Calcul de l'incertitude à distance finie dans les modèles non linéaires à effets mixtes

Mélanie Guhl

Université Paris Cité
Ecole Doctorale 393 Pierre Louis de Santé Publique
Epidémiologie et Sciences de l'Information Biomédicale

Institut national de la santé et de la recherche médicale
Laboratoire IAME UMR 1137- Infection, Antimicrobiens, Modélisation, Evolution
Equipe BIPID - Modélisation Biostatistique, Pharmacométrie et Investigation
Clinique en Maladies Infectieuses

Calcul de l'incertitude à distance finie dans les modèles non linéaires à effets mixtes

Mélanie GUHL

Thèse de doctorat de Biostatistique et Biomathématiques

Dirigée par le Dr Emmanuelle COMETS

Presentée et publiquement défendue le 24 Mai 2024

Pr Chantal GUIHENNEUC	PU, Université Paris Cité	Présidente
Dr Estelle KUHN	DR, INRAE	Rapportrice
Dr David MAKOWSKI	DR, INRAE	Rapporteur
Dr Julie BERTRAND	CR-HDR, Université Paris Cité	Examinatrice
Dr Frédéric BOIS	HDR, Certara	Examinateur
Pr Nicolas GRÉGOIRE	PU-PH, Université de Poitiers	Examinateur
Dr Emmanuelle COMETS	CR-HDR, Université Paris Cité	Directrice

L'incertitude est le pire de tous les maux jusqu'au moment où la réalité vient nous faire regretter l'incertitude.

– Alphonse Karr

Remerciements

Je remercie tout d'abord les membres du jury d'avoir accepté d'évaluer cette thèse, le Dr Frédéric Bois, le Pr Nicolas Grégoire et le Pr Chantal Guihenneuc, et en particulier les rapporteurs, le Dr Estelle Kuhn et le Dr David Makowski, pour leurs retours constructifs sur le manuscrit.

Je remercie évidemment chaleureusement mes encadrantes, le Dr Emmanuelle Comets et le Dr Julie Bertrand, pour ces quatre dernières années. Merci de m'avoir fait confiance et de m'avoir soutenue et guidée à travers les tumultes de la thèse. J'ai beaucoup appris grâce à vous, sur tous les plans. Emmanuelle, pour ta patience et ton humilité. Julie, pour ta curiosité et ta joie de vivre à toute épreuve. Merci pour votre disponibilité et votre implication, c'est un plaisir de travailler avec vous.

Je remercie également le Dr France Mentré et le Dr Jérémie Guedj, pour vos nombreux conseils. France, pour ton aide pendant mon stage et tout le projet MBBE, et ton avis toujours précieux. Jérémie, pour tes retours sur la modélisation des données Discovery, et pour tes conseils sur la vie en général.

Merci à Lucie Fayette, pour ton implication dans les projets de cette thèse. Comme première expérience d'encadrement, travailler avec quelqu'un d'aussi autonome et mature, c'est plutôt agréable ! Merci d'avoir partagé ma frustration pendant ces quelques mois et de m'avoir fait sentir moins seule dans cette galère...

Je remercie aussi le Dr Maud Delattre, pour tes conseils (plus déterminants pour mon travail que tu ne le penses !) et ton expertise. Merci également à tous les membres de la FDA impliqués dans le projet MBBE, en particulier le Dr Satish Sharan et le Dr Mark Donnelly, pour votre collaboration, ainsi qu'au Dr François Mercier, non seulement pour ton implication dans le projet MBBE, mais également pour ton enthousiasme et ta réactivité lorqu'on t'a proposé de poursuivre cette collaboration dans les projets suivants.

J'aimerais aussi remercier le Pr Myriam Fradon et le Pr François Coquet, dont j'ai eu la chance de croiser la route, et sans qui je n'aurais pas pris les mêmes décisions. Vous m'avez encouragée, chacun votre tour, à oser, à ne pas me contenter, quand je n'étais pas encore capable de l'envisager par moi-même, et il est évident pour moi aujourd'hui que vous aviez raison.

Merci à l'ensemble de l'équipe BIPID et de l'unité IAME, à ses membres passés et présents. Merci à Houda, Stephan, Zaïna et Myriam pour l'organisation, pas toujours facile, surtout face à des gens pas toujours très organisés... Merci à Hervé, Lionel et Rémy pour le support informatique et la gestion du centre de calcul, on l'aime autant qu'on le malmène.

A tous les docs, postdocs, stagiaires et autres compagnons de pause, qui dans un bureau, qui autour d'un café, d'un thé, d'une clope, d'une pizza, d'une pinte, ou plutôt maintenant d'une boisson certifiée sans alcool, qui à la Mercerie, au Enkore, au Ground

Control, sur mon canap ou le vôtre, bien cachés dans les remparts de Saint-Malo ou à l'autre bout de l'Europe aux frais du contribuable : Antonio, Jérémy, François, Jinju, Claire, Drifa, Emilie, Coralie, Romain, Cédric, Marion, Guillaume, Nadège, Mathieu, Arthur, Aurélien, Alexandra, Selma, Antoine, Morgane, Lucie, Sophie, Niels, Maxime, Julien, Aloïs, Gaëlle, Ibtissem, Inès, Clément, Aline-Marie, Nicolas, Adrien, Assil, Clarisse, Bach, Carlos, Tom, Davide, j'en oublie certainement, pas trop j'espère.

Marion, merci pour ton humanité, tes contradictions, ton assurance, et les rebondissements perpétuels de ta vie qui me divertissent tant. Alex, Aurélien, merci d'avoir traversé avec moi les saisons 2020 à 2023 de BIPID : beaucoup de cliffhangers, beaucoup de plot twists, un peu déçue quand vous avez quitté la série avant la dernière saison mais bon, vous avez été attirés par les strass et les paillettes, classique, je vous en veux pas.

Selma, il fallait au moins quelqu'un d'aussi sociable que toi (oui oui) pour insister et voir si bien derrière ma carapace de bougonnerie quotidienne. Merci de m'avoir promenée et fait mangé des gâteaux en Suède pendant deux semaines, sans toi je serais encore en train de chercher l'arrêt de bus pour l'aéroport. C'était sympa aussi de me présenter toute l'assistance du PAGE en Slovénie. C'était plus sympa de semer tout le monde pour faire des sudokus dans un vieux café l'année suivante. Merci pour tout, le bureau des plaintes, les fous rires, les conversations sur tout et son contraire. En attendant de t'envoyer les 2000 reels de boomer qui t'attendent dans mes enregistrements (ça me fera des trucs à te dire quand on sera loin), je m'accorde le droit de t'afficher un peu, je fais ce que je veux : bravo pour ta Sheiner et ton fantastic modeling work, t'es une star. Period.

Les copains, les amis, ceux qui sont là, ceux qui sont loin, ceux qui n'en sont plus, ceux qui comprennent, ceux qui essayent, vous avez tous laissé une trace, pris part à mon chemin, été bienveillants. Maxime, toujours un plaisir de te voir évoluer, te retrouver à l'autre bout du monde, recevoir tes cartes postales et chanter du Hubert Lenoir avec toi pendant tes passages éclairs.

Merci à ma famille, de me soutenir depuis toujours. Maman, merci d'avoir toujours tout fait pour que je ne manque jamais de rien et que je sois dans un confort absolu. Et merci de faire semblant de ne pas t'inquiéter maintenant que je passe ma vie à faire exactement le contraire. Papa, merci pour tous les sacrifices que tu as fait pour nous. Si tu pouvais arrêter les blagues à deux balles quand même... ça m'arrangerait. J'espère que vous êtes fiers de moi. Mes grands parents, Violette, Marguerite, Alain, mon cousin Pierre, j'aurais tellement aimé vous montrer tout ça. Jérémie (deuxième du nom), Mathilde, Marie, Malo, Arthur, Charlie, Jérémie (premier du nom), Léo, Noa, Lucie, Sacha, Gaspard (deuxième du nom), Maryse, Jean-Louis, David, Virginie, Gaspard (premier du nom), la famille n'est pas très grande mais qu'est ce que j'ai de la chance de vous avoir, qu'est ce que ça fait du bien à chaque fois que je vous vois. Si je peux me permettre, ça la fout quand même mal d'avoir des gens avec des noms identiques dans la famille, en plus mes collègues savent que je viens du nord, qu'est ce qu'ils vont penser ?

Enfin merci à Tanguy, pour l'aventure. Je crois que je ne cesserai jamais d'être fascinée par ton optimisme, il m'a été d'un grand secours ces dernières années. Merci de me faire rire et de me faire sentir chez moi. Merci aussi à ta famille déjantée, pour leur fougue et leur ouverture d'esprit.

Last but not least ... I wanna thank me.
I wanna thank me for believing in me,
I wanna thank me for doing all this hard work,
I wanna thank me for having no days off,
I wanna thank me for never quitting,
I wanna thank me for always being a giver,
and trying to give more than I receive,
I wanna thank me for trying to do more right than wrong,
I wanna thank me for just being me at all times.
- Snoop Doggy Dogg 2018

Valorisation scientifique

Publications de thèse

— **Guhl Mélanie**, Mercier François, Hofmann Carsten, Sharan Satish, Donnelly Mark, Feng Kairui, Sun Wanjie, Sun Guoying, Grosser Stella, Zhao Liang, Fang Lanyan, Mentré France, Comets Emmanuelle, Bertrand Julie. Impact of model misspecification on model-based tests in PK studies with parallel design : real case and simulation studies. Journal of Pharmacokinetics and Pharmacodynamics, 2022, 49, 557-577. https://doi.org/10.1007/s10928-022-09821-z

— **Guhl Mélanie**, Bertrand Julie, Fayette Lucie, Mercier François,Comets Emmanuelle. Uncertainty computation at finite distance in nonlinear mixed effects models - a new method based on Metropolis Hastings algorithm. AAPS Journal, 2024, 26, 53. https://doi.org/10.1208/s12248-024-00905-x

Communications orales

— **Journées de Biostatistiques**, 1er-2 octobre 2020 (Paris, France), "Approches statistiques par modélisation pour les études de bioéquivalence pharmacocinétique avec données éparses"

— **52èmes Journées de Statistique de la Société Française de Statistique (SFdS)**, 7-11 Juin 2021 (virtuel), "Tests d'équivalence pharmacocinétique par modélisation : impact d'une mauvaise spécification du modèle"

— **Pharmacometrics in France**, 29 Septembre 2023 (Paris, France), "Computation of parameter uncertainty at finite distance in nonlinear mixed effects models"

— **Séminaire invité**, Bayesian Modeling Research Group, Finnish Center for Artificial Intelligence, Aalto University (Aalto, Finlande), 19 Octobre 2023, "Computation of parameter uncertainty at finite distance in nonlinear mixed effects models"

— **Séminaire invité**, Division of Systems Pharmacology and Pharmacy (SPP), Leiden University (Leiden, Pays-Bas), "Computation of parameter uncertainty at finite distance in nonlinear mixed effects models"

Communications affichées

— **42nd Annual Conference of the International Society for Biostatistics (ISCB)**, 18-22 Juin 2021 (virtuel), "Impact of model misspecification on model-based bioequivalence"

— **29th Population Approach Groupe Europe (PAGE) Meeting**, 2-3 et 6-7 Septembre 2021 (virtuel), "Model-based analysis of PK equivalence : assessing the impact of model misspecification"

— **30th Population Approach Groupe Europe (PAGE) Meeting**, 28 Juin - 1er Juillet 2022 (Ljubljana, Slovénie), "Computation of standard errors at finite distance in non linear mixed effects models"

— **31st Population Approach Groupe Europe (PAGE) Meeting**, 28-30 Juin 2023 (La Corogne, Espagne), "Uncertainty computation at finite distance in nonlinear mixed effects models - a new method evaluated on simulations and applied to the evolution of clinical status of patients hospitalised for COVID-19"

— **44nd Annual Conference of the International Society for Biostatistics (ISCB)**, 27-31 Août 2023 (Milan, Italie) : "Uncertainty computation at finite distance in nonlinear mixed models : evaluation of a new Bayesian method"

Table des matières

Liste des acronymes

ABC Approximate Bayesian computation
ANCOVA Analyse de covariance
ANOVA Analyse de variance
ANSM Agence Nationale de Sécurité du Médicament et des produits de santé
Asympt Méthode asymptotique de calcul des SE basée sur l'information de Fisher
AUC Aire sous la courbe
BIC Critère d'information Bayésien
Cmax Concentration maximale
EBE Estimateurs de Bayes empiriques
ECMO Oxygénation par membrane extracorporelle
EM Expectation-Maximisation
EMV Estimateur du maximum de vraisemblance
ESS Taille d'échantillon effective
FDA U.S. Food and Drugs Administration
FIM Matrice d'information de Fisher
GLM Modèles linéaires généralisés
GMR Ratio des moyennes géométriques
HMC Monte Carlo Hamiltonien
IS Importance sampling
MAP Maximum *a posteriori*
MBBE Bioéquivalence par modélisation
MCEM Monte Carlo Expectation-Maximisation
MCMC Monte Carlo par chaînes de Markov
MH Metropolis Hastings
NCA Approche non compartimentale
NEWS National Early Warning Score
NLMEM Modèles non linéaires à effets mixtes
OMS Organisation mondiale de la santé
Post Méthode semi-Bayésienne de calcul des SE basée sur l'algorithme HMC
RSE Erreurs standards relatives
SAEM Stochastic Approximation of Expectation-Maximisation
SE Erreurs standards
SIR Sampling Importance Resampling
SoC Soins standards
TOST Two One Sided Test

Chapitre 1

Introduction

Les modèles non linéaires à effets mixtes (NLMEM) sont des outils fréquemment utilisés pour modéliser les données longitudinales dans la recherche clinique. Les travaux de cette thèse s'inscrivent dans le but global d'améliorer le calcul de l'incertitude sur les paramètres estimés dans ces modèles, et notamment dans le cas de petits échantillons, ce qui en pratique arrive fréquemment dans les jeux de données pour lesquels ces modèles sont adaptés.

Après avoir défini en détails le type de données et de modèles auxquels on s'intéresse, cette introduction expose les différentes méthodes statistiques existantes pour les mettre en oeuvre, dans un contexte Bayésien et fréquentiste, puis présente deux cas de figures dans lesquels le calcul de l'incertitude en particulier est une question primordiale : les études de bioéquivalence et les études d'effet traitement. Ces deux exemples d'application ont un point commun : l'estimation d'un effet traitement et l'importance de quantifier correctement l'incertitude autour de cette estimation, afin de pouvoir utiliser les résultats des analyses effectuées comme aide à la prise de décision dans un contexte de développement de médicaments ou de recherche de traitement efficace pour combattre une maladie.

Enfin, les objectifs ayant structuré les différents projets menés au cours de cette thèse sont présentés à la fin de ce chapitre.

1.1 Modélisation des données longitudinales dans les essais cliniques

1.1.1 Données cliniques longitudinales

Les données longitudinales sont des données répétées dans le temps : pour chaque groupe d'intérêt, en général un individu, on mesure à plusieurs reprises la quantité d'intérêt, ce qui permet d'observer son évolution au cours du temps. Ces données sont de plus en plus collectées dans les essais cliniques car elles permettent de tester l'effet d'un traitement, d'une comorbidité ou d'une autre covariable sur l'évolution observée (Albert, 1999).

Dans la plupart des essais cliniques, le critère de jugement reste une donnée observée ponctuelle, mais de nombreux tests et modèles ont été développés ces dernières décennies pour prendre en compte les données longitudinales. Celles-ci présentent une information supplémentaire permettant d'interpréter voir d'anticiper l'évolution d'un patient ou d'un biomarqueur, la nécessité d'un changement de traitement, d'une sortie d'étude clinique, ou autre. Ces données peuvent être continues (par exemple, des concentrations de médicaments ou des charges virales) ou discrètes (par exemple, des scores catégoriels reflétant le statut clinique du patient ou le stade de la maladie).

La particularité de ces données est qu'elles sont corrélées entre les différents temps de mesure pour chaque individu. De plus, on observe fréquemment des données manquantes et/ou des temps de mesures différents entre les individus, conduisant à un design déséquilibré.

Pour analyser ce type de données, il faut des méthodes capables de prendre en compte les corrélations entre les observations, ce qui permet de minimiser la variabilité non expliquée, maximiser la puissance statistique et donc diminuer la taille d'échantillon requise pour les inférences (Miot, 2023). Pour inférer sur ces données, on peut utiliser des tests (dans le cas où l'on a uniquement deux observations par sujet par exemple,

un test de Student, de Wilcoxon ou de Cochran pour données appariées), ou des analyses de variance (ANOVA) ou de covariance (ANCOVA) répétées. Cependant, ce type d'approches a des limites, notamment en présence de données manquantes ou de design non équilibré (De Livera et al., 2014). L'utilisation des tests statistiques est donc valide si les essais sont rigoureusement contrôlés, avec un design identique pour chaque individu, ce qui se produit rarement en pratique.

Pour pallier le manque de robustesse des tests statistiques aux données manquantes et aux designs non équilibrés, une alternative évidente est l'utilisation de modèles. Cependant, ceux-ci doivent être capables de prendre en compte la structure des corrélations existantes entre les données observées, car dans un modèle de régression classique, on enfreindrait l'hypothèse d'indépendance entre les résidus calculés.

1.1.2 Modèles non linéaires à effets mixtes

Les modèles dits à effets mixtes permettent d'analyser les données longitudinales en évaluant l'évolution observée à l'échelle de la population mais aussi à l'échelle individuelle en introduisant une variabilité inter-individuelle voire intra-individuelle. Les modèles à effets mixtes modélisent la variabilité individuelle de manière interprétable. Cependant, si on utilise un modèle à effets mixtes, une hypothèse sous-jacente à cette méthode est que le modèle structurel de la dynamique est le même dans les différents groupes observés (les différents individus mais également les différents groupes de covariables, par exemple on peut supposer que l'effet traitement n'influence pas la forme du modèle utilisé dans un essai clinique avec une covariable de traitement).

Historiquement, les modèles linéaires à effets mixtes ont été développés dans les années 1980 et 1990 pour pallier les limites des analyses de type ANOVA, notamment en termes de structure de variabilité (Verbeke and Molenberghs, 2000).

Par la suite, on a voulu étendre le concept des modèles mixtes à des dynamiques non linéaires. Le développement des modèles non linéaires à effets mixtes (NLMEM) s'est

fait dans le cadre du suivi de routine des patients, dans lequel beaucoup de données longitudinales étaient recueillies à des fins de suivi thérapeutique et on manquait de méthodes adaptées pour les exploiter au mieux (Sheiner et al., 1977; Lindstrom and Bates, 1990; Pinheiro et al., 1994). Ces modèles ont été développés sur des données issues de la pharmacocinétique, c'est-à-dire de l'étude du devenir du médicament dans l'organisme (absorption, distribution, élimination), puis se sont généralisés à la pharmacométrie, c'est-à-dire l'étude quantitative de la pharmacologie clinique, notamment dans les essais cliniques de phase I et II (Lavielle, 2014).

On peut décrire un modèle non linéaire à effets mixtes comme suit :

$$y_{ijk} \sim D_y(x_{ijk}, \psi_{ik}, \xi) \qquad \text{modèle des observations} \tag{1.1}$$

$$\psi_{ik} \sim D_\phi(C_{ik}, \mu, \beta, \Omega) \qquad \text{modèle des paramètres} \tag{1.2}$$

- D_y : distribution (gaussienne, binomiale, ou autre) des observations
- D_ϕ : distribution (gaussienne, log-normale, ou autre) des paramètres
- x_{ijk} : variable expérimentale (temps, régime de dose en pharmacocinétique, ou autre) j de l'individu i à l'occasion k
- y_{ijk} : réponse de l'individu i spécifique à la variable x_{ijk} (i=1...N, j=1,...n_i)
- ψ_{ik} : vecteur des paramètres pour l'individu i à l'occasion k
- ξ : vecteur des paramètres d'erreur
- C_{ik} : matrice des covariables spécifique à l'individu i et l'occasion k
- μ : vecteur des paramètres moyens
- β : vecteur des effets des covariables
- Ω : matrice de variance covariance des effets aléatoires

Le vecteur des paramètres de population à estimer est $\theta = (\mu, \beta, \Omega, \xi)$.

La vraisemblance marginale des données observées y sous ce modèle général peut

s'écrire par le biais d'une intégration sur les paramètres individuels non observés ψ :

$$p(y;\theta) = \prod_{i,k} \int \left(\prod_{j} p(y_{ijk}|\psi_{ik}) \right) p(\psi_{ik};\theta) d\psi_{ik} \tag{1.3}$$

Cas continu gaussien

On peut écrire une version plus spécifique du modèle non linéaire à effets mixtes dans le cas continu gaussien, avec les hypothèses supplémentaires faites sur la distribution des effets aléatoires et des erreurs résiduelles :

$$y_{ijk} = f(x_{ijk}, \psi_{ik}) + g(x_{ijk}, \psi_{ik}, \xi)\ \epsilon_{ijk} \tag{1.4}$$

$$\phi_{ik} = h(\psi_{ik}) = h(\mu) + C_{ik}\beta + \eta_i + \kappa_{ik} \tag{1.5}$$

- $f()$: modèle structurel continu
- $g()$: modèle d'erreur
- $\epsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$: erreur résiduelle
- $h()$: fonction de transformation pour les paramètres individuels
- ϕ_{ik} : vecteur des paramètres pour l'individu i à l'occasion k sur l'échelle transformée
- $\eta_i \sim \mathcal{N}(0,\Omega)$: effets aléatoires inter-individuels
- $\kappa_{ik} \sim \mathcal{N}(0,\Gamma)$: effets aléatoires intra-individuels inter-occasions
- Ω : matrice de variance covariance des effets aléatoires inter-individuels
- Γ : matrice de variance covariance des effets aléatoires inter-occasions

Dans le cas des essais cliniques en parallèle (avec une seule occasion), on peut omettre la variabilité intra-individuelle inter-occasion, ce qui retire une partie des paramètres individuels (κ) et des paramètres de population (Γ) à estimer.

Dans cette écriture, on a supposé que les effets fixes transformés $h(\mu)$ et les effets de covariables β interviennent linéairement dans la définition du paramètre individuel sur l'échelle des ϕ. En pratique, certains algorithmes ont besoin de cette hypothèse mais

d'autres non et il peut arriver que la relation entre ϕ, $h(\mu)$ et β ne soit pas linéaire.

Cas binaire

Dans le cas d'une variable catégorielle $y_{ijk} \in \{c_1, c_2, ..., c_L\}$, le modèle est défini entièrement par sa fonction de masse : $\big(\mathbb{P}(y_{ijk} = c_l)\big)_{l=1,...,L}$. Les modèles linéaires généralisés (GLM, Nelder and Wedderburn (1972)) permettent d'appliquer le concept de la régression linéaire à ce type de variable en transformant les probabilités estimées grâce à une fonction de lien pour les associer à des paramètres.

Pour une variable binaire par exemple, $y_{ijk} = 0$ ou $y_{ijk} = 1$, il suffit de modéliser $\mathbb{P}(y_{ijk} = 1)$:

$$m(\mathbb{P}(y_{ijk} = 1)) = \psi_{ik} \tag{1.6}$$

$$\phi_{ik} = h(\psi_{ik}) = \mu + C_{ijk}\beta + \eta_i + \kappa_{ik} \tag{1.7}$$

avec

- m() : fonction de lien (logit, probit, ou autre)
- $h()$: fonction de transformation pour les paramètres individuels

Contrairement au modèle continu, il n'y a pas de modèle ni de paramètres d'erreur car on modélise directement la distribution des probabilités.

Pour estimer les paramètres des NLMEM, et particulièrement pour prendre en compte les effets aléatoires non observés, les méthodes utilisées dépendent du paradigme dans lequel on se place.

1.2 Paradigme Bayésien

1.2.1 Inférence Bayésienne

Les modèles mixtes peuvent être considérés comme des modèles hiérarchiques dans lesquels les paramètres individuels sont des variables aléatoires latentes.

Dans le paradigme Bayésien, on considère le vecteur des paramètres d'intérêt comme un vecteur aléatoire dont on essaie d'obtenir la distribution. Pour ce faire, on définit une distribution *a priori* pour chaque paramètre, basée sur la littérature, la connaissance préexistante du type de données étudié, des études préalables, ou arbitrairement (dans ce cas la distribution doit être aussi non informative que possible avec des variances très larges et une densité plate).

On souhaite mettre à jour cette distribution *a priori* grâce à la vraisemblance des données observées y pour obtenir la distribution *a posteriori*. En d'autres termes, on actualise l'information *a priori* avec l'information que l'on peut extraire des données observées pour informer la distribution des paramètres. La formule de Bayes (1763) résume le concept de l'inférence Bayésienne :

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.8}$$

avec $p(\theta)$ la densité *a priori* de θ, $p(\theta|y)$ la densité *a posteriori* de θ et $p(y|\theta)$ la vraisemblance des y conditionnelle à θ. Comme $p(y)$, la vraisemblance marginale des y, est une constante indépendante de θ, on peut juste la considérer comme un facteur d'échelle et l'omettre :

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.9}$$

Dans le cas des NLMEM, utiliser l'inférence Bayésienne revient à ajouter une couche

au modèle général présenté plus haut (équations 1.1 et 1.2) :

$$\theta \sim Prior(\gamma) \qquad \text{loi } a \textit{ priori} \tag{1.10}$$

avec γ le vecteur des hyperparamètres de la loi *a priori* sur les paramètres de population.

Le modèle ainsi défini peut être représenté par un graphe acyclique dirigé qui décrit les différentes couches hiérarchiques obtenues (voir Figure 1.1).

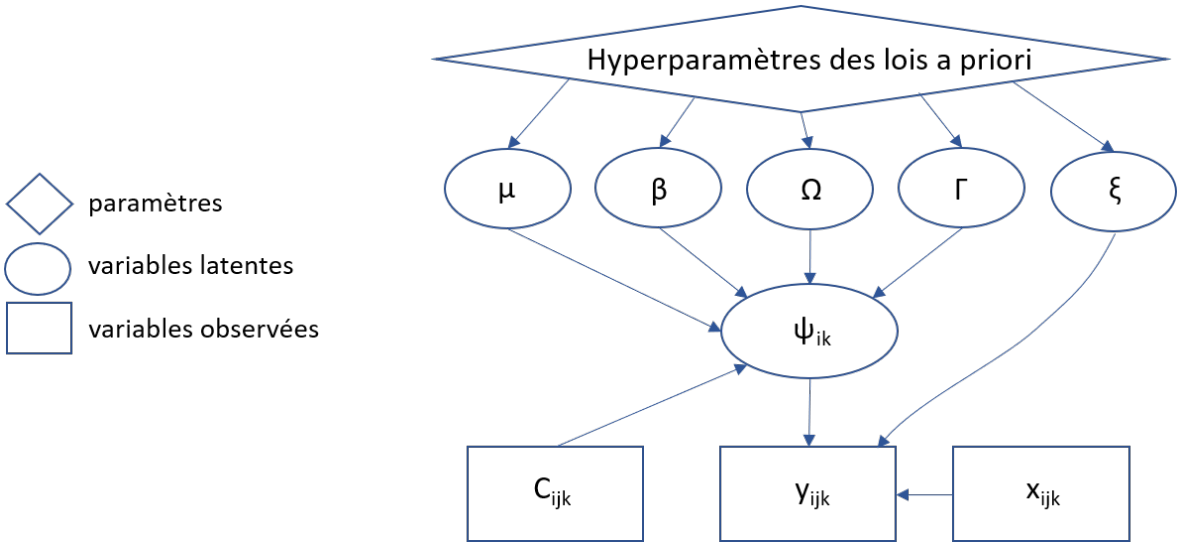


FIGURE 1.1 – Représentation hiérarchique par graphe acyclique dirigé du modèle non linéaire à effets mixtes dans le paradigme Bayésien

1.2.2 Algorithmes Bayésiens

Les algorithmes Bayésiens les plus connus et utilisés sont les méthodes de Monte-Carlo par chaînes de Markov (MCMC) (Robert, 2011). C'est un groupe de méthode qui permet d'estimer la distribution *a posteriori* des paramètres de population à partir d'échantillons tirés selon une chaîne de Markov dont la loi stationnaire est la loi *a posteriori*. Elles permettent d'échantillonner dans des distributions sans définition analytique, en définissant un noyau de proposition qui doit seulement être grossièrement proportionnel à la loi cible, et la propriété de Markov des chaînes assure la convergence

de ces méthodes vers la loi cible sous des conditions très générales.

Par exemple, l'algorithme de Metropolis-Hastings (MH, Metropolis and Ulam (1949); Metropolis et al. (1953)) utilise un noyau de proposition $q()$ pour tirer une réalisation θ_i, qui dépend ou non de la réalisation précédente θ_{i-1}. Un ratio d'acceptation α est ensuite calculé, basé sur le ratio de la distribution *a posteriori* entre le précédent échantillon et le nouveau, divisé par le ratio de densité du noyau :

$$\alpha(\theta_i) = \frac{p(\theta_i|y)q(\theta_{i-1})}{p(\theta_{i-1}|y)q(\theta_i)} \tag{1.11}$$

Selon la formule de Bayes, la distribution *a posteriori* peut être remplacé par le produit de la loi *a priori* par la vraisemblance conditionnelle (ici la vraisemblance des observations $p(y)$ est présente de chaque côté de la fraction et s'élimine). On peut alors écrire le pseudo-code de l'algorithme MH tel que :

Définir une valeur initiale θ_0

Définir une loi *a priori* $p(.)$

Définir un noyau de proposition $q(.)$

Définir la longueur de la chaîne M

Pour i allant de 1 à M :

- Tirer $\theta_i \sim q(.)$
- Calculer le ratio d'acceptation :

$$\alpha(\theta_i) = \frac{p(y|\theta_i)p(\theta_i)q(\theta_{i-1})}{p(y|\theta_{i-1})p(\theta_{i-1})q(\theta_i)}$$

- Accepter θ_i avec probabilité α

L'algorithme Hamiltonian Monte Carlo (HMC), initialement développé en physique quantique (Betancourt, 2017), consiste quant à lui à simuler une trajectoire en utilisant le gradient de la vraisemblance pour orienter les échantillons dans la direction qui maximise la vraisemblance. Cet algorithme est plus efficace car il définit la direction

dans laquelle la vraisemblance est la plus améliorée pour tirer de nouveaux échantillons du paramètre mais il est aussi plus compliqué à implémenter car il y a plusieurs hyperparamètres à calibrer (taille du pas, matrice dite "d'impulsion").

Dans cette thèse a également été considéré un algorithme appelé Approximate Bayesian Computation (ABC, Rubin (1984)), dont la particularité est qu'il se passe du calcul de la vraisemblance pour échantillonner dans la distribution cible. Cette méthode consiste à définir une statistique synthétique, quantité que l'on calcule à partir des observations y, et qui doit refléter l'information contenue dans ces observations. Après avoir calculé cette statistique synthétique sur les données observées, l'algorithme ABC consiste à tirer des échantillons du vecteur de paramètres d'intérêt dans un noyau de proposition, puis à simuler la variable d'intérêt à partir de ces tirages, d'en calculer la statistique synthétique associée, et de la comparer avec celle calculée sur les données observées grâce à la définition d'une fonction de distance appropriée. Les échantillons tirés sont acceptés si la distance calculée est inférieure à un certain seuil qui est également à définir.

Toutes ces méthodes nécessitent un outil de diagnostic fiable pour savoir si les chaînes échantillonnées ont les propriétés nécessaires pour être considérées comme des réalisations de la distribution *a posteriori* cible, et donc pour estimer l'incertitude des paramètres correctement. Parmi les diagnostics classiquement utilisés (Gelman et al., 2013), on retrouve la taille d'échantillon effective (effective sample size, ESS) qui permet d'estimer le nombre d'échantillons indépendants obtenus dans une chaîne :

$$ESS = \frac{M}{1 + 2\sum_k \rho_k} \tag{1.12}$$

avec M le nombre total d'échantillons tirés et ρ_k le k-ième lag d'autocorrélation.

On peut également vérifier, dans le cas où l'on échantillonne plusieurs chaînes parallèlement, que le facteur de réduction d'échelle potentiel, appelé $\hat{R}$ et défini par le ratio de la variance inter-chaînes sur la variance intra-chaîne moyenne, soit assez proche de

1 :

$$\hat{R} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}} \tag{1.13}$$

avec W la variance intra-chaîne moyenne et B la variance inter-chaînes. Cet outil permet de vérifier que les différentes chaînes ont bien "mixé" et formalise le résultat visuel obtenu lorsque l'on trace les différentes chaînes obtenues en les superposant. En pratique, il est conseillé de ne garder que les chaînes avec $\hat{R} < 1.1$ et il est estimé qu'un ESS de 100 est suffisant pour calculer des statistiques sur la distribution *a posteriori*, bien qu'on puisse augmenter ce seuil si on est intéressé par le comportement des queues de distribution (Gelman et al., 2013).

Dans le paradigme Bayésien, on obtient un échantillonnage de la distribution d'intérêt complète. On peut donc utiliser cette distribution et ses caractéristiques afin de définir un estimateur ponctuel pour les paramètres à estimer et de quantifier l'incertitude autour de celui-ci. L'estimateur ponctuel classiquement utilisé est le maximum *a posteriori* (MAP), c'est à dire le mode de la distribution échantillonnée. Pour construire des intervalles de confiance, on peut simplement construire des intervalles de prédiction en utilisant les quantiles correspondants au niveau de confiance choisi.

1.2.3 Implémentation

Il existe plusieurs logiciels spécialisés pour la mise en oeuvre de l'inférence Bayésienne, notamment `Stan` (Carpenter et al., 2017), `BUGS` (Lunn et al., 2000) et `JAGS` (Plummer, 2003), ainsi que des packages dans les logiciels `R` et `python`.

Le logiciel `Stan`, qui implémente l'algorithme HMC, plus adapté aux modèles en grande dimension avec des paramètres corrélés que l'algorithme MH (Gelman et al., 2013), est particulièrement intéressant pour estimer les paramètres de NLMEM. Il existe de plus des interfaces avec le logiciel `R` tels que les packages `rstan` (Stan Development Team), `cmdstanr` (Gabry et al.) et `brms` (Bürkner, 2017) qui facilitent sa prise en main.

1.3 Paradigme fréquentiste

1.3.1 Estimation

Dans le paradigme fréquentiste, on cherche à maximiser directement la vraisemblance pour obtenir l'estimateur du maximum de vraisemblance (EMV) : c'est un estimateur convergent, asymptotiquement efficace et de loi normale (Wasserman, 2004).

Une fois la vraisemblance écrite, on peut utiliser des algorithmes de minimisation classiques, basés sur le gradient de la vraisemblance, ou bien des algorithmes itératifs comme l'algorithme Expectation-Maximisation (EM). Ce dernier, en considérant les paramètres individuels comme des données manquantes, permet d'estimer les paramètres d'un modèle à effets mixtes en évaluant la vraisemblance attendue des observations y intégrée sur les paramètres individuels ψ (étape E) puis en maximisant cette quantité (étape M) à chaque itération (Dempster et al., 1977).

L'utilisation des algorithmes EM et de leurs variantes (par exemple, l'algorithme MCEM (Levine and Casella, 2001) qui remplace l'étape d'intégration par un algorithme de Monte Carlo) a longtemps été limitée à cause de leur lourdeur computationnelle. L'algorithme Stochastic Approximation Expectation-Maximisation (SAEM) a été développé dans les années 1990 (Delyon et al., 1999) pour cette raison. Il s'agit d'une variante de l'algorithme EM qui remplace l'étape E par une étape de simulation des paramètres individuels ψ, puis une approximation stochastique de la log-vraisemblance. Le pseudo-code de l'algorithme SAEM est donné ci-dessous :

Définir une valeur initiale $\hat{\theta}_0$ d'estimation de θ

Pour k allant de 1 à K :

— Simuler m_k matrices d'effets aléatoires ψ selon leur distribution conditionnelle $p(\psi|y;\hat{\theta}_{k-1})$

— Approcher stochastiquement la log-vraisemblance

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k * \left(\frac{1}{m_k} \sum_{j=1}^{m_k} log(p(\psi_j, y; \theta)) - Q_{k-1}(\theta) \right)$$

avec $(\gamma_k)_k$ une séquence de pas positifs convergeant vers 0

— Maximiser cette vraisemblance pour mettre à jour $\hat{\theta}$

$$\hat{\theta_k} = \text{Arg} \max_{\theta} Q_k(\theta)$$

L'étape de simulation des effets aléatoires se fait grâce à un algorithme MCMC ; en pratique, on simule en général une seule matrice pour chaque itération de l'algorithme SAEM.

Le développement de l'algorithme SAEM, beaucoup plus rapide que les algorithmes alors disponibles, a permis de démocratiser l'estimation des NLMEM et a fait du paradigme fréquentiste l'option favorisée pour l'étude de ce type de modèles.

1.3.2 Calcul de la vraisemblance

Au cours de l'estimation des paramètres par SAEM, la vraisemblance est seulement approchée stochastiquement pour trouver l'EMV. Il faut donc également définir une méthode de calcul de la vraisemblance à utiliser à la fin de l'algorithme SAEM, lorsqu'on estime qu'on a atteint l'EMV.

Or, les fonctions utilisées sont non linéaires, ce qui complexifie l'évaluation analytique de la vraisemblance (équation 1.3). Comment alors calculer la vraisemblance marginale des observations y ? Deux approches sont possibles pour obtenir une forme de la vraisemblance :

1) Linéariser le modèle (Lavielle, 2014)

Dans le cas d'un modèle continu gaussien, on peut linéariser le modèle grâce à une approximation de Taylor du premier ordre autour de ψ tous égaux aux paramètres de population ou autour de l'estimateur de Bayes empirique (EBE), c'est-à-dire le mode de la distribution conditionnelle des ψ.

Cette approche a l'avantage d'être rapide et le désavantage d'être d'autant plus biaisée que l'approximation linéaire est erronée ; elle peut également rencontrer des difficultés de convergence en pratique.

2) Approcher numériquement ou stochastiquement la vraisemblance

On peut réaliser une intégration numérique de la vraisemblance : par exemple, la quadrature gaussienne consiste à approximer l'intégrale par une somme finie d'aires approchées.L'approximation de Laplace, pouvant être considérée comme un cas particulier de la quadrature de Gauss, consiste quant à elle en une reparamétrisation de l'intégrale à calculer puis d'une expansion du second ordre de la fonction reparamétrée.

On peut également échantillonner la vraisemblance par une approche stochastique de type Monte Carlo : on échantillonne plusieurs valeurs de la vraisemblance à partir de la distribution des ψ conditionnelle à y et θ puis on en calcule la moyenne empirique, donc la performance de cette méthode dépend de la qualité de l'estimation des paramètres individuels. Il y a plusieurs variantes de cette méthode, qui peut prendre la forme d'un algorithme MCMC ou d'une méthode de type Importance Sampling (IS) dans laquelle on attribue des poids à chaque échantillon en fonction d'un critère défini. Ici il n'y a pas d'approximation du modèle mais une intégration stochastique de la vraisemblance. Elle a l'avantage d'être applicable à des modèles non gaussiens et même non continus.

Le calcul de la vraisemblance à la fin de l'estimation est utilisée pour comparer la qualité de différents modèles directement par comparaison des vraisemblances ou de leurs dérivées (par exemple le BIC) ou via un test du rapport des vraisemblances (voir section 1.3.5).

1.3.3 Calcul de l'incertitude

Information de Fisher

Dans le paradigme fréquentiste, l'outil classique utilisé pour quantifier l'incertitude autour des paramètres estimés est la matrice d'information de Fisher (FIM) $I(\theta)$, que l'on peut exprimer comme l'espérance du produit matriciel de la fonction de score (c'est-à-dire le gradient, vecteur des dérivées premières de la log-vraisemblance) par sa transposée (Fisher, 1922). A l'asymptotique, la FIM peut également être obtenue à partir de la matrice hessienne (c'est à dire la matrice des dérivées secondes), lorsque la vraisemblance du modèle est deux fois dérivable. En effet, la matrice hessienne de la log-vraisemblance est asymptotiquement égale à l'opposé du produit matriciel du score par sa transposée. On peut alors exprimer la FIM comme l'espérance de l'opposé de la matrice hessienne de la log-vraisemblance :

$$I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} log\ p(y;\theta) \right) \tag{1.14}$$

A partir de cette matrice, on définit la borne de Cramér-Rao (Cramér, 1999; Rao, 1992), qui est l'inverse de la FIM, comme la borne inférieure de la matrice de variance-covariance d'un estimateur non biaisé de θ. On peut dériver de cette matrice les erreurs standards (SE) du vecteur de paramètres estimé.

Dans le cas de l'EMV, cette borne est un résultat uniquement valable à l'asymptotique car l'EMV peut être biaisé à distance finie. Pour les modèles mixtes, on définit l'asymptotique à la fois en termes de nombre de sujets N et de nombre d'observations

par sujet n. Les SE dérivées de la FIM sont donc uniquement fiables sous une hypothèse asymptotique, c'est-à-dire qu'à distance finie, elles auront tendance à sous-estimer l'incertitude autout de l'EMV. Il a été montré par divers travaux qu'à la fois l'ordre de N et de n influencent la précision des estimations des SE basées sur la FIM (Panhard and Mentré, 2005; Bertrand et al., 2009; Dubois et al., 2010, 2011; Loingeville et al., 2020).

La formule décrite plus haut implique le calcul d'espérance et permet donc de calculer la FIM attendue. En pratique, il existe plusieurs méthodes de calcul de la FIM observée, estimateur de la FIM attendue :

- on peut calculer la matrice hessienne par différentiation numérique de la log-vraisemblance obtenue selon une des méthodes présentées précédemment,
- ou par une méthode stochastique via un algorithme MCMC (Riviere et al., 2016), ce qui peut être fait par le biais de statistiques exhaustives lorsque la vraisemblance répond à certains critères de forme (Delattre and Kuhn, 2023),
- d'autres méthodes sont basées sur la formule de Louis (Louis, 1982) qui lie la vraisemblance des observations y à la vraisemblance des données complètes (y, ψ), que l'on peut calculer par approximation stochastique (Kuhn and Lavielle, 2005) ou par algorithme MCMC (Savic et al., 2011) par exemple.

A partir des SE obtenues dans la FIM, on peut construire des intervalles de confiance avec une hypothèse de distribution (classiquement gaussienne, ce qui donne des intervalles de confiance symétriques) :

$$IC_{1-\alpha}(\hat{\theta}_i) = [\hat{\theta}_i + q_{\alpha/2} SE(\hat{\theta}_i); \hat{\theta}_i + q_{1-\alpha/2} SE(\hat{\theta}_i)]$$

avec SE l'erreur standard associée à $\hat{\theta}_i$ dérivée de la FIM et q_α le quantile d'ordre α de la distribution choisie.

La méthode basée sur la FIM, appelée Asympt, implique une hypothèse asymptotique qui n'est pas tenable dans de nombreux cas lorsque l'on utilise des modèles mixtes, par

exemple dans certains essais cliniques pour lesquels il est difficile d'enrôler beaucoup de patients (par exemple les maladies rares) ou d'avoir beaucoup d'observations par patients (par exemple les essais pédiatriques). De plus, l'hypothèse de loi nécessaire à la construction d'intervalles de confiance est également restrictive, et potentiellement invalide à distance finie. On a donc besoin de développer des méthodes alternatives de calcul de l'incertitude pour les cas où la méthode Asympt n'est pas fiable.

Méthodes de ré-échantillonnage

Une alternative évidente à la méthode de calcul de l'incertitude basée sur la FIM est d'utiliser une méthode de ré-échantillonnage qui s'affranchit de toute hypothèse asymptotique et qui permet également de se passer d'une hypothèse de loi pour construire des intervalles de confiance.

La méthode bootstrap consiste à répliquer le jeu de données initial et répéter la procédure d'estimation dans chaque réplicat, permettant ainsi d'obtenir une série d'estimations dont on peut utiliser la variabilité comme proxy de l'incertitude sur l'estimation obtenue sur le jeu de données initial. Plusieurs façons de répliquer les données ont été décrites et la méthode optimale dépend du cas de figure dans lequel on se trouve. Dans le cas des NLMEM, il faut prendre en compte les différents niveaux de variabilité présents dans les modèles. Par exemple, le "case bootstrap" consiste à échantillonner avec remise les sujets présents dans le jeu de données, et à répliquer la totalité de leurs observations, afin de conserver la structure de corrélations présente dans les données initiales (Thai et al., 2013). Autre exemple, le bootstrap conditionnel consiste à échantillonner dans la distribution conditionnelle des paramètres individuels (Comets et al., 2021).

Le Sampling Importance Resampling (SIR, Dosne et al. (2016)), méthode de ré-échantillonnage issue de concepts Bayésiens, consiste à échantillonner des vecteurs θ dans une distribution de proposition puis à les ré-échantillonner selon leur ratio d'importance, poids prenant en compte la vraisemblance des échantillons obtenus et la densité de la

distribution de proposition utilisée par rapport à ceux obtenus avec l'EMV. En cela, l'algorithme SIR ressemble beaucoup à une version fréquentiste de l'algorithme MH, dans lequel la définition du ratio d'acceptation se rapproche de la définition du ratio d'importance.

Méthodes semi-Bayésiennes

Les méthodes que l'on appelle ici semi-Bayésiennes consistent en des méthodes basées sur des algorithmes Bayésiens tels que MCMC afin de mesurer l'incertitude autour de l'EMV(Ueckert et al., 2015). On les appelle semi-Bayésiennes car le but est de garder l'EMV comme estimateur ponctuel de θ et d'utiliser le paradigme Bayésien uniquement pour en calculer l'incertitude via des échantillons de sa distribution.

Ces méthodes se basent sur le théorème de Bernstein-von Mises (van der Vaart, 1998) selon lequel les distributions limites de l'EMV et du maximum *a posteriori* (MAP) sont égales. Ce théorème nous donne donc un résultat d'égalité asymptotique entre les distributions de l'EMV fréquentiste et du MAP Bayésien, ce qui signifie qu'échantillonner avec un algorithme Bayésien nous permettrait, à l'asymptotique, d'obtenir des échantillons cohérents avec la distribution de l'EMV.

La question se pose donc de savoir si ce résultat est toujours valable à distance finie, cas dans lequel on manque de méthodes fiables pour mesurer l'incertitude de l'EMV. Les premiers résultats de cette approche montre de bonnes performances sur des petits échantillons (Ueckert et al., 2015; Loingeville et al., 2020).

1.3.4 Implémentation

L'algorithme SAEM présenté plus haut a été implémenté dans différents logiciels comme Monolix (Lavielle, 2014) et NONMEM (Bauer, 2019a,b). Il est aussi disponible dans le logiciel `R` via le package `saemix` (Comets et al., 2017). Ce package est sur le Comprehensive R Archive Network (CRAN) depuis juin 2011, la dernière version étant la 3.2

sortie en juillet 2023 (`https://CRAN.R-project.org/package=saemix`).

Le package `saemix` permet d'estimer les NLMEM avec l'algorithme SAEM, en prenant en compte la variabilité inter-individuelle (mais il ne permet pas pour le moment de prendre en compte la variabilité intra-individuelle). Le package contient de nombreuses fonctionnalités, il permet notamment d'introduire des covariables dans le modèle, de définir la structure de covariance, de prendre en compte différentes formes de transformation des effets individuels (normale, log-normale, logit-normale) et différents modèles d'erreur (additif, multiplicatif, combiné, exponentiel), d'estimer des modèles discrets, et de calculer la vraisemblance par différentes méthodes. Une version de développement disponible sur Github (`https://github.com/saemixdevelopment`) présente également les fonctionnalités en cours d'implémentation qui ne sont pas encore disponibles sur le CRAN (par exemple, la possibilité d'estimer des modèles de survie conjoints (Lavalley-Morelle et al., 2023)).

Dans cette thèse, nous nous sommes concentrées sur le package `saemix` et ses fonctionnalités. En l'état, le package calcule la FIM à la fin de l'algorithme grâce à une méthode par linéarisation : cette méthode n'étant applicable que pour des modèles continus gaussiens, il n'est pour l'instant pas possible d'estimer une FIM pour les autres types de modèles, notamment les modèles catégoriels.

1.3.5 Tests statistiques

Test du rapport de vraisemblance

Dans une procédure de sélection de modèle, pour déterminer la forme structurelle ou les covariables à inclure par exemple, on peut effectuer un test qui compare l'apport de vraisemblance d'un modèle par rapport à un modèle de référence moins complexe, en contre-balançant l'apport de vraisemblance du modèle testé par sa complexité. Les deux modèles doivent être imbriqués l'un dans l'autre pour effectuer ce test.

Soit H_0 l'hypothèse nulle sous laquelle les données suivent le modèle M_0 et l'hypothèse alternative H_1 sous laquelle les données suivent le modèle M_1 plus complexe avec p paramètres supplémentaires ; la statistique du test s'écrit alors :

$$\lambda = -2log\frac{p(y;\hat{\theta_0})}{p(y;\hat{\theta_1})} \tag{1.15}$$

avec θ_0 le vecteur de paramètres estimé pour M_0 et θ_1 le vecteur de paramètres estimé pour M_1. Sous H_0, $\lambda \sim \chi^2(p)$. On rejette donc l'hypothèse nulle selon laquelle les données suivent le modèle M_0 avec un risque α si $\lambda > q_{1-\alpha}$, avec $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ d'une loi $\chi^2(p)$.

Test de Wald

Un test de Wald permet de tester la significativité d'un paramètre, par exemple l'effet d'une covariable sur un paramètre du modèle.

Sous l'hypothèse nulle de ce test, l'effet estimé est considéré comme non significatif. La statistique de Wald peut s'écrire ainsi :

$$W = \frac{(\hat{\theta_i} - \theta_0)^2}{SE(\hat{\theta_i})^2} \tag{1.16}$$

avec $W \sim \chi^2(1)$ sous H_0. On peut aussi utiliser la statistique suivante :

$$T = \sqrt{W} = \frac{\hat{\theta_i} - \theta_0}{SE(\hat{\theta_i})} \tag{1.17}$$

avec SE l'erreur standard associée à $\hat{\theta_i}$ et $T \sim \mathcal{N}(0,1)$ sous H_0.

1.4 Etudes de bioéquivalence

Le premier cas d'étude ayant motivé les travaux de cette thèse est une étude de bioéquivalence.

Ce type d'étude utilise un test appelé le Two One Sided Test (TOST) : c'est un test de Wald qui nécessite donc d'utiliser une méthode de calcul de l'erreur standard associée

au paramètre d'intérêt, parfois suivant une modélisation par les NLMEM, afin d'inférer sur les caractéristiques de la biodisponibilité d'un médicament.

La biodisponibilité est la proportion du principe actif d'un médicament qui atteint la circulation sanguine, celle-ci étant liée à l'action du médicament sur le site cible (Chow and Liu, 2008; Hema Nagadurga, 2019).

Les études de bioéquivalence consistent à tester la similarité de la biodisponibilité, autrement dit l'équivalence pharmacocinétique, de deux formes d'un médicament contenant le même principe actif, en général un médicament princeps et un médicament générique. Ce type d'étude a pour but d'écourter le processus de développement d'un médicament dont le principe actif, c'est à dire sa toxicité et son efficacité, ont déjà été étudiés. Depuis 1984, la U.S. Food and Drug Administration (FDA) peut approuver des médicaments génériques en se basant sur la preuve de la bioéquivalence moyenne avec le médicament princeps (Chow, 2014).

En pratique, pour montrer la bioéquivalence de deux formes d'un médicament, on étudie les concentrations plasmatiques du médicament en fonction du temps. Des variables résumant les profils de concentration plasmatique sont calculées : l'aire sous la courbe de concentration (AUC), qui peut être calculé du temps $t = 0$ à la dernière observation (AUC_{tlast}) ou extrapolée à l'infini (AUC_{∞}), et la concentration plasmatique maximale (C_{max}).

1.4.1 Test de bioéquivalence

Puisque l'on veut tester l'équivalence entre les traitements, le test réalisé doit avoir pour hypothèse nulle la non-égalité de la quantité d'intérêt entre les bras de traitement. Ainsi, un test de bioéquivalence est un test d'effet traitement dont les hypothèses sont inversées. Si l'hypothèse nulle est rejetée, alors on peut conclure à la bioéquivalence entre les deux traitements pour le paramètre considéré.

Les effets traitement sur l'AUC et le C_{max}, respectivement β^T_{AUC} et $\beta^T_{C_{max}}$, sont les différences d'espérance de ces variables sur l'échelle logarithmique dans le bras référence (R) et le bras test (T). Par exemple :

$$\beta^T_{AUC} = \mathbb{E}(log(AUC_T)) - \mathbb{E}(log(AUC_R))$$

Le TOST, proposé par Schuirmann (1987), consiste en une hypothèse alternative sous la forme d'un intervalle, dont les bornes sont à définir. Sous l'hypothèse nulle, l'effet traitement sur le paramètre considéré est en dehors de cet intervalle.

L'hypothèse nulle du TOST peut s'écrire ainsi si on considère un intervalle symétrique :

$$H_0 : \{\beta^T \leq -\delta \text{ or } \beta^T \geq \delta\}$$

Les agences de régulation fixent la borne pour les tests de bioéquivalence à $\delta = log(1.25)$ (U.S. Food and Drug Administration, 2021; European Medicines Agency, 2010). En décomposant l'hypothèse nulle en deux sous-hypothèses, on peut mettre en place deux tests unilatéraux, d'où le nom de la procédure :

$$H_{0,-\delta} : \{\beta^T \leq -\delta\} \text{ and } H_{0,\delta} : \{\beta^T \geq \delta\}$$

Ces deux tests sont réalisés à partir du calcul de statistiques de Wald. Celles-ci sont rejetées au risque $\alpha = 5\%$ si :

$$Z_{-\delta} = \frac{\beta^T + \delta}{SE(\beta^T)} \geq q_{1-\alpha} \text{ et } Z_{\delta} = \frac{\beta^T - \delta}{SE(\beta^T)} \leq q_{\alpha}$$

avec $SE(\beta^T)$ l'erreur standard associée à l'effet traitement β^T et q_α le quantile d'ordre α de la distribution de référence à définir.

De manière équivalente, on peut rejeter l'hypothèse nulle si l'intervalle de confiance de β est compris dans $[-\delta, \delta]$, c'est-à-dire si l'intervalle de confiance de l'exponentielle de β est compris dans $[0.8; 1.25]$. L'exponentielle de β est souvent présentée dans les

résultats du test et est appelé le ratio des moyennes géométriques (geometric mean ratio, GMR). Par exemple :

$$\text{GMR}_{AUC} = exp(\beta^T_{AUC}) = exp\left(\mathbb{E}\left(log\left(\frac{AUC_T}{AUC_R}\right)\right)\right)$$

Appliquer le test sur l'échelle logarithmique a un avantage : les hypothèses de normalité des variables d'intérêt, ainsi que d'égalité des variances entre les bras, nécessaires à l'application et l'interprétation du TOST, sont souvent vérifiées sur cette échelle (Hauck and Anderson, 1984).

Pour mettre en place un test de bioéquivalence, les agences de régulation recommandent un essai clinique croisé lorsque cela est possible. Dans le cas de médicaments à la demi-vie longue par exemple, un essai en parallèle est également accepté. Le nombre de sujets inclus doit permettre d'avoir une puissance minimale fixée *a priori* pour rejeter l'hypothèse nulle, et ce nombre ne doit pas être plus petit que 12.

1.4.2 Approche non compartimentale

A partir des données recueillies, les agences de régulation recommandent l'utilisation du TOST sur l'AUC et le C_{max} dont les effets traitement doivent être calculés grâce à une approche dite non compartimentale (NCA). Dans cette approche, les AUC et C_{max} individuels sont calculés sans modélisation, directement à partir des observations, grâce à la méthode des trapèzes pour l'AUC, éventuellement suivie d'une extrapolation linéaire.

Il est recommandé de collecter 12 à 18 observations pour chaque sujet inclus dans l'étude. Les mesures doivent durer au moins pendant trois (U.S. Food and Drug Administration, 2021) ou cinq (European Medicines Agency, 2010) demi-vies. Il est également recommandé d'avoir au moins trois observations dans la dernière phase d'élimination.

Ensuite, l'effet traitement à l'échelle de la population est estimé par régression linéaire (simple pour une étude en parallèle, mixte pour une étude croisée), et la distribution de référence utilisée dans le TOST est une loi de Student à $N - 2$ degrés de liberté. C'est donc une méthode en deux étapes qui se rapproche d'une ANOVA. Ce type d'analyse ne se généralise pas bien aux cas où le design de l'étude n'est pas équilibré ou au cas où l'on observe des données manquantes. Par ailleurs, une hypothèse sous-jacente de la NCA est que la pharmacocinétique considérée est linéaire, ce qui peut être contestable. De plus, le protocole riche recommandé n'est pas toujours réalisable ou éthique. Un cas très parlant est celui des médicaments ophtalmiques dans lequel on ne peut obtenir qu'une seule observation par sujet. Plus généralement, il est difficile de collecter beaucoup d'observations pour chaque sujet dans certains cas comme les études pédiatriques ou gériatriques.

Dans le cas où l'on réalise une analyse de bioéquivalence sur des données plus éparses que le protocole recommandé, il a été montré dans divers cas que les effets traitement estimés sont parfois biaisés, résultant en une inflation ou une déflation de l'erreur de type I du test de bioéquivalence (Dubois et al., 2010, 2011).

1.4.3 Approche par modélisation

Dans le cas où les données sont trop éparses pour atteindre les recommandations de la FDA, une solution envisagée est d'utiliser les modèles à effets mixtes afin d'informer le modèle grâce aux connaissances déjà acquises sur le médicament (le médicament princeps ayant déjà été développé antérieurement, et notamment la pharmacocinétique de la molécule ayant déjà fait l'objet d'analyses durant les différentes phases de son développement) et d'agréger l'information entre les individus.

Avec un modèle pharmacocinétique, on estime les paramètres directs du modèle structurel choisi, ainsi que les effets traitement associés à chaque paramètre (et à chaque occasion dans le cas d'une étude croisée) à partir desquels on dérive les effets traitement

sur AUC et C_{max}, qui n'apparaissent généralement pas directement dans le modèle. C'est donc une approche en une seule étape contrairement à la NCA qui nécessite la mise en place d'un modèle linéaire, après calcul des AUC et C_{max} individuels, afin d'estimer les effets traitement.

Par exemple, dans un modèle pharmacocinétique à un compartiment, on a la relation $AUC = \frac{D}{Cl}$, avec D la dose administrée et Cl la clairance du médicament, donc on peut calculer l'effet traitement sur AUC à partir de l'effet traitement sur Cl par la delta-méthode (Cramér, 1999). Pour des relations plus complexes, des simulations peuvent être nécessaires pour dériver les effets traitement et surtout leur SE sur les paramètres de biodisponibilité.

Dans cette approche, la distribution de référence utilisée dans le TOST est une loi normale centrée réduite.

Avant aujourd'hui, la FDA et l'European Medicines Agency n'ont jamais évoqué l'utilisation d'approches compartimentales, c'est à dire l'utilisation de modèles, pour tester la bioéquivalence. Cependant, ces dernières années, plusieurs études (Panhard and Mentré, 2005; Dubois et al., 2011; Zhao et al., 2019; Hughes et al., 2017) semblent indiquer qu'une approche par modélisation (MBBE) pourrait améliorer la puissance des tests réalisés et permettre de relâcher les conditions strictes sur le protocole des études cliniques à mettre en place.

La FDA a initié plusieurs collaborations avec des équipes de recherche académiques ces dernières années afin d'explorer cette méthode alternative MBBE pour les études de bioéquivalence sur données éparses (Fang et al., 2018; Lee et al., 2021; U.S. Food and Drug Administration, 2022b,c). Les études publiées montrent les meilleures performances de l'approche MBBE dans certains cas (Dubois et al., 2011, 2012; Tardivon et al., 2023), mais aussi les limites de cette approche, dans laquelle les SE sont calculées par la méthode Asympt et qui donc à distance finie sous estiment parfois l'incertitude autour des effets traitement estimés, entraînant une inflation des erreurs

de type I (Loingeville et al., 2020).

Différentes alternatives ont été évaluées pour pallier cette sous-estimation des SE dans le cadre de la bioéquivalence (Loingeville et al., 2020) :

— une correction du calcul des SE, proposée par Gallant (1975), prenant en compte le nombre de paramètres du modèle et la quantité de données disponible (la distribution de référence utilisée dans le TOST étant également remplacée par une loi de Student) (Bertrand et al., 2012) ;

— une méthode par bootstrap ;

— une méthode semi-Bayésienne appelée Post utilisant l'algorithme HMC pour échantillonner dans la distribution de l'effet traitement (Ueckert et al., 2015).

Ces projets ont donné lieu pour la première fois en 2022 à la publication par la FDA de nouvelles recommandations faisant mention des méthodes MBBE comme alternative envisageable pour les études de bioéquivalence dans le cas de données éparses et/ou de molécules à la demi-vie très longue (U.S. Food and Drug Administration, 2022a).

1.4.4 Données Gantenerumab

Dans le cadre de cette thèse, nous avons utilisé les données d'une étude de bioéquivalence pharmacocinétique provenant d'essais cliniques de phase I investigant la biodisponibilité relative, la tolérabilité et la relation dose-exposition de deux formulations du Gantenerumab, un anticorps monoclonal utilisé dans le traitement de la maladie d'Alzheimer. Cette molécule ayant une demi-vie très longue, ce sont des essais avec bras de traitement parallèles.

Le jeu de données analysé provient de deux essais cliniques randomisés sur des sujets sains entre 40 et 70 ans, et contient des données riches avec beaucoup de mesures par sujet, mais elles nous ont été fournies dans le cadre d'une collaboration avec la FDA pour évaluer l'approche MBBE, et pour cela nous avons également utilisé une sous-partie plus éparse des données disponibles.

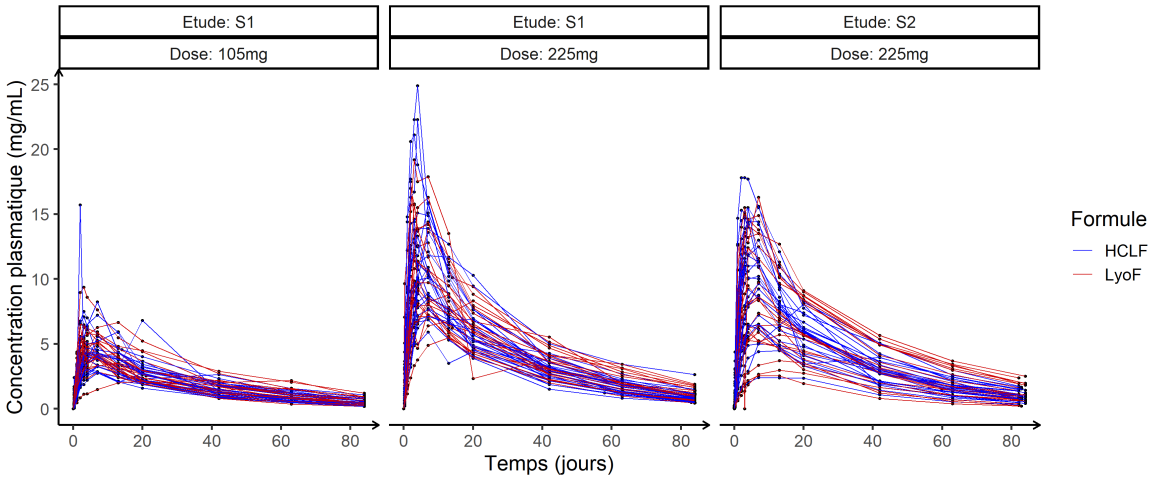


FIGURE 1.2 – Concentrations individuelles, par dose, dans le premier (S1) et le second (S2) essai clinique pour le bras référence (HCLF) en bleu et le bras test (LyoF) en rouge

Dans le premier essai clinique (NCT01636531), on trouve cinq bras parallèles, trois bras pour la forme de référence (forme liquide à autre concentration, HCLF) à trois niveaux de dose unique par injection sous-cutanée (105 mg, 225 mg et 300 mg) et deux bras pour la forme test (forme lyophilisée, LyoF) à deux niveaux de dose (105 mg et 225 mg). Chaque bras contient 24 sujets et la concentration plasmatique de chaque sujet a été observée à 11 temps identiques, jusqu'à 13 semaines après injection de la dose : 6 heures, 1 jour, puis 2, 3, 4, 7, 13, 20, 42, 63 et 84 jours.

Le second essai clinique (NCT02133937) est composé de deux bras parallèles de la forme référence pour 25 sujets et de la forme test pour 23 patients à la dose 225 mg. Les temps d'échantillonnage sont les mêmes que dans l'essai précédemment décrit avec un temps supplémentaire 1 heure après administration de la dose.

Les données utilisées dans la cadre de cette thèse sont représentées sur la Figure 1.2. Les données pour la dose 300mg n'ont pas été exploitées car il n'y avait pas de données dans le bras test pour cette dose.

1.5 Essai clinique Discovery

1.5.1 Covid19

La pandémie de Covid19, provoquée par le virus SARS-CoV-2, a commencé par un cluster de patients à Wuhan en Chine en décembre 2019 (Zhu et al., 2020). Dès les premiers cas, les médecins ont observé des symptômes très variables : certains patients, non détectés à ce moment-là, étaient asymptomatiques, d'autres présentaient une légère infection des voies respiratoires supérieures, et certains une pneumonie sévère conduisant parfois à la mort. Des comorbidités ont été observées chez de nombreux patients développant des formes sévères, principalement l'hypertension, le diabète et les maladies cardiaques chroniques (Zhou et al., 2020).

Plus de trois ans après le début de la pandémie, l'Organisation Mondiale de la Santé (OMS) estime que les 760 millions de cas et 6.9 millions de décès enregistrés sous-estiment les chiffres réels. Plus de 13 milliards de doses de vaccins ont été administrées avant juin 2023 (`https://www.who.int/news-room/fact-sheets/detail/coronavirus-disease-(covid-19)`).

