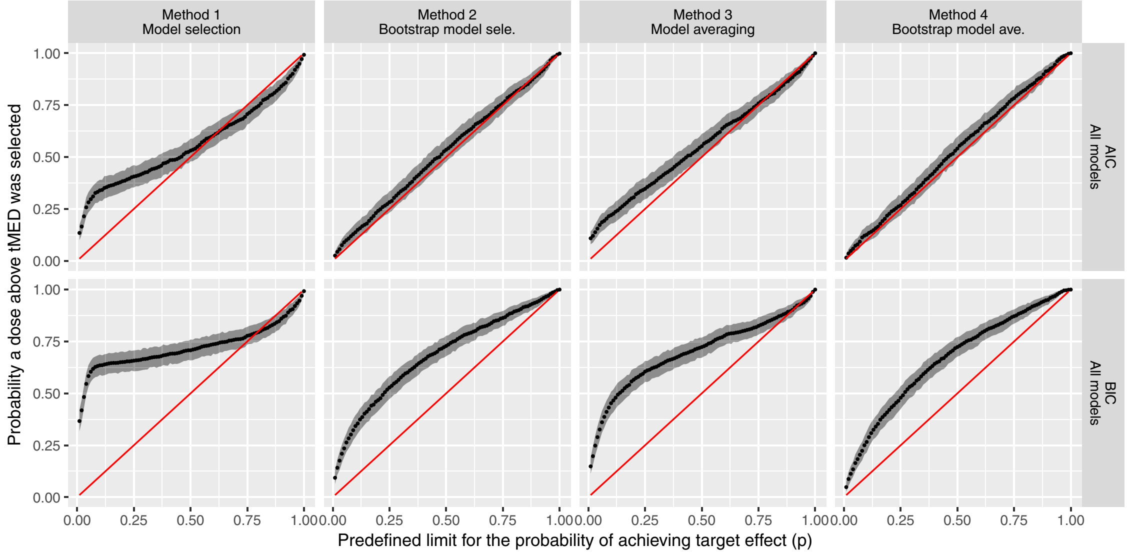

Fig. 9 The accuracy of the calculated probability of achieving a target endpoint. The x-axis is the predefined limit for the probability of achieving target endpoint where the dose was chosen. The y-axis is the probability that the chosen dose by the various methods is above tMED. If the probability of achieving the target endpoint is estimated without bias, the plot should lie on the line of identity (red straight line). Grey shaded areas are 95% confidence intervals calculated by the random sampling with replacement of the 1200 trial simulation datasets

investigated here available as a GUI based open source software (available at www.bluetree.me and the Mac App Store, app name: modelAverage) as well as an R script supplied as the supplementary material of this paper.

Conclusion

We recommend the use of the bootstrap model selection method (Method 2) presented in this paper when conducting model-based decision-making at the end of phase IIb study. The studies here indicate the proposed method reduces the analysis bias originating from model selection bias of single model structure based analyses. As a consequence of including model structure uncertainty, the quantified uncertainty may appear to be larger than single model based uncertainty; however, the method appears to more accurately reflect the true uncertainty of the investigated models and estimated parameters. The proposed method increases the probability of making the correct decisions at the end of phase IIb trial compared to conventional ANOVA-based Study Protocols.

Acknowledgements Authors would like to thank Professor Mats O. Karlsson for useful discussions, the Methodology group of the Uppsala University Pharmacometrics Group for discussions and Ms. Ida Laaksonen for refining the diagrams in this work.

Appendix: Detailed description of the method

In this section, we present step by step explanations of our model selection and averaging methodologies. We have implemented this methodology in the C++ language, and an open source software with a graphical user-interface is available at www.bluetree.me and the Mac App Store (app name: modelAverage). Also, easy to read (computationally not optimized) R script used for the numerical experiments presented in this is available as the supplementary material.

Prior to the analysis

We assume that prior to the analysis there are multiple candidate models pre-specified before the collection of the data, and subsequently collected study data is available.

