STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:5162–5179
Published online 8 May 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2910

Design in nonlinear mixed effects models: Optimization using the Fedorov–Wynn algorithm and power of the Wald test for binary covariates

Sylvie Retout[1,2,3,*,†], Emmanuelle Comets[1,2,3], Adeline Samson[1,2,3] and France Mentré[1,2,3]

[1]*Inserm, U738, Paris, France*
[2]*Université Paris 7, UFR de Médecine, Paris, France*
[3]*AP-HP, Hôpital Bichat, UF de Biostatistiques, Paris, France*

SUMMARY

We extend the methodology for designs evaluation and optimization in nonlinear mixed effects models with an illustration of the decrease of human immunodeficiency virus viral load after antiretroviral treatment initiation described by a bi-exponential model. We first show the relevance of the predicted standard errors (SEs) given by the computation of the population Fisher information matrix using the R function PFIM, in comparison to those computed with the stochastic approximation expectation–maximization algorithm, implemented in the Monolix software. We then highlight the usefulness of the Fedorov–Wynn (FW) algorithm for designs optimization compared to the Simplex algorithm. From the predicted SE of PFIM, we compute the predicted power of the Wald test to detect a treatment effect as well as the number of subjects needed to achieve a given power. Using the FW algorithm, we investigate the influence of the design on the power and show that, for optimized designs with the same total number of samples, the power increases when the number of subjects increases and the number of samples per subject decreases. A simulation study is also performed with the nlme function of R to confirm this result and show the relevance of the predicted powers compared to those observed by simulation. Copyright 2007 John Wiley & Sons, Ltd.

KEY WORDS: nonlinear mixed effects models; Fisher information matrix; population design; Wald test; power; HIV

1. INTRODUCTION

Nonlinear mixed effects models are increasingly used to model longitudinal data. They have been initiated in the context of pharmacokinetics analyses by Sheiner *et al.* in 1972 [1] and are now more

*Correspondence to: Sylvie Retout, Inserm, U738, Université Paris 7, UFR de Médecine, site Bichat, 16, rue Henri Huchard, Paris 75018, France.
†E-mail: sylvie.retout@bichat.inserm.fr

Received 4 August 2006
Accepted 20 March 2007

and more popular in other kinds of longitudinal studies in particular for human immunodeficiency virus (HIV) viral dynamics [2–4]. The purpose of nonlinear mixed effects model approach is to estimate the mean value of the parameters in the studied population and also to estimate their interindividual variability. Another use of this methodology, also called the population approach, is to determine and to quantify the influence of covariates on the parameters, which can help to define groups of population with different levels of response. In the context of HIV viral dynamics for instance, nonlinear mixed effects models can therefore be used to determine a difference in the potency of antiviral treatments, through the use of markers of potency such as the parameter describing the first HIV viral decay rate [5, 6]. Nonlinear mixed effects methodology does not require individual parameter estimation and one of its main advantages is that it can therefore deal with sparse individual data, where individual estimation methods would not be successful. Several estimation methods for those models have been proposed and are now implemented in software which makes them easier to use [7]. Methods have been proposed in the context of maximum likelihood estimation based on an approximation of the log-likelihood using a linearization of the model around a value of the random effects, like the First-Order (FO) method, or the First-Order Conditional Estimation method proposed by Lindstrom and Bates [8] and implemented in the NONMEM software [9] and in the nlme function of Splus and R software [10]. Alternatives to those linearization methods have also been proposed, such as Gaussian quadrature [11] method, which implements the corresponding classical numerical quadrature methods like the NLMIXED procedure of SAS. Nevertheless, those algorithms often imply slow convergence and problems of stability. More recently, a new algorithm has been developed, based on the most commonly used method to estimate models with missing or non-observed data such as random effect, the expectation–maximization (EM) algorithm. Because of the nonlinearity of the model, the new algorithm involves a stochastic approximation version of the EM algorithm, and is thus called the stochastic approximation expectation–maximization (SAEM) algorithm. Proof that the produced estimates are convergent and consistent has been given [12]. The SAEM algorithm has been implemented in the MONOLIX software [13] and a comparison study to other existing algorithms has highlighted the small bias and root mean square errors obtained with this algorithm [14].

Before estimation, the experimenters faced the problem of the determination of the design for collecting data. Indeed, as this approach allows sparse individual data, the data have to be chosen informatively to avoid poor parameter estimates. Population designs are defined by several groups of subjects; each group is composed of a number of samples to be performed on a number of subjects at given times. Simulation studies have shown that the precision of parameter estimates depends on the choice of the balance between the number of groups to include, the number of subjects per group and the number and allocation of the sampling times [15, 16]. The general theory of design determination used for classical nonlinear models [17, 18] has been extended to nonlinear mixed effects models. This theory relies on the Rao–Cramer inequality which states that the inverse of the Fisher information matrix (M_F) is the lower bound of the variance–covariance matrix of any unbiased parameter estimator. This involves the determination of the expression of M_F but, as there is no analytical expression of the likelihood in those models, determination of the exact analytical expression of the population M_F is not possible. An approximation was first proposed by Mentré *et al.* [19], and extended by Retout *et al.* [20, 21]. This approximation uses a FO Taylor expansion of the model around the random effects. Assessing the relevance of determining population designs using this approximation of M_F has been the purpose of several papers, in which the usefulness of this approach has been demonstrated, either by simulation [20–22] or on real pharmacokinetic studies [23, 24]. To facilitate its use, a tool for population

DOI: 10.1002/sim

design evaluation based on the approximation of M_F has been proposed (PFIM) as a generic function developed in the statistical software Splus version 6 but also in its free version, R [25].

To deal with more complex models, the expression of M_F has been extended for models including fixed effects for the influence of covariates on the parameters and for an additional variability of the parameters of a given individual between several occasions, also called intra-individual or inter-occasion variability [21].

A tool has also been proposed for design optimization: PFIMOPT 1.0, implemented also as a generic function in Splus and R [25]. PFIMOPT 1.0 uses the Simplex algorithm to maximize the D-optimal criterion, i.e. to maximize the determinant of M_F for a given *a priori* value of the parameters. For a given total number of samples, the best sampling times as well as the best proportions of subjects per group can be determined within given continuous intervals. However, the Simplex algorithm is a general optimization algorithm, not specifically tailored for complex designs optimization problems that can sometimes occur for population designs. Indeed, for large optimization problems (several groups with a large number of sampling times), we observed in a previous work that the Simplex algorithm sometimes converges to local minima [26]. Moreover, although the optimization may result to proportion of subjects equal to zero for some groups, optimization of the number of groups to include and the number of samples per group cannot be easily considered with the Simplex algorithm. Other general algorithms have been proposed for this purpose, such as nonadaptative random search and simulated annealing but they all appear to be very cumbersome [26]. A more specific design optimization algorithm, the Fedorov–Wynn (FW) algorithm [27, 28], has already been used in this context by Mentré *et al.* [19] and Retout *et al.* [20]; it has the property of converging towards the D-optimal design and it can be used to optimize both the group structure (number of groups, number of subjects per group, number of samples per group) and the sampling times from a finite set of times. However, as far as we know, no comparison of the FW algorithm to other optimization algorithms has been performed and no code is available for a generic use of this algorithm in the context of nonlinear mixed effects models.

