Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm

Julie Bertrand[1], Emmanuelle Comets[1], Céline M. Laffont[2], Marylore Chenel[3], France Mentré[1]

[1] UMR738, INSERM, Paris, France; Université Paris Diderot, Paris, France

[2] UMR181, Physiologie et Toxicologie Expérimentales INRA, ENVT, Toulouse, France

[3] Institut de Recherches Internationales Servier, Courbevoie, France

Correspondence and request for reprint to Julie Bertrand

UMR 738, INSERM, Université Paris Diderot

16 rue Henri Huchard, 75018 Paris, France.

Tel: 33 (0)1 57 27 75 39; fax: 33 (0)1 57 27 75 21

e-mail: julie.bertrand@inserm.fr

Abstract

Pharmacogenetics is now widely investigated and health institutions acknowledge its place in clinical pharmacokinetics. Our objective is to assess through a simulation study, the impact of design on the statistical performances of three different tests used for analysis of pharmacogenetic information with nonlinear mixed effects models: i) an ANOVA to test the relationship between the empirical Bayes estimates of the model parameter of interest and the genetic covariate, ii) a global Wald test to assess whether estimates for the gene effect are significant, and iii) a likelihood ratio test (LRT) between the model with and without the genetic covariate. We use the stochastic EM algorithm (SAEM) implemented in MONOLIX 2.1 software. The simulation setting is inspired from a real pharmacokinetic study. We investigate four designs with N the number of subjects and n the number of samples per subject: i) N=40/n=4, similar to the original study, ii) N=80/n=2 sorted in 4 groups, a design optimized using the PFIM software, iii) a combined design, N=20/n=4 plus N=80 with only a trough concentration and iv) N=200/n=4, to approach asymptotic conditions. We find that the ANOVA has a correct type I error estimate regardless of design, however the sparser design was optimized. The type I error of the Wald test and LRT are moderatly inflated in the designs far from the asymptotic (<10%). For each design, the corrected power is analogous for the three tests. Among the three designs with a total of 160 observations, the design N=80/n=2 optimized with PFIM provides both the lowest standard error on the effect coefficients and the best power for the Wald test and the LRT while a high shrinkage decreases the power of the ANOVA. In conclusion, a correction method should be used for model-based tests in pharmacogenetic studies with reduced sample size and/or sparse sampling and, for the same amount of samples, some designs have better power than others.

Keywords

Pharmacogenetics; Pharmacokinetics; Nonlinear mixed effects models; Test; Design; Single nucleotid polymorphism; SAEM

Introduction

Pharmacogenetics (PG) studies the influence of variations in DNA sequence on drug absorption, disposition and effects (1, 2). This area is now widely investigated and the European Medicines Agency (EMEA) has published in 2007 a reflection paper acknowledging the place of PG in clinical pharmacokinetics (PK) (3).

Pharmacogenetic data are mainly studied using non-compartmental methods followed by a one-way analysis of variance (ANOVA) on the individual parameters of interest (4). More sophisticated approaches have also been used such as NonLinear Mixed Effects Models (NLMEM). These models allow to integrate the knowledge accumulated on the drug PK, and they have the advantage of being applicable with less samples per patient.

Various methods can be used to include pharmacogenetic information in NLMEM. Preliminary screening is usually performed using ANOVA on the individual parameters estimates (5) followed by a stepwise model building approach with the likelihood ratio test (LRT) (6). As an alternative approach, a global Wald test can assess whether estimates for the genetic effect are significant (7).

In a previous work (8), we performed a simulation study to assess the statistical properties of these different approaches. We used the estimation algorithms FO and FOCE interaction (FOCE-I) implemented in the NONMEM software version V (9). In the present work, to avoid the linearisation step we use the Stochastic EM algorithm (SAEM), implemented in the MONOLIX software version 2.1 (10) for the analysis of the simulated data sets with the same three tests. SAEM computes exact maximum likelihood estimates of the model parameters using a stochastic version of the EM algorithm including a MCMC procedure.

In (8), we have simulated a design of 40 subjects inspired from a real pharmacokinetic substudy on indinavir performed during the COPHAR2-ANRS 111 trial in HIV patients (11, 12). We have also simulated the same sampling schedule but with a larger sample size of 200 subjects to be closer to the asymptotic properties of the test. Whereas the estimated type I error of the ANOVA was found to be close to 5% whatever the design, those of the Wald test and the LRT showed for the FOCE-I algorithm a slight and significant increase, respectively, for the first design with 40 subjects. In the present paper, we aim to further investigate the impact of the design on the performances of these three tests in terms of type I error and power. The EMEA has stated that pharmacogenetic studies should include a satisfactory number of patients of each geno- or phenotype in order to obtain valid correlation data (3). Therefore, with the SAEM algorithm, we also consider two other designs with a larger number of subjects but different blood sampling strategies, as extensive sampling on each patient would no longer be practical. One of these designs was optimized using the PFIM interface software version 2.1 (13, 14) and another includes a group with only trough concentrations to explore a design that is easily implemented in practice. These two designs involve the same total number of observations as the original design with 40 subjects, to allow proper comparisons between designs.

In the first section of the article, we introduce the model as well as the notations, the three tests under study and the four designs. Then, we describe the simulation study and how we perform the evaluation. Next the main results of the simulation are exposed. Finally the study results and perspectives are discussed.

Methods

Model and notations

In this work, we consider the effect on a pharmacokinetic parameter of one biallelic Single Nucleotid Polymorphism (SNP), i.e. the existence of 2 variants for a base at a given locus on the gene. We denote, without loss of generality, C the wild allele and T the mutant, leading to k=3 possible genotypes (CC, CT and TT). Let $y_{i,j}$ represents the concentration at time $t_{i,j}$ of a subject i=1,..,N with genotype G_i at measurement j=1,..,n such as:

$$y_{i,j} = f\left(t_{i,j}, G_i, \theta_i\right) + \epsilon_{i,j} \tag{1}$$

with θ_i the subject specific parameters of the nonlinear model function f and $\epsilon_{i,j}$ the residual error normally distributed with zero mean and an heteroscedastic variance $\sigma^2_{i,j}$, with:

$$\sigma^2_{i,j} = \sigma^2(a + bf\left(t_{i,j}, G_i, \theta_i\right))^c \tag{2}$$

This combined error model (additive and proportional) is commonly used in population pharmacokinetics with c fixed to 2. For identifiability purpose

σ^2 is set to one. We assume that the genetic polymorphism G_i for subject i affects θ_p, the p^{th} component of the vector θ through the following relationship:

$$\theta_{p,i} = \mu_p e^{\beta_{G_i}} e^{\eta_{p,i}} \tag{3}$$

where μ_p is the population mean for parameter θ_p and $\eta_{p,i}$ follows a Gaussian distribution with zero mean and variance ω_p^2 the p^{th} diagonal element of matrix Ω. β_{G_i} is the effect coefficient corresponding to the genotype of subject i, we assume $\beta_{G_i} = 0$, β_1 or β_2 for $G_i =$ CC, CT or TT, taking CC as the reference group.

In the following, we note M_{base} the model without a gene effect, where $\{\beta_1 = \beta_2 = 0\}$ i.e. $\{CC = CT = TT\}$, and M_{mult} the model with a multiplicative effect on the population mean of the parameter of interest, where $\{\beta_1 \neq \beta_2 \neq 0\}$ i.e. $\{CC \neq CT \neq TT\}$.

As in NLMEM the integral in the likelihood has no analytical form, specific algorithms are needed to estimate the model parameters and their standard error (SE) (15). Since the beginning of the 21st century, EM-like algorithms appear as a potent alternative to the linearisation used in the earlier approaches. The SAEM algorithm is a stochastic version of EM algorithm where the individual parameter estimates are considered as the missing values (16). The estimation step is decomposed in the simulation of the individual parameters using a Monte Carlo Markov Chain (MCMC) approach followed by the computation of stochastic approximation for some sufficient statistics of the model. The subsequent maximisation step of the sufficient statistics provides an update of the estimates. The estimation variance matrix is deduced from the NLMEM after linearisation of the function f around the conditional expectation of the individual parameters, the gradient of f being numerically computed.

The loglikelihood is obtained through importance sampling once parameter estimation is achieved, as follows. For each subject, s=1,...,T samples of individual parameters are generated from a Gaussian approximation of the subject's individual posterior distribution. These T samples are used to derive T realizations of the loglikelihood, each weighted by the probability of the corresponding sample. The importance sampling estimator is the empirical average over the weighted T realizations. The variability of this approximation decreases when increasing the number of samples T (17).

Tests

Analysis of variance (ANOVA)

The data are analysed with the model not including the gene effect, M_{base}. We used the conditional expectation (mean) of the individual parameters provided by the MCMC procedure in SAEM as the empirical Bayes estimates (EBE). Then, the equality of the mean between the three genotypes is tested with an analysis of variance. The statistic is compared to the critical value of a Fisher distribution (F-distribution) with 3-1=2 numerator degrees of freedom and N-3 denominator degrees of freedom, 3 being the number of genotypes to consider.

