Extension of NPDE for evaluation of nonlinear mixed effect models in presence of data below the quantification limit with applications to HIV dynamic model.

Thi Huyen Tram Nguyen, Emmanuelle Comets, France Mentré

Noname manuscript No.
(will be inserted by the editor)

Extension of NPDE for evaluation of nonlinear mixed effect models in presence of data below the quantification limit with applications to HIV dynamic model

the date of receipt and acceptance should be inserted later

Abstract Data below the quantification limit (BQL data) are a common challenge in data analyses using nonlinear mixed effect models (NLMEM). In the estimation step, these data can be adequately handled by several reliable methods. However, they are usually omitted or imputed at an arbitrary value in most evaluation graphs and/or methods. This can cause trends to appear in diagnostic graphs, therefore, confuse model selection and evaluation. We extended in this paper two metrics for evaluating NLMEM, prediction discrepancies (pd) and normalised prediction distribution errors (npde), to handle BQL data. For a BQL observation, the pd is randomly sampled in a uniform distribution over the interval from 0 to the probability of being BQL predicted by the model, estimated using Monte Carlo (MC) simulation. To compute npde in presence of BQL observations, we proposed to impute BQL values in both validation dataset and MC samples using their computed pd and the inverse of the distribution function. The imputed dataset and MC samples contain original data and imputed values for BQL data. These data are then decorrelated using the mean and variance - covariance matrix to compute npde. We applied these metrics on a model built to describe viral load obtained from 35 patients in the COPHAR 3 - ANRS 134 clinical trial testing a continued antiretroviral therapy. We also conducted a simulation study inspired from the real model. The proposed metrics show better behaviours than naive approaches that discard BQL data in evaluation, especially when large amounts of BQL data are present.

Keywords model evaluation · nonlinear mixed effect models · prediction discrepancies · normalised prediction distribution errors · limit of quantification · HIV dynamic model

Introduction

Nonlinear mixed effect models (NLMEM), also referred to as population analysis, have gained broad acceptance in longitudinal data analysis since their first applications to pharmacokinetic data, introduced by Sheiner *et al* in the late 1970s [1, 2]. NLMEM can help us to understand many complex nonlinear biological processes as well as the mechanisms of drug action, the different sources of variation, e.g., the interindividual variability. NLMEM are also useful to study the progression of chronic diseases under treatment such as chronic infections with hepatitis virus or HIV [3–6]. In HIV infection, the viral load is a widespread marker for the disease progression and decrease of HIV viral load is used to evaluate the efficacy of antiretroviral treatment [4, 7–9]. NLMEM are well adapted to study repeated viral load measurements and different sources of

T.H.T. Nguyen · E. Comets · F. Mentré
INSERM, UMR 738, F-75018 Paris, France
Univ Paris Diderot, Sorbonne Paris Cité, UMR 738, F-75018 Paris, France
F. Mentré
AP-HP, Hosp Bichat, Service de Biostatistique, F-75018 Paris, France
Corresponding author: T.H.T. Nguyen
E-mail: thi-huyen.nguyen@inserm.fr

variation. HIV viral load evolution can be characterised by a complex system of several differential equations [7–11] but in practice, the biphasic decline of HIV viral load under treatment can be simply described by a bi-exponential model. This model is an approximate analytical solution for the complex differential equations system obtained through additional assumptions: a constant concentration of CD4+ cells during treatment and no pharmacokinetic/pharmacodynamic (PK/PD) delay [7, 8].

NLMEM are associated with a number of assumptions with regards to the complex nonlinear model structure, variability distributions etc. It is therefore a crucial part of modeling to assess the validity of these assumptions by evaluating how well the model predicts a validation dataset. The validation dataset can be the original dataset that was used to build the model (internal evaluation) or can be another dataset (external evaluation). We define the null hypothesis, H_0, the fact that the tested model describes adequately the validation dataset. Many evaluation methods have been developed and used for assessing NLMEM. The methods based on residuals and prediction errors are the most classical diagnostic tools. However, in the context of NLMEM, these metrics show poor statistical properties, due to the linearization step required in their computation [12]. Another family of evaluation tools, Posterior Predictive Check (PPC), was proposed by Bayesian statisticians [13, 14]. These PPC metrics are also known under the name "simulation-based metrics" as simulations are required in their computation. These methods consist in comparing a statistic calculated from the validation dataset with the predictive distribution of the chosen statistic under the tested model. They are shown to present better behaviours than residuals-based methods. Several metrics belonging to this family are now widely used for evaluating NLMEM, e.g. the Visual Predictive Check (VPC) [15], the Numerical Predictive Check (NPC), the prediction discrepancies (pd) [12] or the normalised prediction distribution errors (npde) [16, 17]. In this family, the prediction discrepancy (pd), an evaluation metric proposed by Mentré and Escolano, is widely used to evaluate NLMEM in PK/PD modeling [12, 18]. In case of repeated measurements, if modelers want to perform statistical tests, npde, a decorrelated version of the metric, should be used instead of pd because pd are correlated within individuals [16–19]. These metrics are now routinely output in NONMEM and MONOLIX, the two most popular software for PK/PD modelling. A library for R (npde) is also available [17].