In this article, we extend the population design evaluation and optimization methodology and illustrate it on the decrease of HIV viral load after initiation of antiretroviral treatment using the bi-exponential model proposed by Ding and Wu [5]. Despite the large number of ongoing HIV dynamics studies analysed by nonlinear mixed effects models, the design of those population studies is still an issue. Wu and Ding [6] evaluated a finite number of plausible designs by simulations. Han and Chaloner investigated experimental designs for those models using Bayesian evaluations [29]. Wu and Ding [6] also studied the impact of several design strategies on the power for identifying a treatment difference. Indeed, the evaluation of the efficacy of antiviral treatments is a crucial issue as well as designing a clinical trial for treatment comparison. A potential good marker for this efficacy is the first viral decay rate parameter [5]. In the current study, our objective is to propose a method to predict and to improve with respect to the population design, the power of a Wald test to detect a treatment effect. Kang *et al.* have already proposed a method to compute sample sizes to achieve a given power in the context of nonlinear mixed effect models using linearization [30, 31]. However, they do not take into account the influence of the estimation of variances of random effects on the power. Marschner also addressed some statistical issues in the design of such studies by taking into account those variances of the random effects but in the case of linear models [32].

We first present the statistical model used for the modelling of HIV viral load decrease (Section 2). We compare the predicted standard errors (SEs) given by the computation of M_F

DOI: 10.1002/sim

with PFIM to those given by a 'true' evaluation of M_F, obtained by stochastic approximation using MONOLIX (Section 3). We then investigate the usefulness of using the FW algorithm by comparison to the Simplex algorithm and we propose a generic implementation of this algorithm in PFIMOPT, as an alternative to the Simplex algorithm (Section 4). Last, we compute the predicted power of the Wald test on the first decay rate for a comparison of two treatments using the SE predicted by PFIM; we investigate the influence of the design on this power and we evaluate by simulation the relevance of the predicted power (Section 5).

2. MODEL

The decrease of the viral load of HIV-infected patients after initiation of antiretroviral treatment can be modelled by a bi-exponential model as recommended by Ding and Wu [5]. The statistical model for a subject i among N at time t_{ij} is therefore given by

$$y_{ij} = f(\phi_i, t_{ij}) + \varepsilon_{ij}$$

with y_{ij} the $\log_{10}$ viral load of subject i at time t_{ij} and

$$f(\phi_i, t_{ij}) = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{ij}} + P_{2i}e^{-\lambda_{2i}t_{ij}})$$

ϕ_i is the vector of log-parameters for subjects i, composed of the baseline values P_{1i} and P_{2i} and of the two-phase viral decay rates λ_{1i} and λ_{2i}, so that $\phi_i = (\log P_{1i}, \log P_{2i}, \log \lambda_{1i}, \log \lambda_{2i})$. ε_{ij} is the random error and it is assumed that $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$.

Each ϕ_i is assumed to be multi-normally distributed with mean $\mu = (\log P_1, \log P_2, \log \lambda_1, \log \lambda_2)$ and a diagonal variance covariance matrix Ω, such as: $\phi_i = \mu + b_i$ with $b_i \sim \mathrm{N}(0, \Omega)$. We note $\omega^2 = (\omega_1^2, \omega_2^2, \omega_3^2, \omega_4^2)$ the elements of the diagonal of Ω, i.e. the vector of the variances of the random effects.

In case of a study including two groups of treatment A and B, we assume an additional fixed effect β for the antiretroviral treatment effect of treatment B compared to treatment A, added on the first rate constant such as $\log(\lambda_{1i})^B = \log(\lambda_{1i})^A + \beta$, so that $\mu = (\log P_1, \log P_2, \log \lambda_1, \log \lambda_2, \beta)$.

The usual assumptions are made: $\varepsilon_i | b_i$, $i = 1, \ldots, N$, are assumed to be independent from one subject to the other and for each subject, ε_i and b_i are also independent.

The vector Ψ of the population parameters to be estimated in those models, with or without a treatment effect β, is then composed of the fixed effects μ, of ω^2 and of σ^2.

3. EVALUATION OF THE POPULATION FISHER INFORMATION MATRIX USING SAEM

3.1. Method

In this section, we compare for a given design the predicted SE computed by approximation with PFIM as in Retout *et al.* [21] with the 'true' SE obtained with SAEM. Indeed, Samson *et al.* [33] proposed an alternative to the linearization approach for the evaluation of the population Fisher information based on the SAEM algorithm implemented in MONOLIX and using the Louis's principle [34]. Although this evaluation of M_F is much more cumbersome and its use cannot thus

DOI: 10.1002/sim

Table I. Comparison of the relative standard errors (per cent), RSE (per cent), computed for the empirical design with 100 subjects per group, either by the SAEM procedure or by PFIM using the bi-exponential viral decay model with a treatment effect.

	RSE (per cent)	
Parameters	PFIM	SAEM
$\ln P_1$	0.34	0.34
$\ln P_2$	0.52	0.57
$\ln \lambda_1$	7.9	8.1
β	0.079	0.078
$\ln \lambda_2$	1.3	1.3
ω_1^2	10.9	10.8
ω_2^2	11.5	12.9
ω_3^2	10.3	10.4
ω_4^2	10.4	10.8
σ	3.5	2.8

be envisioned for design optimization purpose, it does not require any linearization and can thus be considered as the 'true' population Fisher information matrix. To do that comparison, we use the bi-exponential viral load decrease model, including the treatment effect. We evaluate a population design, called 'Empirical' design, composed of two groups of 100 subjects with the same sampling times (1, 3, 7, 14, 28, 56) days after treatment initiation. We assumed that those two groups differ only by their treatment. *A priori* values of the fixed effects are taken as approximately the same as those used in Ding and Wu [5]: $\log P_1 = 12.0$, $\log P_2 = 8.0$, $\log \lambda_1 = -0.7$ and $\log \lambda_2 = -3.0$. We assume that the treatment effect $\beta = 0$ is estimated and that the variances of the random effects are equal with a variation coefficient of 55 per cent, which corresponds to $\omega_1^2 = \omega_2^2 = \omega_3^2 = \omega_4^2 = 0.3$. Regarding the variance for the error model, we set $\sigma = 0.065$ on $\log_{10}$ viral load, which corresponds to a coefficient of variation of 15 per cent for the viral load.