We denote X_0 to be the independent variables (e.g., patient number, dosage, observation time, covariates) of the original dataset, and $\boldsymbol{y}_0$ be the dependent variable (e.g., observed biomarker values). For simplicity, we consider X_0 to be $N_{\text{records}} \times N_{\text{ind\_variables}}$ matrix and $\boldsymbol{y}_0$ to be a vector of N_{records} elements. We denote candidate model structures as $\text{model}_i\,(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta},\epsilon)$ where $\boldsymbol{x}$ is a vector of independent variables, $\boldsymbol{\theta}$ is a vector of fixed effect parameters, $\boldsymbol{\eta}$ is a vector of random effect variables related to individual variabilities, and $\boldsymbol{\epsilon}$ is a vector of random variable related to residual (unexplainable) variabilities. We denote the dose–effect relationship embedded in model structure model$_i$ by DE$_i$ (dose; $\boldsymbol{\theta}, \boldsymbol{\eta}$). We assume we have N_{model} candidate models and we denote the placebo model by model$_0$ hence we have $N_{\text{model}} + 1$ models in total.

Create bootstrap datasets based on the original data and estimate parameters for each bootstrap dataset

Construct bootstrap datasets based on $(X_0, \boldsymbol{y}_0)$ and we denote them as $\{(X_i, \boldsymbol{y}_i)\}_{i=1}^{N_{\text{bootstrap}}}$. We have used case sampling bootstrap in our numerical experiment; however, it can be extended to other types of bootstrap methods.

Estimate parameters and the maximum likelihood from each bootstrap dataset and estimate maximum likelihood parameters for each model for each bootstrap dataset and denote them as $(\widehat{\boldsymbol{\theta}}_{ij}, \widehat{\Omega}_{ij}, \widehat{\Sigma}_{ij})$ for $i = 0, \ldots, N_{\text{bootstrap}}$ and $j = 0, \ldots, N_{\text{model}}$, i.e.,

$$(\widehat{\boldsymbol{\theta}}_{ij}, \widehat{\Omega}_{ij}, \widehat{\Sigma}_{ij}) = \operatorname{argmax}(\boldsymbol{\theta}, \Omega, \Sigma) l(\text{model}_j(\cdot; \boldsymbol{\theta}, \cdot, \cdot); \Omega, \Sigma; X_i, \boldsymbol{y}_i),$$

where l is a likelihood function for the nonlinear mixed-effect model (we refer the readers to [23] for more detailed discussion and approximation methods for this likelihood function), and we denote the maximum likelihood as a $\widehat{l}_{ij}$, i.e.,

$$\widehat{l}_{ij} = l(\text{model}_j(\cdot; \widehat{\boldsymbol{\theta}}_{ij}, \cdot, \cdot); \widehat{\Omega}_{ij}, \widehat{\Sigma}_{ij}; X_i, \boldsymbol{y}_i).$$

Conduct numerical identifiability test and LRT

In order to have a rigorous Type-I error control in our model selection and averaging methods, each model that we use is subject to the LRT against the placebo model. That is to say, we have imposed the following to the estimated likelihood:

$$\widehat{l}_{ij} = \begin{cases} 0 & \text{if } |\widehat{l}_{ij} - \widehat{l}_{i0}| < \chi^2_{0.05}(\text{df}), \\ \widehat{l}_{ij} & \text{otherwise.} \end{cases}$$

where $\widehat{l}_{i0}$ is the estimated likelihood of the placebo model, df is the degree of freedom of the Chi square distribution that is calculated as the number of the dose–effect relationship related parameters (i.e., linear: df = 1, logLinear: df = 1, emax: df = 2, sigmoidal df = 3). Also, to reduce the chance of contaminating the model averaging and selection by non-identifiable models, we conduct numerical identifiability test to remove the models that are locally-practically non-estimable from a bootstrap dataset. We do so by re-estimating the model parameters using preconditioning [19]. We denote the estimated parameter and maximum likelihood using preconditioning by $(\widetilde{\boldsymbol{\theta}}_{ij}, \widetilde{\Omega}_{ij}, \widetilde{\Sigma}_{ij})$ and $\widetilde{l}_{ij}$, respectively. We reject the model by setting the likelihood to be zero if $\widehat{\boldsymbol{\theta}}_{ij}$ and $\widetilde{\boldsymbol{\theta}}_{ij}$ are significantly different while $\widehat{l}_{ij}$ and $\widetilde{l}_{ij}$ are similar. In particular, for our numerical experiment, we have imposed the following:

$$\widehat{l}_{ij} = \begin{cases} 0 & \text{if } |\widehat{l}_{ij} - \widetilde{l}_{ij}| < 0.1 \text{ and } \left| \dfrac{\widehat{\boldsymbol{\theta}}_{ij} - \widetilde{\boldsymbol{\theta}}_{ij}}{\widehat{\boldsymbol{\theta}}_{ij} + \widetilde{\boldsymbol{\theta}}_{ij}} \right|_{\inf} > 0.10, \\ \widehat{l}_{ij} & \text{otherwise.} \end{cases}$$

where the division of $\dfrac{\widehat{\boldsymbol{\theta}}_{ij} - \widetilde{\boldsymbol{\theta}}_{ij}}{\widehat{\boldsymbol{\theta}}_{ij} + \widetilde{\boldsymbol{\theta}}_{ij}}$ is the elementwise division of the vectors.

We acknowledge that the presented numerical identifiability test can only provide the evidence of non-estimability and does not necessarily prove the estimability of the model parameters; however, we have observed that this simple identifiability test has successfully reduced the number of non-estimable models included in the model averaging and selection schemes.

Simulate the quantities of interest (e.g., dose-endpoint relationships)

In this case, we construct the dose-endpoint relationships based on the estimated parameters in Step 2 and the definition of the model based clinical trial endpoint. Construct the estimated dose-endpoint relationships for each bootstrap dataset for each model and denote them as $h_{ij}(\text{dose})$. For example, if the endpoint is defined as the average effect like this FEV1 case study

$$h_{ij}(\text{dose}) = E_{\boldsymbol{\eta}}(\text{DE}_j(\text{dose}; \boldsymbol{\theta}_{ij}, \boldsymbol{\eta}); \Omega_{ij}),$$

where $E_{\boldsymbol{\eta}}(\cdot; \Omega)$ denotes expectation over $\boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\Omega)$. Other choices of the endpoint definition would be the median or percentile of $(\text{DE}_j(\text{dose}; \boldsymbol{\theta}_{ij}, \boldsymbol{\eta}); \Omega_{ij})$.

Depending on the definition of the endpoint and the structure of the dose–effect relationship with respect to $\boldsymbol{\eta}$, a stochastic simulation may be required to compute $h_{ij}(\cdot)$. The candidate drug effect for this case study is linear with respect to $\boldsymbol{\eta}$ and the end point defined by the study protocol is an average over the population, which we can analytically determine, $h_{ij}(\text{dose})$.

Summarize the simulations

Based on the computed likelihood $\widehat{l}_{ij}$ and the dose-endpoint relationship $h_{ij}(\text{dose})$, we compute the probability of achieving target endpoint versus dose relationship. In this step, we need to choose a weighting scheme where models are selected or averaged. We denote this weight function as w_j and it will depend on the likelihood $\widehat{l}_{ij}$ and the structure of the model (i.e., the number of model parameters). We denote the weight of the ith bootstrap sample with Model j as w_{ij}.

For the weights calculated based on AIC, we let w_{ij} to be the following:

$$w_{ij} = w_j(\widehat{l}_{ij}) = \exp\left(\ln(\widehat{l}_{ij}) - N_{\text{para}j}\right)$$

where $N_{\text{para}\,j}$ is the number of parameters of Model j.

For the weights calculated based on BIC, we let w_{ij} to be the following:

$$w_{ij} = w_j(\widehat{l}_{ij}) = \exp\left(\ln(\widehat{l}_{ij}) - N_{\text{para}\,j}\ln(N_{\text{obs}})/2\right).$$

where N_{obs} is the number of observations (total number of FEV1 measurements in a dataset).

Using this weight function, we can define the probability of achieving the target endpoint $p(\text{dose})$ as follows:

Method 1: model selection

$$p(\text{dose}) = \sum_{i=0,\dots,N_{\text{bootstrap}}} \begin{cases} 1/(1+N_{\text{bootstrap}}) & \text{if } h_{ik}(\text{dose}) > \text{TV} \\ 0 & \text{otherwise.} \end{cases}$$

where $k = \text{argmax}_j(w_{0j})$.

