



HAL
open science

Model-based bioequivalence approach for sparse pharmacokinetic bioequivalence studies: Model selection or model averaging?

Morgane Philipp, Adrien Tessier, Mark Donnelly, Lanyan Fang, Kairui Feng, Liang Zhao, Stella Grosser, Guoying Sun, Wanjie Sun, France Mentré, et al.

► To cite this version:

Morgane Philipp, Adrien Tessier, Mark Donnelly, Lanyan Fang, Kairui Feng, et al.. Model-based bioequivalence approach for sparse pharmacokinetic bioequivalence studies: Model selection or model averaging?. *Statistics in Medicine*, 2024, Online ahead of print. 10.1002/sim.10088 . inserm-04605232

HAL Id: inserm-04605232

<https://inserm.hal.science/inserm-04605232>

Submitted on 7 Jun 2024


HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Model-based bioequivalence approach for sparse pharmacokinetic bioequivalence studies: Model selection or model averaging?

Morgane Philipp¹  | Adrien Tessier² | Mark Donnelly³ | Lanyan Fang³ | Kairui Feng³ | Liang Zhao³ | Stella Grosser⁴ | Guoying Sun⁴ | Wanjie Sun⁴ | France Mentré¹ | Julie Bertrand¹

¹Université Paris Cité, IAME, INSERM, Paris, France

²Clinical Pharmacometrics, Quantitative Pharmacology, Servier, Suresnes, France

³Division of Quantitative Methods and Modeling, Office of Research and Standards, Office of Generic Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

⁴Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Correspondence

Morgane Philipp, Université Paris Cité, IAME, INSERM, 75018 Paris, France.
Email: morgane.philipp@inserm.fr

Funding information

U.S. Food and Drug Administration, Grant/Award Number: 75F40119C10111

Conventional pharmacokinetic (PK) bioequivalence (BE) studies aim to compare the rate and extent of drug absorption from a test (T) and reference (R) product using non-compartmental analysis (NCA) and the two one-sided test (TOST). Recently published regulatory guidance recommends alternative model-based (MB) approaches for BE assessment when NCA is challenging, as for long-acting injectables and products which require sparse PK sampling. However, our previous research on MB-TOST approaches showed that model misspecification can lead to inflated type I error. The objective of this research was to compare the performance of model selection (MS) on R product arm data and model averaging (MA) from a pool of candidate structural PK models in MBBE studies with sparse sampling. Our simulation study was inspired by a real case BE study using a two-way crossover design. PK data were simulated using three structural models under the null hypothesis and one model under the alternative hypothesis. MB-TOST was applied either using each of the five candidate models or following MS and MA with or without the simulated model in the pool. Assuming T and R have the same PK model, our simulation shows that following MS and MA, MB-TOST controls type I error rates at or below 0.05 and attains similar or even higher power than when using the simulated model. Thus, we propose to use MS prior to MB-TOST for BE studies with sparse PK sampling and to consider MA when candidate models have similar Akaike information criterion.

KEYWORDS

bioequivalence, model averaging, model selection, non-linear mixed effect models, two one-sided test

Abbreviations: AIC, Akaike information criterion; AUC, area under the curve; BE, bioequivalence; BSV, between-subject variability; BOT, bioequivalence optimal test; C_{max} , maximum concentration; CI, confidence interval; EMA, European Medicine Agency; FDA, Food and Drug Administration; FIM, Fisher information matrix; GLMEM, general linear mixed effect model; GMR, geometric mean ratio; LOQ, limit of quantification; MA, model averaging; MB, model-based; MC, Monte Carlo; MCMC, Monte Carlo Markov chains; MS, model selection; NCA, non-compartmental analysis; NLMEM, non-linear mixed effect model; PK, pharmacokinetics; R, reference; RUV, residual unexplained variance; SAEM, stochastic approximation of expectation-maximization; SE, standard error; SP, secondary parameter; T, test; TOST, two one-sided test; WSV, within-subject variability.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 Servier and The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

1 | INTRODUCTION

Bioequivalence (BE) studies are key in the development of generic drugs and new formulations of an approved drug product. The U.S. Food and Drug Administration (FDA) states that a proposed drug product, or a test (T) product, is bioequivalent to a reference listed drug, or a reference (R) product, if the rate and extent of absorption of the drug do not show a significant difference from the rate and extent of absorption of the listed drug when administered at the same molar dose of the therapeutic ingredient under similar experimental conditions in either a single dose or multiple doses.¹

Conventionally, the two metrics used for BE evaluation are the area under the curve (AUC) and the maximum plasma concentration (C_{max}), which respectively characterize the extent and rate of absorption, and the statistical test performed is the two one-sided test (TOST) proposed by Schuirmann.² When using average BE analysis, FDA (2001) and European Medicines Agency (EMA, 2010) recommend that if the 90% confidence interval (CI) of the geometric mean ratio (GMR, e.g., the anti-log of the difference of means in log-transformed pharmacokinetic (PK) metric between T and R) is contained within the BE limits 0.8 and 1.25, BE is concluded.^{3,4} The 2001 FDA guidance states that “due to the nature of normal-theory confidence intervals, this is equivalent to carrying out two one-sided tests of hypothesis at the 5% level of significance (Schuirmann, 1987).”³

Regulatory authorities typically recommend BE studies to be conducted using a single-dose, two-period, two-sequence, two-treatment, crossover (two-way crossover) study design and PK data to be analyzed using non-compartmental analysis (NCA).^{4,5} NCA is a model-independent approach which requires few assumptions, but rich data is recommended. For instance, FDA recommends that PK sampling includes 12 to 18 samples with at least three sampling points after the peak.¹ Individual AUC are then derived using the trapezoidal rule and the treatment effects on AUC and C_{max} are obtained using a general linear mixed effect model (GLMEM) on individual log-transformed AUC and C_{max} .

In 2011, our group had proposed a model-based (MB) approach for BE assessment (MB-TOST) where PK data are analyzed using a non-linear mixed effect model (NLMEM) with a treatment effect on all fixed effects. MB-TOST provided comparable results to the NCA-TOST in BE PK studies with rich data.⁶ MB-TOST even outperformed the non-parametric bootstrap NCA-based approach recommended by FDA for PK BE studies on ophthalmic drug products using a parallel design with only one PK sample taken per patient if the underlying PK structural model can be correctly specified.⁷ However, due to under-estimation of the asymptotic standard errors (SE) on sparse data, MB-TOST showed inflated type I error rates.⁶ Thus, we proposed alternative methods for SE calculation at finite distance.⁸ In addition, we proposed a bioequivalence optimal test (BOT), which can be more efficient than TOST.⁹

In a broader context, model-based drug development (MIDD) has demonstrated its ability to improve the effectiveness of drug development and the regulatory decision-making process. Indeed, it efficiently streamlines time and resource allocation during the initial phases of learning, while also providing valuable insights for the subsequent confirmatory stages of development.¹⁰ Recent revisions to *Population Pharmacokinetics Guidance for Industry*, published by FDA in February 2022, note that MB approaches can be an alternative for evaluating the BE of long-acting injectables, products with sparse PK sampling or other scenarios when NCA becomes challenging.¹¹ The absence of harmonization should be contributed to an under utilization of MIDD methods in both drug development and regulatory decision-making.¹⁰ However, the guidance notes that “for such applications, the model’s intended use and its regulatory impact determine the level of robustness needed for model evaluation”.¹¹ Thus, a robust model evaluation may provide opportunities to utilize the insights gained from available data for optimal designs, improve the understanding of confirmatory studies, and diminish reliance on conventional methods during drug development when they are not feasible.¹⁰ One notable challenge with using MB approach to assess BE is the lack of information or consensus on the true PK model of the reference product. In this case, a pool of candidate models may be considered in the MBBE approach. Our previous evaluation of PK BE studies with sparse PK sampling and parallel design on drugs with a long half-life, showed that a model selection (MS) step on the R product arm data could prevent a type I error inflation due to model misspecification when the T product arms have the same PK model as the R product arms.¹² Another method, model averaging (MA), which allows for model uncertainty, has recently showed good statistical properties in dose finding clinical trials¹³⁻¹⁵ as well as model-informed precision dosing.¹⁶ MS and MA have been compared in numerous other situations, but our study investigated the potential utility of these methods in MBBE evaluation of PK studies with sparse sampling, which present certain challenges using conventional BE methods.¹⁷

The aim of the present work was to compare the impact of MS and MA in BE evaluation for PK studies with sparse sampling when a pool of candidate structural PK models is available. Of note, we did not explore misspecification of

the random-effect and/or residual error variability models. First, we studied the concentrations of amlodipine collected in a single-dose, two-way crossover BE study of two formulations of a drug used to treat hypertension, developed at Servier.¹⁸ MS and MA were applied to the real case study data using the original sampling times and a subset of sparse sampling times, with a pool of ten candidate models, then MB-TOST was applied. Second, we performed a simulation study to evaluate the impact of MS and MA in BE evaluation using MB-TOST in terms of type I error rate and power when T and R have the same PK structural model. We simulated with three different structural PK models under the null hypothesis and with one model under the alternative. MS and MA in BE evaluation using MB-TOST were applied to the simulated datasets with a pool of four to five candidate models whether the simulated model was excluded or included in the pool.