We evaluate with PFIM the expected SE for the empirical design on the HIV viral load model. We then use the SAEM algorithm to compute the SE under asymptotic convergence assumption. To do that, we simulate a data set with a large number of subjects per group: $N_{\mathrm{SIM}} = 5000$ subjects, with the sampling times of the empirical design. We estimate the parameters and the observed Fisher information matrix on this simulated data set using the SAEM algorithm. Given the hypothesis of an identical sampling design for each subject, the Fisher information matrix of the complete data set is the sum of the individual Fisher information matrices. The SE of N subjects can then be evaluated from the SE of the simulated data set using: $\mathrm{SE}_N(\beta) = \mathrm{SE}_{N_{\mathrm{SIM}}}(\beta)/\sqrt{N_{\mathrm{SIM}}/N}$. In our case, we set $N = 100$ subjects.

3.2. Results

Results are given in Table I. SEs are reported as relative SEs (per cent), noted RSE (per cent). For all the parameters, the RSE given by PFIM or SAEM are very similar, especially on the treatment

DOI: 10.1002/sim

effect parameter. This confirms the relevance of the predicted SE computed by PFIM. Furthermore, the very low values of the RSE show that the empirical design is very efficient.

4. FEDOROV–WYNN ALGORITHM FOR DESIGN OPTIMIZATION

4.1. Fedorov–Wynn algorithm

The FW algorithm [27, 28] is an iterative algorithm that maximizes the determinant of the Fisher information matrix within a finite set of possible designs. We define an elementary design as a series of sampling times. We briefly describe the algorithm here. Interested readers should consult Walter and Pronzato book (Chapter 6) for details [18]. Let ζ denote, in the following, the set of elementary designs.

A population design Ξ is defined as a series of n_Ξ designs ξ_i in ζ, where design ξ_i has frequency α_i and $\sum_{i=1}^{n_\Xi} \alpha_i = 1$. Let n_Ψ denote the number of parameters (or dimension) of the model. Following Walter and Pronzato [18], we denote $M_{F_\Xi}(\Psi, \Xi)$ the Fisher information matrix corresponding to Ξ evaluated given the n_Ψ parameters Ψ. Let det denote the determinant. For any elementary design ξ in ζ, and any α between 0 and 1, we can define a new design $\Xi' = (1-\alpha)\Xi + \alpha\xi$. This design has information matrix $M_{F_{\Xi'}}(\Psi, (1-\alpha)\Xi + \alpha\xi)$ and we define the distance $d(\Xi, \xi)$ as the derivative of $\log(\det(M_{F_{\Xi'}}))$ taken at $\alpha = 0$.

The FW algorithm relies on the Kiefer–Wolfowitz equivalence theorem [35], which states that the three following proposals are equivalent:

- Ξ is D-optimal;
- $\max_{\xi_i \text{ in } \zeta} d(\Xi, \xi_i) = n_\Psi$;
- Ξ minimizes $\max_{\xi_i \text{ in } \zeta} d(\Xi, \xi_i)$.

The algorithm proposed by Fedorov iteratively improves the population design as follows:

1. we start with an initial guess Ξ_0;
2. at step k, with the current design being Ξ_k, we find $\xi^* = \arg\max_{\xi_i \text{ in } \zeta} d(\Xi_k, \xi_i)$
   - we stop if $\max_{\xi_i \text{ in } \zeta} d(\Xi_k, \xi_i) \leqslant n_\Psi + \varepsilon$ where $\varepsilon \ll 1$ is a predetermined tolerance;
3. otherwise, we update the design to $\Xi_{k+1} = (1-\alpha^*)\Xi_k + \alpha^*\xi^*$, where α^* is chosen over]0, 1[to maximize

$$\det(M_{F_\Xi}(\Psi, \Xi_{k+1})) : \alpha^* = \frac{d(\Xi_k, \xi^*) - n_\Psi}{n_\Psi(d(\Xi_k, \xi^*) - 1)}$$

Because this algorithm can only add elementary designs and is thus prone to include more support points than needed, an additional step is included to optimize the frequencies at step $(k+1)$, $\{\alpha_j^{(k+1)}\}$. We use an active set method with a projected gradient method for the selection of the direction as implemented in Mallet [36] for nonparametric estimation in nonlinear mixed effects models. Elementary designs are removed after this optimization step if their frequency is lower than a predetermined δ (we chose $\delta = 10^{-8}$).

The practical implementation requires to specify the set of possible elementary designs ζ. We assume first a set S_T including n_T possible sampling times; these are given by clinical constraints. Let n_S be the desired number of sampling times. We generate the $N_P = C_{n_T}^{n_S}$ elementary designs

DOI: 10.1002/sim

consisting of n_S different times within S_T using the library `combinat` in the statistical software `R`. We then generate the Fisher matrices for these N_P elementary designs. We start the algorithm with an initial guess m_0 provided by the user.

A program in the `C` language was written to implement the FW algorithm. It is compiled within `R` as a shared library and called through the `dyn.load()` function. Wrapper functions have been created within `PFIM` to handle the call. Users must provide the possible sampling times to use the FW optimization option. Since the elementary designs are generated through combinatorial processes, it is possible to specify for instance that each elementary design should include two points amongst a first set of sampling times (e.g. day one) and a third point amongst a second set of sampling times (e.g. day two). The user therefore specifies a list of sampling windows, and for each window, the set of possible sampling times and the minimum and maximum number of sampling times within that window. Indeed, the number of sampling times n_S need not be identical across subjects; for instance, one may wish to obtain designs with 2–4 observations per subject. In this case, we generate all the elementary designs corresponding to 2, 3 or 4 points and optimize across the set of all the designs.

4.2. Designs optimization

The objective of this section is to compare the FW algorithm to the Simplex algorithm in the context of population designs optimization. We use the example of the bi-exponential model for the viral load decrease, without any treatment effect, and compare the results of optimization of several designs to those obtained with the Simplex algorithm already implemented in PFIMOPT 1.0.

We optimize with both the FW and the Simplex algorithm, four different population designs with 8, 5, 4 and 3 samples per subject, respectively, and for a total number of 480 samples. The purpose is then to optimize the number of groups to be included, the proportions of subjects per group and the sampling times. Optimization is performed using the D-optimality criterion for both algorithms: with the FW algorithm, the determinant of the Fisher matrix is maximized over the space of possible designs Ξ; its inverse is minimized with the Simplex algorithm.

We use the same *a priori* values of the population parameters as in Section 3. We use the same initial designs as starting points for both the FW and the Simplex algorithm. For designs with 5, 4 and 3 samples per subject, they are composed of four groups with one-fourth of the subjects assigned to each. The 8 samples initial design is composed of only one group.