Method 2: model selection using bootstrap maximum likelihood

$$p(\text{dose}) = \sum_{i=0,\dots,N_{\text{bootstrap}}} \begin{cases} 1/(1+N_{\text{bootstrap}}) & \text{if } h_{ik_i}(\text{dose}) > \text{TV} \\ 0 & \text{otherwise} \end{cases}$$

where $k_i = \text{argmax}_j(w_{ij})$.

Method 3: model averaging

$$p(\text{dose}) = \sum_{i=0,\dots,N_{\text{bootstrap}},\, j=0,\dots,N_{\text{model}}} \begin{cases} \dfrac{w_{0j}}{\sum_{j=0,\dots,N\text{model}} w_{0j}} & \text{if } h_{ij}(\text{dose}) > \text{TV} \\ 0 & \text{otherwise.} \end{cases}$$

Method 4: model averaging using bootstrap maximum likelihood

$$p(\text{dose}) = \sum_{i=0,\dots,N_{\text{bootstrap}},\, j=0,\dots,N_{\text{model}}} \begin{cases} \dfrac{w_{ij}}{\sum_{j=0,\dots,N_{\text{model}}} w_{ij}} & \text{if } h_{ij}(\text{dose}) > \text{TV} \\ 0 & \text{otherwise.} \end{cases}$$

Detailed analyses of numerical experiments

In this section, we investigate the numerical computational results presented in the "Results" Section more in detail.

Effect of excluding the simulation model from the set of candidate models

All the numerical experiments presented so far has the simulation model (the model that was used to create a trial simulation dataset) included as one of the candidate

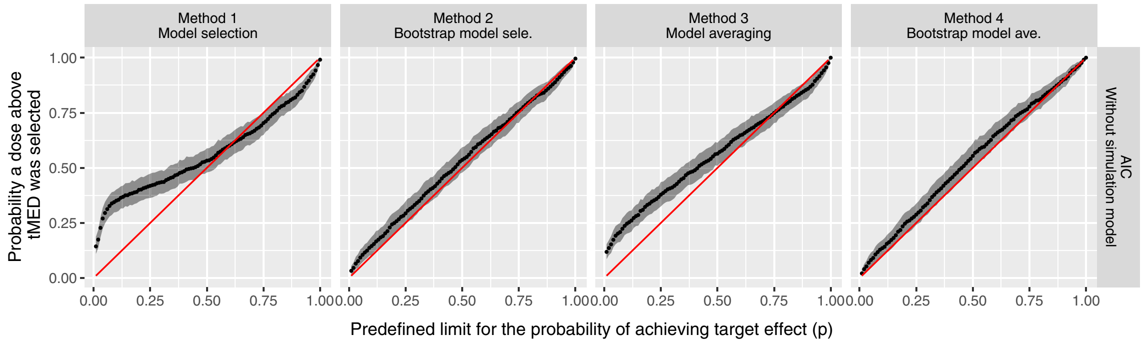

Fig. 10 The accuracy of calculated probability of achieving target endpoint. The Methods 1–4 used in this example did not include the simulation model

models. It is natural to suspect superior performance of the model averaging methods compared to the study protocol can be spoiled if the simulation model is not included in the set of candidate models. Numerical Experiments 1 and 4 were re-run with candidate models excluding the simulation models.

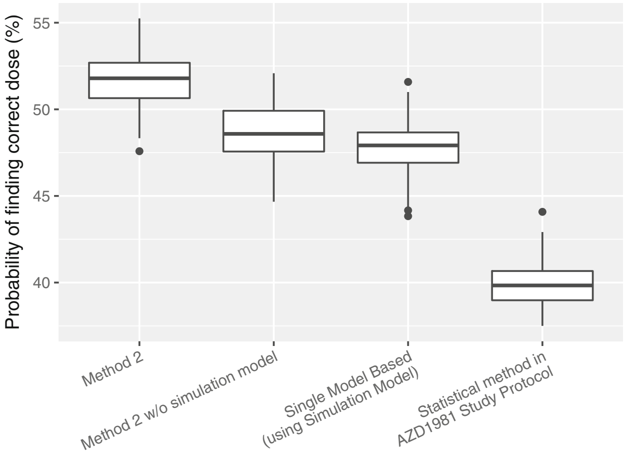

Fig. 11 Probability of finding the correct dose

As can be seen in Fig. 10, Method 2 and 4 can still be used to accurately estimate the probability of achieving target endpoints. As can be seen in Fig. 11, if the simulation model is excluded from the set of candidate models, the probability of finding correct dose decreases; however, it still performs superior to the ANOVA based statistical method used in the Study Protocol.