De nombreux essais cliniques ont été mis en place dès le début de l'année 2020, lorsque le virus commençait à circuler dans le monde entier, selon les recommandations de l'OMS (World Health Organisation, 2020). Ces essais avaient pour but de comprendre la dynamique virale du SARS-CoV-2 et les mécanismes de contamination en jeu, de détecter les populations les plus à risque, les comorbidités et leurs risques associés, de tester des traitements et de développer des vaccins.

1.5.2 Protocole clinique

En France, plusieurs essais cliniques ont été mis en place, dont l'essai clinique Discovery (NCT04315948), promu par l'Inserm et dont l'investigateur principal est le Pr

Florence Ader. C'est un essai clinique européen de phase III, randomisé, controlé, ouvert, multicentrique, et adaptatif. Il fait également partie de l'essai clinique européen du consortium Solidarity (WHO Solidarity Trial Consortium, 2021).

Le but principal de cet essai à sa création était l'évaluation de l'efficacité et de la sécurité de plusieurs médicaments antiviraux dans le traitement de l'infection par le SARS-CoV-2 chez des patients hospitalisés ayant besoin de supplémentation en oxygène. Les cinq traitements initiaux étaient les soins standards (SoC), SoC + lopinavir/ritonavir, SoC + lopinavir/ritonavir avec interferon, SoC + hydroxychloroquine et SoC + remdesivir (Ader, 2020). À ce stade, le remdesivir avait déjà montré une activité in vitro, un bénéfice clinique chez les macaques et une réduction du temps de guérison dans une étude de cohorte (Wang et al., 2020; Williamson et al., 2020; Beigel et al., 2020).

L'inclusion du premier patient a eu lieu le 22 mars 2020, avec un objectif de 3100 inclusions au total. Le critère de jugement principal de l'étude était le score clinique ordinal en 7 catégories de l'OMS, mesuré au jour 15 après l'inclusion. Un test devait être réalisé sur l'odds ratio calculé par un modèle à côtes proportionnelles.

Le score clinique suggéré par l'OMS est un score ordinal en 7 catégories :

- Catégorie 1 - Patient non hospitalisé, sans limitations des activités,
- Catégorie 2 - Patient non hospitalisé, avec limitations des activités,
- Catégorie 3 - Patient hospitalisé, sans supplémentation en oxygène,
- Catégorie 4 - Patient hospitalisé, avec supplémentation en oxygène,
- Catégorie 5 - Patient hospitalisé, sous ventilation non-invasive ou appareil d'oxygénation à haut débit,
- Catégorie 6 - Patient hospitalisé, sous ventilation mécanique invasive ou ECMO,
- Catégorie 7 - Patient décédé.

L'essai comporte certaines limites : pour des raisons éthiques, les différents traitements n'ont pas été comparés à un placebo, et les soins standards peuvent varier d'un centre à l'autre ; l'essai n'a pas été réalisé en double aveugle car les modes d'administration des

différents traitements ne sont pas les mêmes ; les patients inclus avaient des symptômes sévères, ce qui suggère qu'ils étaient infectés depuis un certain temps et que l'essai ne permet pas d'étudier l'effet des traitements à un stade plus précoce de la maladie.

1.5.3 Résultats principaux

En mai 2021, les résultats obtenus sur le bras SoC + hydroxychloroquine, arrêté prématurément le 25 mai 2020 sur demande de l'ANSM, et sur les bras SoC + lopinavir/ritonavir et SoC + lopinavir/ritonavir + interferon, arrêtés prématurément le 29 juin 2020 pour futilité, ont été publiés (Ader et al., 2021). Les analyses ont été faites sur 583 des patients inclus. Les résultats ne montrent pas d'effet traitement dans les bras cités et sont cohérents avec ceux obtenus par le consortium Solidarity.

En septembre 2021, les résultats sur le bras SoC + remdesivir restant ont été publiés (Ader et al., 2022). Les analyses ont été faites sur 832 patients provenant de 48 sites européens différents. Les résultats ne montrent pas d'amélioration clinique, d'accélération de la clairance virale, ni de diminution de la mortalité dûs au remdesivir, mais un allongement du délai avant la détérioration des capacités respiratoires. Ces résultats diffèrent d'études précédentes sur le remdesivir mais sont cohérents avec les résultats du consortium Solidarity.

Au vu de l'ampleur de la propagation de l'infection au niveau mondial et de la difficulté à trouver des traitements efficaces pour traiter les formes aiguës sévères, la précision des effets traitement estimés dans les essais cliniques liés au SARS-CoV-2 est primordiale. L'utilisation de modèles mixtes pourrait permettre de prendre en compte la nature longitudinale des données collectées et d'éventuels effets sur la dynamique virale et/ou l'évolution de l'état clinique du patient.

Des études secondaires ont été réalisées dans ce sens, utilisant des méthodes statistiques différentes de celles du protocole de l'essai, et notamment des méthodes de modélisation. Ainsi, Lingas et al. (2022) ont montré un effet du remdesivir sur la réplication virale,

suggérant que l'utilisation de données de dynamique virale permet de détecter des effets que les méthodes classiquement utilisées dans les essais cliniques ne sont pas assez puissantes pour détecter.

Une autre étude évaluant le lien entre la dynamique virale et l'amélioration clinique a confirmé l'effet du remdesivir sur la clairance du virus, plus large chez les patients avec des charges virales hautes à leur entrée dans l'essai, ce qui confirme que le traitement a plus d'impact lorsqu'il est mis en place le plus tôt possible. Un lien a également été établi entre la dynamique virale et l'évolution du statut clinique. Cependant, l'effet estimé du remdesivir sur la dynamique virale n'était pas assez fort pour avoir un impact sur l'évolution du score et donc l'état clinique des patients (Néant et al., 2023).

1.5.4 Données longitudinales

Outre celle du jour 15, des mesures quotidiennes du score OMS ont été prises chez les patients hospitalisés. Chez les patients sortis de l'hôpital, le score a uniquement été évalué au jour 15 et au jour 29. Les données manquantes ont été imputées selon le protocole statistique de l'essai Discovery.

Un autre score a également été mesuré dans l'essai : le National Early Warning Score 2 (NEWS-2), un score clinique composite basé sur le rythme respiratoire, la saturation en oxygène, le besoin de supplémentation en oxygène, la pression artérielle, le pouls, l'état de conscience et la température du patient. Chaque composante peut ajouter jusqu'à 3 points au score NEWS-2 global, qui peut prendre toutes les valeurs entières entre 0 et 20, 20 étant le score le plus défavorable (voir Figure 1.4).

Ce score est lui aussi mesuré quotidiennement chez les patients hospitalisés. On peut observer sur la Figure 1.5 qu'il y a une quantité non négligeable de données manquantes, surtout à partir du jour 10, atténuée aux jours 15 et 29 : ces données manquantes sont majoritairement dues à la sortie d'hôpital d'une partie des patients, qui, selon le protocole de l'essai Discovery, sont suivis aux jours 15 et 29 afin d'évaluer leur score.

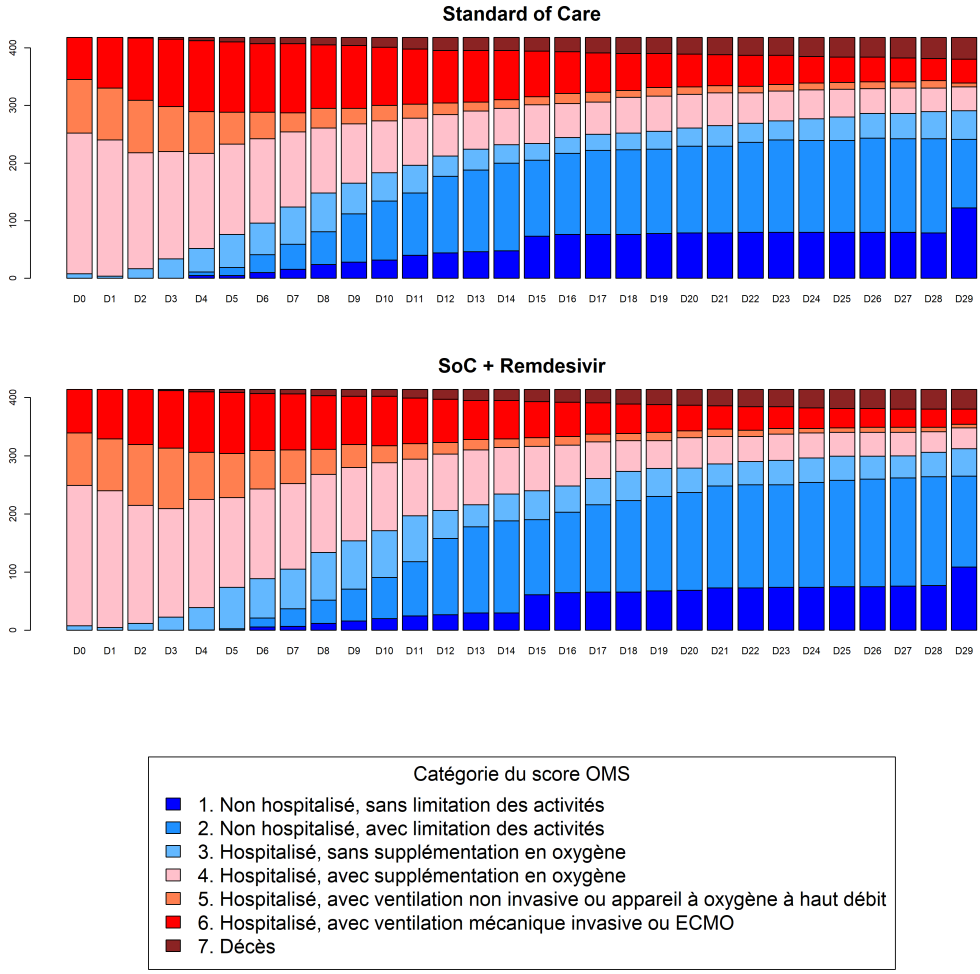


FIGURE 1.3 – Répartition du score OMS dans les bras SoC (408 patients) et SoC+remdesivir (402 patients) de l'essai Discovery, du jour de l'inclusion dans l'essai au jour 29

Physiological parameter	Score 3	2	1	0	1	2	3
Respiration rate (per minute)	≤8		9–11	12–20		21–24	≥25
SpO_2 Scale 1 (%)	≤91	92–93	94–95	≥96			
SpO_2 Scale 2 (%)	≤83	84–85	86–87	88–92 ≥93 on air	93–94 on oxygen	95–96 on oxygen	≥97 on oxygen
Air or oxygen?		Oxygen		Air			
Systolic blood pressure (mmHg)	≤90	91–100	101–110	111–219			≥220
Pulse (per minute)	≤40		41–50	51–90	91–110	111–130	≥131
Consciousness				Alert			CVPU
Temperature (°C)	≤35.0		35.1–36.0	36.1–38.0	38.1–39.0	≥39.1	

FIGURE 1.4 – Table de calcul du score NEWS-2

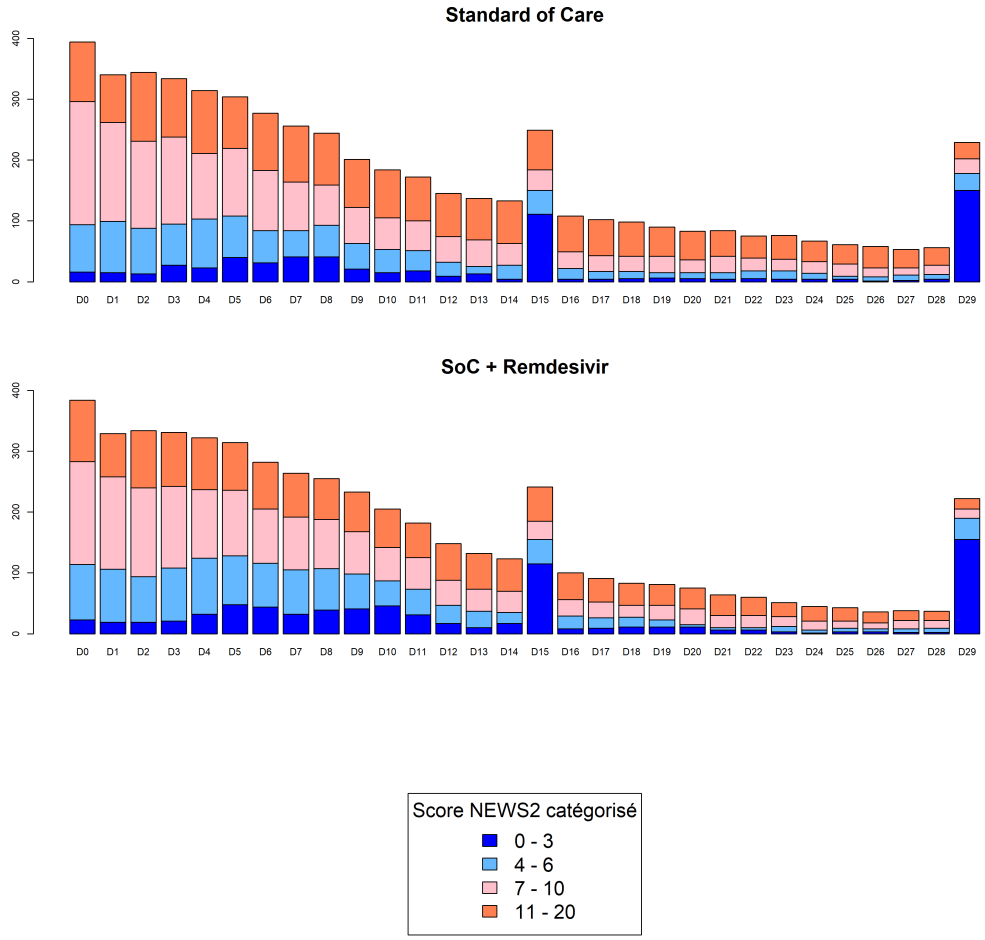


FIGURE 1.5 – Répartition du score NEWS-2 réparti en 4 niveaux de gravité dans les bras SoC (408 patients) et SoC+remdesivir (402 patients) de l'essai Discovery, du jour de l'inclusion dans l'essai à J29

1.6 Objectifs de la thèse

Les différents projets de cette thèse se sont concentrés sur l'incertitude autour de l'EMV obtenu dans les NLMEM avec l'algorithme SAEM, lorsque ces derniers sont utilisés dans le cadre d'étude de bioéquivalence ou de comparaison de traitements. Différentes méthodes de calcul de l'incertitude autour de l'EMV ont été proposées et mises en oeuvre dans des extensions au code du package `saemix`. Ces extensions ont été utilisées pour évaluer les méthodes proposées et les comparer aux méthodes existantes.

1.6.1 Impact du calcul de l'incertitude sur les études de bioéquivalence à distance finie

Une première étude a été réalisée dans le contexte de la bioéquivalence afin d'étudier l'impact du calcul des SE sur les propriétés du test TOST et les risques associés à une mauvaise estimation de la variabilité des paramètres d'intérêt sur les conclusions d'une étude de bioéquivalence par MBBE. Sur les données issues des essais cliniques sur Gantenerumab présentés précédemment, une étude de bioéquivalence par NCA et MBBE a été réalisée, pour comparer les résultats obtenus. Nous avons également étudié la bioéquivalence sur une sous-partie éparse des données pour évaluer les performances de l'approche MBBE dans ce contexte où la NCA n'est pas recommandée, et comparer les différentes méthodes de calcul des SE proposées, l'hypothèse asymptotique de la FIM n'étant plus vraisemblable. Suite à cette étude, nous avons effectué une étude de simulation inspirée des données réelles afin de comparer l'erreur de type I et la puissance du test TOST sur différents designs (riche ou épars), avec différentes approches (NCA ou MBBE) et méthodes de calcul des SE (fréquentistes ou semi-Bayésiennes).

1.6.2 Développement d'une méthode semi-Bayésienne de calcul de l'incertitude dans SAEM

Dans le second travail, après avoir constaté les limites des méthodes existantes pour le calcul des SE à distance finie dans les NLMEM, nous avons développé une méthode semi-Bayésienne basée sur l'algorithme de Metropolis-Hastings afin de s'affranchir de l'hypothèse asymptotique pour calculer l'incertitude sur les paramètres, et ainsi d'obtenir une méthode que l'on espère viable sur des petits échantillons de données lorsque la FIM n'est pas fiable. La méthode est évaluée sur un jeu de simulations relativement simple, et sur une partie des données Gantenerumab.

Le troisième travail explore les limites de la méthode développée ainsi que différentes variantes de cette méthode pour tenter de remédier à ses limites. L'évaluation est également faite sur un jeu de simulations plus complexe inspiré des données Gantenerumab.

1.6.3 Application de la méthode développée aux données de l'essai Discovery

Le quatrième projet a pour enjeu la modélisation des données longitudinales de score clinique recueillies dans l'essai clinique Discovery, ainsi que l'application des méthodes présentées précédemment sur ces scores.

Le score clinique NEWS-2 a été modélisé comme une variable continue, et une procédure de construction du modèle a été mise en place : sélection de la forme du modèle structurel et du modèle d'erreur, de la structure de covariance, des covariables à inclure dans le modèle. Nous avons ensuite étudié l'existence d'un effet traitement du remdesivir sur l'évolution de ce score chez des patients hospitalisés pour la Covid19.

Enfin, nous avons comparé les méthodes que l'on a développées précédemment aux méthodes existantes de calcul des SE, notamment sur les effets traitement estimés, sur le jeu de données complet des scores NEWS-2 avec le modèle sélectionné, ainsi que sur un sous-ensemble de ces données avec le même modèle.

Chapitre 2

Etude de bioéquivalence par modélisation à distance finie

2.1 Résumé

Objectifs

Les études de bioéquivalence sont typiquement réalisées chez un petit nombre de sujets dans des conditions où on peut s'attendre à un écart à la situation asymptotique. Le premier travail de cette thèse a porté sur le calcul des SE utilisées dans les tests de bioéquivalence par une approche de modélisation, et la comparaison des erreurs obtenues par la méthode asymptotique classique et par la méthode semi-Bayésienne Post.

Synthèse

Le premier projet de cette thèse s'inscrit plus largement dans un travail promu par la FDA et en collaboration avec plusieurs laboratoires pharmaceutiques afin d'explorer les méthodes MBBE pour les études de bioéquivalence.

Nous avons travaillé sur les données de deux essais cliniques réalisés par Roche pour étudier la biodisponibilité relative de deux formes du Gantenerumab, un médicament

développé pour le traitement de la maladie d'Alzheimer. Ces données avaient été identifiées par Roche comme d'interêt dans ce projet notamment parce que le Gantenerumab est un anticorps monoclonal à la demi-vie longue et que les essais sont donc en deux bras parallèles.

Nous avons réalisé une étude de bioéquivalence sur les données pharmacocinétiques longitudinales de ces essais, en particulier pour comparer les performances de différentes approches pour réaliser les tests.

Les données de cette étude peuvent être considérées comme riches et sont proches des critères définis par la FDA pour mettre en place une approche NCA (11 prélèvements par sujet).

Sur ces données, nous avons donc mis en place un test de bioéquivalence TOST par NCA. Nous avons également effectué le test de bioéquivalence par une approche de modélisation afin d'évaluer les performances de celle-ci sur des données riches et de comparer les résultats obtenus avec les deux méthodes. Pour ce faire, nous avons sélectionné le modèle pharmacocinétique le plus approprié parmi une sélection de modèles compartimentaux selon un critère de qualité d'ajustement du modèle aux données (BIC) et la structure de variabilité inter-individuelle selon un critère de qualité d'estimation des paramètres (erreurs standards relatives (RSE) $< 50\%$). Nous avons ensuite calculé les SE des paramètres d'intérêt AUC et C_{max} à partir de la FIM.

Nous avons également réalisé toute la procédure de sélection de modèle décrite ci-dessus sur un sous-ensemble des données, en choisissant 5 points de mesure (les modèles sélectionnés sur les données riches comportant 5 paramètres) à l'aide de PFIM, un algorithme d'optimisation de protocoles de population visant à maximiser l'information apportée par un protocole pour l'estimation des paramètres de population. Sur ces données éparses, nous avons mis en place le test de bioéquivalence uniquement par approche de modélisation, car les critères nécessaires à la NCA ne sont pas atteints. Nous avons comparé les résultats obtenus avec les différentes méthodes de calcul des

SE présentées dans la section 1.4.3 : à partir de la FIM, avec la correction de Gallant, et avec la méthode Post implémentée dans le logiciel `Stan`.

Dans cette étude, les résultats obtenus avec le test de bioéquivalence ne sont pas influencés par le design, l'approche utilisée ou la méthode de calcul des SE.

Pour approfondir l'évaluation de l'approche MBBE, nous avons mis en place une étude de simulation dont le design est inspiré de celui des données réelles, afin d'évaluer le contrôle de l'erreur de type I et la puissance du test selon l'approche et la méthode de calcul des SE utilisées.

Sur des jeux de simulation riches, dont le design et le modèle étaient similaires à ceux des données réelles, nous avons comparé l'approche NCA et l'approche MBBE avec calcul des SE à partir de la FIM. Sur des jeux plus épars, une fois encore au design optimisé avec PFIM, nous avons comparé les différentes méthodes de calcul des SE dans l'approche MBBE. Nous avons simulé des effets traitement aux deux bornes de l'hypothèse de rejet afin d'évaluer les erreurs de type I obtenues, et des effets traitement proches de 0 pour évaluer la puissance du test.

L'approche MBBE pour les études de bioéquivalence implique le choix d'un modèle structurel pour décrire les données observées. Pour mesurer l'impact du modèle utilisé sur les résultats, nous avons réalisé les analyses avec un modèle très proche du modèle simulé et un modèle différent (sans l'étape de sélection de modèle).

Apports du travail

Le résultat principal de cette étude de simulation est l'importance de l'étape de sélection du modèle dans l'approche MBBE. En effet, l'utilisation d'un modèle éloigné du modèle simulé a pour conséquence une inflation des erreurs de type I, et la comparaison de la qualité d'ajustement des modèles permet dans la majorité des cas de retenir le modèle le plus approprié, et permet de manière générale de contrôler les erreurs de type I.

Ce travail met également en évidence la pertinence de l'approche semi-Bayésienne pour calculer les SE des paramètres de population dans un modèle non linéaire à effets mixtes. En effet, c'est cette approche qui permet de contrôler au mieux l'erreur de type I du test de bioéquivalence dans cet exemple. Cependant, avec la méthode mise en place dans cette étude, nous avons dû utiliser deux logiciels différents, `R` et `Stan`, pour l'estimation des paramètres de population par l'algorithme SAEM et la construction de la distribution semi-Bayésienne. Dans la suite de la thèse, nous avons donc cherché à développer un algorithme intégré dans SAEM.

2.2 Article 1 (publié)

Ce premier projet a fait l'objet d'un article publié dans Journal of Pharmacokinetics and Pharmacodynamics le 16 septembre 2022.

Journal of Pharmacokinetics and Pharmacodynamics (2022) 49:557–577
https://doi.org/10.1007/s10928-022-09821-z

ORIGINAL PAPER

Impact of model misspecification on model-based tests in PK studies with parallel design: real case and simulation studies

Mélanie Guhl[1] · François Mercier[2] · Carsten Hofmann[3] · Satish Sharan[4] · Mark Donnelly[4] · Kairui Feng[4] · Wanjie Sun[5] · Guoying Sun[5] · Stella Grosser[5] · Liang Zhao[4] · Lanyan Fang[4] · France Mentré[1] · Emmanuelle Comets[1,6] · Julie Bertrand[1]

Received: 1 February 2022 / Accepted: 11 August 2022 / Published online: 16 September 2022

Abstract

This article evaluates the performance of pharmacokinetic (PK) equivalence testing between two formulations of a drug through the Two-One Sided Tests (TOST) by a model-based approach (MB-TOST), as an alternative to the classical non-compartmental approach (NCA-TOST), for a sparse design with a few time points per subject. We focused on the impact of model misspecification and the relevance of model selection for the reference data. We first analysed PK data from phase I studies of gantenerumab, a monoclonal antibody for the treatment of Alzheimer's disease. Using the original rich sample data, we compared MB-TOST to NCA-TOST for validation. Then, the analysis was repeated on a sparse subset of the original data with MB-TOST. This analysis inspired a simulation study with rich and sparse designs. With rich designs, we compared NCA-TOST and MB-TOST in terms of type I error and study power. With both designs, we explored the impact of misspecifying the model on the performance of MB-TOST and adding a model selection step. Using the observed data, the results of both approaches were in general concordance. MB-TOST results were robust with sparse designs when the underlying PK structural model was correctly specified. Using the simulated data with a rich design, the type I error of NCA-TOST was close to the nominal level. When using the simulated model, the type I error of MB-TOST was controlled on rich and sparse designs, but using a misspecified model led to inflated type I errors. Adding a model selection step on the reference data reduced the inflation. MB-TOST appears as a robust alternative to NCA-TOST, provided that the PK model is correctly specified and the test drug has the same PK structural model as the reference drug.

Keywords Equivalence test · Pharmacokinetics · Non-compartmental analysis · Non-linear mixed effects models · Sparse design

✉ Mélanie Guhl
melanie.guhl@inserm.fr

[1] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, IAME, 75018 Paris, France

[2] Department of Biostatistics, Roche Innovation Center Basel, Basel, Switzerland

[3] Department of Clinical Pharmacology, Roche Innovation Center Basel, Basel, Switzerland

[4] Division of Quantitative Methods and Modeling, Office of Research Standards, Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, USA

[5] Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, USA

[6] Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085, 35000 Rennes, France

Introduction

In bioequivalence (BE) studies with pharmacokinetic (PK) endpoints (for generics), or PK similarity studies (for biologicals), we aim to compare the exposure after administration of different drug formulations by comparing two PK parameters of interest: the area under the curve (AUC) of the plasma concentration as a function of time, and the maximal concentration (C_{max}).

BE studies are an essential part of drug development and still an active research field. Currently, a key science and research priority at the U.S. Food and Drug Administration (FDA) is to "improve quantitative pharmacology and BE trial simulation to optimise the design of BE studies for generic drug products and establish a foundation for model-based BE study designs" [1].

The classical statistical test used to assess BE is the Two One Sided Tests (TOST) proposed by Schuirmann in 1987 [2]. It consists of two t tests, on PK parameters of interest, comparing the difference of treatment effects computed to a threshold δ. The FDA as well as the European Medicines Agency (EMA) fix this threshold to $\delta = log(0.8)$ and $\delta = log(1.25)$ [3, 4].

FDA and EMA recommend estimating BE treatment effects via non-compartmental analysis (NCA) for both crossover and parallel study designs [3, 4]. However, assessment of PK equivalence may be challenging for PK BE studies with sparse sampling, such as in participants receiving ophthalmic or oncology drug products. PK BE studies for ophthalmic drug products typically involve a sparse design with one sampling time point per subject (or per treatment group per subject in a crossover design). In such studies, FDA recommends BE to be assessed using a non-parametric bootstrap NCA-based approach or a parametric method [5, 6]. This type of sparse study design may be useful for certain drug products or may occur from study interruptions due to the COVID-19 pandemic or other causes.

An alternative proposed by Dubois et al. [7] is to use a model-based (MB) approach, using the empirical Bayes estimated (EBE) individual parameters of a non-linear mixed effects model instead of NCA parameters. They showed that this method leads to an increase in type I error when the EBE shrinkage is above 20%, which is frequent in case of sparse design. Dubois et al. [8] also proposed a MB approach, this time inferring on the population parameters. They showed that this MB approach works as well as the NCA on rich designs and can be applied on sparser designs. Currently, it is unclear when MBBE methods would be preferred over traditional BE approaches. As such, FDA has actively supported research focused on MBBE approaches for PK BE studies with sparse designs [9–11]. Indeed, MB tests can lead to an inflation of the type I error because of an underestimation of the standard error (SE) of treatment effects on sparse designs in presence of large variability, which led Loingeville et al. to propose and evaluate methods of correction of the standard errors in MB studies [10]. Shen et al. [12] also proposed a MB alternative to traditional BE tests. In this MBBE approach, rich individual PK profiles are simulated from the model and NCA is performed to estimate individual AUC and C_{max} values. Since TOST was based on individual predicted values, the authors assessed distributional assumptions.

MB approaches involve the selection of a PK model to fit the data, which raises the question of the impact of model misspecification on the results of the equivalence tests.

In this study, we define a "sparse" design as any study with only a few sampling points and that challenges the identifiability of the model, which means that the sparse nature of data depends on the complexity of the model of interest.

Our work was based on data collected during the development of gantenerumab, a monoclonal antibody for the treatment of Alzheimer's disease. As this drug has a very long half-life, the clinical trials were conducted using a parallel design (more than 13 weeks of follow up), which is not the classical design for PK equivalence studies that are usually conducted using a crossover design.

In this real case, we compared the PK data gathered in participants treated with two formulations of gantenerumab. Then, we evaluated the performance of the MB approach on simulations based on data from this study and assessed the impact of study design, model misspecification, and the relevance of a model selection step. Although this assessment was based on PK data from a monoclonal antibody, our novel method may potentially be used to evaluate BE studies in generic drug development when there is sparse PK sampling.

We first present the theoretical background, i.e., the NCA and MB approach for equivalence TOST tests. We then describe the observed data, the methodology to analyse it and the results of this real case study. We finally present the design, methods and results of the simulation study, and discuss our findings in the last section.

Theoretical background

Two One-Sided Tests

Showing the PK equivalence of two drug formulations, one reference (R) and one test (T), means showing their exposure is equivalent.

In PK BE studies, drug exposure is typically characterised by two PK parameters, variables of the plasma concentration versus time profiles : the Area Under the Curve (AUC), which can be computed from 0 to the last sampling point (AUC_{tlast}) or extrapolated to infinity (AUC_∞), and the maximum plasma concentration (C_{max}). Treatment effects on AUC and C_{max}, namely θ_{AUC} and $\theta_{C_{max}}$, are defined as the difference of the expectation of the log individual values of these variables under test and reference treatment. For instance:

$$\theta_{AUC} = \mathbb{E}(log(AUC_T)) - \mathbb{E}(log(AUC_R)) \quad (1)$$

Since we wish to reject the assumption that the two formulations have different exposures, we write the null hypothesis as [2]:

$$H_0 : \{\theta \leq -\delta \text{ or } \theta \geq \delta\} \quad (2)$$

where δ is the tolerance. The regulatory guidances for equivalence studies fix the threshold $\delta = log(1.25)$ [3, 4].

By decomposing this null hypothesis in two, we perform Two One-Sided Tests (TOST):

$$H_{0,-\delta} : \{\theta \leq -\delta\} \text{ and } H_{0,\delta} : \{\theta \geq \delta\} \quad (3)$$

The two t test statistics are rejected at $\alpha = 5\%$ if:

$$Z_{-\delta} = \frac{\theta + \delta}{SE(\theta)} \geq q_{1-\alpha} \text{ and } Z_{\delta} = \frac{\theta - \delta}{SE(\theta)} \leq q_{\alpha} \quad (4)$$

with q_α the quantile of order α of a reference distribution.

Equivalently, we can reject the null hypothesis if the confidence interval of θ is within $[-\delta, \delta]$, that is if the confidence interval of the exponential of θ is within [0.8 ; 1.25]. The exponential of θ is often shown in the results of the test and is called the geometric mean ratio (GMR).

Non-compartmental analysis

The standard method for PK equivalence studies is to compute individual AUC and C_{max} and use an ANOVA or a linear mixed model to estimate the treatment effect. AUC_{tlast} can be computed using the trapezoidal method and AUC_∞ can be estimated by linear extrapolation. For this, FDA recommends that sampling continues for at least three or more terminal elimination half-lives of the drug and there are at least three sampling points after the peak [3]. C_{max} is defined as the maximal concentration measured among the study sampling times.

Depending on the study design, there can be a period and a sequence effect on the variables of interest. In parallel studies, there is only one period: each group of participants receives one treatment only. Our present work focuses on a drug with a long half-life which warrants a parallel study design instead of the classical crossover design for PK equivalence studies. In this case, there is no period or sequence effect and intra-individual variability cannot be properly evaluated. The models to fit are simply:

$$log(AUC_i) = \mu_{AUC} + \theta_{AUC} T_i + \epsilon_{AUC_i} \quad (5)$$

$$log(C_{max_i}) = \mu_{C_{max}} + \theta_{C_{max}} T_i + \epsilon_{C_{max_i}} \quad (6)$$

with:

- μ: mean value of variable for the reference treatment;
- T_i: treatment covariate variable for individual i;
- θ: coefficient of treatment effect;
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$: residual error.

The treatment effects on the variables of interest and their standard errors are obtained directly from the linear model inference.

The geometric mean ratio is, e.g. for AUC:

$$\begin{aligned} GMR &= \frac{exp(\mathbb{E}(log(AUC_T)))}{exp(\mathbb{E}(log(AUC_R)))} \\ &= \frac{exp(\mu_{AUC} + \theta_{AUC})}{exp(\mu_{AUC})} \\ &= exp(\theta_{AUC}) \end{aligned}$$

In non-compartmental PK equivalence analyses (hereafter called NCA-TOST), the standard error is obtained with the Fisher Information Matrix (FIM), which is asymptotically the inverse of the lower bound of the variance-covariance matrix of regression coefficients. With balanced groups, the reference distribution to use in NCA-TOST is a Student's t distribution with N-2 degrees of freedom, N being the number of participants in the study.

Model-based approach

Regulatory requirements may not be met in studies with sparse sampling design, and NCA-TOST may then become less accurate. Indeed, it can be hard to compute individual AUC and C_{max} if we only have a few points per subject. In an effort to leverage population data over time to inform predictions for individuals, a model-based alternative has been proposed [8, 10], in which we build a structural PK model and use a non-linear mixed effect model (NLMEM) to estimate the treatment effect. The corresponding statistical model can be written as follows in the case of parallel studies:

$$y_{ij} = f(t_{ij}, \phi_i) + g(t_{ij}, \phi_i)\epsilon_{ij} \quad (7)$$

$$log(\phi_{il}) = log(\mu_l) + \theta_l T_i + \eta_{il} \quad (8)$$

with:

- t_{ij}: time j for individual i;
- y_{ij}: concentration for individual i at time t_{ij};

- ϕ_i: vector of parameters for individual i (typically of size 3 to 10);
- $f(t_{ij}, \phi_i)$: non-linear structural PK model depending on ϕ_i;
- $g(t_{ij}, \phi_i)$: error model;
- $\epsilon_{ij} \sim \mathcal{N}(0, 1)$: residual error;
- μ_l: fixed effect for parameter l;
- T_i: treatment covariate variable;
- θ_l: coefficient of treatment effect for parameter l;
- $\eta_{il} \sim \mathcal{N}(0, \omega_l)$: between subject random effect for parameter l;
- ω_l: standard deviation of the inter-individual random effect for parameter l.

$g()$ describes the error model, with usual models being:

- Additive error model: $g(t_{ij}, \phi_i) = \sigma_a$;
- Multiplicative error model:$g(t_{ij}, \phi_i) = \sigma_b f(t_{ij}, \phi_i)$;
- Combined error model: $g(t_{ij}, \phi_i) = \sigma_a + \sigma_b f(t_{ij}, \phi_i)$.

In the context of BE studies, we usually have previous knowledge on the underlying PK characteristics of the reference product, which could be described by a subset of structural PK models $f()$.

In this study, we only fitted and compared PK models that differed in terms of number of compartments, order of absorption, and presence of an absorption delay. A description of all the models used in this study can be found in Appendix 1, defining the vector μ of l parameters related to each model.

Computation of standard errors

In this study, we used and compared three different methods of computation of SE in the MB approach, that are described below, and called "Asympt", "Gallant" and "Post". These three methods have also been evaluated in the context of BE studies by Loingeville et al. [10].

Asympt

AUC and C_{max} are secondary PK parameters of the models, i.e., functions derived from the PK model direct parameters, and their treatment effects are also functions of the PK model direct parameters and treatment effect: $\theta = h(\mu_{PK}, \theta_{PK})$. For instance, for all PK models with a linear elimination, $AUC_\infty = \frac{FD}{CL}$, where D is the dose administered, F the bioavailability of the drug and CL the clearance, so the treatment effect on AUC_∞ can be simply derived from the model as $\theta_{AUC_\infty} = -\theta_{CL/F}$ and $SE(\theta_{AUC_\infty}) = SE(\theta_{CL/F})$. In one compartment models, there are analytical solutions for all secondary PK parameters, so the delta-method can be used to compute the standard errors of treatment effects. In two-compartment models, there is no analytical solution for C_{max}, so we need to compute $\theta_{C_{max}}$ and its standard error by simulation. This method consists of sampling parameters from a multi-normal distribution with maximum likelihood estimates as the mean vector and the inverse of the FIM as the variance-covariance matrix, to simulate rich concentration profiles for reference and test treatments (see Appendix 2 for a more precise description of the method).

In this approach (which will be designated hereafter by MB-TOST Asympt), the standard error computed in NLMEM is also obtained with the FIM, using a linearisation of the PK model.

The reference distribution we use in MB-TOST Asympt is a Gaussian distribution with zero mean and a standard deviation equal to 1.

In the MB approach, an underestimation of the asymptotic standard errors of the treatment effects has been observed which resulted in an inflation of type I error when performing PK equivalence tests [8]. To address this, several methods of correction of the asymptotic standard errors have been suggested. Here, we use two methods of correction, designated Gallant and Post, which were proposed for equivalence tests by Loingeville et al. [10].

Gallant

The Gallant correction [13] (MB-TOST Gallant) aims to take into account the number of parameters estimated towards the available data to correct for the underestimation of the standard errors of treatment effects. It involves re-weighting the standard errors using the following formula:

$$SE_{Gallant} = SE \sqrt{\frac{N}{N-p}} \tag{9}$$

with N the number of participants in the study and p the number of fixed and covariate effects (here, we only have the treatment as a covariate).

We also switch the reference distribution used in the tests from a Gaussian distribution to a Student's t distribution with $N - p$ degrees of freedom.

Post

This method (MB-TOST Post) uses posterior distribution samples to compute the standard errors of treatment effects [10].

Samples of population parameters are generated by Bayesian inference, with the Hamiltonian Monte Carlo algorithm. Maximum likelihood estimates obtained with NLMEM are used as initial values. Uniform priors are used

for the fixed and treatment effects and Half-Cauchy distributions with zero mean and a standard deviation equal to 1 for the random effects and residual error variance parameters.

When the data are not informative enough given the number of model parameters to estimate, these priors can result in chains with low N_{eff} and high $\hat{R}$. When $N_{eff} \leq 400$ and $\hat{R} \geq 1.05$, log normal priors can be used for the fixed effects, with mean equal to the maximum likelihood estimation and a standard deviation equal to 0.5 and normal priors with zero mean and standard deviation equal to 0.5 for the treatment effects as in [10].

The standard errors of treatment effects are computed using samples from the posterior distribution.

The reference distribution, as for MB-TOST Asympt, is a Gaussian distribution with zero mean and a standard deviation equal to 1.

Case study: gantenerumab

Data

In our analysis, PK data was collected from two phase I randomised clinical trials on healthy male or female subjects between 40-70 years of age. These trials investigated the relative bioavailability, tolerability, and dose-exposure relationship of a high concentration liquid formulation (HCLF G3) versus a lyophilised formulation (LyoF G2) of gantenerumab, a monoclonal antibody used for the treatment of Alzheimer's disease. Hereafter we considered the high concentration liquid formulation as the reference formulation. Both formulations were administered by subcutaneous injection. The first study (NCT01636531, here called S1) was composed of five parallel arms with 24 participants each: three reference arms at different dose levels (105, 225 and 300 mg) and two test arms (105 and 225 mg). In the second study (NCT02133937, here called S2), composed of one reference arm of 25 participants and one test arm of 23 participants, the dose tested was 225 mg. PK sampling was performed in participants for up to 13 weeks using the following scheme: 0.25, 1, 2, 3, 4, 7, 13, 20, 42, 63, and 84 days post dose. There was one additional sampling time in S2, one hour post dose (0.04 days). We evaluated PK equivalence of the two formulations in terms of C_{max} and AUC_{∞}.

Methods

We performed separate analyses for each study and dose tested, hereafter called S1-105, S1-225 and S2-225, discarding the 300 mg arm of S1 as this study did not include a test treatment arm at this dose.

On the original rich design data (11 sampling points per subject), different structural PK models and residual error models were fitted on the reference arms, and compared for selection purposes. The structural PK models tested differed in terms of number of compartments (one or two), order of absorption (zero or one) and presence of an absorption delay. A description of all these models can be found in Appendix 1. As we work on a drug administered by sub-cutaneous injection, the parameters of the PK models used are apparent parameters scaled by the bioavailability of the drug F. Inter-individual variability followed a log-normal distribution for all parameters. Three types of error models were tested: additive, multiplicative and combined. Models were compared using the Bayesian Information Criterion (BIC) computed by Importance Sampling, combined with a second criteria of a relative SE (RSE) below 50% for all parameters. Inter-individual variability parameters that did not meet this second criteria were removed. We also explored the relevance of adding a correlation between the inter-individual variabilities. Goodness of fit was assessed with Visual Predictive Checks (VPC) and Normalised Prediction Distribution Errors (NPDE) [14]. The selected PK model was then fitted on both the reference and test arms and treatment effects were estimated on all parameters. We compared the results of MB-TOST, using only the Asympt computation method for the SE, with results obtained with NCA-TOST which usually performs well on such rich designs.

MB analyses were also run on a sparse subset of the data to explore the impact of the study design. The sparse subset for each study contained 5 points per subject because it is the maximum number of population parameters that we needed to estimate, in order to make the model identifiable. These points were obtained by optimisation of the design with *PFIM* [15] (Population Fisher Information Matrix, an algorithm for the evaluation and optimisation of designs), using the model fitted on the rich reference and test arms. Given that this manuscript focuses on the investigation of MB methods as an alternative for sparse design, we tested the PK equivalence only with MB-TOST, selecting again the PK structural model on the reference arm. Three methods to compute the SE were used: Asympt, Gallant and Post.

Implementation

Analyses were run on R version 4.0.2. Parameters of the PK models were estimated by maximising the likelihood using the Stochastic Approximation of Expectation Maximisation algorithm (SAEM) [16], in the *saemix* R package [17]

(development version: https://github.com/saemixdevelopment/saemixextension). For NCA-TOST, AUC_{∞} was computed by extrapolation with the *PKNCA* R package [18] version 0.9.4, using the observed concentration at t_{last}. Sampling points for the sparse designs were chosen with the *PFIM* [15] R package version 4.0 which enables to optimise population design using the Fedorov–Wynn algorithm.

Results

Figure 1 shows spaghetti plots of the plasma concentrations of gantenerumab versus time in log-scale, for the two lower doses in each study.

The same model, a two-compartment model (V_1/F: apparent volume of the principal compartment, V_2/F:

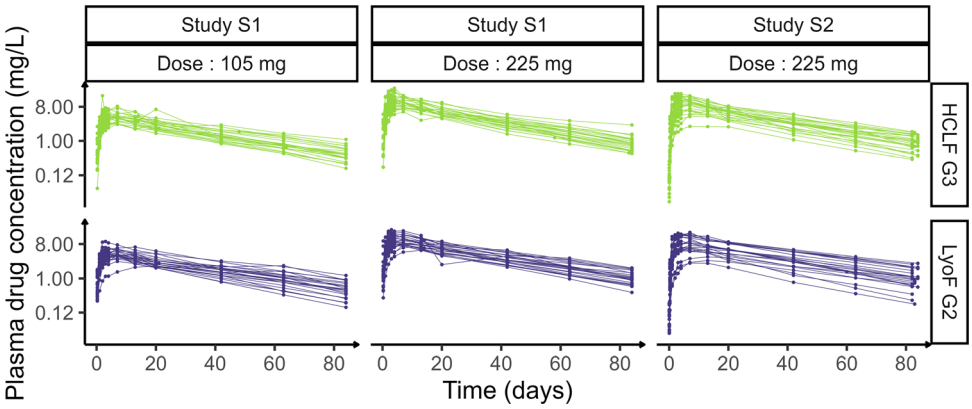


Fig. 1 Individual concentration versus time profiles, in log scale, in studies S1 and S2 per dose (105 and 225 mg), in the reference (HCLF G3) and test (LyoF G2) treatment arms (colour figure online)

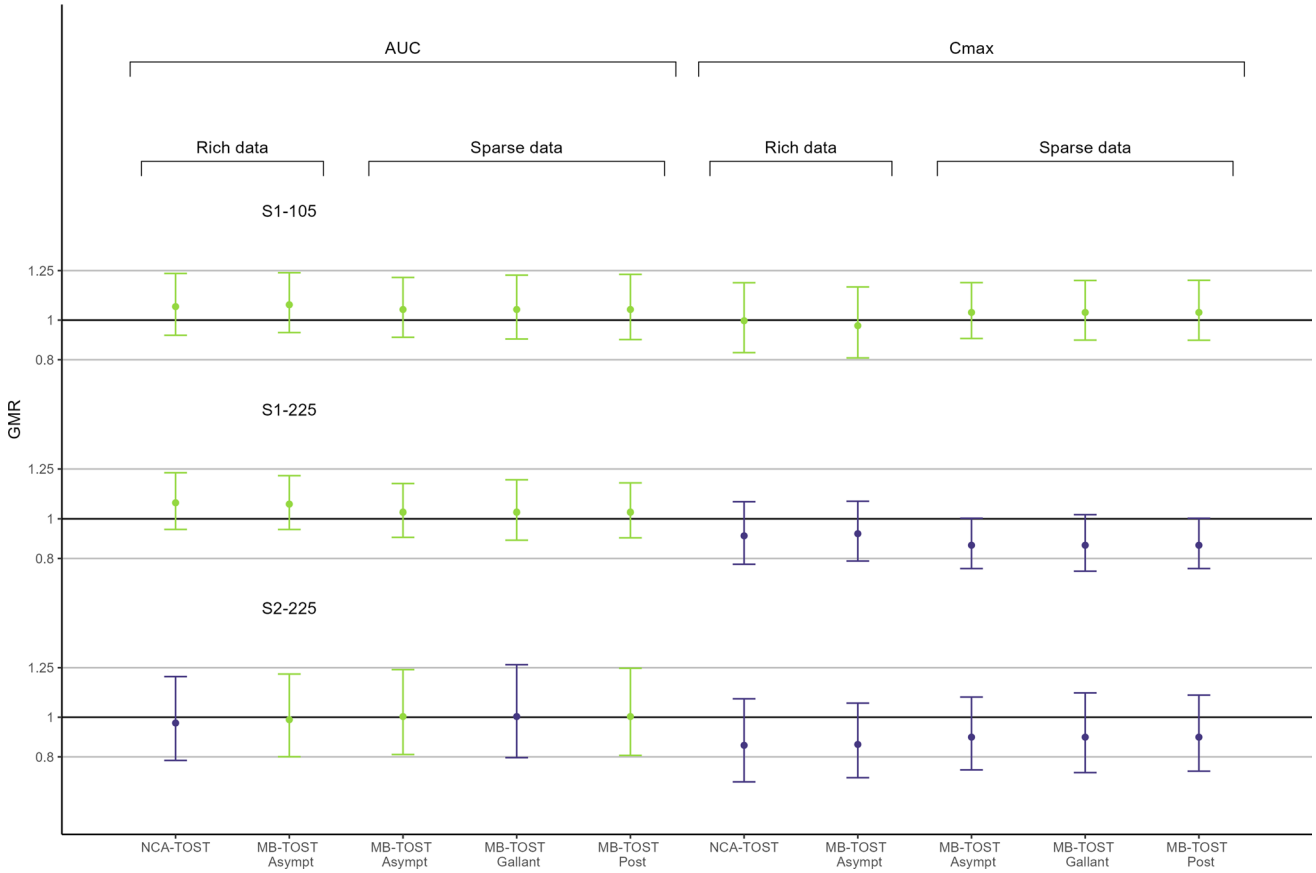


Fig. 2 Geometric mean ratios (GMR) and their 90% confidence intervals for *AUC* and C_{max}, with NCA-TOST and MB-TOST Asympt on observed data and with MB-TOST Asympt, Gallant and Post on sparse data S1-105 denotes Study 1 with dose=105mg reference and treatment arms and similarly for S1-225 and S2-225. *Grey lines are the limits of the null hypothesis interval, GMR* = 0.8 *and GMR* = 1.25, *and the black line represents GMR* = 1. *PK equivalence is shown as green intervals while blue intervals highlight the parameters and datasets for which PK equivalence was not established*

apparent volume of the peripheral compartment, *Q*/*F*: apparent inter-compartmental clearance) with linear absorption (*ka*: absorption constant) and elimination (*CL*/*F*: apparent clearance constant) with an absorption delay (T_{lag}), was selected to be the best (among the considered candidates) at describing the drug PK across studies/arms (taken as three separate datasets). A treatment effect was estimated on all 6 parameters (θ_{Tlag}, θ_{ka}, $\theta_{CL/F}$, $\theta_{V1/F}$, $\theta_{Q/F}$, and $\theta_{V2/F}$). On all datasets, based on BIC, the inter-individual random effect on V_2/F was withdrawn, and a correlation between the inter-individual random effects of *CL*/*F* and V_1/F was estimated. On S1-105 and S1-225, the error model was multiplicative. On S1-225, no inter-individual random effect was kept on *Q*/*F*. On S2-225, the error model was combined. The models selected were therefore very similar. Table 4 in Appendix 3 gives the parameter estimates obtained across datasets. As shown in Fig. 2, illustrating the GMR and their confidence intervals in the different datasets investigated, the different methods gave consistent results: for S1-105, with both NCA-TOST and MB-TOST Asympt, the 90% confidence interval of the GMR of *AUC* and C_{max} fell within [0.8; 1.25], but for S1-225, equivalence could not be shown on C_{max} with either of the two methods. On S2-225, equivalence could not be shown on C_{max} with both methods. For *AUC*, equivalence was shown using MB-TOST but not using NCA-TOST, although the estimates were close (MB-TOST Asympt: 90% CI=[0.801;1.218], p-value=0.049; NCA-TOST: 90% CI=[0.782;1.205], p-value=0.070). The data used to produce Fig. 2 are provided in Table 5 in Appendix 3.

The sparse design optimised using *PFIM* led to the following sampling scheme: 0.25, 3, 7, 20, 84 days post dose for S1-105, 0.25, 4, 20, 42, 84 days for S1-225, and 0.04, 4, 13, 42, 84 days post dose for S2-225. The selected PK model was a one compartment model with linear absorption and an absorption delay on the two S1 datasets, and a one compartment model with zero order absorption and no absorption delay on S2. Again, a treatment effect was estimated on all apparent parameters in each case. On all datasets, a correlation between the inter-individual random effects of *CL*/*F* and *V*/*F* was selected. On S1-105 and S1-225, the error model selected was multiplicative. On S2-225, the error model selected was combined. On S1-225 and S2-225, no inter-individual random effect was kept on T_{lag}. Table 4 in Appendix 3 gives the parameters estimated on all these subsets. Although the PK models selected on the sparse data were different from the ones selected on the observed data, the results of the equivalence study using MB-TOST were consistent, across all computation methods of SE, and comparable to those obtained on rich design (Fig. 2).

Fig. 6 shows the VPC and Fig. 7 reports the normality of residuals for S1-225 original and sparse design. These goodness of fit plots have also been checked for S1-105 and S2 (not shown).

Simulation study

Methods

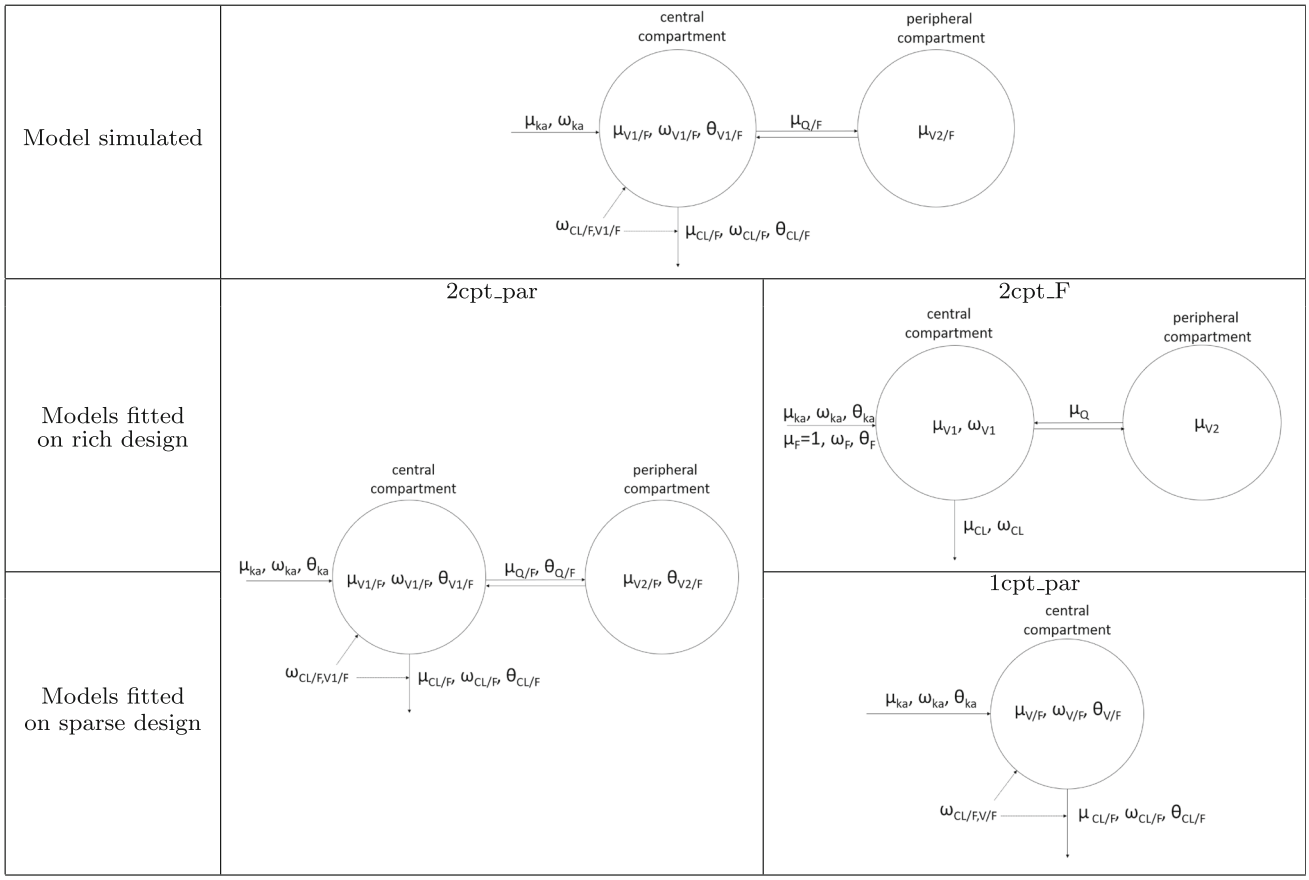
The real case study inspired our simulation settings with rich and sparse design. We simulated parallel studies with reference and test treatment arms, 24 participants per arm. The vector of rich sampling times was taken from S1-225 : 0.25, 1, 2, 3, 4, 7, 13, 20, 42, 63, and 84 days post dose.

The PK model used to simulate data was the one selected to describe the data of the reference arm of S1-225, corresponding to a two-compartment model with linear absorption and elimination. We removed the absorption delay. Moreover, the simulation study was performed prior to the availability of the data for publication. At the time, we only had access to scaled values of the doses that were divided by 15. Table 1 gives a graphical representation of the model simulated, and Table 2 gives the values of the fixed, random and error parameters simulated that were taken from the fit of S1-225.

Different levels of treatment effects were simulated on the apparent parameters, in order to get a treatment effect on *AUC* and C_{max} at the desired levels. To compute type I errors, we simulated data with treatment effects on *AUC* and C_{max} at boundaries of the null hypothesis, log(0.8) and log(1.25). These scenarios are denoted as $H_{0:0.8}$ and $H_{0:1.25}$, respectively. To study the power, we simulated data with treatment effects on *AUC* and C_{max} at and close to 0 (log(0.9), log(1) and log(1.11)). These scenarios are denoted as $H_{1:0.9}$, $H_{1:1}$ and $H_{1:1.11}$. The treatment effects were simulated on clearance (*CL*/*F*) and central volume (V_1/F), with no treatment effect on *ka*, *Q*/*F* and V_2/F. In practice, the treatment effect on *CL*/*F* was fixed (e.g. $\theta_{CL/F} = log(0.8)$ to get $\theta_{AUC} = log(1.25)$) and then the treatment effect on *V*1/*F* was varied to obtain the desired treatment effect on C_{max} without impacting the treatment effect on *AUC*. Table 3 gives the values of the different levels of treatment effects simulated. For each of the 5 treatment effects, 1000 datasets were simulated.

On rich design simulations, we compared the performances of NCA-TOST and MB-TOST Asympt in terms of type I error and study power. We first fitted the simulated structural PK model, estimating treatment effects on all 5 apparent parameters (referred to as model 2cpt_par). We also explored the performance of MB-TOST Asympt when modeling the treatment effects differently, i.e., two-

Table 1 Graphical representation of the model simulated and the models fitted on the rich and sparse design simulations, with the corresponding fixed and treatment effects and inter-individual variability parameters



The graphical representation 1*cpt_par* corresponds to the third model presented in Appendix 1 (one compartment model with linear absorption and elimination) and the three other graphical representations correspond to the fifth model presented in Appendix 1 (two compartment model with linear absorption and elimination)

Table 2 Fixed coefficient values for fixed effects and standard deviations of the inter-individual random effects and residual errors, under which data were generated in the simulation study

μ_{ka} (d)	$\mu_{CL/F}$ (L.d^{-1})	$\mu_{V_1/F}$ (L)	$\mu_{Q/F}$ (L.d^{-1})	$\mu_{V_2/F}$ (L)
0.45	0.04	0.96	0.03	0.34
ω_{ka} (%)	$\omega_{CL/F}$ (%)	$\omega_{V_1/F}$ (%)	ρ_{CL/V_1}	σ_b (%)
57	26	36	0.8	15

compartment model with treatment effects estimated on the absorption parameter only, i.e., *ka*, and an additional scale/bioavailability parameter defined by F, with μ_F fixed to 1, and ω_F estimated (called hereafter 2cpt_F). Table 1 represents the structure of both models fitted to the rich design data.

Table 3 Treatment effects simulated on CL/F and V_1/F and GMR obtained on AUC and C_{max} on each simulation scenario

Scenario	Treatment effect on		GMR on	
	CL/F	V_1/F	AUC	C_{max}
$H_{0:0.8}$	log(1.25)	log(1.279)	0.8	0.8
$H_{1:0.9}$	log(1.11)	log(1.124)	0.9	0.9
$H_{1:1}$	log(1)	log(1)	1	1
$H_{1:1.11}$	log(0.9)	log(0.889)	1.11	1.11
$H_{0:1.25}$	log(0.8)	log(0.778)	1.25	1.25

In a second step, we also analysed sparse optimal design subsets, using PFIM: we selected 5 time-points, assuming 2cpt_par was true. The same 5 time points were selected

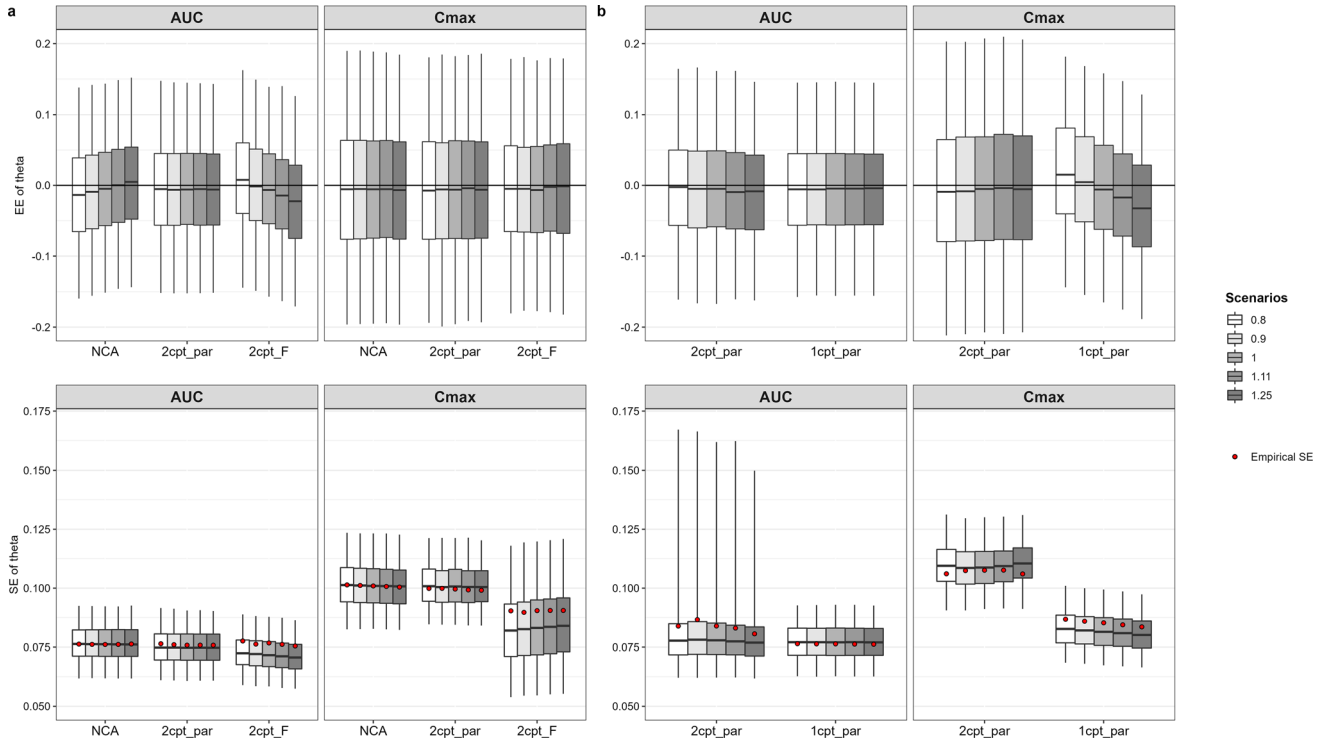


Fig. 3 Boxplots of estimation errors (EE) (top row) and standard errors (SE) (bottom row) of the treatment effects estimated on *AUC* and C_{max}, on (a) rich design simulations with NCA-TOST and MB-TOST Asympt, using the simulated PK structural model and treatment effects estimated and all apparent parameters (2cpt_par) or only on *ka* and *F* (2cpt_F), and (b) sparse design simulations with MB-TOST Asympt using the simulated PK structural model (2cpt_-par) or a misspecified one compartment model (1cpt_par), with treatment effects estimated on all apparent parameters

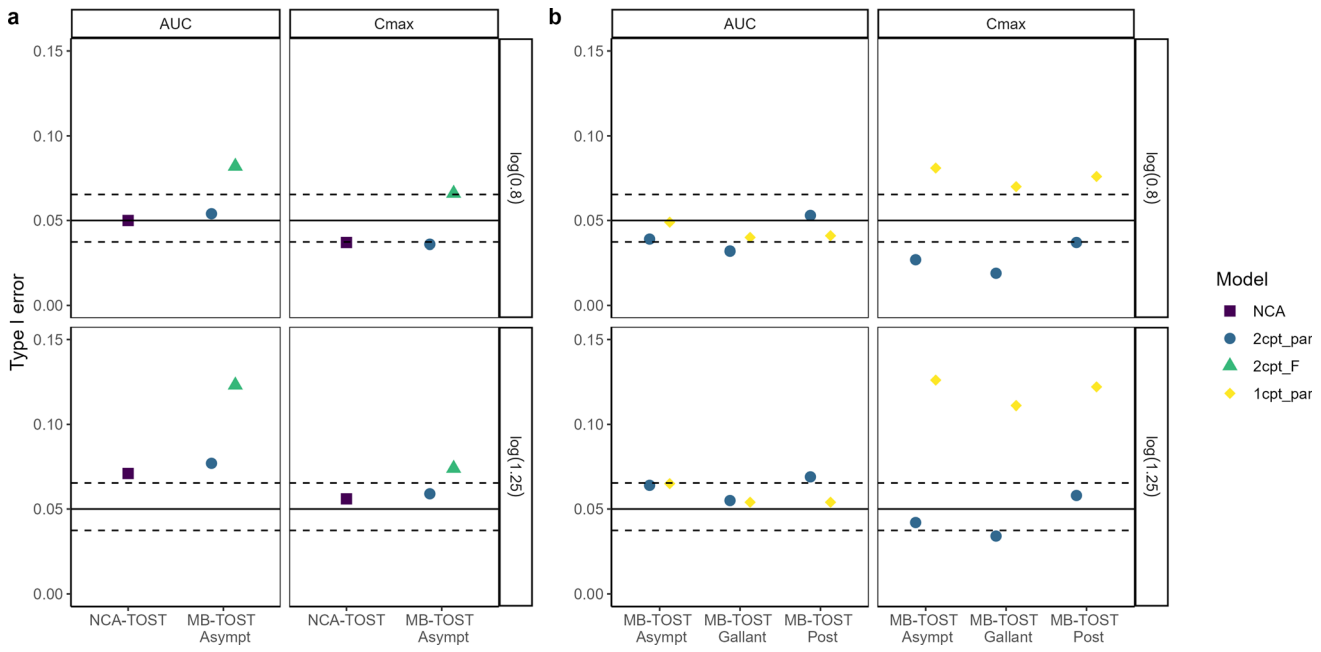


Fig. 4 Type I errors for *AUC* and C_{max}, under $H_{0:0.8}$ and $H_{0:1.25}$, on (a) rich design simulations with NCA-TOST and MB-TOST Asympt, and on (b) sparse design simulations with MB-TOST Asympt, Gallant and Post

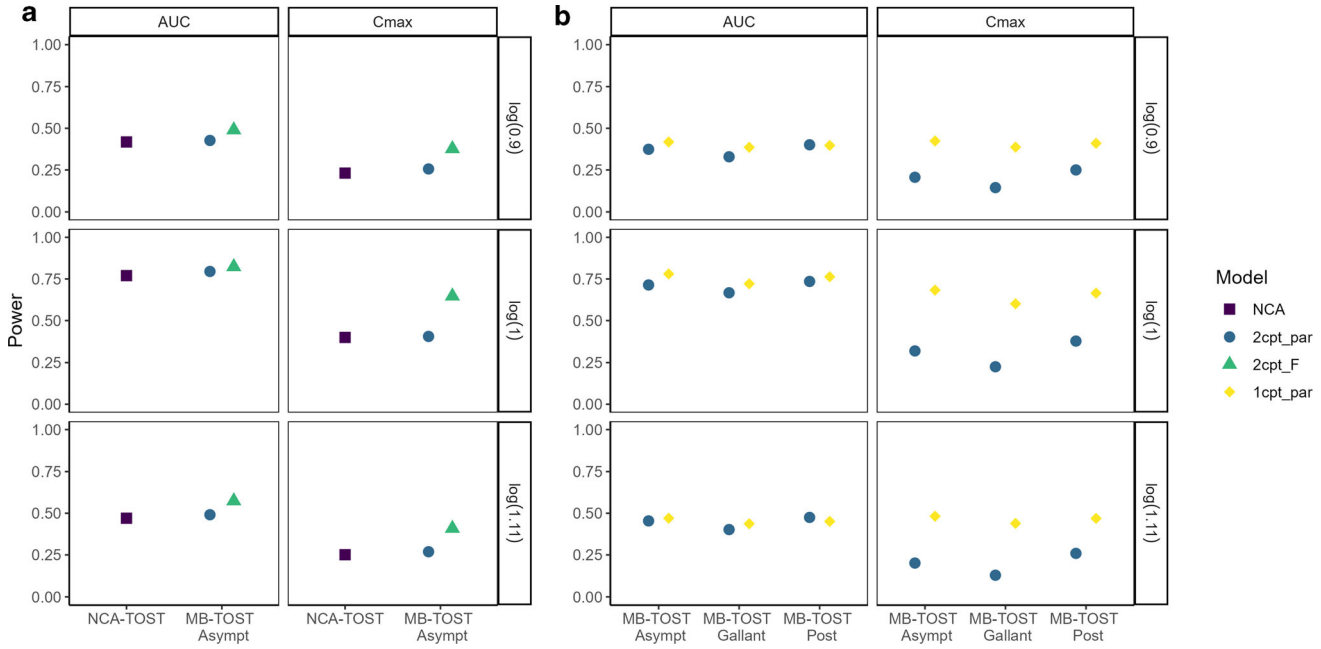


Fig. 5 Study power for AUC and C_{max}, under $H_{1:0.9}$, $H_{1:1}$ and $H_{1:1.11}$, on (a) rich design simulations with NCA-TOST and MB-TOST Asympt, and on (b) sparse design simulations with MB-TOST Asympt, Gallant and Post

regardless of the level of treatment effect considered: 0.25, 7, 20, 42, and 84 days post dose. On these sparse design simulations, we challenged MB-TOST by exploring the impact of a structural PK model misspecification: the model used to fit the data was either 2cpt_par or a misspecified one-compartment model with treatment effects estimated on all apparent parameters (1cpt_par). Table 1 represents the models fitted on sparse design simulations. As on the case study, three methods of computation of the SE were used on the sparse design simulations: Asympt, Gallant and Post.