Regarding the optimization with the FW algorithm, we fix a set of 12 allowed sampling times similar to those used in Wu and Ding [6]: 0, 1, 2, 3, 5, 7, 10, 14, 21, 28, 42 and 56 days after treatment initiation. These sampling times were chosen by the authors because of their clinical feasibility.

Regarding the Simplex algorithm, we optimize the sampling times in a continuous interval from 0 to 56 days. Note that in the case of the Simplex algorithm, optimization of the number of groups is performed through the optimization of proportion of subjects: this number decreases when one, or more, proportion of subjects is optimized to 0; it follows that the number of groups cannot increase. In order that the optimized designs be clinically more relevant, we impose a minimum delay of 1 h between two successive sampling times. Because of the known problems of local minima, two other initial designs are also tested; moreover, for each of these initial designs, optimization is performed several times, using iteratively the obtained optimal designs as the new initial designs for the next optimization. The optimized design with the best criterion is then selected as the result for the optimization with the Simplex algorithm. From this optimized design, we derive the

DOI: 10.1002/sim

Table II. Optimized designs with the Fedorov–Wynn and the Simplex algorithms and corresponding efficiency criterion Φ_D for the model without treatment effect. All designs are given as groups of sampling times in brackets and the corresponding numbers of subjects. The total number of samples is fixed to 480 and different numbers of samples per subject are considered.

Total number of subjects	Number of samples per subject	Initial population design {(sampling times), number of subject}	Optimized design	
			Fedorov–Wynn	Simplex
60	8	$\Phi_D = 471.1$	$\Phi_D = 478.0$	$\Phi_D = 483.3$
		{(0, 1, 7, 10, 21, 28, 42, 56), 60}	{(0, 1, 2, 7, 14, 21, 28, 56), 60}	{(0, 1, 7, 16, 17, 18, 54, 55), 60}
96	5	$\Phi_D = 599.5$	$\Phi_D = 644.1$	$\Phi_D = 647.6$
		{(0, 1, 2, 14, 56), 24} {(0, 1, 7, 28, 56), 24} {(0, 5, 7, 21, 56), 24} {(0, 7, 10, 42, 56), 24}	{(0, 7, 14, 21, 56), 96}	{(0, 7, 16, 17, 56), 96}
120	4	$\Phi_D = 602.2$	$\Phi_D = 645.9$	$\Phi_D = 646.2$
		{(0, 1, 7, 56), 30} {(0, 1, 21, 56), 30} {(0, 7, 10, 56), 30} {(0, 7, 21, 56), 30}	{(0, 7, 21, 56), 70} {(0, 21, 28, 56), 15} {(10, 14, 21, 56), 15} {(0, 1, 2, 5), 12} {(0, 5, 14, 56), 8}	{(0, 7, 17, 56), 73} {(0, 19, 55, 56), 24} {(9, 18, 19, 56), 17} {(8, 35, 56, 18), 6}
160	3	$\Phi_D = 493.2$	$\Phi_D = 639.8$	$\Phi_D = 642.7$
		{(0, 3, 10), 40} {(1, 7, 28), 40} {(2, 14, 56), 40} {(5, 21, 42), 40}	{(0, 1, 5), 53} {(0, 21, 56), 52} {(10, 21, 56), 52} {(21, 28, 56), 3}	{(9, 17, 56), 55} {(0, 1, 5), 54} {(0, 20, 56), 51}

DOI: 10.1002/sim

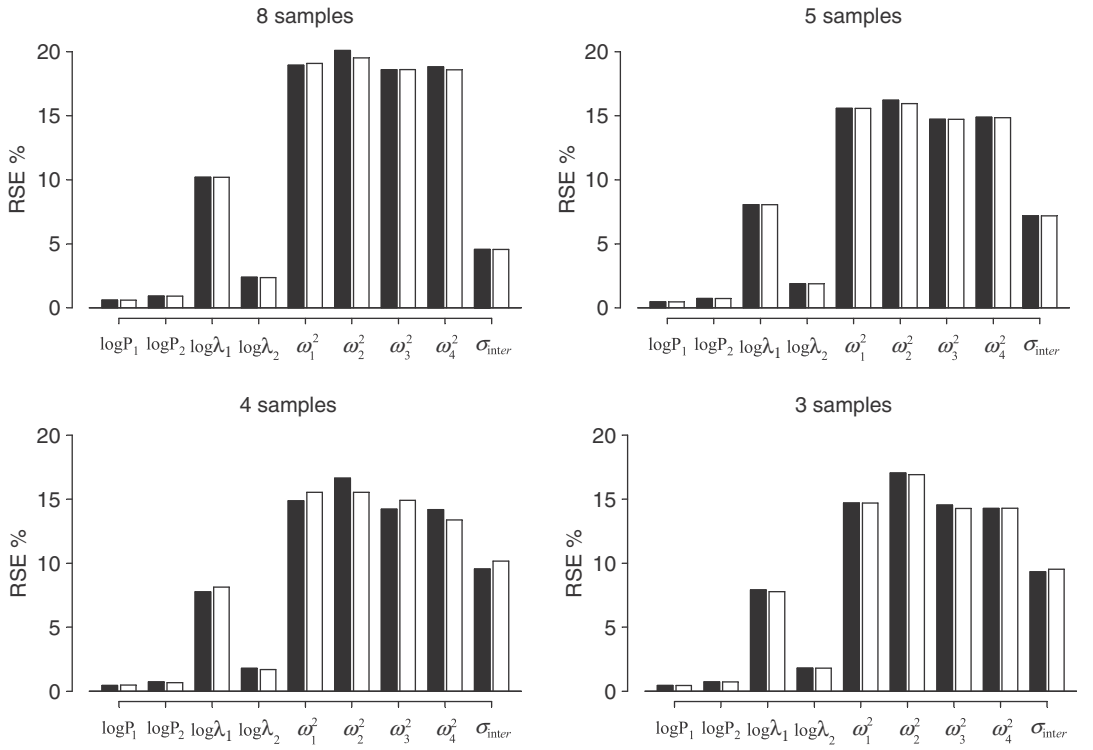

Figure 1. Comparison of the predicted RSE for designs with either 8, 5, 4 or 3 samples and optimized with either the Fedorov–Wynn algorithm (full bar) or the Simplex algorithm (open bar).

number of subjects from the optimized proportions and round them to the nearest integer. Last, in order to get more practical designs, we round the sampling times to the nearest day.