2 | MODEL-BASED BIOEQUIVALENCE

2.1 | Model-based two one-sided tests

MB-TOST has been previously studied in sparse sampling PK BE studies using parallel^{6,8,12} and crossover^{6,8} designs.

In a cross-over study with two treatments (R and T), two periods and two sequences (RT and TR), the NLMEM describing the drug concentration y_{ijk} of individual $i \in \{1, \dots, N\}$, at sampling time t_{ijk} with $j \in \{1, \dots, n_{ik}\}$, for the period/occasion $k \in \{1, 2\}$ can be written as follows:

$$y_{ijk} = f(t_{ijk}, \phi_{ik}) + (a + b \times f(t_{ijk}, \phi_{ik})) \times \epsilon_{ijk}$$

with f the non-linear structural model depending on ϕ_{ik} the p -dimensional vector of individual parameters for subject i at occasion k . $\epsilon_{ijk} \sim N(0, 1)$ refers to the measurement error for the individual i , at the time t_{ijk} , for the period k with a the intercept and b the proportional term of the residual unexplained variability (RUV) model. To ensure positiveness, we use log-normally distributed parameters. Therefore, all the elements of the vector ϕ_{ik} can be detailed as follows, here we consider the case of the l^{th} element:

$$\log(\phi_{ikl}) = \log(\mu_l) + \beta_l^T \times T_{ik} + \beta_l^P \times P_k + \beta_l^S \times S_i + \eta_{il} + \kappa_{ikl}$$

with μ_l the fixed effect of the reference product for the l^{th} PK parameter. T_{ik} , P_k and S_i correspond to the two-dimensional vector of treatment, period and sequence covariates with the first element being considered as the reference. Consequently β_l^T , β_l^P and β_l^S are two-dimensional vectors of treatment, period and sequence effect coefficients for the l^{th} PK parameter with a 0 as first element. η_{il} and κ_{ikl} correspond to random-effects of individual i at occasion k for the parameter l respectively capturing the between and within subject variability (BSV and WSV) where $\eta_i \sim N(0, \Omega)$ and $\kappa_{ik} \sim N(0, \Gamma)$ are the p -dimensional random-effect vectors and Ω and Γ are the $p \times p$ variance covariance matrices. Finally, $\theta = (\mu, \beta^T, \beta^P, \beta^S, \Omega, \Gamma, a, b)$ is the vector of parameters to be estimated with AUC and C_{\max} being secondary parameters (SP) from the PK model. Additional steps are necessary to obtain the treatment effect estimate of the SP, $\hat{\beta}_{SP}^T$ (ie, $\hat{\beta}_{AUC}^T$ and $\hat{\beta}_{C_{\max}}^T$), and their associated SE; calculation details are provided in Appendix A. For a full explanation of this method, please refer to appendix 2 of Guhl et al.¹²

The null hypothesis of the TOST, $H_0: \{\beta_{SP}^T \leq -\delta \text{ or } \beta_{SP}^T \geq \delta\}$ is decomposed in two one-sided hypotheses: $H_{0,-\delta} : \{\beta_{SP}^T \leq -\delta\}$ and $H_{0,\delta} : \{\beta_{SP}^T \geq \delta\}$ and these two hypotheses can respectively be tested with the following statistics:

$$W_{-\delta} = \frac{\hat{\beta}_{SP}^T + \delta}{SE(\hat{\beta}_{SP}^T)} \text{ and } W_{\delta} = \frac{\hat{\beta}_{SP}^T - \delta}{SE(\hat{\beta}_{SP}^T)}$$

where $\hat{\beta}_{SP}^T$ is the treatment effect estimate for the secondary parameter of interest and $SE(\hat{\beta}_{SP}^T)$ its standard error estimate. The null hypothesis is rejected and BE is established if $W_{-\delta} \geq z_{1-\alpha}$ and $W_{\delta} \leq -z_{1-\alpha}$ where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the normal distribution $N(0, 1)$ with $\alpha = 5\%$. Similarly, BE can be concluded when the 90% CI of the GMR = $\exp(\hat{\beta}_{SP}^T)$ is included in $[\exp(-\delta), \exp(\delta)] = [0.8, 1.25]$.

2.2 | Model selection

We considered a pool of M structural models with $m \in \{1, \dots, M\}$ to select from. The reference product arm data are fitted with all M structural PK candidate models. In these models, the fixed effects, the variance-covariance matrix of the random-effects capturing the BSV, as well as the error model terms were estimated, that is, $\hat{\theta}_R = (\hat{\mu}, \hat{\Omega}, \hat{a}, \hat{b})$. Indeed, no treatment, period, and sequence effects were included and WSV was not estimated.

Here, the model selection is performed according to the Akaike information criterion $AIC = -2 \times \log(\hat{L}) + 2 \times \dim(\hat{\theta}_R)$ where \hat{L} is the maximized likelihood. The model with the smallest AIC is selected.

The selected structural PK model is, then, fitted to both arm data and MB-TOST is applied using this model estimates. Here, we make the assumption that R and T products have the same structural PK model.

We believe it is preferable to apply MS of the structural PK model on R arms only and not to use data from the T product arms at that stage of the analysis. Nonetheless, we also investigated MS of the structural PK model on both the R and T product arms.

2.3 | Model averaging

Both reference and test product arm data are fitted with all M structural PK candidate models. In these models, the fixed effects, the period, sequence and treatment effects, the variance-covariance matrix of the random-effects capturing the BSV and the WSV, as well as the error model terms were estimated, that is, $\hat{\theta} = (\hat{\mu}, \hat{\beta}^T, \hat{\beta}^P, \hat{\beta}^S, \hat{\Omega}, \hat{\Gamma}, \hat{a}, \hat{b})$. A weight is then associated to each of the M candidate models. The weight is not a known constant but is derived from the model's AIC , as Buatois et al¹⁵ showed this criterion to have the best predictive performance for dose-response modelling:

$$w_m = \frac{\exp(-\Delta AIC_m/2)}{\sum_{m'=1}^M \exp(-\Delta AIC_{m'}/2)}$$

where AIC_m is the $AIC = -2 \times \log(\hat{L}) + 2 \times \dim(\hat{\theta})$ of the m^{th} model and $\Delta AIC_m = AIC_m - \min(AIC_1, \dots, AIC_M)$. The formula is derived from the Schwarz's (1978)¹⁹ approximation of the Bayes factor,²⁰ such that the two models with the same AIC value are given the same weight whether or not they have the same penalty ($2 \times \dim(\hat{\theta})$).

Then, with w_m we derive the $\hat{\beta}_{SP}^T$ estimate as follows:

$$\hat{\beta}_{SP}^T = \sum_{m=1}^M w_m \times \hat{\beta}_{SP_m}^T$$

with $\hat{\beta}_{SP_m}^T$ the treatment effect estimate on the secondary parameter of interest of the model m . The formula by Turek et al²¹ is used to obtain the $SE(\hat{\beta}_{SP}^T)$:

$$SE(\hat{\beta}_{SP}^T) = \sum_{m=1}^M w_m \sqrt{\left(SE(\hat{\beta}_{SP_m}^T) \times \frac{t_{v_m, 1-\alpha}}{z_{1-\alpha}} \right)^2 + \left(\hat{\beta}_{SP_m}^T - \hat{\beta}_{SP}^T \right)^2}$$

with $SE(\hat{\beta}_{SP}^T)$ the SE of the treatment effect estimate of the m^{th} model, $t_{v_m, 1-\alpha}$ the $(1 - \alpha)$ quantile of the t-distribution with $v_m = 2N - \dim(\mu_m)$ degrees of freedom⁸ and $z_{1-\alpha}$ the $(1 - \alpha)$ quantile of the standard normal distribution.