Accuracy of the probability of achieving target endpoint estimation for each simulation study

Q–Q plots similar to Fig. 9 are plotted for each Simulation Study in Fig. 12. Surprisingly, the calculated probability of achieving the target endpoint based on the model-based approach using the simulation model (i.e., using the model structure that was used to simulate the drug effect) was not very accurate especially in Simulation Studies 1 and 4. Further investigation on this simulation study has shown that the values of ED50 used for Emax or Sigmoidal models in Simulation Study 1 were below 10 mg and for the Simulation Study 4 were near 400 mg hence the design of the experiment was poor for these simulation studies. Due to the uninformative design of the study, the model-

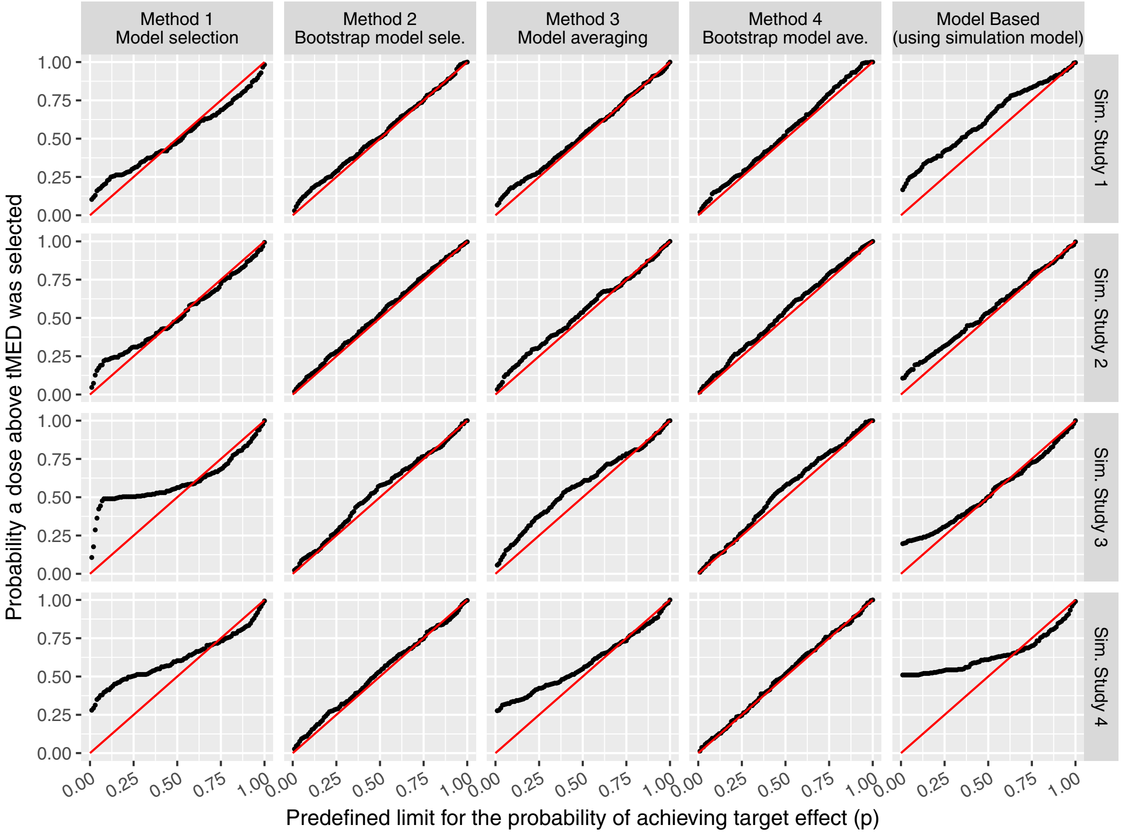

Fig. 12 The accuracy of calculated probability of achieving target endpoint. The x-axis is the predefined limit for the probability of achieving target endpoint where the dose was chosen. The y-axis is the probability that the chosen dose by the various methods is above tMED. If the probability of achieving the target endpoint is estimated without bias, the plot should lie on the red straight line

based analysis of data with simpler models provided more accurate predictions than with the model used to simulate the drug effect.