The comparison of the optimized designs with both algorithms is based on the comparison of their efficiency and the comparison of their group structure and the optimal sampling times. The usual efficiency criterion $\Phi_D(\Xi)$ is defined as the determinant of $M_F(\Psi, \Xi)$ normalized by n_Ψ, the dimension of Ψ, the vector of the population parameters to be estimated, i.e. $\Phi_D(\Xi) = [\det(M_F(\Psi, \Xi))]^{1/n_\Psi}$. The efficiency of a design Ξ_1 compared to a design Ξ_2, $\mathrm{Eff}(\Xi_1, \Xi_2)$, is then computed by the ratio of the efficiency criteria for both designs, i.e.:

$$\mathrm{Eff}(\Xi_1, \Xi_2) = \frac{\Phi_D(\Xi_1)}{\Phi_D(\Xi_2)}$$

4.3. Results

Optimized designs with their corresponding efficiency criterion $\Phi_D(\Xi)$ are given in Table II for both algorithms; the corresponding RSE are reported in Figure 1. All the results with the FW algorithm have been obtained after only one optimization run while the Simplex algorithm required to be run several times (about 2 or 3 times) for each initial population design to converge towards a minimum.

DOI: 10.1002/sim

Whatever the constraints on the number of subjects, both algorithms give optimized designs with similar criteria, although slightly lower for the FW algorithm, meaning that both algorithms generate designs supporting globally the same information. This is confirmed by the comparison of the predicted RSE given in Figure 1: whatever the optimization case (8, 5, 4 or 3 samples per subject) and the parameter, the expected RSE are very similar for the two optimization algorithms.

Optimized designs with either 5, 4 or 3 samples per subject allow globally the same level of information with a range of efficiency criteria from 639.8 to 647.6. Optimized designs with 8 sampling times are less efficient with an efficiency criterion of 478.0 for the FW algorithm and 483.3 for the Simplex algorithm. Globally, all these designs are very informative with expected RSE lower than 10 per cent for the fixed effects and lower than 20 per cent for the random effects parameters.

For the optimized sampling times (Table II), there are some differences between both algorithms. This was foreseeable due to the finite set of allowed sampling times for the FW algorithm: for instance the 17 days sampling time occurs in all the designs optimized with the Simplex algorithms but was not in the set of times used for the FW.

Regarding the group structure for designs with 5 samples per subject, both algorithms reduce the initial 4 groups design to only one group. The same group structure is obtained for designs with 8 samples per subject. When the number of samples per subject decreases (4 and 3), this group structure is more complex with at least 3 groups for the Simplex algorithm, and 4 for the FW algorithm.

5. INFLUENCE OF THE DESIGN ON THE POWER OF THE WALD TEST OF A TREATMENT EFFECT

5.1. Computation of the power of the Wald test and number of subjects needed to treat

Nonlinear mixed effects models also allow comparison of two treatment groups, testing a treatment effect on some of the fixed effects of the model. The Wald test can be used to assess this difference between the two groups. In this section, we aim at computing the power of the Wald test using the predicted SE of PFIM. We use the example of the bi-exponential viral load decrease with a treatment effect β on the first slope. We assume that the clinical trial aims to detect a minimum difference of at least β_1 between the two treatment groups on the parameter λ_1. Hence, the null hypothesis to test is $H_0: \{\beta = 0\}$ while the alternative hypothesis is $H_1: \{\beta \geqslant \beta_1\}$. The statistic of the Wald test is $S_W(\hat{\beta}) = \hat{\beta}/\mathrm{SE}(\hat{\beta})$ where $\hat{\beta}$ is an estimate of β and $\mathrm{SE}(\hat{\beta})$ the corresponding SE. To ensure a type I error α for the Wald test under H_0, the rejection region is $\{|S_W| > z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}$ is the critical value of a standard normal distribution. Under H_1, the statistic $S_W(\hat{\beta})$ is asymptotically distributed with a normal distribution centred on $\beta_1/\mathrm{SE}(\beta_1)$ where $\mathrm{SE}(\beta_1)$ is the predicted SE for the treatment effect β when $\beta = \beta_1$. Therefore, the power P of a Wald test is defined as

$$P = 1 - \Phi\left(z_{\alpha/2} - \frac{\beta_1}{\mathrm{SE}(\beta_1)}\right)$$

for $\beta_1 > 0$ where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution and $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

DOI: 10.1002/sim

Using the extension of PFIM for discrete covariate [21] as in Section 2, we derive the expected SE of β under the alternative hypothesis $\beta = \beta_1$ and can thus evaluate the expected power of this Wald test for a given design and a given value β_1 of the alternative hypothesis H_1.

Using the predicted SE of PFIM, we can also derive the number of subjects needed to achieve a power P to detect a treatment effect with the Wald test. To do that, we first compute the SE needed on β to obtain a power of P, called NSE(P), using the following relation:

$$\mathrm{NSE}(P) = \frac{\beta_1}{z_{\alpha/2} - \Phi^{-1}(1-P)}$$

We then compute the number of subjects needed to be included to obtain a power of P, called NNI(P) using

$$\mathrm{NNI}(P) = N \times \left(\frac{\mathrm{SE}(\beta_1)}{\mathrm{NSE}(P)}\right)^2$$

with N the initial number of subjects in the design and SE(β_1) the corresponding predicted SE of β for the same design with N subjects.

As an illustration of this method, we consider as in Section 2 two groups of 100 patients with the same elementary design (1, 3, 7, 14, 28 and 56) days after treatment initiation and derive the power of the Wald test for an alternative hypothesis increasing the first slope by 30 per cent, i.e. $\beta_1 = 0.262$. We found that $\mathrm{SE}(\beta_1) = 0.079$ and the expected power computed for a type I error $\alpha = 5$ per cent is thus of 92 per cent. We also investigate the influence of the number of subjects (40 or 100 per group) and of the value of the alternative hypothesis on the power: increase of the first slope by 30 per cent ($\beta_1 = 0.262$) or 50 per cent ($\beta_1 = 0.405$). According to the value of the increase, the influence of the number of subjects is different. Indeed, for an increase of 30 per cent, the predicted powers are of 55 per cent for 40 subjects per group and 92 per cent for 100 subjects per group, whereas an increase of 50 per cent involves less difference: a predicted power of 90 per cent for 40 subjects per group *versus* 99 per cent for 100 subjects.

For a same type I error $\alpha = 5$ per cent and an increase of the first slope of 30 per cent, we compute that only 72 subjects per group are needed to achieve a power of 80 per cent.

5.2. Design optimization to improve the power

In this section, we investigate the influence of the design on the expected SE of the treatment effect and thus on the expected power.