Data below the limit of quantification (LOQ) is a common challenge in data analysis using NLMEM. In fact,the decline of viral loads below the detection limit is a marker of the efficiency of HIV treatments. Consequently, the proportion of BQL data observed during HIV clinical studies increases with the appearance of more efficient treatments. This is a concern for parameter estimation, since it was shown that common naive approaches such as discarding all these BQL data or imputing them at an arbitrary value (LOQ or LOQ/2) could lead to biased and inaccurate estimates if the percentage of BQL data is high [20–23]. More sophisticated approaches that account for BQL data in the likelihood function were also developed and evaluated such as the M3, M4 methods [22–25], the MCEM algorithm [26] or the extended SAEM algorithm [6]. These methods were shown to allow more accurate and less biased estimates than the naive approaches [6, 22, 23, 25]. However, although BQL data are correctly handled during the estimation step, they are often omitted or imputed at an arbitrary value in most of the current evaluation methods. In this paper, we show how trends can appear in diagnostic graphs when BQL data are omitted from the plots and we extend two evaluation metrics (pd and npde) to take into account these data. As the bi-exponential HIV dynamic model has been used as illustrated example in another study on BQL data [6], we decided to use that model to illustrate the use of the new metrics and to evaluate their properties. We first applied the new metrics to evaluate a HIV dynamic model describing the HIV viral load evolution under a new anti-retroviral treatment. The HIV viral load data were obtained in the COPHAR 3 - ANRS 134 trial, a phase II clinical trial supported by the French Agency for AIDS Research [27]. A simulation study was also carried out to evaluate the graphical use of the two new metrics using diagnostic tools as well as their statistical properties (type I error and power) using statistical tests.

Statistical methods

Notations

Let i denote the i^{th} individual ($i = 1, \ldots, N$) and j the j^{th} measurement of an individual ($j = 1, \ldots, n_i$, where n_i is the number of observations for individual i). The statistical model for the observation y_{ij} in individual i at time t_{ij} is given by:

$$y_{ij} = f(t_{ij}, \theta_i) + \varepsilon_{ij} \tag{1}$$

where the function f is a (nonlinear) structural model supposed to be identical for all individuals, θ_i is the vector of the individual parameters and ε_{ij} is the residual error, which is assumed to be normal with zero mean. We assume that the variance of the error follows a combined error model:

$$Var(\varepsilon_{ij}) = (\sigma_{inter} + \sigma_{slope} \times f(t_{ij}, \theta_i))^2 \tag{2}$$

where σ_{inter} and σ_{slope} are two parameters characterizing the error model. We can also assume a constant error model in which $\sigma_{slope} = 0$ or a proportional error model where $\sigma_{inter} = 0$.

The individual parameters θ_i can be decomposed into fixed effects μ representing mean effects of the population and random effects η_i specific for each individual. We assume frequently an additive effect, e.g., for the q^{th} component of the vector θ_i:

$$\theta_{iq} = \mu_q + \eta_{iq} \tag{3}$$

or an exponential effect:

$$\theta_{iq} = \mu_q \times \exp(\eta_{iq}) \tag{4}$$

It is assumed that $\eta_i \sim N(0, \Omega)$ with Ω defined as the variance - covariance matrix so that each diagonal element ω_q^2 represents the variance of the q^{th} component of the random effect vector η_i.

We define Ψ the vector of population parameters, including the vector of fixed effect parameters μ, parameters characterizing the distribution of random effect (unknown elements of the variance - covariance matrix Ω) and parameters for residual errors (σ_{slope}, σ_{inter}), $\Psi = \{\mu, \Omega, \sigma_{slope}, \sigma_{inter}\}$.

Prediction discrepancies

The null hypothesis assumes that the validation dataset can be described by the model being tested. First, to compute the prediction discrepancy, let $p_i(y|\Psi)$ be the whole marginal predictive distribution of observations for the individual i predicted by the tested model, it is defined as:

$$p_i(y|\Psi) = \int p(y|\theta_i, \Psi) p(\theta_i|\Psi) d\theta_i \tag{5}$$

Let F_{ij} denote the cumulative distribution function of the predictive distribution $p_i(y|\Psi)$. The prediction discrepancy is defined as the percentile of an observation in the predictive distribution $p_i(y|\Psi)$, as given by:

$$pd_{ij} = F_{ij}(y_{ij}) = \int^{y_{ij}} p_i(y|\Psi) dy = \int^{y_{ij}} \int p(y|\theta_i, \Psi) p(\theta_i|\Psi) d\theta_i dy \tag{6}$$

The discrepancy can be considered as a new "type" of residuals: unlike the "classical" residual or prediction error which is the difference between the observation and a fitted or predicted value, the prediction discrepancy evaluates the position of the observation in its predictive distribution. In NLMEM, the predictive distribution $p_i(y|\Psi)$ has no analytical expression and can be approximated by MC simulation. In this method, we simulate K datasets using the design of the validation dataset and the parameters of the tested model. The prediction discrepancy can then be calculated as:

$$pd_{ij} = F_{ij}(y_{ij}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{y_{ij}^{sim(k)} < y_{ij}} \tag{7}$$

where $y_{ij}^{sim(k)}$ is the simulated observation at time t_{ij} for the i^{th} individual. Note that if $y_{ij} \leq y_{ij}^{sim(k)}$ for all $k = 1 \dots K$, then $pd_{ij} = \frac{1}{2K}$. Similarly, if $y_{ij} \geq y_{ij}^{sim(k)}$ for all $k = 1 \dots K$, then $pd_{ij} = 1 - \frac{1}{2K}$.