In our model, the log-parameters are normally distributed and the natural parameters, which have a biological meaning, are log-normally distributed. We apply the ANOVA on both the log-parameters and the natural parameters, but it is usually considered that ANOVA is rather insensitive to departure from the normal assumption as long as the observations have the same non-normal parent distribution with possibly different means (18).

Global Wald test

The data are analysed with the model including the gene effect, M_{mult}. The significance of the gene effect coefficient is assessed by the following statistic :

$$W = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}^T V^{-1} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \tag{4}$$

where V is the block for β_1 and β_2 of the estimation variance matrix. The statistic W is compared to the critical value of a χ^2 with 2 degrees of freedom.

Likelihood ratio test (LRT)

The data are analysed with M_{base} and M_{mult}. These two models are nested, thus the LRT can be used. The test statistic $-2 \times (L_{base} - L_{mult})$, where L_{base} and L_{mult} are the loglikelihood of respectively M_{base} and M_{mult}, is compared to the critical value of a χ^2 with 2 degrees of freedom, corresponding to the difference in the number of population parameters between the two models.

Study designs

We simulated data according to four designs. The first three have the same total number of observations and represent different trade-offs between the sample size N and the number of samples per patient n. The fourth design contains more subjects with many observations per patient to be closer to asymptotic conditions. Fig. 1 illustrates the differences between the four designs regarding the samples allocation in time and the sampling size. The graph is composed of four rows (one per design) on top of the pharmacokinetic profile. Within each design, the sampling times of a group are represented as linked circles of size proportional to the number of subjects in the group with this sampling time.

1) N=40/n=4

The first design is inspired from a real world example, the PK sub-study from the group of subjects receiving indinavir boosted with ritonavir b.i.d. in the COPHAR 2-ANRS 111 study, a multicentre non-comparative pilot trial of early therapeutic drug monitoring in HIV positive patients naïve of treatment (11, 12). This design includes 40 subjects with 4 samples at time 1, 3, 6 and 12 hours after the drug intake, which leads to a total of 160 observations. At the time of the study, these sampling times were empirically determined.

2) N=80/n=2

In the second design, we require 80 subjects with two samples per patient and sampling times within the set of the original design. We used the Federov-Wynn algorithm that maximizes the determinant of the Fisher information matrix within a finite set of possible designs and which is implemented in the PFIM Interface 2.1 software (13). We had to set the regression function f, the error model and a priori values of the population parameters (see Simulation study) as well as an initial guess for the population design. Regarding these constraints, the optimal design consists of 80 subjects sorted in two groups of 30 and two groups of 10 with two samples per subject respectively scheduled at 1 and 3 hours, 6 and 12 hours, 3 and 12 hours and 1 and 12 hours. This configuration provides a rather sparse design keeping a total number of observations of 160.

3) N=100/n=4,1

Third, we consider a pragmatic design with 20 subjects with the original set of sampling times (1, 3, 6 and 12 hours) and 80 subjects with only a trough concentration (12 hours) potentially collected in clinical routine. This combined design also contains a total number of observations of 160.

4) N=200/n=4

The last design includes 200 subjects having the original set of sampling times.

Simulation study

The model and parameters used for the pharmacokinetic settings come from a preliminary analysis without covariates of the indinavir data described above using the FO algorithm implemented in NONMEM (see details in (8)). The concentrations are simulated using a one compartment model at steady state with first order absorption (k_a), first order elimination (k), a diagonal matrix for the random effects and a proportional error model (a fixed to 0). The dose is set to 400 mg. The fixed effects are k_a=1.4 h^{-1}, the apparent volume of distribution V/F=102 L and k=0.2 h^{-1}, this parameterization was chosen to have only one parameter linked to the bioavailability, F. The between subjects variabilities on these parameters are respectively set to 113%, 41.3% and 26.4%. The coefficient of variation for the residual error is set to 20% (a=0, b=0.2). The first value in a series of simulated concentration below the limit of quantification (LOQ=0.02 mg/L, according to the indinavir measurement technique in the COPHAR2 trial) is set to LOQ/2 and the remaining values are discarded (19).

The genetic framework is inspired from two SNPs of the ABCB1 gene coding for the P-glycoprotein, found to have an influence on the PK of protease inhibitors (20, 21). We simulate a diplotype of SNP_1 and SNP_2 with C and G respectively the wild-type allele for the 2 exons and T the mutant allele. Their distribution mimic that of exon 26 and exon 21 of the ABCB1 gene as reported by Sakaeda et al. (22) yielding for SNP_1 unbalanced frequencies of 24%, 48% and 28% respectively for CC, CT and TT genotypes. As in the intestine, the P-glycoprotein restricts drug entry into the body we consider an effect on the drug bioavailability through the volume of distribution V/F, so that :

$$V/F_i = V/F e^{\beta_{G1_i}} e^{\delta_{G2_i}} e^{\eta_{V/F,i}} \tag{5}$$

where $G1_i$ denotes the genotype for SNP_1 and $G2_i$ the genotype for SNP_2, β_{G1_i} is 0, β_1 or β_2 if $G1_i$=CC, CT or TT and δ_{G2_i} is 0, δ_1 or δ_2 if $G2_i$=GG, GT or TT. Under the null hypothesis both $e^{\beta_{G1_i}}$ and $e^{\delta_{G2_i}}$=1, 1, 1, whereas under the alternative hypothesis, we set a genetic model of co-dominance and multiplicative effects: $e^{\beta_{G1_i}}$=1, 1.2, 1.6 and $e^{\delta_{G2_i}}$=1, 1.1, 1.3. These values were chosen to be consistent with results found in the literature for ABCB1 polymorphisms on drugs disposition (23) and provide clinically relevant effect, with V/F and CL/F ($=k \times V/F$) increasing from 105.4 to 200.5 L and 21.1 to 40.1 L/h respectively between wild and mutant homozygotes for SNP_1. In the following, tests focus on the effect of SNP_1 even if we simulated diplotypes.

For the three designs 1), 2) and 3) with the same total number of observations, 1000 data sets are simulated both under the null (H_0) and the alternative hypothesis (H_1). The design 4) with N=200/n=4 is simulated only under H_0, providing evaluation of the type I error on 1000 data sets in conditions close to asymptotic to verify the convergence of the estimation algorithm. The technical description of the simulations is given in (8). Fig. 2 represents spaghetti plots of simulated concentrations *versus* time for the three designs with a total number of observations of 160, for one simulated data set respectively under H_0 and under H_1. According to their genotype for SNP_1=CC, CT or TT, subjects curves are represented in plain, dashed or dotted lines, respectively, as well as the 12 hours sample with circles, triangles or plus for subjects of the N=100/n=4,1 design. It is not readily apparent within each column which of the two data sets includes the gene effect.

Evaluation

In this work we use the SAEM algorithm implemented in the MONOLIX software version 2.1 (10). The number of iterations during the two estimation phases and the number of Markov chains are set to provide fine convergence on one representative data set for each design under both hypotheses. Other parameters of the estimation algorithm are left to the default values.

On a given data set, the same seed is used to estimate parameters from M_{base} and M_{mult} but two different seeds are used for the importance sampling in the computation of the likelihood. A preliminary work was also performed to set the number of samples T of this importance sampling for each design. We considered 6 different values of T = 1000, 3000, 5000, 7000, 10000, 15000, 300000. For each value of T, the log-likelihood was estimated 25 times on one representative data set with both M_{base} and M_{mult} and the corresponding LRT was computed. The 25 estimations allowed us to discard any bias related to the choice of a seed as we used 5 different seeds for the random number generator at the estimation step and 5 different seeds for the random number generator at the importance sampling step. In the rest of the study, the number of samples T was set to a value that provides both a relative standard deviation on the 25 LRT estimates below 15% and moderate computing times.

Our work aims to evaluate the tests for the different designs dealing with statistical significance issues, which not necessarily imply clinical relevance (24). First, the three tests are used to detect an effect of the SNP_1 (the effect of SNP_2 is not included in these analyses) on the bioavailability through the apparent volume of distribution parameter (V/F) in the 1000 data sets simulated under H_0 for the four designs. Then, the type I error of each test is computed as the percentage of data sets where the corresponding test was significant. Based on the central limit theorem and with 5% the expectation for this percentage under H_0 the predicted interval around the type I error estimate is $[0.05\text{-}/{+}1.96 \times \sqrt{\frac{0.05\times(1-0.05))}{1000}}]$=[3.6 ; 6.4]. To ensure a type I error of 5%, we define a correction threshold as the 5^{th} percentile of the distribution of the p-values of the test under H_0.