We optimize with the FW algorithm designs with either 6, 5 or 4 samples per subject, called Opt6, Opt5 and Opt4, respectively, for the bi-exponential HIV viral load decrease model including the treatment effect β. We fix the total number of samples to 480 and we impose identical designs in both the groups. We assume for the alternative hypothesis an increase of the first slope by 30 per cent, i.e. $\beta_1 = 0.262$. For each optimized design, the corresponding power of the Wald test for β is computed from the SE predicted by PFIM. We also compute the number of subjects and the total number of samples needed to achieve a power of 80 per cent for a type I error $\alpha = 5$ per cent. We then compare these numbers between the designs to investigate the influence of the group structure on the number of subjects needed.

Designs optimized with the FW algorithm for the model with the treatment effect are given in Table III, as also the number of samples and the number of subjects needed per group. Globally,

DOI: 10.1002/sim

Table III. Optimized designs with the Fedorov–Wynn algorithm, corresponding predicted power and total number of samples and subjects needed to achieve a power of 80 per cent. Whatever the design, optimization is performed for a total number of 480 samples and the efficiency criterion Φ_D is reported.

Design	Total number of subjects	Number of samples per subject	Optimized design {(sampling times), number of subjects}	Φ_D	SE (β)	Predicted power (per cent)	To achieve a power of 80 per cent	
							Total number of samples needed	Total number of subjects needed
Opt6	40	6	{(0, 1, 5, 14, 21, 56), 40}	471	0.124	55	845	71
Opt5	48	5	{(7, 14, 21, 56), 48}	523	0.113	64	702	71
Opt4	60	4	{(0, 5, 14, 56), 40 (0, 1, 2, 3), 10 (0, 14, 21, 56), 10}	536	0.102	73	572	72

DOI: 10.1002/sim

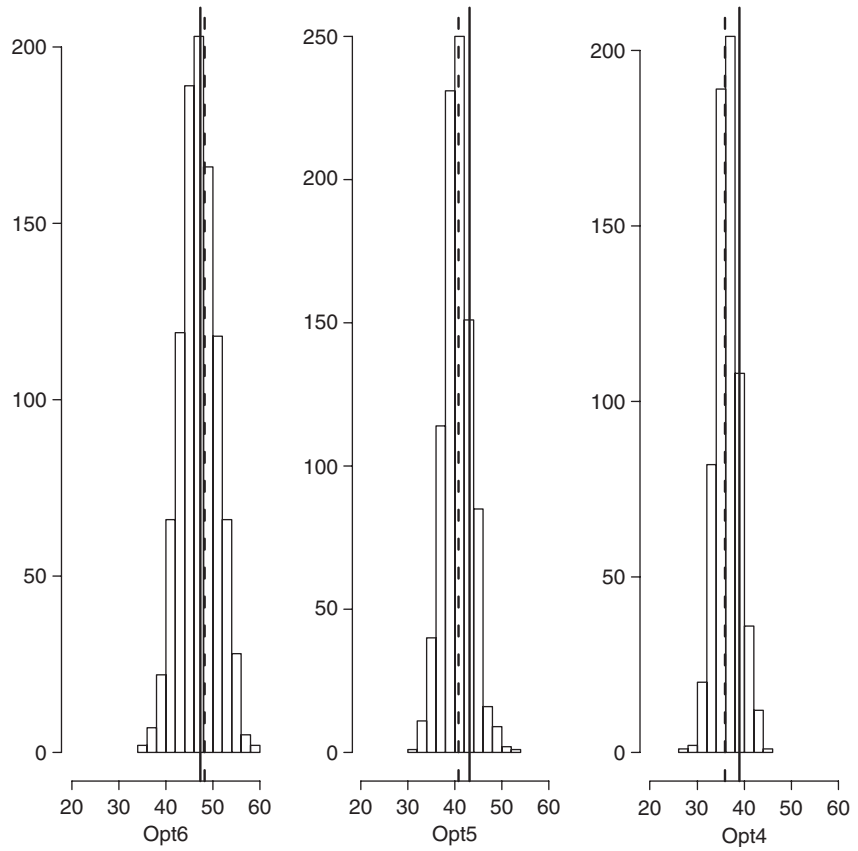

Figure 2. Histograms of the relative SE (per cent) estimated by nlme on the treatment effect β on the 997, 950 and 727 converged data files for Opt6, Opt5 and Opt4, respectively. The dotted line represents the empirical RSE, and the full line the expected RSE from PFIM.

for the same total number of 480 samples, designs with less samples per individual but more subjects are more efficient, even if the gain in efficiency is small. Moreover, as in Section 4, the group structure is more complex when the number of individual samples decreases with only one group for Opt6 and Opt5 to three groups for Opt4.

Regarding the estimation of the treatment effect, the SE predicted from PFIM decreases when the number of subjects increases. The same pattern is then observed for the power computed from this predicted SE: design optimized with 40 subjects (Opt6) leads to a predicted power of the Wald test of only 55 per cent, whereas design optimized with 60 subjects (Opt4) leads to a predicted power of 73 per cent.

To achieve the power of 80 per cent, the Opt6 design would require nearly two (about 1.8) times more samples than the initial total number of samples (845 *versus* 480 samples). Moreover, for a similar number of subjects per group (about 70), design Opt4 could achieve this power of 80 per cent but with a lower total number of samples compared to Opt6 or Opt5.

DOI: 10.1002/sim

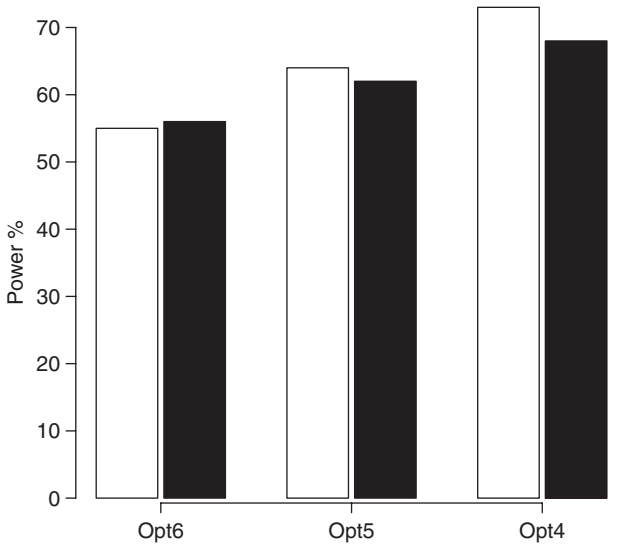

Figure 3. For each optimized designs, comparison between the predicted power obtained by PFIM (full bar) and the power of the Wald test estimated by simulation (open bar) on the converged data files (997, 950 and 727 for Opt6, Opt5 and Opt4, respectively).