BQL observations are referred to as left-censored observation, denoted y_{ij}^{cens}. Using the predictive distribution, we can evaluate its probability of being under LOQ at time t_{ij} for the i^{th} individual, $Pr(y_{ij}^{cens} \leq \text{LOQ})$ predicted from the model:

$$Pr(y_{ij}^{cens} \leq \text{LOQ}) = F_{ij}(\text{LOQ}) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{y_{ij}^{sim(k)} \leq \text{LOQ}} \tag{8}$$

We propose to compute the pd for a left-censored observation y_{ij}^{cens}, pd_{ij}^{cens}, as a random sample from a uniform distribution over the interval $[0, Pr(y_{ij}^{cens} \leq \text{LOQ})]$, assuming that the model is correct.

By construction, pd are expected to follow a uniform distribution $\mathcal{U}[0, 1]$. pd can also be transformed to a normal distribution using the inverse function of the cumulative distribution function Φ of $\mathcal{N}(0, 1)$:

$$npd_{ij} = \Phi^{-1}(pd_{ij}) \tag{9}$$

where the npd are the normalised prediction discrepancies.

In case of repeated measurements, npd within an individual are correlated which increases the type I error of the test.

Prediction distribution errors

In order to decorrelate pd, Brendel *et al* [16] proposed to decorrelate the observations and predictions before computing pd by evaluating the mean $E(y_i)$ and the variance - covariance matrix $Var(y_i)$ over the K simulations under the tested model. The mean is approximated by:

$$E(y_i) \approx \frac{1}{K}\sum_{k=1}^{K} y_i^{sim(k)} \tag{10}$$

and the variance - covariance matrix is approximated by:

$$V_i = Var(y_i) \approx \frac{1}{K}\sum_{k=1}^{K} (y_i^{sim(k)} - E(y_i))(y_i^{sim(k)} - E(y_i))' \tag{11}$$

Both observed and simulated data are decorrelated as follows:

$$y_i^* = V_i^{-\frac{1}{2}}(y_i - E(y_i)) \tag{12}$$

$$y_i^{sim(k)*} = V_i^{-\frac{1}{2}}(y_i^{sim(k)} - E(y_i)) \tag{13}$$

and are then used to compute the decorrelated metric, pde, with the same formula as Eq.8

$$pde_{ij} = F_{ij}^*(y_{ij}^*) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{y_{ij}^{sim(k)*} < y_{ij}^*} \tag{14}$$

The decorrelated pd are called prediction distribution error (pde). The normalised version of this metric, denoted npde, is more frequently used in model evaluation:

$$npde_{ij} = \Phi^{-1}(pde_{ij}) \tag{15}$$

In the presence of BQL data, we cannot simply decorrelate the vector of observed data y_i as it also contains censored observations. One possible approach is to impute these censored data to a value between 0 and LOQ using the distribution of the observations predicted by the model. However, this distribution is unknown and for this reason, we use the uniform distribution of pd to impute values for BQL data as described as followed. We first impute a pd for the censored

observation by randomly sampling in $\mathcal{U}(0, p_{\text{LOQ}})$, then using the inverse function of the predictive distribution F_{ij} to get the imputed value for the BQL observation:

$$y_{ij}^{cens(new)} = F_{ij}^{-1}(pd_{ij}^{cens}) \tag{16}$$

The new vector of observations y_i^{new} contains both observed values, for non-censored data, and imputed values for censored data. As for other simulation-based evaluation methods, simulation must be performed by taking into account all factors influencing the observations, we propose to treat simulated and observed data identically. BQL values in the MC samples are therefore replaced by the same imputation method: we calculate pd_{ij}^{sim} for each y_{ij}^{sim} below LOQ in the simulated data and these y_{ij}^{sim} are replaced using the same imputation method applied on the observed data.

$$y_{ij}^{sim(new)} = F_{ij}^{-1}(pd_{ij}^{sim}) \text{ if } y_{ij}^{sim} \leq LOQ \tag{17}$$

As a result, we obtain, after this imputation step, a new vector of observations y_i^{new} and new simulated data $y_i^{sim_{new}}$. The complete data are then decorrelated using the same technique as described above.

Under the null hypothesis, we expect pde to follow the distribution $\mathcal{U}[0, 1]$ and npde to follow the distribution $\mathcal{N}(0, 1)$. However, one should bear in mind that npde are uncorrelated but not totally independent as observations in NLMEM are not Gaussian [18, 28]. To use statistical tests on npde, we have to assume that the decorrelation step renders the pde independent. This can sometimes cause type I errors to be higher than the nominal level [28]. Also, it should be noted that with the proposed method, left censored observations in the validation datasets are imputed using the model to be evaluated and this could lead to loss of power if the model is wrong.

Alternative approaches for computing pd & npde in presence of BQL data

In the previous sections, we proposed a new method for handling BQL data when computing pd and npde. One can think of other approaches to calculate pd and npde in presence of BQL observations. For instance, we could omit all BQL observations in the validation datasets and also discard all the simulated values in the MC samples corresponding to these BQL data. This approach is named "omit obs" method in the paper. Another possible approach, denoted "omit both" method, consists in discarding BQL values in both the validation dataset and the MC samples. In this method, we remove also BQL values in the simulated data in order to compare observations and simulations under the same conditions. However, removing BQL values in the simulated datasets makes the decorrelation become much more complex and the usual decorrelation method cannot be applied as the variance - covariance matrix cannot be easily calculated. For this reason, we only evaluate this method on the sparse design, where decorrelation step is not required.