In a second step, for the designs N=40/n=4, N=80/n=2 and N=100/n=4,1 the tests are performed using the 1000 data sets simulated under H_1. Then, the power is defined as the percentage of data sets where the corresponding test was significant. We use the corrected threshold to compute the corrected powers, to allow comparison of the different tests taking into account the type I error different from 5%. In a third step, we have computed the data sets simulated under H_1 where the test was significant and at least one of the gene effect coefficient estimates (the absolute value) was clinically relevant i.e. greater than 20%. This calculation provided us with an estimate of each test ability to detect a clinically relevant effect on V/F (and thus CL/F)(24). For the ANOVA only, one data set under H_0 and two data sets under H_1 where the number of subjects with a given SNP_1 was less than 2 were discarded from the analysis.

The ANOVA is based on the EBE for the parameter of interest, here the volume of distribution V/F. To assess the quality of the individual estimates from M_{base}, we compute the extent of the shrinkage on V/F for the four designs. A measure of the shrinkage of empirical Bayes estimates has been proposed by Savic et al. as 1 minus the ratio of the empirical standard deviation of η over the estimated standard deviation of the corresponding random effect (25). Shrinkage estimators in literature are computed with a ratio of variances shrinking the observation toward the common mean (26, 27). By analogy with these shrinkage estimators, in the present work, we define shrinkage on V/F as:

$$Sh_{\eta_{V/F}} = 1 - \frac{var\left(\eta_{V/F,i}\right)}{\omega^2_{V/F}} \tag{6}$$

where $var\left(\eta_{V/F,i}\right)$ is the empirical variance of η for the volume of distribution and $\omega^2_{V/F}$ is the estimated variance of the corresponding random effect. A shrinkage, computed on standard deviation, over 30% is considered to potentially impact on covariates testing according to (25), therefore here we consider a threshold of 50%.

We also compare the empirical SE and the distribution of the SE obtained with SAEM for β_1 and β_2 for the different designs under both hypotheses. The empirical SE is defined as the sample estimate of the standard deviation from the β_1, β_2 estimates respectively on the 1000 simulated data sets.

To adress point estimate and bias and how it may impact on the tests type I error and power, we compute the relative bias and relative root mean square error (RMSE) for V/F, $\omega^2_{V/F}$ and the residual error parameter b from M_{base} on the data sets simulated under H_0 and V/F, β_1, β_2, $\omega^2_{V/F}$ and b from M_{mult} on the data sets simulated under H_1. In addition, we have computed the relative bias and relative RMSE on the estimates obtained with FOCE-I in (8) on the N=40/n=4 and N=200/n=4 designs.

Results

The number of samples for the importance sampling, T, was set to 10000 and 15000 for the designs N=40/n=4 and N=80/n=2 and 20000 for both designs N=100/n=4,1 and N=200/n=4. SAEM achieves convergence on all data sets simulated with the four designs and each hypothesis.

Table I reports the estimated type I error for the three tests performed on the four designs. ANOVA has a correct type I error estimate for all designs with a value for the design at N=80/n=2 although close to the upper boundary. The results are analogous whether we consider the log-parameters or the natural parameters of the apparent volume of distribution (V/F), 5.5% and 5.3% respectively on the original design. The Wald test and the LRT, which are asymptotic tests, have significantly increased type I error in the three designs with a total number of observations equal to 160. Yet, the inflation remains moderate as all the estimates are below 10%. On the N=200/n=4 design, the Wald test and the LRT type I error returns to the nominal level of 5%.

The estimates for the power and the corrected power are given in Table II, for the three designs N=40/n=4, N=80/n=2 and N=100/n=4,1. Before the correction, the Wald test and the LRT appeared wrongly more powerful than ANOVA. The ability to detect a clinically relevant effect is lower than the power to detect a statistically significant effect for the ANOVA, but identical for the Wald test and the LRT. In the following, we consider only the corrected power for comparisons across tests and designs as it accounts for the type I error inflation (or reduction for the ANOVA). For each design, the corrected power is rather analogous for the three tests within each design. For the three tests, the corrected power is greater for the design optimized using PFIM, with more subjects and less sample per subjects. In classical analysis increasing N improves the power and this also applies in longitudinal data analysis up to a point. Not only must N increase, but n also should be considered as well as the sampling schedule. This trade-off was achieved through optimal design and led to a satisfactory sparse design that even ANOVA, based on EBE, can handle.

Fig. 3a displays the shrinkage for the apparent volume of distribution estimated using M_{base} on data sets simulated under H_0 and H_1 for the four designs under study. In Fig. 3b and 3c, the type I error of the ANOVA on the log-parameters is plotted versus the median shrinkage for V/F under H_0 and the power of ANOVA on the log-parameters is plotted versus the median shrinkage for V/F under H_1. The median shrinkage is lower than

40% for the design N=200/n=4 under H_0 and for the designs N=40/n=4 and N=80/n=2 under both hypotheses. Only the design with N=100/n=4,1 subjects shows shrinkage with a potential impact on covariates testing, i.e. greater than 50%. This high value of shrinkage is essentially due to the 80 subjects with one sample (median value of shrinkage around 75% for these subjects versus 21% for the other subjects with 4 samples in this design). Under the alternative hypothesis, we simulated a mixture of normals with similar variance but three different means for the individual parameters of V/F. Under both hypothesis, the shrinkage is computed using the estimates from M_{base}. Under H_1, both the empirical variance of $\eta_{V/F,i}$ and the $\omega^2_{V/F}$ estimates are larger compared to the estimates under H_0. However, the empirical variance of $\eta_{V/F,i}$ increased more than $\omega^2_{V/F}$, thus the shrinkage estimates appeared to be consistently lower under H_1. For all designs under study, the type I error estimates of ANOVA remain within the prediction interval around 5% whereas the shrinkage estimates range from 19 to 64%. We do not observe a clear relationship between the power of ANOVA and the shrinkage on V/F, but the power decreases between the sparse and the combined design. Indeed, the ANOVA obtains a corrected power of 58% when performed only on the 80 subjects with one sample from the combined design, while on the optimized design with the same N but n=2 its power was of 92.5%.

The relative Bias and RMSE for the estimated parameters are displayed in table III. SAEM and FOCE-I obtained unbiased estimates on both designs and similar relative RMSE except for V/F on the N=200/n=4 design where the expected improvement was observed only with SAEM. As the bias were null the discrepancies in RMSE across the designs arised only from the precision of estimation and the SE predicted by PFIM matched the lowest RMSE. Regarding the precision of estimation on β_1 and β_2 under both hypotheses for the designs under study in Fig. 4a, the SAEM algorithm shows good statistical properties: as expected, lower SE are observed for the design closer to asymptotic and the SE obtained with SAEM are close to their empirical value, albeit lightly under-estimated. Among the three designs with a total of 160 observations, the design N=80/n=2 provided the best performances; i.e., its empirical SE for estimates of the gene effect coefficients are the lowest. In Fig. 4b, the type I error of the Wald test is plotted versus the ratio of the median SE over the empirical SE for β_2 estimated under H_0. The under-estimation of the SE appears to be related to the type I error inflation of the Wald test as the three designs with a ratio below 0.98 have type I error estimates significantly above the nominal level. In Fig. 4c, the corrected power of the Wald test is plotted versus the empirical SE for β_2 estimated under H_1. The SE appears to be related to the power of the Wald test as it decreases as the SE increases with the highest power for the N=80/n=2 design.

Fig. 5 represents the density function of a χ^2 with 2 degrees of freedom along with a focus on the values above 5.99 (the theoretical threshold) overlaid on a histogram of the LRT statistics obtained with the four designs simulated under H_0. For the first three designs, the density curve is slightly shifted to the left compared to the histogram obtained under H_0 while for the N=200/n=4 design the superposition is complete.

Here, the corrected power of the Wald test is about 70% for the design N=40/n=4. In our previous work, we used the FOCE-I algorithm implemented in NONMEM version V (9) and we observed, for this design, a much lower corrected power of the Wald test (25%). Fig. 6 displays, the standard errors of the gene effect coefficients β_1 (left) and β_2 (right) versus their estimates when using FOCE-I (top) or SAEM (bottom). With the FOCE-I algorithm, we observe a correlation between the estimate of the gene effect coefficients and its estimation error, that we do not observe with the SAEM algorithm. Such relationship leads to decreased values of the Wald statistic and therefore reduces the power to detect a gene effect.

Discussion

In the present study, we describe the impact of four designs on the performances of three tests for a pharmacogenetic effect in NLMEM using an exact maximum likelihood approach, the SAEM algorithm.