2.4 | Implementation

We used the stochastic approximation of expectation-maximization (SAEM) estimation algorithm to estimate all model parameters. SE were derived from the variance-covariance estimation matrix ($\widehat{VAR}(\hat{\theta})$) estimated as the inverse of the Fisher information matrix (FIM) obtained by linearization. The log-likelihood was estimated by importance sampling with the concentrations below the limit of quantification (LOQ) contributing as left-censored data. For numerical reasons, models with $w_m < 0.005$ on a dataset were removed from the candidates pool for this dataset and new weights were calculated with the reduced pool.

The auto-stop criteria, which enables switching automatically from the exploratory phase to the smoothing phase, was deactivated to have similar conditions between the runs. Similarly, the simulated annealing, which allows a wider exploration of the estimate space for a longer time, was deactivated as there is accumulated knowledge prior to BE studies.²² The numbers of exploratory and smoothing phase iterations were fixed to $K_1 = 1000$ and $K_2 = 500$, respectively. Initial values for the parameters fixed effects, a and b can be found in Supplementary Material (Table S1 in Supplementary Data S1), between and within subject standard deviations were initialised at 1 (to enable a wide exploration of the parameter estimate space) and treatment, period and sequence effect coefficients were initialised at 0 (as a neutral start for the exploration). The number of Monte Carlo Markov chains (MCMC) of the SAEM algorithm was set to 5 and increased to 10 when difficulties were met to obtain the SE of the fixed effects and treatment effects required to calculate the SE of the AUC and C_{\max} treatment effects.

Data management and visualization were performed with R version 4.0.3 using the packages dplyr, tidyverse and ggplot2. PK modelling was performed with Monolix 2020R1 and automatized using the Monolix application programming interface for R with the lixoftConnector package, the corresponding R code ran on the INSERM, Université Paris Cité, UMR 1137, computing center with R version 4.1.2. The MB-TOST statistics were calculated using R version 4.0.3.

3 | REAL CASE STUDY

3.1 | Data

The real case study was based on data from a Servier phase I, open-label, randomised, two-way, single-dose, crossover study in healthy volunteers,¹⁸ conducted in 2006, comparing two treatments for hypertension; a reference product with two tablets: perindopril ter-butylamine (8 mg) + amlodipine (10 mg) and a test product with one tablet fixed combination of perindopril arginine (10 mg) + amlodipine (10 mg). In this work, only the amlodipine concentrations were considered for the BE assessment.

During each period, the 36 healthy volunteers enrolled in the study received 10 mg of amlodipine at time $t = 0$ and 22 PK samples were collected pre-dose and at 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8, 10, 12, 16, 24, 36, 48, 72, 96, 144, 192 and 240 h after the dose. The amlodipine concentrations were determined by a validated LC-MS/MS method, with a LOQ at $0.05 \mu\text{g/L}$.¹⁸ A 3-week washout separated the two periods. Two sequences were defined (sequence 1: RT and sequence 2: TR). Half of the patients were assigned to sequence 1 and the other half to sequence 2.

Our group proposed MB-TOST as an alternative to NCA for PK data with sparse sampling.⁶ So, to challenge MB-TOST on the real case study data, we selected six sampling times out of the 22 from the original study empirically based on the observation of the PK profiles, to create a sparse dataset. We selected the sampling time of 0.25 h to capture a potential delay in the absorption phase, sampling times at 3, 6, and 12 h to capture the C_{\max} , and sampling times at 72 and 144 h to capture the elimination phase.

3.2 | Analysis

In the work by Rohatagi et al,²³ amlodipine PK was best described by a one-compartment model with a delayed first order absorption and a linear elimination. They explored different absorption models and number of compartments for distributions were considered. As such, we defined a pool of $M = 10$ candidate PK models detailed in Table 1.

BE was assessed on the rich and sparse datasets using MS and MA.

3.3 | Results

The amlodipine individual concentration profiles in the rich and sparse datasets are displayed in Figure 1.

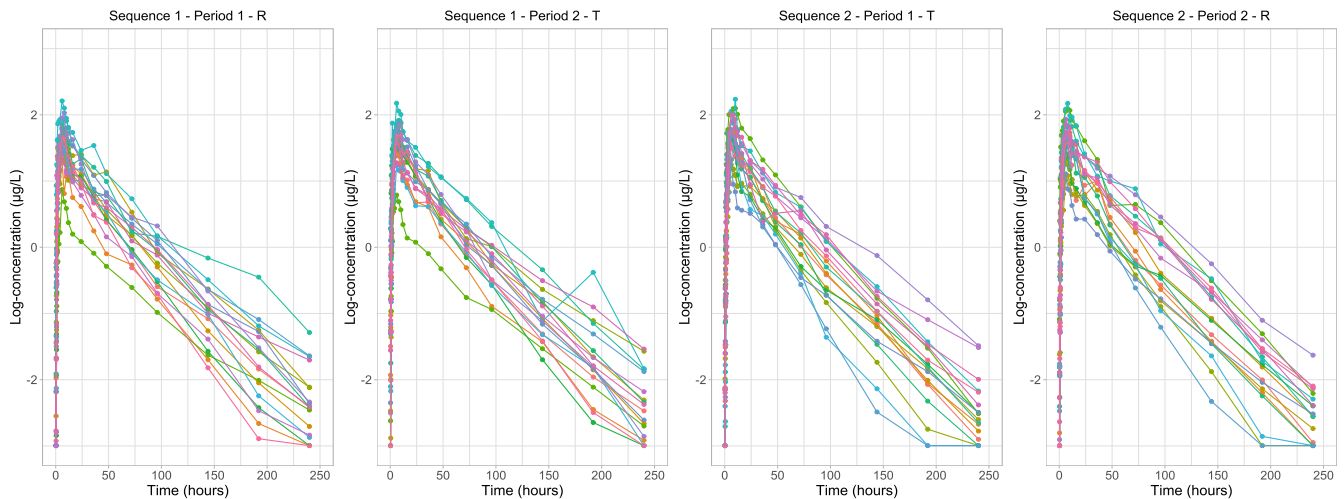
The model which best described the concentrations in the reference product arms of the rich and sparse datasets was the TRANSIT_2-COMPT model (Table S2 in Supplementary Data S1), with the highest number of parameters, and MA on both arm data of the rich and sparse datasets gave a weight = 1 to this model. Consequently, MB-TOST was applied using the parameter estimates of the TRANSIT_2-COMPT model following both MS and MA. The parameter estimates of the model selected on the R arms, with their SE, can be found in Table S3 in Supplementary Data S1 for the rich and

TABLE 1 Description, name and list of parameters for the ten structural PK models in the pool of candidates.

Description	Name	Parameters
One compartment model with first order absorption	1-ORDER_1-COMPT	ka, V/F, Cl/F
Two compartments model with first order absorption	1-ORDER_2-COMPT	ka, V1/F, Cl/F, V2/F, Q/F
One compartment model with zero order absorption	0-ORDER_1-COMPT	Tk0, V/F, Cl/F
Two compartments model with zero order absorption	0-ORDER_2-COMPT	Tk0, V1/F, Cl/F, V2/F, Q/F
One compartment model with delayed first order absorption	LAG_1-ORDER_1-COMPT	Tlag, ka, V/F, Cl/F
Two compartments model with delayed first order absorption	LAG_1-ORDER_2-COMPT	Tlag, ka, V1/F, Cl/F, V2/F, Q/F
One compartment model with delayed zero order absorption	LAG_0-ORDER_1-COMPT	Tlag, Tk0, V/F, Cl/F
Two compartments model with delayed zero order absorption	LAG_0-ORDER_2-COMPT	Tlag, Tk0, V1/F, Cl/F, V2/F, Q/F
One compartment model with transit absorption	TRANSIT_1-COMPT	ktr, Mtt, ka, V/F, Cl/F
Two compartments model with transit absorption	TRANSIT_2-COMPT	ktr, Mtt, ka, V1/F, Cl/F, V2/F, Q/F

Abbreviations: Cl/F, apparent clearance; ka, absorption constant rate; ktr, transit rate; Mtt, mean transit time; Tk0, absorption duration; V/F, apparent volume of the central compartment; V1/F, apparent volume of the compartment 1; V2/F, apparent volume of the compartment 2; Q/F, apparent inter-compartmental clearance.