Graphs & Tests

Graphs can be used to evaluate visually the distribution of npde. The npde package for R provides 4 types of graphs: (i) q-qplot of the interested metric versus the expected distribution; (ii) histogram; scatterplots (iii) versus time and (iv) versus predicted values (calculated as the mean $E(y_{ij}^{sim(k)})$ over the K simulated value of y_{ij}). Under the null hypothesis, no trend is expected to be present in the last two plots. To facilitate visual interpretation, we added in this paper 95% prediction intervals around some selected percentiles (e.g. the 10^{th}, 50^{th}, 90^{th} percentiles as in the VPC plot) of the observed npde/pd as proposed in [18]. The prediction intervals of the interested metric can be calculated using K MC datasets or by simulating K samples following the expected distribution, e.g., the standard normal distribution. Prediction band can also be added to the q-qplot of npde (pd) versus theorical quantiles of the expected distribution. As pd are correlated, prediction band for the q-qplot must be calculated from the K simulated MC datasets. If we assume that decorrelation renders npde independent, then the

prediction band for the q-qplot of npde can be calculated from K samples simulated from the $\mathcal{N}(0,1)$ distribution.

We can use a statistical test to evaluate the distribution of npde with respect to the expected distribution $\mathcal{N}(0,1)$. To test if a sample follows the $\mathcal{N}(0,1)$ distribution or not, Brendel *et al* proposed three tests [16]: (i) Wilcoxon signed rank test, to test if the mean is significantly different from 0; (ii) a Fisher test, to test whether the variance is significantly different from 1; (iii) a Shapiro-Wilk test, to test whether the distribution is significantly different from a normal distribution. The global test combines these 3 tests with a p-value corrected using Bonferroni correction for multiple tests: the p-value reported in the global test is the minimum of the p-value of three component tests multiplied by 3. We can also use an "omnibus" test such as the Kolmogorov-Smirnov to compare the distribution of npde with $\mathcal{N}(0,1)$. It should be noted that the type I error will increase when testing the npd because of correlation between individuals [12, 19].

Clinical data and HIV dynamic model

Data

The data used in this paper were collected from the COPHAR 3 - ANRS 134 clinical trial, a phase II multicentric study supported by the French Agency for AIDS Research [27]. In this study, 35 patients infected with HIV and naive to antiretroviral treatment were included and followed up for a period of 24 weeks. All patients received the same treatment with a once daily dose containing atazanavir (300 mg), ritonavir (100 mg), tenofovir disoproxil (245 mg) and emtricitabine (200 mg) during 24 weeks. Viral load was measured at the first day of treatment and at weeks 4, 8, 12, 16 and 24 (corresponding to days 28, 56, 84, 112, 168) after the initiation of treatment. If the viral load at the week 16 was higher than 200 copies/mL, another measurement was added at the week 20. The HIV RNA assays used in this multicentric study had LOQ of 40 or 50 copies/mL. Only viral load under treatment were kept in the analysis.

Methods

We used the bi-exponential model which was proposed by Ding *et al* [7, 8] and previously used in other studies to describe the dynamic of the viral load obtained in COPHAR 3 - ANRS 134 trial:

$$f(t_{ij},\theta_i) = \log_{10}(P_{1i}\mathrm{e}^{-\lambda_{1i}t_{ij}} + P_{2i}\mathrm{e}^{-\lambda_{2i}t_{ij}}) \tag{18}$$

where f is the $\log_{10}$-transformed viral load. This model contains four individual parameters θ_i: P_{1i}, P_{2i} are the baseline values of viral load and the λ_{1i}, λ_{2i} represent the biphasic viral decline rates. These parameters are positive and assumed to follow a log-normal distribution with the fixed effects $\mu = (P_1, P_2, \lambda_1, \lambda_2)$. We tested different error models in order to choose the most appropriate one. Model selection was based on the log-likelihood ratio test (for nested models) or the Bayesian information criterion (for non nested models). The parameters of the dynamic model were estimated using the extended SAEM algorithm to take into account BQL data (see more about this algorithm in [6]). An additional variable were used to indicate censoring. Parameters were estimated using the MONOLIX software, version 3.2.

Results

A total of 211 measurements of viral load obtained in 35 patients were used for model building, of which, 102 observations (48.3%) were below the LOQ (40 or 50 copies/mL). The spaghetti plot of viral load data in logarithmic scale versus time is presented in Fig. 1.

We first tested and compared several residual error models on the simplest interindividual variability model where no random effect correlation was present. A constant error model was

finally selected. Secondly, we examined different interindividual variability model with or without correlation terms between random effects and a correlation coefficient between the random effects of the parameters P_1, P_2 was found to be significant (with a gain in BIC = 24.3). The optimal statistical model is composed of (i) a bi-exponential structural model, (ii) an exponential model for interindividual variability with correlation between the random effects of P_1, P_2, and (iii) a constant error model on $\log_{10}$ viral load. The parameter estimates for the final model are presented in Table 1. Parameters were well estimated with the relative standard errors about 30% for fixed effect and smaller than 50% for variability. The relative standard error for the fixed effect of P_1 was a little higher than 30% but this can be explained by its large interindividual variability. We also found a high correlation between the random effects of the two parameters P_1 and P_2 ($\rho = 0.76$).

Figure 2 presents some examples for individual fits provided by MONOLIX 3.2. In general, we obtain very good fits for observed data. However, we cannot evaluate how well the model works for BQL data region as in the plots of individual fits, all these data are imputed at the LOQ value.

Some of goodness-of-fit plots provided by MONOLIX of the final model are shown in Fig. 3. The plots of observations versus population predictions and versus individual predictions indicate that the model adequately describes data above LOQ. However, clear trends which is a consequence of imputing BQL data at an arbitrary value can be observed in the BQL data region and prevent to evaluate any model misspecification.