We also explored the relevance of a PK model selection step, on the reference arm, on the BIC, prior to the equivalence test, on rich and sparse design simulations (two models to compare in each case). We observed the impact of this approach in terms of type I error.

Estimation Errors (EE) and Standard Errors (SE) of treatment effects were computed to evaluate the agreement between the estimations of NLMEM and the real values under which we simulated the data. Empirical SE were computed as the standard deviation on the 1000 estimates of each parameter in each scenario.

Implementation

A script detailing the analysis of one simulated dataset with *saemix* and *stan* is available on Zenodo (https://doi.org/10.5281/zenodo.6500556).

Results

Rich design

Figure 3a shows the boxplots of estimation errors (EE, top) and standard errors (SE, bottom) of the treatment effects on AUC and C_{max} in the different simulation scenarios with a rich design. We see that the treatment effects estimated with 2cpt_par (the structure of which is similar to the one of the model we simulated except treatment effects are estimated on all parameters) showed no bias and good precision : the EE were close to 0 and the estimated SE were close to the empirical SE. As expected on this rich design, NCA also provided good estimations of the treatment effects.

Figure 4a shows the type I errors of the TOST for AUC and C_{max} using NCA or a MB approach on rich design. The type I errors obtained with MB-TOST Asympt, using 2cpt_par, were similar to those obtained with NCA-TOST and close to the nominal value of 5%.

When we modelled the treatment effects differently from how they were simulated (i.e., using the misspecified model 2cpt_F), the model misspecification led to unsatisfactory results: the graph of EE (Fig. 3a Top) shows that the treatment effect on AUC was underestimated. In the scenario $H_{0:0.8}$, the relative bias in the estimation of the treatment effect on AUC is − 0.038, 0.016, and 0.016 for 2cpt_F, 2cpt_par, and NCA, respectively. In the scenario $H_{0:1.25}$, the relative bias in the estimation of the treatment effect on AUC is − 0.104, − 0.021, and − 0.030 for

2cpt_F, 2cpt_par, and NCA, respectively. The asymptotic SE boxplots appear lower than the empirical SE, though the relative root mean square errors (RMSE) are approximately − 0.35 for both 2cpt_par, 2cpt_F, and NCA, respectively. Increasing bias led to inflated type I errors we see in Fig. 4a.

A selection step using the BIC, prior to the test, on reference data helped in correcting the bias. Indeed, the difference of BIC between 2cpt_par and 2cpt_F ranged from − 22.1 to 10.3 with a median of − 3.4. The simulation model was found in 85% of the cases thanks to the selection procedure. Consequently, the type I error of MB-TOST was within the 95% prediction interval of the nominal value of 0.05 for each simulated level of treatment effect .

Study power for each study design, with NCA or MB-TOST, was low due to the parallel design of the clinical trial and the sample size (N=24 per arm, see Fig. 5).

Sparse design

On the simulations with sparse design, the treatment effects were still well estimated using 2cpt_par (Fig. 3b). Figure 4b shows the type I errors on sparse simulations with the MB approach where MB-TOST Asympt led to type I errors close to the 95% prediction interval of the nominal value of 0.05 with 2cpt_par.

When the structural PK model was misspecified, with only one compartment for the drug to distribute to, we observed a large inflation of the type I error on C_{max}, which we infer from Fig. 3b to be due to an underestimation of both the treatment effect and its SE. Indeed, in the scenario $H_{0:0.8}$, the relative bias in the estimation of the treatment effect on C_{max} is − 0.079 and 0.037 for 1cpt_par and 2cpt_par, respectively. In the scenario $H_{0:1.25}$, the relative bias in the estimation of the treatment effect on C_{max} is − 0.139 and − 0.015 for 1cpt_par and 2cpt_par, respectively. In the scenario $H_{0:0.8}$, the relative RMSE in the estimation of the treatment effect on C_{max} is − 0.40 and − 0.48 for 1cpt_par and 2cpt_par, respectively. In the scenario $H_{0:1.25}$, the relative RMSE in the estimation of the treatment effect on C_{max} is 0.40 and 0.47 for 1cpt_par and 2cpt_par, respectively.

MB-TOST Post gave results similar to MB-TOST Asympt (Fig. 4b). MB-TOST Gallant corrected the inflation of type I errors partly but could not correct for the bias in the estimations.

The numbers used to produce Figs. 4 and 5 are provided in Tables 6 and 7 in Appendix 3.

Here, a selection step using the BIC, prior to the test, to choose the number of compartments of the structural PK model on reference data, led to the selection of the simulated structural model in most cases (at least 99.0%). The difference of BIC between 2cpt_par and 1cpt_par ranged from -69.1 to 6.0 with a median of − 20.8. This allowed for control of type I error with MB-TOST.

We checked the assumption of normality of the test statistics under the null with Asympt in both rich and sparse design (data not shown).

Discussion

In this article, we compare the PK data gathered in participants treated with two formulations of gantenerumab, a monoclonal antibody for the treatment of Alzheimer's disease. The data used was originally collected to study the relative bioavailability of these two formulations. In this work, we use the data to compare the conventional NCA-TOST to the MB-TOST approach for PK equivalence testing. The data evaluated in our study is based on a parallel design instead of the more conventional crossover design in equivalence studies. The data is then used to generate a simulation study to explore the impact of sparse design and of model misspecification on the MB approach to test for PK equivalence.

After finding a dose effect on the pooled data, we performed the analyses separately on each study and dose evaluated. In our evaluation of these PK BE studies, we assume that the PK characteristics of the reference drug are well-known and the change of treatment does not affect the underlying PK structural model. In our simulations, we assumed the residuals are independent of the treatment covariate in the model. Also, we assumed that the study population would be adequately randomised to avoid imbalance between the treatment arms, so we did not evaluate the impact of covariates in our MB approaches. However, it is important to acknowledge that covariates would likely have a greater impact on a PK BE study with a parallel study design as compared to a crossover design. Moreover, adding covariates that affect the PK would decrease the between subject variability. Thus, future research may be warranted in this area. Using the real data, we evaluated the models selected to assess the assumptions made on the residuals as part of model building process. Also, the distributions of the MB-TOST statistics under the null from our simulations were verified as recommended by Shen et al. [12].

Using the original data, the NCA and MB-TOST approaches generally provide consistent results with the original rich design and the MB-TOST approach provides consistent results after sparsifying the data.

Previous studies by Dubois et al. and Reijers et al. have shown that MB approaches evaluating studies with a crossover design [19] and a parallel design [20], respectively, have performed as well as NCA methods for

biosimilarity studies in the case of rich sampling. Dubois et al. also explored MB approaches on a sparse version of their data.

In the present study, we performed a simulation study to explore the influence of the design and model specification on the performance of the approaches and the relevance of the model selection.

Here, as in the previous works [8, 10], we only considered average BE. With average BE, by contrast with individual and population BE [21], we only take into account the average treatment effect at population level. Population BE would also take into account the variability of this effect, and individual BE would take into account the within-subject and subject-by-formulation variabilities. In this parallel study, population BE could be done, because variability is not correctly accounted for. Individual BE requires replicated cross-over studies so this approach would not be feasible on our data.

In the simulation study, when using the simulated model, MB-TOST Asympt achieved controlled type I errors that were similar to those obtained with NCA-TOST on rich designs. These results complement previous studies showing the efficiency of MB approaches for equivalence tests [8]. In general, regulatory authorities recommend that PK sampling includes 12–18 samples with at least three sampling points after the peak [3, 4]. These recommendations present unique challenges for PK studies with sparse designs. Indeed, the sparse design we extracted from the full design did not comply with those requirements, consequently we did not apply NCA-TOST to the datasets simulated with the sparse design. In this setting, we used MB-TOST as it relies on NLMEM which demonstrated improved accuracy of the estimates in particular when dealing with sparse designs [22]. However, Dubois et al. [8] showed that MB tests can lead to an inflation of the type I error because of an underestimation of the standard error of treatment effects when it is estimated asymptotically on sparse design with high variability. As such, Loingeville et al. proposed and evaluated methods of correction of the standard errors in MB studies, with satisfying results [10]. Notably, they compared the three methods we present here, along with a bootstrap method, but considering one model only (one-compartment) and without exploring the interest of model selection. One of the correction methods for SE in MB studies, Gallant, has been used outside the context of BE. To illustrate, Bertrand et al. [23] considered various methods of correcting the number of degrees of freedom in a Student distribution and found that the Gallant correction was a good compromise in NLMEM to handle the information carried by the number of subjects. In this research, the use of Gallant leads to the same reference distribution in MB-TOST Gallant as in NCA-TOST, instead of the Gaussian distribution used in MB-TOST Asympt and MB-TOST Post.

Our results showed that MB-TOST Asympt was adequate with sparse designs, with a slightly conservative type I error for C_{max} that was corrected using MB-TOST Post. Here, the Post method was used only as an alternative to produce SE. This algorithm is sensitive to the choice of prior distribution, and this could be further investigated. Nevertheless, the performance of the different MB methods were very similar on the sparse design in our work. Actually, we obtained asymptotic SE close to the empirical SE which explains that the results of the tests were not affected by the correction methods.

As the treatment effect on C_{max} is not directly linked to the parameters, we estimated it via simulations. We used an approximation simulating the treatment effect on a profile using the mean parameters; in Appendix 2, we provide a more computationally intense method. In this example, the first approximation gave equivalent results, but the second approximation should be used in the presence of higher variability.

A sparse design is commonly seen in PK BE studies for ophthalmic drug products where only one sample of aqueous humor is collected from the eye at a single time point. Currently, FDA recommends a non-parametric bootstrap NCA-based approach or a parametric method in the BE assessment for these drug products [5, 6]. In our assessment, we evaluated a study design with only five sampling points, which were optimally selected using PFIM. One limitation of this work is that we did not evaluate the performance of the classical NCA-TOST approach on sparse design as our focus was to evaluate the MB-TOST approach. The limitation of few sampling points per subject apply to both approaches as the NCA-TOST approach may become less accurate when there are few sample points whereas the MB-TOST approach may select a wrong PK structure model-based. Indeed, in our application study, the model parameter estimates varied considerably between the rich and sparse design (see Table 4 in Appendix 3).

The MB approach was previously evaluated only in simulations assuming the true model to be known [8–10]. In our present study, we investigate this question by fitting PK models different from the one used to simulate the data. The two-compartment model with treatment effects estimated on ka and F only, fitted on the rich designs, is the same structural PK model as the simulated one but with an alternative way of parameterising the treatment effects. It has already been used in other studies as the simulated model [11]. Here, it cannot properly fit the data as θ_F reflects a treatment effect on all distribution and elimination parameters, which does not agree with the way we simulated the data (i.e., without an effect on the peripheral

clearance and volume). This explains why the effect on AUC is underestimated. With biosimilars, differences in the PK characteristics of a drug may be due to factors other than differences in the absorption phase. In contrast, the misspecified one-compartment model with treatment effects estimated on all apparent parameters, fitted on sparse designs, is a different PK structural model than the simulated one. The choice of the number of compartments is an essential step in structural PK model building. It is very sensitive to the study design and is therefore highly susceptible to misspecification. A less complex model would more likely be selected on a real study in case of non-optimised sparse design because of the lack of information. The treatment effect on AUC is still quite well estimated because it is a mean PK parameter, unlike C_{max} which is more sensitive to the misspecification because it is driven by only one point.

Adding a step of model selection on the reference data allowed to select the simulated model in most cases. When the simulated model is not selected, the difference of BIC between the models is very low. In this case, we assume that the misspecified model can adequately describe the data because the overall type I errors are controlled after the selection step. Most importantly, we mimic a real model development setting, where model selection is always part of the PK analysis. The selection of the model is based on data from the reference product only in order to avoid a bias in the MBBE evaluation from using test product data to fit the model used in the BE assessment. However, it is possible that using the reference arm for the model selection, and then for the assessment of a treatment effect, could inflate the type I error of the BE assessment. Therefore, this issue may warrant further investigation. Moreover, this can cause a problem if the underlying PK model is different in the test arm. Another limit of our simulation study is that we only selected between two different PK models. We could extend this approach to test and compare more features of the PK structural (absorption and elimination phases) and/or variability (random effects and residual errors) models as we performed in the real case study. We could also consider more complex data exhibiting, for example, double peaks which can be very challenging to evaluate, or that the magnitude of the variability depends on the treatment arm. It is likely that, in this case, the simulated model would not be recovered as often, potentially affecting the type I errors. However, the impact may not be very large if there were more candidate models in the selection step, as the models retained would have adequate goodness of fit. Hence, the estimated AUC and C_{max} would all be acceptable despite the diversity of underlying structural PK models. It would therefore be interesting to further evaluate the impact of small model variations on the model selection process and the ensuing ability to estimate C_{max} and AUC and the associated treatment effects. Competing models could also be taken into account via model averaging, which has been shown to work at least as well as model selection in dose finding studies using NLMEM [24, 25], as it allows to take into account the uncertainty on the model.

The methods presented in our study may be applied to PK similarity for large molecules (i.e., biologics) as well as PK BE studies for small molecules. By re-scaling the time frame, we could transpose our simulation settings and results to a BE study framework. In both cases, the test product or new drug contains the same active substance as the reference product, for which the PK is likely well characterised. To shorten the development phase of the new drug, it is recommended to demonstrate that there is no difference of treatment effect on the PK. In both cases, MB approaches may serve as an alternative method to NCA for sparse designs, and thus, are increasingly explored [26]. However, it is acknowledged that the performance of NCA and MB methods will drop in case of large inter-individual variability in PK or deviations from working assumptions.

Thus, we propose the use of MB-TOST when NCA-TOST may not be feasible or reasonable, as MB approaches are more informative and flexible than NCA.

This is consistent with recent proposals for MB approaches to serve as an alternative BE approach in generic drug development in situations for which conventional BE approaches are not feasible [27].

Conclusions

Our novel MB BE approach appears to be a robust alternative to the conventional NCA approach provided that the PK model is correctly specified and the test drug has the same PK structural model as the reference drug. Our simulation studies show that the selection of the PK model is a key step in the implementation of a model-based approach for PK equivalence studies. However, MB methods rely on numerous assumptions which need further investigation to determine when MB could offer a viable alternative to NCA in the context of PK BE studies.

Appendix 1 : Pharmacokinetic models equations and parameters

One compartment model with zero-order absorption, linear elimination

$$C(t) = \begin{cases} \frac{D}{Tk_0 CL}\left(1 - \exp(-\frac{CL}{V_1}t)\right) & \text{if } t \le Tk_0 \\ \frac{D}{Tk_0 CL}\left(1 - \exp(-\frac{CL}{V_1}Tk_0)\right)\exp\left(-\frac{CL}{V_1}(t - Tk_0)\right) & \text{if } t > Tk_0 \end{cases}$$

with:

- $C(t)$ the concentration at time t;
- D the dose administered;
- Tk_0 the absorption duration;
- V_1 the volume of distribution of the compartment;
- CL the clearance of the drug;
- $k = \frac{CL}{V_1}$ the elimination rate constant.

Here, there are $l = 3$ parameters: $\mu = c(Tk_0, V_1, CL)$.

One compartment model with zero-order absorption, linear elimination, with a lag time

$$C(t) = \begin{cases} 0 & \text{if } t \le T_{lag} \\ \frac{D}{Tk_0 CL}\left(1 - \exp(-\frac{CL}{V_1}(t - T_{lag}))\right) & \text{if } T_{lag} < t \le T_{lag} + Tk_0 \\ \frac{D}{Tk_0 CL}\left(1 - \exp(-\frac{CL}{V_1}Tk_0)\right)\exp\left(-\frac{CL}{V_1}(t - T_{lag} - Tk_0)\right) & \text{if } t > Tk_0 \end{cases}$$

The lag time T_{lag} adds a period of latency before the concentration starts rising. It works the same for all models.

Here, there are $l = 4$ parameters: $\mu = c(Tk_0, V_1, CL, T_{lag})$.

One compartment model with first-order absorption, linear elimination

The model $1cpt\_par$ represented in Table 1 corresponds to the equation:

$$C(t) = \frac{D}{V_1}\frac{ka}{\frac{CL}{V_1} - ka}\left(\exp(-kat) - \exp(-\frac{CL}{V_1}t)\right) \tag{10}$$

with ka the absorption constant rate.

Here, there are $l = 3$ parameters: $\mu = c(ka, V_1, CL)$.

Two compartment model with zero-order absorption, linear elimination

$$C(t) = \begin{cases} \frac{D}{Tk_0}\left(\frac{A}{\alpha}(1 - \exp(-\alpha t)) + \frac{B}{\beta}(1 - exp(-\beta t))\right) & \text{if } t \le Tk_0 \\ \frac{D}{Tk_0}\left(\frac{A}{\alpha}(1 - \exp(-\alpha Tk_0))\exp(-\alpha(t - Tk_0)) \right. & \text{if } t > Tk_0 \\ \quad \left. + \frac{B}{\beta}(1 - exp(-\beta Tk_0))\exp(-\beta(t - Tk_0))\right) & \end{cases}$$

with:

- $A = \frac{1}{V_1}\frac{k_{21} - \alpha}{\beta - \alpha}$ the first macro-constant;
- $B = \frac{1}{V_1}\frac{k_{21} - \beta}{\alpha - \beta}$ the second macro-constant;
- $\alpha = \frac{k_{21}k}{\beta}$ the first rate constant;
- $\beta = \frac{1}{2}\left(k_{12} + k_{21} + k - \sqrt{(k_{12} + k_{21} + k)^2 - 4k_{21}k}\right)$ the second rate constant;
- $k_{12} = \frac{Q}{V_1}$ the distribution rate constant between the principal and the peripheral compartment;
- $k_{21} = \frac{Q}{V_2}$ the distribution rate constant between the peripheral and principal compartment;
- Q the inter-compartmental clearance;
- V_1 the volume of distribution of the principal compartment;
- V_2 the volume of distribution of peripheral compartment.

Here, there are $l = 5$ parameters: $\mu = c(Tk_0, V_1, CL, V_2, Q)$.

Two compartment model with first-order absorption, linear elimination

The model used to generate the data in the simulation, and the models $2cpt\_par$ and $2cpt\_F$, are represented in Table 1 and correspond to the equation:

$$C(t) = D\left(A\exp(-\alpha t) + B\exp(-\beta t) - (A + B)\exp(-kat)\right) \tag{11}$$

with:

- $A = \frac{ka}{V_1}\frac{k_{21} - \alpha}{(ka - \alpha)(\beta - \alpha)}$;
- $B = \frac{ka}{V_1}\frac{k_{21} - \beta}{(ka - \beta)(\alpha - \beta)}$.

Here, there are $l = 5$ parameters: $\mu = c(ka, V_1, CL, V_2, Q)$.

Note: in the two compartment models under study here, the clearance occurs only from the central compartment via the clearance constant CL. The drug in the peripheral compartment can only return to the central compartment via the inter-compartmental clearance constant Q.

Parameterisation with *F*

Implicit in the equations above is the notion of bioavailability, defined as the fraction of dose reaching the system. Including bioavailability as an explicit parameter F corresponds to replacing D with $D \times F$ in the equations above. We can easily see from equations 10 and 11 that this is equivalent to dividing both CL and V_1 by F, so that the

latter, when estimated from data collected after oral absorption, are called apparent clearance and volume, and sometimes denoted CL/F and V_1/F to show their dependency on F.

This leads to an alternative way of parameterising the model, by including F in the model. Because F cannot be identified without intravenous data, we fix the population value at F=1 and only allow for some inter-individual variability. We put a treatment effect only on the absorption parameters and F. Also, no correlation is allowed between the random effects of volumes and clearances, as these correlations are assumed to be carried by F. This parameterisation allows to compute fewer treatment effect coefficients.

Appendix 2: Method to compute the treatment effect on C_{max} and its SE

As part of a PK equivalence analysis, after fitting a NLMEM, we want to compute treatment effects on the PK parameters of interest and their SE when there is no explicit relationship with the direct parameters of the model. We simulate typical concentration versus time profiles, taking into account the variance covariance matrix of the fixed parameters.

Let $c(\mu, \theta)$ be the vector of fixed effects and treatment effects obtained with the NLMEM and M_F^{-1} the asymptotic variance-covariance matrix of the fixed effects and treatment effects, obtained by solving the Fisher Information Matrix of the model.

We simulate K parameter sets with a multivariate normal distribution.

$$c(\mu_k, \theta_k) \sim \mathcal{N}(c(\mu, \theta), M_F^{-1})$$

with k=1,...,K, here K=1000.

For each parameter set, we compute a profile of concentrations with the population parameters and a short time step, under reference treatment and under test treatment. For example, with a two-compartment model with first order absorption:

$$C_k^R = C(time, ka_k, CL_k, V1_k, Q_k, V2_k)$$

and

$$C_k^T = C(time, ka_k\ e^{\theta_{ka,k}}, CL_k\ e^{\theta_{CL,k}}, V1_k\ e^{\theta_{V1,k}}, Q_k\ e^{\theta_{Q,k}}, V2_k\ e^{\theta_{V2,k}})$$

Then we compute the treatment effect as the log ratio of the PK parameter of interest under test and reference treatment. For instance, with $C_{max,k}^R$ the maximum over the vector C_k^R and $C_{max,k}^T$ the maximum over the vector C_k^T:

$$\theta_{C_{max,k}} = log\left(\frac{C_{max,k}^T}{C_{max,k}^R}\right)$$

We obtain a vector of K estimated treatment effects. We estimate the global treatment effect as the mean of this vector $mean(\theta_{Cmax,k})$, and its standard error as the standard deviation of this vector $sd(\theta_{Cmax,k})$.

Consequently, the geometric mean ratio is computed as the exponential of the point estimate of θ computed, $GMR_{Cmax} = exp(mean(\theta_{Cmax,k}))$ for instance.

We evaluate the performance of this method on the estimated standard error of the treatment effect on AUC, because it has an explicit formulation with direct parameters:

$$\begin{aligned} AUC &= \frac{D}{CL} \\ AUC_k^R &= \frac{D}{CL_k} \\ AUC_k^T &= \frac{D}{CL_k\ e^{\theta_{CL,k}}} \\ \theta_{AUC_k} &= log\left(\frac{AUC_k^T}{AUC_k^R}\right) = -\theta_{CL,k} \\ \mathbb{E}(sd(\widehat{\theta_{AUC}})) &= \mathbb{E}[sd((\theta_{CL,k}))] = sd(\theta_{CL}) \end{aligned}$$

That shows that the method would give an estimate of the standard error of θ_{AUC} consistent with the method based on the explicit link with the direct parameters. However, we compute θ_k as the treatment effect on a concentration profile in the mean parameters. The definition of θ is the mean of the treatment effect on each individual profile. In this example, the relationship between θ_{AUC} and θ_{CL} is linear, so these two quantities are equal, but this does not apply for C_{max}. A more accurate simulation method would take into account the interindividual variability by simulating individual time-concentration profiles using the variance-covariance matrix of the random effects. This method would be much more computationally intensive because simulating too few participants would lead to the poor estimation of the variability of the treatment effect, even more if the random variability of the direct parameters influencing C_{max} is high.

In this study, a comparison between the two approximations showed us that they gave similar results in terms of estimation of θ for C_{max} and its SE: it seems its relationship with the direct parameters treatment effect is close enough to linear, so we decided to use it for its computational conservativeness. It is likely that in a study with more random effects, for instance if there is some intra-individual variability, the second method would be preferable.

We also checked that this simulation method gave results similar from the FIM-based method for θ_{AUC}, but decided to keep the FIM-based method for θ_{AUC} because it is less time-consuming.

Appendix 3: Tables

Table 4 Parameters estimates and treatment effect coefficients (relative standard errors), given by *saemix* on all separate studies, with the original and sparse design

	Parameters (RSE, %)					
	S1-105 Rich	S1-105 Sparse	S1-225 Rich	S1-225 Sparse	S2-225 Rich	S2-225 Sparse
μ_{Tk0}						3.892 (7.6)
θ_{Tk0}						0.076 (141.7)
μ_{ka} (d)	0.327 (16.0)	0.844 (18.0)	0.469 (15.9)	2.947 (57.5)	0.361 (12.6)	
θ_{ka}	0.418 (54.0)	0.271 (110.7)	0.272 (82.6)	-1.220 (50.9)	0.019 (953.5)	
$\mu_{CL/F}$ ($L.d^{-1}$)	0.632 (6.0)	0.622 (6.1)	0.632 (5.4)	0.621 (5.6)	0.681 (8.8)	0.698 (8.9)
$\theta_{CL/F}$	− 0.075 (113.7)	− 0.052 (166.3)	− 0.070 (108.6)	− 0.032 (245.9)	0.013 (1013.4)	− 0.003 (3806.2)
$\mu_{V_1/F}$ (L)	11.611 (14.8)	21.858 (6.0)	14.698 (9.1)	19.709 (6.1)	15.615 (12.8)	23.181 (8.6)
$\theta_{V_1/F}$	0.200 (102.4)	− 0.014 (622.2)	0.194 (64.3)	0.077 (113.6)	0.095 (200.2)	0.108 (114.4)
$\mu_{Q/F}$ ($L.d^{-1}$)	1.882 (30.2)		0.415 (30.9)		0.601 (28.9)	
$\theta_{Q/F}$	0.421 (114.3)		− 0.460 (140.4)		0.773 (51.1)	
$\mu_{V_2/F}$ (L)	9.343 (11.5)		4.828 (15.0)		6.234 (13.3)	
$\theta_{V_2/F}$	− 0.235 (90.9)		− 0.444 (67.1)		0.126 (166.2)	
$\mu_{T_{lag}}$ (d)	0.062 (26.5)	0.129 (13.0)	0.037 (31.8)	0.208 (11.1)	0.050 (23.9)	
$\theta_{T_{lag}}$	− 0.744 (68.9)	− 0.209 (149.1)	− 0.079 (608.5)	− 0.974 (37.5)	− 0.529 (69.7)	
ω_{Tk0}						0.226 (23.4)
ω_{ka}	0.492 (13.1)	0.548 (15.1)	0.659 (11.2)	0.807 (11.0)	0.448 (11.9)	
$\omega_{CL/F}$	0.287 (10.7)	0.283 (11.6)	0.257 (10.7)	0.261 (11.3)	0.438 (10.4)	0.441 (10.4)
$\omega_{V_1/F}$	0.430 (12.2)	0.255 (13.1)	0.328 (11.4)	0.268 (13.0)	0.526 (10.8)	0.419 (10.7)
$\omega_{Q/F}$	1.129 (17.0)				0.412 (41.0)	
$\omega_{T_{lag}}$	0.926 (19.8)	0.258 (42.6)	0.869 (25.0)		1.005 (14.8)	
ρ_{CL/V_1}	0.694 (34.1)	0.843 (31.5)	0.762 (30.5)	0.844 (31.3)	0.931 (26.7)	0.940 (26.3)
σ_a					0.064 (11.8)	0.048 (14.6)
σ_b	0.168 (3.9)	0.194 (7.4)	0.153 (3.7)	0.171 (7.2)	0.112 (4.9)	0.099 (9.5)

Table 5 Gantenerumab analysis—Geometric mean ratios (GMR), their 90% confidence interval and the $p-value$ of the test, for AUC and C_{max}, with NCA-TOST and MB-TOST Asympt on original data and with MB-TOST Asympt, Gallant and Post on sparse data

Dataset	Design	Method	PK parameter	GMR	90% CI	p
S1-105	Rich	NCA-TOST	AUC	1.068	[0.924 ; 1.236]	**0.038**
			C_{max}	0.997	[0.836 ; 1.189]	**0.021**
		MB-TOST Asympt	AUC	1.077	[0.937 ; 1.239]	**0.040**
			C_{max}	0.972	[0.809 ; 1.167]	**0.040**
	Sparse	MB-TOST Asympt	AUC	1.054	[0.913 ; 1.215]	**0.025**
			C_{max}	1.039	[0.907 ; 1.189]	**0.012**
		MB-TOST Gallant	AUC	1.054	[0.905 ; 1.227]	**0.033**
			C_{max}	1.039	[0.899 ; 1.200]	**0.018**
		MB-TOST Post	AUC	1.054	[0.902 ; 1.231]	**0.035**
			C_{max}	1.039	[0.898 ; 1.201]	**0.018**
S1-225	Rich	NCA-TOST	AUC	1.080	[0.947 ; 1.231]	**0.034**
			C_{max}	0.914	[0.771 ; 1.085]	0.098
		MB-TOST Asympt	AUC	1.073	[0.946 ; 1.216]	**0.023**
			C_{max}	0.925	[0.787 ; 1.087]	0.070
	Sparse	MB-TOST Asympt	AUC	1.033	[0.906 ; 1.177]	**0.008**
			C_{max}	0.867	[0.749 ; 1.003]	0.184
		MB-TOST Gallant	AUC	1.033	[0.892 ; 1.196]	**0.017**
			C_{max}	0.867	[0.736 ; 1.021]	0.208
		MB-TOST Post	AUC	1.033	[0.904 ; 1.180]	**0.009**
			C_{max}	0.867	[0.749 ; 1.003]	0.184
S2-225	Rich	NCA-TOST	AUC	0.971	[0.782 ; 1.205]	0.070
			C_{max}	0.858	[0.674 ; 1.093]	0.314
		MB-TOST Asympt	AUC	0.988	[0.801 ; 1.218]	**0.049**
			C_{max}	0.863	[0.695 ; 1.071]	0.284
	Sparse	MB-TOST Asympt	AUC	1.003	[0.812 ; 1.240]	**0.044**
			C_{max}	0.899	[0.734 ; 1.102]	0.171
		MB-TOST Gallant	AUC	1.003	[0.796 ; 1.265]	0.059
			C_{max}	0.899	[0.720 ; 1.123]	0.190
		MB-TOST Post	AUC	1.003	[0.807 ; 1.247]	**0.048**
			C_{max}	0.899	[0.728 ; 1.111]	0.182

Significant p-values are highlighted in bold

Table 6 Type I errors for AUC and C_{max}, under $H_{0:0.8}$ and $H_{0:1.25}$, on rich (R) design simulations with NCA-TOST and MB-TOST Asympt, and on sparse (S) design simulations with MB-TOST Asympt, Gallant and Post

			Type I error	
			AUC	C_{max}
$R_{0.8}$	NCA-TOST		0.05	0.037
	2cpt_par	MB-TOST Asympt	0.054	0.036
	2cpt_F	MB-TOST Asympt	0.082	0.066
$R_{1.25}$	NCA-TOST		0.071	0.056
	2cpt_par	MB-TOST Asympt	0.077	0.059
	2cpt_F	MB-TOST Asympt	0.123	0.074
$S_{0.8}$	2cpt_par	MB-TOST Asympt	0.039	0.027
		MB-TOST Gallant	0.032	0.019
		MB-TOST Post	0.054	0.038
	1cpt_par	MB-TOST Asympt	0.049	0.081
		MB-TOST Gallant	0.040	0.070
		MB-TOST Post	0.041	0.077
$S_{1.25}$	2cpt_par	MB-TOST Asympt	0.064	0.042
		MB-TOST Gallant	0.055	0.034
		MB-TOST Post	0.069	0.058
	1cpt_par	MB-TOST Asympt	0.065	0.126
		MB-TOST Gallant	0.054	0.111
		MB-TOST Post	0.055	0.123

Table 7 Study power to detect a treatment effect on AUC and C_{max}, under $H_{1:0.9}$, $H_{1:1}$ and $H_{1:1.11}$, on rich (R) design simulations with NCA-TOST and MB-TOST Asympt, and on sparse (S) design simulations with MB-TOST Asympt, Gallant and Post

			Power	
			AUC	C_{max}
$R_{0.9}$	NCA-TOST		0.418	0.231
	2cpt_par	MB-TOST Asympt	0.427	0.256
	2cpt_F	MB-TOST Asympt	0.490	0.377
R_1	NCA-TOST		0.770	0.401
	2cpt_par	MB-TOST Asympt	0.795	0.407
	2cpt_F	MB-TOST Asympt	0.823	0647
$R_{1.11}$	NCA-TOST		0.470	0.251
	2cpt_par	MB-TOST Asympt	0.491	0.269
	2cpt_F	MB-TOST Asympt	0.574	0.409
$S_{0.9}$	2cpt_par	MB-TOST Asympt	0.374	0.206
		MB-TOST Gallant	0.329	0.144
		MB-TOST Post	0.409	0.251
	1cpt_par	MB-TOST Asympt	0.418	0.424
		MB-TOST Gallant	0.386	0.387
		MB-TOST Post	0.4399	0.411
S_1	2cpt_par	MB-TOST Asympt	0.714	0.320
		MB-TOST Gallant	0.667	0.225
		MB-TOST Post	0.739	0.384
	1cpt_par	MB-TOST Asympt	0.780	0.683
		MB-TOST Gallant	0.721	0.601
		MB-TOST Post	0.762	0.667
$S_{1.11}$	2cpt_par	MB-TOST Asympt	0.454	0.201
		MB-TOST Gallant	0.402	0.128
		MB-TOST Post	0.4731	0.255
	1cpt_par	MB-TOST Asympt	0.470	0.482
		MB-TOST Gallant	0.437	0.439
		MB-TOST Post	0.450	0.467

Appendix 4: Figures

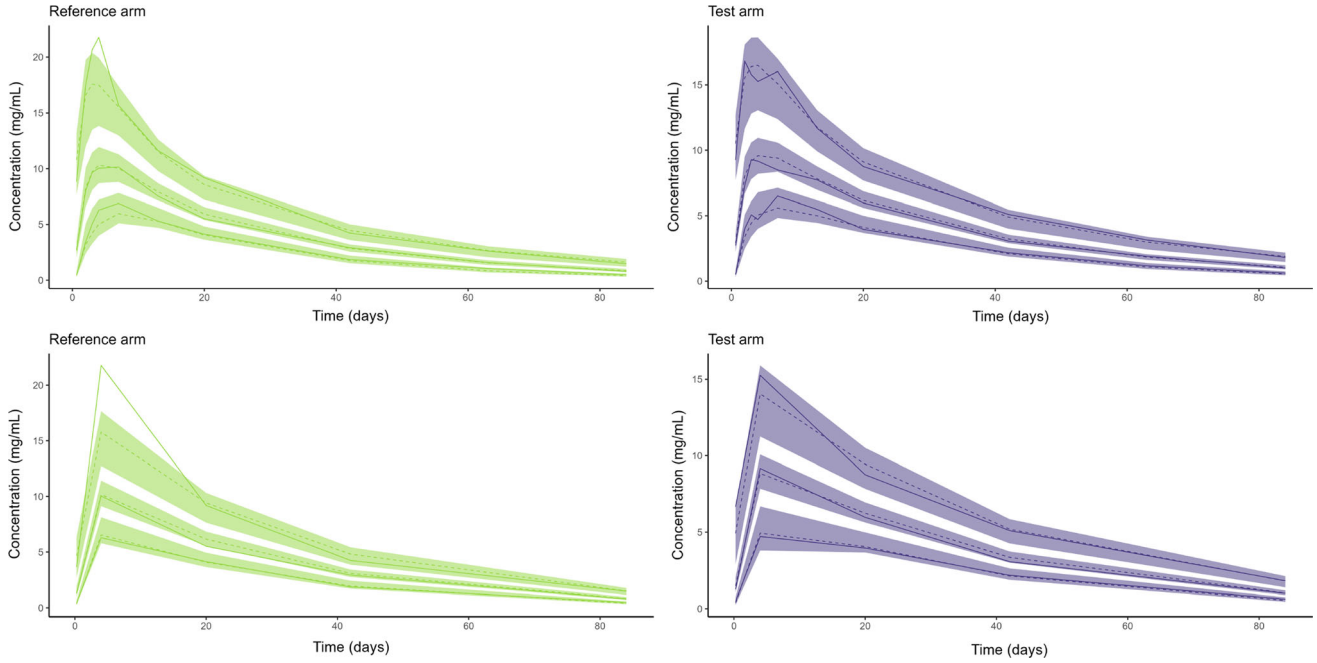


Fig. 6 Visual predictive check for the S1-225 study reference (left) and test (right) arm, on original (top) and sparse (bottom) design. Note: the predicted 5%, 50% and 95% percentiles are shown as dashed lines; the observed percentiles as solid lines (colour figure online)

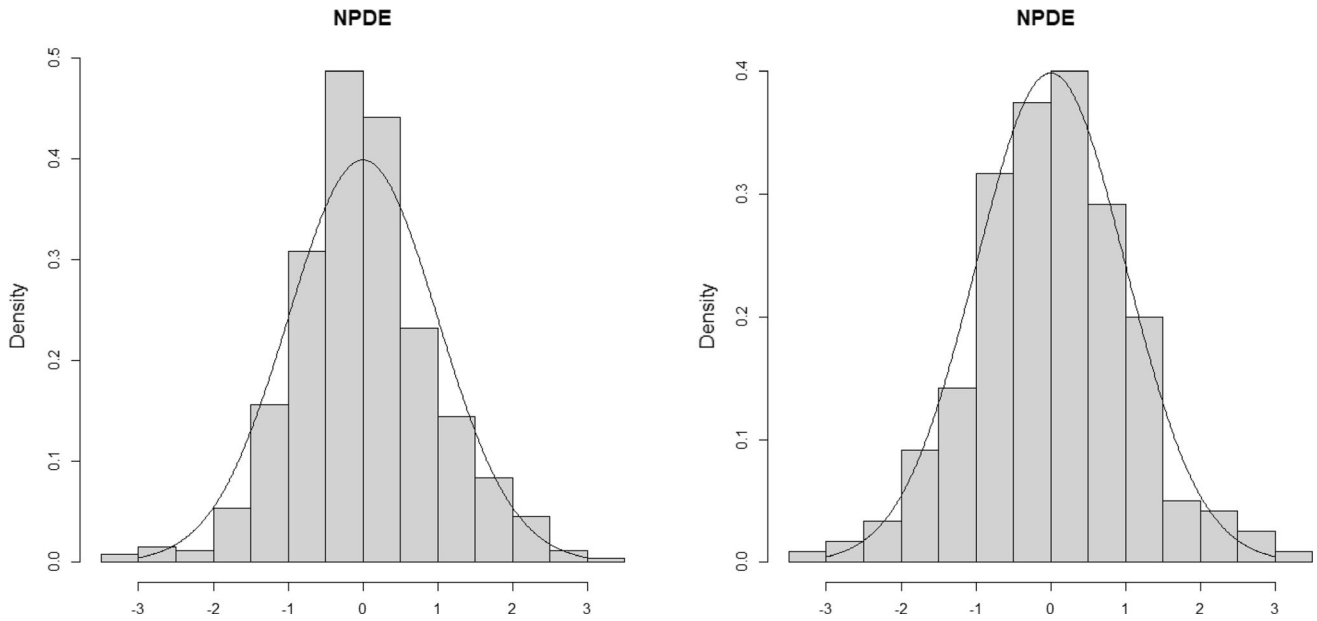


Fig. 7 Distributions of normalised predictive distribution errors (NPDE) for the S1-225 study on original (left) and sparse (right) design

Definition of Visual Predictive Checks (from Monolix documentation) :

The VPC (Visual Predictive Check) offers an intuitive assessment of misspecification in structural, variability, and covariate models. The principle is to assess graphically whether simulations from a model of interest are able to reproduce both the central trend and variability in the observed data, when plotted versus an independent variable (typically time). It summarises in the same graphic the structural and statistical models by computing several quantiles of the empirical distribution of the data after having regrouped them into bins over successive intervals. More precisely, the goal is to compare the two following elements:

Empirical percentiles: percentiles of the observed data, calculated either for each unique value of time, or pooled by adjacent time intervals (bins).

Theoretical percentiles: percentiles of simulated data are computed from multiple Monte Carlo simulations with the model of interest and the design structure of the original dataset (i.e., dosing, timing, and number of samples). For each simulation, the same percentiles are computed across the same bins as for empirical percentiles. Prediction intervals for each percentile are then estimated across all simulated data and displayed as colored areas.

If the model is correct, the observed percentiles should be close to the predicted percentiles and remain within the corresponding prediction intervals.

Definition of normalised prediction distributions errors (NPDE) from Comets et al. [28]:

The cumulative distribution function (cdf) of the predictive distribution of the concentrations observed can be computed using Monte–Carlo simulations.

We define the prediction discrepancies (pd) as the value of this cdf at each observation.

pd are computed as the percentiles of each observation in the empirical distribution of the simulations.

By construction, pd are expected to follow $\mathcal{U}(0,1)$, but only in the case of one observation per subject; within-subject correlations introduced when multiple observations are available for each subject induce an increase in the type I error of the test. To correct for this correlation, we compute the empirical mean and empirical variance-covariance matrix over the simulations.

Decorrelation is performed simultaneously for simulated data and for observed data. Decorrelated pd are then obtained using the same formula but with the decorrelated data, and we call the resulting variables prediction distribution errors (pde).

If the number of Monte-Carlo simulations is large enough, the distribution of the prediction distribution errors should follow a uniform distribution over the interval [0,1] by construction of the cdf. Normalised prediction distribution errors can then be obtained using the inverse function of the normal cumulative density function. By construction, NPDE follow the $\mathcal{N}(0,1)$ distribution without any approximation and are uncorrelated within an individual.

Acknowledgements The authors would like to thank all their collaborators in the project "Evaluation of model-based bioequivalence (MBBE) statistical approaches for sparse design PK studies" under the FDA contract 75F40119C10111. They also thank Hervé Le Nagard and Lionel de la Tribouille for the use of the CATIBioMed calculus facility. The illustrative example data were obtained from studies sponsored by F Hoffmann-La Roche. We thank the participants and investigators who participated in these studies. We are grateful to the gantenerumab project team, which allowed us to use these data and provided comments to improve the quality of this manuscript.

Disclaimer This work reflects the views of the author and should not be construed to represent the FDA's views or policies.

Author contributions MG, EC and JB wrote the manuscript. FM, CH, SS, MD, KF, WS, GS, SG, LZ, LF and FM criticially revised the manuscript. FM and JB designed the research. MG, EC and JB performed the research. MG simulated and analysed the data.

Funding This work was supported by the U.S. Food and Drug Administration (FDA) under contract 75F40119C1011. The authors thank the FDA for this funding.

Declarations

Conflict of interest Francois Mercier and Carsten Hofmann are Roche employees. Francois Mercier and Carsten Hofmann declare no conflict of interest. IAME laboratory has PhD students funded by Roche but not for this project.

References

1. GDUFA (2022) Generic Drug User Fee Amendments (GDUFA) science and research priority initiatives for Fiscal Year (FY) 2022. https://www.fda.gov/media/154487/download
2. Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15:657–680. https://doi.org/10.1007/BF01068419
3. U.S. Food and Drug Administration (2021) Bioequivalence studies with pharmacokinetic endpoints for drugs submitted under an ANDA guidance for industry. https://www.fda.gov/media/87219/download
4. European Medicines Evaluation Agency (2010) Guideline on the investigation of bioequivalence. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf
5. U.S. Food and Drug Administration (2016) Draft guidance on dexamethasone; tobramycin. https://www.accessdata.fda.gov/drugsatfda_docs/psg/Dexamethasone;%20Tobramycin_ophthalmic%20ointment_RLD%20050616_RV06-16.pdf
6. U.S. Food and Drug Administration (2018) Draft guidance on loteprednol etabonate. https://www.accessdata.fda.gov/

drugsatfda_docs/psg/Loteprednol%20Etabonate_draft_Ophthalmic%20drops%20susp_RLD%2020583_RC02-18.pdf
7. Dubois A, Gsteiger S, Pigeolet E, Mentré F (2010) Bioequivalence tests based on individual estimates using non-compartmental or model-based analyses: evaluation of estimates of sample means and type I error for different designs. Pharm Res 27:92–104. https://doi.org/10.1007/s11095-009-9980-5
8. Dubois A, Lavielle M, Gsteiger S, Pigeolet E, Mentré F (2011) Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. Stat Med 30:2582–2600. https://doi.org/10.1002/sim.4286
9. Möllenhoff K, Loingeville F, Bertrand J, Nguyen TT, Sharan S, Sun G, Grosser S, Zhao L, Fang L, Mentré F et al (2022) Efficient model-based bioequivalence testing. Biostatistics 23. https://doi.org/10.1093/biostatistics/kxaa026
10. Loingeville F, Bertrand J, Nguyen T, Sharan S, Feng K, Sun W, Han J, Grosser S, Zhao L, Fang L, Möllenhoff K, Dette H, Mentré F (2020) New model-based bioequivalence statistical approaches for pharmacokinetic studies with sparse sampling. AAPS J 22:141. https://doi.org/10.1208/s12248-020-00507-3
11. Hooker A (2022) Model averaging for model-based bioequivalence design and analysis. In: World Conference of Pharmacometrics WCOP 2022
12. Shen M, Russek-Cohen E, Slud E (2016) Distributional assumptions for pharmacokinetic summary statistics based on simulations with compartmental models. J Biopharm Stat. https://doi.org/10.1080/10543406.2016.1222535
13. Gallant AR (1975) Seemingly unrelated nonlinear regressions. J Econom 3:35–50. https://doi.org/10.1016/0304-4076(75)90064-0
14. Nguyen T, Mouksassi M, Holford N, Al-Huniti N, Freedman I, Hooker A, John J, Karlsson M, Mould D, Ruixo JP, Plan E, Savic R, van Hasselt J, Weber B, Zhou C, Comets E, Mentré F (2017) Model evaluation of continuous data pharmacometric models: metrics and graphics. CPT Pharmacomet Syst Pharmacol 6:87–109. https://doi.org/10.1002/psp4.12161
15. Dumont C, Lestini G, Le Nagard H, Mentré F, Comets E, Nguyen TT (2018) the PFIM group, PFIM 4.0, an R program for design evaluation and optimisation in nonlinear mixed effect models. Comput Methods Prog Biomed 156:217–229. https://doi.org/10.1016/j.cmpb.2018.01.008
16. Delyon B, Lavielle M, Moulines E (1999) Convergence of a stochastic approximation version of em algorithm. Ann Stat 27:94–128. https://doi.org/10.1214/aos/1018031103
17. Comets E, Lavenu A, Lavielle M (2017) Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. J Stat Softw 80:1–41. https://doi.org/10.18637/jss.v080.i03
18. Denney W, Duvvuri S, Buckeridge C (2015) Simple, automatic noncompartmental analysis: the PKNCA R package. J Pharmacokinet Pharmacodyn 42(11–107):S65. https://doi.org/10.1007/s10928-015-9432-2
19. Dubois A, Gsteiger S, Balser S, Pigeolet E, Steimer J, Pillai G, Mentré F (2012) Pharmacokinetic similarity of biologics: analysis using nonlinear mixed-effects modeling. Clin Pharmacol Ther 91:234–242. https://doi.org/10.1038/clpt.2011.216
20. Reijers J, van Donge T, Schepers F, Burggraaf J, Stevens J (2016) Use of population approach non-linear mixed effects models in the evaluation of biosimilarity of monoclonal antibodies. Eur J Clin Pharmacol 72:1343–1352. https://doi.org/10.1007/s00228-016-2101-6
21. U.S. Food and Drug Administration (2001) Statistical approaches to establishing bioequivalence. https://www.fda.gov/media/70958/download
22. Hu C, Moore KHP, Kim YH, Sale ME (2004) Statistical issues in a modeling approach to assessing bioequivalence or PK similarity with presence of sparsely sampled subjects. J Pharmacokinet Pharmacodyn 31:321–339. https://doi.org/10.1023/B:JOPA.0000042739.44458.e0
23. Bertrand J, Comets E, Chenel M, Mentré F (2012) Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. Biometrics 68:146–155. https://doi.org/10.1111/j.1541-0420.2011.01665.x
24. Buatois S, Ueckert S, Frey N, Retout S, Mentré F (2018) Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. AAPS J 20:56. https://doi.org/10.1208/s12248-018-0205-x
25. Gonçalves A, Mentré F, Lemenuel-Diot A, Guedj J (2020) Model averaging in viral dynamic models. AAPS J 22:48. https://doi.org/10.1208/s12248-020-0426-7
26. Yue C, Ozdin D, Selber-Hnatiw S, Ducharme M (2019) Opportunities and challenges related to the implementation of model-based bioequivalence criteria. Clin Pharm Ther 105:350–362. https://doi.org/10.1002/cpt.1270
27. Zhao L, Kim MJ, Zhang L, Lionberger R (2018) Generating model integrated evidence for generic drug development and assessment. Clin Pharmacol Ther 105:338–349. https://doi.org/10.1002/cpt.1282
28. Comets E, Brendel K, Mentré F (2008) Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the NPDE add-on package for R. Comput Methods Programs Biomed. https://doi.org/10.1016/j.cmpb.2007.12.002

Chapitre 3

SAEM_MH : méthode de calcul des erreurs standards semi-Bayésienne et application aux données Gante-nerumab

3.1 Résumé

Objectifs

Le premier travail ayant montré l'intérêt de la méthode semi-Bayésienne évaluée (Post), nous nous sommes dans la suite de la thèse attachés à intégrer une méthode similaire de calcul des SE dans le package `saemix`. L'algorithme SAEM intégrant un algorithme MH pour le tirage des paramètres individuels, nous avons cherché à tirer parti d'une approche similaire pour construire la distribution des paramètres de population lors de la phase de convergence de l'algorithme.

Ceci permettrait notamment d'éviter de devoir faire l'analyse en deux étapes et de changer de logiciel et de langage. Ce deuxième travail de thèse a donc eu pour objectif de développer cette méthode en l'appliquant aux données Gantenerumab présentées dans le chapitre précédent, puis de l'évaluer en mettant en œuvre une étude de simulation inspirée du premier travail.

Synthèse

Nous avons développé une méthode appelée SAEM_MH qui consiste à utiliser un algorithme MH dans l'algorithme SAEM, afin de tirer dans la distribution *a posteriori* de l'EMV à chaque itération de SAEM. On utilise l'algorithme MH qui est déjà intégré dans le package `saemix` pour échantillonner les paramètres individuels.

Cette méthode s'appuie sur un résultat théorique, le théorème de Bernstein-von Mises, selon lequel la distribution de l'estimateur Bayésien tend asymptotiquement vers une loi normale dont la variance est l'inverse de la FIM. Cela veut donc dire que les distributions asymptotiques de l'EMV et du MAP sont identiques. La question que l'on se pose ici est de savoir si ce résultat tient toujours à distance finie, et dans quelle mesure.

Nous avons évalué cette méthode en simulant des données pharmacocinétiques selon un modèle à un compartiment (donc relativement simple), avec un design riche (150 sujets et 10 points par sujet) et épars (12 sujets et 3 points par sujet). Nous l'avons comparé à deux méthodes fréquentistes (Asympt et SIR) et à la méthode semi-Bayésienne Post déjà présentée dans le premier projet. Les méthodes Gallant, qui est uniquement une méthode de correction pour les tests, et Bootstrap, qui est très lourde en caluls, n'ont pas été évaluées dans ce travail.

Nous avons comparé les résultats en termes de taux de couverture des différents paramètres du modèle (effets fixes, effets des covariables, variabilités inter-individuelles, paramètres du modèle d'erreur). Nous avons également étudié les zip plots, des graphiques permettant de représenter la distribution des intervalles de confiance calculés sur un ensemble de jeux de données.

Les méthodes fréquentistes et semi-Bayésiennes ont donné des résultats similaires et satisfaisants sur les données riches, ce qui confirme le théorème de Bernstein-von Mises. Sur les données éparses, la méthode Asympt a sous estimé les SE comme attendu. L'autre méthode fréquentiste, SIR, présente les mêmes limites que la méthode Asympt,

alors que les méthodes semi-Bayésiennes ont donné des résultats plus robustes aux données éparses.

Suite à cette étude de simulation, nous avons appliqué SAEM_MH aux données de Gantenerumab riches et éparses. Comme sur les simulations, les résultats étaient similaires sur les données riches entre les différentes méthodes, mais sur les données éparses, il y avait des discordances entre les méthodes fréquentistes et semi-Bayésiennes, confirmant une déviation de l'asymptotique dans ce cas. Notamment, la méthode SAEM_MH a donné des résultats insatisfaisants, avec des taux d'acceptation trop bas pour être fiable.

Pour comprendre ce résultat, un second jeu de simulations a été réalisé avec des caractéristiques plus proches des données de Gantenerumab, notamment une matrice de variabilités inter-individuelles plus complexe, avec des hautes variabilités et des corrélations entre les différents paramètres. Sur ce jeu de données, SAEM_MH a également obtenu des taux d'acceptation proches de zéro, ce qui a mis la méthode en défaut. Nous soupçonnons donc que c'est la complexité de la matrice de variabilité et/ou le grand nombre de paramètres à estimer qui met la méthode en difficulté.

Apports du travail

Ces résultats nous ont permis d'avoir un aperçu des avantages de la méthode développée : SAEM_MH donne de meilleurs résultats que les méthodes fréquentistes sur les données éparses simulées, notamment en termes de taux de couverture des paramètres de population, et permet même d'obtenir de meilleures performances que la méthode basée sur l'algorithme HMC. SAEM_MH présente néanmoins des limites : l'application à des données réelles a montré sa sensibilité à la dimension du vecteur de paramètres de population, notamment en présence d'une structure de variabilité inter-individuelle complexe, ce qui fait drastiquement chuter les taux d'acceptation des chaînes échantillonnées. Des pistes sont à explorer pour dépasser ces difficultés : les

perspectives d'amélioration de la méthode SAEM_MH impliquent de tester des variations de la méthode permettant de diminuer la dimension du vecteur de paramètres à échantillonner et ainsi d'augmenter les taux d'acceptation des chaînes obtenues.

3.2 Article 2 (publié)

The AAPS Journal (2024) 26:53
https://doi.org/10.1208/s12248-024-00905-x

RESEARCH ARTICLE

Uncertainty Computation at Finite Distance in Nonlinear Mixed Effects Models—a New Method Based on Metropolis-Hastings Algorithm

Mélanie Guhl[1] · Julie Bertrand[1] · Lucie Fayette[1] · François Mercier[2] · Emmanuelle Comets[1,3]

Received: 6 December 2023 / Accepted: 29 February 2024

Abstract
The standard errors (SE) of the maximum likelihood estimates (MLE) of the population parameter vector in nonlinear mixed effect models (NLMEM) are usually estimated using the inverse of the Fisher information matrix (FIM). However, at a finite distance, i.e. far from the asymptotic, the FIM can underestimate the SE of NLMEM parameters. Alternatively, the standard deviation of the posterior distribution, obtained in Stan via the Hamiltonian Monte Carlo algorithm, has been shown to be a proxy for the SE, since, under some regularity conditions on the prior, the limiting distributions of the MLE and of the maximum a posterior estimator in a Bayesian framework are equivalent. In this work, we develop a similar method using the Metropolis-Hastings (MH) algorithm in parallel to the stochastic approximation expectation maximisation (SAEM) algorithm, implemented in the `saemix` R package. We assess this method on different simulation scenarios and data from a real case study, comparing it to other SE computation methods. The simulation study shows that our method improves the results obtained with frequentist methods at finite distance. However, it performed poorly in a scenario with the high variability and correlations observed in the real case study, stressing the need for calibration.

Keywords finite distance · Metropolis-Hastings · nonlinear mixed effect models · standard errors · uncertainty

Introduction

Nonlinear mixed effect models (NLMEM) are powerful tools to model longitudinal data with nonlinear trajectories, and these are now routinely collected in clinical trials, investigating pharmacokinetics (PK) or pharmacodynamics (PD), viral dynamics, evolution of clinical scores, etc. Assessing the magnitude and statistical significance of a treatment effect is often the primary outcome of clinical trial analyses, and the uncertainty of the model parameters, i.e. the standard errors (SE) of the estimators, is needed to compute the test statistic determining the conclusion of the clinical trial. Therefore, when the treatment effect is a parameter of a NLMEM, it is important to have a reliable computation method of the parameter uncertainty.

Several algorithms using gradient-based approaches, such as the first-order conditional estimation (FOCE), or expectation maximisation (EM) approaches, such as the stochastic approximation EM (SAEM), and software (NONMEM (1, 2), Monolix (3), `saemix` (4)) have been developed in the past 30 years to handle NLMEM. The methods used in these algorithms are typically frequentist, where the population parameter vector θ is considered a point estimate, and the SE of the maximum likelihood estimator (MLE) can be computed asymptotically based on the Fisher information matrix (FIM). However, at a finite distance, that is, when the number of subjects N and/or the number of sampling points per subject n decrease, the FIM can underestimate the SE of NLMEM, and this underestimation can notably result in an inflation of type I error when performing tests. This has been shown for example in bioequivalence studies, using a one-compartment PK model and a smaller number of subjects (down to N=12) and/or number of samples per subject (down to n=2) (5, 6). This likely contributes to limiting the use of NLMEM when making clinical decisions.

Mélanie Guhl
melanie.guhl@inserm.fr

1 Université Paris Cité, Inserm, IAME, F-75018 Paris, France

2 Department of Biostatistics, Roche Innovation Center Basel, Basel, Switzerland

3 Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085, F-35000 Rennes, France

Approaches to address this issue include methods to correct the asymptotic SE (7) and resampling methods like bootstrap (8) or sampling importance resampling (SIR) (9). Alternatively, Ueckert *etal*. (10) proposed to borrow from the Bayesian framework to compute the SE of NLMEM parameters. In a Bayesian framework, the population parameter vector θ is considered a random variable and assigned a prior, and this method, called hereafter Post, aims to characterise the posterior distribution from which we compute the maximum a posteriori (MAP). Under some regularity conditions on the prior, the limiting distributions of the MLE and the MAP estimator are equivalent (Bernstein-von Mises theorem (11)). This leads to the idea of using the standard deviation of the posterior distribution of θ as a proxy for the SE of the MLE, making the assumption that the equivalence of distributions still holds at finite distance. The Post method has been successfully implemented in Stan (12) via the Hamiltonian Monte Carlo (HMC) algorithm (6, 13).

Our objective in this work is to develop a similar method, SAEM_MH, based on a Metropolis-Hastings algorithm, within the SAEM algorithm and implement it in the `saemix` package. This would avoid implementing the model again in a different software, as the Bayesian algorithm is included directly in the frequentist estimation with SAEM_MH: we run SAEM in parallel to a Bayesian algorithm initialised after K1 iterations to obtain on one side the frequentist estimates from SAEM as usual, and on the other side a Bayesian estimate of the whole posterior distribution.

We implement SAEM_MH and evaluate using simulations inspired from a PK study setting and compare to using the expected asymptotic FIM, the SIR and Post approaches mentioned above. Finally, we apply the four methods in a real case study analysis of two formulations of Gantenerumab, a monoclonal antibody developed for the treatment of Alzheimer's disease for which the data and model used have been described previously (13).

Methods

NLMEM

A typical formulation of a nonlinear mixed effect model (NLMEM) (3) for a continuous response y can be written as follows:

$$y_{ij} = f(t_{ij}, \psi_i) + g(t_{ij}, \psi_i, \xi)\,\epsilon_{ij} \quad (1)$$

$$\phi_i = h(\psi_i) = h(\bar{\psi}_i) + \eta_i \quad (2)$$

with y_{ij} the outcome for individual i at time t_{ij} (i=1...N, j=1,...n_i), $f()$ the structural model, ψ_i the vector of parameters for individual i, $g()$ the error model, ξ the error parameter vector and $\epsilon_{ij} \sim \mathcal{N}(0,1)$ the residual error. $h()$ is a transformation function for individual parameters, $\bar{\psi}_i$ the typical value of ψ_i, η_i the between subject random effect - $\eta \sim \mathcal{N}(0, \Omega)$ and Ω the variance-covariance matrix of the inter-individual random effects.

The typical value $\bar{\psi}_i$ for subject i is a combination of the population value μ and optionally covariate effects β. We add the random effect η_i to the transform of $\bar{\psi}_i$ to obtain the transform of the individual parameter ϕ_i.

To describe the variance of the measurement error, three error models are classically used: additive error $g(t_{ij}, \psi_i) = a$, multiplicative error $g(t_{ij}, \psi_i) = b\ f(t_{ij}, \psi_i)$ and combined error $g(t_{ij}, \psi_i) = a + b\ f(t_{ij}, \psi_i)$. So $\xi \in \{a, b, (a, b)\}$.

SAEM algorithm

The EM algorithm is a powerful solution to estimate the population parameters $\theta = (\mu, \beta, \Omega, \xi)$. It iterates between an expectation step computing the conditional expectation of the complete likelihood of the data and the (unknown) individual parameters, and a maximisation step updating θ, until convergence. In this work, we use the SAEM algorithm, which replaces the expectation step by a simulation step followed by a stochastic approximation of the likelihood. The algorithm is divided into two phases: an exploratory phase (K_1 iterations) and a smoothing phase (K_2 iterations) where population parameters are forced to converge by using decreasing increments in the stochastic approximation. This algorithm has been proven to be very efficient for NLMEM and has been implemented in several software. A more detailed description is available in Appendix 1 (14).

Bayesian approach

In a Bayesian paradigm, θ is considered a random vector and not a fixed parameter. From a prior distribution $p(\theta)$ and the information from the data y, we want to compute the posterior distribution $p(\theta|y)$. Bayes theorem gives the following:

$$p(\theta|y) \propto L(y|\theta)p(\theta) \quad (3)$$

where $L()$ denotes the likelihood function. In the following, we will use $l()$ to denote the loglikelihood.

We can define estimators of θ from the posterior distribution: an example is the mode of the distribution, called the maximum a posteriori (MAP). Different Bayesian algorithms such as the HMC algorithm (implemented in the Stan software for example) or the Metropolis-Hastings (MH) algorithm can be used to draw in the posterior distribution of θ.

SE Computation in NLMEM by Frequentist Approach

FIM

To compute the SE of the population parameter vector $\theta = (\mu, \beta, \Omega, \xi)$, the classical approach in NLMEM is to use the expected Fisher information matrix (FIM) of the model.

This matrix cannot be computed in a closed form in NLMEM and is usually approximated by the FIM of a Gaussian model derived from the NLMEM after linearisation of the regression function f around the conditional expectation of the individual Gaussian parameters (method hereafter called Asympt):

$$l(y|\theta,\psi) \simeq -N/2\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log(det\ \Gamma_i) - \frac{1}{2}\sum_{i=1}^{N}(y_i-\bar{y}_i)^t\ \Gamma_i^{-1}\ (y_i-\bar{y}_i) \quad (4)$$

where $\bar{y}_i = f(t_i,\psi_i) + \partial_\psi f(t_i,\psi_i)(\theta-\bar{\psi}_i)$

$$\Gamma_i = \partial_\psi f(t_i,\psi_i)\ \Omega\ \partial_\psi f(t_i,\psi_i)^t + g(t_i,\psi_i)\ I_{n_i}\ g(t_i,\psi_i))^t$$

with $y = (y_i)_{i=1,...,N}$, where y_i is the vector of n_i observations for individual i, I_{n_i} is the identity matrix of size n_i, $det\ \Gamma_i$ is the determinant of the matrix Γ_i and $\partial_\psi f(t_i,\psi_i)$ is the gradient of f with respect to ψ calculated in (t_i,ψ_i).