Rich (original) data



Sparse data

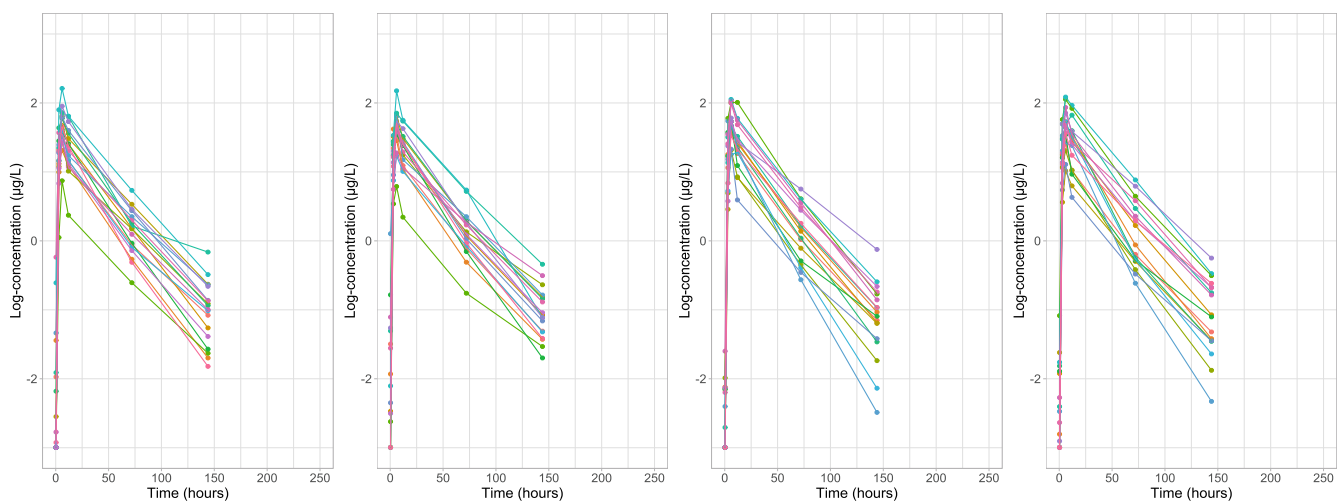


FIGURE 1 Spaghetti plots of amlodipine individual log-concentration versus time profiles in the original data with rich sampling (top) and sparse sampling (bottom) ordered by sequence and period (R: reference and T: test product).

TABLE 2 Geometric mean ratio [90% confidence interval] using a model-based (MB) and non-compartmental analysis (NCA) for AUC and C_{\max} on the original data with rich sampling and using a MB analysis only on the data with sparse sampling.

Method	Rich		Sparse	
	AUC	C_{\max}	AUC	C_{\max}
MB*	1.00 [0.97, 1.04]	1.04 [0.99, 1.10]	1.00 [0.96, 1.04]	0.97 [0.85, 1.12]
NCA	1.00 [0.97, 1.04]	1.03 [0.99, 1.07]		

*Because the weight associated to the TRANSIT_2-COMPT model was equal to 1, MS and MA led to the same results.

sparse data sets, respectively, and on both arms R and T in Table S4 in Supplementary Data S1. Parameter estimates were similar with larger SE on the sparse dataset for the variances of the absorption parameters and the additive term of the RUV model.

For both rich (original) and sparse datasets, MB-TOST led to the conclusion that the R and T products were bioequivalent with GMR 90% CI for AUC and C_{\max} included in the [0.8, 1.25] range (Table 2). On the rich (original) dataset, MB-TOST and the traditional BE method (NCA-TOST calculated from observation data) led to the same result.¹⁸

MS on both the R and T product arms led to the same results (Table S2 in Supplementary Data S1).

4 | SIMULATION STUDY

4.1 | Settings

We simulated a PK BE study using a single-dose, two-way crossover design with 40 subjects based on the real case study data. Twenty subjects were assigned to sequence 1 (RT) and 20 subjects were assigned to sequence 2 (TR). Each patient received a 10 mg dose of amlodipine at time $t = 0$ of each period. All subjects were sampled at 0.3, 3, 6, 12, 72, 144 h post-dose for both periods.

Under the null hypothesis of bioequivalence, we set $\beta^T = \log(1.25)$ on apparent clearance and volume parameters and we simulated three scenarios, each with a different structural PK model. We used the three PK models which best described the real case study original data in the reference product arms: TRANSIT_1-COMPT, TRANSIT_2-COMPT and LAG_0-ORDER_2-COMPT (see Table S2 in Supplementary Data S1). The simulated PK parameter fixed effects are reported in Table 3. They are derived from the fit of the R arms of the real case study data (Table S3 in Supplementary Data S1). We simulated only one scenario under the alternative hypothesis of bioequivalence of the R and T products, with $\beta^T = \log(1.05)$ using the TRANSIT_2-COMPT model. Under both hypotheses, no sequence and period effects (ie, $\beta^S = 0$ and $\beta^P = 0$) were simulated.

We simulated high levels of variability (ie, BSV = 50% and WSV = 30%) for all PK parameters. Of note, BSV and WSV were estimated at 10% to 60% and 5% to 80%, respectively, on the real case study data (Table S4 in Supplementary Data S1). The additive term of the residual error variance model was simulated at 0.02 $\mu\text{g/L}$ and the proportional term at 30% ($b = 14\%$ on the real case study data (Tables S3 and S4 in Supplementary Data S1)).

For each scenario, $S = 200$ datasets were simulated (ie, 800 in total) using the `simulx` function of the `mlxR` package in R version 4.0.3.

4.2 | Analysis

We considered a pool of $M = 5$ candidate models to fit the simulated data: TRANSIT_1-COMPT, TRANSIT_2-COMPT, LAG_0-ORDER_2-COMPT, LAG_1-ORDER_2-COMPT and LAG_0-ORDER_1-COMPT models (cf. Table 1 for model parameters). The first three models described correspond to the simulated models. The latter two models described were included because they provided a simpler model of absorption and combined absorption and distribution, respectively, when compared to the simulated models.

BE was assessed for each of the 800 simulated datasets using each of the five PK models in the pool of candidates, MS, and MA. Both MS and MA were performed with and without the simulated model in the pool of candidates.

TABLE 3 Parameter fixed effect values for the three simulated structural PK models.

TRANSIT_1-COMPT		TRANSIT_2-COMPT		LAG_0-ORDER_2-COMPT	
Parameter (units)	Value	Parameter (units)	Value	Parameter (units)	Value
μ_{ktr} (/h)	1.2	μ_{ktr} (/h)	1.4		
μ_{Mit} (h)	1.2	μ_{Mit} (h)	0.9	μ_{Tlag} (h)	0.3
μ_{ka} (/h)	1.3	μ_{ka} (/h)	0.6	μ_{Tko} (h)	3.7
$\mu_{Cl/F}$ (L/h)	40	$\mu_{Cl/F}$ (L/h)	41	$\mu_{Cl/F}$ (L/h)	41
$\mu_{V1/F}$ (L)	2130	$\mu_{V1/F}$ (L)	1660	$\mu_{V1/F}$ (L)	1890
		$\mu_{Q/F}$ (L/h)	42	$\mu_{Q/F}$ (L/h)	18
		$\mu_{V2/F}$ (L)	600	$\mu_{V2/F}$ (L)	400

4.3 | Evaluation

We evaluated the type I error rate of the MB-TOST for AUC and C_{max} on the three scenarios simulated under H_0 and the power on the scenario simulated under H_1 . The 95% CI around the estimated type I error rates and powers were calculated assuming a binomial distribution. Type I error rates were compared to 0.05 with an exact two-sided binomial test at the level 5%.

4.4 | Results

Under H_0 , the simulated profiles were very similar across the three structural PK models simulated (Figure S1 in Supplementary Data S1). Further, the spaghetti plots were more scattered than the real data (Figure 1) due to the larger simulated RUV (30%).

Of note, BE could not always be assessed when using one model from the pool of candidates, especially when fitting the LAG_0-order_2-COMPT model and/or fitting data simulated with the TRANSIT_1-COMPT model. Indeed, the additional steps to calculate $\hat{\beta}_{C_{max}}^T$ and the associated SE relies on Monte Carlo (MC) calculations (see Appendix A) and numerical issues arised especially in these two cases (Table S5 in Supplementary Data S1).

Figure 2 displays the estimation errors of the treatment effect on AUC and C_{max} under H_0 and H_1 . For both AUC and C_{max} , the treatment effect coefficients were estimated without bias using each model from the pool of candidates, MS or MA with and without the simulated model included in the pool (ie, TRANSIT_1-COMPT, TRANSIT_2-COMPT or LAG_0-ORDER_2-COMPT), with the exception of $\hat{\beta}_{C_{max}}^T$ using TRANSIT_1-COMPT and TRANSIT_2-COMPT models to analyze the data even when these two models were the models that were used to generate the PK data. These results could be explained by the sparse design, which may be sub-optimal for TRANSIT models. To illustrate, the TRANSIT_2-COMPT model was used to estimate seven PK parameters using only six sampling times. Conversely, a rich design with more sampling times in the absorption phase performed better with the TRANSIT models (see Figure S3 in Supplementary Data S1). Indeed, the estimation errors for $\hat{\beta}_{C_{max}}^T$ using the TRANSIT models varied from -109 to 38 .