We can observe that Methods 2 and 4 slightly underestimate the probability of achieving target endpoint for Simulation Study 3; however, the inaccuracy of these methods are significantly less than that of other methods. Methods 2 and 4 are consistently more accurate than the other methods, hence, they can help reduce the risk of inaccurate estimation of the probability of achieving target endpoint by properly averaging over multiple possible model structures.

Precision of the estimation of the MED

To quantify the precision of the proposed methods, for each method and simulation, we have calculated the difference between the estimated dose that achieves a target effect with 70% probability and the estimated dose that achieves a target effect with 50% probability (we refer to this as the 'estimated MED range'). For comparison, the estimated MED range obtained using Methods 1, 3, 4 and Single Model Based method are compared against Method 2. The differences of the estimated MED range of various methods and Method 2 are depicted in Fig. 13. As can be seen, the estimated MED range is usually wider when estimated using Method 4 when comparing with Method 2. That is to say, Method 2 usually estimates the MED more precisely than Method 4. Although Methods 2 and 4 are similarly accurate, as demonstrated in Figs. 9, 10, 11 and 12, since Method 4 is less precise than Method 2, we have observed worse performance in Experiment 3.

Method 1 is typically more precise than Method 2 and the Single Model Based method is often more precise than

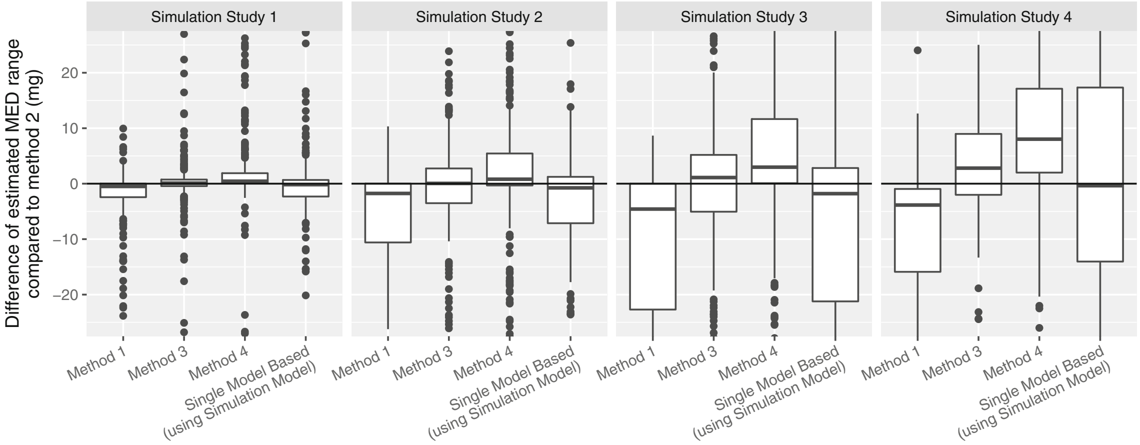

Fig. 13 Precision of the estimation of the MED compared to Method 2. The various methods were compared against Method 2 for the estimated MED range. A positive difference indicates the method has a larger estimated MED range, hence poorer precision

Table 3 Probability of selecting the correct dose (either 10, 40, 100, or 400 mg)