SIR

In this work, we also explored the approach called sampling importance resampling (SIR) as a frequentist alternative to the FIM. In this method, a large number M_S of parameter vectors $\hat{\theta}_{SIR,s}$, $s = 1,...,M_S$ are simulated from a proposal distribution, generally Gaussian. For each vector, the 'importance' is computed as the ratio between the objective function, which is proportional to the likelihood of the data given the simulated vector $L(y|\hat{\theta}_{SIR,s})$, and the density of the simulated vector given the proposal distribution $p(\hat{\theta}_{SIR,s})$. Then, an importance ratio is computed as the ratio of this importance over the importance of the MLE estimated by `saemix`:

$$\mathrm{IR}_s = \frac{L(y|\hat{\theta}_{\mathrm{SIR,s}})/p(\hat{\theta}_{\mathrm{SIR,s}})}{L(y|\theta_{\mathrm{MLE}})/p(\theta_{\mathrm{MLE}})} \quad (5)$$

In a second step, among the M_S samples, a subset of m_S parameter vectors are resampled (without replacement) according to probabilities proportional to their IR. The SE are computed from the variance-covariance matrix of these m_S samples. This method is hereafter called SIR.

SE Computation in NLMEM by Bayesian Approach

In a full Bayesian framework, the SE of the MAP estimator would be computed using the standard deviation (SD) of posterior distribution samples. Here, we propose a hybrid approach where we use the MLE as the point estimate and draw in the posterior distribution $p(\theta|y)$ at finite distance to evaluate its standard deviation as a proxy for the SE of the MLE.

Post

As a benchmark, we used the HMC algorithm, implemented in Stan, to obtain the posterior distribution using the MLE obtained with SAEM as initial values and semi-informative prior distributions. This method is hereafter called Post.

SAEM_MH

Our proposal, which we hereafter call SAEM_MH, is to apply a similar method within SAEM during the smoothing phase, modifying the embedded Metropolis-Hastings algorithm as follows:

At iteration K_1+1, set a prior $p(.)$ on θ.
At every iteration $k = K_1+1, ..., K_1+K_2$ of the SAEM algorithm:

- Compute FIM_k, by drawing z samples of ψ conditionally to θ_k with the SAEM machinery and compute the linearised FIM around their mean.
- Draw $\theta_k^{(MH)}$ with M_k iterations of the Metropolis-Hastings algorithm.

For each MH chain, the initial value is $\theta^{(0)} = \theta_k$. At each m_k iteration:

- Draw $\theta^{(m_k)}$ from a kernel $q(.)$.
- Draw z samples of $\psi^{(m_k)}$ from the conditional distribution $p(\cdot|y;\theta^{(m_k)})$ and obtain their mean $\bar{\psi}^{(m_k)}$.
- Compute the likelihood of $\theta^{(m_k)}$ by linearisation around $\bar{\psi}^{(m_k)}$.
- Compute the acceptance ratio α as

$$\alpha(\theta_k^{(m_k-1)},\theta^{(m_k)}) = \min\left(1, \frac{p(\theta^{(m_k)}|y,\bar{\psi}^{(m_k)})q(\theta_k^{(m-1)})}{p(\theta_k^{(m_k-1)}|y,\bar{\psi}_k^{(m_k-1)})q(\theta^{(m_k)})}\right) \quad (6)$$

$$= \min\left(1, \frac{L(y|\theta^{(m_k)},\bar{\psi}^{(m_k)})p(\theta^{(m_k)})q(\theta_k^{(m_k-1)})}{L(y|\theta_k^{(m_k-1)},\bar{\psi}_k^{(m_k-1)})p(\theta_k^{(m_k-1)})q(\theta^{(m_k)})}\right) \quad (7)$$

- Accept $\theta_k^{(m_k)} = \theta^{(m_k)}$ and $\psi_k^{(m_k)} = \psi^{(m_k)}$ with probability $\alpha(\theta_k^{(m_k-1)},\theta^{(m_k)})$, otherwise, $\theta_k^{(m_k)} = \theta_k^{(m_k-1)}$ and $\psi_k^{(m_k)} = \psi_k^{(m_k-1)}$with probability $1-\alpha(\theta_k^{(m_k-1)},\theta^{(m_k)})$
- At the end of the M_k iterations, $\theta_k^{(MH)} = \theta_k^{(M_k)}$, $\psi_k^{(MH)} = \psi_k^{(M_k)}$.

At the end of the SAEM algorithm, use the SD of the samples $(\theta_k^{(MH)})_{k=K_1+1,...,K_1+K_2}$ to compute the SE of $\hat{\theta}$.

Simulation Study

Settings

The settings for our first simulation scenario are derived from the theophylline dataset presented in the `saemix` package, and have already been used in a previously published simulation study focusing on alternative SE computations for model-based bioequivalence tests (6). The pharmacokinetics (PK) are described through a one-compartment model with linear absorption and elimination:

$$C(t) = \frac{D}{V}\frac{ka}{\frac{CL}{V} - ka}\left(\exp(-ka\ t) - \exp(-\frac{CL}{V}t)\right) \quad (8)$$

This model represents a plasmatic concentration curve with first an exponential absorption phase and then an exponential elimination phase. $C(t)$ is the plasmatic concentration of the drug measured at time t, D represents the dose administered, ka is the absorption rate, V is the volume of the plasma compartment and CL is the clearance of the drug.

We used this model to simulate clinical trials with two treatment arms (T=0, placebo; T=1, treatment arm). We simulated a dose of $D = 4mg$ for all subjects. The mean population values for each parameter were $\mu_{ka} = 1.5h^{-1}$, $\mu_{CL} = 0.04mg.L^{-1}$ and $\mu_V = 0.5L$, which were obtained for the theophylline real dataset (15). We simulated treatment effects on CL and V: β^T_{ka}=0, β^T_{CL}=log(1.25), β^T_V=log(1.25). The variance-covariance matrix of the inter-individual random effects was diagonal with ω_{ka}=0.22, ω_{CL}=0.11, and ω_V=0.22. We used a proportional error model with b=0.1.

Here, all the parameters are log-normal, which means that, using the notations from Eq. 1, $h() = \log()$. We now have the following:

$$\begin{aligned} \log(ka_i) &= \log(\mu_{ka}) + \beta^T_{ka}\ \mathbb{1}_{T_i=1} + \eta_{ka,i} \\ \log(CL_i) &= \log(\mu_{CL}) + \beta^T_{CL}\ \mathbb{1}_{T_i=1} + \eta_{CL,i} \\ \log(V_i) &= \log(\mu_V) + \beta^T_V\ \mathbb{1}_{T_i=1} + \eta_{V,i} \end{aligned} \iff \begin{aligned} ka_i &= \mu_{ka}\exp(\beta^T_{ka}\ \mathbb{1}_{T_i})\exp(\eta_{ka,i}) \\ CL_i &= \mu_{CL}\exp(\beta^T_{CL}\ \mathbb{1}_{T_i})\exp(\eta_{CL,i}) \\ V_i &= \mu_V\exp(\beta^T_V\ \mathbb{1}_{T_i})\exp(\eta_{V,i}) \end{aligned}$$

with $\eta_i = (\eta_{ka,i}, \eta_{CL,i}, \eta_{V,i}) \sim \mathcal{N}(0, \Omega)$, $\Omega = diag(\omega_{ka}, \omega_{CL}, \omega_V)$ and $\mathbb{1}^{T_i} = 1$ if subject i received the treatment, 0 otherwise.

To evaluate our method in an asymptotic setting, we used a rich design with N = 150 patients (75 in each treatment arm) and n = 10 sampling times ($t \in \{0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12, 24\}$ hours). Finite conditions were simulated using a sparse design with N = 12 patients (6 in each treatment arm) and n = 3 sampling times ($t \in \{0.25, 3.5, 24\}$ hours). These two designs are hereafter called rich and sparse.

Evaluation

For each design, 1000 datasets were simulated. We fitted the true structural and statistical model in `saemix`, but estimated an additional effect on ka to mimic what would be done on a real dataset when testing for treatment effects.

We evaluated the capacity of our method to estimate uncertainty and confidence intervals in these simulation scenarios, comparing the SE obtained with SAEM_MH to those obtained with Asympt, SIR and Post. Our main criterion was the 95% coverage rate (CR) which we defined as the percentage of datasets in which the simulated value of the parameter was within the parameter estimate 95% confidence interval. For Asympt, the confidence intervals were computed symmetrically as follows: mean±qnorm(0.975)×SE. For other methods (SIR, Post and SAEM_MH), we used the 2.5th and 97.5th quantiles of the sampled distributions. Our target was the 95% prediction interval around 0.95 which we computed using a binomial distribution.

We also checked acceptation rates (AR) which we defined as the percentage of accepted new values in the MH chains. An AR of 0.234 has been suggested to be a reasonable value in multiple dimensions (16), avoiding both low AR indicating insufficient sampling of the posterior distribution and high AR showing a lack of discrimination.

We also used zip plots (see Appendix 2) (17) which help to interpret the coverage rates obtained over all 1000 confidence intervals.

Implementation

This work was performed using R and Stan. The `saemix` package was used to estimate the parameters in the NLMEM, and *ad hoc* code was written to implement the MH algorithm and its variants. The initial values for parameter estimation were the true values for fixed effects, 0 for treatment effects and the diagonal identity matrix for the inter-individual variance matrix (100% variability, the default value). We ran three chains with $K_1 = 300$ and $K_2 = 100$. For SIR, we used the SIR code available in the development version of `saemix` on its GitHub (https://github.com/saemixdevelopment/saemixextension). The proposal distribution was a multivariate normal distribution centred on the MLE, and the variance was the inverse of the FIM multiplied

by an inflation factor (we kept the default of 1). For each simulated dataset, we ran M_S=1000 samples and m_S=500 resamples.

For Post, we ran three chains of 1500 iterations (including 500 warm-up iterations). The initial values were the estimates returned by `saemix`, and the prior distributions were Gaussian, centred on the true values of θ, with 30% coefficient of variation (CV) for fixed effects and 50% for treatment effects, variance matrix and residual error parameter. We kept the results only if the highest $\hat{R}$ was lower than 1.2 to ensure the chains had mixed enough.

For SAEM_MH, we used the same Gaussian prior. The kernel was also Gaussian: $q(.) = \mathcal{N}(\hat{\theta}_k, \text{inf} * \widehat{FIM}_k^{-1})$ with three different inflation factors inf=1, 1.5, 2. As $K_2 = 100$, we ran 100 MH chains of length M_k=100. The number of ψ drawn to compute the different likelihoods and FIM matrices was set to $z = 50$.

A script detailing the analysis of one simulated dataset with `saemix` and Stan is available on a Zenodo depot (https://doi.org/10.5281/zenodo.8190068).

Results

Figure 1 shows the coverage and its 95% confidence interval obtained for the different model parameters and scenarios in the simulation study. Each panel represents one parameter, with the different methods given on the X-axis. The target CR of 0.95 is shown as a horizontal line. The crosses represent the results for the rich design, on which all methods converged and could be evaluated on all datasets, whereas on the sparse design, represented on Fig. 1 by bullets, with Post, 84 datasets out of the 1000 were removed due to $\hat{R}$ being $\geq$ 1.2 for at least one model parameter. Although Fig. 1 represents all the model parameters, in the following, we focus on the results for the treatment effect coefficients.

On the rich design, all methods gave satisfactory CR. The similar results between Asympt and Post validate the Bernstein-von Mises theorem and support our approach. On the sparse design, CR estimates on treatment effect coefficients were below the target with Asympt and SIR and above the target with Post. CR estimates with SAEM_MH were below the target for the treatment effect coefficients but better than with Asympt. Inflating the variance kernel by a factor of 2 restored appropriate CR estimates. Figure 1 shows that the AR with SAEM_MH decreased when inflating the kernel variance. If we refer to the optimal AR given in the literature of 0.234 (16), we would choose an inflation factor of 2 on the rich design and 1 on the sparse design; however, the best coverage rates were obtained with inflation rates of 1 on the rich design and 2 on the sparse design.

Zip plots summarising the distributions of confidence intervals over the different datasets are shown in Appendix 2. On the sparse design, they indicated that with Asympt and SIR the undercoverage was due to an underestimation of the SE. With Post, the parameters were slightly biased but the 95% CI were large enough to capture the simulated value. With SAEM_MH, treatment effect coefficient estimates were less biased, but the 95% CI were smaller. SAEM_MH performed better when inflating the variance kernel by a factor of 2.

In terms of computation times, Asympt computes within 1 s for both designs. Post is also relatively fast, with a mean computation time of 90 s on the rich design and 8 s on the sparse design, which is not surprising as the stan software is well optimised and efficient. SAEM_MH is much slower but is currently an *ad hoc* code and could be optimised: on the rich design, it takes 90 min on average and 32 min on the sparse design. The SIR method was the slowest, taking on average 2.5 h on the rich design and 15 min on the sparse design.

Application to Gantenerumab Data

Data

The data used comes from a phase I clinical trial (NCT01636531) investigating the relative bioavailability, tolerability, and dose-exposure relationship of a high-concentration liquid formulation (HCLF G3) *versus* a lyophilised formulation (LyoF G2) of Gantenerumab, a monoclonal antibody developed for the treatment of Alzheimer's disease with a very long half-life. We used the data from the reference G3 and test G2 treatment arms with a subcutaneous dose of D=225 mg. There were 24 subjects in each arm, with 11 samples per subjects, taken after 6 h then on days 1, 2, 3, 4, 7, 13, 20, 42, 63 and 84. These data have already been presented elsewhere (13).

Methods

A model selection procedure has been performed previously on these data (13), comparing different structural models (one or two compartments, delay or not, zero-order or first-order absorption), error models (additive, multiplicative or combined) and selecting the covariance structure. We kept the model selected in the previous analysis, a two-compartment model with delayed first-order absorption and linear elimination:

$$C(t) = D\big(A\exp(-\alpha\ (t - T_{lag})) + B\exp(-\beta\ (t - T_{lag})) - (A + B)\exp(-ka\ (t - T_{lag}))\big) \quad (9)$$

with:

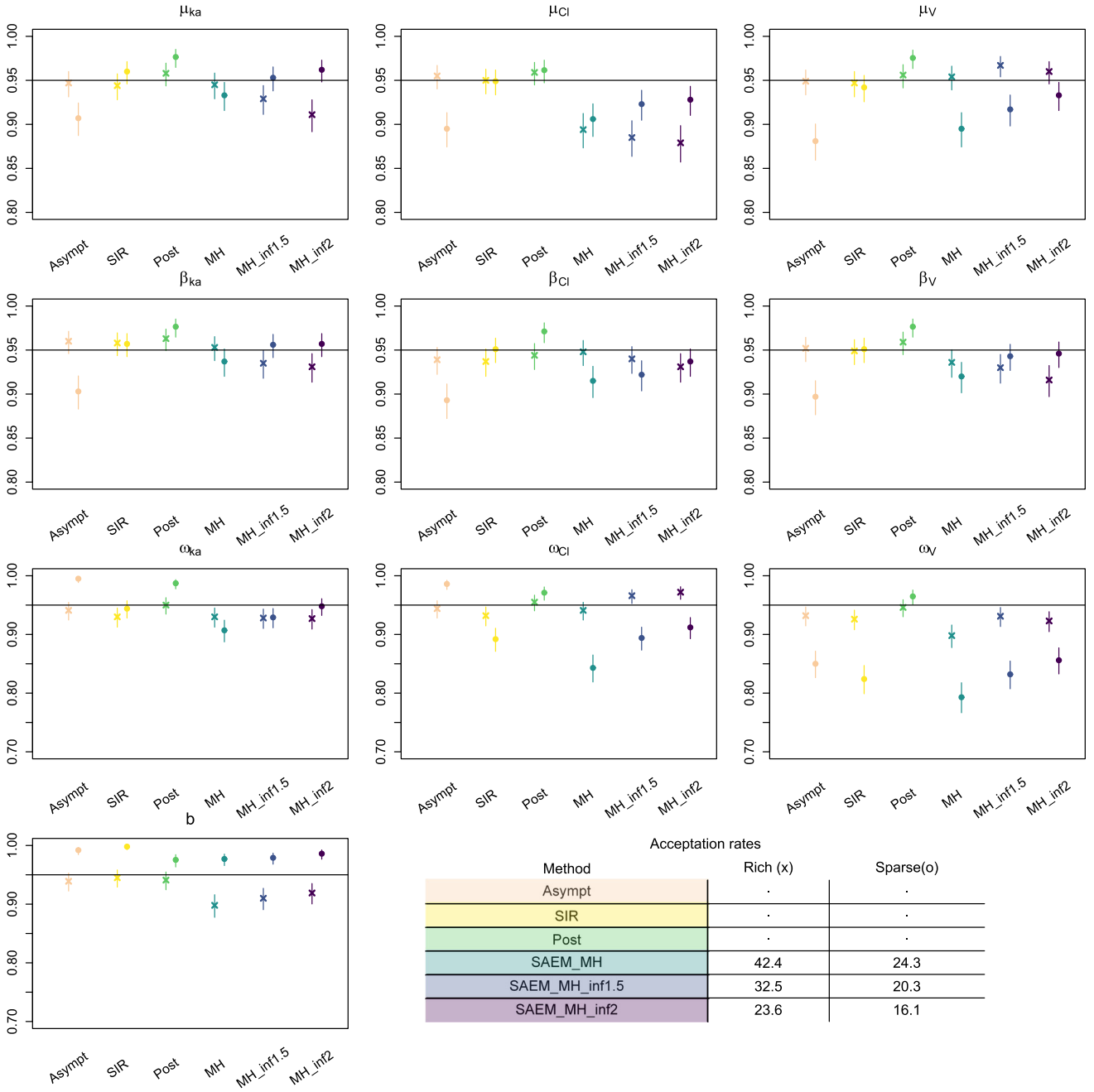


Fig. 1 95% coverage rates obtained with Asympt ( ) , SIR ( ), Post ( ), SAEM_MH with an inflation factor for the variance kernel equal to 1 ( ), 1.5 ( ) and 2 ( ) on the rich (x) and sparse (o) scenarios. Each panel represents a parameter of the model. The different methods are given on the *X*-axis. 95% coverage rates were defined as the proportion of datasets in which the 95% confidence interval of a parameter recovers its true value. The target value for 95% coverage rates was 0.95. Here, the coverage rates are represented with their 95% confidence intervals. The number of datasets used was 1000 for FIM, SIR and SAEM_MH for both scenarios. For Post, the number of datasets included was 1000 in the rich scenario and 916 in the sparse scenario

- T_{lag} the period of latency before the concentration starts rising
- ka the absorption constant rate
- $A = \frac{ka}{V_1}\frac{k_{21}-\alpha}{(ka-\alpha)(\beta-\alpha)}$
- $B = \frac{ka}{V_1}\frac{k_{21}-\beta}{(ka-\beta)(\alpha-\beta)}$
- $\alpha = \frac{k_{21}k}{\beta}$ the first rate constant
- $\beta = \frac{1}{2}\left(k_{12}+k_{21}+k-\sqrt{(k_{12}+k_{21}+k)^2-4k_{21}k}\right)$ the second rate constant
- $k_{12} = \frac{Q}{V_1}$ the distribution rate constant between the principal and the peripheral compartment
- $k_{21} = \frac{Q}{V_2}$ the distribution rate constant between the peripheral and principal compartment
- $k = \frac{CL}{V_1}$ the elimination rate constant
- Q the inter-compartmental clearance
- V_1 the volume of distribution of the principal compartment
- V_2 the volume of distribution of peripheral compartment

The six parameters of interest were ka, V_1, Q, V_2, CL and T_{lag}. All parameters were assumed to have a log-normal distribution. We estimated a treatment effect on the fixed effect of each parameter. Random effects were kept on ka, CL, V_1 and T_{lag}, with one correlation between the random effects of CL and V_1. The error model was proportional.

We compared the RSE obtained on these six parameters with the asymptotic FIM method (Asympt), SIR, Post and SAEM_MH. We also compared the RSE obtained with these different methods on a subset of the data where we only kept six random subjects in each arm.

Implementation

Asympt, SIR and SAEM_MH were run using the `saemix` package in R. For the SAEM estimation, we ran six chains with $K_1 = 500$ and $K_2 = 300$. The number of iterations and chains was higher than in the simulation study to ensure convergence of the algorithm, which was checked through convergence plots, as this model was more complex. The initial values used were 0.5 for μ_{ka}, 0.6 for μ_{CL}, 15 for μ_{V_1}, 0.5 for μ_Q, 5 for μ_{V_2} and 0.05 for μ_{Tlag}, 0 for all β, 1 for all ω, 0 for ρ_{CL,V_1} and 1 for b. For the SIR algorithm, we ran $M_S = 1000$ samples and $m_S = 500$ resamples, modifying the code available on the GitHub to account for correlations in the inter-individual variance-covariance matrix.

Post was run using Stan (12). In Post, we ran three chains of 1500 iterations (including 500 warm-up iterations). The initial values were the estimations of `saemix` and the prior distribution was Gaussian, centred on 0.46 for μ_{ka}, 0.63 for μ_{CL}, 15 for μ_{V_1}, 0.42 for μ_Q, 5 for μ_{V_2} and 0.04 for μ_{Tlag}, 0 for all β, 1 for all ω, 0 for ρ_{CL,V_1}, 1 for b, with 30% CV for fixed effects and 50% for treatment effects, variance matrix and error parameter.

In SAEM_MH, we used a Gaussian prior centred on the MLE with 30% CV for fixed effects and 50% for treatment effects, variance matrix and residual error parameter. The kernel was also Gaussian: $q(.) = \mathcal{N}(\hat{\theta}_k, \text{inf} \times \widehat{FIM}_k^{-1})$. As $K_2 = 300$, we ran 300 MH chains of length M_k=100. The number of ψ drawn to compute the different likelihoods and FIM matrices was set to $z = 50$.

On the sparse subset, we ran ten chains in the SAEM algorithm to account for the reduction of N. The other settings were the same as on the complete dataset.

Results

The parameters estimated on the full dataset and the sparse subset of the data are shown in Table I. Taking only a subset of the data does not noticeably change the point estimates, and at an individual level, the predictions computed for the 12 subjects present in both datasets are very similar (Fig. 2). However, the RSE computed with the FIM increase, as expected. Aside from nonsignificant treatment effects, the

Table I Parameters Estimates and their Relative Standard Errors (RSE %) Given by `saemix` for the Modelling of the Two Formulations of Gantenerumab, in the Full Dataset of N=48 Subjects (Left) and a Subset of N=12 Subjects (Right)

Parameter	Estimation (RSE,%) Full dataset	Estimation (RSE, %) N=12
$\mu_{ka}(d)$	0.46 (16)	0.57 (30)
β^T_{ka}	0.29 (78)	0.26 (162)
$\mu_{CL/F}$ (L.d^{-1})	0.63 (5)	0.63 (8)
$\beta^T_{CL/F}$	−0.07 (108)	0.003 (4,017)
$\mu_{V_1/F}$ (L)	14.63 (9)	13.36 (17)
$\beta^T_{V_1/F}$	0.20 (61)	0.28 (77)
$\mu_{Q/F}$ (L.d^{-1})	0.42 (31)	0.54 (67)
$\beta^T_{Q/F}$	−0.51 (129)	−0.85 (151)
$\mu_{V_2/F}$ (L)	4.88 (15)	4.78 (31)
$\beta^T_{V_2/F}$	−0.48 (63)	−0.43 (136)
$\mu_{T_{lag}}$ (d)	0.04 (33)	0.04 (80)
$\beta^T_{T_{lag}}$	−0.03 (1430)	−3.05 (674)
ω_{ka}	0.65 (11)	0.59 (22)
$\omega_{CL/F}$	0.26 (11)	0.20 (22)
$\rho_{CL/F,V_1/F}$	0.76 (9)	0.96 (5)
$\omega_{V_1/F}$	0.33 (11)	0.25 (24)
$\omega_{T_{lag}}$	0.90 (24)	1.18 (55)
b	0.15 (4)	0.16 (7)

Note 1 RSE were computed with the asymptotic FIM method

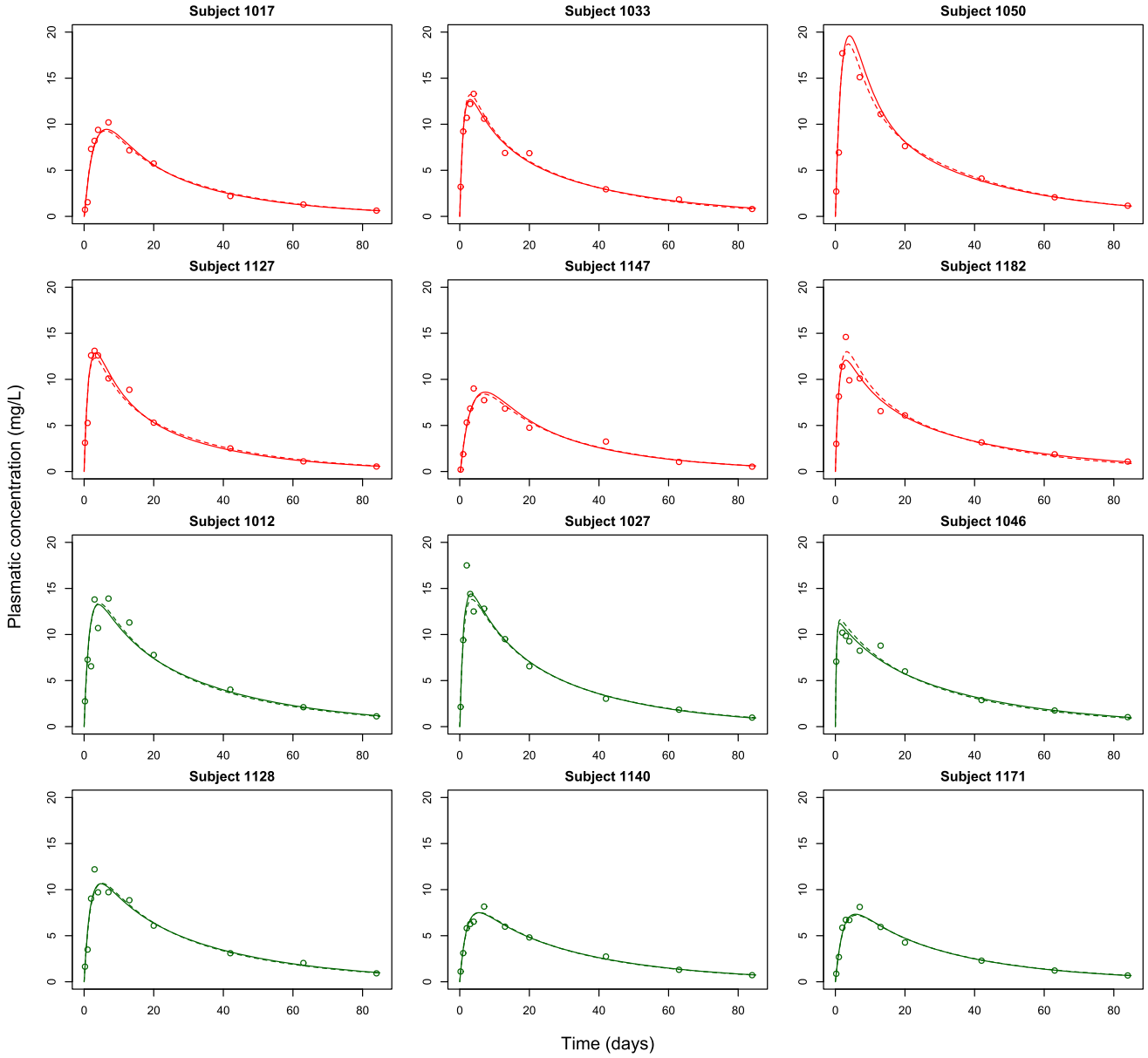


Fig. 2 Individual fits with the model fitted on the full dataset (full line) and the model fitted on the sparse subset of the data (dotted line) for the 12 subjects selected in the sparse dataset. Points represent the concentrations measured for each subject, and lines represent the individual predicted fit of the PK model. Subjects from the reference arm are shown in green and subjects from the test arm in red

increase in RSE was roughly of a factor 2 which is concordant with the theoretical computation of SE being proportional to $\sqrt{(N)}$, as we divided the number of subjects by 4.

On the full dataset, the RSE computed with the different methods are generally concordant (see Fig. 3). This confirms that we are in an asymptotic situation. The acceptation rate for SAEM_MH was around 37%, and we also ran the method with an inflation of 1.5 and 2 on the kernel variance, which lowered the acceptation rates to 14% and 3%, respectively, but did not have an impact on the RSE computed. The numerical results used to produce the plot are given in Table II in Appendix 4.

On the sparse subset of the data, we see more discrepancies between the asymptotic methods (FIM/SIR) and the Bayesian method (Post), which seems to indicate that we are no longer in an asymptotic regime. The RSE for the SAEM_MH method were lower than all other methods and likely unreliable because the acceptation rates were close to

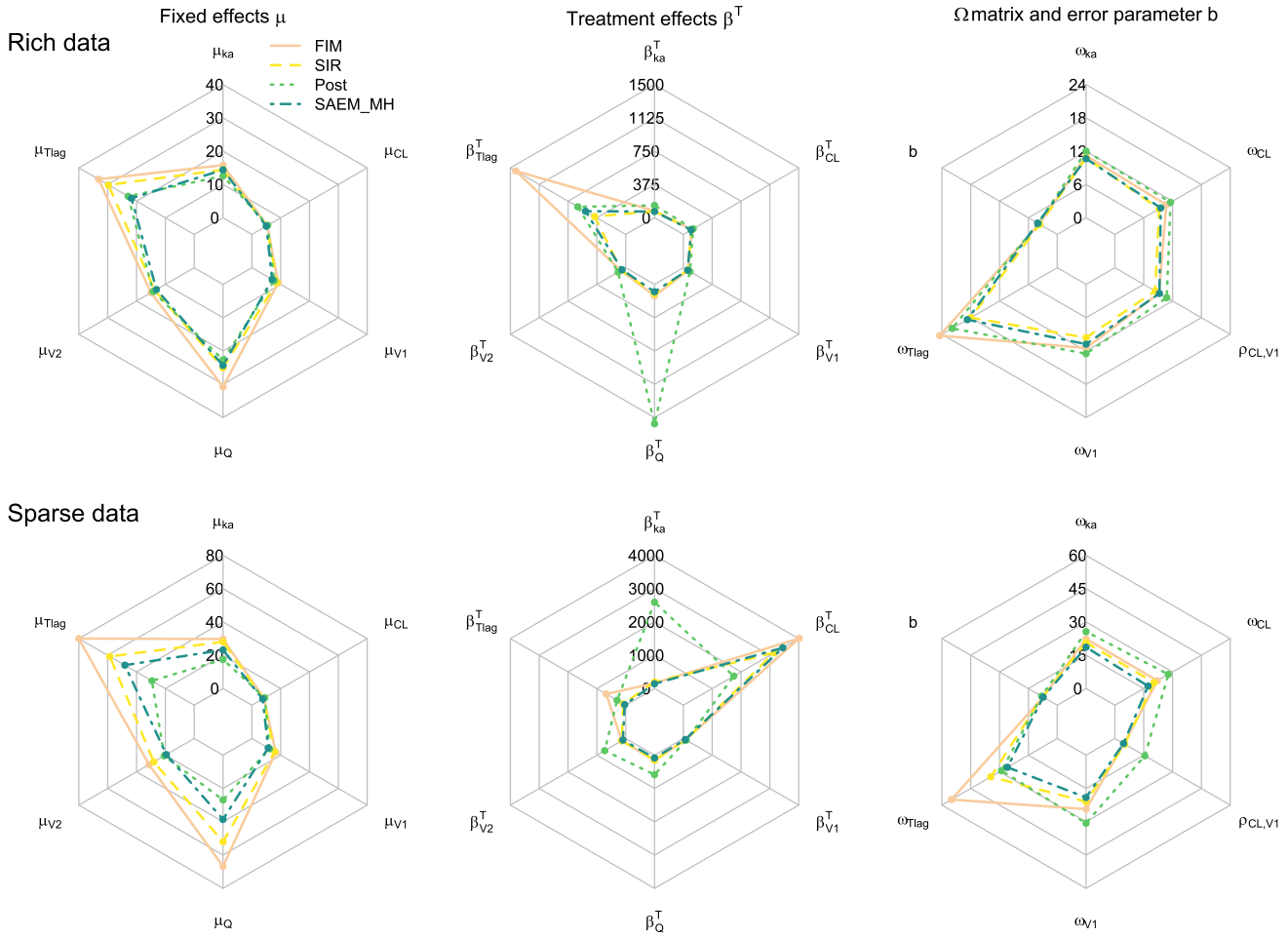


Fig. 3 Starplots of the RSE obtained with Asympt, SIR, Post and SAEM_MH with an inflation factor of the variance kernel equal to 1, for fixed effects (left), treatment effects (middle) and random effects + error parameter (right) of the model estimated on the full Gantenerumab dataset (top) and the sparse subset of N=12 subjects (bottom)

0, indicating the algorithm seemed unable to sample in the posterior distribution correctly. The corresponding numerical results are given in Table III in Appendix 4.

Discussion

In the present work, we built a MH algorithm to sample from the posterior distribution of the population parameters during the smoothing phase of the SAEM algorithm in the `saemix` package for R. We evaluated the algorithm in a simulation study using a simple PK model and two scenarios, a rich design where the asymptotic equivalence between the Bayesian posterior distribution and the distribution of the maximum likelihood estimator is expected to hold, and a sparse design to represent finite distance conditions. We compared the coverage rates for model parameters obtained with our semi-Bayesian algorithm and with other approaches including the asymptotic method based on the FIM, the SIR and the semi-Bayesian Post approach implemented in Stan with the HMC algorithm.

On the rich design, good coverage rates were obtained for Asympt and Post. The performance of our algorithm was reasonable but failed to achieve proper coverage for some parameters. SIR also showed a slight undercoverage for the variance of the clearance parameter, but less than SAEM_MH.

The Asympt method showed its limits when working on a sparse design: as reported in the literature, it underestimated the uncertainty in the parameters. SIR does not improve on those results. Post shows the satisfactory estimation of SE, supporting our assumption that the Bernstein-von Mises theorem still holds at finite distance, and is a strong argument in favour of a Bayesian method of SE computation at finite distance. However, based on $\hat{R}$ as a tool to diagnose the poor performance of the Post algorithm, we could not use the results of Post in 8.4% of datasets. This shows that Post was

also challenged by the sparse design of this study, though it worked well when the available diagnostic criteria were met. We also ran a fully Bayesian algorithm in stan, using the true simulated parameters as initial values instead of the frequentist estimates and lengthening the HMC chains, and we obtained similar results to Post (see Appendix 3), which shows that Post approaches the Bayesian posterior distribution well. Of note, a larger number of sparse datasets gave large $\hat{R}$, which suggests that using `saemix` estimates as initial values helped the HMC algorithm to converge.

Our implementation with a MH algorithm was not 100% successful, showing undercoverage for some parameters. Inflating the variance of the kernel improved the coverage rates and allowed our approach to outperform Asympt and SIR, but even with a factor 2 of inflation, the coverage rates for ω_{CL} and ω_V were too low. We explored various calibration settings, e.g. different prior distributions, which did not change the results on SE computation. We also explored increasing the length of the chains, the number of samples for individual parameters at each step of the algorithm, or running a single MH chain at the end of the smoothing phase of SAEM which did not change the results either. Acceptation rates seemed promising but not sufficient to guide the choice of the inflation rate on the variance of the kernel. Perspectives for a better diagnostic tool for SAEM_MH could include separate AR for fixed and random parameters with different inflation factors.

We then evaluated and compared our method on Gantenerumab real data. The model was a slightly more complex PK model than the one used in the simulation study, and the variability structure was more complex, with higher inter-individual variabilities and correlation between random effects. On the full dataset, Asympt and SIR gave results concordant to Post: following the Bernstein-von Mises theorem, we can assume that this design was close enough to the asymptotic. This application also shows it is hard to identify where finite distance starts, as this depends on the structural and variability model used. Therefore, we applied the methods to a sparse subset on this data, to find a departure between the frequentist methods (Asympt/SIR) and the Bayesian one (Post). SAEM_MH gave satisfactory results on the full dataset, but performed poorly at finite distance due to low acceptation rates. Our assumption is that it was too challenged by the high inter-individual variation coefficients and the correlation between random effects, as well as the overall higher number of parameters to estimate.

Following this application, we extended our simulation study with an additional scenario presenting higher variances and correlation in the random effects, some challenging features encountered in the application and likely to appear in real clinical data. In this situation, acceptation rates for SAEM_MH collapsed to zero alerting on the method's unsuitability for SE computation (see Fig. 17 in Appendix 5), despite the fact that we used the same prior distributions in both Bayesian methods. Stan also has more reliable diagnostic tools and is faster in drawing samples as HMC is more efficient and stan is well optimised (12, 18). Comparing SIR and SAEM_MH, SIR performed better in the more challenging setting, despite using the same proposal distribution as the kernel distribution in the SAEM_MH method, and could be tuned further by inflating the variances.

Perspectives of this work include evaluating variations of the SAEM_MH method: some solutions would be to sample parameters per blocks (e.g. of fixed and random effects following the structure of the FIM) or use conditional univariate kernels or a Gibbs sampler. This would overcome the high-dimensionality problem that challenges the MH algorithm (18) and could help with increasing the acceptation rates. Indeed, in our first simulation study, θ was of size 10 $versus$ 18 and 13 in the real case and the second simulation study, and our results suggest that the acceptation rates were sensitive to the dimension of θ. We also suspect that SAEM_MH is sensitive to the complexity of the Ω matrix. Therefore, by using block or univariate kernels, we could adapt multiple acceptation rates separately and dynamically during the algorithm. Another perspective would be to move away from the MH algorithm and use the approximate Bayesian computation (ABC) algorithm which would allow us to base our sample acceptation on criteria other than the likelihood. Indeed, in NLMEM, the likelihood is non-tractable, and even linearising the model remains computationally heavy and can be inaccurate, especially if the non-linearity of the model is strong. Using the ABC algorithm means we could avoid computing the likelihood and define instead a suitable distance to minimise, which would not be sensitive to the dimensionality of θ.

Here, we focused on parameter SE computation, and we did not extend the evaluation to type I error of Wald tests for treatment effects. Our simulation setting was not designed to assess type I errors on the treatment effect on CL and V, and testing an effect on ka with this model and design would not be realistic clinically. However, several studies have demonstrated the link between underestimated uncertainty around the parameter of interest and the inability to control the type I error of tests (5, 6, 15, 19).

Conclusion

We developed a new method of uncertainty computation to handle sparse data design in NLMEM that showed improve-

ment compared to the classical SE computation methods on a simple PK design. The Bayesian paradigm seems promising to overcome the lack of information in sparse data to assess parameter uncertainty, and an integrated method in the SAEM algorithm was appealing. However, SAEM_MH found more complex data challenging, especially involving complex variability structures, which often occurs in real case studies as we found in the Gantenerumab study. Further work is needed to calibrate the method and overcome its limitations.

Appendix

SAEM

- The probability $p(y_i|\psi_i)$ of the observations $y_i = \{y_{ij}\}$ for subject i is supposed to depend on a known distribution (mostly Gaussian) and on the individual parameters ψ_i
- Modelling of the joint distribution $p(y, \psi, \theta; A)$ (for several subjects)

$$p(y, \psi, \theta; A) = p(y|\psi, \theta)\ p(\psi|\theta)\ p(\theta|A) \tag{10}$$

  - $\theta = (M, B, \Omega)$ is the population parameter vector
  - ψ is the individual parameter matrix
  - A is the (potential) hyperparameter vector of the prior on θ

The EM algorithm is a two-step iterative method developed to compute the likelihood in case of missing data (here, the individual parameters which are not observed):

- Computation of the conditional expectation of the log-likelihood l of the complete data $u = (y, \psi)$ knowing the incomplete data y and the current estimation θ^k (E):

$$Q(y; \theta|\theta^k) = \mathbb{E}(l(\theta; u)|y; \theta^k) = \int l(\theta; u)p(\psi|y; \theta^k)d\psi \tag{11}$$

- Maximisation of the conditional expectation of the log-likelihood l of the complete data (M):

$$\theta^{k+1} = \operatorname{Arg}\max_{\theta} Q(y; \theta|\theta^k) \tag{12}$$

In the SAEM algorithm, the step E is approximated by simulating a single value in the conditional law of the complete data $p(u|y; \theta^k)$ at each step, which is considerably faster than a stochastic integration.

At iteration k of SAEM:

- *Simulation step*: draw ψ_k from the conditional distribution $p(\cdot|y; \theta_k)$.
  - because we do not have access to the conditional distribution, the simulation-step is replaced at iteration k by m iterations of the Metropolis-Hastings algorithm
  - in default algorithm, 2 iterations of each of 3 successive kernels
- *Stochastic approximation*: update conditional loglikelihood l at iteration k, $l_k(\theta)$, according to:

$$l_k(\theta) = l_{k-1}(\theta) + \gamma_k(\log p(y, \psi_k; \theta) - l_{k-1}(\theta)) \tag{13}$$

  where (γ_k) is a decreasing sequence of positive numbers such that $\gamma_1 = 1$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.
- *Maximisation step*: update θ_k as

$$\theta_k = \operatorname{Arg}\max_{\theta} l_k(\theta) \tag{14}$$

The first K_1 iterations of SAEM are the exploratory phase, and the K_2 following iterations of SAEM are the smoothing phase.

MH algorithm in simulation step:

- for $i = 1, 2, \ldots, N$ and denoting $\psi_{i,0} = \psi_i^{(k-1)}$, for $p = 1, 2, \ldots, m$, the MH algorithm consists in:
  1. draw $\tilde{\psi}_{i,p}$ using the proposal kernel $q_{\theta_k}(\psi_{i,p-1}, \cdot)$
  2. set $\psi_{i,p} = \tilde{\psi}_{i,p}$ with probability

$$\alpha(\psi_{i,p-1}, \tilde{\psi}_{i,p}) = \min\left(1, \frac{p(\tilde{\psi}_{i,p}|y_i; \theta_k)q_{\theta_k}(\tilde{\psi}_{i,p}, \psi_{i,p-1})}{p(\psi_{i,p-1}|y_i; \theta_k)q_{\theta_k}(\psi_{i,p-1}, \tilde{\psi}_{i,p})}\right) \tag{15}$$

  3. set $\psi_{i,p} = \psi_{i,p-1}$ with probability $1 - \alpha(\psi_{i,p-1}, \tilde{\psi}_{i,p})$.
- let $\psi_i^{(k)} = \psi_{i,m}$

In both Monolix (3) and `saemix` (4), 2 iterations of 3 successive transition kernels are used.

Maximum Likelihood Estimator of the Population Parameters θ: θ is not considered a random variable but as a fixed parameter.

$$\hat{\theta}_{ML} = \operatorname{Arg}\max_{\theta} Q(y; \theta) \tag{16}$$

Simulation Study—Zip Plots

Zip plots are a way to visualise why coverage is controlled or not in a simulation study by displaying all confidence interval (CI) estimates (17). For any estimated parameter,

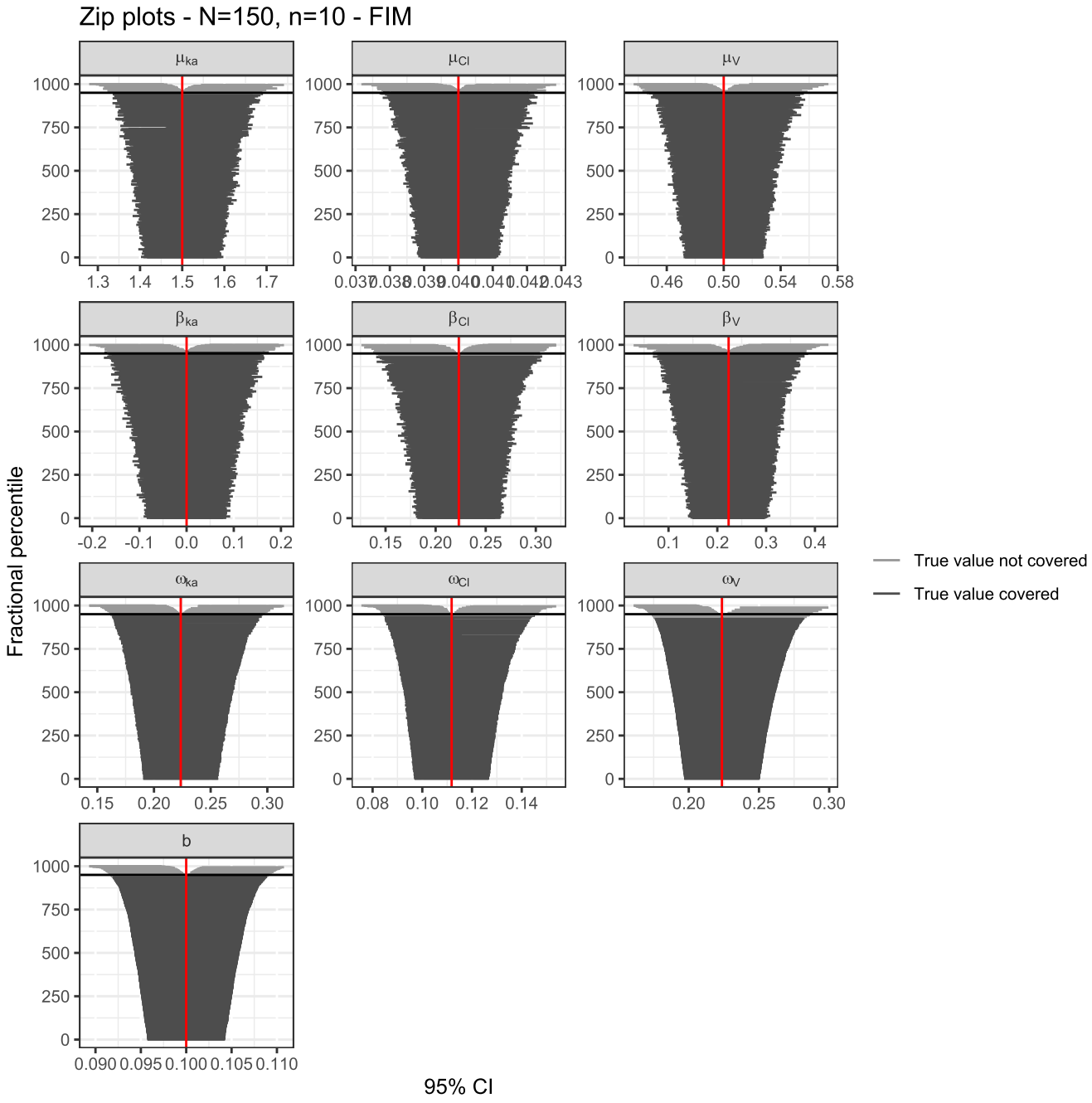


Fig. 4 Zip plot obtained with Asympt on the rich design

confidence intervals are displayed for each simulated dataset, ranked according to the ratio of bias over SE. Ranks are then plotted against the confidence intervals, emphasizing in grey those who do not cover the simulated value. A horizontal line represents the 0.95 target coverage rate.

This kind of plot allows to show bias (when the deviation of CI from the true value is heavier on one side) and SE over/underestimation (when the grey CI not covering the true value are below/above the target value) at the same time.