The residual plots are shown in Fig. 4. The residuals (weighted residuals calculated using population predictions WRES, weighted residuals calculated using individual predictions IWRES and npde) seem to distribute homogenously around 0 in the early times when the viral load remains detectable by the HIV RNA assay. However, at later times, where BQL data appear more frequently, we can see important departures of the residuals from 0, that could be explained by the omission of BQL data in residual plots because trends only appear in later times where BQL data are present at high proportions.

The model is also assessed by a visual predictive check (VPC) (Fig.5). Two types of VPC were provided by MONOLIX 3.2 to evaluate the model with respect to both data above and under LOQ. Figure 5(a) is the classical VPC plot, showing the 10^{th}, 50^{th}, 90^{th} percentiles for observed data over time and their corresponding 90% prediction intervals calculated from K Monte Carlo samples (simulated using the model, the parameter estimates and the design of the building dataset). In the VPC plot, the percentiles of observed data are expected to be within the corresponding prediction bands. This VPC graph shows a good predictive performance of the model for observed data above LOQ. However, as BQL data in the building dataset as well as in K MC samples are imputed at LOQ values, the three observed percentiles and their corresponding prediction bands cannot be correctly calculated. This renders model evaluation in BQL data regions meaningless. Figure 5(b) shows the 90% prediction interval for the observed cummulative proportion of BQL data over time. The fact that this proportion of BQL data remains within the 90% prediction interval supports the choice of final model.

We next illustrated the use of new approach to take into account BQL data when calculating npde on the real data for the final model. Figure 6 shows different graphs, npde versus time and q-qplot of npde versus the distribution $\mathcal{N}(0, 1)$, calculated by the "omit obs" or new methods. Once again, we can see that omission of BQL observations in the validation dataset introduces trends in the scatterplot of npde versus time and the model begin to fail in regions where BQL data are frequently observed. Taking into account BQL data by the new approach, trends found in the scatterplot of classical npde disappear in the graph of new npde versus time. Q-qplot and results of different tests performed on the two npde samples show no departure of npde from the normal distribution $\mathcal{N}(0, 1)$ in either case; with the original npde however, we could not rule out a loss of power since about half of the observations were omitted (48.3%).

Evaluation by simulation

Simulation setting

The model used in the simulation study is inspired from the bi-exponential model built for real data obtained in the COPHAR 3 - ANRS trial. To simulate data under alternative hypotheses (i.e., under false models), we modified the value of the fixed effect and the variability of the parameter λ_2, a parameter having direct impact on the percentage of BQL data, of the "true" model: two false models were obtained by multiplying or dividing λ_2 by 2 and two other false models were generated by multiplying or dividing the variability of λ_2 (ω_{λ_2}) by 3. The parameters used for simulating these models are shown in Table 2.

We chose the main sampling times of the COPHAR 3 - ANRS 134 trial in our simulation study: viral loads were measured on day 0, 28, 56, 84, 112, 168 after the beginning of treatment. In this simulation, we did not simulate the adaptive design which was used in the COPHAR 3 - ANRS 134 trial, where the measurements at week 20 (day 140) depend on the viral load measured at week 16 (day 112). Therefore, we removed the measurement at week 20 in our sampling schedule. We used two designs, both with a total number of 300 observations and with the same number of observations (50 observations) at each time. The first design, called "sparse" design D_{sparse}, is composed of a total of 300 observations in 300 patients, 1 observation per individual. This design, which was first proposed and used by Mentré *et al* in a publication on pd [12], is certainly not realistic but is interesting to study the statistical properties of pd in the absence of within-subject correlations. The second, termed "rich" design D_{rich}, is closer to the real life data with 50 patients, each of them having 6 sampling times. This design was used to study the impact of within individual correlations on the type I errors of pd and to evaluate the properties of the decorrelated metric, npde. Datasets were simulated under H_0 with the true model and under alternative assumptions H_1 with the false models described in the previous paragraph. To study the impact of intra-individual correlation caused by random effects on type I errors, two variability settings S were used: in the first setting, S_{high}, the variability of the two parameters P_1, P_2 are close to those obtained from real data (2.1 and 1.4, respectively); in the second setting, S_{low}, the variability for these parameters are both decreased to 0.3. For datasets simulated with the sparse design, because only one observation is measured for each patient, we cannot study how the two variability settings affect within individual correlation using this design. Therefore, for the sparse design, we chose to use only the high variability setting, which is closer to real life data. Contrarily, for datasets generated with the rich design, both variability settings were used to study the impact of within individual correlation on the new metrics.

In pratical modeling, we can encounter another type of model misspecification: the structural misspecification. For example, data characterising a one-compartment model can be modelled using a two-compartment model. This situation can become true in HIV treatment if a very effective treatment is discovered someday. For this reason, we also want to evaluate the ability of the new method to detect this kind of false model. For this purpose, we chose to simulate a HIV mono-exponential model, V_{mono}, with the following parameters: $P_1 = 10000$ copies/mL, $\lambda_1 = 0.1\ day^{-1}$, $\omega_{P_1} = 2.1$, $\omega_{\lambda_1} = 0.3$ and $\sigma_{inter} = 0.14$. These parameters are close to those obtained by fitting the real COPHAR 3 data with the mono-exponential model with ω_{P_1}, ω_{λ_1} and σ_{inter} fixed at the values used for simulating bi-exponential model.