	Sim. study 1 (%)	Sim. study 2 (%)	Sim. study 3 (%)	Sim. study 4 (%)	Average (%)
Study protocol	48.0	32.7	30.0	49.0	39.9
Method 1	67.3	46.7	21.0	60.3	48.8
Method 2	66.0	48.3	26.7	65.7	51.7
Method 3	65.3	42.7	25.7	59.7	48.4
Method 4	66.0	44.0	31.3	60.0	50.3
Method 1 (w/o sim. model)	65.0	41.7	18.3	61.3	46.6
Method 2 (w/o sim. model)	64.0	44.7	20.0	66.7	48.9
Method 3 (w/o sim. model)	63.0	38.3	21.0	61.7	46.0
Method 4 (w/o sim. model)	63.0	42.0	25.0	63.0	48.3
Single model based (using simulation model)	68.0	45.0	42.0	49.0	51.0

For the model averaging and selection methods as well as the single model based analysis, the minimum dose with more than a 50% probability of achieving the target endpoint was selected

Table 4 Number of times a "correct" decision was made (dose selected above the theoretical MED). For the model-based methods the dose was chosen with at 70% of confidence level

	Sim. study 1 (%)	Sim. study 2 (%)	Sim. study 3 (%)	Sim. study 4 (%)	Average
Study protocol	92.00	78.30	70.70	49.00	72.50
Method 1	95.33	84.70	73.60	58.00	77.91
Method 2	95.00	83.70	76.30	57.00	78.00
Method 3	94.63	83.70	75.70	55.00	77.26
Method 4	92.97	83.00	73.60	52.30	75.47
Method 1 (w/o sim. model)	94.27	81.60	71.70	58.30	76.47
Method 2 (w/o sim. model)	93.67	80.90	73.00	56.70	76.07
Method 3 (w/o sim. model)	92.03	80.10	72.30	55.70	75.03
Method 4 (w/o sim. model)	90.67	80.30	70.40	52.70	73.52
Single model based (using simulation model)	96.30	83.0	76.30	50.30	76.47

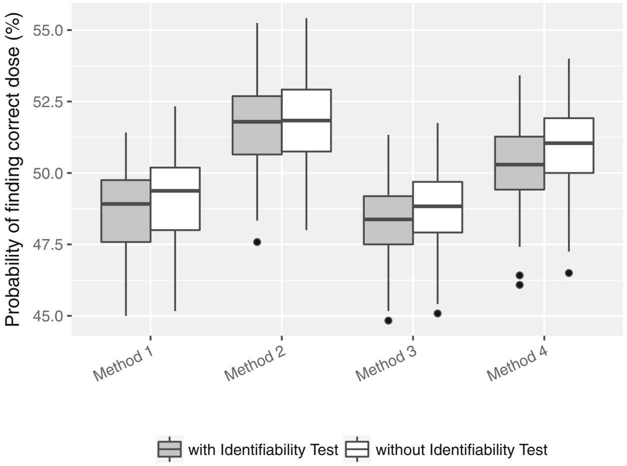

Fig. 14 Probability of finding correct dose with and without numerical identifiability test

Method 2. Both Method 1 and the Single Model Based method only used one model to simulate the endpoint hence more precision; however, as can be seen in Figs. 9, 10, 11 and 12 these methods are not accurate and, hence, not desirable methods.

Dose finding accuracy for each simulation study

In Table 3, the probability of choosing the correct dose was tabulated. By using Method 2, the probability of choosing the correct dose has increased from 39.92 to 51.67% compared to the study protocol. What is particularly noteworthy is that the dose finding accuracy has increased from 49 to 65.7% for Simulation Study 4 where the highest tested dose was the correct dose choice.

For all simulation studies except for simulation study 1, the bootstrap model selection and averaging methods (Methods 2 and 4) outperformed simple model selection and averaging methods (Methods 1 and 3). For all simulation studies except for simulation study 3, the model selection and averaging methods outperformed the study protocol even if the simulation model is not included in the candidate models. The overall performance of Methods 2 and 4 are similar to the case where idealized single model based analysis was done using the simulation model.

Decision-making accuracy for each simulation study

As can be seen in Table 4, Method 2 (bootstrap model selection) consistently outperforms Method 4. As discussed in the previously, Method 4 is generally less precise than 2. As a result, Method 4 does not perform as well as Method 2 when a dose is selected not based on the median.