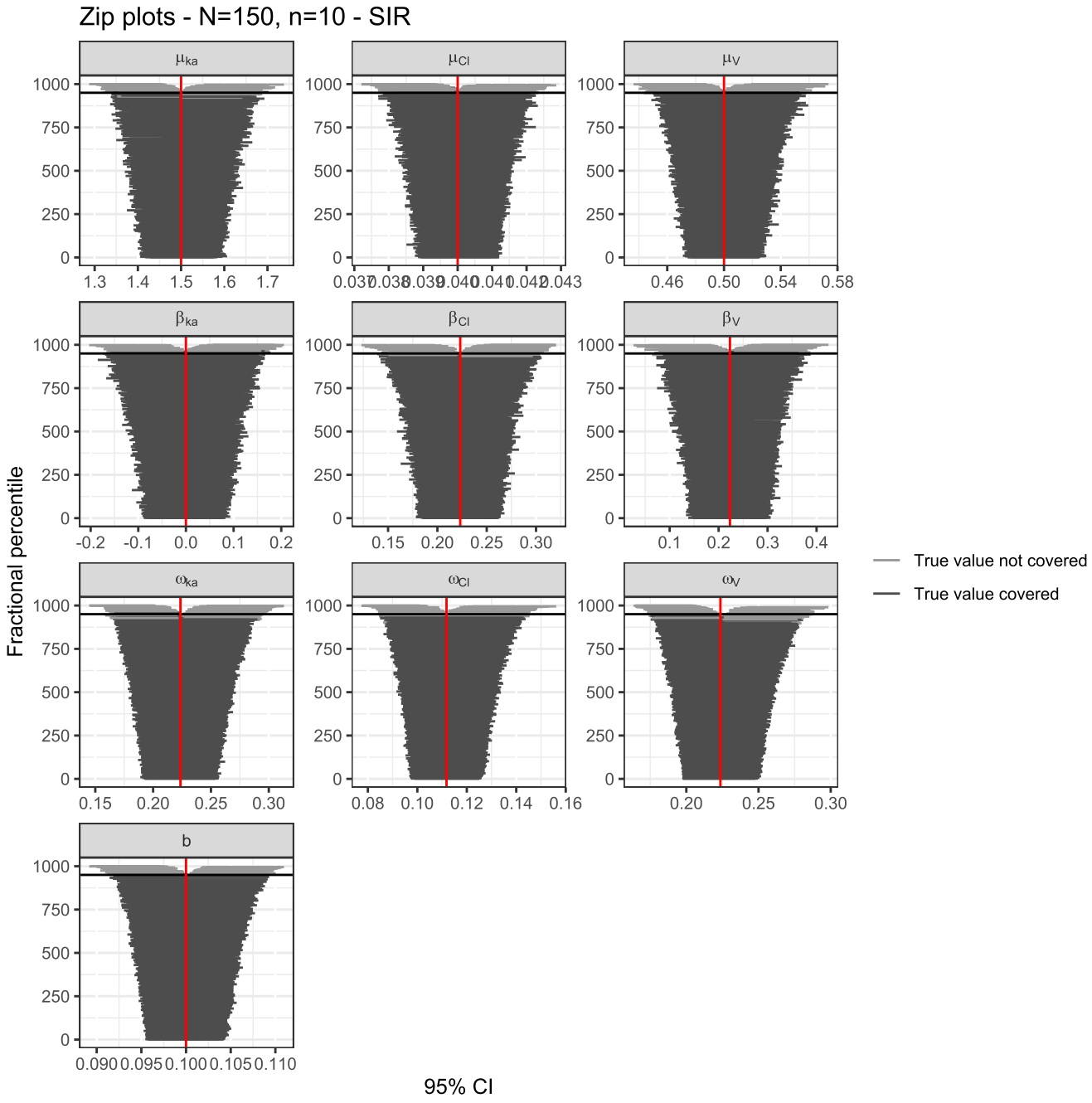


Fig. 5 Zip plot obtained with SIR on the rich design

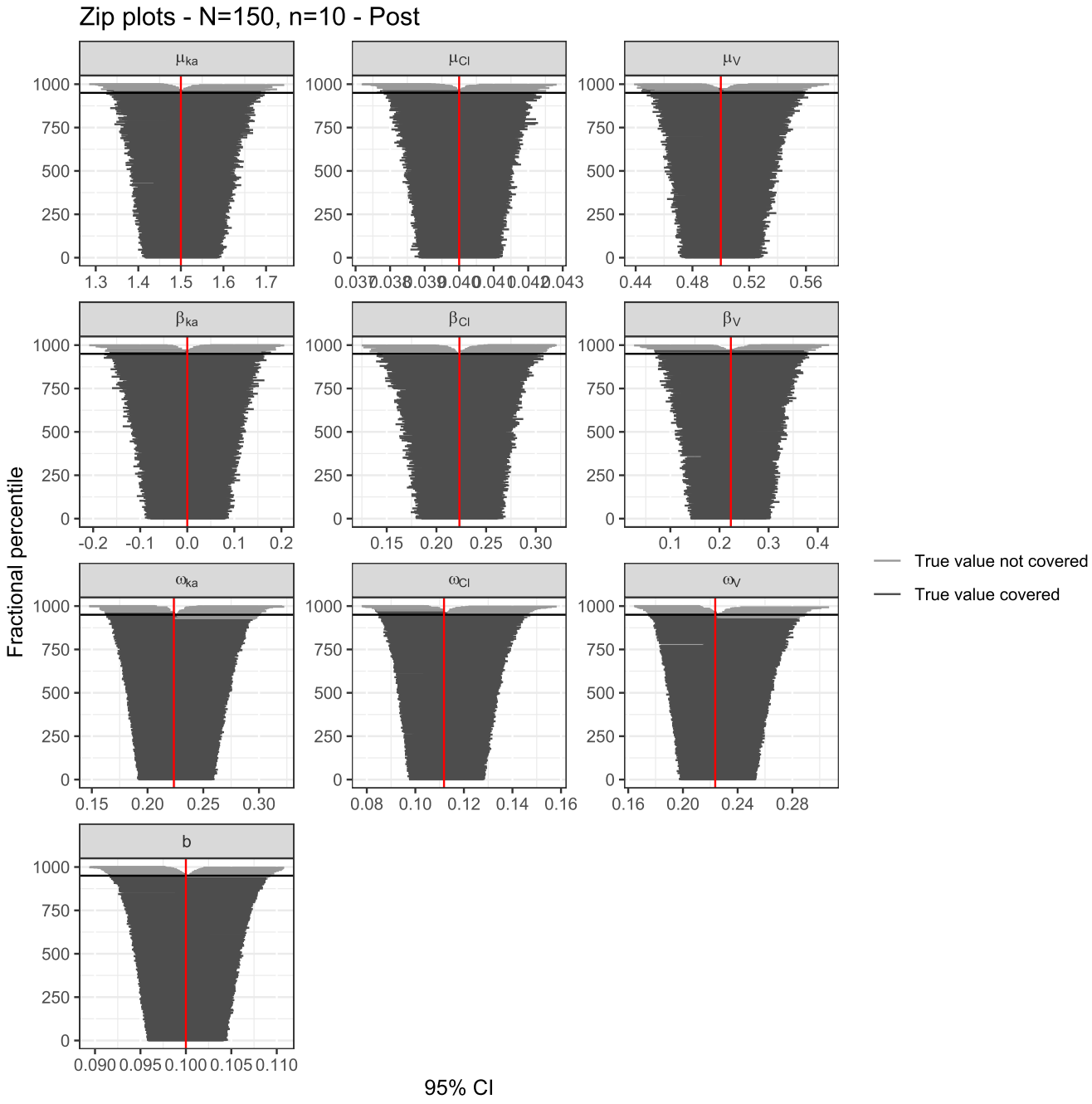


Fig. 6 Zip plot obtained with Post on the rich design

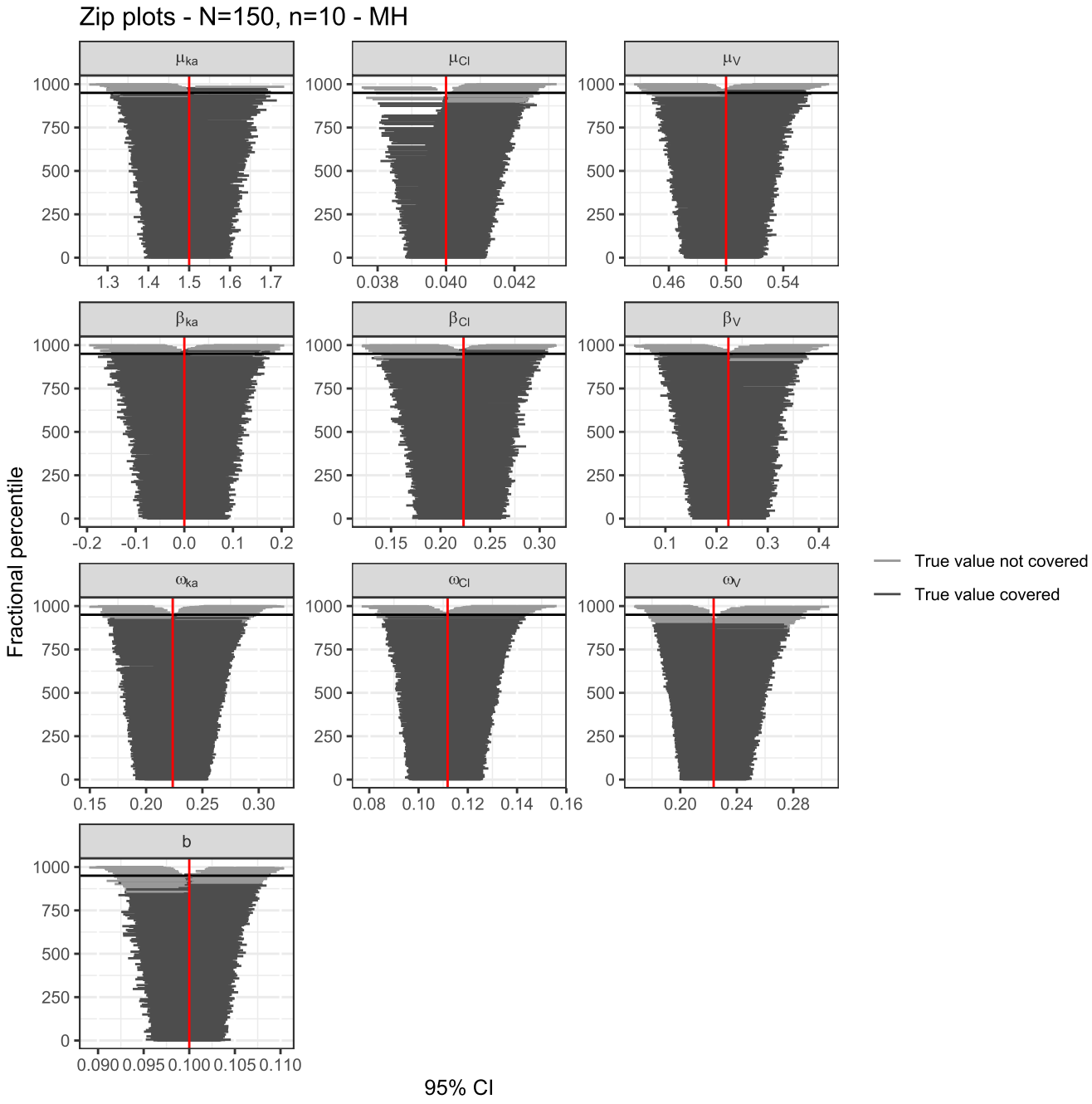


Fig. 7 Zip plot obtained with MH on the rich design

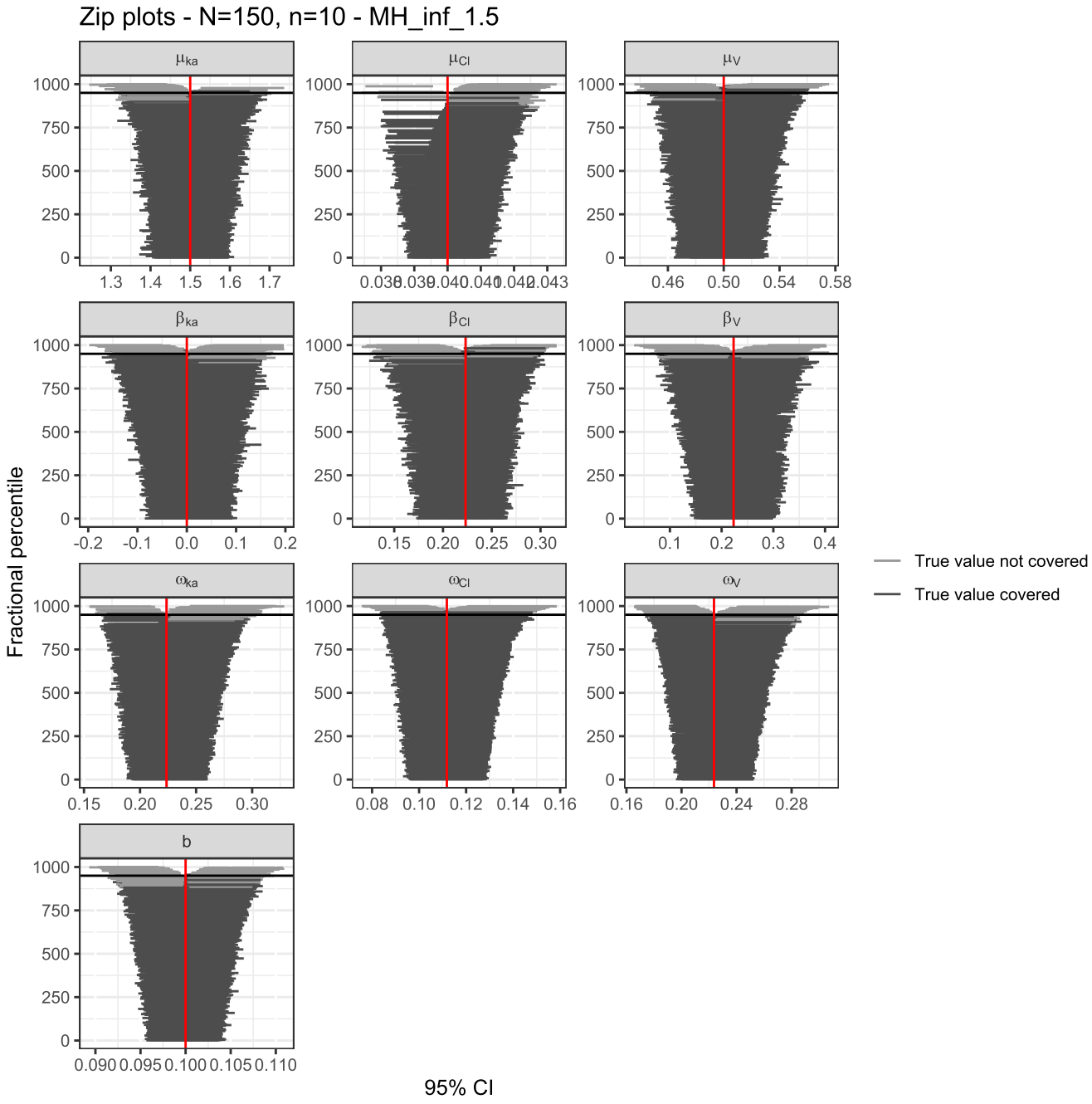


Fig. 8 Zip plot obtained with MH—variance inflated by 1.5—on the rich design

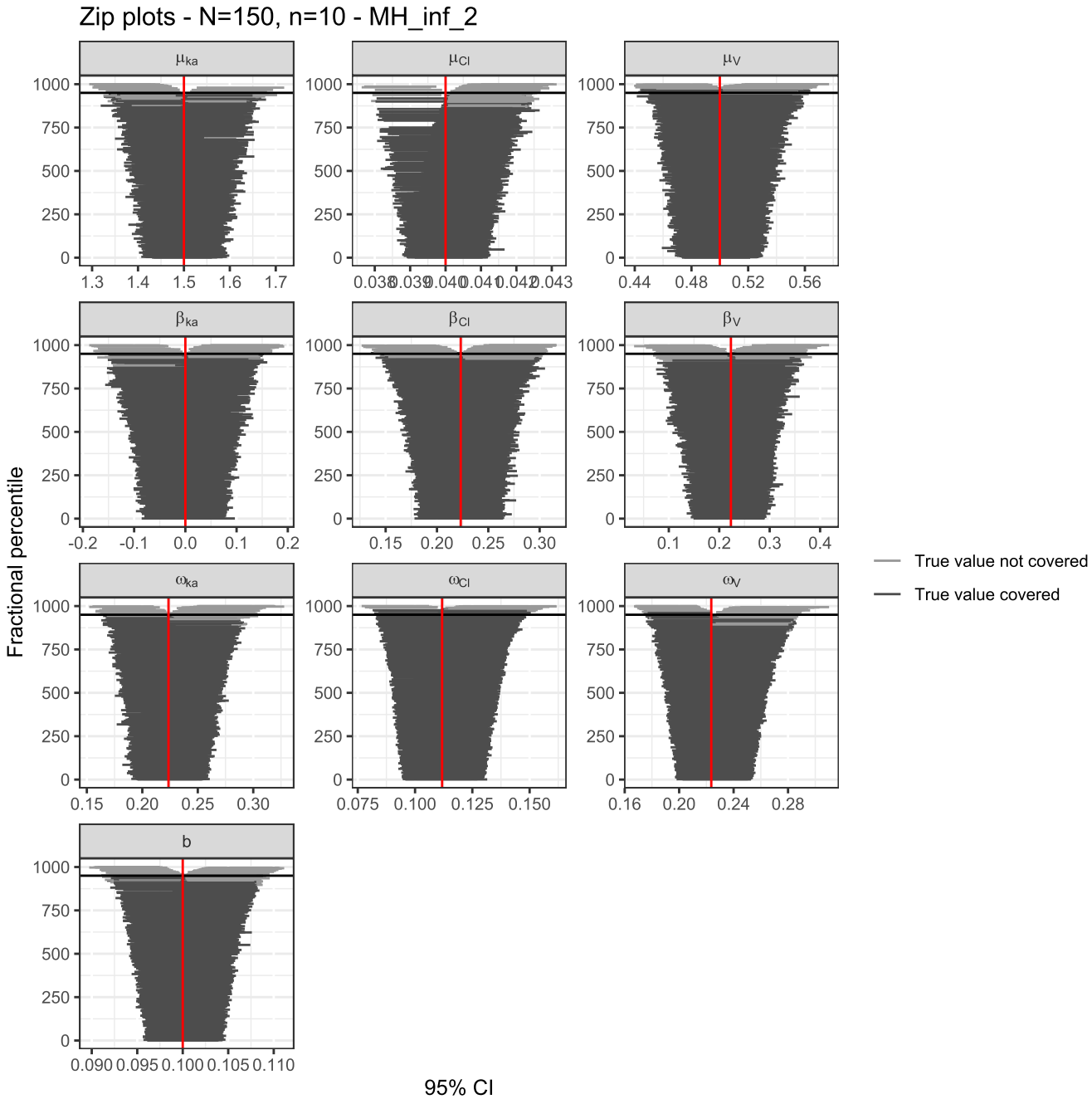


Fig. 9 Zip plot obtained with MH—variance inflated by 2—on the rich design

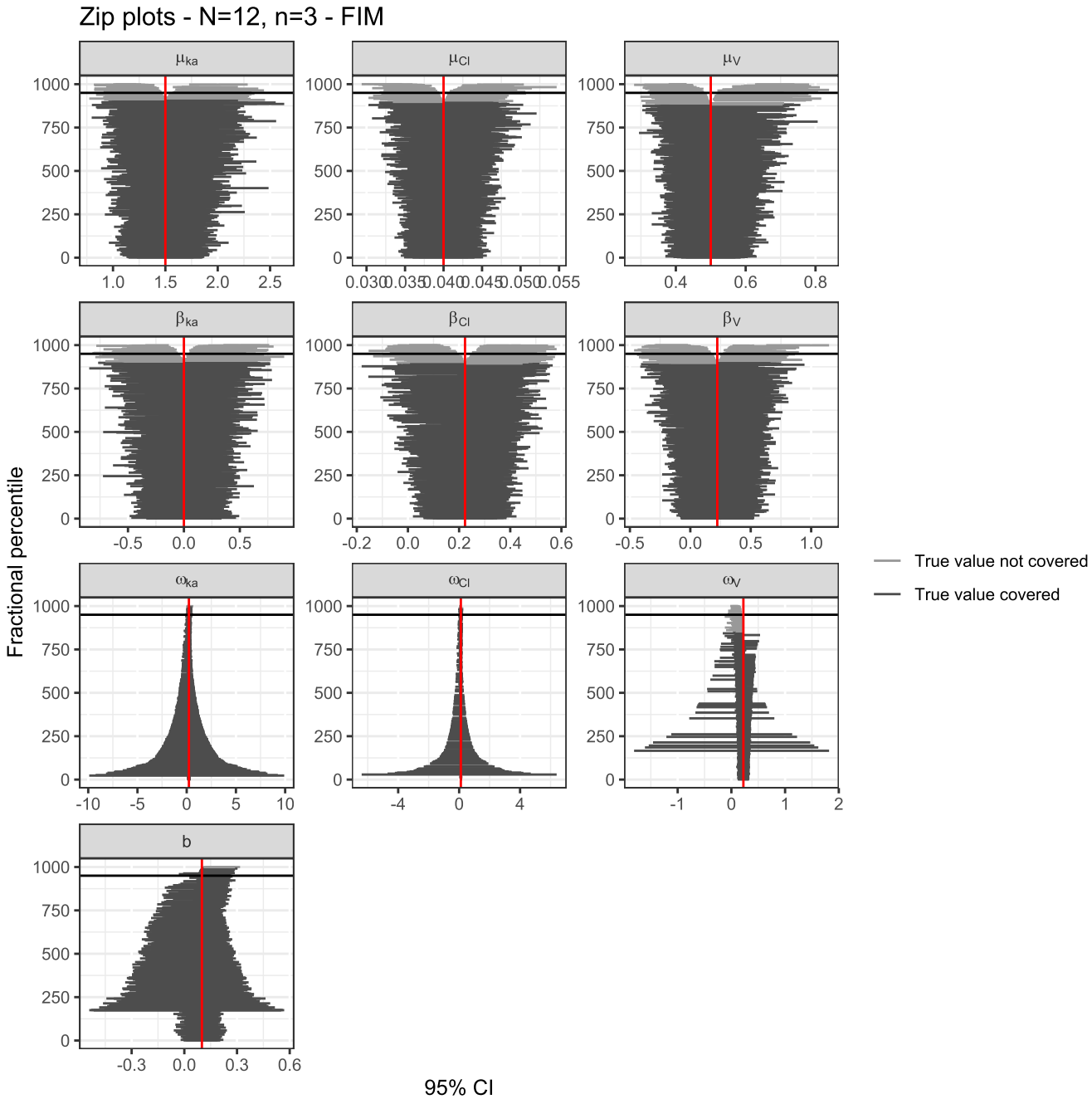


Fig. 10 Zip plot obtained with Asympt on the sparse design

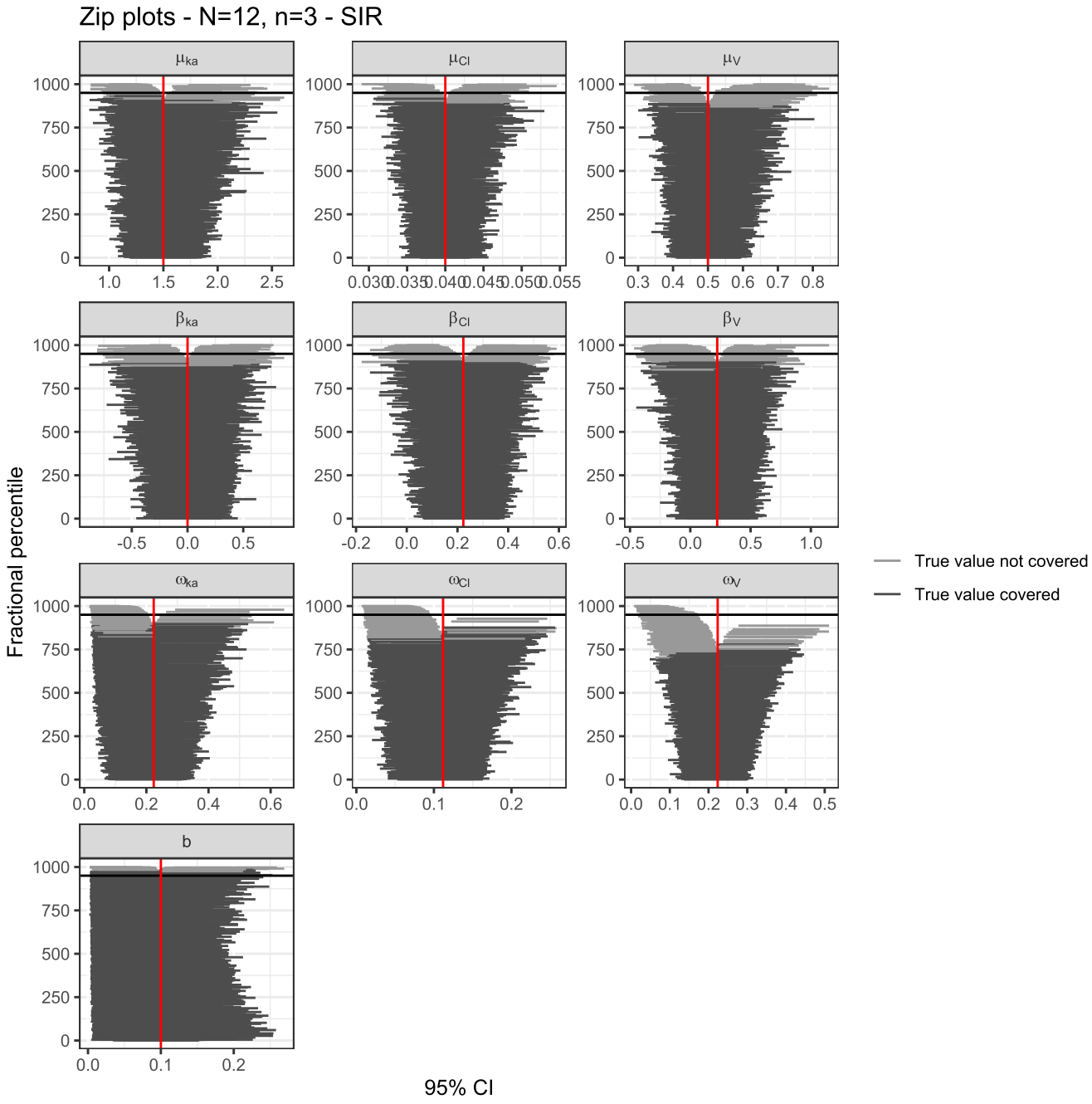


Fig. 11 Zip plot obtained with SIR on the sparse design

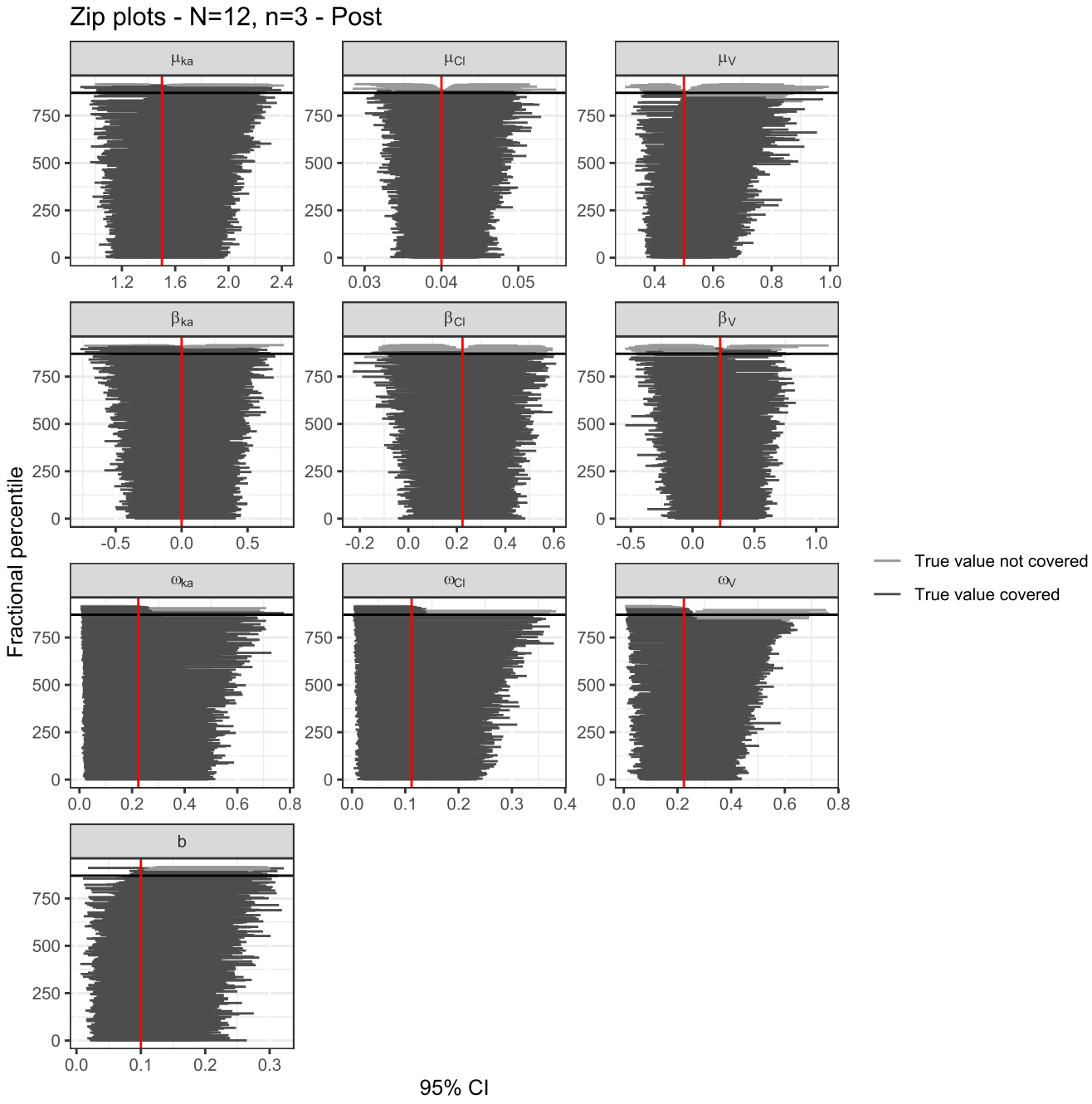


Fig. 12 Zip plot obtained with Post on the sparse design

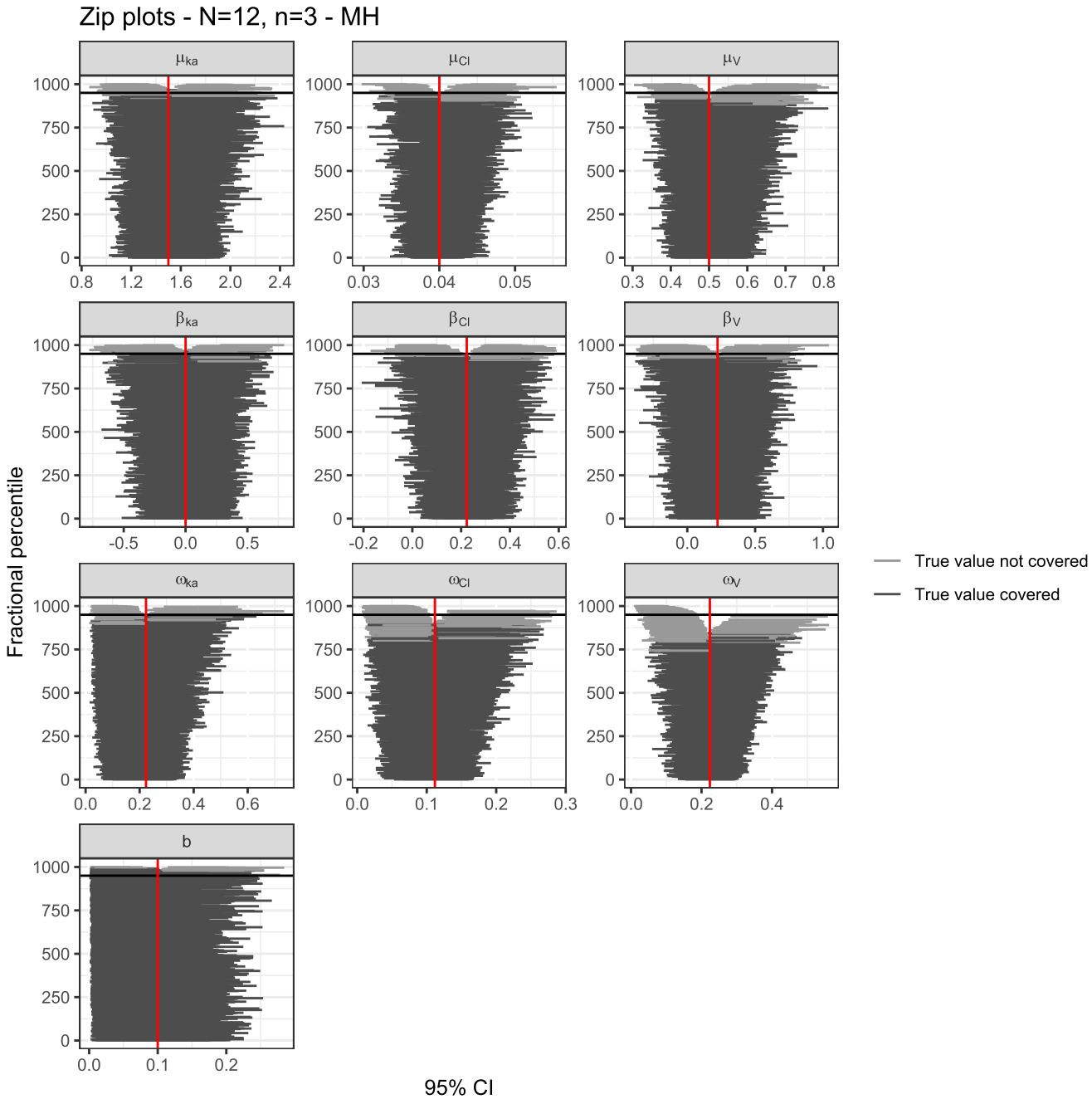


Fig. 13 Zip plot obtained with MH on the sparse design

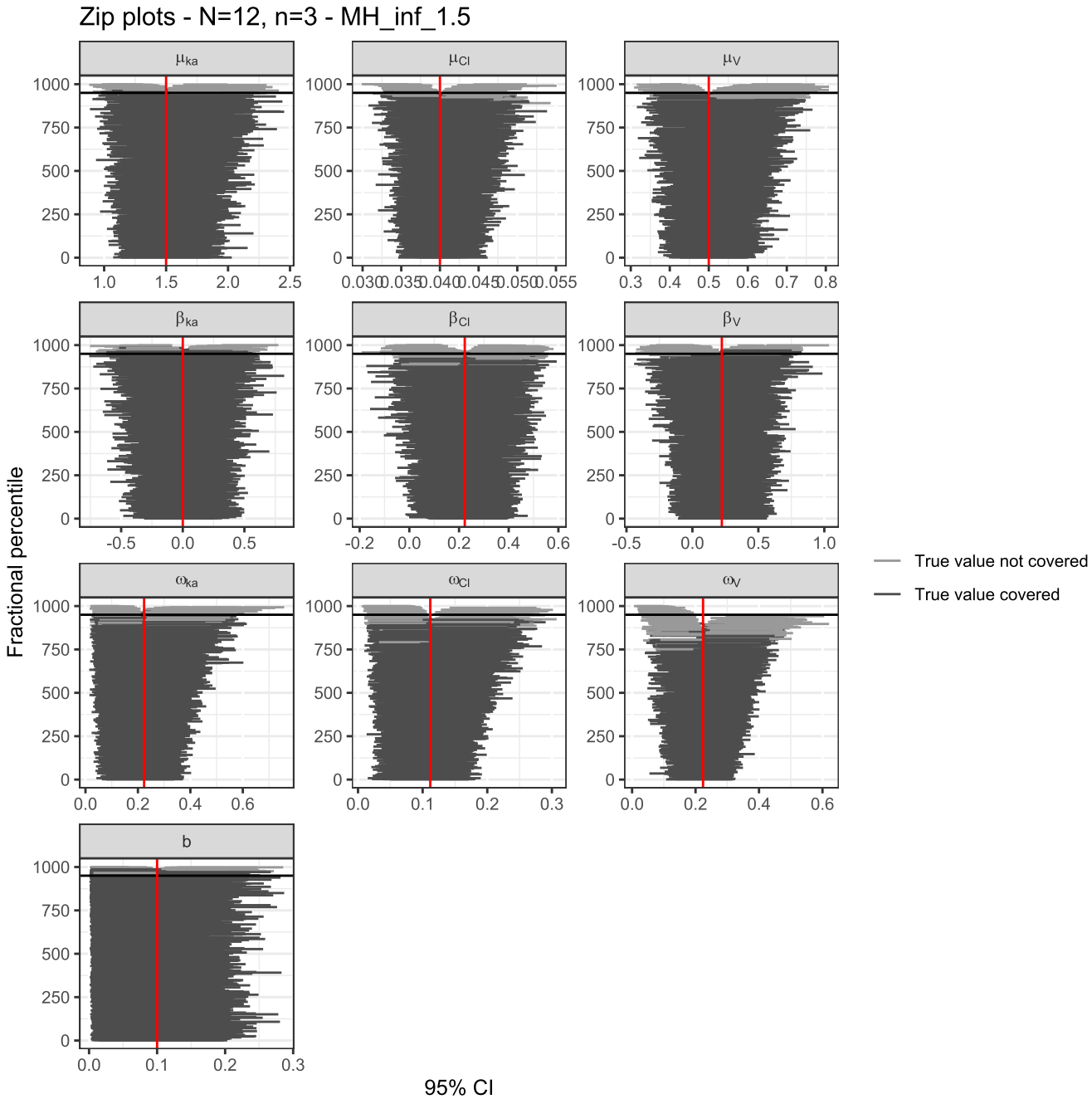


Fig. 14 Zip plot obtained with MH—variance inflated by 1.5—on the sparse design

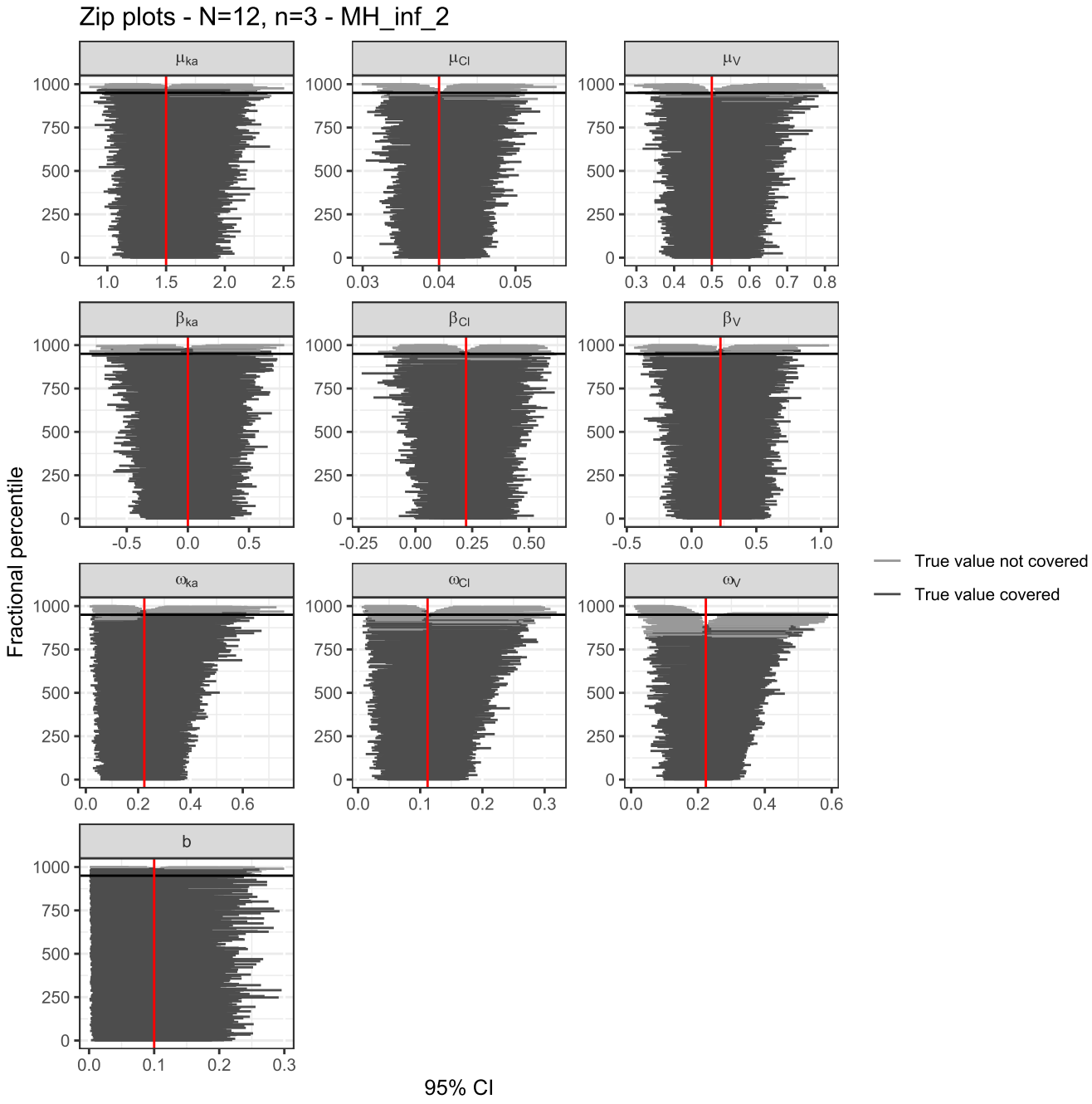


Fig. 15 Zip plot obtained with MH—variance inflated by 2—on the sparse design

Simulation Study—Comparison of Post with a Full Bayesian Method

We compared the RSE obtained with Post and a full Bayesian approach as implemented through the HMC algorithm in Stan (12), hereafter called HMC, on a subset of 100 datasets. HMC was run using Stan (12). We ran three chains of 11,000 iterations (including 1000 warm-up iterations). The prior distribution was the same as for Post.

The initial values were the true values on the rich data, and lower Ω values on the sparse datasets to help the chains converge. Even with that help, 23% of the datasets gave too high

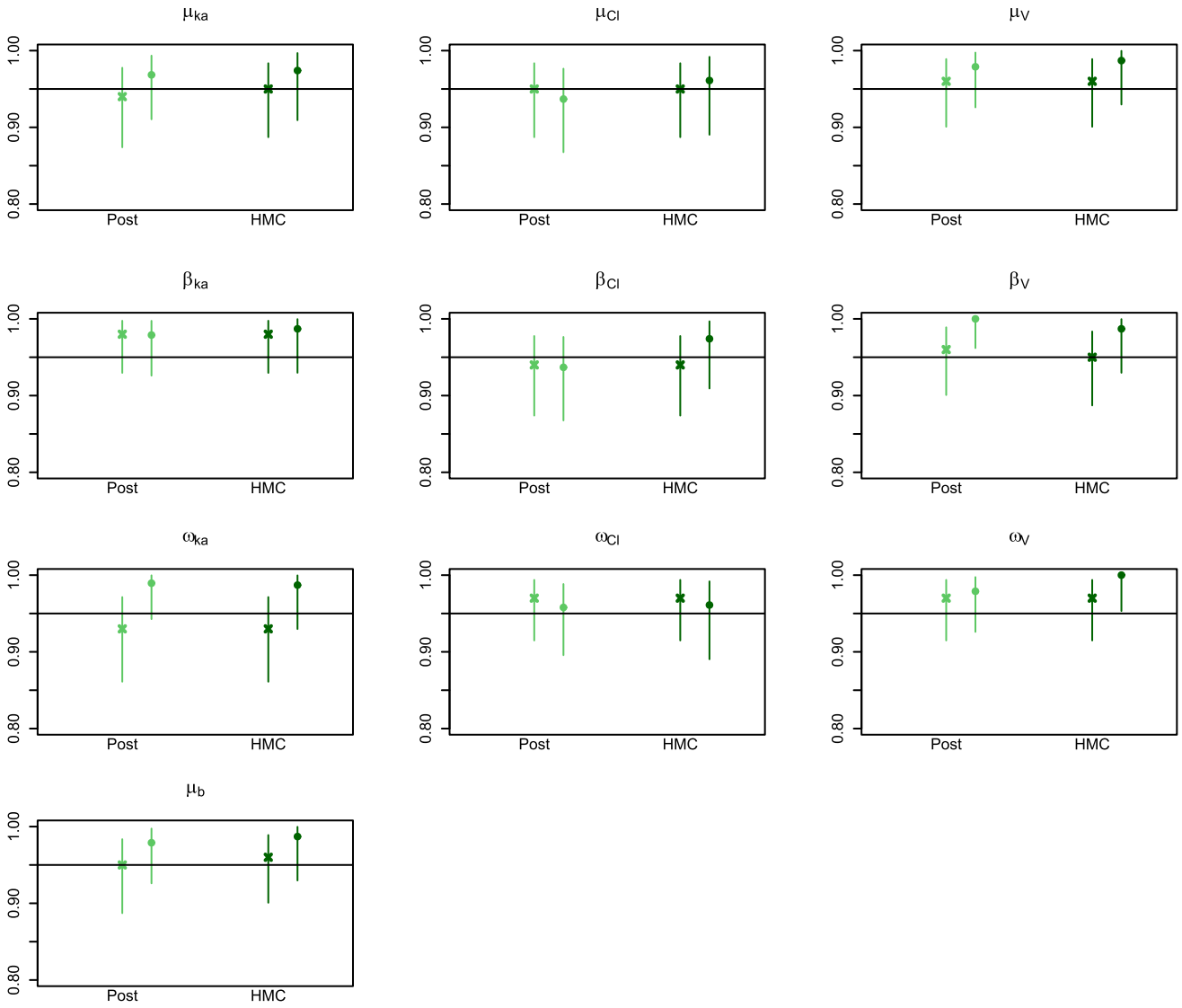


Fig. 16 Comparaison of RSE obtained for the different model parameters with stan (■) and HMC (■) methods on 100 rich (X) and sparse (O) datasets

$\hat{R}$ for the results to be used, which suggests that the HMC algorithm is sensitive to initial values and that in the Post method, initial values taken from the `saemix` fit helped get better performances.

In Fig. 16, we compare the results obtained with a full Bayesian algorithm (method called HMC) to those obtained with Post and show that they are very similar.

Tables of RSE on Gantenerumab Real Data

In Tables II and III, we present the results shown on the star plots of Fig. 3 with an additional column showing that once again the results obtained with a full Bayesian algorithm (method called HMC) are very similar to the Post results.

Table II Relative Standard Errors Computed for the Parameters of the Model Fitted on the Full Dataset (N=48) with Asympt, SIR, Post, HMC and SAEM_MH Methods

Parameter	Asympt	SIR	Post	HMC	SAEM_MH
μ_{ka}	16	14	13	13	15
β_{ka}	77	75	140	134	73
$\mu_{CL/F}$	5	5	6	5	5
$\beta_{CL/F}$	108	109	127	120	97
$\mu_{V_1/F}$	9	9	8	8	9
$\beta_{V_1/F}$	61	56	86	86	65
$\mu_{Q/F}$	31	25	23	23	29
$\beta_{Q/F}$	129	108	1,570	1,257	127
$\mu_{V_2/F}$	15	14	14	14	14
$\beta_{V_2/F}$	63	59	99	99	64
$\mu_{T_{lag}}$	33	30	23	23	27
$\beta_{T_{lag}}$	1430	404	623	518	697
ω_{ka}	11	11	12	12	11
$\omega_{CL/F}$	11	9	12	11	11
$\rho_{CL/F,V_1/F}$	9	8	11	11	9
$\omega_{V_1/F}$	11	10	12	13	11
$\omega_{T_{lag}}$	24	18	22	22	20
b	4	4	4	4	4

Table III Relative Standard Errors Computed for the Parameters of the Model Fitted on the Sparse Subset of the Data (N=12) with Asympt, SIR, Post, HMC and SAEM_MH Methods

Parameter	Asympt	SIR	Post	HMC	SAEM_MH
μ_{ka}	30	28	17	18	17
β_{ka}	162	199	2594	1990	84
$\mu_{CL/F}$	8	8	9	9	5
$\beta_{CL/F}$	4017	3144	1747	2007	2461
$\mu_{V_1/F}$	17	16	12	12	8
$\beta_{V_1/F}$	77	81	100	102	40
$\mu_{Q/F}$	67	52	27	26	30
$\beta_{Q/F}$	151	149	585	593	60
$\mu_{V_2/F}$	31	28	21	21	14
$\beta_{V_2/F}$	136	137	729	706	63
$\mu_{T_{lag}}$	70	59	29	29	45
$\beta_{T_{lag}}$	674	83	297	277	28
ω_{ka}	22	21	26	25	12
$\omega_{CL/F}$	22	20	28	29	13
$\rho_{CL/F,V_1/F}$	5	5	16	16	3
$\omega_{V_1/F}$	24	21	31	31	14
$\omega_{T_{lag}}$	55	35	29	29	19
b	7	8	8	8	5

Extension of the Simulation Study

Following the real case study, we extended our simulation study with more challenging features encountered in the application, i.e. higher variances and correlation in the random effects. We simulated 100 sparse datasets with the same settings as the first part of the simulation study except for the inter-individual variability matrix Ω: $\omega_{ka} = \omega_{CL} = \omega_V = 1.1$ and $\rho_{ka,CL} = 0.9$, $\rho_{ka,V} = 0.98$, $\rho_{CL,V} = 0.9$. Figure 17 shows the coverage rates obtained on this extension of the simulation study. Asympt gave more accurate coverage rates because the high correlations between parameters compensated for the sparsity of data, although uncertainty was still underestimated, as it was with SIR. Post method coverage rates were below the target for all parameters although $\hat{R}$ were not particularly higher and the proportion of datasets

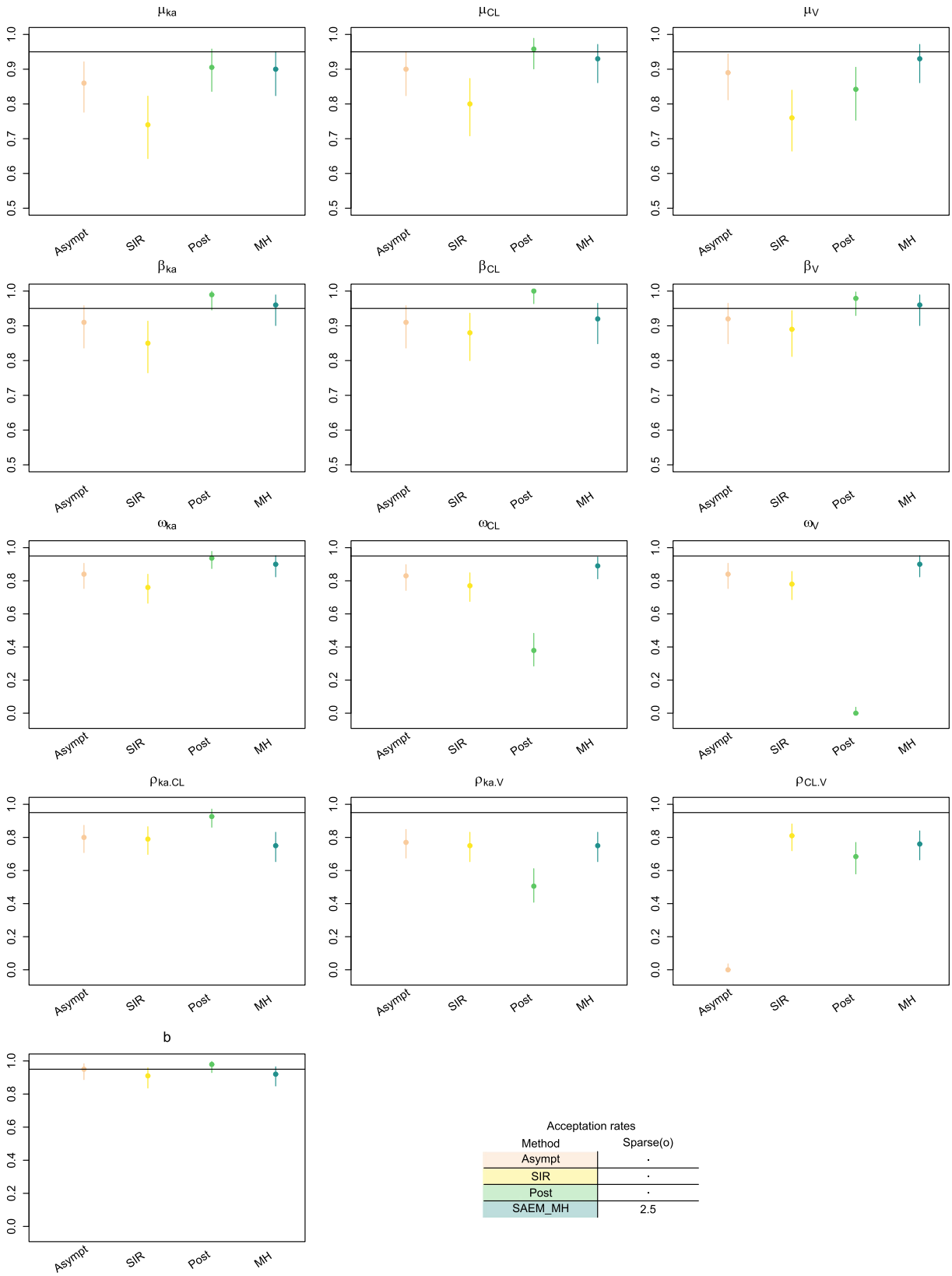


Fig. 17 95% coverage rates obtained with Asympt (), SIR (), Post (), SAEM_MH with inflation factor of the variance kernel at 1 ()

failing to converge was stable. SAEM_MH acceptation rates were very low (2.5%), indicating that the method was unable to sample sufficiently to assess the uncertainty. Indeed all coverage rates were below the target.
Of note, five datasets could not be used for the Post method due to $\hat{R}$ being too high.

Acknowledgements The authors would like to thank Dr. Maud Delattre for her valuable input on the methods of this article, and Hervé Le Nagard, Lionel de la Tribouille and Rémy Bertino for the use of the CATIBioMed calculus facility. The illustrative example data were obtained from studies sponsored by F Hoffmann-La Roche. We thank the participants and investigators who participated in these studies.

Author Contribution MG, JB and EC designed the study. MG and LF performed the analyses. MG wrote the article. JB and EC critically revised it. FM provided the data and approved the final version of the article.

Data Availability The real data used in this article comes from a clinical trial conducted by F Hoffmann-La Roche and is not publicly available. Requests for access should be made to François Mercier.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Bauer RJ. NONMEM Tutorial Part I: description of commands and options, with simple examples of population analysis. CPT: Pharmacometrics Syst Pharmacol. 2019;8(8):525–37.
2. Bauer RJ. NONMEM Tutorial Part II: estimation methods and advanced examples. CPT: Pharmacometrics Syst Pharmacol. 2019;8(8):538–56.
3. Lavielle M. Mixed effects models for the population approach: models, tasks, methods and tools. Chapman and Hall/CRC; 2014.
4. Comets E, Lavenu A, Lavielle M. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. J Stat Softw. 2017;80:1–41.
5. 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):2582–600.
6. Loingeville F, Bertrand J, Nguyen T, Sharan S, Feng K, Sun W, et al. New model-based bioequivalence statistical approaches for pharmacokinetic studies with sparse sampling. AAPS J. 2020;22(6):141.
7. Bertrand J, Comets E, Chenel M, Mentré F. Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. Biometrics. 2012;68(1):146–55.
8. Thai H, Mentré F, Holford N, Veyrat-Follet C, Comets E. Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. J Pharmacokinet Pharmacodyn. 2013;41(1):15–33.
9. Dosne A, Bergstrand M, Harling K, Karlsson M. Improving the estimation of parameter uncertainty distributions in nonlinear mixed effects models using sampling importance. J Pharmacokinet Pharmacodyn. 2016;43(6):583–96.
10. Ueckert S, Riviére M, Mentré F. Alternative to resampling methods in maximum likelihood estimation for NLMEM by borrowing from Bayesian methodology. 2015. p 24. https://www.page-meeting.org/?abstract=3632.
11. Vaart A. Asymptotic statistics (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press; 1998.
12. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan : a probabilistic programming language. J Stat Softw. 2017;76(1).
13. Guhl M, Mercier F, Hofmann C, Sharan S, Donnelly M, Feng K, 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):557–77.
14. Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of EM algorithm. Ann Stat. 1999;27(1):94–128.
15. Panhard X, Mentré F. Evaluation by simulation of tests based on non-linear mixed-effects models in pharmacokinetic interaction and bioequivalence cross-over trials: evaluation of tests based on NLMEM. Stat Med. 2005;24(10):1509–24.
16. Gelman A, Gilks W, Roberts G. Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann Appl Probab. 1997;7(1).
17. Morris T, White I, Crowther M. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102.
18. Neal R. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1. Department of Computer Science, University of Toronto; 1993.
19. Bertrand J, Comets E, Laffont C, Chenel M, Mentré F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. J Pharmacokinet Pharmacodyn. 2009;36:317–39.

Corrigendum

Page 3 of 27 : FIM

$$l(y|\theta,\psi) \simeq -\frac{\boldsymbol{log(2\pi)}}{\mathbf{2}}\sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{n_i} - \frac{1}{2}\sum_{i=1}^{N} log(det\ \Gamma_i) - \frac{1}{2}\sum_{i=1}^{N}(y_i-\bar{y}_i)^t\ \Gamma_i^{-1}\ (y_i-\bar{y}_i)$$

$$\text{where } \bar{y}_i = f(t_i,\widehat{\boldsymbol{\psi_i}}) + \partial_\psi f(t_i,\psi_i)(\boldsymbol{\psi_i} - \widehat{\boldsymbol{\psi_i}})$$
$$\Gamma_i = \partial_\psi f(t_i,\psi_i)\ \Omega\ \partial_\psi f(t_i,\psi_i)^t + g(t_i,\psi_i)\ I_{n_i}\ g(t_i,\psi_i))^t$$

with $y = (y_i)_{i=1,...,N}$, where y_i the vector of n_i observations for individual i, I_{n_i} the identity matrix of size n_i, $det\ \Gamma_i$ the determinant of the matrix Γ_i, $\partial_\psi f(t_i,\psi_i)$ is the gradient of f with respect to ψ calculated in (t_i,ψ_i) **and $\widehat{\boldsymbol{\psi_i}}$ is the conditional expectation of the individual parameters for individual $\boldsymbol{i}$.**

Page 10 of 27 : Discussion
In this situation, acceptation rates for SAEM_MH **were lower** alerting on the method's unsuitability for SE computation (see Fig.17 in Appendix 5), despite the fact that we used the same prior distributions in both Bayesian methods.

Page 11 of 27 : SAEM
— $\boldsymbol{\theta}$ is the population parameter vector
(...)

~~**Maximum likelihood estimator of the population parameters** θ : θ is not considered as a random variable but as a fixed parameter.~~

~~$$\boldsymbol{\hat{\theta}_{ML} = \text{Arg}\max_{\theta} Q(y;\theta)}$$~~

Page 26 of 27 : Figure 17
Acceptation rate for SAEM_MH is **6%**.

Page 27 of 27 : Extension of the Simulation Study
SAEM_MH acceptation rates were **low (**6%**)**, indicating that the method was unable to sample sufficiently to assess the uncertainty.

Chapitre 4

Variations de SAEM_MH et exploration de l'algorithme ABC (Approximate Bayesian Computation)

4.1 Résumé

Objectifs

L'objectif de ce travail était d'explorer les pistes d'amélioration de notre méthode SAEM_MH, et nous avons aussi cherché à déterminer des diagnostics pour la calibrer. L'algorithme MH étant intrinsèquement sensible à la dimension du vecteur de paramètres à estimer, nous avons également considéré une autre alternative basée sur l'algorithme ABC.

Synthèse

Dans ce travail, nous avons mis en oeuvre différentes variations de l'algorithme SAEM_MH permettant de diminuer la dimension du noyau de proposition de l'algorithme MH et de déterminer l'inflation de la variance du noyau de façon dynamique et différenciée selon les

paramètres. Nous avons également utilisé des outils Bayésiens tels que l'échantillonnage de Gibbs et la marche aléatoire. Ces variations avaient pour but d'améliorer les taux d'acceptation très bas observés dans l'algorithme SAEM_MH en présence de modèles complexes.

La première variation explorée est d'échantillonner selon les deux blocs de la FIM, le bloc pour les effets fixes et les effets de covariables, et le bloc pour les éléments de la matrice de variabilité et les paramètres du modèle d'erreur. Cela permet de réduire la dimension des vecteurs à échantillonner tout en conservant la structure des corrélations de la FIM, donc sans perdre d'information. Cela permet aussi de pouvoir calibrer séparément les deux blocs pour ajuster les taux d'acceptation en cas de comportement différent entre les deux blocs. Une version plus extrême de cette idée est d'échantillonner les éléments du vecteur d'intérêt un à un, en conditionnant le noyau de proposition au reste du vecteur fixé. Cela permet de revenir à un problème d'échantillonnage univarié, de calibrer l'algorithme individuellement pour chaque élément du vecteur, mais cela augmente aussi le temps de calcul de la méthode. Nous avons également exploré l'échantillonnage de Gibbs qui consiste à échantillonner les paramètres de population après avoir échantillonné les paramètres individuels, ainsi que l'idée d'échantillonner selon un processus de marche aléatoire. Enfin, l'influence de l'inflation de la variance du noyau de proposition, définie selon des méthodes de calibration automatiques venant de la littérature, a été évaluée.

Ces différentes variations ont été testées, seules et en combinaison, avec différentes étapes de calibration des paramètres de l'algorithme, notamment le nombre de chaînes tirées. L'évaluation de toutes ces méthodes a été faite sur la même étude de simulation que dans le précédent projet. Nous avons évalué les taux de couverture ainsi que les RSE et les biais. Nous avons comparé les différentes méthodes proposées aux méthodes Asympt, SIR, Bootstrap et Post.

L'algorithme Bayésien ABC a également été évalué : il a l'avantage d'être plus rapide en temps de calcul car dans cet algorithme, la vraisemblance n'a pas besoin d'être calculée. Cet algorithme permet de définir soi-même le critère d'acceptation des échantillons remplaçant la vraisemblance, c'est à dire la distance que l'on souhaite garder sous un certain seuil dans les échantillons acceptés. Cela permet, dans le cas de modèles complexes, de définir un critère non sensible à la dimension du vecteur de paramètres à échantillonner.

Certaines variations de SAEM_MH ont permis d'améliorer les taux d'acceptation mais globalement, l'algorithme MH atteint ses limites lorsque le modèle utilisé est trop complexe, c'est-à-dire lorsque le nombre de paramètres de population à estimer augmente (il montre de bons résultats jusqu'à 10 paramètres) et/ou en présence de fortes corrélations entre les paramètres et de forte variabilité de ces derniers. La méthode ABC a donné les résultats les plus intéressants : elle permet vraiment de s'écarter des résultats de la FIM et garde des taux d'acceptation élevés même sur des jeux de données très épars et complexes.

Apports du travail

Ce travail a permis de mettre en évidence le potentiel de l'algorithme ABC pour estimer l'incertitude des paramètres des modèles non linéaires à effet mixtes à distance finie et/ou en présence de structures complexes dans les paramètres. En effet, elle combine les avantages du paradigme Bayésien avec celui de s'affranchir de la vraisemblance comme quantité d'intérêt, tout en dépassant les limites liées à la dimensionnalité de l'algorithme MH, et celles liées à la complexité de la mise en oeuvre de l'algorithme HMC.

4.2 Article en préparation pour une soumission dans Statistics in Medicine

Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

ARTICLE TYPE

Development of a semi-Bayesian SAEM algorithm for finite-distance estimation of parameter uncertainty in nonlinear mixed-effects models

Fayette Lucie*[1] | Bertrand Julie[1] | Comets Emmanuelle[1,2] | Guhl Mélanie[1]

[1]Université Paris Cité , Inserm, IAME, F-75018 Paris, France
[2]Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085, F-35000 Rennes, France

Correspondence
*Fayette Lucie Email: lucie.fayette@inserm.fr

Abstract

This work focuses on the calculation at finite distance of standard errors (SE) on parameters in the context of nonlinear mixed effects models where classical methods based on the Fisher Information Matrix (FIM) underestimate the uncertainty when the asymptotic conditions are not verified. Several methods have already been introduced to overcome this problem, in particular semi-Bayesian methods that approximate the SE by the standard deviations of the posterior distributions. Here, we propose and evaluate variations around the Metropolis Hastings (MH) and the Approximate Bayesian Computation (ABC) algorithms. These methods are evaluated by simulations and on a real dataset. Methods based on MH algorithm are challenged in situations where the variability structure is complex whereas ABC shows promising results.

KEYWORDS:
Nonlinear mixed effect models, finite distance, uncertainty, standard errors, Metropolis Hastings, Approximate Bayesian Computation

1 | INTRODUCTION

Nonlinear mixed-effects models (NMLEM) are increasingly used in the analysis of clinical trials. Their use now extends well beyond pharmacokinetic (PK) and pharmacodynamic studies[1]. In particular, they are increasingly used in the analysis of longitudinal data collected during clinical trials. Indeed, they complement the analysis of the primary endpoint and can provide a greater power to highlight the effect of a treatment[2]. Also, downstream, they allow to anticipate individual trajectories enabling therapeutic monitoring where needed[3]. In all these applications arises the problem of parameter uncertainty estimation and how to account for it adequately. Indeed, all parameters of a NLMEM require a reliable computation method of their uncertainty, but it is actually even more important for the treatment effect as its standard error (SE) will be used to compute the test statistic leading to the clinical trial conclusion on the treatment efficacy.

For parameter estimation in NLMEM, frequentist approach through maximum likelihood estimation is commonly used, although the likelihood for these models has no analytical solution. The Stochastic Approximation Expectation-Maximisation (SAEM) algorithm[4] has been proven to be very efficient to obtain the Maximum Likelihood Estimator (MLE) for NLMEM. Various methods have been proposed for quantifying uncertainty in frequentist inference. The most widely used are based on the calculation of the Fisher Information Matrix (FIM). Indeed, according to the Cramér–Rao bound, the variance of any unbiased estimator is bounded by the inverse of the FIM. The MLE being efficient, this lower bound is reached when the sample size and

observations per subject tend to infinity. The FIM is generally calculated by linearising the model. This approximation has been extensively validated and used, in particular, for protocol optimisation[5]. However, at finite distance, the FIM can underestimate the SE of NLMEM[6]. This underestimation can notably result in an inflation of type I error when performing tests[7,8]. This is likely why the use of NLMEM is still limited at the pivotal decision level. Different methods have been explored not to use the FIM, such as the bootstrap[9,10] or likelihood profiling[11] while other methods have tried to correct the FIM[12] or resample from it as the Sampling Importance Resampling (SIR) method[13], but with limited success[14,15] and a high computational cost.

In a Bayesian framework, population parameters are considered as random variables with some prior distribution, and the aim is to characterise the posterior distribution from which the maximum a posteriori (MAP) is computed. Under some regularity conditions on the prior, the limit distributions of the MLE and the MAP estimator are equivalent (Bernstein-von Mises theorem[16]). This leads to the idea of using the standard deviation of the posterior distribution of population parameters as a proxy for the SE of the MLE, which we define as a semi-Bayesian approach. Ueckert et al. (2015)[17] proposed to do so to compute the SE of NLMEM parameters and Loingeville et al. (2020)[8] implemented this method. They proposed to run an Hamiltonian Monte-Carlo (HMC) algorithm in Stan[18]software following the frequentist inference in order to obtain a posterior distribution from which they derived SE.

Recently, Guhl et al. (2024)[19] proposed to obtain the posterior distribution using a Metropolis Hastings (MH) algorithm during the frequentist inference, within the SAEM algorithm. In the cited paper, the proposed method, called SAEM_MH, was evaluated and compared to other frequentist and semi-Bayesian SE computation methods. On a simulation study of sparse PK data, it showed improvements in terms of coverage of the simulated parameters compared to frequentist methods. However, the results differed from those obtained with the method based on HMC algorithm. Moreover, when applied to real data, the method was challenged by the high and correlated inter-individual variability (IIV) of the fitted model. We assume this is linked to the dimension of the parameter vector to estimate, which is an inherent limitation of the MH algorithm.

The aim of the present work is to explore variations that would overcome this limitation. We also explored the Approximate Bayesian Computation (ABC) method[20], which bypasses the evaluation of the likelihood function. The different methods and their variations were compared in a simulation study, with a scenario similar to that explored in Guhl et al.[19], an a second scenario presenting more complex features challenging the SAEM_MH method. Then, they were applied to a real case study investigating the relative bioavailability of two formulations of Gantenerumab, a monoclonal antibody developed by Roche and used for the treatment of Alzheimer's disease.

2 | METHODS

2.1 | Statistical model in NLMEM

We introduce here some notations in NLMEM, with the general form of the models we are considering as follows:

$$\begin{aligned} \forall 1 \leq i \leq N,\ 1 \leq j \leq n_i \quad & y_{ij} = f(t_{ij}, \psi_i) + g(t_{ij}, \psi_i, \xi)\varepsilon_{ij} \\ & \phi_i = h(\psi_i) = h(\bar{\psi}_i) + \eta_i \end{aligned} \tag{1}$$

N denotes the total number of subjects, n_i the number of observations for the i-th subject, $y_{ij} \in \mathbb{R}$ the j-th observation for the i-th subject, f the structural model, $t_{ij} \in \mathbb{R}$ the j-th observation time for the i-th subject, $\psi_i = (\psi_{i,\ell},\ 1 \leq \ell \leq n_\psi) \in \mathbb{R}^{n_\psi}$ the vector of unobserved individual parameters for the i-th subject, with n_ψ the number of parameters of the structural model, g the residual error model depending on parameters ξ (e.g. if the error model is proportional, $g(.) = b \times f(.)$ and $\xi = b$) and $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ the residual error.

$h()$ refers to a transformation function (.g. identity, log, logit, etc) used to obtain individual parameters ϕ_i on a different scale.

$\bar{\psi}_i$ refers to the typical value of ψ_i, which depends on μ and β, unknown vectors of fixed effects and associated fixed covariate effects, and some known individual covariates. η_i is the IIV with $\eta \sim \mathcal{N}(0, \Omega)$ where Ω is the variance covariance matrix of the inter-individual random effects.

In the following, $\theta = (\mu, \beta, \text{vec}(\Omega), \xi)$ denotes the P-vector of the model population parameters, where $\text{vec}(\Omega)$ is the vector of the components of the variance-covariance matrix Ω.

In this article, the estimation of θ is done by maximising the likelihood of the model, using the SAEM algorithm, to obtain the MLE. The SAEM algorithm[21] is an iterative algorithm consisting in K_1 exploratory iterations followed by K_2 smoothing iterations of a procedure based on the likelihood of the completed data conditionally to the current parameter. It is a stochastic

approximation because it samples individual parameters at each iteration via a Monte Carlo Markov Chain (MCMC) algorithm, namely Metropolis Hastings (MH)[22], to compute the conditional likelihood.

2.2 | SE calculation in NLMEM

In the following various methods for estimating uncertainty on the MLE are presented. Of note, they are not used here to obtain point estimates.

2.2.1 | Frequentist methods

Fisher Information Matrix[23]

To compute the SE of a population parameter vector θ, the classical method in NLMEM is to use the FIM. Indeed, denoting $\theta^\star$ the true unknown value of θ, and $\widehat{\theta}$ the maximum likelihood estimate of θ, if the likelihood function L is sufficiently smooth, asymptotic theory for maximum-likelihood estimation holds:

$$\sqrt{N}(\widehat{\theta} - \theta^\star) \underset{\mathrm{N}\to\infty}{\to} \mathcal{N}(0, I(\theta^\star)^{-1}) \tag{2}$$

where $I(\theta^\star) = \mathbb{E}\left[\left(\partial_\theta \log L(y;\theta^\star)\right)\left(\partial_\theta \log L(y;\theta^\star)\right)^T\right]$ is the expected Fisher information matrix (FIM). Thus, an estimate of the asymptotic covariance of $\widehat{\theta}$ is the inverse of the Fisher information matrix $I(\widehat{\theta}) = \mathbb{E}\left[\left(\partial_\theta \log L(y;\widehat{\theta})\right)\left(\partial_\theta \log L(y;\widehat{\theta})\right)^T\right]$. Nevertheless, the FIM cannot be computed in closed form in NLMEM. An option is thus to approximate it by the FIM of a Gaussian model obtained by linearising the regression function f around the conditional expectation $\bar{\psi}$ of the individual Gaussian parameters, as given in equation (3). The FIM of this Gaussian model is a block matrix (no correlations between the estimated fixed effects and the estimated variances). The gradient of f is numerically computed.

$$log L(y|\theta,\psi) \simeq -\frac{N}{2} log(2\pi) - \sum_{i=1}^{N} log(det\ \Gamma_i) - \frac{1}{2}\sum_{i=1}^{N}(y_i - Y_i)^t\ \Gamma_i^{-1}\ (y_i - Y_i) \tag{3}$$

with

$$\begin{aligned} Y_i &= f(t_i,\bar{\psi}_i) + \partial_\psi f(t_i,\psi_i)(\theta - \bar{\psi}_i) \\ \Gamma_i &= \partial_\psi f(t_i,\bar{\psi}_i)\ \Omega\ \partial_\psi f(t_i,\bar{\psi}_i)^t + g(t_i,\bar{\psi}_i,\xi)\ I\ g(t_i,\bar{\psi}_i,\xi)^t \end{aligned} \tag{4}$$

Once a FIM estimator has been computed, the estimate of the SE for the p^{th} component of θ, denoted θ_p, can be used to derive a confidence interval (CI) through the normal approximation, giving standard α-level CI. Formula is as follows:

$$\theta_p \in \left[\hat{\theta}_p - z_{\alpha/2}\widehat{SE}(\hat{\theta}_p); \hat{\theta}_p + z_{1-\alpha/2}\widehat{SE}(\hat{\theta}_p)\right] \tag{5}$$

where z_q is the q-quantile of the normal distribution. In the following, this method is called Asympt.

Sampling Importance Resampling

The Sampling Importance Resampling (SIR)[24] aims to sample in the parameter distribution. This noniterative and universally applicable method of obtaining samples from an unknown distribution is based on draws from an approximation of this distribution.

First, a high number M_S of parameter vectors $\hat{\theta}_{SIR,s}$, $s = 1, ..., M_S$ are sampled from a kernel distribution $p(\theta)$, and for each of them, the importance ratio is computed. This quantity measures the agreement between the approximated distribution and the data and is given in the following equation

$$\mathrm{IR}_s = \frac{L(y|\hat{\theta}_{\mathrm{SIR,s}})/L(y|\theta_{\mathrm{MLE}})}{p(\hat{\theta}_{\mathrm{SIR,s}})/p(\theta_{\mathrm{MLE}})} \tag{6}$$

Thereafter, these IR are normalised and used as probability to resample m_S vectors among the M_S. The standard deviation of the distribution obtained can be used as a proxy for the SE of the MLE, and the percentiles can also be used to construct confidence intervals.

Case bootstrap

In case bootstrap, M resampled datasets are built by resampling the vectors of individual observations with replacement among the original data and the whole estimation procedure is redone for each of these datasets. The uncertainty of the MLE is thereafter computed over these M estimations.

2.2.2 | Semi-Bayesian methods

Here we consider hybrid approaches where MLE is used as the point estimate, and its SE is approximated by the standard deviation of its posterior distribution drawn.

The HMC algorithm implemented in the Stan software allows to obtain this distribution, using the MLE obtained with SAEM as initial values and semi-informative prior distributions.[8] This method was used as a benchmark and is called hereafter Post.

SAEM_MH

The SAEM_MH method[19] consists in sampling the population parameters using the MH algorithm implemented in the smoothing phase of the SAEM algorithm as described in algorithm 1. At each iteration, the current frequentist estimations are used as parameters of the kernel for the MH algorithm, and the last iteration of the chain is kept as a realisation of the posterior distribution. In this algorithm, acceptation rates are computed based on the densities of the kernel and prior distributions, and the likelihood of y conditional on ψ.

Algorithm 1. SAEM_MH

At iteration K_1+1, set a prior $p(.)$ on θ.

At every iteration $k = K_1 + 1, ..., K_1 + K_2$ of the SAEM algorithm:

1. Compute the linearised FIM at iteration k by drawing a z-sample of ψ from the conditional distribution $p\left(.|y;\theta_k\right)$ using MH algorithm in SAEM machinery and compute the linearised FIM around their mean
2. Metropolis-Hastings algorithm - For $m = 1, \dots, M$:
   (a) Draw $\theta^{(m)}$ from the kernel $q(.) \sim \mathcal{N}\left(\theta_k, FIM_k^{-1}\right)$
   (b) Draw a z-sample of individual parameters $\psi^{(m)} = \left(\psi^{(m,1)}, \dots, \psi^{(m,z)}\right)$ conditional to $\theta^{(m)}$ using SAEM machinery
   (c) Set $\bar{\psi}^{(m)} = \frac{1}{z}\sum_1^z \psi^{(m,.)}$
   (d) Compute the loglikelihood of $\theta^{(m)}$ conditional on $\bar{\psi}^{(m)}$ by linearisation
   (e) Compute the likelihood ratio

$$\alpha\left(\theta^{(m-1)}, \theta^{(m)}\right) = \min\left(1, \frac{p\left(\theta^{(m)}|y, \bar{\psi}^{(m)}\right) q\left(\theta^{(m-1)}\right)}{p\left(\theta^{(m-1)}|y, \bar{\psi}^{(m-1)}\right) q\left(\theta^{(m)}\right)}\right)$$

$$\min\left(1, \frac{L\left(y|\theta^{(m)}, \bar{\psi}^{(m)}\right) p\left(\theta^{(m)}\right) q\left(\theta^{(m-1)}\right)}{L\left(y|\theta^{(m-1)}, \bar{\psi}^{(m-1)}\right) p\left(\theta^{(m-1)}\right) q\left(\theta^{(m)}\right)}\right)$$

   (f) Set

$$\theta^{(m)} = \begin{cases} \theta^{(m)} & \text{with probability } \alpha\left(\theta^{(m-1)}, \theta^{(m)}\right) \\ \theta^{(m-1)} & \text{with probability } 1 - \alpha\left(\theta^{(m-1)}, \theta^{(m)}\right) \end{cases}$$

3. Set $\theta_k^{(MH)} = \theta^{(M)}$

Use the chain of samples $\left(\theta_k^{(MH)}\right)_{k=K_1+1,\dots,K_1+K_2}$ standard deviation to estimate the SE of the ME

Several variations around SAEM_MH method were explored as described in Table 1 to try to find a method more robust to data with high and/or correlated IIV. Briefly, we tried: (i) using Gibbs sampling i.e. sampling in the individual parameters ψ before sampling in the population pramater θ, (ii) adopting a random walk (RW) behavior i.e. centering the kernel on the previous sample instead of the current frequentist estimation, (iii) reducing the kernel dimension, which we suspect is challenging the

Number of MH chains	Kernel distribution	within Gibbs	Random walk	Adaptive kernel	Method name
K2 chains within SAEM	One block Gaussian	No	No	No	MH 5 (SAEM_MH)
	Two-block Gaussian	No	Yes	No	MH 20
				Garthwaite method	MH 21
One chain after SAEM	One-block Gaussian	No	No	No	MH 6
				Ad hoc method	MH 11
			Yes	No	MH 9
				Ad hoc method	MH 12
				Haario method	MH 13
				Garthwaite method	MH 14
	Two-block Gaussian	No	No	No	MH 7
			Yes	No	MH 10
		Yes	No	No	MH 15
			Yes	No	MH 16
				Ad hoc method	MH 18
				Garthwaite method	MH 19
	Univariate Gaussian	No	No	No	MH 8
		Yes			MH 17

TABLE 1 Variations of SAEM_MH explored

SAEM_MH method, by (a) sampling from the two blocks of the FIM (one for the fixed effects, one for the components of the random effects variance matrix and the error parameters) separately, which reduces the dimension without loss of information, (b) sampling univariately, conditionally on the rest of the vector θ which is more computationally intensive but reduces the problem to one dimension, (iv) adapting the kernel variance within the MH chain, whether by an ad hoc procedure based on the empirical covariance of the chain already drawn, or by methods found in the literature, namely the Haario method[25], a more sophisticated version of our ad hoc procedure, and the Garthwaite method[26] which targets a specific acceptation rate, (v) combining the variations mentioned beforehand, (vi) drawing one long chain at the end of SAEM instead of K_2 chains within SAEM to be more computationally efficient and save time during the evaluation of all these variations. More details for each of the variations are given in Appendix B.

Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is an algorithm to compute posterior distributions that has been introduced in the 80s[20]. One of its goals is to bypass the evaluation of the likelihood function by using data simulations and summary statistics. It is called "approximate" because the posterior distribution $p(\theta|y)$ is approximated by $p\left(\theta|d\left(S(\hat{y}), S(y)\right) \leq \varepsilon\right)$, where $S(.)$ denotes some predefined summary statistics, y the observed data, $\hat{y}$ simulated data, $d(.)$ a distance, and ε an acceptance threshold. Summary statistics are used because the more the dimension of the data increases, the more the probability of simulating data that are close to the observations decreases[27]. Therefore summary statistics are built in order to represent the maximum amount of information in the simplest possible form. If the latter were sufficient, the distributions $p\left(\theta|d\left(S(\hat{y}), S(y)\right) \leq \varepsilon\right)$ and $p\left(\theta|d\left(\hat{y}, y\right) \leq \varepsilon\right)$ would be the same.

The rejection-based ABC is given in algorithm 2. In our context of mixed effects, we need to add a step in which we compute individual parameters before simulating the observations. To that purpose, we use the MH algorithm embedded in the SAEM algorithm.

Algorithm 2. ABC

1. Compute summary statistics of the observed data $S(y)$
2. For $m = 1, \ldots, M$:
   (a) Draw $\theta^{(m)}$ from the kernel $q(.)$
   (b) Draw $\psi^{(m)}|y, \theta^{(m)}$ using MH

(c) Simulate $y^{(m)}|\psi^{(m)}, \theta^{(m)}$

(d) Compute summary statistics $S(y^{(m)})$

(e) Compute $d\left(S(y^{(m)}), S(y)\right)$

(f) If $d\left(S(y^{(m)}), S(y)\right) \leq \varepsilon$, accept $\theta^{(m)}$

The distance was chosen as the relative root mean square difference to the reference (observed) summary statistics, as follows:

$$d(S(\widehat{y}), S(y)) = \frac{1}{r}\sum_{s=1}^{r}\sqrt{\left(\frac{S(\widehat{y})_s - S(y)_s}{S(y)_s}\right)^2} \quad \text{where } r \text{ is } S(y)\text{'s dimension} \tag{7}$$

Rejection-sampling based ABC has the known limitation that only a small dimension summary statistics vector can be used because otherwise, either the acceptance rate is very low, or the tolerance ε must be very high[28]. Methods have been proposed to reduce the dimension of the summary statistics[29]. One of these methods is to build statistics semi-automatically[30]. In this method, the summary statistics vector has the same dimension P as the vector of parameter to be estimated; and those statistics are chosen "automatically" through a linear regression that weighs the different candidates. The procedure is given in algorithm 3.

Algorithm 3. Semi-automatic ABC

1. First define some initial summary statistics $h(y)$ (here individual predictions and their squares).
2. Run the ABC given in algorithm 2 with these summary statistics (the number of ABC iterations M has to be at least two times the dimension of $h() + 1$)
3. Fit a linear regression model linking the posterior samples obtained $\theta^1, \ldots, \theta^{M_{ss}}$ and the associated simulated data $y^1, \ldots, y^{M_{ss}}$. For the p^{th} parameter: $\theta_p = \beta_{0,p} + \beta_p h(y) + \varepsilon_p$
   where $\beta_{0,p}, \varepsilon_p \in \mathbb{R}$, $\beta_p, h(y) \in \mathbb{R}^r$.
4. Because only differences in summary statistics are evaluated, the constant term can be neglected, and the p^{th} summary statistics is $\widehat{\beta}_p h(y)$. The semi-automatically built summary statistics is thus $S(y) = \left(\widehat{\beta}_1 h(y), \ldots, \widehat{\beta}_P h(y)\right)$.
5. Run ABC again with the semi-automatically built summary statistics.

As for SAEM_MH, several variations of the ABC method were explored and are summarised in table 2 . We considered i) different kernel distributions (a one-block multivariate Gaussian distribution or a three-block multivariate Gaussian distribution for the fixed effects, the random effects and the error parameters), ii) inflation of the kernel variance ($\times$ 1 or 2), iii) different summary statistics ($S(y) = y$ (i.e. no summary) and $S(y) = \left(y, y^2\right)$ or the semi-automatic ABC with $h(y) = \left(y, y^2\right)$) and iv) combinations of these variations. Details on the variations and their comparison are given in Appendix 2.