This work follows a previous study (8) which evaluated those three tests on two designs (N=40/n=4 and N=200/n=4) using the estimation algorithms FO and FOCE-I in NONMEM version V (9). Type I error and power of Table 1 and 2 in (8) can be compared to those in Table I and II of the present paper respectively for the designs N=40/n=4 and N=200/n=4. The ANOVA in (8) was performed on the natural parameters. That simulation study has shown poor performances with the FO algorithm. The results obtained here with SAEM, in terms of type I error and power are rather similar to those obtained previously using FOCE-I, except for the Wald test. Indeed, with FOCE-I the type I error of the Wald test was still inflated on the design N=200/n=4 and the power was much lower. We hypothesised that the reduced power of the Wald approach could result partly from a poor estimation of the estimation variance matrix of the fixed effects due to the log-likelihood function linearisation, as we observed with FOCE-I a high correlation between the estimate and its estimation error. We did not meet this problem with SAEM. Besides, both algorithms obtained unbiased estimates with a similar improvement in relative RMSE on design N=200/n=4 except for V/F with FOCE-I. Moreover, FOCE-I had convergence problems for several data sets or did not provide the estimation variance matrix on design N=40/n=4 under H_1, while SAEM achieved convergence on all data sets whatever the design with the estimation variance matrix always provided. In the evaluation of model selection strategies in (8), we underlined the very poor performance of the Akaike criteria (AIC). This finding remains with SAEM (data not shown).

Other studies have evaluated by simulation the performance of tests for discrete covariate on continuous responses using NLMEM with various designs and estimation methods. The articles reporting these studies are summarized and sorted by year of publication in Table IV. Linearization based algorithms were mostly used with the exception of two recent works also using SAEM (17, 28). Furthermore, categorical covariates were always simulated in two classes, apart from one study where it was up to three classes (29) and one study with continous covariate (30).

In the present study, the ANOVA obtains the best performances with respect to type I error as no inflation is observed on the four designs, so there is no need in practice to correct the threshold for the test based on the EBE. This finding is in accordance with the results from Bonate et al. (31). Considering t-tests on individual estimates, Comets et al. observed no inflation either (32). Panhard et al. (33) obtained inflated type I error for t-tests for small n, however they studied cross-over trials where the model is fitted for each treatment separately and then the EBE are derived. With small n, the individual parameters estimates are thus shrunk toward the mean within each group, artificially increasing the statistic of the test. Analysing the whole data set, we thought that the ANOVA would be conservative in presence of sparse data, because shrinkage leads to regression of the individual parameters estimates towards the mean. Indeed, this phenomenon appears likely to reduce the test ability to discriminate means between the genotypes. In our study, the shrinkage may not have been strong enough as the sparse design was an optimal design and the one with more shrinkage had some subjects with rich design. Another advantage of ANOVA is that it requires only the model with no covariate to converge. It is noteworthy though that with unequal sample size within groups ANOVA is sensitive to heterogeneity of variances (34), this feature has not been studied in this simulation setting.

We explain the type I error inflation observed for the Wald test and the LRT by the designs with a total of 160 observations being far from the asymptotic. This result differ from those of Panhard et al. (33), Gobburu et al. (30) and Wählby et al. (29) which had similar trade-off in N an n given the number of model parameters with less that 160 observations (Table III) as well as similar interindividual variability for the parameter of interest (≈30%) and residual error variability (20-10%). Besides, Samson et al. (17) and Panhard et al. (28) observed no inflation of the type I error for these tests using SAEM for a covariate simulated in two classes with equivalent group size and at least n=6. We hypothese therefore that the departure from the asymptotic found here is related to the covariate distribution, with only 11 mutant homozygotes in average for the design N=40/n=4. Distribution of genetic covariate (from a biallelic SNP with C and T, the wild and the mutant allele) is indeed very specific; the Hardy-Weinberg proportions (35) lead to proportions of 1/4, 1/2, 1/4 for CC, CT, TT being the less unbalanced of the possible distributions. Thus, we recommend to correct the type I error of asymptotic tests for genetic polymorphism with unbalanced genotypes including small number of subjects. Furthermore, such recommendation is relevant for any other covariate with several classes and very unbalanced distribution, such as disease status or tumor classes.

For the Wald test, we relate this inflation to the under-estimation of the SE of the gene effect coefficients. Indeed, when we performed the Wald test using the empirical SE rather than the estimated SE, we observed that the type I error was then no longer significantly different from the nominal level for all designs. Panhard et al. (36) observe this relationship with FOCE-I as well and show that modelling interoccasion variability in cross-over trials leads to a better estimation of the SE of the covariate effect coefficients providing type I errors of the Wald test and the LRT close to the nominal level. Here, the SE are obtained by MONOLIX after the estimation with SAEM using a linearization of the model around the conditional expectation of the individual parameters, yet Dartois et al. (37) have also observed under-estimated SE when using the computation approach based on Louis' principle (38). With SAEM, as expected, the inflation did not worsen when increasing the number of samples per subjects as reported for FO, FOCE-I in NONMEM (39, 29, 31, 30, 32) or FOCE-I in nlme (40, 33, 36). This slight inflation can be handled using randomisation tests (41), computing the true distribution of the statistic for the data set under study and deriving a P-value. Approximate tests could also be used with degrees of freedom derived from the information in the design i.e. accounting for k, n and N (42), although there is no real consensus on how to do it for nonlinear mixed effect models. An additional advantage of the Wald test is that only the model including the covariate is required and, assuming symmetric confidence intervals, it is not a problem to test if the gene effect coefficients equal 0.

To assess the power, we have simulated a 60% increase in V/F which leads to a relevant adjustement in the dose in the TT genotypes for SNP_1; a 40% increase. There was no or slight changes in the proportion of data sets simulated under H_1 where the three tests were significant when considering for a clinically relevant genetic effect, with the exception of the ANOVA on the design N=100/n=4,1. We show the impact of the shrinkage due to the subjects with only one sample in the design N=100/n=4,1 on the ANOVA performance. In our simulation setting, the reduction in the test ability to discriminate means between the genotypes is more pronounced under the alternative hypothesis. For the N=40/n=4 design the median shrinkage was 14.2% for V/F (Fig. 3) and 30.5% and 36.2% for k_a and k respectively, thus the shrinkage should not have impacted on the power of the ANOVA had the effect been assessed on those parameters. Besides, the shrinkage was also found to be lower under the alternative hypothesis, further research on this trend would be interesting. For the Wald test, we show a direct relationship between the design, the precision of estimation for the covariate effect and the power. Indeed the design N=80/n=2 optimised using PFIM has both the lowest SE on the gene effect coefficients (β_1 and β_2) and the highest power. Our previous results with FOCE-I also underline that unbiased SE estimates are required to perform the Wald test. We should note however that we used the population model without covariate for design optimisation. Our results are in accordance with the work performed by Retout et al. (14). Indeed, they studied design optimization to improve the power of the Wald test using a model including the covariate and also found that the power increases when the number of subjects increases and the number of samples per subject decreases. For this work, Retout et al. developped the Fisher information matrix for population model with covariate. But this developpement has not yet been implemented in the available version of the PFIM software. One extension of the present work would be to investigate other criteria such as D_S-optimality criterion to design pharmacogenetic studies specifically focusing on gene effect coefficients.

In the choice of the two additional designs compared to (8) used for this simulation study, we account for practical considerations. Basically, we increased the number of subjects to fit the requirements of the EMEA (3). However, increasing the number of subjects can lead to practical issues in terms of blood sampling, as extensive sampling can not be performed in all subjects for practical reasons. Therefore, we consider two designs. First, an exploratory study where we use PFIM to define different groups with two samples per subjects within a predefined set of sampling times. This approach could be used in studies with pharmacogenetics as primary endpoint when the population pharmacokinetic model is already known; for instance, studies on pharmacokinetic evaluation of a chemical entity when the genetic variation is likely to translate into important differences in the systemic exposure. Second, a more practical study in which we use trough concentrations collected during routine monitoring as well as a small group of subjects with more extensive sampling. The latter could be a phase III or IV clinical study where genotyping will support recommendations for use in genetic subpopulations (43).

In this work, we assume that the gene effect only acts on a single parameter, the bioavailability, so we use k (the elimination constant rate) rather than CL/F in order to have only one parameter related to F, the oral volume of distribution V/F. However, population models are more commonly parameterized using CL/F, thus another perspective of this work would be to consider a gene effect on several parameters : CL/F and V/F. Besides, more than one exon control the complex pathway leading from DNA to metabolic activity. Thus, it would be interesting to investigate how model-based tests handle haplotypes(44) which lead to a larger number of unbalanced classes. Here, we could hardly consider haplotypes due to the small sample sizes. Finally, investigating genes not on the same chromosome will also raise the issue of multiple covariates.

In conclusion, the ANOVA can be applied easily and performs satisfactorily as long as the design provides low shrinkage on the parameter of interest. Whereas for asymptotic tests, a correction has to be performed on designs with unbalanced genotypes including small number of subjects. Design optimization algorithms for models with covariate are well suited and offer perspectives to handle pharmacogenetic studies but have still to be implemented in the available softwares.

Acknowledgments

We would like to thank the COPHAR 2-ANRS 111 scientific committee (investigators: Pr. D. Salmon and Dr X. Duval, pharmacology: Pr JM. Tréluyer, methodology: Pr F. Mentré) for giving us access to the pharmacogenetic data of the indinavir arm in order to build our simulations. We would also like to thank the IFR02 of INSERM and Hervé Le Nagard for the use of the "centre de biomodélisation" as well as the Pr. Marc lavielle for the precious help he provided in using MONOLIX.