Figure 3 displays the proportion of datasets for each scenario where models from the pool of candidates (including or excluding the simulated model) are selected. The probability to select the simulated model was 0.28, 0.41 and 0.48 when simulating with the TRANSIT_1-COMPT, TRANSIT_2-COMPT and LAG_0-ORDER_2-COMPT model under H_0 , respectively, and 0.39 when simulating with the TRANSIT_2-COMPT model under H_1 . MS appeared driven first by the number of compartments and second by the type of absorption. Indeed, when data were simulated with the TRANSIT_2-COMPT model and the later was not included in the pool of candidates, models with two compartments were preferentially selected whereas when data were simulated with the TRANSIT_1-COMPT model and the later was not included in the pool of candidates, then the TRANSIT_2-COMPT model was preferentially selected.

Figure 4 displays the distribution of model averaging weights for the five models whether the simulated model is included or excluded from the candidates pool. In 68% and 73% of the datasets simulated under H_0 and H_1 , respectively, MA was equivalent to MS with one of the model from the pool of candidates having a weight close to 1; a mixture of two models was observed in 25% and 24% of the datasets and a mixture of more than two models in 7% and 4%.

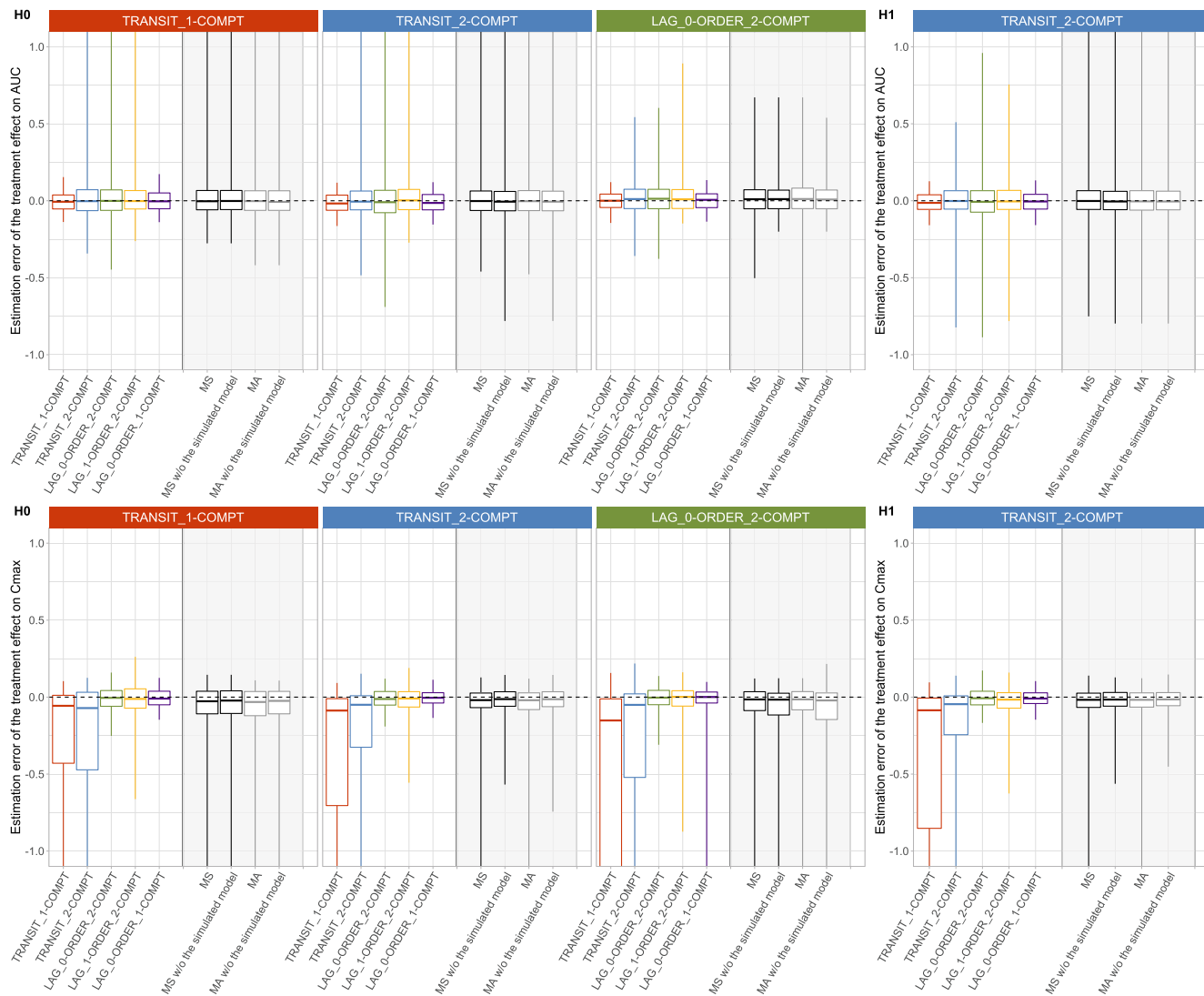
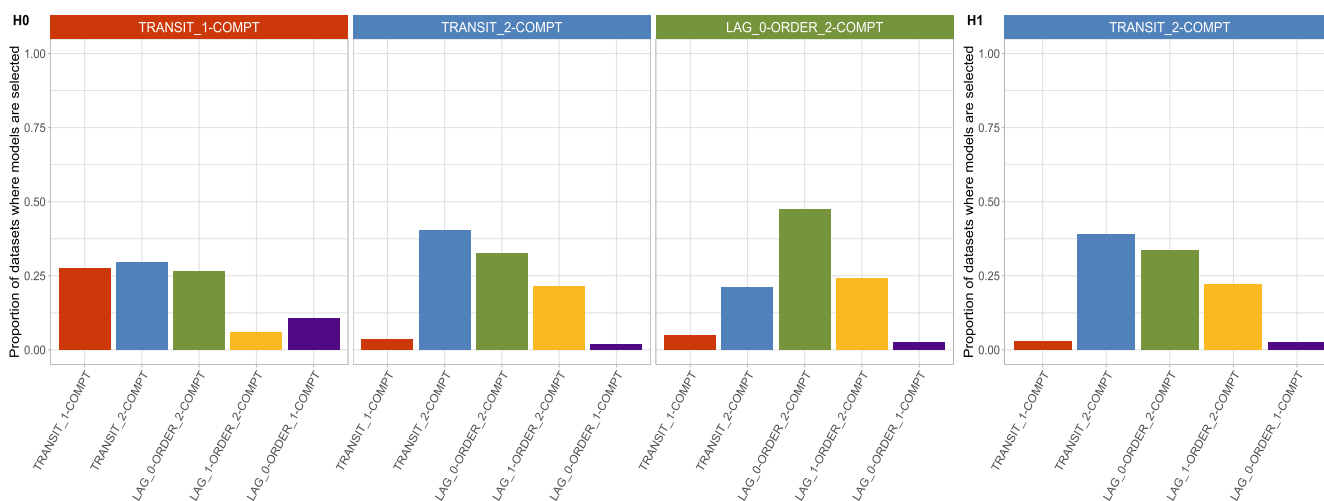


FIGURE 2 Boxplot (2.5th, 25th, 50th, 75th and 97.5th percentiles) of the estimation errors of the treatment effect on AUC (top) and C_{max} (bottom), for the five models from the pool of candidates, model selection (MS) and averaging (MA) with and without (w/o) the simulated model when simulating with the TRANSIT_1-COMPT, the TRANSIT_2-COMPT and the LAG_0-ORDER_2-COMPT model under H_0 and with the TRANSIT_2-COMPT model under H_1 . Note: 3.6% and 9.6% of the estimation errors were > 1 for AUC and C_{max} , respectively, on datasets simulated with the TRANSIT_1-COMPT model under H_0 and 3.4% and 6.7% on datasets simulated with the TRANSIT_2-COMPT model under H_1 .

Table 4 displays MB-TOST type I error rates when simulating with the TRANSIT_1-COMPT, the TRANSIT_2-COMPT and the LAG_0-ORDER_2-COMPT model under H_0 for AUC and C_{max} when analyzing the data using each of the five models, MS or MA with or without the simulated model in the pool of candidates. Likewise, Table 4 also displays the power estimates when simulating with the TRANSIT_2-COMPT model under H_1 . For AUC, the type I error rates were not significantly different from 0.05 regardless of the structural PK model used to fit the data. However, the type I error rates for C_{max} were sometimes significantly lower than 0.05. Overall, MA and MS, with the simulated model included or excluded from the pool of candidate, in general performs similarly in terms of type I error rate and power. The highest power estimates were obtained with the LAG_0-order_1-COMPT model even if it was not the simulated model. Using MS and MA led to similar power as using the simulated model for AUC and greater power for C_{max} .