3 | SIMULATION STUDY

3.1 | Settings

3.1.1 | Model

We performed a simulation study using a one-compartment PK model with linear absorption and elimination, previously used in a published simulation study focusing on alternative SE calculations for model-based bioequivalence tests[8]. The equation for the time course of concentrations is:

$$C(t) = \frac{D}{V}\frac{ka}{\frac{Cl}{V} - ka}\left(\exp(-ka \times t) - \exp(-\frac{Cl}{V}t)\right) \tag{8}$$

where ka (h^{-1}) is the absorption rate constant, Cl ($L.h^{-1}$) the clearance, V (L) the volume of distribution and D (mg) the drug dose. Here we set $D = 4$ mg. Simulation values were $\mu_{ka} = 1.5\ h^{-1}$, $\mu_{Cl} = 0.04\ L.h^{-1}$, $\mu_V = 0.5\ L$ for the fixed effects, $\beta_{ka} = 0$, $\beta_{Cl} = log(1.25)$ and $\beta_V = log(1.25)$ for the associated treatment effects (with a log-normal distribution on all parameters, so that e.g $ka_i = \mu_{ka}\ e^{\beta^T * T_i}\ e^{\eta_{ka,i}}$), and the variance matrix of the random effects varies amongst different scenarios (see below). A proportional error model with $b = 0.1$ was used.

Kernel distribution	Summary statistics	Variance inflation	Method name
One-block Gaussian	$S(y) = (y)$	None	ABC 2
		2	ABC 3
	$S(y) = \left(y, y^2\right)$	None	ABC 6
		2	ABC 7
	$h(y) = \left(y, y^2\right)$	None	ABC 10
		2	ABC 11
Three-block Gaussian	$S(y) = (y)$	None	ABC 4
		2	ABC 5
	$S(y) = \left(y, y^2\right)$	None	ABC 8
		2	ABC 9
	$h(y) = \left(y, y^2\right)$	None	ABC 12
		2	ABC 13

TABLE 2 Variations of ABC explored
With $S(y) = y$ the individual predictions are used directly as summary statistics, with $S(y) = \left(y, y^2\right)$ the individual predictions and their squares are used as summary statistics, with $h(y) = \left(y, y^2\right)$ we use semi-automatic summary statistics obtained from a regression on these quantities.

3.1.2 | Scenarios

First, a rich design was considered corresponding to an asymptotic setting, as a preliminary verification that all methods behave the same in that case, with $N = 150$ patients (75 in each treatment arm) and $n = 10$ sampling times ($t \in \{0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12, 24\}$ hours).

Then, a sparse design was simulated to evaluate the methods at finite distance, with N = 12 patients (6 in each treatment arm) and n = 3 sampling times ($t \in \{0.25, 3.5, 24\}$ hours). The latter were chosen using the R package PFIM[31] for optimal design among $n = 10$ sampling times ($t \in \{0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12, 24\}$ hours) and not accounting for correlations in the random effects as it is not possible to implement it in PFIM.

Two sparse scenarios were implemented: as in Guhl et al., the first scenario presented no correlations between the random effects (Ω_1 in Table 3), while in the second we simulated correlations ($\rho_{ka,Cl} = 0.9$, $\rho_{ka,V} = 0.98$, $\rho_{V,Cl} = 0.9$) and high IIV (Ω_2 in Table 3).

$$\Omega_1 = \begin{pmatrix} 0.0500 & 0.0000 & 0.0000 \\ 0.0000 & 0.0125 & 0.0000 \\ 0.0000 & 0.0000 & 0.0500 \end{pmatrix} \quad \Omega_2 = \begin{pmatrix} 1.2000 & 1.0800 & 1.1760 \\ 1.0800 & 1.2000 & 1.0800 \\ 1.1760 & 1.0800 & 1.2000 \end{pmatrix}$$

TABLE 3 Simulation values for the Ω matrix in the two sparse scenarios

3.2 | Evaluation

For each scenario, 500 datasets were simulated and parameters were estimated using the SAEM algorithm. Thereafter, the different SE calculation methods were applied to each dataset.

The first evaluation criteria was the 95% coverage rate (CR), defined as the proportion of datasets for which the simulated value of the parameter is within the parameter estimate 95% confidence or credible interval. Confidence interval on the CR was computed using the Clopper–Pearson for binomial confidence intervals[32].

Estimated RSE, computed as the standard deviation in the posterior samples divided by its mean and multiplied by 100, were also compared between the different methods and to the empirical RSE. The latter was considered our target and obtained as the RSE of the parameter estimates on the 500 datasets.

Also, based on these estimates, the empirical distribution for the MLE was estimated using the R function *density()* from the package *stats*. A complementary evaluation criteria was considered as the match between this empirical distribution and the posterior density estimated with the different methods. This comparison was first made by visual comparison, for a few data sets, then according to the Kullback–Leibler divergence of the posterior from the empirical distribution.

For MH based methods, the acceptance rates were compared to the recommended value of 0.234 [33].

Run times were evaluated by the median based on all simulated datasets.

3.3 | Implementation

Parameter estimation was performed using the R package `saemix` [34]. The initial values for the parameter estimation were the simulated values for fixed effects, 0 for treatment effects and the diagonal identity matrix for the IIV matrix. We ran 3 chains with $K_1 = 300$ and $K_2 = 100$.

Minor changes have been made to the existing implementation of the SIR algorithm in *saemix* to make it compatible with the latest version of the package and to apply it to models with parameters of complex variability structure. The code is available on github (https://github.com/saemixdevelopment/saemixextension/tree/master/SIR_Lucie). The proposal distribution was a multivariate normal distribution centered on the MLE with the variance being the inverse of the FIM. To be consistent with the sampling algorithm, the likelihood used in the algorithm is obtained by importance sampling instead of linearisation. M_S=1000 samples were drawn and m_S=500 resampled.

For the case bootstrap, the function *saemix.bootstrap* from the R package `saemix` was used and $M = 500$ datasets were resampled.

For Post, 3 chains of 1500 iterations (including 500 warm up iterations not used to compute standard deviations) were run. The initial values were the estimations obtained with `saemix` and the prior distributions were gaussian, centered on the simulated values of θ with 30% variation coefficient for fixed effects and 50% for treatment effects, IIV and residual error parameters. Results were kept only if the highest $\hat{R}$ was lower than 1.05.

For SAEM_MH, the prior was set to a Gaussian distribution centered on the simulated values of θ with 30% variation coefficient for fixed effects and 50% for treatment effects, IIV and residual error parameters. As $K_2 = 100$, 100 MH chains of length M=100 were run. The number of ψ drawn to compute the different likelihoods and FIM matrices was set to $z = 50$. We used the FIM obtained by linearisation. Based on previous work [19], the kernel distribution used was $\text{infl} \times FIM_k^{-1}$ with $infl = 2$ at each iteration k of the SAEM algorithm. For all variations using only one chain after SAEM, the chain length was set to $M = 1000$.

For ABC algorithm, $M = 1000$ samples were drawn from the kernel, which was a multivariate Gaussian distribution as in SAEM_MH and its variations. The tolerance ϵ on distance for sample acceptation was chosen among $0.5, 1, 2, 5, 10, 15, 30$ to maximise the number of parameters properly covered.

3.4 | Results

3.4.1 | Rich design

On the rich design, all the methods performed similarly and satisfactorily (see Figures E7 and E8 in Appendix E which display the performance of a subset of the methods).

3.4.2 | Sparse design

On the sparse design, the Post method obtained $\hat{R} \geq 1.05$ for at least one model parameter on 137 (27.4%) and 169 (33.8%) data sets simulated respectively without and with correlation. Therefore, the evaluation was performed only on the remaining datasets i.e. 363 and 331 datasets respectively, on which all the methods ran successfully.

In the following figures we focused on a subset of the SAEM_MH and ABC method variations selected according to their performance on the first scenario. For the SAEM_MH algorithm, we selected the variation that draws one chain after the SAEM algorithm, using a two-block Gaussian kernel with covariance matrix set to the inverse of the FIM, inflated by a factor 2, and a random walk. This method is thereafter referred as MH_after_SAEM and corresponds to MH10 in table 1 . For the ABC method, we selected the variation with a three-block Gaussian kernel and a variance inflation of 2, a regression on individual simulated observations and their squares as summary statistics. This method is thereafter referred as ABC and corresponds to ABC 13 in table 2 .

In the following, these SAEM_MH and ABC variations are compared to Asympt, SIR, bootstrap, Post and SAEM_MH. However, in Appendix C and D respectively, figures can be found displaying the performances of all variations of SAEM_MH and all variations of ABC on a subset of 100 data sets.

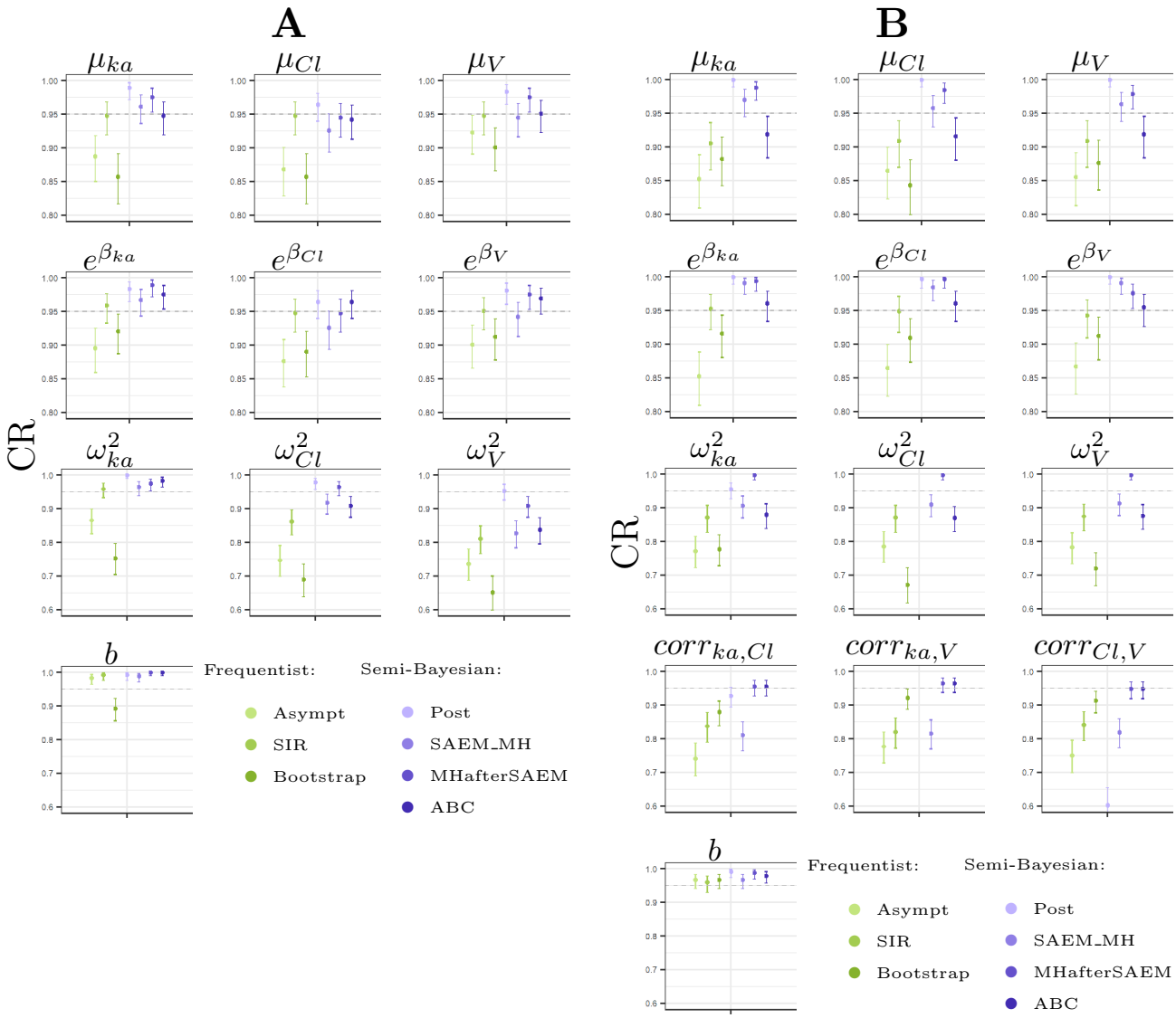


FIGURE 1 95% coverage rates and their 95% confidence interval
A: Scenario without correlations
B: Scenario with correlations - Some coverage rate estimates are outside the plot for Post: ω^2_{Cl}: CR=0.18 [0.14; 0.23], ω^2_V: CR=0.00 [0; 0.01], $\rho_{ka,V}$: CR=0.38 [0.33; 0.43] and $\rho_{Cl,V}$: CR=0.6 [0.55; 0.65]
The target CR at 0.95 is shown as an horizontal line.

3.4.3 | Scenario without correlations

On this scenario, SAEM_MH has an average acceptation rate of 0.16 (sd = 0.03) which is below the recommended 0.234 versus 0.33 (sd = 0.05) for MH_after_SAEM.

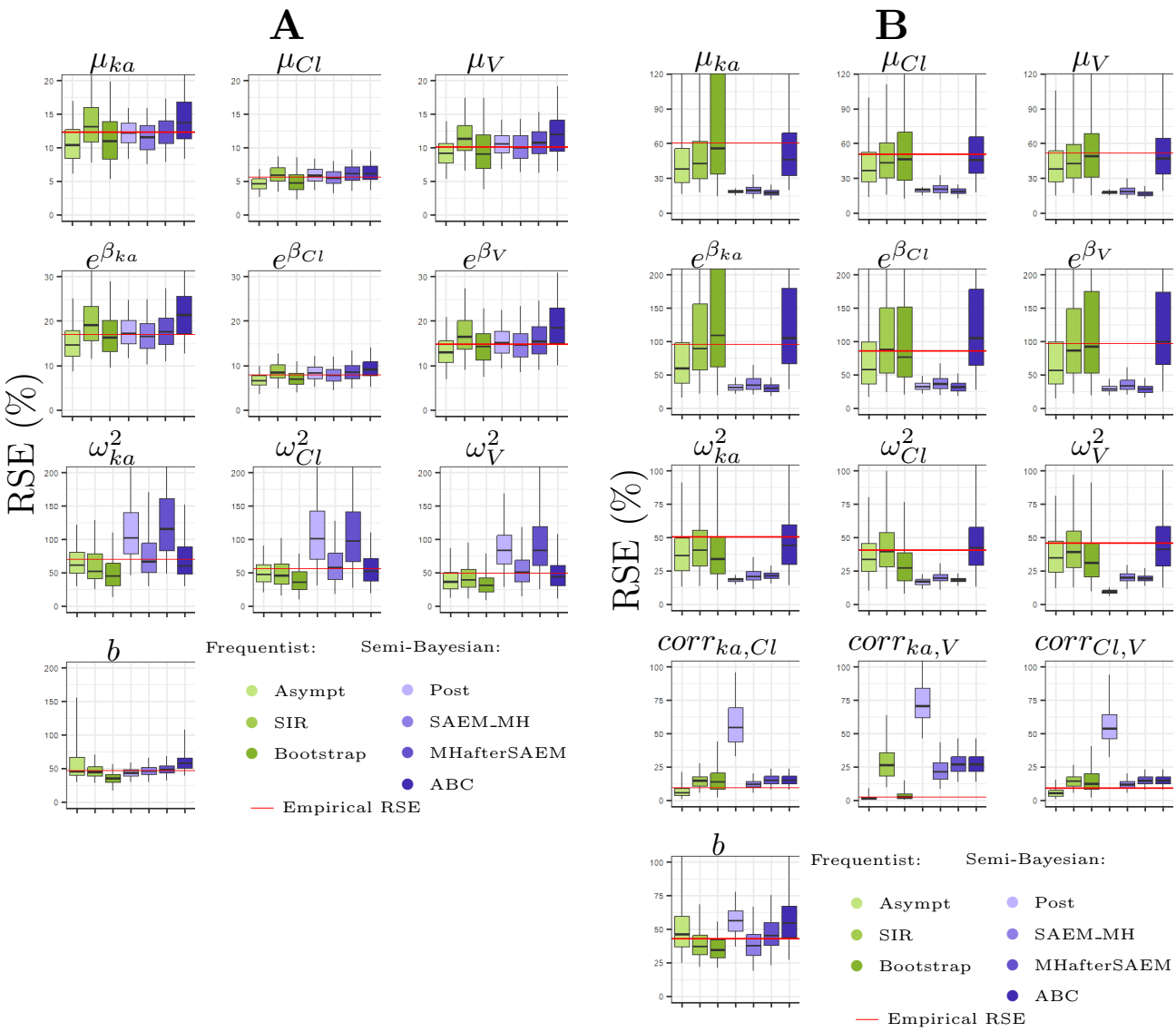


FIGURE 2 Estimated Relative Standard Error
A: Scenario without correlations
B: Scenario with correlations
Each panel represents one model parameter. Our target, the empirical SE from SAEM estimates is shown as an horizontal red line, and its 95% confidence interval as dotted red lines. The box represent the interval between the 25% and 75% quartiles with a mark at the median; while the whiskers represent the interval between the 2.5% and 97.5% quantiles.

Figure 1 A shows the coverage rates (CR) obtained and their 95% confidence interval for the different model parameters and methods to be compared to a target CR at 0.95. For the fixed effects and the treatment effects, both Asympt (●) and bootstrap (●) gave CR below the target. CR were generally above the target for Post (●), as well as for MH_after_SAEM (●) to a lesser extent. These parameters are properly covered with SIR (●), SAEM_MH (●) and ABC (●). For the variance matrix, all frequentist methods (Asympt, SIR and bootstrap) gave overall CR below the target. Of note, Boostrap was also biased on these parameters (not shown), which contributes to explain the very low CR obtained. Post CR are still above the target, while SAEM_MH and ABC undercovered these parameters. MH_after_SAEM gave the most satisfying results, though it still undercovered one

variance parameter. The error parameter is overcovered with all methods except bootstrap, for which this parameter estimation was also biased.

Figure 2 A shows the estimated relative standard errors. The box represent the the 25%, 50% and and 75% quantiles while the whiskers represent the 2.5% and 97.5% quantiles. The target, being the empirical RSE from SAEM estimates, is shown as an horizontal red line, and its 95% confidence interval as dotted red lines. On the fixed and treatment effects, Asympt and Bootstrap gave underestimated RSE, while SIR slightly overestimated them. Post, SAEM_MH and MH_after_SAEM gave RSE close to the target. ABC gave slighlty overestimated RSE.

On the variance parameters, all frequentist methods were slighlty below the target, which for bootstrap is emphasised by the bias obtained on the point estimates. Post and MH_after_SAEM gave overestimated RSE while SAEM_MH and ABC gave satisfying results.

On the error parameter, Asympt showed a high variability in RSE estimates, though the median was on target. SIR, SAEM_MH and MH_after_SAEM gave controlled RSE. Bootstrap and, to a lesser extent, Post underestimated the RSE, while ABC overestimated it.

SIR took a median time of 35 minutes on this scenario, while bootstrap took a median time of 45 minutes. Post was the quickest method, taking on average less than 3 minutes, followed by MH_after_SAEM (15 minutes) and ABC (20 minutes). SAEM_MH was the most computation intensive method, taking on average 70 minutes to run on one dataset with a high variability across datasets.

3.4.4 | Scenario with correlations

On this scenario, SAEM_MH had very low AR (0.04 (sd = 0.01)), mechanically leading to posterior distributions with small standard deviations. MH_after_SAEM had higher AR (0.14 (sd = 0.06)).

CR for the scenario with high IIV and correlations are given in Figure 1 B. For the fixed effects, all frequentist methods gave CR below the target, as well as ABC. Post and MH_after_SAEM overcovered these parameters. The most satisfying results were obtained with SAEM_MH. For the treatment effects, Asympt and bootstrap once again systemically undercovered the parameters, while SIR and ABC gave satisfying CR. Post, SAEM_MH and MH_after_SAEM gave CR above the target. For the variance and correlation parameters, all frequentist methods were below the target. Post and SAEM_MH were especially challenged and gave highly underestimated CR. MH_after_SAEM gave overcovered variances and well covered correlations, while ABC gave undercovered variances and well covered correlations. The error parameter was well covered with all frequentist methods and SAEM_MH but was overcovered with all the other semi-Bayesian methods.

Estimated RSE for the scenario with high IIV and correlations are given in Figure 2 B. For the fixed effects, all frequentist methods gave controlled RSE, as well as ABC, while all other semi-Bayesian methods gave highly underestimated RSE. Of note, the empirical RSE are between 60 and 100% for these parameters. We displayed the relative SE for the exponent of the treatment effects as $\beta_{ka} = 0$. For the variance parameters, all frequentist methods underestimated the RSE, but not as much as Post, SAEM_MH and MH_after_SAEM. ABC was the only semi-Bayesian method to give satisfying results. For the correlation parameters, Asympt gave slightly underestimated RSE, while SIR, bootstrap and all semi-Bayesian methods gave overestimated RSE with Post far above the others. For the error parameter, Asympt gave unbiased but imprecise RSE. SIR, bootstrap and SAEM_MH gave underestimated RSE. ABC gave overestimated RSE. Overall, only MH_after_SAEM gave satisfying RSE for this parameter.

In this scenario, semi-Bayesian methods, except ABC, gave good CR associated with very small RSE. Indeed, the likelihood-based acceptation procedure does not accept samples far from the MLE. As a consequence, the posterior distribution is narrow around the MLE, the latter being close to the simulated value as these methods are unbiased. Therefore, the credible interval computed based on the quantiles of the posterior distribution almost always contains the simulated value so the coverage is close to 1. Plots of the posterior distributions are given in Appendix F, and illustrate that phenomenon on one dataset (Figure F11). This information is summarised over the 500 datasets by the Kullback-Leibler (KL) divergence, in Figure F9 of Appendix F. According to the KL divergence, for all the parameters, ABC most often gave the closest distribution to the empirical distribution.

On this scenario, SIR was the most computation intensive method, with a median time of 110 minutes, and was also the method with the highest variability in run times. Post was still the fastest method with run times below 5 minutes on average, followed by ABC (15 minutes) and MH_after_SAEM (25 minutes). SAEM_MH was slower, with an average time of 90 minutes, almost as long as bootstrap with 100 minutes.

4 | APPLICATION TO GANTENERUMAB DATA

4.1 | Methods

We applied Asympt, SIR, bootstrap, Post, SAEM_MH, MH_after_SAEM and ABC on PK data from a phase I parallel clinical trial, conducted by Roche, comparing two formulations of Gantenerumab, a monoclonal antibody used in the treatment of Alzheimer's disease. The data design consists of two arms of N=24 subjects with 10 observations per subject (at 0.25, 1, 2, 3, 4, 7, 13, 20, 42, 63, and 84 days). The dataset and model selection process have been described elsewhere[35]: the model selected to estimate the treatment effect on the PK of the drug is a two-compartment model with a delayed first order absorption. Treatment effects are estimated on all PK parameters with four random effects variances estimated in the Ω matrix (varying from 26% to 112%) as well as a correlation of 0.76 between two random effects. The residual error mode was proportional.

For the `saemix` fit, 3 chains of $K_1 = 500$ exploratory iterations and $K_2 = 300$ smoothing iterations were run, with initial values being $\mu_{ka} = 0.5\ d^{-1}$; $\mu_{CL/F} = 0.6\ L.d^{-1}$, $\mu_{V_1/F=15\ L}$, $\mu_{Q/F} = 0.5\ L.d^{-1}$, $\mu_{V_2/F} = 5\ L$, $\mu_{Tlag} = 0.05\ d$, 0 for the treatment effects, a diagonal matrix of 1 for variance covariance, and $b = 1$. For SIR, $M_S = 1000$ samples and $m_S = 500$ resamples were drawn. For Bootstrap, 500 datasets were resampled. For Post, we used normal prior distributions on all parameters. For the fixed effects, it was centered on the `saemix` estimation with a coefficient of variation of 30%. For the treatment effects, it was centered on 0 with standard deviation of 0.5. For the variance matrix, it was centered on a diagonal matrix of 1 with standard deviation of 0.5 and for the residual error parameter it was centered on 1 with standard deviation of 0.5. The initial values were the `saemix` estimations. 500 warm up iterations and 1000 sampling iterations were drawn. We used the same prior distributions for SAEM_MH and $M = 100$ iterations were run in each of the $K_2 = 300$ MH chains. For ABC, 1000 iterations were run and an inflation factor of 2 was used. The summary statistics and distance function used were the same as in the simulation study. The threshold ϵ was set to 10.

We also analysed a sparse subset of 12 random patients (6 per arm). On the latter, the settings of some methods were slightly adapted, i.e. `saemix` was run with 10 chains to ensure convergence and in MH_after_SAEM, we inflated the kernel variance by a factor 2.

4.2 | Results

Figure G1 in Appendix G shows the parameter estimates obtained on the full data and the sparse subset. As in the simulation study with correlations, this model presents high IIV as well as correlations between the random effects.

Figure 3 shows the different RSE obtained on radar plots. The fixed effects μ and treatment effects β^T are represented separately, with a third graph presenting the components of Ω and b, for both settings (full and sparse subset of the data). The corresponding values are given in Tables G2 and G3 in Appendix G.

On the full dataset, RSE were in general concordance. For fixed effects as well as variance matrix and error parameters, they were similar across methods. All treatment effects were non significant (RSE>50%) which explains the higher variability observed across methods. Of note, SAEM_MH acceptation rates were at a median of 9%. We explored an inflation factor on the kernel variance of 1.5 or 2, which lowered the AR to 7% and 5% respectively but did not change the results in terms of RSE (not shown). With MH_after_SAEM, AR were at 8% for fixed effects and 29% for random effects. The AR of the ABC method was at 26%.

On the sparse subset of the data, Asympt and SIR gave similar results, whereas for some parameters, bootstrap gave much higher RSE, Post and SAEM_MH lower RSE and ABC moderately higher RSE. SAEM_MH had AR at a median of 3%, and even lower when inflating the kernel variance (not shown). MH_after_SAEM had AR very close to 0 for fixed effects making it impossible to compute reliable RSE and at 19% for variances which is more reasonable. ABC had an AR at 25%.

The concordant results on the full data suggest we reached an asymptotic setting, therefore the Asympt method, which is the fastest, could be used or any of the other frequentist or semi-Bayesian methods. On the sparse subset of the data, however, discordances between methods suggest we were far from the asymptotic bound and Asympt should not be trusted, nor should the SIR method which gives similar results. Bootstrap behaved erratically, and methods based on the MH algorithm gave low AR indicating they were not reliable either. According to our simulation study, ABC, which had an acceptable AR, might be the most reliable method in this example.

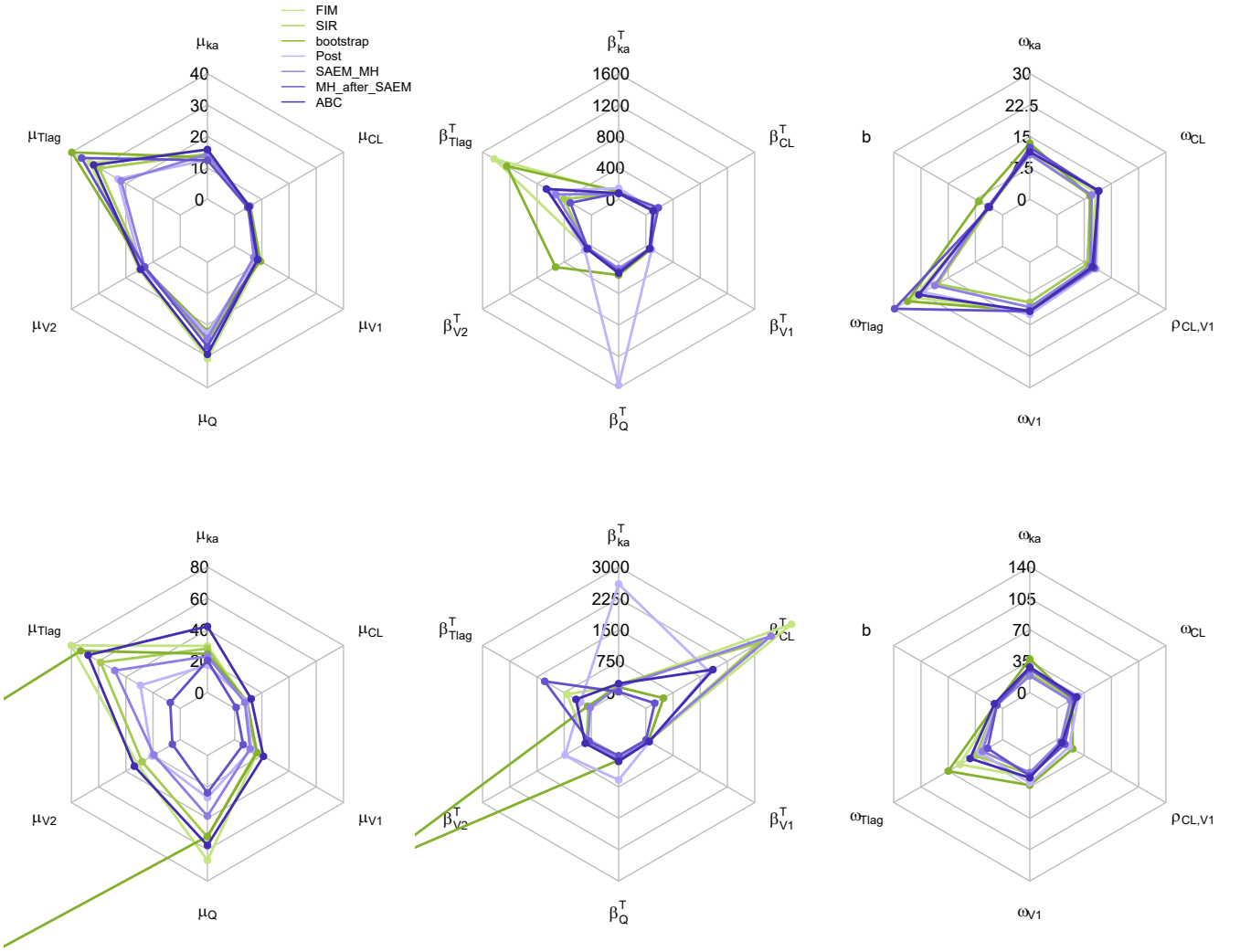


FIGURE 3 Radar plots of relative standard errors computed for the parameters of the model fitted on the full (top) and sparse (bottom) dataset with Asympt (●), SIR (●), bootstrap (●), Post (●), SAEM_MH (●), MH_after_SAEM (●) and ABC (●) methods

5 | DISCUSSION

Our objective in this work was to propose a method to assess the uncertainty of parameter estimates of NLMEM in sparse design studies. We built different variations around the MH and ABC algorithms to sample from the posterior distribution of the population parameters using as prior information the estimates obtained with the SAEM algorithm. These methods were evaluated in a simulation study using a simple PK model, on two sparse design scenarios. The methods were also evaluated on a real clinical trial considering the full dataset and a sparse subset. They were compared to several methods including the asymptotic method using the FIM, SIR, and bootstrap in terms of CR and SE.

As expected at finite distance, Asympt underestimated the uncertainty[6,7]. Case bootstrap showed poor performances and was moreover biased on the variances. Those results were consistent with previous work[15] showing that with very few patients bootstrap fails to estimate the uncertainty correctly. SIR showed better performances but still failed for some parameters. It should also be noted that the SIR method performed better with an inflation factor of 2, despite the fact that the recommended diagnostic graphs based on differences in the objective function[13] suggested that inflation was not appropriate. This last point is therefore a limiting factor for the performance of the SIR method, since it does not follow the usual diagnostics.

The Post method comes with diagnostic tools that alerted us on the lack of convergence of the HMC algorithm on some datasets of the simulation study, therefore we excluded them from analysis. However, we checked that the overall results on CR and RSE were the same for all other methods when using all datasets, excluding the risk of dataset selection bias. This method was challenged by the second scenario, as well as SAEM_MH. Both appeared not to be robust to high IIV and correlations. MH_after_SAEM ran faster than SAEM_MH and overcame some of its limitations on the complex scenario but was still challenged by high IIV and correlations. The selected ABC method showed promising results. Indeed, on both scenarios, most of the parameters were properly covered with RSE well or over estimated, which is preferable to an under-estimation of RSE as it avoids type I error inflation or over confident predictions. Notably, for the treatment effect coefficients leading pivotal decisions, it can be noted that the SIR and ABC methods were the only methods robust to sparse design with high IIV and correlations. The latter should therefore be preferred.

We would like to stress how the sparse scenario with high IIV and correlation represent a very difficult estimation framework as evidenced by the very high empirical RSE. Such high IIV and correlation in random effects are realistic and indeed in our application study we estimated variances of this magnitude and a correlation between random effects. Nevertheless, in combination with the chosen study design, it led to a very challenging scenario, with only 36 points per dataset to inform 13 population parameters with complex correlations. Therefore, it is necessary to qualify the argument according to which some of the methods over or underestimate the SE since the order of magnitude of the RSE, for example on the treatment effects, is about one hundred.

On the data from the Gantenerumab clinical trial, similarly, all treatment effect estimates had very large RSE, not allowing for a fine comparison between methods. On the full data, results indicated that the asymptotic assumption was verified, whereas on the sparse subset of the data, results differed across methods and ABC seemed to be the preferable method.

Of note, all the methods relied on the observed data but Asympt which used the expected FIM estimator, computed by linearisation. We compared Asympt to a method using an observed FIM estimator computed with a stochastic approximation (implementation within the *saemix* package available on github https://github.com/saemixdevelopment/saemixextension/tree/master/FIM_stochastique). Detailed methods and results are shown in Appendix A. Briefly, Asympt gave similar SE than the method using an observed FIM estimator at the asymptotic, but faced computational limits at finite distance.

SIR and ABC gave the most satisfying results in this work. However, we note that when they use the FIM inflated by a factor 2 in the proposal distribution, they have RSE close to Asympt RSE $\times\sqrt{2}$, so these methods may not be sensitive enough to select draws that belong to the true MLE distribution. More work is required to understand the good performance of these two methods and a number of challenges need to be overcome before the use of ABC can be widely adopted.

Overall, ABC is a very vague term for a set of methods, and it is necessary to consider its use on a case-by-case basis. For instance, the choice of summary statistics is one of the thorniest issues of the ABC. Here we explored a regression based methodology to reduce the dimension of summary statistics which requires a first pilot run of the ABC and this choice can be questioned. In addition, we tested only rejection-sampling based ABC, in which all accepted vectors are treated the same regardless of their distance to the MLE summary statistics, which can damage the approximation if the choice for the tolerance is not restrictive enough. To bypass this issue, smooth weighting methods associated with local linear regression have been proposed[28], or automatic sampling schemes using sequential Monte Carlo (SMC)[36].We could also have explored a kernel density estimation method for estimating the posterior distribution and used it as summary statistics[37]. Finally, in our implementation of the ABC method, the tolerance ϵ was chosen based on simulation study results which will not be possible in practice. However, methods that determine automatically sequence of tolerance levels could be explored[38,39].

6 | CONCLUSION

In this work, we investigated methods of uncertainty computation to handle sparse data design in NLMEM where using the FIM fails. SIR, bootstrap and Semi-Bayesian methods using MH and ABC algorithms were explored with several variations. MH algorithm remained challenged when working on data with high IIV and correlations while ABC algorithm and SIR showed good performances. Further work is needed to limit SE overestimation and improve computational cost.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

References

1. Jönsson S, Henningsson A, Edholm M, Salmonson T. Role of modelling and simulation: a European regulatory perspective. *Clinical Pharmacokinetics* 2012; 51: 69–76.

2. Jonsson EN, Sheiner LB. More efficient clinical trials through use of scientific model-based statistical tests. *Clinical Pharmacology & Therapeutics* 2002; 72(6): 603–614.

3. Desmée S, Mentré F, Veyrat-Follet C, Sébastien B, Guedj J. Nonlinear joint models for individual dynamic prediction of risk of death using Hamiltonian Monte Carlo: application to metastatic prostate cancer. *BMC Medical Research Methodology* 2017; 17(1): 1–12.

4. Lavielle M. *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press . 2014.

5. Nyberg J, Bazzoli C, Ogungbenro K, et al. Methods and software tools for design evaluation in population pharmacokinetics–pharmacodynamics studies. *British Journal of Clinical Pharmacology* 2015; 79(1): 6–17.

6. Bertrand J, Comets E, Laffont CM, Chenel M, Mentré F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. *Journal of Pharmacokinetics and Pharmacodynamics* 2009; 36: 317–339.

7. Dubois A, Lavielle M, Gsteiger S, Pigeolet E, Mentré F. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. *Statistics in Medicine* 2011; 30(21): 2582–2600.

8. Loingeville F, Bertrand J, Nguyen TT, et al. New Model–Based Bioequivalence Statistical Approaches for Pharmacokinetic Studies with Sparse Sampling. *The AAPS journal* 2020; 22: 1–8.

9. Efron B. Bootstrap methods: another look at the jacknife. *The Annals of Statistics* 1979; 7(1): 1–26.

10. Parke J, Holford NH, Charles BG. A procedure for generating bootstrap samples for the validation of nonlinear mixed-effects population models. *Computer Methods and Programs in Biomedicine* 1999; 59(1): 19–29.

11. Broeker A, Wicha SG. Assessing parameter uncertainty in small-n pharmacometric analyses: value of the log-likelihood profiling-based sampling importance resampling (LLP-SIR) technique. *Journal of Pharmacokinetics and Pharmacodynamics* 2020; 47(3): 219–228. doi: 10.1007/s10928-020-09682-4

12. Bertrand J, Comets E, Chenel M, Mentré F. Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. *Biometrics* 2012; 68(1): 146–155.

13. Dosne AG, Bergstrand M, Karlsson MO. An automated sampling importance resampling procedure for estimating parameter uncertainty. *Journal of Pharmacokinetics and Pharmacodynamics* 2017; 44: 509–520.

14. Thai HT, Mentré F, Holford NH, Veyrat-Follet C, Comets E. Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. *Journal of Pharmacokinetics and Pharmacodynamics* 2014; 41: 15–33.

15. Broeker A, Wicha SG. Assessing parameter uncertainty in small-n pharmacometric analyses: value of the log-likelihood profiling-based sampling importance resampling (LLP-SIR) technique. *Journal of Pharmacokinetics and Pharmacodynamics* 2020; 47: 219–228.

16. Vaart A. Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics). *Cambridge University Press* 1998.

17. Ueckert S, Rivière M, Mentré F. Alternative to Resampling Methods in Maximum Likelihood Estimation for NLMEMs by Borrowing from Bayesian Methodology. *PAGE Meeting* 2015.

18. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. *Journal of Statistical Software* 2017; 76(1).

19. Guhl M, Fayette L, Mercier F, Bertrand J, Comets E. Uncertainty computation at finite distance in nonlinear mixed effects models - a new method based on Metropolis Hastings algorithm. *The AAPS Journal* (in press).

20. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* 1984: 1151–1172.

21. Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* 1999; 27(1). doi: 10.1214/aos/1018031103

22. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 1953; 21(6): 1087–1092. doi: 10.1063/1.1699114

23. Fisher R. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 1922; 222(594-604): 309–368. doi: 10.1098/rsta.1922.0009

24. Rubin DB. Using the SIR algorithm to simulate posterior distribution. *Bayesian Statistics* 1988; 3: 395–402.

25. Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. *Bernoulli* 2001: 223–242.

26. Garthwaite PH, Fan Y, Sisson SA. Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Communications in Statistics-Theory and Methods* 2016; 45(17): 5098–5111.

27. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate bayesian computation. *PLoS computational biology* 2013; 9(1): e1002803.

28. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics* 2002; 162(4): 2025–2035.

29. Prangle D. Summary statistics. In: Chapman and Hall/CRC. 2018 (pp. 125–152).

30. Fearnhead P, Prangle D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2012; 74(3): 419–474.

31. Leroux R, Seurat J, Nagard H, Mentré F, group P. New features in PFIM for optimal design in nonlinear mixed effects model using the R Package PFIM5.0. *PAGE 30* 2022.

32. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; 26(4): 404–413.

33. Gelman A, Gilks WR, Roberts GO. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* 1997; 7(1): 110–120.

34. Comets E, Lavenu A, Lavielle M. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software* 2017; 80(3): 1–41. doi: 10.18637/jss.v080.i03

35. 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. *Journal of Pharmacokinetics and Pharmacodynamics* 2022; 49(5): 557–577. doi: 10.1007/s10928-022-09821-z

36. Bonassi FV, West M. Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *arXiv:1503.07791* 2015.

37. Irvine MA, Hollingsworth TD. Kernel-density estimation and approximate Bayesian computation for flexible epidemiological model fitting in Python. *Epidemics* 2018; 25: 80–88.

38. Beaumont MA, Cornuet JM, Marin JM, Robert CP. Adaptive approximate Bayesian computation. *Biometrika* 2009; 96(4): 983–990.

39. Lenormand M, Jabot F, Deffuant G. Adaptive approximate Bayesian computation for complex models. *Computational Statistics* 2013; 28(6): 2777–2796.

40. Riviere MK, Ueckert S, Mentré F. An MCMC method for the evaluation of the Fisher information matrix for non-linear mixed effect models. *Biostatistics* 2016; 17(4): 737–750.

41. Ueckert S, Mentré F. A new method for evaluation of the Fisher information matrix for discrete mixed effect models using Monte Carlo sampling and adaptive Gaussian quadrature. *Computational Statistics & Data Analysis* 2017; 111: 203–219.

42. Delattre M, Kuhn E. Estimating Fisher information matrix in latent variable models based on the score function. *arXiv preprint arXiv:1909.06094* 2019.

43. Lavalley-Morelle A, Mentré F, Comets E, Mullaert J. Extending the code in the open-source saemix package to fit joint models of longitudinal and time-to-event data. *Computer Methods and Programs in Biomedicine* 2024; 247. doi: 10.1016/j.cmpb.2024.108095

44. Gelman A, Roberts GO, Gilks WR. Efficient Metropolis jumping rules. *Bayesian Statistics* 1996; 5(599-608): 42.

How to cite this article: Fayette L., Bertrand J., Comets E. and Guhl M. (), Development of a semi-Bayesian SAEM algorithm for finite-distance estimation of parameter uncertainty in nonlinear mixed-effects models, *Stat Med*, .

APPENDIX

A APPENDIX: STOCHASTIC OBSERVED FISHER INFORMATION MATRIX

The linear approximation of the FIM has been extensively validated and used, in particular, for protocol optimisation[5].

Exact calculations, with a higher computational cost, have also been proposed recently, in particular for calculating estimation errors in discrete data models for which linearisation of the model is not adequate[40,41].

The FIM can also be estimated by the observed FIM and it has been shown[42] that the latter can be stochastically approximated. Previous work[43] has illustrated its effectiveness in the context of NLMEM. Therefore, we generalized the implementation of a stochastic approximation for FIM computation in the *saemix* package for exploration.

A.1 Method

The (expected) Fisher information is defined as the variance of the score as given by equation (A1).

$$I(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ln l(Y;\theta)}{\partial \theta} \frac{\partial \ln l(Y;\theta)}{\partial \theta^T} \right] \tag{A1}$$

$\mathbb{E}_\theta(x)$ denotes the expectation of x with respect to its probability distribution indexed by θ ($\mathbb{E}_\theta(x) = \int x f(x|\theta) dx$).

If the loglikelihood is twice differentiable, then the FIM may also be written as equation (A2).

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \log l(Y;\theta)}{\partial \theta^2} \right] \tag{A2}$$

As developped in[42], equations (A1) and (A2) lead to two different estimators for the expected FIM, given in equations (A3) and (A4), based on a n-sample $\left(y_1, \dots, y_n\right)$ of observations.

$$I_n^{sco}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln l(y_i;\theta)}{\partial \theta} \frac{\partial \ln l(y_i;\theta)}{\partial \theta^T} \tag{A3}$$

$$I_n^{obs}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log l(y_i;\theta)}{\partial \theta^2} \tag{A4}$$

$I_n^{obs}(\theta)$ is usually referred to as the observed FIM.

It has been shown that if $\left(Y_1, \dots, Y_n\right)$ are independent non identically distributed random variables whose density function depends on a common parameter θ, $I_n^{sco}(\theta)$ converges almost surely toward $I(\theta)$ and is asymptotically normal. The same results hold for $I_n^{obs}(\theta)$, on condition that the loglikelihood is twice differentiable.

In our setting with N subjects, the population expected FIM is the sum of the individual expected FIM given by equation (A5).

$$FIM(\theta) = \sum_{i=1}^{N} \mathbb{E}_\theta \left(\frac{\partial \ln l(y_i;\theta)}{\partial \theta} \frac{\partial \ln l(y_i;\theta)}{\partial \theta^T} \right) \tag{A5}$$

In the same way, two estimators can be derived:

$$FIM_n^{sco}(\theta) = \sum_{i=1}^{N} \frac{\partial \ln l(y_i;\theta)}{\partial \theta} \frac{\partial \ln l(y_i;\theta)}{\partial \theta^T} \tag{A6}$$

$$FIM_n^{obs}(\theta) = -\sum_{i=1}^{N} \frac{\partial^2 \log l(y_i;\theta)}{\partial \theta^2} \tag{A7}$$

$FIM_n^{sco}(\theta)$ can be stochastically approximated[42] by equation (A8).

$$I_{n,sco}^K = \sum_{i=1}^{N} \Delta_i^K \left(\Delta_i^K\right)^T \quad \text{with } \Delta_i^k = (1-\gamma_k)\Delta_i^{k-1} + \gamma_k \partial_\theta \ln l(y_i, \phi_i^k; \theta_{k-1}) \tag{A8}$$

Otherwise, according to Louis's missing information principle, $FIM_n^{obs}(\theta)$ is also given by equation A9

$$-\partial_\theta^2 \ln l(y;\theta) = -\mathbb{E}\Big[\partial_\theta^2 \ln l(y,\phi;\theta)|y;\theta\Big] - Cov\Big[\partial_\theta \ln l(y,\phi;\theta)|y;\theta\Big] \tag{A9}$$

where

$$Cov\Big[\partial_\theta \ln l(y,\phi;\theta)|y;\theta\Big] = \mathbb{E}\Big[\Big(\partial_\theta \ln l(y,\phi;\theta)\Big)\Big(\partial_\theta \ln l(y,\phi;\theta)\Big)^T|y;\theta\Big] - \mathbb{E}\Big[\Big(\partial_\theta \ln l(y,\phi;\theta)\Big)|y;\theta\Big]\mathbb{E}\Big[\Big(\partial_\theta \ln l(y,\phi;\theta)\Big)|y;\theta\Big]^T \tag{A10}$$

$FIM_n^{obs}(\theta)$ can thus be stochastically approximated through the sequence $\left(H_k\right)$ given in (A11).

$$\begin{aligned} H_k &= -\sum_{i=1}^{N}\left(D_i^K + G_i^K - \Delta_i^K\left(\Delta_i^K\right)^T\right) \\ &\text{with} \\ \Delta_i^k &= (1-\gamma_k)\Delta_i^{k-1} + \gamma_k \ln l(y,\phi^{(k)};\theta) \\ D_i^k &= (1-\gamma_k)D_i^{k-1} + \gamma_k \partial_\theta^2 \ln l(y,\phi^{(k)};\theta) \\ G_i^k &= (1-\gamma_k)G_i^{k-1} + \gamma_k\left(\partial_\theta \ln l(y,\phi^{(k)};\theta)\right)\left(\partial_\theta \ln l(y,\phi^{(k)};\theta)\right)^T \end{aligned} \tag{A11}$$

A.2 Results

A.2.1 Rich scenario without correlation

On rich datasets, because asymptotic conditions were reached, the observed FIM computed with stochastic approximation gave RSE close to those obtained with the expected FIM, computed by linearisation. The observed FIM based on the Hessian formula gave smaller RSE than the observed FIM based on the score formula and than the expected FIM computed by linearisation. We also note that the diagonal term relative to b in the inverse of the FIM is always negative, so its RSE cannot be calculated.

A.2.2 Sparse scenario without correlation

On sparse datasets, the observed FIM based on the score formula gave much higher RSE than the expected FIM and failed on some datasets because of non invertible matrices. The observed FIM based on the Hessian formula gave RSE close to those obtained with the expected FIM computed by linearisation; but on some datasets, RSE could not be computed for the variance and the residual error parameter because of the inverse of the FIM having negative diagonal elements.

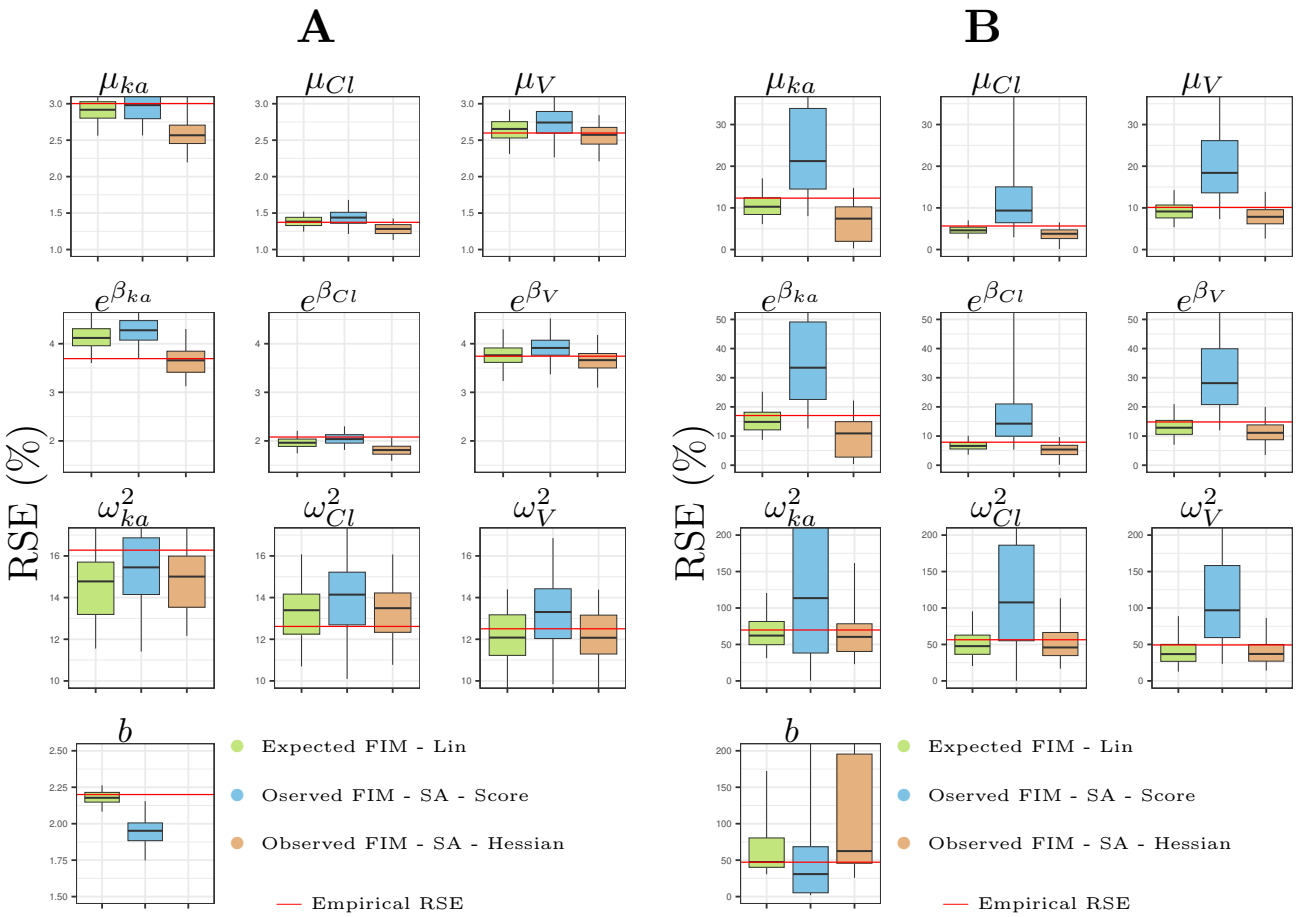


FIGURE A1 RSE
A: Rich scenario ($N = 150,\ n = 10$) without correlation
B: Sparse scenario ($N = 12,\ n = 3$) without correlation
Each panel represents one model parameter. The different methods are given on the X-axis. The target being the empirical SE from SAEM estimates is shown as an horizontal red line, and its 95% confidence interval as dotted red lines. The box represent the interval between the 25% and 75% quartiles with a mark at the median; while the whiskers represent the interval between the 2.5% and 97.5% quantiles.

A.2.3 Scenario with high IIV and correlation

On the sparse scenario with high IIV and correlation, the method using a stochastic approximation of the FIM failed due to non invertible matrices.

B APPENDIX: VARIATIONS AROUND SAEM_MH - METHODS

Several variations around SAEM_MH method were explored.

We first evaluated if drawing $K2$ MH chains during SAEM smoothing phase was more valuable than drawing only one long chain at the end of the algorithm and considering it as a sample of the posterior distribution.

Different kernel distributions, especially random walk, univariate kernels, and adaptive kernels were explored.

Another variation was to replace the MH algorithm by a Gibbs sampler; actually in our case it consisted in drawing the individual parameters once at each iteration of the chain, before drawing $\theta^{(m)}$.

B.1 MH after SAEM

Instead of drawing K2 samples in parallel of the smoothing phase of the SAEM, this algorithm draw one sample using the final estimates of the SAEM to build the a posteriori distribution.The kernel distribution was still a multivariate Gaussian, centered on the MLE and with the covariance matrix set to the inverse of the FIM evaluated at the MLE and inflated by a factor denoted $infl$.

B.2 Two-block Gaussian kernel distribution

In this variation, the kernel was conditional and by block, meaning that fixed effects were first drawn conditionally to the components of the covariance matrix and the residual error parameters from the previous iteration, then the components of the covariance matrix and the residual error parameters were drawn conditionally to the new fixed effects. Denoting $\theta_F^{(m)} = \left(\theta_1^{(m)}, \ldots, \theta_f^{(m)}\right)$ and $\theta_R^{(m)} = \left(\theta_{f+1}^{(m)}, \ldots, \theta_p^{(m)}\right)$ the vector subset of $\theta^{(m)}$ containing in one hand the fixed effects and in the other hand the components of the covariance matrix and the residual error parameters, it consisted in replacing the Metropolis Hastings algorithm by two consecutive Metropolis Hastings algorithms drawing $\theta_F^{(m)}$ and $\theta_R^{(m)}$ subsequently.

B.3 Three-block Gaussian kernel distribution

This variation consisted in three MH algorithms: first the fixed effects were drawn conditionally to the covariance matrix of the random effects and the residual error parameters from the previous iteration of the Markov chain, and accepted based on the MH rule, then the same was repeated for the covariance matrix of the random effects conditionally to the new fixed effects and the previous residual error parameters, and finally for the residual error parameter, conditionally to the new fixed effects and to the new covariance matrix of the random effects.

B.4 Univariate Gaussian kernel distribution

In this variation, the kernel was univariate and conditional, meaning that at each iteration, the parameters were drawn and accepted (or not) one after the other, and conditionally to the parameter already drawn.

B.5 Random walk

In this variation, the chain drawn was a random walk, meaning that the multivariate Gaussian kernel was not centered on the MLE anymore, but on the vector parameter from the previous chain iteration $\theta^{(m-1)}$.

B.6 Adaptive kernel - ad hoc method

In this method, the kernel distribution was a Gaussian distribution, with an adaptive covariance matrix set at each iteration m to the covariance of the $m-1$ previous states of the chain (after a "burn-in" period).

B.7 Adaptive kernel - Haario method

To improve the performance of the MH algorithm and to speed up its convergence, several methods have been proposed in the literature for 'on the fly' adaptation according to accepted samples.

In the method proposed by Haario et al.[25], the kernel distribution is a Gaussian distribution centered on the current state, and the covariance is said 'adaptive' because calculated accounting for all of the previous states. An advantage of this kernel is that it uses the cumulative information since the beginning of the simulation and the fast beginning of the adaptation guarantees that the exploration turns out to be more successful reducing the number of function evaluation required. Moreover, the authors proved that this adaptive version of MH has the correct ergodicity properties.

The kernel writes $\theta^{(m)} \sim \mathcal{N}\left(\theta^{(m-1)}, \Sigma_m\right)$ with Σ_m given in equation (B12).

$$\Sigma_m = \begin{cases} \Sigma_0 & \text{if } m < m_0 \\ sd\ Cov\left(\theta^{(0)}, \ldots, \theta^{(m-1)}\right) + sd\varepsilon Id & \text{if } m \geq m_0 \end{cases} \tag{B12}$$

The parameter m_0 is defined by the user and corresponds to the iteration of the chain from which the kernel is adapted. sd is a scaling parameter that may be chosen equal to $sd = \frac{2.4^2}{P}$ where P is the number of population parameters in the model. Indeed, this choice optimises the mixing properties of the Metropolis search in the case of Gaussian targets and Gaussian kernels[44]. $Cov(.)$ is the empirical covariance matrix. The parameter ε ensures that the matrix will not become singular.

The initial variance covariance matrix Σ_0 can be set according to prior information, therefore in our semi-Bayesian framework we used $\Sigma_0 = FIM^{-1}_{K1+K2}$.

B.8 Adaptive kernel - Garthwaite method

Garthwaite et al.[26] proposed to use a stochastic search algorithm to automatically tune the scale parameter of the Gaussian random walk MH algorithm to target a pre-specified value of the acceptance probability. Theoretical arguments suggest that optimal acceptance rate for Gaussian multivariate distribution is 0.234[33].

Hence, the kernel variance covariance matrix was the same as the previous one, but with an adaptive scale σ_m^2: $\theta^{(m)} \sim \mathcal{N}\left(\theta^{(m-1)}, \sigma_m^2\Sigma_m\right)$. The formula for σ_m^2 in given in equation (B13) with P the number of population parameters in the model.

$$\sigma_m = \sigma_{m-1} \exp\left(\delta\ \frac{\alpha\left(\theta^{(m-1)}, \theta^{(m)}\right) - \alpha^*}{P}\right) \tag{B13}$$

where

$$\delta = \left(1 - \frac{1}{P}\right)\frac{\sqrt{2\pi}\exp\left(\frac{a^2}{2}\right)}{2a} + \frac{1}{P\ \alpha^*(1-\alpha^*)} \tag{B14}$$
$$a = -\Phi^{-1}\left(\frac{\alpha^*}{2}\right) \quad \text{and} \quad \alpha^* = 0.234$$

B.9 Metropolis-Hastings within Gibbs

In Metropolis-Hastings within Gibbs, individual parameters were jointly considered with the population parameters within a Gibbs sampling procedure. A MH acceptation procedure was still needed for the population parameters.

C APPENDIX: VARIATIONS AROUND SAEM_MH - RESULTS

Regarding the variations of SAEM_MH, having only one long chain did not seem to worsen the CR. Using a 2-block or a univariate Gaussian kernel distribution (7th and 8th methods respectively) or using a 2-block random walk (10th) improved the CR for the fixed effects. Note that in our case, the FIM was block diagonal, meaning that drawing from the conditional by block distribution was the same as drawing the two blocks independently. The difference came from the acceptation rate: the AR was higher for the fixed effects blocks, suggesting that in the case where all the parameters were drawn at the same time, the "bad" values of the random effects caused the whole vector to be rejected even if the fixed effects were plausible. No improvement was observed using a three-block kernel (results not shown). With or without random walk, all the adaptive methods tested for the variance covariance of the kernel worsened the CR (methods 11 to 14) despite the promise of a better exploration of the parameter space. Within Gibbs methods with two-block kernel or univariate kernel showed good CR for all the fixed effects and the residual error parameter, and adding RW with two-blocks kernel allowed to better perform for the Ω with CI containing the target value or at least being closer, at the expense of the residual error parameter, whose CR was above the target. Using a three-block kernel did not improve the results.

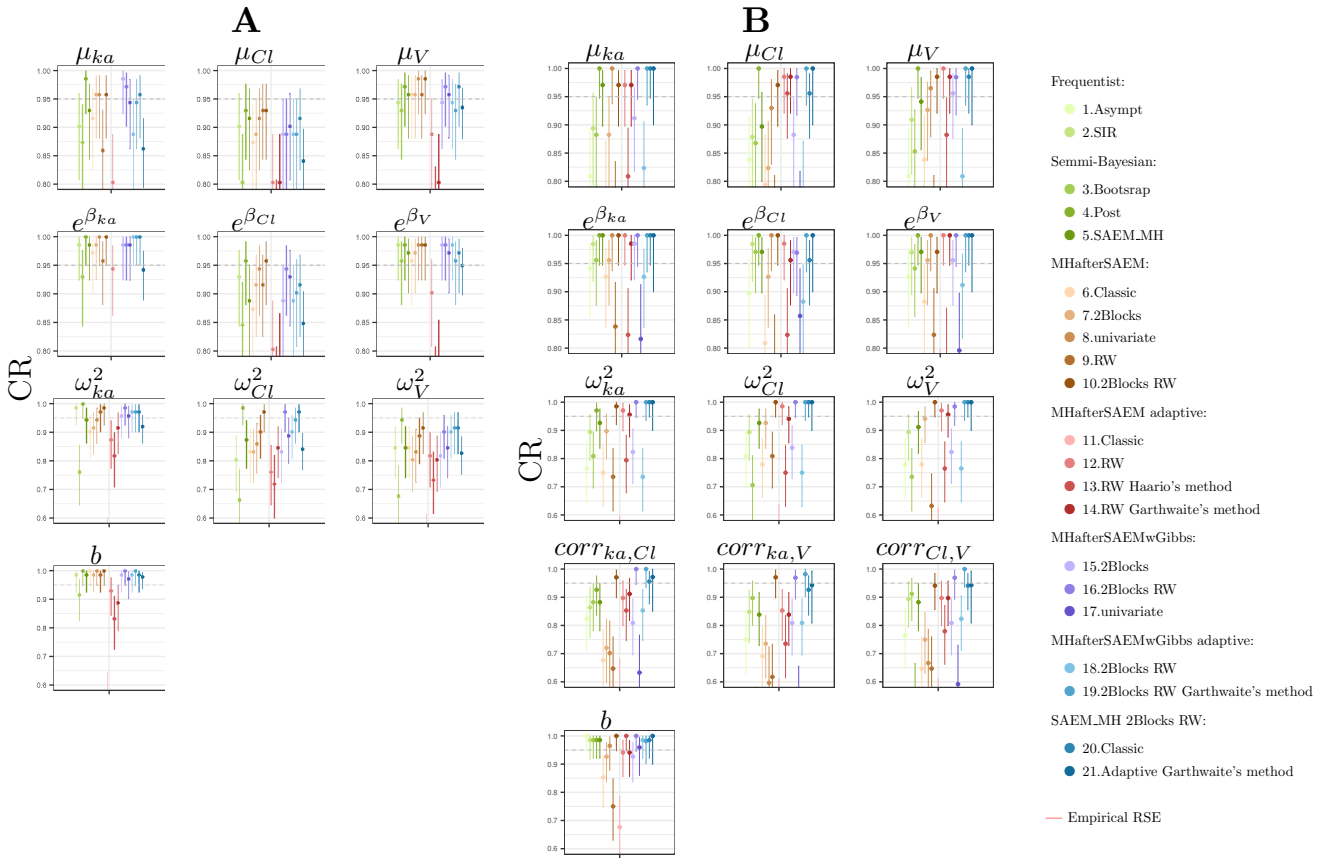


FIGURE C2 95% coverage rate and its 95% confidence interval
A: Scenario without correlation
B: Scenario with high IIV and correlation
Each panel represents one model parameter. The different methods are given on the X-axis. The target CR at 0.95 is shown as an horizontal line.

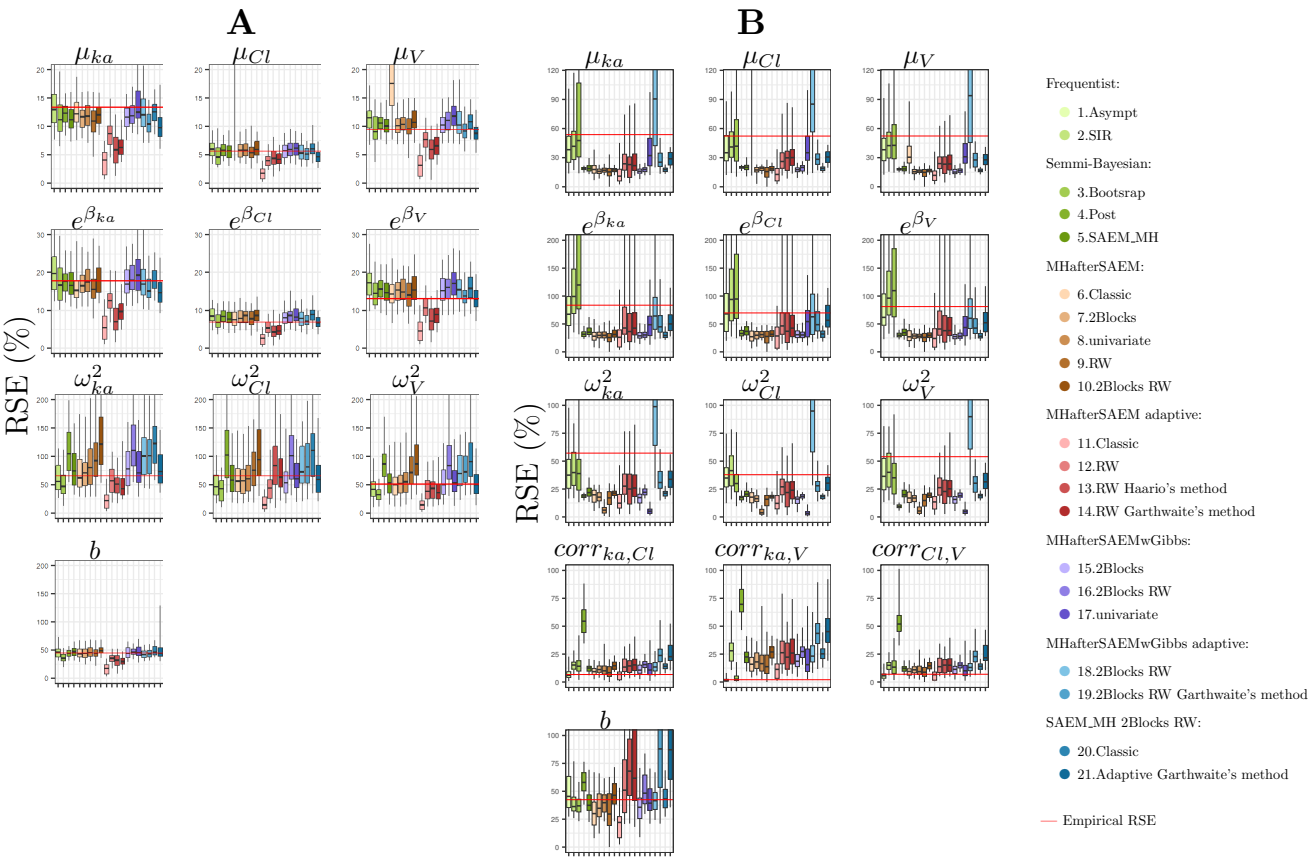


FIGURE C3 Estimated Relative Standard Error
A: Scenario without correlation
B: Scenario with high IIV and correlation
Each panel represents one model parameter. The different methods are given on the X-axis. The target being the empirical SE from SAEM estimates is shown as an horizontal red line, and its 95% confidence interval as dotted red lines. The box represent the 25%, 50% and 75% quantiles and the whiskers the 2.5% and 97.5% quantiles.