5.3. Evaluation of the expected powers by simulation

For each optimized design of the previous section, a simulation is performed to investigate the relevance of the power predicted using PFIM. To do that, we simulate with R (version 1.9.0) 1000 data sets using the bi-exponential HIV viral load decrease model under H_1 with $\beta_1 = 0.262$. We then use the nlme function of Splus to estimate the population parameters on each replicated data sets, using two different sets of initial parameters. The first set is the same as the one used for the simulation except for the treatment effect which is set to 0; the second one is $\log P_1 = 5$, $\log P_2 = 2$, $\log \lambda_1 = -1$, $\log \lambda_2 = -0.1$ and $\beta = 0$. Selection of the best run from those two initial sets of parameter values is performed based on the best log-likelihood. For each simulated design, we also recorded the number of data sets for which nlme failed to converge.

From the K converged nlme estimations, we compare the expected SE given by PFIM to the distribution of the K SE computed by nlme on each population parameter, with a focus on the treatment effect parameter β. We also compare the expected SE to the empirical SE, defined as the sample estimate of the standard deviation from the K parameter estimates.

Last, the observed power of the Wald test for the treatment effect is computed on the simulations as the proportion of trials for which H_0 is rejected, and is compared to the power computed from the expected SE of PFIM.

Convergence is obtained for 99.7 and 95.0 per cent of the files for designs Opt6 and Opt5. This rate is much lower for designs Opt4 with 72.7 per cent, even using the two sets of initial parameter values.

The predicted RSE, the empirical RSE and the distribution of the RSE computed by nlme for the treatment effect parameter β are reported in Figure 2 for each optimized design. For each design, the distribution of the RSE obtained by nlme and the empirical RSE are in the same range; moreover, empirical RSE and predicted RSE from PFIM are very close. Globally, empirical RSE

DOI: 10.1002/sim

and observed RSE confirm the influence of the design on the SE already showed with PFIM: for the same total number of samples, lower SE are obtained with higher number of subjects.

The power observed by simulation and the predicted power computed by PFIM are reported in Figure 3 for each optimized design. Globally, they are similar, especially for Opt6 and Opt5. There is more discrepancy for Opt4 with fewer samples per subject but the prediction is still very good; it should be noticed that for this design, because of the lower rate of convergence of estimation on the simulated data files, the observed power is may be less reliable, not taking into account runs without convergence and may be thus overpredicting the power. Moreover, for the same total number of samples, the increase in the power, observed with PFIM when the number of subjects increases, is confirmed by the simulation.

6. DISCUSSION

In this paper, we first showed the relevance of the predicted SE computed by PFIM by comparison to those given by the SAEM algorithm, even on the treatment effect parameter. We thus illustrated the usefulness of using PFIM to predict SE and to optimize designs, as an alternative to the much more cumbersome SAEM algorithm. Indeed, the evaluation of the expected SE for the empirical design was performed in a very short computing time with PFIM (less than 1 s) compared to SAEM (about 42 min). All the computations of this paper were performed on a Pentium 4 3.20 GHz PC with the Windows operating system.

We then proposed a generic implementation of the FW algorithm and demonstrated its usefulness on the bi-exponential viral load model. Compared to the Simplex algorithm, optimizations were faster and much more robust, without any need for several runs with different initial designs. For example, optimization of the design with 5 samples per subject took about 175 s with the Simplex algorithm for one initial design whereas the same optimization requires about only 12 s with the FW algorithm. For this algorithm, the larger the set of admissible sampling times, the longer the time required for the method but not for the optimization process itself. Indeed, the computing time is almost proportional to the time needed to compute the Fisher information matrix for each possible elementary design, defined as any combination of the sampling times; the time for the optimization algorithm itself being negligible. For the design with 5 samples per subject and 12 admissible sampling times, 792 Fisher information matrices were computed in 12 s, i.e. computation of one matrix in 0.015 s. Optimization with a larger set of admissible sampling times would then require an addition of about 0.015 s for the computation of the Fisher matrix of each new possible elementary design. Here, for instance, for 16 admissible sampling times the total run-time was 67 s, which is still less than the time for the Simplex algorithm.

The specification of a finite set of sampling times with the FW algorithm can be viewed as a great advantage from a clinical point of view. Indeed, the user can specify the clinically feasible sampling times and optimize a design among them. Optimization within continuous intervals of times as with the Simplex algorithm can be more questionable: the algorithm can, sometimes during a long time, pursue the optimization by exchanging at each iteration one or several sampling times by some others which differ from the previous by only few seconds. This greatly slows down the whole optimization process and has absolutely no relevance in clinical practice.

However, although the designs obtained with both the Simplex and the FW algorithms carried a similar amount of information, the group structure was sometimes more complex with the FW algorithm, involving an additional group in the case of 4 and 3 samples per subject. One explanation

DOI: 10.1002/sim

could be the lack of flexibility in the admissible sampling times compared to the Simplex algorithm. To compensate the lack of some crucial times, such as the 17 days sampling times contained in the four optimized designs with the Simplex algorithm, the FW algorithm may increase the number of groups, adding sampling times surrounding the optimal one. Nevertheless, this 17 days time was not in the set of feasible sampling times and although the FW algorithm may involve optimized designs with a slightly more complex group structure and a minor loss of information compared to the Simplex algorithm, it may have a major gain of clinical feasibility.

Using the predicted SE of PFIM we computed also the predicted power of the Wald test to detect a treatment effect as well as the number of subjects needed to treat to achieve a given power, using the same method as in [30, 31] but taking into account the whole population parameters, including parameters for the variances of the random effects. We showed the influence of the design on the power, emphasizing that, for optimized designs, the power increases when the number of subjects increases and the number of samples per subject decreases. We confirmed these results by simulation and showed the relevance of the predicted power computed from the SE of PFIM. In practice, optimization of both the design (group structure and sampling times) and the number of subjects needed to achieve a given power can be computed in two steps. Indeed, the FW algorithm performs a statistical optimization, involving those optimal group structures and sampling times that are independent from the total number of subjects. Users can first optimize the design, and then, based on the obtained SE for the treatment effect parameter, derive the number of subjects needed to obtain the needed SE to achieve the required power.

We focused on the computation of the power for the detection of a treatment effect on the first viral decay rate parameter in the context of HIV viral load decrease using a bi-exponential model. In this context, tests on other parameters could of course be performed; for instance, Samson *et al.* [33] found a significant difference between two treatments given in a real clinical trial; the significant treatment effect was detected on the second viral decay rate although a treatment effect on the first rate was also tested. Moreover, simple values of the parameters were taken in this illustration, and real applications should consider much more relevant values obtained from previous studies. Last, although this study demonstrates that the optimal population design method allows derivation of efficient designs even with a low number of samples per subjects, further illustrations of the interest of that method would be to derive optimized designs for more complex and realistic HIV dynamics models using differential equations systems [37, 38].