MS when performed on both the R and T product arms obtained comparable results to MS on the R product arms only in terms of proportion of selected model (Figure S2 in Supplementary Data S1), type I error rates and powers (Table S6 in Supplementary Data S1) in our simulations where T was simulated to have the same structural model as R.

With the simulated model included in the pool of candidate models



With the simulated model excluded from the pool of candidate models

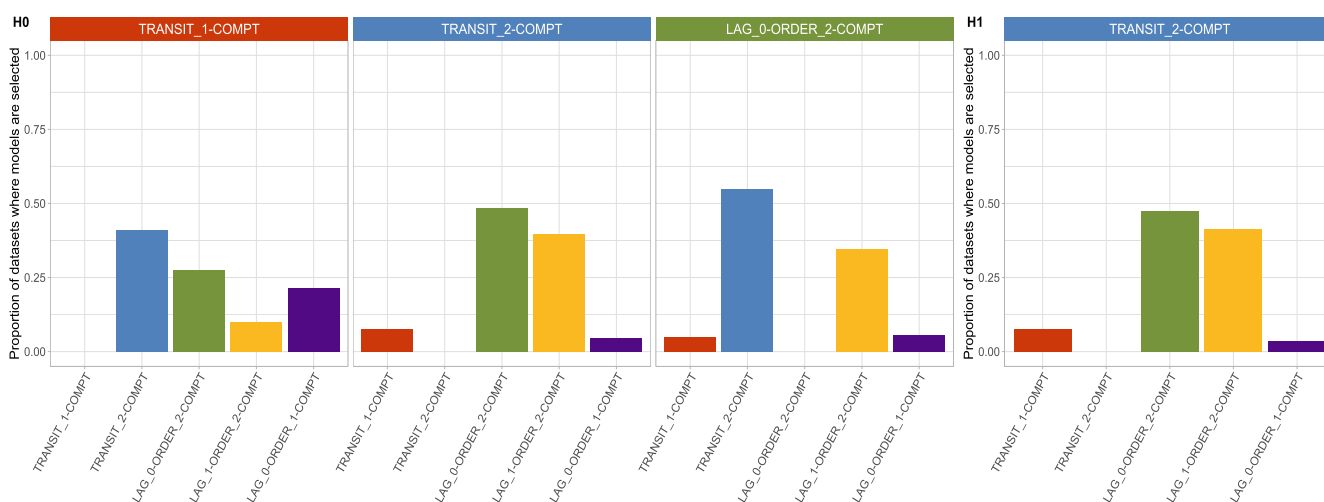


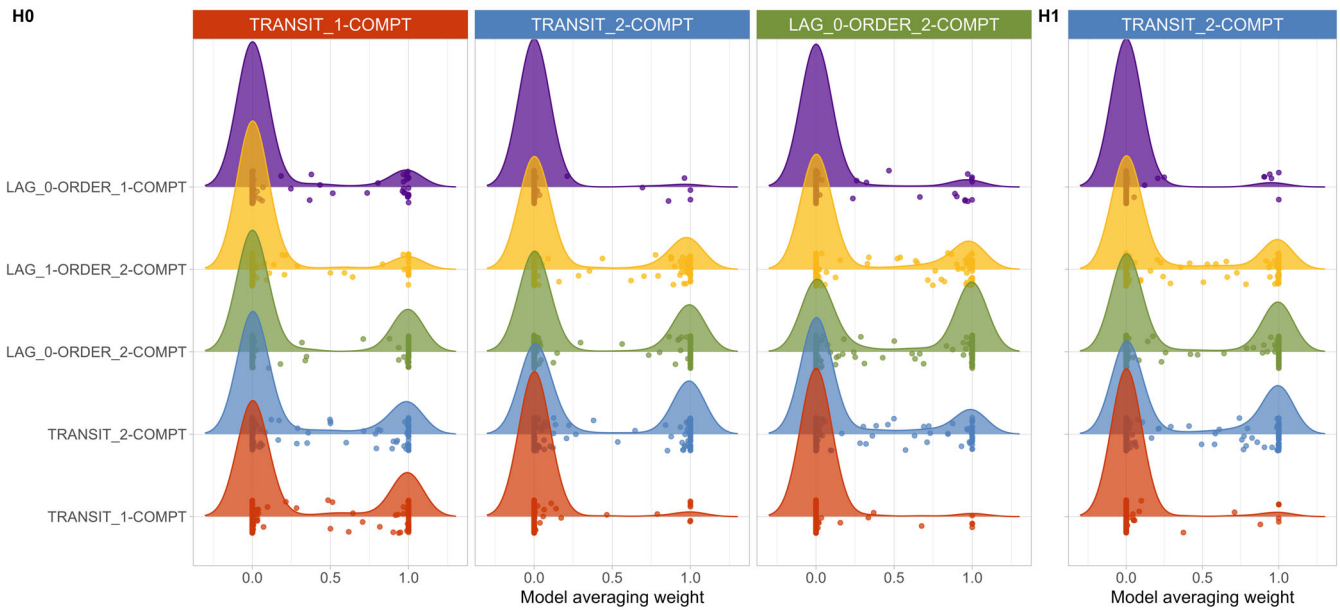
FIGURE 3 Proportion of datasets where the five candidate models were selected, with the simulated model included or excluded from the pool of candidates, when simulating with the TRANSIT_1-COMPT, the TRANSIT_2-COMPT and the LAG_0-ORDER_2-COMPT model under H_0 and with the TRANSIT_2-COMPT model under H_1 .

5 | DISCUSSION

In the present work, we applied MS and MA approaches to the MBBE analysis of a two-way crossover study conducted by Servier. The MB analysis of the original data with rich and sparse sampling led to consistent estimates and conclusions compared to the NCA analysis. Of note, this BE phase I study was meant for illustrative purposes. Indeed, the use of the MBBE approach is intended to be an alternative when the NCA-based approach may not be feasible, such as in BE studies involving special populations. For instance, the collection of rich PK sampling may not be feasible or ethical in BE studies involving children, oncology patients, or immunocompromised patients.²⁴

In our simulations, we evaluated the performance of MB-TOST in presence of a pool of five structural PK model candidates. We considered the combinations of different types of distribution (one vs. two compartments) but also different types of absorption (first or zero-order and delay captured with a lag time or transit compartments), the later being of interest when comparing different drug formulations. First, we showed that MS and MA led to type I errors not significantly different from 0.05 while ensuring a reasonable power. Actually, we did not observe the type I error inflation reported by Dubois et al⁶ using the simulated model with a similar sample size or reported by Guhl et al¹² in the case of model misspecification. Both attributed the type I error inflation to an underestimation of the standard error of the treatment effect. Here, we did not use D-optimal design to define the sampling times in the sparse scenario as performed by Dubois et al⁶