During this work, Céline M. Laffont was working at the Institut de Recherches Internationales Servier as pharmacometrician and Julie Bertrand was supported by a grant from the Institut de Recherches Internationales Servier.

References

1. EMEA. ICH topic E15 definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories. Technical report, EMEA, 2008.

2. FDA. E15 definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories. Technical report, FDA, 2008.

3. EMEA. Reflection paper on the use of pharmacogenetics in the pharmacokinetic evaluation of medicinal products. Technical report, EMEA, 2007.

4. XP Hu, JM Xu, YM Hu, Q Mei, and XH Xu. Effects of CYP2C19 genetic polymorphism on the pharmacokinetics and pharmacodynamics of omeprazole in chinese people. *J. Clin. Pharmacol. Ther.*, 32:517–524, 2007.

5. D Hirt, F Mentré, A Tran, E Rey, S Auleley, D Salmon, X Duval, JM Tréluyer, and the COPHAR2-ANRS Study Group. Effect of

CYP2C19 polymorphism on nelfinavir to M8 biotransformation in HIV patients. *Br. J. Clin. Pharmacol.*, 65:548–57, 2008.

6. Y Yamasaki, I Ieiri, H Kusuhara, T Sasaki, M Kimura, H Tabuchi, Y Ando, S Irie, JA Ware, Y Nakai, S Higuchi, and Y Sugiyama. Pharmacogenetic characterization of sulfasalazine disposition based on NAT2 and ABCG2 (Bcrp) gene polymorphisms in humans. *Clin. Pharmacol. Ther.*, 84:95–103, 2008.

7. D Li, W Lu, JY Zhu, J Gao, YQ Lou, and GL Zhang. Population pharmacokinetics of tacrolimus and CYP3A5, MDR1 and IL-10 polymorphisms in adult liver transplant patients. *Clin. Pharm. Ther.*, 32:505–515, 2007.

8. J Bertrand, E Comets, and F Mentré. Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters. *J. Biopharm. Stat.*, 18:1084–1102, 2008.

9. L Sheiner and S Beal. *NONMEM Version 5.1.* University of California, NONMEM Project Group, San Francisco, 1998.

10. M Lavielle. *MONOLIX (MOdèles NOn LInéaires à effets miXtes).* MONOLIX group, Orsay, France, 2008. http://software.monolix.org/index.php.

11. X Duval, F Mentré, E Rey, S Auleley, G Peytavin, M Biour, A Métro, C Goujard, AM Taburet, C Lascoux, X Panhard, JM Tréluyer, and D Salmon. Benefit of therapeutic drug monitoring of protease inhibitors in HIV-infected patients depends on PI used in HAART regimen - ANRS 111 trial. *Fundam. Clin. Pharmacol.*, in press, 2009.

12. J Bertrand, JM Treluyer, X Panhard, A Tran, S, E Rey, D Salmon-Céron, X Duval, F Mentré, and the COPHAR2-ANRS 111 study group. Influence of pharmacogenetics on indinavir disposition and short-term response in HIV patients initiating HAART. *Eur. J. Clin. Pharmacol.*, 65:667–678, 2009.

13. S Retout, E Comets, H Le Nagard, C Bazzoli, and F Mentré. *PFIM Interface 2.1.* UMR738, INSERM, Université Paris 7, Paris, France, 2007. www.pfim.biotstat.fr.

14. S Retout, E Comets, A Samson, and F Mentré. Design in nonlinear mixed effects models: optimization using the Fedorov-Wynn algorithm

and power of the Wald test for binary covariates. *Stat. Med.*, 26:5162–5179, 2007.

15. GC Pillai, F Mentré F, and JL Steimer. Non-linear mixed effects modeling - from methodology and software development to driving implementation in drug development science. *J. Pharmacokinet. Pharmacodyn.*, 32:161–183, 2005.

16. B Deylon, M Lavielle, and E Moulines. Convergence of a stochastic approximation version of EM algorithm. *Ann. Stat.*, 27:94–128, 1999.

17. A Samson, M Lavielle, and F Mentré. The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Stat. Med.*, 26:4860–4875, 2007.

18. GEP Box and SL Andersen. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. R. Stat. Soc. Ser. B Meth.*, 17:1–34, 1955.

19. SL Beal. Ways to fit a PK model with some data below the quantification limit. *J. Pharmacokinet. Pharmacodyn.*, 28:481–504, 2001.

20. J Fellay, C Marzolini, E Meaden, D Back, T Buclin, and J Chave. Response to antiretroviral treatment in HIV-1-infected individuals with allelic variants of the multidrug resistance transporter gene MDR1: a pharmacogenetic study. *Lancet*, 359:30–36, 2002.

21. C Solas, N Simon, NP Drogoul, S Quaranta, V Frixon-Marin, V Bourgarel-Rey C Brunet, JA Gastaut JA, A Durand, B Lacarelle, and I Poizot-Martin. Minimal effect of MDR1 and CYP3A5 genetic polymorphisms on the pharmacokinetics of indinavir in HIV-infected patients. *Br. J. Clin. Pharmacol.*, 64:353–362, 2007.

22. T Sakaeda, T Nakamura, and K Okumura. MDR1 genotype-related pharmacokinetics and pharmacodynamics. *Biol. Pharm. Bull.*, 25:1391–1400, 2002.

23. C Marzolini, E Paus, T Buclin, and R B Kim. Polymorphisms in human MDR1 (p-glycoprotein): recent advances and clinical relevance. *Clin. Pharmacol. Ther.*, 75:13–33, 2003.

24. DL Sackett, RB Haynes, GH Guyatt, and P Tugwellrada. *Clinical Epidemiology. A basic Science for Clinical Medecine.* Little Brown, Boston, 2 edition, 1991.

25. MO Karlsson and RM Savic. Diagnosing model diagnostics. *Clin. Pharm. Ther.*, 82:17–20, 2007.

26. CP Robert. *The Bayesian Choice. A Decision-Theoretic Motivation.* Springer-Verlag, New York, 1994.

27. G Verbeke and G Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, 2000.

28. X Panhard and A Samson. Extension of the SAEM algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, 10:121–135, 2009.

29. U Wählby, EN Jonsson, and MO Karlsson. Assessment of actual significance levels for covariate effects in NONMEM. *J. Pharmacokinet. Pharmacodyn.*, 28:231–252, 2001.

30. JV Gobburu and J Lawrence. Application of resampling techniques to estimate exact significance levels for covariate selection during nonlinear mixed effects model building: some inferences. *Pharm. Res.*, 19:92–98, 2002.

31. PL Bonate. Covariate detection in population pharmacokinetics using partially linear mixed effects models. *Pharm. Res.*, 22:541–549, 2005.

32. E Comets and F Mentré. Evaluation of tests based on individual versus population modeling to compare dissolution curves. *J. Biopharm. Stat.*, 11:107–123, 2001.

33. 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. *Stat. Med.*, 24:1509–1524, 2005.

34. GEP Box. Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *Ann. Math. Stat.*, 25:290–302, 1954.

35. JF Crow. Hardy, Weinberg and language impediments. *Genetics*, 152:821–825, 1999.

36. X Panhard, AM Taburet, C Piketti, and F Mentré. Impact of modelling intra-subject variability on tests based on non-linear mixed-effects models in cross-over pharmacokinetic trials with application to the interaction of tenofovir on atazanavir in HIV patients. *Stat. Med.*, 26:1268–1284, 2007.

37. C Dartois, A Lemenuel-Diot, C Laveille, B Tranchand, M Tod, and P Girard. Evaluation of uncertainty parameters estimated by different population pk software and methods. *J. Pharmacokinet. Pharmacodyn.*, 34:289–311, 2007.

38. E Kuhn and M Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Stat. Data Anal.*, 49:1020–38, 2005.

39. DB White, CA Walawander, DY Liu, and TH Grasela. Evaluation of hypothesis testing for comparing two populations using NONMEM analysis. *J. Pharmacokinet. Biopharm.*, 20:295–313, 1992.

40. PI Lee. Design and power of a population pharmacokinetic study. *Pharm. Res.*, 18:75–82, 2001.

41. BFJ Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman & Hall, London, 2 edition, 1998.

42. DA Elston. Estimation of denominator degrees of freedom of F-distributions for assessing Wald statistics for fixed-effect factors in unbalanced mixed models. *Biometrics*, 54:1085–1096, 1998.

43. M Chou, J Bertrand, O Segeral, C Verstuyft, L Borand, E Comets, L Becquemont, V Ouk, F Mentré, and AM Taburet. Inter- and intra-patient variabilities in nevirapine plasma concentrations in HIV-infected cambodian patients and the effect of CYP2B6 genetic polymorphism: ANRS 12154 study. 16th Conference on Retroviruses and Opportunistic Infections, Montreal, Canada, 2009.