We defined names for different validation datasets, simulated from true model V_t, and from false models V_{fix_1}, V_{fix_2}, V_{var_1}, V_{var_2}, V_{mono}. The last false model, V_{mono}, is only simulated with the rich design. A set of 1000 validation datasets were then simulated for each scenario. To calculate pd and npde, K = 1000 Monte Carlo samples were simulated with the true model and the design of the underlying validation dataset.

Each simulated dataset was further studied under 3 settings: uncensored, and with two levels of censoring (LOQ = 20 or LOQ = 50 copies/mL), thus generating 3 datasets with different percentages of BQL observations. This allows us to study the behaviour of these new metrics in different conditions. Table 3 shows the proportion of BQL data under the three censoring schemes.

Evaluation of the new npde

To illustrate visually the use of new metric, we first used one randomly selected dataset simulated with the rich design for each scenario under the null hypothesis. The spaghetti plots of these validation datasets are shown in Fig. 7. The figure also shows a randomly selected dataset simulated under each alternative assumption.

We then evaluated the type I errors of chosen statistical tests on the new metric, using 1000 simulated datasets for each scenario. Similarly, under alternative assumptions, we estimated the power of the different tests on new npde using 1000 simulated datasets to detect the corresponding model misspecification. All computations were performed using the statistical software R, version 2.12.2.

Results

Model evaluation under null hypothesis H_0

Graphical illustration Scatterplots of npde calculated using the "omit obs" and new methods vs time are shown in Fig. 8 for one dataset simulated under the null hypothesis H_0, i.e. with the true model, for each of the two settings S_{high} (high variability, top two lines) and S_{low} (low variability, bottom two lines). For each dataset the value of the LOQ used for censoring increases from left (no censoring) to right (LOQ = 20 then 50 copies/mL). This figure clearly shows the necessity of taking these data into consideration in the evaluation step. Indeed, while the model does not show misspecification on the full dataset, trends start appearing in the scatterplots with the "omit obs" npde when a part of the dataset is omitted. These trends disappear with the new npde when BQL data are imputed (red dots).

Figure 9 shows the result of the imputation directly on the data, for the same datasets as in Fig. 7. The red dots represent imputed data for observations which were censored, and comparing both figures shows that the imputation step allows to reconstruct a dataset very similar to the original.

Type I error For the sparse design ($N = 300$, $n = 1$), npd are identical to npde. For the rich design ($N = 50$, $n = 6$), performing tests on npd calculated for rich design results in elevated type I errors: 64.7% (S_{high} setting) and 35.0% (S_{low} setting) compared to 5%. It is not surprising that the type I error is higher with the S_{high} setting as larger values of interindividual variability used in this setting induce a high correlation within individuals. Table 4 presents the type I error of several statistical tests (Wilcoxon, Fisher, Shapiro-Wilk, global and KS tests) performed for the npde calculated using different methods, under the three designs D_{sparse}, D_{rich} (S_{high} and S_{low}). As expected, in the absence of BQL data, the type I errors of different statistical tests for npde (or npd for sparse design) are close to the significant level (5%). In the presence of BQL data, omitting BQL values (with "omit obs" method) results in a significant increase of type I errors for all designs. The increase in the type I error for the npde obtained with the "omit obs" method under censoring is particularly spectacular under D_{sparse}, because discarding one BQL observation means discarding a whole individual. Using the "omit both" method, the type I error is satisfactory. For the new npde, the type I error only increases under D_{rich} in the S_{high} setting, where the interindividual variability is highest. The inflation of the type I errors of the Fisher tests are thought to be the result of the imputation method and the decorrelation step. However, type I errors of the global test remain much lower than those obtained by omitting all BQL data and is very close to the theoretical level.

Model evaluation under alternative assumptions

Graphical illustration The scatterplots of npde versus time, calculated by the new methods, for different false models (V_{fix_1}, V_{fix_2}, V_{var_1}, V_{var_2}), are shown in Fig. 10. In absence of BQL data (first column), under the two alternative hypotheses basing on modifications of fixed effect (V_{fix_1}, V_{fix_2}), the scatterplots of npde versus time show a clear trend, which becomes more important

in later times. These trends allow us to detect model misspecification in the second decay phase, which is mainly characterised by parameter λ_2 (upper two lines). Under the hypothesis V_{var_1}, npde are able to detect model deficiencies caused by an increase of the variability of λ_2 as their 10^{th} percentile falls out of its prediction interval. However, npde appear less sensitive to detect a decrease of variability, even when there is no BQL data (last line).

These trends in the scatterplots of the new npde versus time remain in presence of BQL data, under the two assumptions V_{fix_1}, V_{fix_2}. Nevertheless, as the BQL fraction increases over the time course, the percentiles of npde tend to return to their prediction intervals, especially in regions with a high fraction of BQL data. Thus a high level of BQL data is thought to cause some loss of power. Also, we find in the datasets simulated with V_{var_1} that trends can be seen in the S_{low} setting but not in the S_{high} setting; this figure is not shown here as the difference was only observed under the assumption V_{var_1}. In fact, it is logic that an increase of variability can be detected more easily on a system of lower interindividual variability. Moreover, the dataset V_{var_1} simulated using S_{low} setting contains less BQL data than those obtained with S_{high} (see Table 3). Therefore, model deviation due to modification of variability is more evident with S_{low} setting. Finally, on the datasets simulated assuming a decrease of variability of λ_2 (V_{var_2}), the new npde is not able to detect model misspecification, as their observed percentiles are within the corresponding prediction intervals, but this is also the case in the absence of BQL data. Spaghetti plots of imputed data for false models were shown in the last four rows of Fig. 9. As BQL data are imputed using the true model, it is not suprising that the imputed datasets are more comparable to datasets simulated under true model than the corresponding validation datasets. The difference between imputed data and the initial datasets can be more clearly observed in datasets simulated with increased variability.