With the simulated model included in the pool of candidate models



With the simulated model excluded from the pool of candidate models

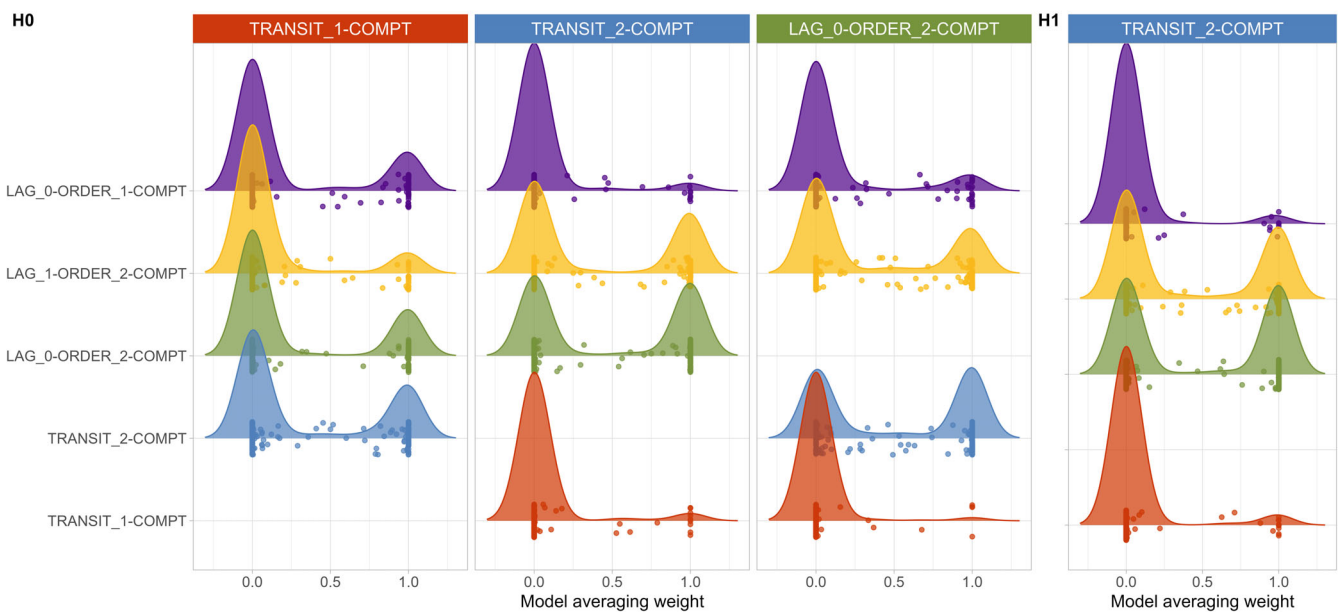


FIGURE 4 Distribution of model averaging weights for the five models, with the simulated model included or excluded from the pool of candidates, when simulating with the TRANSIT_1-COMPT, the TRANSIT_2-COMPT and the LAG_0-ORDER_2-COMPT model under H_0 and with the TRANSIT_2-COMPT model under H_1 .

and Guhl et al.¹² This choice likely led to large standard errors preventing a type I error inflation and reducing power. The absence of type I error inflation may also be due to the high variability of the simulated model (WSV = 30%). Indeed, the power of a two-way crossover study is linked to WSV, likewise type I error rate. Second, we showed that the power of MS and MA was similar to that of the simulated model for AUC and better for C_{max} , even though it was lower than using a couple of PK models (with fewer parameters) which were not used to simulate the data. Actually, when using the simulated model we obtained the lowest power because the sampling scheme was sub-optimal for the TRANSIT_2-COMPT model which requires the estimation of seven PK parameters. In addition, the power of limited (sparse) data of a small study (40 subjects) for estimation of seven PK parameters is lower than that when estimating fewer parameters. Indeed, when simulating with a rich sampling design (as in the real case study data) the TRANSIT_2-COMPT model led to

TABLE 4 Type I error rates and powers with their 95% confidence interval for AUC and C_{\max} for the five models, model selection (MS) and averaging (MA) with and without (w/o) the simulated model in the pool of candidates, when simulating with the TRANSIT_1-COMPT, the TRANSIT_2-COMPT and the LAG_0-ORDER_2-COMPT model under H_0 and with the TRANSIT_2-COMPT model under H_1 .

	Under H_0						Under H_1	
	TRANSIT_1-COMPT		TRANSIT_2-COMPT		LAG_0-ORDER_2-COMPT		TRANSIT_2-COMPT	
	AUC	C_{\max}	AUC	C_{\max}	AUC	C_{\max}	AUC	C_{\max}
TRANSIT_1-COMPT	5.5 [2.8, 9.7]	<i>1.5 [0.3, 4.3]</i>	4.0 [1.7, 7.7]	<i>1.5 [0.3, 4.3]</i>	3.0 [1.1, 6.4]	<i>0.5 [0.0, 2.8]</i>	79.0 [72.7, 84.4]	36.5 [29.8, 43.6]
TRANSIT_2-COMPT	5.6 [2.8, 9.8]	<i>0.5 [0.0, 2.8]</i>	3.5 [1.4, 7.1]	<i>1.0 [0.1, 3.6]</i>	5.5 [2.8, 9.6]	<i>1.5 [0.3, 4.3]</i>	55.5 [48.3, 62.5]	38.0 [31.2, 45.1]
LAG_0-ORDER_2-COMPT	4.5 [1.8, 9.0]	2.6 [0.7, 6.4]	3.2 [1.2, 6.8]	3.2 [1.2, 6.8]	4.1 [1.8, 8.0]	<i>1.0 [0.1, 3.7]</i>	46.2 [38.9, 53.7]	61.3 [53.9, 68.3]
LAG_1-ORDER_2-COMPT	3.8 [1.5, 7.6]	2.2 [0.6, 5.4]	4.0 [1.8, 7.8]	2.5 [0.8, 5.8]	3.6 [1.5, 7.3]	2.6 [0.8, 5.9]	52.0 [44.8, 59.2]	52.5 [45.3, 59.6]
LAG_0-ORDER_1-COMPT	6.0 [3.1, 10.2]	4.2 [2.1, 8.4]	7.0 [3.9, 11.5]	5.0 [2.4, 9.0]	6.0 [3.2, 10.3]	2.0 [0.6, 5.1]	80.2 [74.3, 85.8]	82.5 [76.5, 87.5]
MS	5.5 [2.8, 9.6]	2.0 [0.5, 5.0]	4.5 [2.1, 8.4]	<i>1.5 [0.3, 4.3]</i>	4.5 [2.1, 8.4]	<i>1.5 [0.3, 4.3]</i>	56.0 [48.8, 63.0]	57.5 [50.3, 64.4]
MS w/o the simulated model	5.5 [2.8, 9.6]	2.0 [0.5, 5.0]	4.0 [1.7, 7.7]	3.0 [1.1, 6.4]	4.5 [2.1, 8.4]	<i>1.5 [0.3, 4.3]</i>	56.5 [49.3, 63.5]	63.0 [55.9, 69.7]
MA	6.0 [3.1, 10.2]	<i>0.5 [0.0, 2.8]</i>	4.0 [1.7, 7.7]	<i>1.0 [0.1, 3.6]</i>	4.0 [1.7, 7.7]	<i>1.5 [0.3, 4.3]</i>	54.5 [47.3, 61.5]	52.0 [44.8, 59.1]
MA w/o the simulated model	5.5 [2.8, 9.6]	<i>0.5 [0.0, 2.8]</i>	4.0 [1.7, 7.7]	2.5 [0.8, 5.7]	3.5 [1.4, 7.1]	<i>1.5 [0.3, 4.3]</i>	55.5 [48.3, 62.5]	58.5 [51.3, 65.4]

Note: Estimates with their 95% confidence interval in italic indicate that the type I error is significantly lower than 0.05 according to an exact two-sided binomial test with 5% type I error.

unbiased and precise estimates (Figure S3 in Supplementary Data S1) and similar power estimates to that of the other models (71% for AUC and 80% for C_{\max} in Table S8 in Supplementary Data S1). Of note, two compartment models with rich sampling better described C_{\max} , but bias at the parameter level due to model misspecification was similar in both arms, consequently the estimated $\hat{\beta}_{C_{\max}}^T$ were hardly impacted (see Figure S4 in Supplementary Data S1).

The challenging simulation study design (sampling scheme and high variabilities) also led to the selection of the simulated model in only 28%, 41% and 48% of the cases when simulating with the TRANSIT_1-COMPT, TRANSIT_2-COMPT and LAG_0-ORDER_2-COMPT model under H_0 , respectively, and 39% of the cases when simulating with the TRANSIT_2-COMPT model under H_1 . With regard to the BSV and WSV, our simulation settings were close to the threshold for highly variable drugs set by the regulatory agencies with WSV of approximately 30% for AUC and C_{\max} (Table S9 in Supplementary Data S1). However, the reference scaled BE approach¹ for highly variable drugs is not applicable for the two-way crossover study design considered in our simulations. Moreover, our simulation settings of RUV at 30% led to low power estimates notably for C_{\max} , which estimation depends on a subset of samples, whereas AUC is based on an average over all concentrations. We chose a challenging combination of high values for the variabilities but, arguably, we did not consequently increase the number of subjects to achieve the usually targeted power of 80%.

Here, as in previous works, we simulated and fitted the data with the same model for the R and T product arms. We deem this hypothesis reasonable as the treatments under comparison in BE studies are expected to behave similarly in terms of processes of absorption (delayed or not), elimination (linear or non linear) and distribution (number of compartments) at least for small molecules. However, we recognise the interest of exploring the performance of MS and MA prior to MB-TOST when simulating the R and T product arms with different PK models. Yet, we believe the PK model selection should only be based on R product arm data. If MS is also performed on T product arm data, it could inflate the overall BE assessment type I error rate. With MA, no model is selected, therefore the overall BE assessment type I error rate is mostly spent on MB-TOST when R and T have the same PK model. It would also be interesting to investigate the impact of other model misspecifications, error, and variability models on the performance of MS and MA as well as BE assessment.