44. SB Gabriel, SF Schaffner, H Nguyen, JM Moore, J Roy, B Blumenstiel, J Higgins, M DeFelice, A Lochner, M Faggart, SN Liu-Cordero, C Rotimi, A Adeyemo, R Cooper, R Ward, ES Lander, MJ Daly, and D Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

45. L Zhang, SL Beal, and LB Sheiner. Simultaneous vs. sequential analysis for population PK/PD data I: best-case performance. *J. Pharmacokinet. Pharmacodyn.*, 30:387–404, 2003.

Table I. Type I error estimates (for 5% level test) on the N=40/n=4, N=80/n=2, N=100/n=4,1 and N=200/n=4 designs for each of the three tests using 1000 replicated data sets

		N=40/n=4	N=80/n=2	N=100/n=4,1	N=200/n=4
ANOVA	Log-parameters	5.5	6.2	3.8	4.2
	Natural parameters	5.3	6.4	4.3	5.0
Wald		8.9*	8.7*	8.4*	5.1
LRT		7.6*	7.8*	6.8*	5.9

*Outside the prediction interval for 5% = [3.6 − 6.4]

Table II. Power estimates without and with (Power$_{corr}$) correction for the type I error inflation under H_0 on the N=40/n=4, N=80/n=2 and N=100/n=4,1 designs for each of the three tests using 1000 replicated data sets

		N=40/n=4		N=80/n=2		N=100/n=4,1	
		Power	Power$_{corr}$	Power	Power$_{corr}$	Power	Power$_{corr}$
ANOVA	Log-parameters	75.6	74.2	93.6	92.5	80.8	82.2
	Natural parameters	71.1	70.9	93.4	91.5	78.3	79.5
Wald		81.8	73.0	95.5	92.5	85.7	81.8
LRT		78.6	73.3	94.6	92.2	82.9	79.7

Table III. Relative Bias and root mean square error (RMSE) in % evaluated from 1000 simulated data sets with M_{base} under H_0 for the volume of distribution (V/F), its interindividual variance ($\omega^2_{V/F}$) and the residual error parameter (b) for the N=40/n=4, N=80/n=2, N=100/n=4,1 and N=200/n=4 designs and from 1000 simulated data sets with M_{mult} under H_1 for V/F, β_1, β_2, $\omega^2_{V/F}$ and b for the N=40/n=4, N=80/n=2 and N=100/n=4,1 designs, using estimates from SAEM and FOCE-I when available in (8)

	Parameter	N=40/n=4		N=80/n=2	N=100/n=4,1	N=200/n=4	
		SAEM	FOCE-I	SAEM	SAEM	SAEM	FOCE-I
M_{base} **under** H_0							
Biais (%)	V/F	0.23	2.9	0.04	0.62	0.08	1.4
	$\omega^2_{V/F}$	-2.8	-0.6	0.2	-4.2	-0.8	0.7
	b	-0.3	-1.9	-3.8	-0.9	0.008	-1.8
RMSE (%)	V/F	8.6	9.5	8.5	11.8	3.8	11.1
	$\omega_{V/F}$	28.1	28.9	27.8	38.5	13.4	13.3
	b	8.8	10.3	15.8	12.4	4.0	4.8
M_{mult} **under** H_1							
Biais (%)	V/F	4.1	6.7	3.9	5.1		
	β_1	-1.0	-0.8	-2.5	-1.4		
	β_2	-1.0	-1.0	-1.8	-1.3		
	$\omega_{V/F}$	-7.5	-5.2	-1.3	-7.1		
	b	-0.6	-2.2	-3.5	-0.02		
RMSE (%)	V/F	17.9	19.2	15.0	19.9		
	β_1	19.9	20.0	15.3	18.1		
	β_2	21.7	21.7	16.5	21.3		
	$\omega_{V/F}$	29.7	29.6	26.7	39.1		
	b	9.22	10.1	16.8	13.0		

With FOCE-I, convergence was achieved and thus estimates were obtained from 969 and 950 data sets under H_0 and H_1respectively for N=40/n=4 and 978 data sets under H_0 for N=200/n=4

Table IV. Published simulation studies evaluating the performances of tests on discrete covariate effect for differents designs using nonlinear mixed effects models. ANOVA, Student and Wilcoxon tests are based on empirical Bayes estimates of the individual parameters

First author (reference)	Year	Algorithm	Software	Number of PK parameters	Design	Test	Covariate Distribution (%)	Covariate Effect	Type I error
Bertrand (present study)	2009	SAEM	MONOLIX	3	N={40,200}/n=4 N=80/n=2 N=100/n=4,1	ANOVA Wald LRT	24:48:28	1:1.2:1.6	No inflation for ANOVA Slight inflation for Wald and LRT when N={40,80,100} corrected on N=200
Panhard[1] (28)	2009	SAEM	MONOLIX	3	N={40,24}/n=10	Wald LRT	50:50	-	No inflation
Bertrand (8)	2008	FO FOCE-I	NONMEM	3	N={40,200}/n=4	ANOVA Wald LRT	24:48:28	1:1.2:1.6	No inflation for ANOVA Strong inflation for Wald and LRT with FO when N={40,200} Slight inflation for Wald and LRT when N=40 corrected on N=200 with FOCE-I
Samson (17)	2007	SAEM	MONOLIX	4	N={40,80,200}/n=6	Wald LRT	50:50	1:{1.3,1.5}	No inflation
Panhard[1] (36, 33)	2007/2005	FOCE-I	R (nlme)	3	N=12/n=10 N=24/n=5 N=40/n=3 N={24,40,60}/n=10	Student Wilcoxon Wald LRT	50:50	1:{0.8,0.875, 0.9,1.1, 1.125,1.25}	Inflation for Student and Wilcoxon when n=3, but not for n={5,10} Inflation for Wald and LRT when N={24,12}, but not for N=40 No inflation for Wald and LRT when modelling IOV

Bonate (31)	2005	FOCE	NONMEM	2	All combinations of N={50,100,150,200} /n={2,4,6}	ANOVA LRT	50:50	1:1.25	No inflation for ANOVA Inflation for LRT when n={4,6}
Zhang (45)	2003	FOCE-I	NONMEM	4	N=30/n=5	LRT	50:50	-	No inflation
Gobburu (30)	2002	FO FOCE FOCE-I	NONMEM	3	N=30/{n=5,2,(5,2)}	LRT	Continuous	-	Strong inflation with FO Sligh inflation with FOCE when n={5,(5,2)} No inflation with FOCE-I
Comets (32)	2001	FOCE FOCE-I	NONMEM	3	N=20/n=7	Wilcoxon LRT	50:50	1:1.2	No inflation for Wilcoxon (individual fits) Inflation for LRT with FOCE No inflation for LRT with FOCE-I
Lee (40)	2001	FOCE-I	R (nlme)	3	N=200/n=2 N=100/n={2,3,(5,2)}	Student LRT	90:10 80:20 70:30 60:40 50:50	1:1.3	Inflation for Student when n=5,2 Inflation for LRT when n={2,(5,2)} Inflation increases when the proportion of the subpopulation increases
Wählby (29)	2001	FO FOCE FOCE-I Laplacian Laplacian-I	NONMEM	2	All combinations of N={10,25,50,250,1000} /n={2,4,19}	LRT	98:2 90:10 75:25 33:33:33	-	Strong inflation with FOCE and Laplacian, when n={4,19} and when N={10,25,50}/n=2 with addition of N={250}/n=2 for FO Slight inflation with -I method, when N={10,25} No impact of the covariate distribution, when N=50/n=2
White (39)	1992	FO	NONMEM	2	All combinations of N={60,75,100}[2] /n={10,2}	Wald LRT	85:15[4] 70:30 50:50	1:{0.9,0.8, 0.7,0.6}	Inflation for both Wald and LRT on all designs Inflation increases when n and/or the proportion of the subpopulation increases

[1] Cross-over trials

[2] The size of the control group remains 50, only the size of the comparison group varies