Figure 11 displays the spaghetti plots of simulated and imputed data as well as the scatterplots of the new npde versus time for the false structure model V_{mono}. In absence of BQL data, npde allow us to correctly detect a misspecification of the structural model. When the BQL data fraction increases, the new npde still allow us to identify this kind of misspecification but they seem to be less sensitive. This phenomenon is also observed under other false models. It is also predictable that the spaghetti plots of the imputed data (upper pattern, two last columns) appear to be different to the original simulated dataset as the true model was used to impute BQL data.

Power We examined next the powers of statistical tests for the new metrics under several alternative assumptions (V_{fix_1}, V_{fix_2}, V_{var_1}, V_{var_2}, V_{mono}). We also evaluated the power of the "omit both" method under the assumption V_{fix_1} as the type I error obtained with this method is also satisfactory. The results are given in Table 5.

In the absence of BQL data, for all the designs, systematic deviations due to changes in fixed effect parameter (V_{fix_1}, V_{fix_2}) can be detected by the global test and the KS test with very high powers (100%). The increase in the variability of λ_2 (V_{var_1}) can also be detected by KS or global tests on npd or npde with very satisfactory powers when there are no BQL data (100% with the global test and greater than 93% with the KS test). On the contrary, the powers to detect model deficiencies due to a decrease of variability of λ_2 (V_{var_2}) are much lower even in absence of BQL observations. The power of the five tests used for detecting a false structure model, V_{mono}, is very high, which indicates that npde are a good tool for evaluating this type of model misspecification.

In presence of BQL data, using the new method, the powers of the global and KS tests to detect model deviations caused by changes in fixed effect parameter (V_{fix_1}, V_{fix_2}) remain very satisfactory (greater than 99%), even when the fraction of BQL data becomes more important. Nevertheless, a very slight loss of power (around 1%) can be observed when LOQ value is raised from 20 to 50 copies/mL. Under the third alternative assumption (V_{var_1}), the power to detect a modification of variability of λ_2 in datasets simulated using S_{high} setting decreases rapidly when BQL data fraction increases. The more BQL observations become frequent in the validation datasets, the more important is the loss of power. This can be clearly observed for both D_{sparse} and D_{rich}, S_{high}. For D_{rich}, S_{low}, the power to detect the increase of ω_{λ_2} remains very satisfactory, even when BQL data are present at very high percentages (100% at LOQ = 20 copies/mL and 99.4% at LOQ = 50 copies/mL). This illustrates that an important change of variability can be more easily detected when interindividual variability is low. Under the assumption V_{var_2}, the unsatisfactory powers to detect a model deviation could be expected, especially since even when

no BQL data are observed, the power was already very low. Under the alternative assumption based on a false structure model V_{mono}, the power remains very high even when BQL data are present at a very high proportion. This high power indicates that the new npde can be used to detect structure model misspecification, a common issue in practical modeling when performing internal evaluation. We notice that in all cases, as the BQL data fraction increases, the power to detect model misspecification decreases, as expected.

Using the "omit both" method, the power for detecting an important change in fixed effect parameter of the false model V_{fix_1} in presence of BQL data is much lower than the one obtained with the proposed method for all the test (Table 5). The loss of power when the BQL data proportion increases is also larger in comparison with the new method. Therefore, we did not try to evaluate power of the "omit both" method under other alternative assumptions or under the two rich designs.

Discussion

Evaluation of NLMEM is a crucial issue in population modelling. Therefore, numerous diagnostic tools have been developed and used to examine the adequation of these models. However, most recent evaluation methods are not yet developed to take into account BQL data, even though they are correctly handled in the modelling step by reliable estimation methods. Frequently, in many "goodness-of-fit" graphs, BQL data are discarded or imputed at an arbitrary value such as the LOQ. We have demonstrated in this paper that omitting BQL observations in the validation dataset can introduce patterns or trends in diagnostic graphs when BQL data are present at non-ignorable proportions and thus, could lead to wrong conclusions concerning model adequacy. It is therefore essential to develop new approaches to account for BQL data in the evaluation step.

We focused here on prediction discrepancies (pd) and normalised prediction distribution errors (npde). The two metrics are now widely used to evaluate PK/PD models. Like other diagnostic tools, the recent version of these metrics does not take BQL data into consideration. We proposed in this paper new methods to extend the two metrics to handle BQL data.

We first applied the new methods to evaluate a model obtained from real data. Using the new methods, trends in scatterplots of the classical npde (calculated using the "omit obs" method) disappear from the graph of the new metrics. However, as always in real life, we do not know whether the model is in fact correct for describing the datasets. For this reason, we carried out a simulation study in order to correctly evaluate the properties of the two extended metrics.