Regarding the optimization methodology, we used the D-optimality criterion, looking at an efficient design for the whole parameter estimates. However, in the case of population designs optimization to improve the power, other criteria could be used to answer to this precise objective, such as the D_S-optimality criterion for the precision of estimation of some of the parameters only, for example, the treatment effect, and putting less weight on other less interesting parameters.

Last, we considered only a diagonal variance of the random effects; however, in practice, one may want to allow correlation between the random effects. Expression of the Fisher information matrix has been proposed in this case [19] but has not been yet implemented in PFIM and would certainly considerably increase the computing time because of the increased dimension of the Fisher matrix due to the estimation of these covariances.

A new version, PFIM 2.1, developed in R 2.4.1, is available at www.bichat.inserm.fr/equipes/Emi0357/download.html. It includes PFIM 1.2 and PFIMOPT 1.0 for population design evaluation and optimization, as well as for new extensions: optimization can be performed using either the Simplex algorithm or the FW algorithm; models can be entered either with analytical equations or using a system of differential equations; finally, a library of pharmacokinetics models is provided.

DOI: 10.1002/sim

To conclude, the present study further illustrates the relevance of the Fisher information matrix in nonlinear mixed effects models, even for models including treatment effect parameters. We show the great potential of PFIM to optimize population designs, as well as to control and improve the power of the Wald test.

REFERENCES

1. Sheiner LB, Rosenberg B, Melmon KL. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research* 1972; **5**:411–459.
2. Wu H, Ding AA, De Gruttola V. Estimation of HIV dynamic parameters. *Statistics in Medicine* 1998; **17**: 2463–2485.
3. Wu H, Ding AA. Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* 1999; **55**:410–418.
4. Wu H. Statistical methods for HIV dynamic studies in AIDS clinical trials. *Statistical Methods in Medical Research* 2005; **14**:171–192.
5. Ding AA, Wu H. Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics* 2001; **2**:13–29.
6. Wu H, Ding AA. Design of viral dynamics studies for efficiently assessing potency of anti-HIV therapies in AIDS clinical trials. *Biometrical Journal* 2002; **44**:175–196.
7. Pillai GC, Mentré F, Steimer JL. Non-linear mixed effects modeling—from methodology and software development to driving implementation in drug development science. *Journal of Pharmacokinetics and Pharmacodynamics* 2005; **32**:161–183.
8. Lindstrom ML, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**:673–687.
9. Beal SL, Sheiner LB. *NONMEM Users Guides*. University of California: San Francisco, 1992.
10. Pinheiro JC, Bates MD. *Mixed-Effects Models in S and S-Plus*. Springer: New York, 2000.
11. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.
12. Kuhn E, Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis* 2005; **49**:1020–1038.
13. http://www.math.u-psud.fr/∼lavielle/monolix/
14. Girard P, Mentré F. A comparison of estimation methods in nonlinear mixed effects models using a blind analysis. *Abstracts of the Annual Meeting of the Population Approach Group in Europe*, 2005; 14 [www.page-meeting.org/?abstract=834].
15. Al-Banna MK, Kelman AW, Whiting B. Experimental design and efficient parameter estimation in population pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics* 1990; **18**:347–360.
16. Jonsson EN, Wade JR, Karlsson MO. Comparison of some practical sampling strategies for population pharmacokinetic studies. *Journal of Pharmacokinetics and Biopharmaceutics* 1996; **24**:245–263.
17. Atkinson AC, Donev AN. *Optimum Experimental Designs*. Clarendon Press: Oxford, 1992.
18. Walter E, Pronzato L. *Identification of Parametric Models from Experimental Data*. Springer: New York, 1997.
19. Mentré F, Mallet A, Baccar D. Optimal design in random-effects regression models. *Biometrika* 1997; **84**:429–442.
20. Retout S, Mentré F, Bruno R. Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics. *Statistics in Medicine* 2002; **21**: 2623–2639.
21. Retout S, Mentré F. Further developments of the Fisher information matrix in nonlinear mixed effects models with evaluation in population pharmacokinetics. *Journal of Biopharmaceutical Statistics* 2003; **13**:209–227.
22. Nedelman JR. On some 'disadvantages' of the population approach. *American Association of Pharmaceutical Scientists* 2005; **7**:374–382.
23. Green B, Duffull SB. Prospective evaluation of a D-optimal designed population pharmacokinetic study. *Journal of Pharmacokinetics and Pharmacodynamics* 2003; **30**:145–161.
24. Mentré F, Dubruc C, Thenot JP. Population pharmacokinetic analysis and optimization of the experimental design for mizolastine solution in children. *Journal of Pharmacokinetics and Pharmacodynamics* 2001; **28**:299–319.
25. Retout S, Mentré F. Optimisation of individual and population designs using Splus. *Journal of Pharmacokinetics and Pharmacodynamics* 2003; **30**:417–443.

DOI: 10.1002/sim

26. Duffull S, Retout S, Mentré F. The use of simulated annealing for finding optimal population designs. *Computer Methods and Programs in Biomedicine* 2002; **69**:25–35.
27. Fedorov VV. *Theory of Optimal Experiments*. Academic Press: New York, 1972.
28. Wynn HP. Results in the construction of D-optimum experimental designs. *Journal of the Royal Statistical Society B* 1972; **34**:133–147.
29. Han C, Chaloner K. D- and c-optimal designs for exponential regression models used in viral dynamics, and other applications. *Journal of Statistical Planning and Inference* 2003; **115**:585–601.
30. Kang D, Schwartz JB, Verotta D. A sample size computation method for non-linear mixed effects models with applications to pharmacokinetics models. *Statistics in Medicine* 2004; **23**:2551–2566.
31. Kang D, Schwartz JB, Verotta D. Sample size computations for PK/PD population models. *Journal of Pharmacokinetics and Pharmacodynamics* 2005; **32**:685–701.
32. Marschner IC. Design of HIV viral dynamics studies. *Statistics in Medicine* 1998; **17**:2421–2434.
33. Samson A, Lavielle M, Mentré F. Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: application to HIV dynamics model. *Computational Statistics and Data Analysis* 2006; **51**:1562–1574.
34. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B* 1982; **44**:226–233.
35. Kiefer J, Wolfowitz J. Optimum designs in regression problems. *Annals of Mathematical Statistics* 1959; **30**: 271–294.
36. Mallet A. A maximum likelihood estimation method for random coefficient regression models. *Biometrika* 1986; **73**:645–656.
37. Neumann AU, Lam NP, Dahari H *et al*. Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science* 1998; **282**:103–107.
38. Wu H, Huang Y, Acosta EP *et al*. Modeling long-term HIV dynamics and antiretroviral response: effects of drug potency, pharmacokinetics, adherence, and drug resistance. *Journal of Acquired Immune Deficiency Syndromes* 2005; **39**:272–283.

DOI: 10.1002/sim