D APPENDIX: ABC ALGORITHM VARIATIONS

For the variations of the ABC method, inflating the inverse of the FIM by two allowed better coverage, which makes sense as we have seen that the Asympt method underestimated the SE. Further, if CR results were similar across the methods for the fixed effects and the Ω, only the use of squared individual predictions in the linear regression allowed to cover the residual error parameter properly. Actually, considering the relative estimation error (REE), only the latter summary statistics associated with the three-block Gaussian prior was unbiased for the residual error parameter.

D.1 Relative Estimation Error

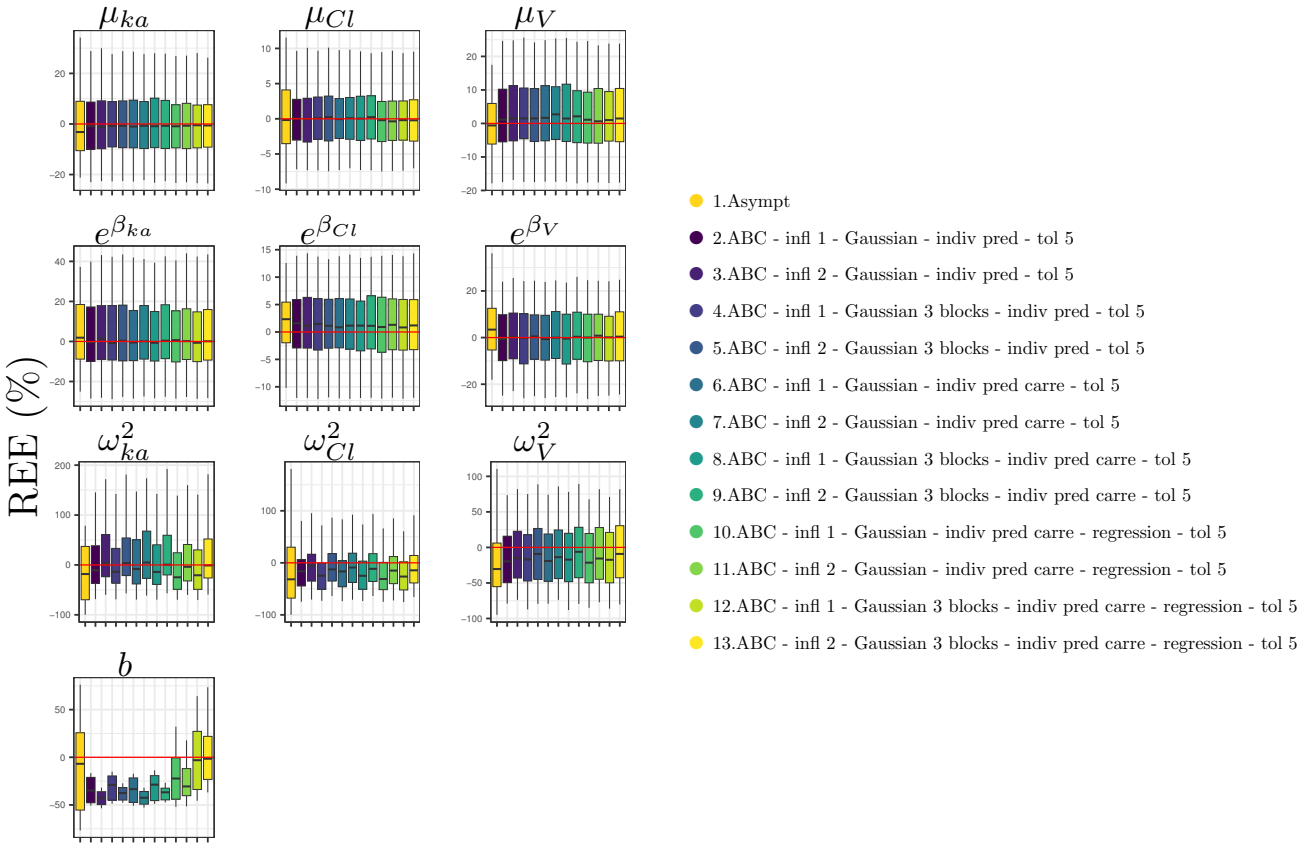


FIGURE D4 ABC comparison - Scenario without correlation - Relative Estimation Error

D.2 Coverage Rate

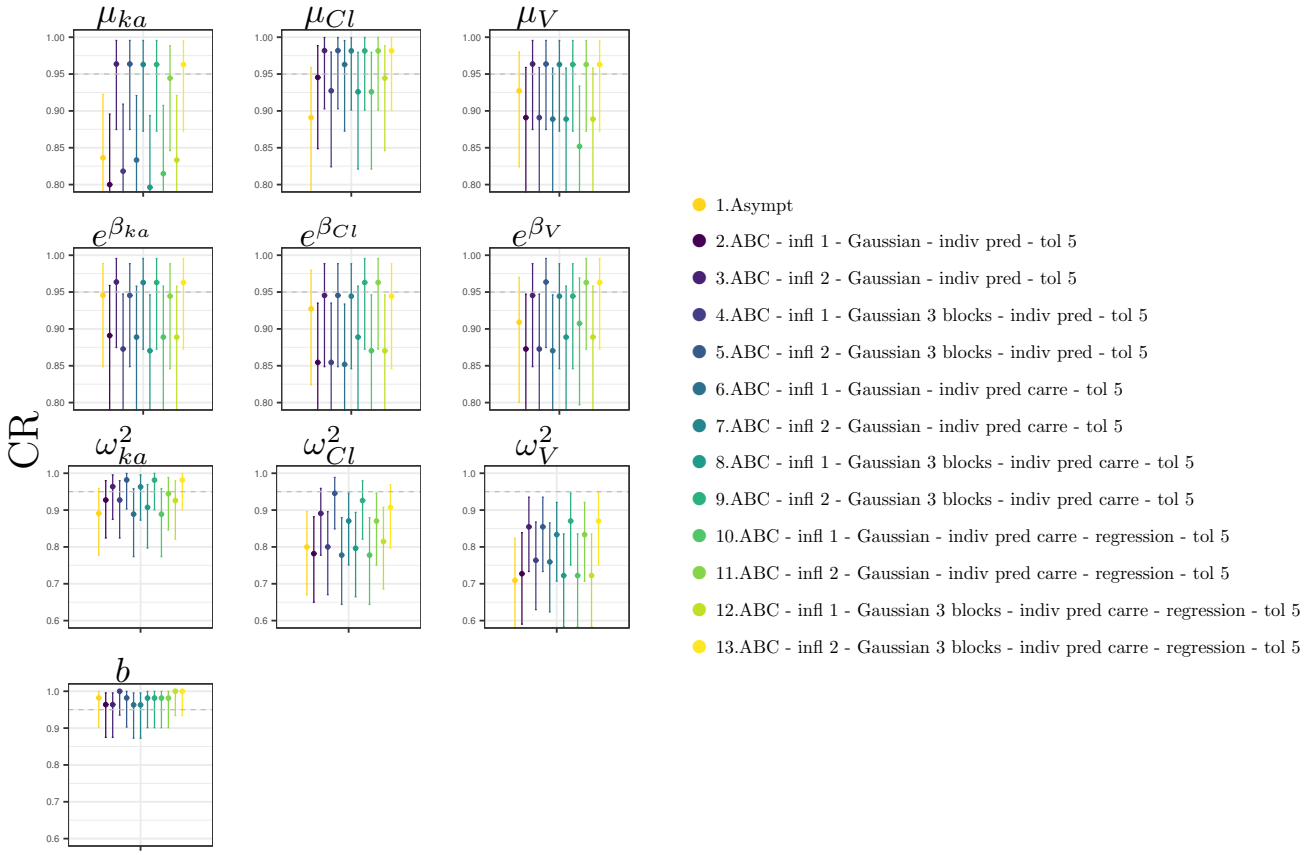


FIGURE D5 ABC comparison - Scenario without correlation - 95% coverage rate and its 95% confidence interval
Each panel represents one model parameter. The different methods are given on the X-axis. The target CR at 0.95 is shown as an horizontal line.

D.3 Estimated SE

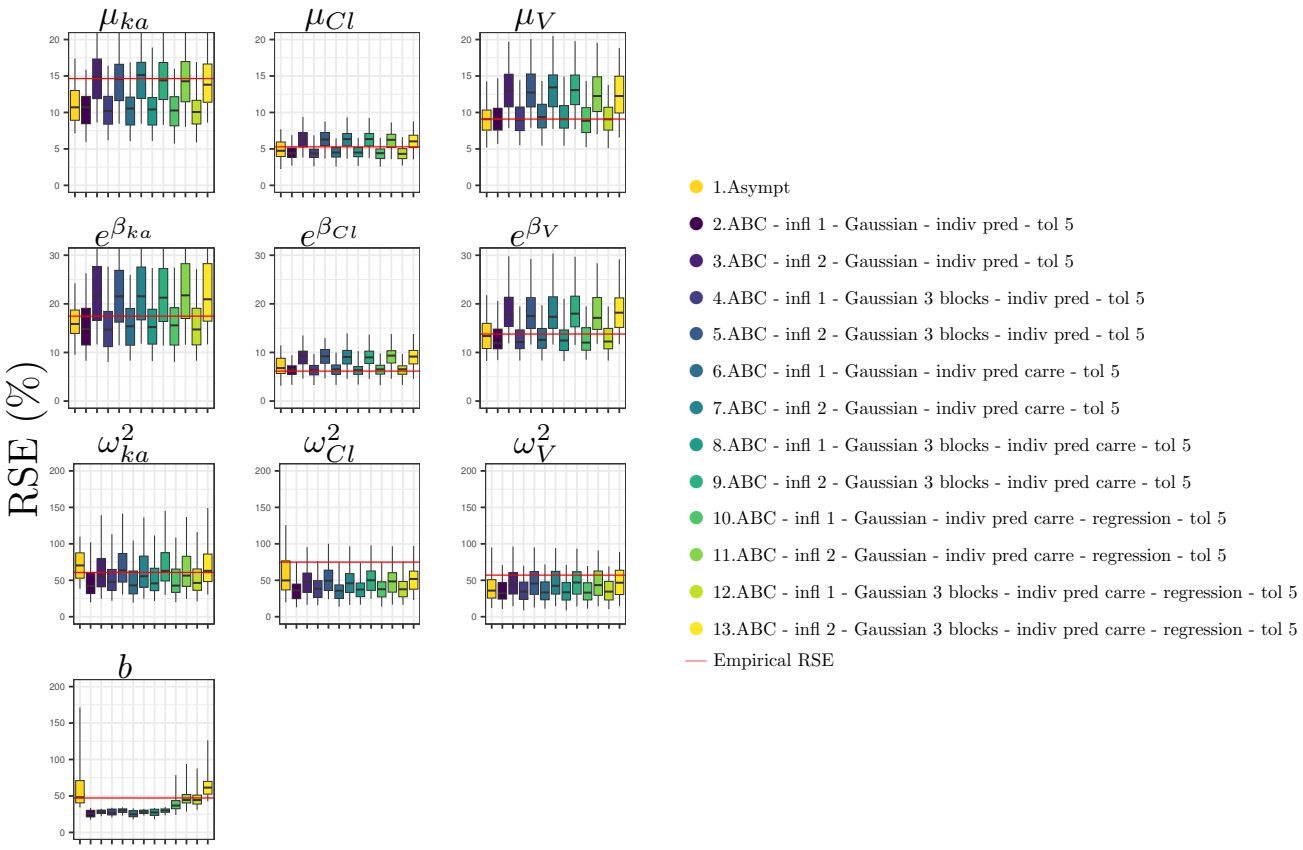


FIGURE D6 ABC comparison - Scenario without correlation - Estimated Standard Error
Each panel represents one model parameter. The different methods are given on the X-axis. The target SE from the `saemix` fits is shown as an horizontal red line. The box represent the 25%, 50% and 75% quantiles and the whiskers the 2.5% and 97.5% quantiles.

E APPENDIX: RESULTS FOR RICH DESIGN WITHOUT CORRELATION

On this scenario, Bootstrap was the most computation intensive method, with an average time of 6 hours , and was also the method with the highest variability in run times. Post was the fastest method, taking 10 minutes on average, followed by MH_after_SAEM (25 minutes). ABC was slower, with an average time of 4.5 hours , almost as long as SAEM_MH with 5.5 hours.

E.1 Rich design without correlation - Coverage

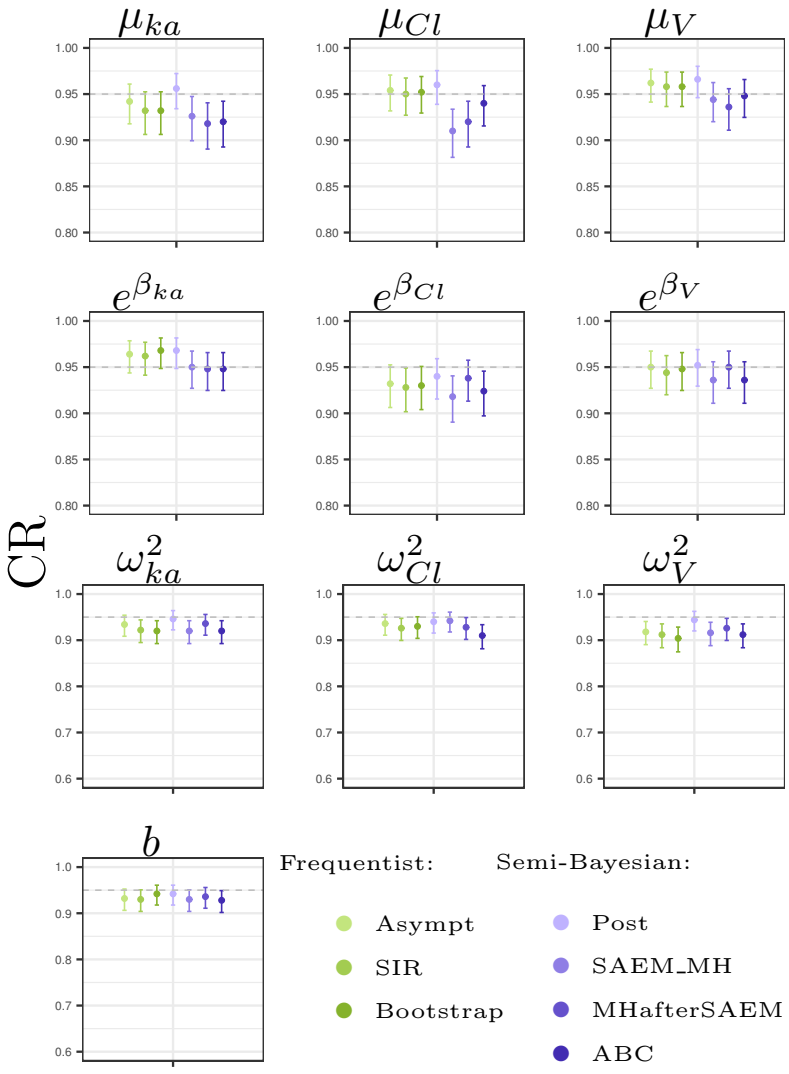


FIGURE E7 Scenario without correlation - Rich data - 95% coverage rates and their 95% confidence interval
Each panel represents one model parameter. The different methods are given on the X-axis. The target CR at 0.95 is shown as an horizontal line.
Asympt method uses the expected FIM computed by linearisation ; SIR uses a proposal two times inflated ; Bootstrap refers to case bootstrap with 500 resamples ; SAEM_MH uses a kernel two times inflated ; MH after SAEM uses a two times inflated 2-Block kernel with Random Walk ; ABC uses a two times inflated 3-Block kernel and tolerance was set to 5.

E.2 Rich design without correlation - RSE

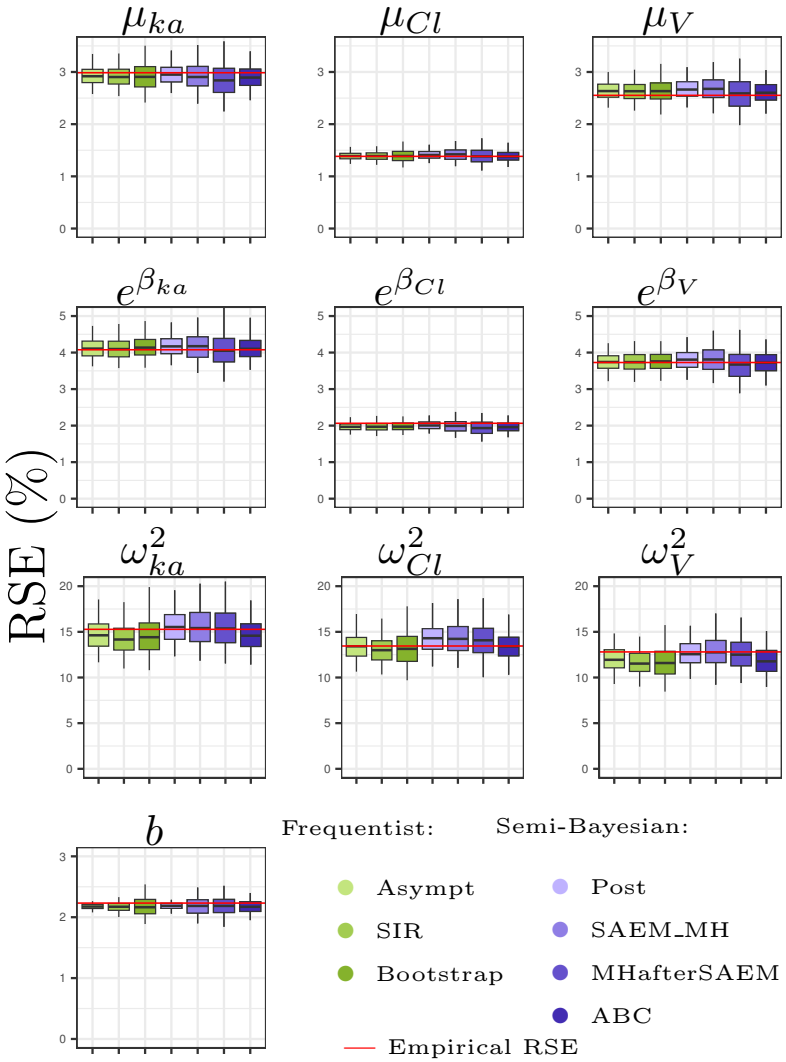


FIGURE E8 Scenario without correlation - Rich data - Estimated Relative Standard Error
Each panel represents one model parameter. The different methods are given on the X-axis. The target being the empirical RSE from SAEM estimates is shown as an horizontal red line, and its 95% confidence interval as dotted red lines. The box represent the interval between the 25% and 75% quartiles with a mark at the median; while the whiskers represent the interval between the 2.5% and 97.5% quantiles.
Asympt method uses the expected FIM computed by linearisation ; SIR uses a proposal two times inflated ; Bootstrap refers to case bootstrap with 500 resamples ; SAEM_MH uses a kernel two times inflated ; MH after SAEM uses an two times inflated 2-Block kernel with Random Walk ; ABC uses an two times inflated 3-Block kernel and tolerance was set to 5.

F APPENDIX: DENSITY AND KL DIVERGENCE

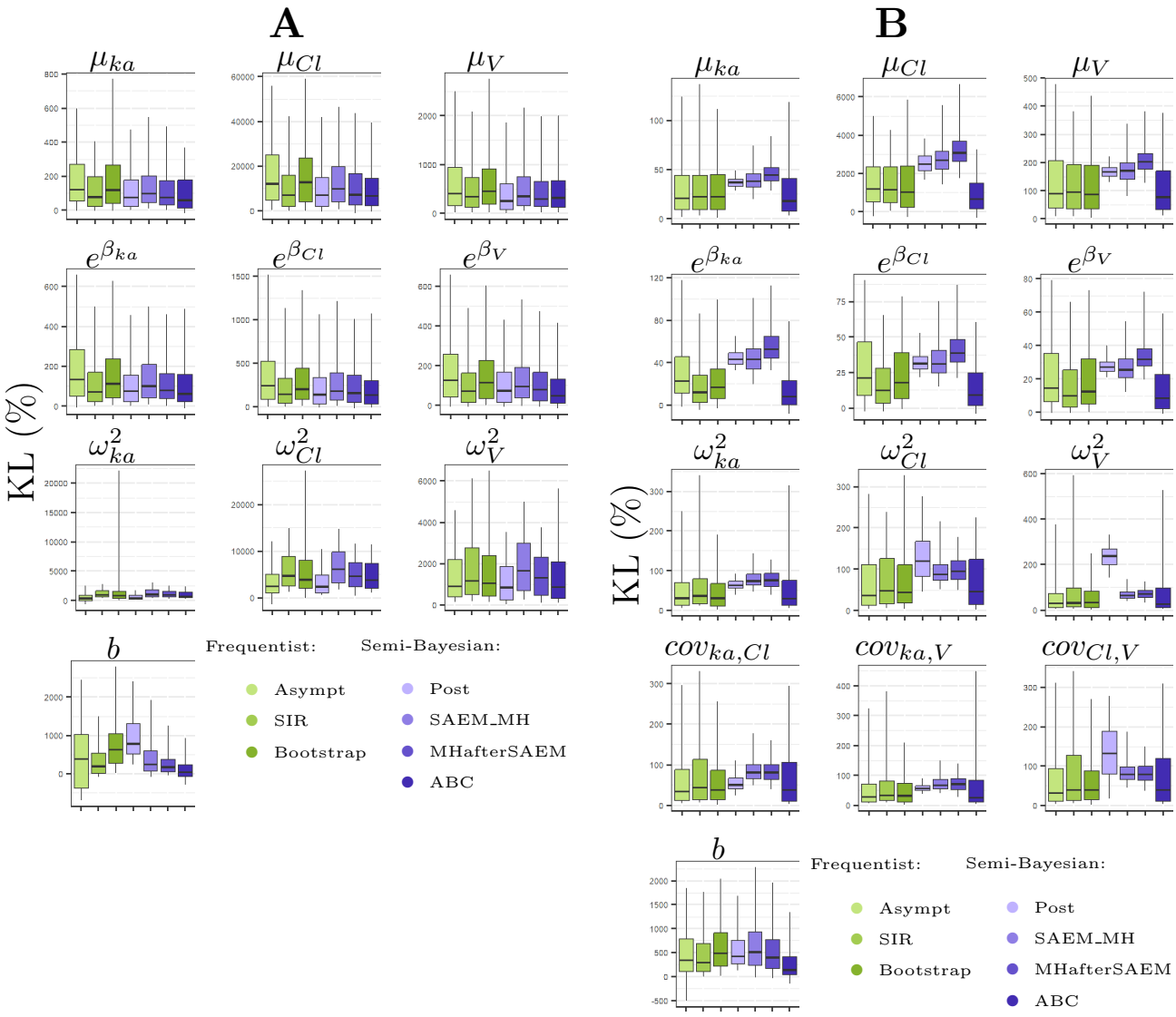


FIGURE F9 Kullback Leibler divergence
A: Scenario without correlation
B: Scenario with high IIV and correlation
Each panel represents one model parameter. The different methods are given on the X-axis. The box represent the interval between the 25% and 75% quartiles with a mark at the median; while the whiskers represent the interval between the 2.5% and 97.5% quantiles.

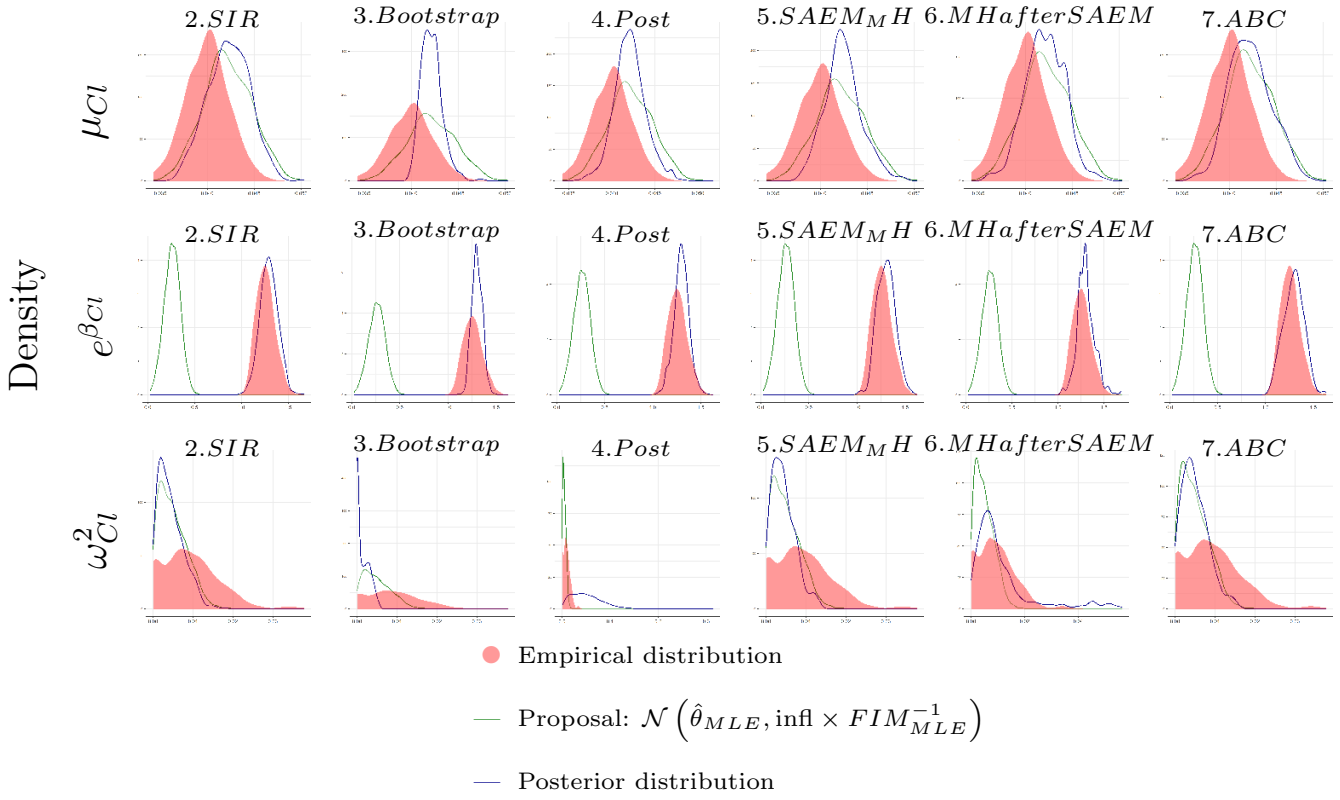


FIGURE F10 Scenario without correlation - Posterior density for third dataset
Each line corresponds to a parameter and each column to a method. The red area corresponds to the empirical distribution, the green line to the inflated Gaussian kernel and the blue line to the posterior distribution.

F.1 Scenario without correlation

Figure F10 shows the comparison between the empirical distribution, the kernel distribution $\mathcal{N}\left(\hat{\theta}_{MLE}, 2FIM^{-1}_{MLE}\right)$ and the posterior distribution obtained with SIR, bootstrap, Post, SAEM_MH, MH_after_SAEM and ABC for μ_{ka}, β_{ka} and ω^2_{ka}. The more the posterior distribution covers the empirical one, the better is the inference.

F.2 Scenario with high IIV and correlation

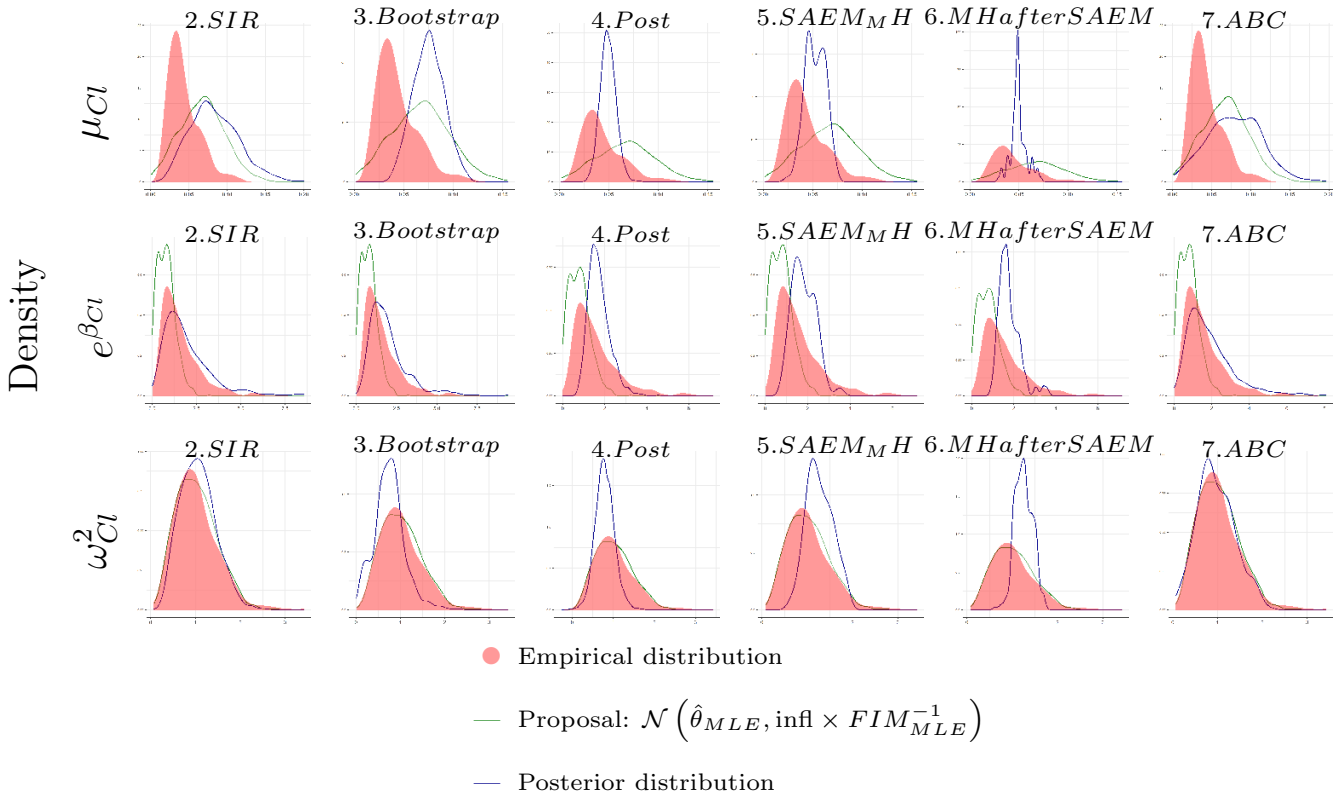


FIGURE F11 Scenario with high IIV and correlation - Posterior density for third dataset
Each line corresponds to a parameter and each column to a method. The red area corresponds to the empirical distribution, the green line to the inflated Gaussian kernel and the blue line to the posterior distribution.

G APPENDIX : DETAILED RESULTS ON GANTENERUMAB DATA

Parameter	Estimation (RSE,%) Full dataset	Estimation (RSE,%) N=12
$\mu_{ka}\ (d^{-1})$	0.46 (16)	0.57 (30)
β^T_{ka}	0.29 (77)	0.26 (162)
$\mu_{CL/F}\ (L.d^{-1})$	0.63 (5)	0.63 (8)
$\beta^T_{CL/F}$	-0.07 (108)	0.003 (4017)
$\mu_{V_1/F}\ (L)$	14.63 (9)	13.36 (17)
$\beta^T_{V_1/F}$	0.20 (61)	0.28 (77)
$\mu_{Q/F}\ (L.d^{-1})$	0.42 (31)	0.54 (67)
$\beta^T_{Q/F}$	-0.51 (129)	-0.85 (151)
$\mu_{V_2/F}$ (L)	4.88 (15)	4.78 (31)
$\beta^T_{V_2/F}$	-0.48 (63)	-0.43 (136)
$\mu_{T_{lag}}\ (d)$	0.04 (33)	0.04 (80)
$\beta^T_{T_{lag}}$	-0.03 (1430)	-3.05 (674)
ω_{ka}	0.65 (11)	0.59 (22)
$\omega_{CL/F}$	0.26 (11)	0.20 (22)
$\rho_{CL/F,V_1/F}$	0.76 (9)	0.96 (5)
$\omega_{V_1/F}$	0.33 (11)	0.25 (24)
$\omega_{T_{lag}}$	0.90 (24)	1.18 (55)
b	0.15 (4)	0.16 (7)

TABLE G1 Parameters estimates and their relative standard errors (RSE %) given by `saemix` for the modelling of the two formulations of Gantenerumab, in the full dataset of N=48 subjects (left) and a subset of N=12 subjects (right)

Parameter	Asympt	SIR	Bootstrap	Post	SAEM_MH	MH after SAEM	ABC
μ_{ka}	16	14	13	13	14	12	16
β_{ka}	77	75	101	140	70	80	77
$\mu_{CL/F}$	5	5	6	6	5	5	5
$\beta_{CL/F}$	108	109	122	127	101	184	110
$\mu_{V_1/F}$	9	9	9	8	7	8	8
$\beta_{V_1/F}$	61	56	75	86	55	59	58
$\mu_{Q/F}$	31	25	22	23	24	27	29
$\beta_{Q/F}$	129	108	166	1570	83	120	142
$\mu_{V_2/F}$	15	14	14	14	13	13	15
$\beta_{V_2/F}$	63	59	527	99	47	54	64
$\mu_{T_{lag}}$	33	30	40	23	22	36	32
$\beta_{T_{lag}}$	1430	404	1244	623	523	309	663
ω_{ka}	11	11	13	12	11	12	11
$\omega_{CL/F}$	11	9	9	12	10	11	12
$\rho_{CL/F,V_1/F}$	9	8	9	11	9	10	10
$\omega_{V_1/F}$	11	10	12	12	11	12	12
$\omega_{T_{lag}}$	24	18	26	22	19	30	23
b	4	4	6	4	4	4	4

TABLE G2 Relative standard errors computed for the parameters of the model fitted on the full dataset (N=48) with Asympt, SIR, bootstrap, Post, SAEM_MH, MH_after_SAEM and ABC methods

Parameter	Asympt	SIR	Bootstrap	Post	SAEM_MH	MH after SAEM	ABC
μ_{ka}	30	28	25	17	29	20	42
β_{ka}	162	199	131	2594	235	19	215
$\mu_{CL/F}$	8	8	9	9	9	1	12
$\beta_{CL/F}$	4017	3144	488	1747	950	246	1850
$\mu_{V_1/F}$	17	16	16	12	13	6	21
$\beta_{V_1/F}$	77	81	67	100	61	5	90
$\mu_{Q/F}$	67	52	52	27	42	24	57
$\beta_{Q/F}$	151	149	113	585	97	6	139
$\mu_{V_2/F}$	31	28	2236	21	21	6	34
$\beta_{V_2/F}$	136	137	5962	729	95	66	172
$\mu_{T_{lag}}$	70	59	73	29	53	7	68
$\beta_{T_{lag}}$	674	83	103	297	41	1289	426
ω_{ka}	22	21	37	26	21	26	29
$\omega_{CL/F}$	22	20	20	28	18	23	25
$\rho_{CL/F,V_1/F}$	5	5	20	16	6	10	6
$\omega_{V_1/F}$	24	21	33	31	21	23	25
$\omega_{T_{lag}}$	55	35	70	29	27	19	42
b	7	8	9	8	9	7	10

TABLE G3 Relative standard errors computed for the parameters of the model fitted on the sparse subset of the data (N=12) with Asympt, SIR, bootstrap, Post, SAEM_MH, MH_after_SAEM and ABC methods

Chapitre 5

Modélisation des données de score de l'essai Discovery

5.1 Résumé

Objectifs

Dans ce dernier projet, nous nous sommes intéressées à la modélisation des données longitudinales de score clinique recueillies dans l'essai Discovery.

L'objectif principal était d'appliquer les méthodes précédemment évoquées sur ces données, après avoir modélisé l'évolution du score clinique chez les patients hospitalisés pour infection au SARS-CoV-2 et testé l'effet du remdesivir sur cette évolution.

Synthèse

L'essai Discovery, présenté dans le chapitre 1.5 de ce manuscrit, inclut plusieurs centaines de patients suivis quotidiennement pendant 30 jours. L'idée initiale était donc de modéliser l'ensemble des données des bras SoC et SoC+remdesivir, afin de tester un effet traitement

sur l'évolution clinique des patients, puis d'appliquer les méthodes de calcul de l'incertitude précédemment développées sur un sous-ensemble des données qui permettrait de s'éloigner du cas asymptotique. L'évolution de l'état des patients était suivie notamment au travers de deux scores cliniques, le score composite NEWS-2 combinant plusieurs paramètres physiologiques, et le score OMS représentant l'état clinique général.

Les scores NEWS-2 ont été considérés comme des données continues, et le modèle a été sélectionné sur le bras SoC (modèle structurel et modèle d'erreur, structure de la matrice de variance-covariance des effets aléatoires inter-individuels, effets des covariables), puis appliqué aux deux bras de traitement SoC et SoC+remdesivir, sur lesquels nous avons alors ajouté des effets traitement à chaque paramètre et testé leur significativité via un test du rapport de vraisemblance. Le modèle retenu était une forme limite du modèle de Bateman, avec des variabilités inter-individuelles très fortes (>100%) et plusieurs corrélations proche de 1 entre les effets aléatoires. L'effet du traitement estimé n'était pas significatif.

Bien que les données utilisées soient riches, la complexité du modèle obtenu a mis en défaut la méthode SAEM_MH, dont les taux d'acceptation étaient trop bas pour être fiables, comme on a pu l'observer dans les projets précédents. C'est une des raisons pour lesquelles nous avons été amenées à explorer la méthode ABC présentée dans le chapitre 4, qui a donné des résultats plus probants.

Apports du travail

La modélisation de l'évolution des scores cliniques NEWS-2 recueillis dans l'essai Discovery, avec la construction d'un modèle continu, a permis d'obtenir des estimations précises et de décrire les profils variés observés. Le test d'effet traitement faisant suite à cette modélisation n'a pas permis de mettre en évidence un effet du remdesivir sur l'amélioration clinique des patients.

Les données longitudinales de score de l'essai Discovery, malgré leur design riche, présentent des structures de variabilité complexes et mettent en difficulté la méthode SAEM_MH peu adaptée à ces caractéristiques.

Le calcul des SE sur un sous-ensemble des données a montré des disparités entre méthodes fréquentistes et semi-Bayésiennes. Sur ce sous-ensemble, visiblement éloigné de l'asymptotique, l'algorithme ABC a donné de bonnes performances et semble plus robuste à ce type de données comme dans le travail précédent.

5.2 Article en préparation

Modelling the evolution of clinical scores in the Discovery trial using nonlinear mixed effects models

Melanie Guhl, Julie Bertrand, Jérémie Guedj[1] and Emmanuelle Comets[1,2]

[1]Université Paris Cité, Inserm, IAME, F-75018 Paris, France
[2]Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085, F-35000 Rennes, France

Abstract

This work aimed to model the longitudinal clinical scores collected in the Discovery trial to test for an effect of remdesivir on the clinical evolution of the patients hospitalised with SARS-CoV-2 infection. To do so, we modelled the longitudinal 20-categories NEWS-2 ordinal score with a continuous nonlinear mixed effects model (NLMEM). The covariate model was built on 414 patients from the Standard of Care arm. A treatment effect was added to all model parameters when adding the 418 patients also treated with remdesivir, but the log-likelihood ratio test showed no evidence of a remdesivir effect. We applied different semi-Bayesian standard errors (SE) computation methods previously developed in the context of sparse longitudinal data on the final model, confirming the limits of some of these methods in the presence of complex variability structures.

1 Introduction

The Discovery trial [1] (NCT04315948) is a European multicentric randomised clinical trial promoted by Inserm (principal investigator: Pr Florence Ader) aiming to evaluate antiviral drugs for the treatment of SARS-CoV-2. Patients included in the trial were followed until discharge for at most 29 days. There were five treatment arms: standard of care (SoC), SoC + hydroxychloroquine, SoC + remdesivir, SoC + lopinavir/ritonavir, SoC + lopinavir/ritonavir + interferon. The primary endpoint used in this study was the WHO 7-categories ordinal clinical score at day 15. A number of clinical variables were also recorded daily during hospital stay, including the WHO score, naso-pharyngeal viral load and the composite clinical score NEWS-2, and these variables were obtained at day 15 and day 29 for all patients including those discharged. The arms evaluating hydroxychloroquine, lopinavir/ritonavir and lopinavir/ritonavir + interferon were prematurely stopped for futility [2]. The final results published showed no effect of any treatment on the primary endpoint, although remdesivir significantly delayed the worsening of patients clinical status (adjusted hazard ratio 0.63, 95% confidence interval [0.45–0.88], p=0.010) [3].

A secondary study on longitudinal outcomes showed an effect of remdesivir on viral loads [4]. Another secondary study highlighted a link between viral loads and evolution of the clinical score. It also confirmed the effect of remdesivir on viral loads, even though the effect of remdesivir was not strong enough, according to simulation studies, to affect the clinical status of the patients studied [5]. These results show that modelling the longitudinal data is powerful and can lead to detect effects that do not show on the primary endpoints. Here our goal was to model the clinical status evolution of the patients included in the clinical trial between March 2020 and January 2021 in the SoC and SoC+remdesivir arms, by considering the NEWS-2 score as a continuous response. To model this evolution, we used

non linear mixed effect models (NLMEM), to account for inter-individual variability. Finally, we tested a treatment effect of remdesivir on the parameters of the NEWS-2 model.

Testing for a treatment effect of remdesivir on the evolution of the scores can be done based on the likelihood of the model or the significance of the treatment effect coefficient estimate via a Wald test. For the latter, we need to estimate the standard error (SE) of the treatment effect coefficient estimate. The classical method is obtained from the Fisher Information Matrix (FIM), which is derived from the likelihood of the model, and allows to compute the Cramér-Rao bound, which is asymptotically valid for the MLE. In previous work [6, 7, 8], semi-Bayesian methods to compute SE in NLMEM were also proposed.

In this work we applied these different methods to compute the SE of the continuous model parameter estimates describing longitudinal NEWS-2 scores. We used the full dataset and a subset of the data. We present perspectives to extend these methods for categorical models to be used on the longitudinal WHO scores and where SE are trickier to compute.

2 Methods

2.1 Data

We used data from the Discovery trial describing the daily clinical status of the patients included in the trial between March 2020 and January 2021, in the SoC (N=414 patients) and SoC+remdesivir (N=418 patients) arms.

The NEWS-2 score combines six physiological parameters (respiratory rate, oxygen saturation, systolic blood pressure, heart rate, level of consciousness and temperature), and ranges from 0 (best) to 20 (worst) and was developed to assess the urgency of treating a patient [9]. We have daily measures of this score for the patients hospitalised, and at least two measures at days 15 and 29 for those discharged before the end of follow up at day 29. In total, 408 patients of the SoC arm and 402 patients of the SoC+remdesivir arm have at least one recorded NEWS-2 score.

As in [5], the covariates tested were sex, age (dichotomised at 65 years old), and comorbidities with prevalence at least 10% and missing data at most 10%: chronic cardiac disease, chronic pulmonary disease, obesity, diabetes and smoking status.

2.2 Non linear mixed effects models (NLMEM)

The aim of NLMEM is to analyse longitudinal data by taking into account the individual variability, and therefore computing mean population parameters as well as individual parameters. For a continuous and normally distributed outcome and without intra-individual variability, the NLMEM can be written as follows [10]:

$$y_{ij} = f(t_{ij}, \psi_i) + g(t_{ij}, \psi_i, \xi)\ \epsilon_{ij} \tag{1}$$

$$\phi_i = h(\psi_i) = h(\mu) + C_i\beta + \eta_i \tag{2}$$

- t_{ij}: time j of subject i
- y_{ij}: outcome of subject i at time t_{ij} (i=1,...,N, j=1,...,n_i)
- $f()$: continuous structural model
- ψ_i: parameter vector of subject i
- $g()$: error model
- ξ: error parameter vector
- $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$: residual error

- ϕ_i: individual parameter vector of subject i on the transformed scale
- $h()$: transformation function for individual parameters (log, logit, etc)
- μ: mean parameter vector
- C_i: covariate matrix of subject i
- β: covariate effect vector
- $\eta_i \sim \mathcal{N}(0, \Omega)$: inter-individual random effects of subject i
- Ω: variance-covariance matrix of the inter-individual random effects

The vector of population parameters to estimate is $\theta = (\mu, \beta, \Omega, \xi)$. Here we assumed that the transformed fixed effects $h(\mu)$ and covariate effects β enter linearly in the definition of ϕ.

To compute the maximum likelihood estimator (MLE) of the parameters of a NLMEM, we use the Stochastic Approximation of the EM algorithm (SAEM) [11]. It is an iterative algorithm of K_1 exploratory iterations and K_2 smoothing iterations, replacing at each iteration the classical Expectation (E) step by the simulation of a vector of individual parameters ψ and the computation of the complete likelihood $l(y, \psi|\theta)$, then updating the population parameter vector estimate $\hat{\theta}$ by maximising the likelihood (M step).

2.3 Standard error (SE) computation

The classical standard error (SE) computation method of the MLE is based on the Fisher Information Matrix (FIM). For Gaussian models, the FIM of NLMEM is computed by linearisation [10]. This method is hereafter called Asympt.

Sampling Importance Resampling (SIR) [12] is an alternative method that consists in drawing M_S samples from a proposal distribution and resampling a subset of m_S of these samples according to weights, named importance ratios, computed based on the density of the proposal and the likelihood of each sample. This method, as all the ones presented below, builds the distribution of the estimator and estimate the SE by its standard deviation, as well as others statistics such as quantiles.

Post [13] is a semi-Bayesian method that consists in drawing samples of the parameter estimates using the Hamiltonian Monte Carlo (HMC) algorithm, after reaching convergence with SAEM to estimate the MLE. It is called semi-Bayesian because we keep the MLE as the point estimate and use the samples drawn by HMC only to compute the uncertainty around it. As we use a Bayesian algorithm, we need to define a prior distribution on the parameter vector.

SAEM_MH [14] is also a semi-Bayesian method, that we embedded in the SAEM algorithm, which consists in drawing one chains of samples in the distribution of the MLE at each iteration k of the smoothing phase of SAEM using a Metropolis Hastings (MH) algorithm. For this method, we need to define a kernel distribution $q()$ to sample from, as well as a prior distribution and the length of each MH chain M_k. The last sample of each chain obtained is kept and the SE is estimated by the standard deviation of the final chain obtained.

MH_after_SAEM [8] is a variation of SAEM_MH where we only draw one chain of samples at the end of the SAEM algorithm. The kernel distribution is different than the one used in SAEM_MH (see details in implementation).

Approximate Bayesian Computation (ABC)[8] is another semi-Bayesian method that, in contrast to the previous methods, does not rely on computing the likelihood of the model: instead, it compares directly the closeness of observed data to data simulated according to parameters drawn using a predefined distance function, and accepts these draws following a given threshold.

2.4 Modelling strategy

2.4.1 Structural model: Bateman equation

In this analysis, we considered the NEWS-2 score as continuous. Because this score is positive and may take the value 0, we chose to log-transform the scores and model $x = log(1 + NEWS2)$, and we used a model derived from the Bateman equation [15] mathematically similar to an equation used for pharmacokinetic models, but with parameters defined for a broader context:

$$y(t) = a\frac{\lambda_2}{\lambda_2 - \lambda_1}(exp(-\lambda_1 t) - exp(-\lambda_2 t))$$

The function describes a curve with an exponential ascending phase parameterised by a rate λ_1 and an exponential descending phase parameterised by a rate λ_2, with a being a scale parameter. The area under the Bateman curve A is equal to a/λ_1. We can re-parameterise the function to estimate A directly. Finally, when $\lambda_1 = \lambda_2 = \lambda$, we obtain the limit form of the function by using a limited development:

$$y(t) = A\lambda^2 t\ exp(-\lambda t)$$

In this study, we used this last equation to model the change in the NEWS-2 score from a baseline value of N_0 that we assumed could vary from 0 to 3 (as in [5]). As the two rates were very close and hard to differentiate from one another, the model included the same slope λ for the exponential increase and decrease. We also added a negative time lag T_{lag} parameter representing the delay between symptom onset and inclusion. Taking A to now represent the area under the NEWS-2 curve and above N_0, so that we can test for a treatment effect on A, our model equation becomes:

$$log\big(1 + \text{NEWS2}(t)\big) = log\big(1 + N_0 + A\ \lambda^2\ (t + T_{lag})\ exp(-\lambda(t + T_{lag}))\big)$$

A, λ and T_{lag} were assumed to follow a log-normal distribution (i.e. $h() = log()$ for these parameters) and N_0 a logit-normal distribution between 0 and 3. We also assumed a constant error model.

2.4.2 Model selection procedure

The parameters were estimated using the maximum likelihood estimator (MLE) in a frequentist paradigm, via the SAEM algorithm [11].

Modelling was performed in two stages. In the first stage, we used the data from the SoC arm only to select the structural model, inter individual variance-covariance structure and covariates (among demographics and comorbidities). We checked that relative standard errors (RSE) were below 50% for all parameters. We started by estimating random effects on all parameters with a full correlation structure in the inter-individual variance-covariance matrix, and only kept variances with RSE below 50% and correlations estimated ≥ 0.7. We then performed covariate selection as follows: first the empirical Bayes estimates (EBE) of the individual parameters were screened with Wilcoxon tests (threshold p=0.1) and then a forward selection was done based on likelihood ratio tests (LRT) with p=0.05. Missing covariates for binary comorbidities were affected to the most represented category. Missing age was affected to the median of the observations and patients with missing sex were remvoed from the dataset.

In the second stage, the selected model was fitted to the data from both treatment arms, adding treatment effects on all parameters except N_0 which should not be affected by a treatment. The treatment was found to have an effect if the global LRT test comparing this full model to the model with only the covariates selected in the first stage was significant with a p-value of less than 0.05.

2.4.3 Evaluation

Visual Predictive Checks (VPC) and individual fits for the final model were checked. VPC were binned specifically to account for the data attrition due to the collection process: patients recovering from SARS-CoV-2 were discharged from the hospital, yet discharged subjects still provided NEWS-2 scores at day 15 and 29. Therefore, the VPC plots were produced by binning the 8 first days and then day 15 and 29.

We compared the RSE obtained with Asympt, SIR, Post, SAEM_MH, MH_after_SAEM and ABC on the full data set and on a random sparse subset of $N = 40$ patients (20 in each arm) stratified on all relevant covariates with the model selected on the complete dataset.

2.5 Implementation

The models were fitted using the `saemix` package in R[16]. For the estimation, we ran 3 chains with $K_1 = 2000$ exploratory iterations and $K_2 = 500$ convergence iterations.

Logit-normal distribution can only be parameterised by [0,1] in `saemix` so in practice, the model used was:

$$log\big(1 + \text{NEWS2}(t)\big) = log\big(1 + 3 * N_0 + A\ \lambda^2\ (t + T_{lag})\ exp(-\lambda(t + T_{lag}))\big)$$

so that $3 * N_0$ follows a logit-normal distribution between 0 and 3.

For the SIR algorithm, we ran $M_S = 1000$ samples and $m_S = 500$ resamples, modifying the code available on the github (`https://github.com/saemixdevelopment/saemixextension/tree/master/SIR`) to account for correlations in the inter-individual variance-covariance matrix.

For Post, we ran 3 chains of 1500 iterations, including 500 warm up iterations not used for standard deviation computation, using Stan [17]. The initial values were the estimates obtained by `saemix` and the prior distribution was Gaussian, centered on rounded versions of the `saemix` estimations for all μ, 0 for all β, a diagnoal matrix of 1 for Ω and 1 for b, with 30% variation coefficient for fixed effects and 50% for treatment effects, variance matrix and error parameter.

For SAEM_MH, we put the same Gaussian prior as in Post. The kernel was also Gaussian, centered on the current MLE with variance equal to the inverse of the current estimated FIM, up to an inflation constant : $q(.) = \mathcal{N}(\hat{\theta}_k, \text{inf} \times \widehat{FIM}_k^{-1})$ with several inflation factors inf=1, 1.5, 2. As $K_2 = 500$, we ran 500 MH chains of length M_k=100.

For MH_after_SAEM, only one chain is drawn after the SAEM algorithm with a two-block Gaussian kernel distribution centered on the previous sample, so that the chain has a random walk behaviour, with the same variance as in SAEM_MH. See Fayette et al. [8] for more details on the kernel distribution.

For ABC, we drew chains of 1000 iterations with a proposal distribution in three blocks, with the same parameters as in the SAEM_MH algorithm. See Fayette et al. [8] for more details on the proposal distribution and distance function used. No inflation was applied to the proposal distribution variance on the full dataset.

On the sparse subset, we ran 10 chains in the SAEM algorithm to account for the reduction in N. The proposal variance in ABC method was also inflated by 2 on the sparse subset.

3 Results

3.1 Effect of remdesivir on the evolution of NEWS-2 scores

Figure 1 shows the distribution of the NEWS-2 score with time, discretised in 4 ad-hoc categories corresponding to worsening clinical status (0-3: healthy, 4-6: mildly ill, 7-10: moderately ill, and 11-20: severely ill). We can observe a substantial amount of missing data, especially in the second half of the study, except at day 15 and day 29 as discharged patients were contacted for follow-up. However the dataset is still very rich and these missing data should not limit the ability to model the score

dynamic, though we need to be mindful of a potential attrition bias. Missing data for NEWS-2 scores was not imputed.

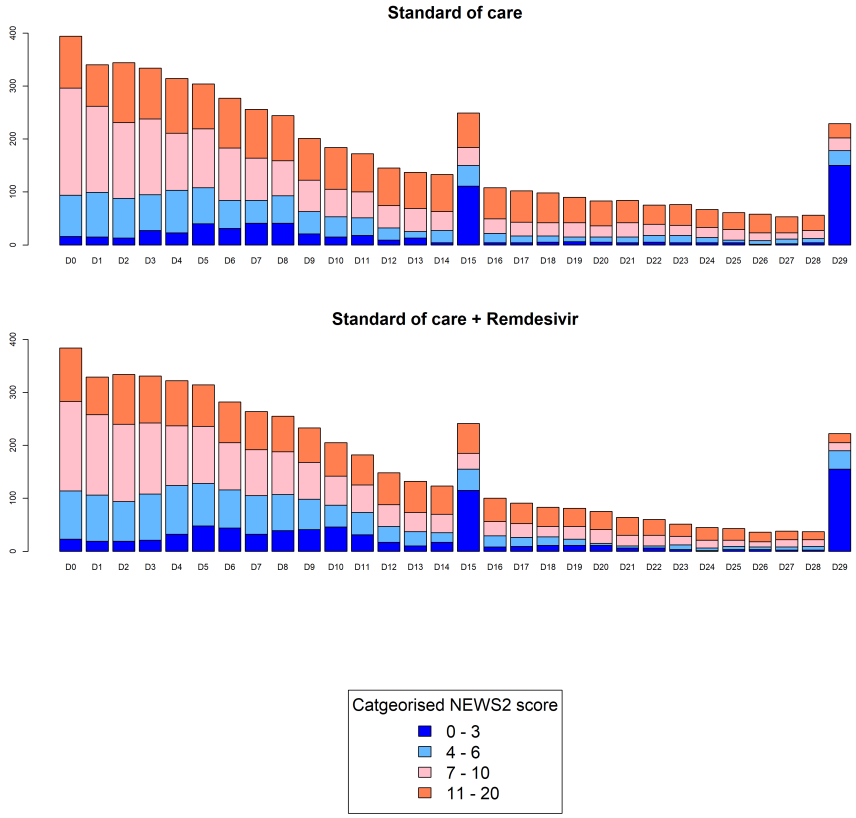


Figure 1: Distribution of NEWS-2 scores, for patients from the standard of care (top) and standard of care +remdesivir (bottom) arms of the Discovery trial, from day of inclusion in the trial to end of study (day 29), discretised in 4 ordered categories

Missing data for comorbidities, which were all binary, were affected to the most represented category. Other covariates tested (age and sex) had no missing data.

Identifiability issues with the standard Bateman model led us to select the form of the Bateman model with identical increase and decrease slope. Parameter estimates all had RSE below 50% on the SoC arm. We estimated very high correlations between the variances of random effects on T_{lag}, A and λ, up to $\rho_{A,\lambda} = 0.98$, as shown in Table 1. Covariate selection led to keeping an effect of age on T_{lag}, A and λ, and an effect of chronic cardiac disease (CCD) and diabetes on N_0.

When applied to the data from both arms, the covariate effect estimates were similar and the higher number of subjects decreased the RSE in most parameters, but the RSE of cardiac disease on N_0 increased to above 50%. The likelihood ratio test (LRT) when adding treatment effects on T_{lag}, A and λ was non-significant (p=0.18) so that we could not conclude to an effect of remdesivir on the evolution of NEWS-2 score. Table 1 shows the parameter estimates and their RSE using the Asympt method obtained with `saemix` at the three key steps of the analysis.

Individual predictions (Figure 2) show that the final model fits well the diversity of profiles ob-

Table 1: Parameter estimates and their relative standard errors (RSE %) of the final model of the evolution of NEWS-2 in the SoC arm after covariate selection (left) , both SoC and SoC+remdesivir arms without treatment effects estimated (center) and both SoC and SoC+remdesivir arms with treatment effects estimated (right)

	SoC arm		SoC and SoC+remdesivir arms			
Parameter	Estimation	RSE (%)	Estimation	RSE (%)	Estimation	RSE (%)
$\mu_{T_{lag}}$ (days)	2.77	11	3.19	8	2.95	10
$\beta^{age}_{T_{lag}}$	0.65	24	0.50	23	0.51	23
$\beta^{T}_{T_{lag}}$					0.15	77
μ_A (days)	102.40	10	108.42	7	108.93	10
β^{age}_{A}	0.71	21	0.57	18	0.57	18
β^{T}_{A}					-0.01	1073
μ_{N_0}	0.13	30	0.16	20	0.17	20
$\beta^{CCD}_{N_0}$	1.20	43	0.19	200	0.22	170
$\beta^{diab}_{N_0}$	1.14	44	1.07	34	0.99	37
μ_λ (days^{-1})	0.21	9	0.20	6	0.20	7
β^{age}_{λ}	-0.69	19	-0.53	16	-0.53	16
β^{T}_{λ}					-0.04	212
$\omega_{T_{lag}}$	1.13	7	1.15	5	1.16	5
ω_A	1.34	4	1.27	3	1.27	3
ω_{N_0}	2.85	8	2.81	6	2.84	6
ω_λ	1.11	5	1.04	3	1.04	3
$\rho_{Tlag,A}$	0.83	3	0.82	3	0.82	3
$\rho_{Tlag,\lambda}$	-0.89	2	-0.89	2	-0.89	2
$\rho_{A,\lambda}$	-0.98	0.4	-0.97	0.3	-0.97	0.3
a	0.24	1	0.25	1	0.25	1

$Tlag$ is the delay between symptom onset and inclusion, 3*N_0 the baseline score value, λ the slope for both exponential increase and decrease, A the area under the NEWS curve and above 3*N_0.
Covariate effect estimates are in grey except treatment effect estimates in red.
$Tlag$, λ and A are log-normal and N_0 is logit-normal. By using the notations of Equation (1), we can write, e.g. for the model with treatment effects fitted on both treatment arms:

$$log(T_{lag,i}) = log(\mu_{T_{lag}}) + \beta^{age}_{T_{lag}} \mathbb{1}^{age_i>65} + \beta^{T}_{T_{lag}} \mathbb{1}^{T_i} + \eta_{T_{lag},i}$$
$$log(A_i) = log(\mu_A) + \beta^{age}_{A} \mathbb{1}^{age_i>65} + \beta^{T}_{A} \mathbb{1}^{T_i} + \eta_{A,i}$$
$$logit(N_{0,i}) = logit(\mu_{N_0}) + \beta^{CCD}_{N_0} \mathbb{1}^{CCD_i} + \beta^{diab}_{N_0} \mathbb{1}^{diab_i} + \eta_{N_0,i}$$
$$log(\lambda_i) = log(\mu_\lambda) + \beta^{age}_{\lambda} \mathbb{1}^{age_i>65} + \beta^{T}_{\lambda} \mathbb{1}^{T_i} + \eta_{\lambda,i}$$

with $\mathbb{1}^{T_i} = 1$ if patient i was in the SoC+remdesivir arm, 0 if he was in the SoC arm, $\mathbb{1}^{age_i>65} = 1$ if patient i was more than 65 years old, 0 otherwise, $\mathbb{1}^{CCD_i} = 1$ if patient i had chronic cardiac disease, 0 otherwise, $\mathbb{1}^{diab_i}$=1 if patient i had diabetes, 0 otherwise. Here $\eta_i = (\eta_{T_{lag},i}, \eta_{A,i}, \eta_{N_0,i}, \eta_{\lambda,i} \simeq \mathcal{N}(0, \Omega)$ with

$$\Omega = \begin{pmatrix} \omega_{Tlag} & \rho_{Tlag,A} & 0 & \rho_{Tlag,\lambda} \\ \rho_{Tlag,A} & \omega_A & 0 & \rho_{A,\lambda} \\ 0 & 0 & \omega_{N_0} & 0 \\ \rho_{Tlag,\lambda} & \rho_{A,\lambda} & 0 & \omega_\lambda \end{pmatrix}$$

All covariates are binary. On log-normal parameters (T_{lag}, λ and A), we interpret a binary covariate effect by multiplying the population value by e^β, e.g. using estimates from the model developed on the SoC arm only, a patient younger than 60 years will have a typical T_{lag} of 2.77 days, and a patient older than 60 years will have a typical T_{lag} of $2.77 \times e^{0.65} = 5.31$ days. On logit-normal parameters like N_0, we compute the covariate effect by adding it to the logit of the parameter and applying the inverse logit function to the result, e.g. using estimates from the model developed on the SoC arm only, a patient with no chronic cardiac disease will have a typical 3*N_0 of 0.39, and a patient with chronic cardiac disease will have a typical 3*N_0 of 1.03.

served. Visual predictive checks (VPC, Figure 3) also show a good overall fit.

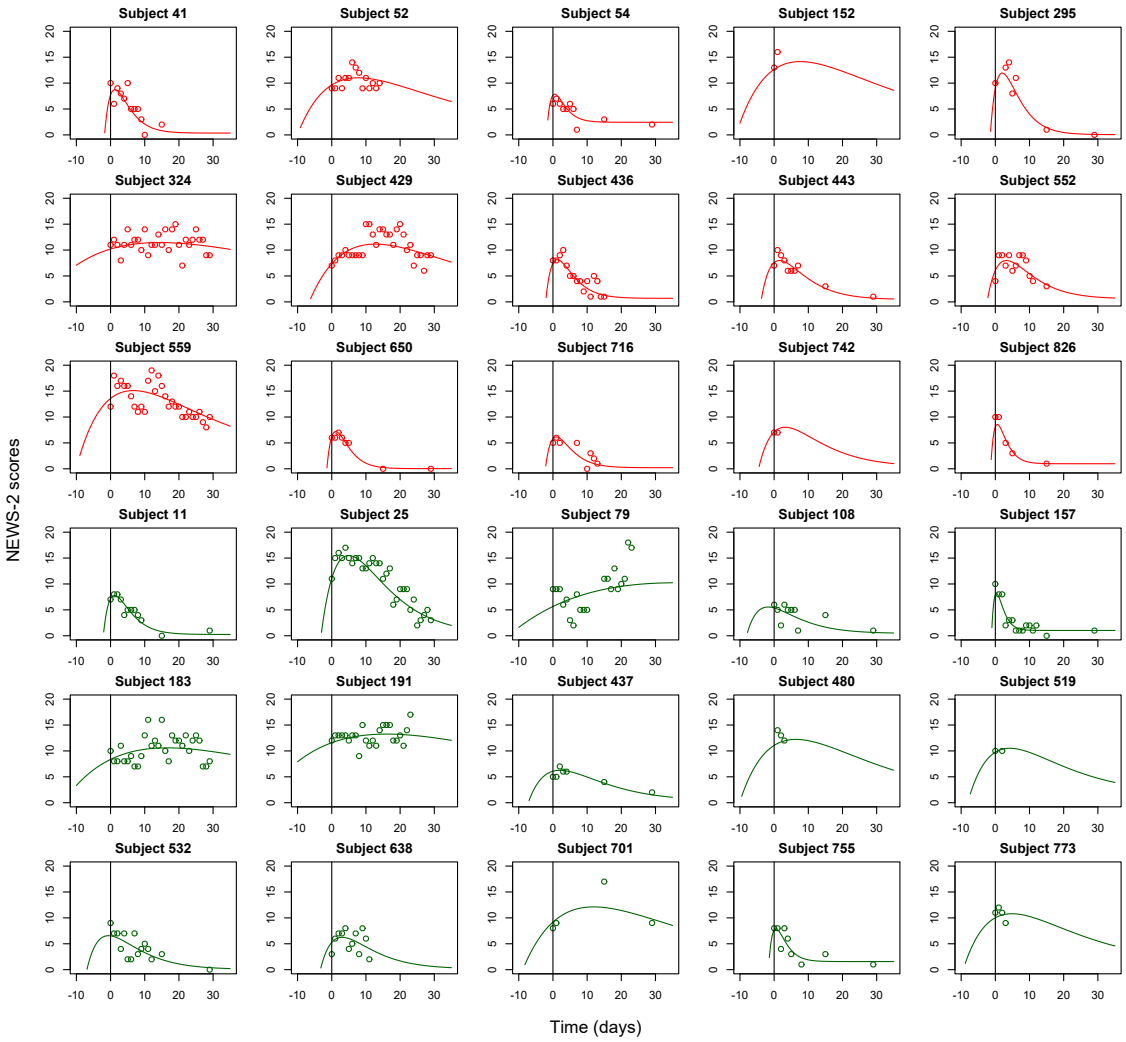


Figure 2: Individual fits of the NEWS-2 trajectories of a random sample of 30 patients using the model with treatment effects fitted on both SoC (red) and SoC+remdesivir (green) arms.
The points represent the observations and the curves represent the model individual predictions computed in `saemix` at the end of the SAEM algorithm using empirical Bayes estimates (EBE) of the individual parameters. The vertical lines represents the time at inclusion (t=0 in the model).