Further, we estimated a treatment effect on all absorption parameters, apparent volumes and clearances instead of using a scaling parameter capturing the drug bioavailability (F). This approach is more costly in terms of number of parameters to estimate but we showed it can be more flexible.¹²

Finally, no added value was observed with MA compared to MS because, in most datasets, one model had a weight equal to 1. Mathematically, a model is assigned a weight of 1 (and the others a weight of 0.05% or less, the threshold used in this work see Section 2.4) when it has an AIC lower by ten points. We hypothesize that such small AIC differences can be expected when comparing structural models with one parameter fixed to different values. Here the PK models were too different with regard to the absorption and distribution processes yielding non negligible differences in log likelihood.

Further for any additional structural parameter, we added six estimates (fixed effect, treatment, period and sequence coefficients, BSV and WSV). Thus, prior to MB-TOST, we propose to use MS on the R product arm data in the first place and consider MA only if the differences in AIC are below ten points, presuming differences of the same magnitude will be observed on R and T product arm data. Buatois et al¹⁵ showed the superiority of MA for model-based dose-response studies in at least one of their simulation study scenarios. The discrepancy between our conclusions and those of Buatois et al¹⁵ emphasizes the need for further assessment of MA and MS in MIDD under various scenarios and for different purposes in drug development.¹⁰

ACKNOWLEDGEMENTS

This project was funded by the FDA under the contract 75F40119C10111. The data came from a study funded by Servier. The authors are grateful to all the collaborators of the project “Evaluation of model-based bioequivalence (MBBE) statistical approaches for sparse design PK studies” under the contract 75F40119C10111 for their expertise.

CONFLICT OF INTEREST STATEMENT

This work reflects the views of the authors and should not be construed to represent the FDA’s views or policies.

DATA AVAILABILITY STATEMENT

Individual clinical data are not shared. The R codes used to perform the simulation study are provided in Supplementary Data S2. The real data that support the findings of this study are not publicly available due to privacy restrictions.

ORCID

Morgane Philipp  <https://orcid.org/0000-0001-9827-316X>

REFERENCES

1. FDA. Bioequivalence studies with pharmacokinetic endpoints for drugs submitted under an ANDA. <https://www.fda.gov/media/87219/download> 2021
2. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm.* 1987;15(6):657-680. doi:10.1007/bf01068419
3. FDA. Guidance for industry-Statistical approaches to establishing bioequivalence. <https://www.fda.gov/media/163638/download> 2022
4. EMA. Guideline on the investigation of bioequivalence. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf 2010
5. FDA. Guidance for industry-Bioequivalence: blood level bioequivalence study. <https://www.fda.gov/media/89840/download> 2016
6. Dubois A, Lavielle M, Gsteiger S, Pigeolet E, Mentré F. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. *Stat Med.* 2011;30(21):2582-2600. doi:10.1002/sim.4286
7. Tardivon C, Loingeville F, Donnelly M, et al. Evaluation of model-based bioequivalence approach for single sample pharmacokinetic studies. *CPT Pharmacometrics Syst Pharmacol.* 2023;12(7):904-915. doi:10.1002/psp4.12960
8. Loingeville F, Bertrand J, Nguyen TT, et al. New model-based bioequivalence statistical approaches for pharmacokinetic studies with sparse sampling. *Am Assoc Pharm Scient J.* 2020;22(6):141. doi:10.1208/s12248-020-00507-3
9. Möllenhoff K, Loingeville F, Bertrand J, et al. Efficient model-based bioequivalence testing. *Biostatistics.* 2020;23(1):314-327. doi:10.1093/biostatistics/kxaa026
10. ICH. Output from ICH model-informed drug development (MIDD)-discussion group (DG) 2021. <https://www.ich.org/page/reflection-papers> 2021
11. FDA. Population pharmacokinetics-Guidance for industry. <https://www.fda.gov/media/128793/download> 2022
12. Guhl M, Mercier F, Hofmann C, et al. Impact of model misspecification on model-based tests in PK studies with parallel design: real case and simulation studies. *J Pharmacokinet Pharmacodyn.* 2022;49(5):557-577. doi:10.1007/s10928-022-09821-z
13. Pinheiro J, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. *Stat Med.* 2014;33(10):1646-1661. doi:10.1002/sim.6052
14. 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;44(6):581-597. doi:10.1007/s10928-017-9550-0
15. Buatois S, Ueckert S, Frey N, Retout S, Mentré F. Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. *Am Assoc Pharm Scient J.* 2018;20(3):56. doi:10.1208/s12248-018-0205-x
16. Uster DW, Stocker SL, Carland JE, et al. A model averaging/selection approach improves the predictive performance of model-informed precision dosing: vancomycin as a case study. *Clin Pharmacol Therapeut.* 2021;109(1):175-183. doi:10.1002/cpt.2065
17. Zhao L, Kim M, Zhang L, Lionberger R. Generating model integrated evidence for generic drug development and assessment. *Clin Pharmacol Therapeut.* 2019;105(2):338-349. doi:10.1002/cpt.1282

18. Donazzolo Y, Lehnick D, Dilger C, Kärcher U. Bioequivalence study of one tablet of the fixed combination of perindopril arginine 10 mg/amlodipine 10 mg versus one tablet of perindopril tert-butylamine 8 mg plus one tablet of amlodipine 10 mg, after single oral dose, in healthy volunteers. An open-label randomised two-period cross-over study in healthy volunteers. 2006 Protocol number: PKH-05985-001. Servier internal report
19. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464. doi:10.1214/aos/1176344136
20. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics*. 1997;53(2):603-618. doi:10.2307/2533961
21. Turek D, Fletcher D. Model-averaged wald confidence intervals. *Comput Stat Data Anal*. 2012;56(9):2809-2815. doi:10.1016/j.csda.2012.03.002
22. Lavielle M, Moulines E. A simulated annealing version of the EM algorithm for non-Gaussian deconvolution. *Stat Comput*. 1997;7(4):229-236. doi:10.1023/a:1018594320699
23. Rohatagi S, Carrothers TJ, Kshirsagar S, Khariton T, Lee J, Salazar D. Evaluation of population pharmacokinetics and exposure-response relationship with coadministration of amlodipine besylate and olmesartan medoxomil. *J Clin Pharmacol*. 2008;48(7):823-836. doi:10.1177/0091270008317847
24. Panetta JC, Iacono LC, Adamson PC, Stewart CF. The importance of pharmacokinetic limited sampling models for childhood cancer drug development. *Clin Cancer Res*. 2003;9(14):5068-5077. PMID:14613983

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Philipp M, Tessier A, Donnelly M, et al. Model-based bioequivalence approach for sparse pharmacokinetic bioequivalence studies: Model selection or model averaging?. *Statistics in Medicine*. 2024;1-14. doi: 10.1002/sim.10088

APPENDIX A. ESTIMATION OF THE TREATMENT EFFECT ON AUC AND C_{max} AND THEIR ASSOCIATED SE

We calculated $\hat{\beta}_{SP}^T$ as a function $h(\hat{\mu}^-, \hat{\beta}^{T-})$ with μ^- a subset of the vector of structural PK model parameter fixed effects and β^{T-} a subset of the vector of treatment effect coefficients.

When h had an analytical form, $\hat{\beta}_{SP}^T$ was derived analytically and we used the delta method to calculate $SE(\hat{\beta}_{SP}^T)$. Otherwise, we performed Monte Carlo (MC) calculations. A total of $k_{MC} = 1, \dots, K_{MC}$ parameters values were sampled from a multi-normal with mean $(\hat{\mu}^-, \hat{\beta}^{T-})$ and variance $\widehat{VAR}(\hat{\mu}^-, \hat{\beta}^{T-})$. Then K_{MC} reference and treatment concentration profiles were calculated to derive K_{MC} reference and treatment SP of interest. Minimum values for $\hat{\mu}^-$ (μ_{min}) were set to enable realistic concentration profiles. Finally, $\hat{\beta}_{SP}^T$ and $SE(\hat{\beta}_{SP}^T)$ were estimated by computing the mean and the standard deviation of the K_{MC} differences in the log SP of interest for the reference and test product arm data ($= \log(SP^T) - \log(SP^R)$). A minimum number of MC samples $K_{MC,min}$ was required, otherwise BE can not be assessed.

We performed $K_{MC} = 1000$ MC samples using the function `mvrnorm` of the package `MASS`, setting the minimum number of samples $K_{MC,min}$ to 800, μ_{min} was set to 1 for apparent volumes and to 0.01 for other parameters.