Effect of the identifiability test

We have repeated all of the numerical experiments without identifiability tests. No significant difference in the results was observed. In order to correctly count the degree of freedom for AIC, we need to reject the models that are not identifiable; however, in practice, the inclusion of non-identifiable models did not influence the analysis results within the scope of this investigation. Figure 14 shows the probability of finding the correct dose for Numerical Experiment 1 both with and without the identifiability test. As can be seen, the identifiability test does not significantly influence the dose finding accuracy.

References

1. Karlsson KE, Vong C, Bergstrand M, Jonsson EN, Karlsson MO (2013) Comparisons of analysis methods for proof-of-concept trials. CPT 2(1):1–8
2. Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. Biometrics 53:603–618

3. Leeb H, Pötscher BM (2005) Model selection and inference: facts and fiction. Econom Theory 21(1):21–59
4. Bretz F, Pinheiro JC, Branson M (2005) Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics 61(3):738–748
5. Bornkamp B, Bretz F, Dmitrienko A, Enas G, Gaydos B, Hsu CH, Franz K, Krams M, Liu Q, Neuenschwander B, Parke T, Pinheiro J, Roy A, Sax R, Shen F (2007) Innovative approaches for designing and analyzing adaptive dose-ranging trials. J Biopharm Stat 17(6):965–995
6. Verrier D, Sivapregassam S, Solente AC (2014) Dose-finding studies, MCP-Mod, model selection, and model averaging: two applications in the real world. Clin Trials 11(4):476–484
7. Mercier F, Bornkamp B, Ohlssen D, Wallstroem E (2015) Characterization of dose-response for count data using a generalized MCP-Mod approach in an adaptive dose-ranging trial. Pharm Stat 14(4):359–367
8. Wählby U, Jonsson EN, Karlsson MO (2001) Assessment of actual significance levels for covariate effects in NONMEM. J Pharmacokinet Pharmacodyn 28(3):231–252
9. Aoki Y, Hooker AC (2016) Model averaging and selection methods for model structure and parameter uncertainty quantification. PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe. ISSN 1871-6032
10. Aoki Y, Hamrén B, Röshammar D, Hooker AC (2014) Averaged model based decision making for dose selection studies. PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe. ISSN 1871-6032
11. Martin MA, Roberts S (2006) Bootstrap model averaging in time series studies of particulate matter air pollution and mortality. J Eposure Sci Environ Epidemiol 16(3):242
12. Roberts S, Martin MA (2010) Bootstrap-after-bootstrap model averaging for reducing model uncertainty in model selection for air pollution mortality studies. Environ Health Perspect 118(1):131
13. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, Pangalos MN (2014) Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov 13(6):419
14. Wang X, Shang D, Ribbing J, Ren Y, Deng C, Zhou T et al (2012) Placebo effect model in asthma clinical studies: longitudinal meta-analysis of forced expiratory volume in 1 second. Eur J Clin Pharmacol 68(8):1157–1166
15. Stuart B, Lewis S, Boeckmann A, Ludden T, Bachman W, Bauer R, ICON plc (1980–2013). NONMEM 7.3. http://www.iconplc.com/technology/products/nonmem/
16. Lindbom L, Pihlgren P, Jonsson N (2005) PsN-Toolkit—a collection of computer intensive statistical methods for non-linear mixed effect modeling using NONMEM. Comput Methods Programs Biomed 79(3):241–257
17. R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
18. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York
19. Aoki Y, Nordgren R, Hooker AC (2016) Preconditioning of nonlinear mixed effects models for stabilisation of variance-covariance matrix computations. AAPS J 18(2):505–518
20. Lalonde RL, Kowalski KG, Hutmacher MM, Ewy W, Nichols DJ, Milligan PA, Corrigan BW, Lockwood PA, Marshall SA, Benincosa LJ, Tensfeldt TG (2007) Model-based drug development. Clin Pharmacol Ther 82(1):21–32
21. Bretz F, Pinheiro JC, Branson M (2005) Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics 61(3):738–748
22. Pinheiro J, Bornkamp B, Glimm E, Bretz F (2014) Model-based dose finding under model uncertainty using general parametric models. Stat Med 33(10):1646–1661
23. Wang Y (2007) Derivation of various NONMEM estimation methods. J Pharmacokinet Pharmacodyn 34(5):575–593