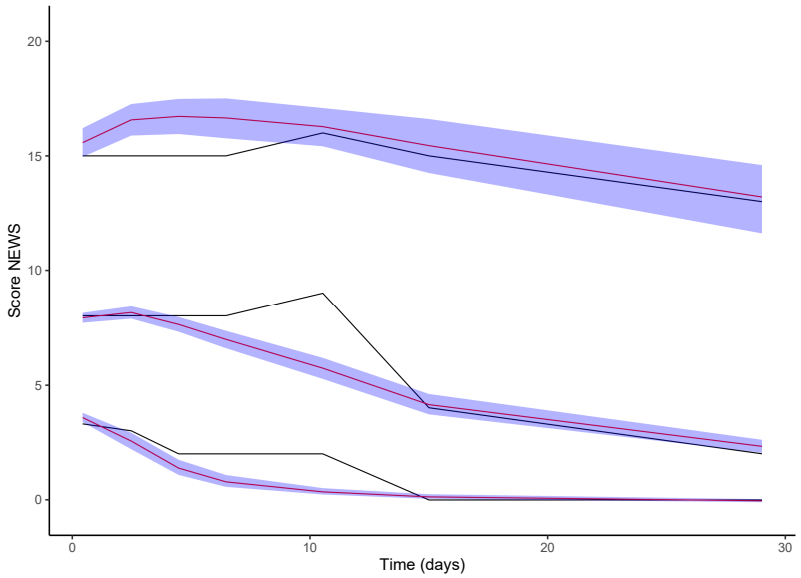


Figure 3: Visual predictive check of the model with treatment effects fitted on both arms
The predicted 5%, 50% and 95% percentiles are shown as red lines and their corresponding 90% prediction interval as blue areas; the observed percentiles as black solid lines. 1000 replicates were simulated; data from day 16 to 28 was not used to compute observed percentiles due to bias induced by the discharge of recovering patients, and the data kept was binned every two days except for day 15 and 29 when the data is more informative.

3.2 SE computation on the modelling of NEWS-2 scores

We applied the different methods to compute the SE on the parameter estimates of the model with treatment effects fitted on both arms (Table 2). The results are represented as radar plots in Figure 4.

On this design, all methods were highly concordant on all parameters except the covariate effects that had RSE above 50%. This confirms that we are in an asymptotic regime. Of note, the graphical diagnostic of the SIR algorithm (dOFV distribution plot [12]) did not give satisfying results even though RSE are concordant with other methods. The Post diagnostic tool $\hat{R}$ was at 1.02 which is acceptable. The median acceptation rate for SAEM_MH was at 6% versus 1.4% for MH_after_SAEM, both being very low. We tried inflating the kernel variance which only lowered the acceptation rates further (not shown). ABC acceptation rate was around 60% when choosing $\epsilon = 50$ which was done in the results shown. Of note, choosing $\epsilon = 20$ lowered the acceptation rate to 25% but did not change the RSE estimated (not shown). Figures used to obtain the radar plots in Figure 4 are shown in Table 2 in Appendix.

We also applied the different SE computation methods on a sparse subset of 40 patients, stratified on treatment and age category, using the model selected on the full dataset and refitted on the sparse subset of the data. This time, we could see discrepancies between the RSE (Table 3 and Figure 5), which confirmed that we were no longer in an asymptotic regime. As expected, the reduction of N induced an increase of RSE. SIR had good diagnostic plots and Post had a $\hat{R}$ lower than 1.02. However, SAEM_MH and MH_after_SAEM acceptation rates were at 4 and 1.6% respectively ABC presented acceptation rates at 25% which is better and appropriate to compute RSE.

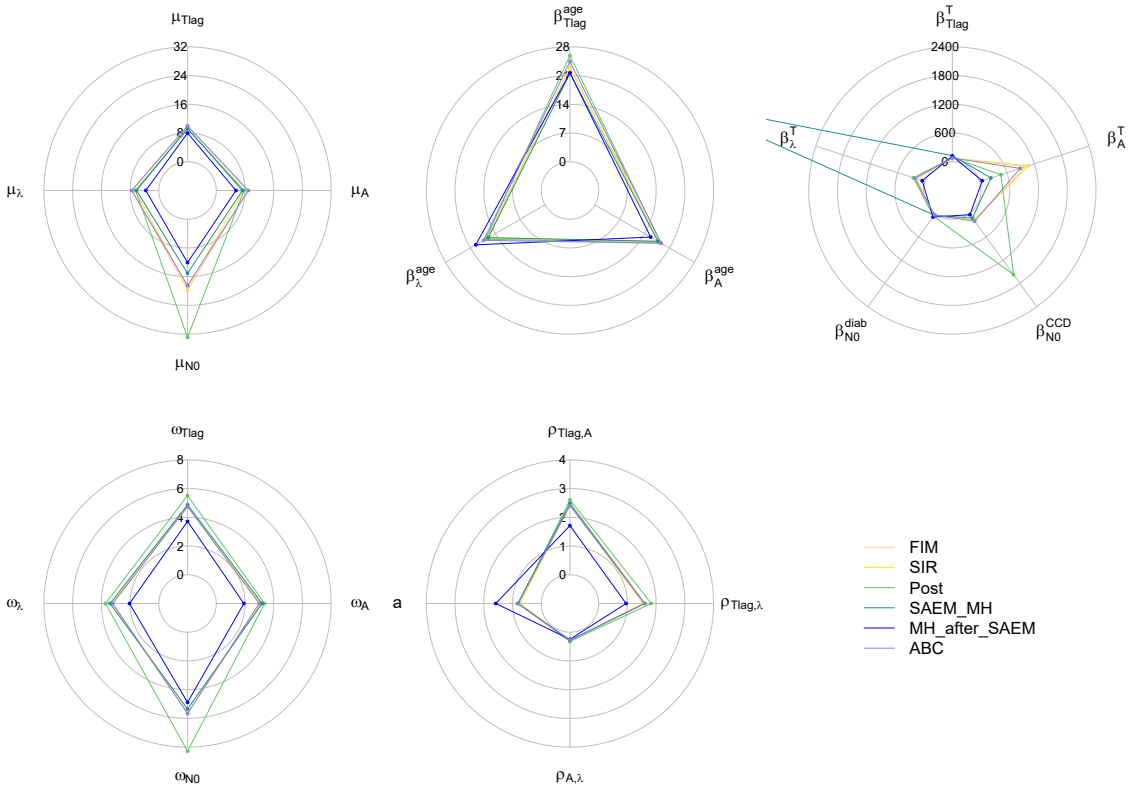


Figure 4: Radar plots of the RSE of the parameters from the model of NEWS-2 trajectory with treatment effects fitted to the SoC and SoC+remdesivir arms (N=832) using Asympt, SIR, Post, SAEM_MH, MH_after_SAEM and ABC

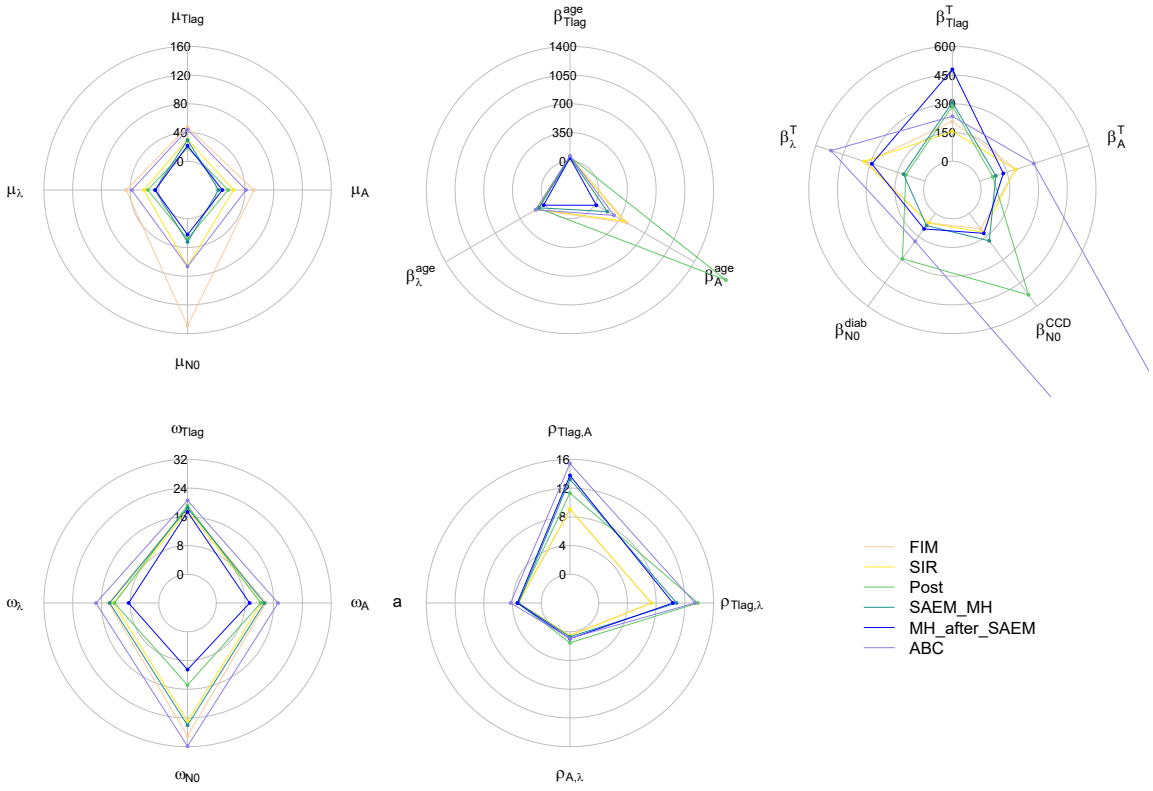


Figure 5: Radar plots of RSE of parameters from the model of NEWS-2 trajectory with treatment effects fitted to the SoC and SoC+remdesivir arms (N=40) using Asympt, SIR, Post, SAEM_MH, MH_after_SAEM and ABC

4 Discussion

In this work we modelled the evolution of the NEWS-2 clinical score using longitudinal data collected in patients hospitalised for SARS-CoV-2 and included in the Discovery trial. The first goal of the modelling was to test for an effect of remdesivir on the evolution of the clinical status of the patients. The second objective of this work was to compare different methods of SE computation in this context. We applied the methods of interest on the full data and on a more sparse and challenging subset.

The results showed that the full dataset satisfied the asymptotic assumption, with frequentist methods giving RSE concordant with semi-Bayesian methods, but the subset that we also analysed was departing from the asymptotic. In both cases, the MH algorithm was too sensitive to be able to sample in the posterior distribution correctly, showing low acceptation rates. This was probably due to the high dimension of the parameter vector to estimate, as well as the complexity of the variability matrix, despite the high number of subjects. A previous study on data with similar characteristics, i.e. inter-individual variabilities up to 100% and a correlation of 0.76 [14], and a sparse simulation study inspired from this data, simulating 3 observations for 12 subjects with variabilities up to 150% and correlations up to 0.98 [8], also illustrated this limit.

The semi-Bayesian method based on the HMC algorithm performed adequately on this dataset, even though the acceptation rates were very low and these methods has already shown limits in the presence of correlations in the random effects[8]. The ABC algorithm does not have a proper diagnostic tool to determine the reliability of its results on a real dataset and this needs to be addressed with further investigations.

When modelling the NEWS-2 score trajectories, we kept the same method of covariate selection as previously done in a study investigating the link between NEWS-2 score evolution with viral dynamics on the same data by Néant et al. [5]. In that study, which investigated the same set of covariates, only age was found to significantly slow down the decrease of NEWS-2 : the decrease rate of NEWS-2 after viral load peak was reduced by 55% in older patients in this model. We found a similar effect plus longer delay before inclusion, slower increase and higher AUC in older patients: according to the model fitted on both treatment arms without treatment effects estimated, the delay between symptom onset and inclusion was multiplied by 1.7 in patients over 65 years, the increase and decrease rate λ reduced by 41% and the relative area under the score curve multiplied by 1.8, which means that the progression was slower in older patients but the global burden of symptoms was higher. Chronic cardiac disease and diabetes were also found to higher the baseline score, which was expected: according to the same model, the baseline score N_0 was multiplied by 1.2 in patients with chronic cardiac disease, by 2.2 in diabetic patients, and by 2.5 in patients having both comorbidities.

If previous studies could show an effect of remdesivir on the viral load dynamics [4] and a link between viral loads and NEWS-2 score trajectories [5], no treatment effect was found on the viral load-NEWS-2 score association parameter. As Néant et al., we could not show a treatment effect directly on the NEWS2 clinical score trajectories. We also explored the consistency of our strategy to select covariate effects and test for a treatment effect, comparing the results of our LRT-based forward selection to a Wald test-based forward selection. We ended up with almost the same selection: all effects presented here are conserved with an additional effect of smoking status on N_0. Further, the multivariate Wald test for treatment effects on T_{lag}, A and λ was also non-significant (p=0.14).

Here, we considered the NEWS-2 score as a continuous Gaussian variable, given its high number of categories. This approximation enabled us to compute the FIM by linearisation in SAEM. However, this assumption has some limitations, notably the model used is bounded at 0 but can predict NEWS-2 scores higher than 20; an IRT or bounded integer model [18] could be used instead. Given that the patients left the hospital when their symptoms improved enough, we checked that the dropout, discharge and death did not affect the parameter estimates using a simple joint model with an exponential survival risk function and the predicted current value of NEWS2 as link function in Monolix software (not shown). This absence of attrition bias can be attributed to the data design of the Discovery trial which collected data for most patients at days 15 and 29 and therefore allowed for a robust estimation.

Our short term perspective is to adapt the semi-Bayesian methods to categorical data as the so-called WHO score, a composite clinical score measuring clinical improvement and survival which was developed and recommended by the WHO to be used as primary endpoint in SARS-CoV-2 clinical trials [19], and which was indeed used as primary endpoint in the Discovery trial. To model this data, we could use different structural models including proportional odds model, multistate models and bounded integer model, with different time effect functions. If the variability structure observed in the NEWS-2 scores is as complex in the WHO scores, it might be more difficult to model as discrete models are more sensitive to random effects on several parameters. Moreover, it is not possible for categorical models to compute the FIM by linearisation, so we could use a stochastic approximation method that was presented in Delattre and Kuhn[20] and applied to joint models in Lavalley et al.[21]. Methods implying repeated computation of the likelihood, as SIR and SAEM_MH, might be challenged by the computational burden of likelihood computation in discrete models, that has to be done by importance sampling. The ABC algorithm does not suffer from that burden, but we would need to define a suitable distance function taking into account the discrete nature of the score.

5 Conclusion

The evolution of NEWS-2 clinical scores collected in the Discovery trial was successfully modelled using a continuous model and allowed to correctly fit various profiles. No effect of remdesivir could be shown on the evolution of the NEWS2 score. This analysis illustrated the limits of SAEM_MH to compute parameters SE on complex data. Further developments are needed to adapt the semi-Bayesian methods for categorical models.

6 Acknowledgements

We would like to thank the Discovery group for sharing the application data.

7 Data availability

The real data used in the application was made available to us under license according to the data sharing policy of the Discovery trial. Requests for access should be made to the Discovery group.

8 Conflicts of interest

The authors declare no conflict of interest.

9 References

References

[1] Florence Ader. Protocol for the DisCoVeRy trial: multicentre, adaptive, randomised trial of the safety and efficacy of treatments for COVID-19 in hospitalised adults. *BMJ Open*, 10(9):e041437, 2020.

[2] Florence Ader, Nathan Peiffer-Smadja, Julien Poissy, Maude Bouscambert-Duchamp, Drifa Belhadi, Alpha Diallo, Christelle Delmas, Juliette Saillard, Aline Dechanet, Noémie Mercier, Axelle Dupont, Toni Alfaiate, François-Xavier Lescure, François Raffi, François Goehringer, Antoine Kimmoun, Stéphane Jaureguiberry, Jean Reignier, Saad Nseir, François Danion, Raphael Clere-Jehl, Kévin Bouiller, Jean-Christophe Navellou, Violaine Tolsma, André Cabié, Clément

Dubost, Johan Courjon, Sylvie Leroy, Joy Mootien, Rostane Gaci, Bruno Mourvillier, Emmanuel Faure, Valérie Pourcher, Sébastien Gallien, Odile Launay, Karine Lacombe, Jean-Philippe Lanoix, Alain Makinson, Guillaume Martin-Blondel, Lila Bouadma, Elisabeth Botelho-Nevers, Amandine Gagneux-Brunon, Olivier Epaulard, Lionel Piroth, Florent Wallet, Jean-Christophe Richard, Jean Reuter, Thérèse Staub, Bruno Lina, Marion Noret, Claire Andrejak, Minh Patrick Lê, Gilles Peytavin, Maya Hites, Dominique Costagliola, Yazdan Yazdanpanah, Charles Burdet, and France Mentré. An open-label randomized controlled trial of the effect of lopinavir/ritonavir, lopinavir/ritonavir plus IFN-β-1a and hydroxychloroquine in hospitalized patients with COVID-19. *Clinical Microbiology and Infection*, 27(12):1826–1837, 2021.

[3] Florence Ader, Maude Bouscambert-Duchamp, Maya Hites, Nathan Peiffer-Smadja, Julien Poissy, Drifa Belhadi, Alpha Diallo, Minh-Patrick Lê, Gilles Peytavin, Thérèse Staub, Richard Greil, Jérémie Guedj, José-Artur Paiva, Dominique Costagliola, Yazdan Yazdanpanah, Charles Burdet, France Mentré, and the Discovery Study Group. Remdesivir plus standard of care versus standard of care alone for the treatment of patients admitted to hospital with COVID-19 (DisCoVeRy): a phase 3, randomised, controlled, open-label trial. *The Lancet Infectious Diseases*, 22(2):209–221, 2022.

[4] Guillaume Lingas, Nadège Néant, Alexandre Gaymard, Drifa Belhadi, Gilles Peytavin, Maya Hites, Thérèse Staub, Richard Greil, Jose-Artur Paiva, Julien Poissy, Nathan Peiffer-Smadja, Dominique Costagliola, Yazdan Yazdanpanah, Florent Wallet, Amandine Gagneux-Brunon, France Mentré, Florence Ader, Charles Burdet, Jérémie Guedj, and Maude Bouscambert-Duchamp. Effect of remdesivir on viral dynamics in COVID-19 hospitalized patients: a modelling analysis of the randomized, controlled, open-label DisCoVeRy trial. *Journal of Antimicrobial Chemotherapy*, 77(5):1404–1412, 2022.

[5] Nadège Néant, Guillaume Lingas, Alexandre Gaymard, Drifa Belhadi, Maya Hites, Thérèse Staub, Richard Greil, Jose-Artur Paiva, Julien Poissy, Nathan Peiffer-Smadja, Dominique Costagliola, Yazdan Yazdanpanah, Maude Bouscambert-Duchamp, Amandine Gagneux-Brunon, Florence Ader, France Mentré, Florent Wallet, Charles Burdet, Jérémie Guedj, and the DisCoVeRy study group. Association between SARS-CoV-2 viral kinetics and clinical score evolution in hospitalized patients. *CPT: Pharmacometrics & Systems Pharmacology*, 2023.

[6] Florence Loingeville, Julie Bertrand, Thu Thuy Nguyen, Satish Sharan, Kairui Feng, Wanjie Sun, Jing Han, Stella Grosser, Liang Zhao, Lanyan Fang, Kathrin Möllenhoff, Holger Dette, and France Mentré. New model-based bioequivalence statistical approaches for pharmacokinetic studies with sparse sampling. *The AAPS Journal*, 22(6):141, 2020.

[7] Mélanie Guhl, Lucie Fayette, François Mercier, Julie Bertrand, and Emmanuelle Comets. Uncertainty computation at finite distance in nonlinear mixed effects models - a new method based on metropolis hastings algorithm. *The AAPS Journal*, (in press).

[8] Lucie Fayette, Mélanie Guhl, Jérémie Guedj, Julie Bertrand, and Emmanuelle Comets. Development of a semi-bayesian saem algorithm for finite-distance estimation of parameter uncertainty in non-linear mixed-effects models. *to be submitted in Statistics in Medicine*, -.

[9] Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.

[10] Marc Lavielle. *Mixed effects models for the population approach: models, tasks, methods and tools.* Chapman and Hall/CRC, 2014.

[11] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of EM algorithm. *Annals of Statistics*, 27(1):94–128, 1999.

[12] Anne-Gaëlle Dosne, Martin Bergstrand, Kajsa Harling, and Mats O Karlsson. Improving the estimation of parameter uncertainty distributions in nonlinear mixed effects models using sampling importance. *Journal of Pharmacokinetics and Pharmacodynamics*, 43(6):583–596, 2016.

[13] Sebastian Ueckert, Marie Karelle Rivière, and France Mentré. Alternative to resampling methods in maximum likelihood estimation for NLMEM by borrowing from Bayesian methodology. *PAGE 24*, 2015. `https://www.page-meeting.org/?abstract=3632`.

[14] Mélanie Guhl, Lucie Fayette, François Mercier, Julie Bertrand, and Emmanuelle Comets. Uncertainty computation at finite distance in nonlinear mixed effects models - a new method based on metropolis hastings algorithm. *The AAPS Journal*, (in press).

[15] Edward R. Garrett. The Bateman function revisited: A critical reevaluation of the quantitative expressions to characterize concentrations in the one compartment body model as a function of time with first-order invasion and first-order elimination. *Journal of Pharmacokinetics and Biopharmaceutics*, 22(2):103–128, 1994.

[16] Emmanuelle Comets, Audrey Lavenu, and Marc Lavielle. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software*, 80:1–41, 2017.

[17] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. *Stan* : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.

[18] Gustaf J Wellhagen, Maria C Kjellson, and Mats O Karlsson. A bounded integer model for rating and composite scale data. *The AAPS Journal*, 21(4):74, 2019.

[19] World Health Organisation. Covid-19 therapeutic trial synopsis. Technical report, 2020. `https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis`.

[20] Maud Delattre and Estelle Kuhn. Estimating Fisher Information Matrix in Latent Variable Models based on the Score Function. *Publisher: arXiv, Version Number: 2*, 2019.

[21] Alexandra Lavalley-Morelle, France Mentré, Emmanuelle Comets, and Jimmy Mullaert. Extending the code in the open-source *saemix* package to fit joint models. *Computer Methods and Programs in Biomedicine (Accepted)*, 2023.

10 Appendix

Parameter	FIM	SIR	Post	SAEM_MH	MH_after_SAEM	ABC
$\mu_{T_{lag}}$ (days)	10	9	9	9	8	10
$\beta^{age}_{T_{lag}}$	23	23	26	22	22	24
$\beta^{T}_{T_{lag}}$	77	71	87	128	116	78
μ_A (days)	9	9	8	7	5	9
β^{age}_{A}	18	18	18	18	16	19
β^{T}_{A}	1073	1012	378	247	57	890
μ_{N_0}	20	19	32	15	12	18
$\beta^{CCD}_{N_0}$	170	165	1792	115	24	194
$\beta^{diab}_{N_0}$	37	38	65	38	87	35
μ_λ (days^{-1})	7	7	6	6	4	8
β^{age}_{λ}	16	16	16	16	19	17
β^{T}_{λ}	212	193	289	5280	62	223
$\omega_{T_{lag}}$	5	5	5	5	4	5
ω_A	3	3	3	3	2	3
ω_{N_0}	6	6	8	5	5	6
ω_λ	3	3	4	3	2	3
$\rho_{T_{lag},A}$	3	2	2	2	2	2
$\rho_{T_{lag},\lambda}$	2	2	3	2	1	2
$\rho_{A,\lambda}$	0.3	0.3	0.3	0.3	0.2	0.3
a	1	1	1	1	2	1

Table 2: Relative standard errors of parameters from the model of NEWS-2 trajectory with treatment effects fitted to the SoC and SoC+remdesivir arms (N=832) using Asympt, SIR, Post, SAEM_MH, MH_after_SAEM and ABC

Parameter	FIM	SIR	Post	SAEM_MH	MH_after_SAEM	ABC
$\mu_{T_{lag}}$ (days)	48	31	19	30	22	44
$\beta^{age}_{T_{lag}}$	57	56	55	49	41	66
$\beta^{T}_{T_{lag}}$	208	156	287	308	480	235
μ_A (days)	52	24	16	3	9	41
β^{age}_{A}	448	402	1844	177	22	269
β^{T}_{A}	199	196	72	89	130	297
μ_{N_0}	149	66	28	32	22	67
$\beta^{CCD}_{N_0}$	100	123	527	178	130	3452
$\beta^{diab}_{N_0}$	62	63	294	80	101	181
μ_λ (days^{-1})	45	21	15	5	5	38
β^{age}_{λ}	127	118	94	85	20	132
β^{T}_{λ}	322	335	105	118	292	516
$\omega_{T_{lag}}$	19	18	19	18	17	21
ω_A	14	13	12	13	9	17
ω_{N_0}	29	25	15	26	11	32
ω_λ	14	13	12	14	8	17
$\rho_{T_{lag},A}$	7	7	14	13	14	13
$\rho_{T_{lag},\lambda}$	9	9	11	11	10	15
$\rho_{A,\lambda}$	0.5	0.5	2	0.7	1	1
a	3	3	3	3	3	4

Table 3: Relative standard errors of parameters from the model of NEWS-2 trajectory with treatment effects fitted to the SoC and SoC+remdesivir arms (N=40), with Asympt, SIR, Post, SAEM_MH, MH_after_SAEM and ABC

Chapitre 6

Discussion et conclusion

L'enjeu principal de cette thèse était le développement de méthodes de calcul de l'incertitude robustes pour les NLMEM, en particulier dans le cas d'un éloignement aux conditions asymptotiques, c'est-à-dire pour les essais cliniques pour lesquels on a peu de sujets et/ou peu d'observations par sujets. Les études comportant peu d'observations par sujet concernent généralement des patients fragiles (essais néonataux, pédiatriques, gériatriques) ou des essais impliquant le prélèvement de tissus (essais ophtalmiques, concentrations dans le tissu synovial) (Aarons, 1993). Les études comportant peu de patients concernent généralement des maladies rares ou pour lesquelles les traitements sont personnalisés. Cela concerne aussi les essais cliniques arrêtés prématurément en raison d'un manque de recrutement : en général, les analyses prévues dans le protocole n'anticipent pas ce cas de figure et c'est l'argument éthique qui nous pousse à essayer de tirer de l'information de cet essai malgré son arrêt prématuré (Billingham et al., 2012).

Dans tous ces cas de figure, lorsque nous cherchons à mettre en évidence l'effet d'un traitement, ou d'une autre covariable, le manque d'information dans les données peut engendrer une sous-estimation de l'incertitude autour de l'effet estimé, et donc une augmentation des erreurs liées aux tests effectués. C'est pourquoi le développement de méthodes robustes pour calculer l'incertitude dans ce contexte est primordial.

6.1 Discussion

Le premier travail de cette thèse a permis de mettre en évidence un des intérêts de développer des méthodes robustes de calcul de l'incertitude dans les NLMEM : jusqu'à présent, les protocoles statistiques font plutôt confiance à des méthodes plus classiques, nécessitant moins d'hypothèses mais plus de données. En effet, dans le cadre de la bioéquivalence, les agences de régulation telles que la FDA n'ont pour l'instant pas assez confiance dans les méthodes par modélisation existantes pour les données longitudinales et recommandent des méthodes sans modélisation avec des données riches (12 à 18 prélèvements par patient), ce qui alourdit la mise en place d'essais cliniques. Développer des méthodes robustes de calcul des SE dans les NLMEM permettrait de valider leur utilisation dans ce contexte, ce qui reste très marginal pour le moment. Cet enjeu a également été mis en évidence avec les données longitudinales de score de l'essai Discovery, étudiées dans le dernier projet de cette thèse, même si la Covid19 n'est pas un cas où la richesse des données est une limite à la mise en place d'essais cliniques. En effet, il n'a pas été envisagé dans les recommandations de l'OMS (World Health Organisation, 2020) de modéliser des données longitudinales pour en inférer le critère de jugement principal, les dynamiques de score clinique et de charge virale restant limitées aux critères secondaires (Ader, 2020). La mise en place de projets de recherche pour développer, évaluer et améliorer les méthodes utilisées dans les NLMEM permet de promouvoir leur utilisation dans les protocoles statistiques et de faire évoluer les recommandations des agences de régulation. Dans le cas de la bioéquivalence par exemple, la FDA a mis à jour ses recommandations (U.S. Food and Drug Administration, 2022b) afin de communiquer sur leur acceptation de l'utilisation des NLMEM pour les données éparses suite aux résultats mis en évidence dans le cadre d'une collaboration avec des équipes de recherche académique, dont fait partie le premier projet de cette thèse.

De précédents travaux ont mis en évidence que les méthodes classiques de calcul des SE dans les NLMEM, basés sur la FIM et une hypothèse asymptotique, ne sont pas robustes au design épars (Panhard and Mentré, 2005; Bertrand et al., 2009; Dubois et al., 2010, 2011; Loingeville et al., 2020). La collaboration entre la FDA et des équipes académiques a permis de mettre

en évidence de bons résultats empiriques pour des méthodes semi-Bayésiennes de calcul de l'incertitude pour les données éparses (Ueckert et al., 2015; Loingeville et al., 2020), ce qui confirme l'intérêt théorique de l'approximation de la distribution de l'EMV par une distribution Bayésienne *a posteriori*. En effet, ces méthodes présentent l'avantage, en comparaison aux méthodes classiques fréquentistes, de s'affranchir de l'hypothèse asymptotique et de construire des intervalles de confiance sans faire d'hypothèse de loi sur la distribution des paramètres. Dans ces travaux, une méthode semi-Bayésienne basée sur l'algorithme HMC a montré de bons résultats sur des simulations éparses obtenues à partir d'un modèle pharmacocinétique à un compartiment avec un design croisé.

Les différents travaux de cette thèse ont permis de confirmer le potentiel des méthodes semi-Bayésiennes, cette fois-ci évaluées sur des simulations avec un design parallèle. La méthode basée sur l'algorithme HMC permet une bonne estimation des SE et un contrôle des erreurs de type I dans les tests de bioéquivalence pour un modèle relativement simple. Elle est toutefois mise en défaut en présence de fortes corrélations dans la matrice de variabilité inter-individuelle. Nous avons également constaté que dans le cas où les méthodes semi-Bayésiennes donnent des résultats satisfaisants, l'algorithme HMC est mis un peu plus en difficulté lorsqu'il est utilisé seul : en effet, les outils diagnostiques disponibles nous ont montré que les critères de convergence sont moins souvent atteints lorsque toute l'inférence est faite avec une approche Bayésienne, ce qui est potentiellement dû à la sensibilité de ce type d'algorithme aux valeurs initiales utilisées. Cette constatation est un argument en faveur de l'utilisation de méthodes semi-Bayésiennes, qui combinent la rapidité de convergence de l'algorithme SAEM avec la capacité des méthodes Bayésiennes à échantillonner autour d'une valeur optimale en explorant correctement l'espace des paramètres.

Le second projet propose une nouvelle méthode semi-Bayésienne appelée SAEM_MH basée sur l'algorithme MH qui présente l'avantage, par comparaison à la méthode semi-Bayésienne basée sur l'algorithme HMC déjà implémentée dans un logiciel spécifique, d'être intégrée à l'algorithme SAEM et implémentée dans le package `saemix`. En effet, l'algorithme MH est utilisé dans l'algorithme SAEM au moment de l'estimation afin d'échantillonner dans la loi des paramètres individuels, conditionnellement aux données et aux paramètres de population

estimés. Cependant, elle présente également certaines limites : la relative simplicité de son implémentation par rapport à HMC est contre-balancée par sa lenteur computationnelle. En effet, l'algorithme MH se comporte par définition comme une marche aléatoire, il est donc plus lourd en calculs que l'algorithme HMC qui optimise les sauts entre chaque itération. Par ailleurs, l'algorithme MH est plus sensible à la dimension du vecteur des paramètres d'intérêt que l'algorithme HMC, et ce projet nous a montré que la méthode SAEM_MH donnait des résultats satisfaisants à distance finie sur des modèles relativement simples, avec un vecteur de paramètres de population de petite dimension, mais était également mise en défaut par des modèles plus complexes qui sont également plus susceptibles d'être utilisés en pratique.

Dans le troisième travail, différentes pistes d'amélioration de la méthode SAEM_MH ont été explorées, afin de réduire la dimension des paramètres à échantillonner et/ou de calibrer l'algorithme de façon dynamique pour adapter la méthode au cas rencontré, et ainsi de dépasser les limites constatées lors de l'évaluation initiale de la méthode, mais aucune variante ne semble surpasser les difficultés liées à la complexité de la structure de la matrice de variabilité inter-individuelle. Ces résultats sont cependant à relativiser au vu de l'extrême complexité des données et du modèle simulés : en effet, les différences constatées entre les différentes méthodes sont à mettre en perspective avec les RSE empiriques calculées dans l'étude de simulation qui sont très élevées. Dans ce travail nous avons également implémenté un autre algorithme Bayésien, ABC, dont l'avantage principal est de comparer directement les observations à des simulations obtenues à partir d'échantillons des paramètres. Ainsi, il n'est plus nécessaire de calculer la vraisemblance, puisqu'une statistique synthétique est définie, ainsi qu'une fonction de distance adéquate, afin de mesurer la proximité des données simulées par rapport aux données observées. Cette méthode a montré des résultats prometteurs sur les différentes études de simulation et de données réelles de cette thèse. Cependant, les définitions de la statistique synthétique, de la fonction de distance et du seuil optimal pour accepter les tirages sont peu documentées dans la littérature et potentiellement difficile à déterminer en pratique.

Dans le dernier projet de cette thèse, la modélisation effectuée sur les données de score longitudinales de l'essai Discovery nous a conduit une fois encore à estimer des structures de

variabilité complexes sur des données réelles. Même si les données complètes sont riches et donc les méthodes discutées dans cette thèse y sont peu appropriées, l'utilisation d'un sous-ensemble des données nous a donné des résultats similaires à ceux obtenus sur les données Gantenerumab. On peut également noter que le modèle structurel utilisé pour les données NEWS-2, considérées comme continues, est proche des modèles utilisés en pharmacocinétique dans les travaux précédents, puisqu'ils sont tous dérivés de la fonction de Bateman.

La modélisation des données OMS, considérées comme catégorielles, permettrait d'étendre les méthodes développées à ce type de données, fréquemment utilisées en recherche clinique. Lorsque l'on traite de données discrètes longitudinales, les approches disponibles, aussi bien fréquentistes que Bayésiennes, ont plus de difficulté à converger vers une estimation stable, car elles sont plus appropriées à des modèles continus. Le calcul de la vraisemblance est lui aussi plus lourd en calculs qu'avec des modèles continus, ce qui rend toutes les méthodes de calcul de l'incertitude basée sur la vraisemblance difficiles à mettre en oeuvre. Le calcul de la FIM par linéarisation n'étant valable que pour les modèles gaussiens, l'incertitude pourrait dans ce cas être obtenue à partir de la FIM calculée par approximation stochastique. La méthode ABC parait également une alternative intéressante dans ce cas, sous condition de pouvoir définir une statistique synthétique et une fonction de distance adéquate.

6.2 Perspectives

La perspective immédiate de cette thèse est l'extension des méthodes semi-Bayésiennes pour le calcul de l'incertitude dans le cas des modèles catégoriels.

Par ailleurs, la méthode SAEM_MH a montré des améliorations de la prise en compte de l'incertitude sur certains designs, mais pour pouvoir l'utiliser en pratique, il faudrait disposer d'un outil diagnostique permettant de détecter les cas sur lesquels la méthode est mise en défaut. Le taux d'acceptation, lorsqu'il est trop proche de 0, permet d'identifier les cas extrêmes, et la cohérence entre les méthodes fréquentistes et semi-Bayésiennes permet d'identifier les cas asymptotiques, mais il est nécessaire de développer un outil qui puisse être fiable dans les cas moins évidents. Une perspective serait d'évaluer plus exhaustivement les outils diag-

nostiques utilisés dans les approches Bayésiennes, notamment la comparaison des variances inter et intra-chaînes (Vehtari et al., 2021) mais également les outils basés sur des simulations permettant de diagnostiquer les distributions *a posteriori* obtenues (Gelman et al., 1996).

Certains éléments permettant de diagnostiquer un algorithme HMC, l'échelle de réduction potentielle $\hat{R}$ et la taille d'échantillon effective ESS (Gelman et al., 2013), peuvent aussi être appliqués à un algorithme MH. Dans la méthode SAEM_MH, il n'est cependant pas possible de calculer un $\hat{R}$ car une seule chaîne de valeurs est obtenue. Une perspective serait de vérifier les $\hat{R}$ entre les différentes chaînes échantillonnées dans la méthode. Pour ce qui est de l'ESS, il est techniquement possible de le calculer mais pas forcément pertinent, étant donné la structure de notre méthode : une seule valeur de chaque chaîne échantillonnée est conservée donc le problème de l'autocorrélation ne se pose pas. On pourrait également garder plusieurs échantillons par chaînes. Pour vérifier que le nombre d'échantillons est assez élevé pour estimer la déviation standard de la chaîne, on peut aussi tracer des plots de convergence de la déviation standard. L'existence de taux d'acceptation nous met à disposition un outil de diagnostic supplémentaire par rapport à l'algorithme HMC qui par construction a un taux d'acceptation de 100%.

L'utilisation de noyaux gaussiens peut aussi être remise en cause et une fois encore on peut s'inspirer des méthodes développées dans l'inférence Bayésienne : par exemple la loi *a priori* dite en fer à cheval ("horseshoe prior", Piironen and Vehtari (2017)) a été développée dans le cadre de modèles à haute dimension avec corrélations et pourrait s'appliquer à la méthode SAEM_MH dans le cas de modèles avec structure de variabilité complexe.

6.3 Conclusion générale

Pour améliorer la prise en compte de l'incertitude dans les NLMEM, une piste prometteuse est de combiner les approches fréquentiste et Bayésienne afin de pallier les limites des méthodes classiques et d'obtenir une méthode robuste aux différentes structures de données et de modèles que l'on peut rencontrer en pratique. En particulier, au cours de ce travail, l'exploration de l'algorithme de Metropolis-Hastings en combinaison avec l'algorithme SAEM a montré,

dans une certaine mesure, une amélioration du calcul de l'incertitude pour les modèles continus, mais également des limites dues à la complexité des modèles utilisés et à la dimension des vecteurs de paramètres à estimer, alors que l'algorithme ABC se montre plus robuste. De plus amples recherches sont nécessaires afin de calibrer cet algorithme et de l'adapter à des données catégorielles.

Bibliographie

L. Aarons. Sparse data analysis. European Journal of Drug Metabolism and Pharmacokinetics, 18(1) :97–100, Mar. 1993. doi : 10.1007/BF03220012.

F. Ader. Protocol for the DisCoVeRy trial : multicentre, adaptive, randomised trial of the safety and efficacy of treatments for COVID-19 in hospitalised adults. BMJ Open, 10(9) : e041437, 2020. doi : 10.1136/bmjopen-2020-041437.

F. Ader, N. Peiffer-Smadja, J. Poissy, M. Bouscambert-Duchamp, D. Belhadi, A. Diallo, C. Delmas, J. Saillard, A. Dechanet, N. Mercier, A. Dupont, T. Alfaiate, F.-X. Lescure, F. Raffi, F. Goehringer, A. Kimmoun, S. Jaureguiberry, J. Reignier, S. Nseir, F. Danion, R. Clere-Jehl, K. Bouiller, J.-C. Navellou, V. Tolsma, A. Cabié, C. Dubost, J. Courjon, S. Leroy, J. Mootien, R. Gaci, B. Mourvillier, E. Faure, V. Pourcher, S. Gallien, O. Launay, K. Lacombe, J.-P. Lanoix, A. Makinson, G. Martin-Blondel, L. Bouadma, E. Botelho-Nevers, A. Gagneux-Brunon, O. Epaulard, L. Piroth, F. Wallet, J.-C. Richard, J. Reuter, T. Staub, B. Lina, M. Noret, C. Andrejak, M. P. Lê, G. Peytavin, M. Hites, D. Costagliola, Y. Yazdanpanah, C. Burdet, and F. Mentré. An open-label randomized controlled trial of the effect of lopinavir/ritonavir, lopinavir/ritonavir plus IFN-β-1a and hydroxychloroquine in hospitalized patients with COVID-19. Clinical Microbiology and Infection, 27(12) : 1826–1837, 2021. doi : 10.1016/j.cmi.2021.05.020.

F. Ader, M. Bouscambert-Duchamp, M. Hites, N. Peiffer-Smadja, J. Poissy, D. Belhadi, A. Diallo, M.-P. Lê, G. Peytavin, T. Staub, R. Greil, J. Guedj, J.-A. Paiva, D. Costagliola, Y. Yazdanpanah, C. Burdet, F. Mentré, A. Egle, R. Greil, M. Joannidis, B. Lamprecht, A. Altdorfer, L. Belkhir, V. Fraipont, M. Hites, G. Verschelden, J. Aboab, F. Ader, H. Ait-Oufella, C. Andrejak, P. Andreu, L. Argaud, F. Bani-Sadr, F. Benezit, M. Blot, E. Botelho-Nevers, L. Bouadma, O. Bouchaud, D. Bougon, K. Bouiller, F. Bounes-Vardon, D. Boutoille, A. Boyer, C. Bruel, A. Cabié, E. Canet, C. Cazanave, C. Chabartier, C. Chirouze, R. Clere-Jehl, J. Courjon, F. Crockett, F. Danion, A. Delbove, J. Dellamonica, F. Djossou, C. Dubost, A. Duvignaud, O. Epaulard, L. Epelboin, M. Fartoukh, K. Faure,

E. Faure, T. Ferry, C. Ficko, S. Figueiredo, B. Gaborit, R. Gaci, A. Gagneux-Brunon, S. Gallien, D. Garot, G. Geri, S. Gibot, F. Goehringer, M. Gousseff, D. Gruson, Y. Hansmann, O. Hinschberger, S. Jaureguiberry, V. Jeanmichel, S. Kerneis, A. Kimmoun, K. Klouche, M. Lachâtre, K. Lacombe, F. Laine, J.-P. Lanoix, O. Launay, B. Laviolle, V. Le Moing, J. Le Pavec, Y. Le Tulzo, P. Le Turnier, D. Lebeaux, B. Lefevre, S. Leroy, F.-X. Lescure, H. Lessire, B. Leveau, P. Loubet, A. Makinson, D. Malvy, C.-H. Marquette, G. Martin-Blondel, M. Martinot, J. Mayaux, A. Mekontso-Dessap, F. Meziani, J.-P. Mira, J.-M. Molina, X. Monnet, J. Mootien, B. Mourvillier, M. Murris-Espin, J.-C. Navellou, S. Nseir, W. Oulehri, N. Peiffer-Smadja, T. Perpoint, G. Pialoux, B. Pilmis, V. Piriou, L. Piroth, J. Poissy, V. Pourcher, J.-P. Quenot, F. Raffi, J. Reignier, M. Revest, J.-C. Richard, B. Riu-Poulenc, C. Robert, P.-A. Roger, C. Roger, E. Rouveix-Nordon, Y. Ruch, N. Saidani, N. Sayre, E. Senneville, A. Sotto, F. Stefan, C. Tacquard, N. Terzi, J. Textoris, G. Thiery, J.-F. Timsit, V. Tolsma, J.-M. Turmel, F. Valour, F. Wallet, G. Wattecamps, Y. Yazdanpanah, Y. Zerbib, M. Berna, J. Reuter, T. Staub, S. Braz, J.-M. Ferreira Ribeiro, J.-A. Paiva, R. Roncon-Albuquerque, M. Bouscambert-Duchamp, A. Gaymard, M.-P. Lê, B. Lina, G. Peytavin, S. Tubiana, S. Couffin-Cadièrgues, H. Esperou, D. Belhadi, C. Burdet, D. Costagliola, A. Dechanet, C. Delmas, A. Diallo, C. Fougerou, J. Guedj, F. Mentré, N. Mercier, M. Noret, J. Saillard, and P. Velou. Remdesivir plus standard of care versus standard of care alone for the treatment of patients admitted to hospital with COVID-19 (DisCoVeRy) : a phase 3, randomised, controlled, open-label trial. The Lancet Infectious Diseases, 22(2) :209–221, 2022. doi : 10.1016/S1473-3099(21)00485-0.

P. S. Albert. Longitudinal data analysis (repeated measures) in clinical trials. Statistics in Medicine, 18(13) :1707–1732, 1999. doi : 10.1002/(SICI)1097-0258(19990715)18:13<1707::AID-SIM138>3.0.CO;2-H.

R. J. Bauer. NONMEM Tutorial Part I : Description of Commands and Options, With Simple Examples of Population Analysis. CPT : Pharmacometrics & Systems Pharmacology, 8(8) : 525–537, 2019a. doi : 10.1002/psp4.12404.

R. J. Bauer. NONMEM Tutorial Part II : Estimation Methods and Advanced Examples.

CPT : Pharmacometrics & Systems Pharmacology, 8(8) :538–556, Aug. 2019b. doi : 10.1002/psp4.12422.

T. Bayes. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Philosophical Transactions of the Royal Society of London, 53 :370–418, 1763. doi : 10.1098/rstl.1763.0053.

J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez De Castilla, R. W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T. F. Patterson, R. Paredes, D. A. Sweeney, W. R. Short, G. Touloumi, D. C. Lye, N. Ohmagari, M.-d. Oh, G. M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M. G. Kortepeter, R. L. Atmar, C. B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J. D. Neaton, T. H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H. C. Lane. Remdesivir for the Treatment of Covid-19 — Final Report. New England Journal of Medicine, 383(19) :1813–1826, 2020. doi : 10.1056/NEJMoa2007764.

J. Bertrand, E. Comets, C. M. Laffont, M. Chenel, and F. Mentré. Pharmacogenetics and population pharmacokinetics : impact of the design on three tests using the SAEM algorithm. Journal of Pharmacokinetics and Pharmacodynamics, 36(4) :317–339, 2009. doi : 10.1007/s10928-009-9124-x.

J. Bertrand, E. Comets, M. Chenel, and F. Mentré. Some Alternatives to Asymptotic Tests for the Analysis of Pharmacogenetic Data Using Nonlinear Mixed Effects Models. Biometrics, 68(1) :146–155, 2012. doi : 10.1111/j.1541-0420.2011.01665.x.

M. Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. 2017. doi : 10.48550/ARXIV.1701.02434.

L. Billingham, K. Malottki, and N. Steven. Small sample sizes in clinical trials : a statistician's perspective. Clinical Investigation, 2(7) :655–657, 2012. doi : 10.4155/cli.12.62.

P.-C. Bürkner. **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. Journal of Statistical Software, 80(1), 2017. doi : 10.18637/jss.v080.i01.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. *Stan* : A Probabilistic Programming Language. Journal of Statistical Software, 76(1), 2017. doi : 10.18637/jss.v076.i01.

S. C. Chow. Bioavailability and bioequivalence in drug development. Wiley Interdisciplinary Reviews. Computational Statistics, 6(4) :304–312, 2014. doi : 10.1002/wics.1310.

S.-C. Chow and J.-P. Liu. Design and analysis of bioavailability and bioequivalence studies (3rd ed.). In Pharmaceutical Formulation Design - Recent Practices. Chapman and Hall/CRC, 2008. doi : 10.1201/9781420011678.

E. Comets, A. Lavenu, and M. Lavielle. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation. Journal of Statistical Software, 80 :1–41, 2017.

E. Comets, C. Rodrigues, V. Jullien, and M. Ursino. Conditional Non-parametric Bootstrap for Non-linear Mixed Effect Models. Pharmaceutical Research, 38(6) :1057–1066, 2021. doi : 10.1007/s11095-021-03052-6.

H. Cramér. Mathematical methods of statistics. Princeton landmarks in mathematics and physics. Princeton University Press, Princeton, 1999. ISBN 978-0-691-00547-8.

A. M. De Livera, S. Zaloumis, and J. A. Simpson. Models for the analysis of repeated continuous outcome measures in clinical trials. Respirology, 19(2) :155–161, 2014. doi : 10.1111/resp.12217.

M. Delattre and E. Kuhn. Computing an empirical fisher information matrix estimate in latent variable models through stochastic approximation. Computo, 2023. doi : 10.57750/r5gx-jk62.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. The Annals of Statistics, 27(1), 1999. doi : 10.1214/aos/1018031103.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1) :1–38, 1977. doi : 10.1111/j.2517-6161.1977.tb01600.x.

A. Dosne, M. Bergstrand, K. Harling, and M. Karlsson. Improving the estimation of parameter uncertainty distributions in nonlinear mixed effects models using sampling importance. J Pharmacokinet Pharmacodyn, 43(6) :583–596, 2016. doi : 10.1007/s10928-016-9487-8.

A. Dubois, S. Gsteiger, E. Pigeolet, and F. Mentré. Bioequivalence Tests Based on Individual Estimates Using Non-compartmental or Model-Based Analyses : Evaluation of Estimates of Sample Means and Type I Error for Different Designs. Pharmaceutical Research, 27(1) : 92–104, 2010. doi : 10.1007/s11095-009-9980-5.

A. Dubois, M. Lavielle, S. Gsteiger, E. Pigeolet, and F. Mentré. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. Statistics in Medicine, 30(21) :2582–2600, 2011. doi : 10.1002/sim.4286.

A. Dubois, S. Gsteiger, S. Balser, E. Pigeolet, J. L. Steimer, G. Pillai, and F. Mentré. Pharmacokinetic Similarity of Biologics : Analysis Using Nonlinear Mixed-Effects Modeling. Clinical Pharmacology & Therapeutics, 91(2) :234–242, 2012. doi : 10.1038/clpt.2011.216.

European Medicines Agency. Guideline on the investigation of bioequivalence. Technical report, 2010. `https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf`.

L. Fang, M.-J. Kim, Z. Li, Y. Wang, C. DiLiberti, J. Au, A. Hooker, M. Ducharme, R. Lionberger, and L. Zhao. Model-informed drug development and review for generic products : Summary of fda public workshop. Clinical Pharmacology & Therapeutics, 104 :27–30, 2018. doi : 10.1002/cpt.1065.

R. Fisher. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594-604) :309–368, 1922. doi : 10.1098/rsta.1922.0009.

J. Gabry, R. Češnovar, and A. Johnson. cmdstanr : R Interface to 'CmdStan'. https ://mc-stan.org/cmdstanr/, https ://discourse.mc-stan.org.

A. Gallant. Seemingly unrelated nonlinear regressions. Journal of Econometrics, 3(1) :35–50, 1975. doi : https://doi.org/10.1016/0304-4076(75)90064-0.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4) :733–760, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian Data Analysis Third Edition. Chapman and Hall/CRC, 3 edition, 2013. doi : 10.1201/b16018.

W. W. Hauck and S. Anderson. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. Journal of Pharmacokinetics and Biopharmaceutics, 12(1) :83-91, 1984. doi : 10.1007/BF01063612.

D. Hema Nagadurga. Bioavailability and bioequivalence studies. In U. Ahmad and J. Akhtar, editors, Pharmaceutical Formulation Design - Recent Practices, chapter 4. IntechOpen, Rijeka, 2019. doi : 10.5772/intechopen.85145.

J. Hughes, R. Upton, and D. Foster. Comparison of non-compartmental and mixed effect modelling methods for establishing bioequivalence for the case of two compartment kinetics and censored concentrations. Journal of Pharmacokinetics and Pharmacodynamics, 44 : 233–244, 2017. doi : 10.1007/s10928-017-9511-7.

E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. Computational Statistics & Data Analysis, 49(4) :1020–1038, 2005. doi : 10.1016/j.csda.2004.07.002.

A. Lavalley-Morelle, F. Mentré, E. Comets, and J. Mullaert. Extending the code in the open-source *saemix* package to fit joint models. submitted to Computer Methods and Programs in Biomedicine, 2023.

M. Lavielle. Mixed effects models for the population approach : models, tasks, methods and tools. Chapman and Hall/CRC, 2014.

J. Lee, Y. Gong, S. Bhoopathy, C. E. DiLiberti, A. C. Hooker, A. Rostami-Hodjegan, S. Schmidt, S. Suarez-Sharp, V. Lukacova, L. Fang, and L. Zhao. Public workshop summary report on fiscal year 2021 generic drug regulatory science initiatives : Data analysis and

model-based bioequivalence. Clinical Pharmacology & Therapeutics, 110(5) :1190–1195, 2021. doi : https://doi.org/10.1002/cpt.2120.

R. A. Levine and G. Casella. Implementations of the Monte Carlo EM Algorithm. Journal of Computational and Graphical Statistics, 10(3) :422–439, 2001. doi : 10.1198/106186001317115045.

M. L. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. Biometrics, 46(3) :673–687, 1990.

G. Lingas, N. Néant, A. Gaymard, D. Belhadi, G. Peytavin, M. Hites, T. Staub, R. Greil, J.-A. Paiva, J. Poissy, N. Peiffer-Smadja, D. Costagliola, Y. Yazdanpanah, F. Wallet, A. Gagneux-Brunon, F. Mentré, F. Ader, C. Burdet, J. Guedj, and M. Bouscambert-Duchamp. Effect of remdesivir on viral dynamics in COVID-19 hospitalized patients : a modelling analysis of the randomized, controlled, open-label DisCoVeRy trial. Journal of Antimicrobial Chemotherapy, 77(5) :1404–1412, 2022. doi : 10.1093/jac/dkac048.

F. Loingeville, J. Bertrand, T. T. Nguyen, S. Sharan, K. Feng, W. Sun, J. Han, S. Grosser, L. Zhao, L. Fang, K. Möllenhoff, H. Dette, and F. Mentré. New Model–Based Bioequivalence Statistical Approaches for Pharmacokinetic Studies with Sparse Sampling. The AAPS Journal, 22(6) :141, 2020. doi : 10.1208/s12248-020-00507-3.

T. A. Louis. Finding the Observed Information Matrix When Using the *EM* Algorithm. Journal of the Royal Statistical Society : Series B (Methodological), 44(2) :226–233, 1982. doi : 10.1111/j.2517-6161.1982.tb01203.x.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. Winbugs - a bayesian modelling framework : Concepts, structure, and extensibility. Statistics and Computing, 10(4) :325–337, 2000. doi : 10.1023/A:1008929526011.

N. Metropolis and S. Ulam. The Monte Carlo Method. Journal of the American Statistical Association, 44(247) :335–341, 1949. doi : 10.1080/01621459.1949.10483310.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation

of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 21 (6) :1087–1092, 1953. doi : 10.1063/1.1699114.

H. A. Miot. Analysis of data with dependent measures in clinical and experimental studies. Jornal Vascular Brasileiro, 22 :e20220150, 2023. doi : 10.1590/1677-5449.202201502.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3) :370, 1972. doi : 10.2307/2344614.

N. Néant, G. Lingas, A. Gaymard, D. Belhadi, M. Hites, T. Staub, R. Greil, J. Paiva, J. Poissy, N. Peiffer-Smadja, D. Costagliola, Y. Yazdanpanah, M. Bouscambert-Duchamp, A. Gagneux-Brunon, F. Ader, F. Mentré, F. Wallet, C. Burdet, J. Guedj, and the DisCoVeRy study group. Association between SARS-CoV-2 viral kinetics and clinical score evolution in hospitalized patients. CPT : Pharmacometrics & Systems Pharmacology, 2023. doi : 10.1002/psp4.13051.

X. Panhard and F. Mentré. Evaluation by simulation of tests based on non-linear mixed-effects models in pharmacokinetic interaction and bioequivalence cross-over trials. Statistics in Medicine, 24(10) :1509–1524, 2005. doi : 10.1002/sim.2047.

J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics, 11(2), 2017. doi : 10.1214/17-EJS1337SI.

J. Pinheiro, D. Bates, and M. Lindstrom. Model building in nonlinear mixed effects models. 1994.

M. Plummer. Jags : A program for analysis of bayesian graphical models using gibbs sampling. 2003.

C. R. Rao. Information and the Accuracy Attainable in the Estimation of Statistical Parameters. In Breakthroughs in Statistics, pages 235–247. Springer New York, 1992. doi : 10.1007/978-1-4612-0919-5_16. Series Title : Springer Series in Statistics.

M.-K. Riviere, S. Ueckert, and F. Mentré. An MCMC method for the evaluation of the Fisher information matrix for non-linear mixed effect models. Biostatistics, 17(4) :737–750, 2016. doi : 10.1093/biostatistics/kxw020.

C. Robert. Monte Carlo Methods in Statistics. In International Encyclopedia of Statistical Science, pages 854–858. Springer Berlin Heidelberg, 2011. doi : 10.1007/978-3-642-04898-2_376.

D. B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. The Annals of Statistics, 12(4), 1984. doi : 10.1214/aos/1176346785.

R. M. Savic, F. Mentré, and M. Lavielle. Implementation and Evaluation of the SAEM Algorithm for Longitudinal Ordered Categorical Data with an Illustration in Pharmacokinetics–Pharmacodynamics. The AAPS Journal, 13(1) :44–53, 2011. doi : 10.1208/s12248-010-9238-5.

D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6) :657-680, 1987. doi : 10.1007/BF01068419.

L. B. Sheiner, B. Rosenberg, and V. V. Marathe. Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. Journal of Pharmacokinetics and Biopharmaceutics, 5(5) :445–479, 1977. doi : 10.1007/BF01061728.

Stan Development Team. RStan : the R interface to Stan. URL `https://mc-stan.org/rstan`. R package version 2.26.24.

C. Tardivon, F. Loingeville, M. Donnelly, K. Feng, W. Sun, G. Sun, S. Grosser, L. Zhao, L. Fang, F. Mentré, and J. Bertrand. Evaluation of model-based bioequivalence approach for single sample pharmacokinetic studies. CPT : Pharmacometrics & Systems Pharmacology, pages 904–915, 2023. doi : 10.1002/psp4.12960.

H. Thai, F. Mentré, N. Holford, C. Veyrat-Follet, and E. Comets. Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models : a simulation study in population pharmacokinetics. J Pharmacokinet Pharmacodyn, 41(1) : 15–33, 2013.

S. Ueckert, M. Riviere, and F. Mentré. Alternative to resampling methods in maximum

likelihood estimation for nlmems by borrowing from bayesian methodology. https://www.page-meeting.org/?abstract=3632, 2015.

U.S. Food and Drug Administration. Bioequivalence studies with pharmacokinetic endpoints for drugs submitted under an anda - guidance for industry. Technical report, 2021. https://www.fda.gov/media/87219/download.

U.S. Food and Drug Administration. Statistical approaches to establishing bioequivalence - guidance for industry. Technical report, 2022a. https://www.fda.gov/media/163638/download.

U.S. Food and Drug Administration. Population pharmacokinetics - guidance for industry. Technical report, 2022b. https://www.fda.gov/media/128793/download.

U.S. Food and Drug Administration. Generic drug user fee amendments science and research priority initiatives for fiscal year 2022. Technical report, 2022c. https://www.fda.gov/media/154487/download.

A. van der Vaart. Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, 1998.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-Normalization, Folding, and Localization : An Improved R^ for Assessing Convergence of MCMC (with Discussion). Bayesian Analysis, 16(2), 2021. doi : 10.1214/20-BA1221.

G. Verbeke and G. Molenberghs. Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. Springer New York, 2000. doi : 10.1007/b98969.

M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, Z. Shi, Z. Hu, W. Zhong, and G. Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Research, 30(3) :269–271, 2020. doi : 10.1038/s41422-020-0282-0.

L. Wasserman. All of Statistics : A Concise Course in Statistical Inference. Springer New York, 2004. doi : 10.1007/978-0-387-21736-9.

WHO Solidarity Trial Consortium. Repurposed Antiviral Drugs for Covid-19 — Interim WHO Solidarity Trial Results. New England Journal of Medicine, 384(6) :497–511, 2021. doi : 10.1056/NEJMoa2023184.

B. N. Williamson, F. Feldmann, B. Schwarz, K. Meade-White, D. P. Porter, J. Schulz, N. Van Doremalen, I. Leighton, C. K. Yinda, L. Pérez-Pérez, A. Okumura, J. Lovaglio, P. W. Hanley, G. Saturday, C. M. Bosio, S. Anzick, K. Barbian, T. Cihlar, C. Martens, D. P. Scott, V. J. Munster, and E. De Wit. Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. Nature, 585(7824) :273–276, 2020. doi : 10.1038/s41586-020-2423-5.

World Health Organisation. Covid-19 therapeutic trial synopsis. Technical report, 2020. https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis.

L. Zhao, M.-J. Kim, L. Zhang, and R. Lionberger. Generating model integrated evidence for generic drug development and assessment. Clinical Pharmacology & Therapeutics, 105(2) : 338–349, 2019. doi : https://doi.org/10.1002/cpt.1282.

F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen, and B. Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China : a retrospective cohort study. The Lancet, 395(10229) :1054–1062, 2020. doi : 10.1016/S0140-6736(20)30566-3.

N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, and W. Tan. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine, 382(8) : 727–733, 2020. doi : 10.1056/NEJMoa2001017.

Calcul de l'incertitude à distance finie dans les modèles non linéaires à effets mixtes

L'utilisation des modèles non linéaires à effets mixtes sur des données longitudinales éparses (peu de sujets et/ou d'observations par sujet) est limitée par le manque de méthodes robustes de calcul de l'incertitude à distance finie, c'est-à-dire loin de l'asymptotique. Les différents travaux de cette thèse confirment la pertinence des méthodes semi-Bayésiennes, mise en évidence ces dernières années, pour calculer l'incertitude autour de l'estimateur du maximum de vraisemblance dans le cas épars. Dans un premier travail, nous avons montré que l'utilisation d'une méthode basée sur l'algorithme Hamiltonian Monte Carlo (HMC), combinée à une procédure de sélection du modèle pharmacocinétique utilisé, permettait de contrôler l'erreur de type I des tests de bioéquivalence mis en place sur des données éparses grâce à une meilleure estimation de l'erreur standard de l'effet traitement. Dans la suite de cette thèse, nous avons développé une méthode basée sur l'algorithme de Metropolis-Hastings intégrée à l'algorithme SAEM et implémentée dans le package R `saemix`. Cette méthode ainsi que celle basée sur l'algorithme HMC ont montré de bonnes performances sur données éparses, mais également présenté des limites en présence de structure de variabilité inter-individuelle complexe. L'algorithme Approximate Bayesian Computation (ABC) n'ont pas présenté pas les mêmes limites et ont donné des résultats prometteurs. Les différents projets de cette thèse ont également permis, grâce à l'utilisation de données d'études cliniques, d'illustrer la possibilité de rencontrer de telles structures de variabilité complexes en pratique, et de mettre en avant la difficulté de calibrer les méthodes semi-Bayésiennes. De plus amples évaluations sont nécessaires pour calibrer les paramètres de l'algorithme ABC et développer un outil diagnostique permettant de détecter la mise en difficulté des méthodes présentées en pratique.

Mots-clés : *modèles non linéaires à effets mixtes, données éparses, incertitude, erreurs standards, inférence semi-Bayésienne*

Computation of uncertainty at finite distance in nonlinear mixed effects models

The use of nonlinear mixed effects models on sparse longitudinal data (meaning few subjects and/or few observations per subject) is limited by the lack of robust methods to compute the uncertainty at finite distance, which means far from the asymptotic. The different projects in this thesis confirm the relevance of semi-Bayesian methods, highlighted by previous works published in recent years, to compute the uncertainty around the maximum likelihood estimator in the sparse case of nonlinear mixed effects models. In a first project, we showed that the use of a method based on the Hamiltonian Monte Carlo (HMC) algorithm, combined with a selection procedure of the pharmacokinetic model, allowed to control the type I error of bioequivalence tests on sparse data by improving the estimation of the standard error on the treatment effect. In the rest of this thesis, we developed a method based on the Metropolis-Hastings algorithm integrated in the SAEM algorithm and implemented in the `saemix` R package. This method, as well as the one based on HMC, showed good performances on sparse data but also limitations in the presence of complex structure of inter-individual variability. The Approximate Bayesian Computation (ABC) algorithm did not present the same limitations and gave promising results. The different projets of this thesis also enabled, thanks to the use of clinical data, to show that such complex variability structures exist in practice, and to highlight the difficulty to calibrate the semi-Bayesian methods in practice. Further evaluation is necessary to calibrate parameters of the ABC algorithm and develop a diagnostic tool allowing to detect when these methods fail in practice.

Keywords : *nonlinear mixed effects models, sparse data, uncertainty, standard errors, semi-Bayesian inference*