Figure 1

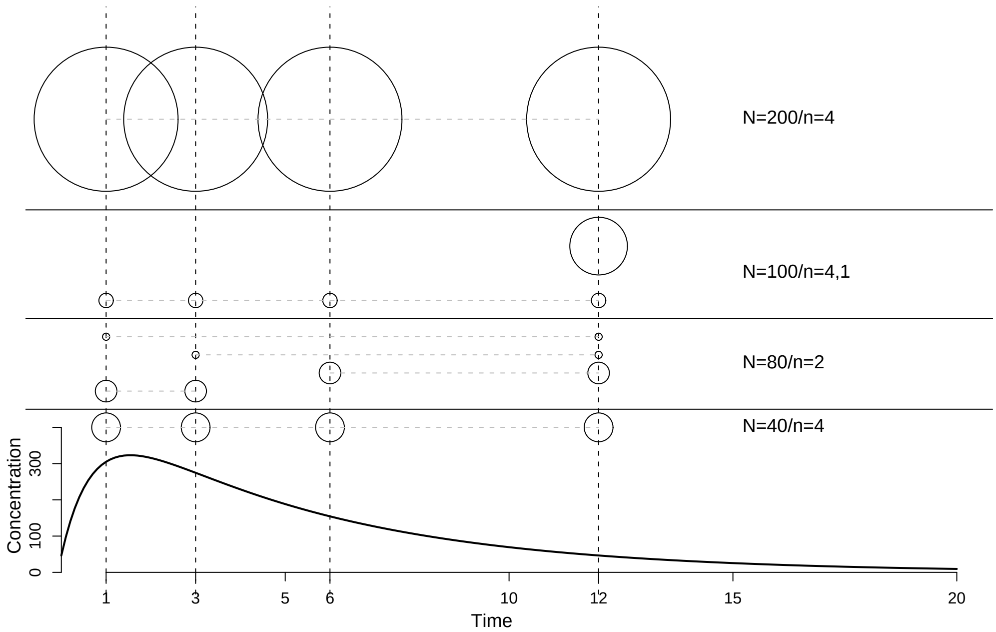
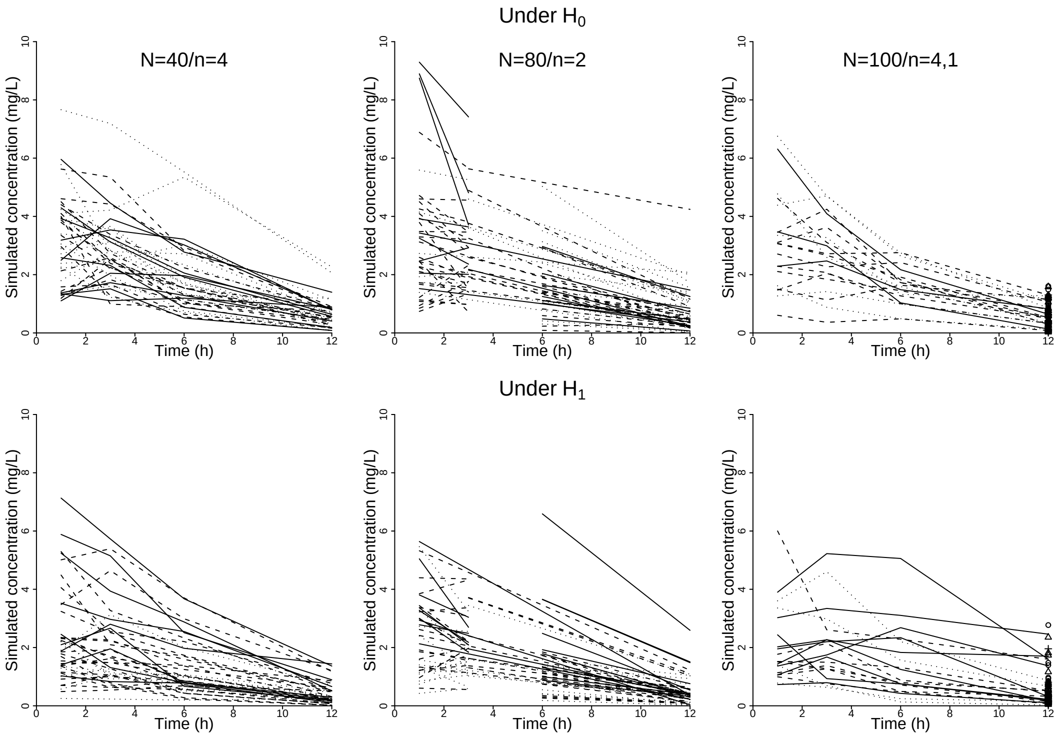
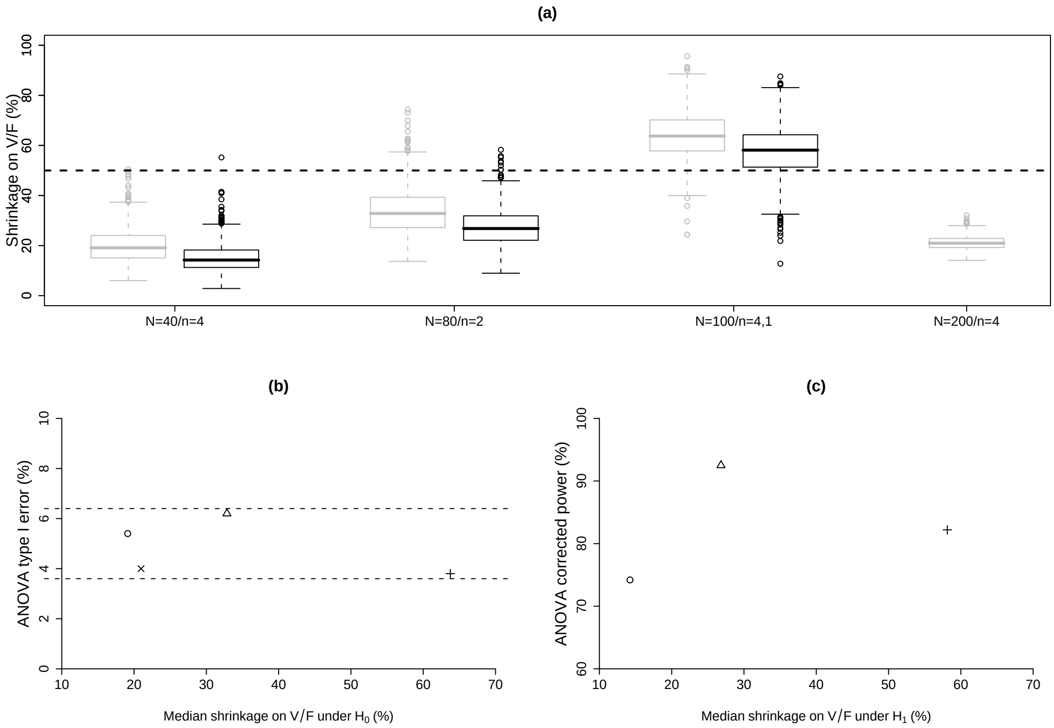

N=200/n=4
N=100/n=4,1
N=80/n=2
N=40/n=4
Concentration
0
100
300
1
3
5
6
10
12
15
20
Time

Figure 2



Under H_0
N=40/n=4
N=80/n=2
N=100/n=4,1
Under H_1
Simulated concentration (mg/L)
Time (h)

Figure 3



(a)
Shrinkage on V/F (%)
0
20
40
60
80
100
N=40/n=4
N=80/n=2
N=100/n=4,1
N=200/n=4
(b)
ANOVA type I error (%)
0
2
4
6
8
10
10
20
30
40
50
60
70
Median shrinkage on V/F under H0 (%)
(c)
ANOVA corrected power (%)
60
70
80
90
100
10
20
30
40
50
60
70
Median shrinkage on V/F under H1 (%)

Figure 4

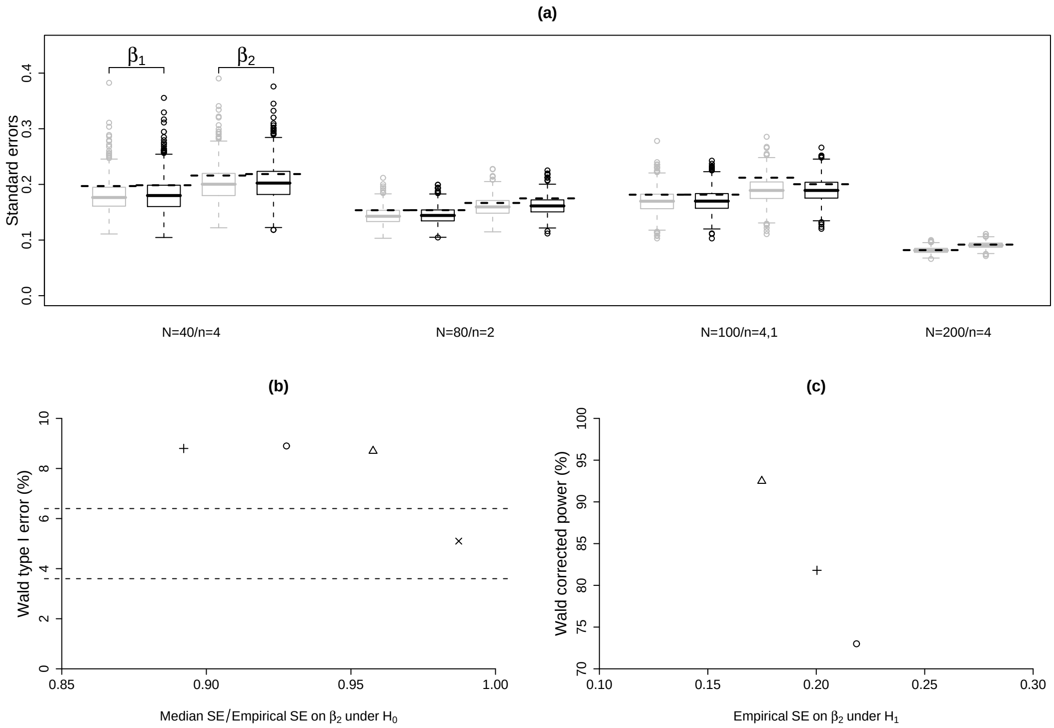
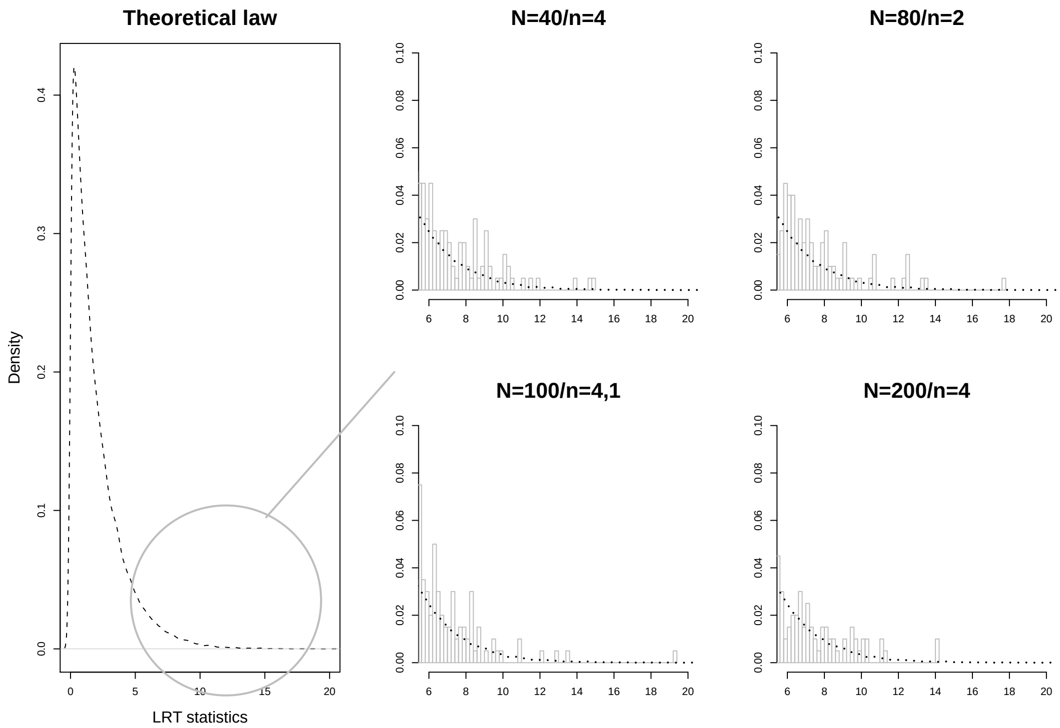

(a)
Standard errors
β1
β2
N=40/n=4
N=80/n=2
N=100/n=4,1
N=200/n=4
(b)
Wald type I error (%)
Median SE/Empirical SE on β2 under H0
(c)
Wald corrected power (%)
Empirical SE on β2 under H1

Figure 5



Theoretical law
Density
LRT statistics
N=40/n=4
N=80/n=2
N=100/n=4,1
N=200/n=4

Figure 6

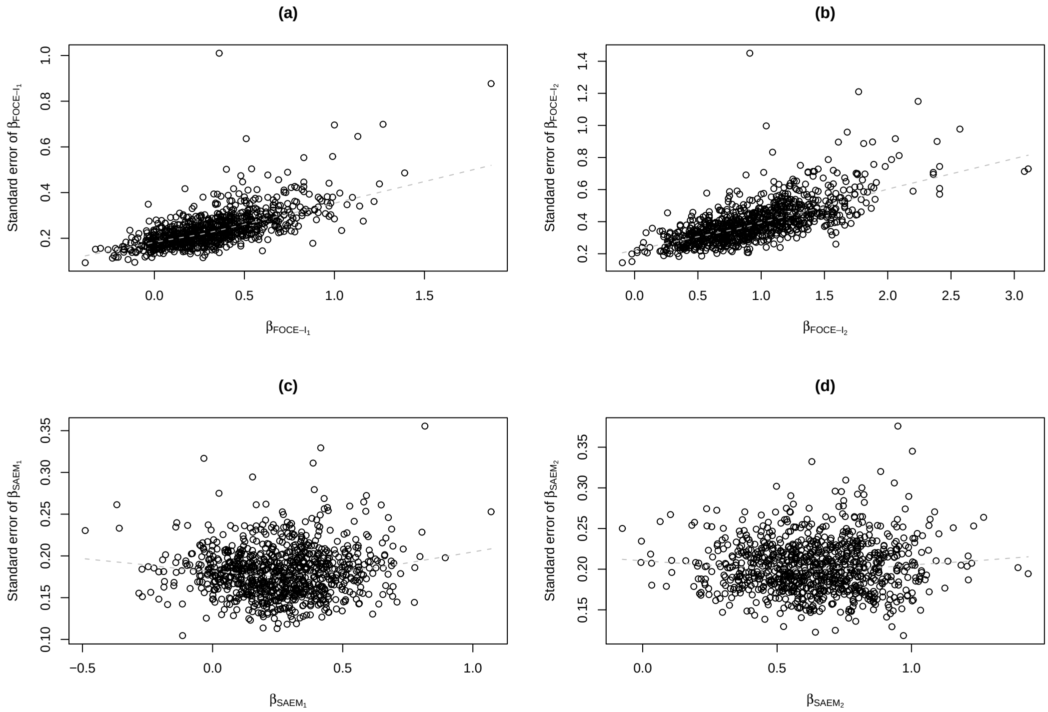


Figure captions

Fig 1. Mean simulated concentration-time curve and allocation of the sampling times within each of the designs N=40/n=4, N=80/n=2, N=100/n=4,1 and N=200/n=4 (separated by solid horizontal lines): the vertical lines denote the four possible sampling times, the dashed horizontal lines join samples within the same group and the circles size is proportional to the sample size within each elementary design.

Fig 2. Concentrations (ng/mL) simulated for the designs N=40/n=4 (left), N=80/n=2 (center) and N=100/n=4,1 (right) for a representative data set under H_0 (top) and a representative one under H_1 (bottom). Solid lines represent the subjects CC while dashed and dotted lines represent the subjects CT and TT for the exon SNP_1, respectively. For the N=100/n=4,1 design, circles represent the subjects CC while triangles and plus represent the subjects CT and TT for the exon SNP_1, respectively.

Fig 3. (a) Boxplot of shrinkage on V/F from M_{base} obtained with SAEM on the 1000 data sets simulated under H_0 (grey) and H_1 (black) for the designs N=40/n=4, N=80/n=2, N=100/n=4,1 and N=200/n=4, (b) type I error for the ANOVA on the log-parameters versus the empirical shrinkage on V/F for the designs N=40/n=4 ($\circ$), N=80/n=2 ($\triangle$), N=100/n=4,1 (+) and

N=200/n=4 ($\times$) simulated under H_0, (c) Corrected power of the ANOVA on the log-parameters versus the empirical shrinkage on V/F for the designs N=40/n=4 ($\circ$), N=80/n=2 ($\triangle$) and N=100/n=4,1 (+) simulated under H_1.

Fig 4. (a) Boxplot of the estimated standard errors (SE) and corresponding empirical SE (dotted line) obtained with SAEM for β_1 and β_2 on the 1000 data sets simulated under both H_0 (grey) and H_1 (black) for the N=40/n=4, N=80/n=2, N=100/n=4,1 and N=200/n=4 designs, (b) Wald test type I error versus the ratio of the median SE over the empirical SE for β_2 for the designs N=40/n=4 ($\circ$), N=80/n=2 ($\triangle$), N=100/n=4,1 (+) and N=200/n=4 ($\times$) simulated under H0, (c) Wald test corrected power versus the empirical SE for β_2 for the designs N=40/n=4 ($\circ$), N=80/n=2 ($\triangle$) and N=100/n=4,1 (+) simulated under H1.

Fig 5. Histograms of the likelihood ratio test (LRT) statistics above the theoretical threshold (5.99) obtained with SAEM under H_0 for the N=40/n=4, N=80/n=2, N=100/n=4,1 and N=200/n=4 designs. The dotted curve corresponds to the density of a χ^2 with 2 degrees of freedom.

Fig 6. Standard errors versus the estimates for β_1 and β_2 obtained with FOCE-I in NONMEM version V (a) and (b) and SAEM in MONOLIX version 2.1 (c) and (d) for the design N=40/n=4 simulated under H_1. Note that β_{FOCE-I_1} and β_{FOCE-I_2} correspond respectively to $e^{\beta_{SAEM_1}}$ and $e^{\beta_{SAEM_2}}$, therefore the scales are different.