The new npde appear to be a promising diagnostic tool in the presence of BQL observations. Indeed, the simulation under H_0 shows very satisfactory type I errors that are close to the significant level for validation datasets with uncorrelated observations. Using the new npde on the datasets simulated with S_{high}, we obtain type I errors that are close to 5%, except for some values that are slightly higher than 5%. Unexpectedly, for about 9% of 1000 datasets, the Fisher test become significant and this value (9%) falls out of the 95% confidence interval of 5%. One possible explanation for the increase of type I errors of the Fisher test is the use of the proposed imputation method. In this method, BQL values if exist in the simulated data (MC samples) are also replaced using the same imputation method applied on the validation data and this step can cause the variance - covariance terms of the observed data to be modified. Indeed, simulated data in each of K MC samples, which correspond to data above LOQ in the validation dataset, may contain some values below LOQ and these BQL values are replaced independently using the imputation method. Thus, when a large number of values in MC samples corresponding to the observed data in the validation dataset are being imputed, the variance - covariance matrix estimated by the imputed MC samples may not reflex correctly the dispersion of the validation data, and change the distribution of the resulting npde. To test that explanation, we used a lower variability setting to study whether within subject correlation could have influence on the variance of npde. For datasets obtained with S_{low} setting, due to low variability data, MC samples are much closer to the validation dataset or, in other words, the MC samples are more likely to have a similar number of data above or below LOQ as those of the validation dataset, in comparison with those simulated with S_{high} setting, and we obtained lower type I errors of the Fisher test

with S_{low} setting. Hence, the increase of type I errors of the Fisher test is probably a result of the imputation method. However, in our opinion, we should keep treating identically the validation data and simulated data, i.e., applying the same imputation method on both observations and simulations. In fact, in any simulation-based method, we should account for all the factors that can have impact on the generation of the validation dataset during the simulation process. For example, if an adaptive design was used then this step should be reproduced during simulation; otherwise, model evaluation could be misleading [29]. In the proposed approach, we did impute BQL values in the validation dataset using a special method. For this reason, we should reproduce the same process on the simulated data. One possible approach to overcome the limitation of the present imputation method (high type I errors of the Fisher test) is to develop a new imputation method that can take into account correlations with data above LOQ within subject.

It should also be noted that the design used in the COPHAR 3 - ANRS 134 trial is, in fact, an adaptive design as the observations at week 20 are conditionnal on the observations at week 16. However, in this trial, there were only 5 over 35 patients having an additional measurement at week 20 (i.e. 5 samples out of 211 observations). Therefore, we considered that the adaptive design has little impact on the dataset. For this reason, we did not try to reproduce this process during simulation step to compute npde and VPC for the real data. Moreover, the objective of the study is to evaluate a new method for BQL data and hence, we did not simulate adaptive design (as in the COPHAR 3 trial) and assumed that there was no measurement at week 20 in our simulation study.

In spite of some problems of the Fisher tests, type I errors for the npde computed by the new method, especially those of the global test, are very close to 5% and much more lower than those obtained by the "omit obs" method or than those of tests performed on non-decorrelated npd. The type I error of the global test obtained with the "omit both" method is also satisfactory in the presence of BQL data. However, the power of this method to dectect an increased fixed effect is much lower than those obtained with the new method. Contrarily, the power to detect changes in fixed effect parameters of the new npde is very high, even at very high proportion of BQL data. Moreover, we observed an more important loss of power with the "omit both" method when the fraction of BQL data inscreases. The low power of the "omit both" method is the consequence of the loss of available information due to removing BQL data. In the new method, we proposed a method to impute values for BQL data, which means, to add a certain quantity of information into the dataset. This can lead to a gain of power when testing. However, as the information added is based on the use of the model to be validated, a loss of power is expected when BQL data increase. In addition to that, the "omit both" method is likely to have some disadvantages. First of all, a part of simulations which contains BQL values is discarded in the censoring step and only the remaining part of the simulation is used to calculate pd. As a consequence, if a model predicts many BQL values, then we lose an important number of simulations. In this case, pd may not be correctly calculated as the predictive distribution is not well approximated if we do not increase the number of MC samples. Another question is how many simulations we have to increase to assure that pd is correctly calculated. This should not only depend on the number of subjects, of observations but also depends on models and LOQ levels. The second disadvantage is that, when calculating npde for an individual, by removing BQL values in the simulations, we may not obtain vectors of the same length for each time point (each vector corresponds to the simulations at each time point). Thus, the variance - covariance matrix cannot be easily calculated and another method for decorrelating observations/simulations must be investigated.

The high power to detect modifications in fixed effects (V_{fix_1}, V_{fix_2}) and also misspecifications in structure model (V_{mono}) of the new npde indicates that the extended metrics are probably good diagnostic tools for checking the structural model in presence of BQL data. On the contrary, the powers to detect deficiencies of variability model are lower, especially much lower for detecting a decreased variability. This is consistent with the results of previous studies concerning PPC metrics in general, showing that it is difficult to detect a decrease of variability of a parameter [12, 16]. In all cases, we observed a loss of power (more or less important) as the BQL data fraction increases. This behaviour can be anticipated from the proposed methods. Indeed, our method uses the model to compute pd and/or npde: first, to compute the probability to be below LOQ which is used to quantify pd and/or to obtain the predictive distribution for each observation and the inverse of that distribution for imputing BQL observations. This dependency on the model

could account for the loss of power when increasing the proportion of BQL data. In addition to that, the more BQL data are present in the dataset, the less informative data we possess and this affects not only the evaluation step but also the estimation step. Despite this property, the new metrics show better performance than those computed by omitting BQL data in validation data ("omit obs" method) or also in simulations ("omit both" method). Therefore, the new npde is useful in evaluating models with not too high variability parameters and is acceptable to evaluate models with high variability parameters in presence of BQL data.

In the proposed methods, new pd and npde are quantified using a stochastic approach with a single imputation: pd of BQL observation is randomly drawn from a uniform distribution from 0 to the probability for the observation to be under LOQ. Consequently, using the same validation dataset and the same Monte Carlo samples, we could obtain different results (for instance, p-value) when performing the imputation several times. However, in a small simulation study (results not shown), the p-values obtained from different computations were not very different. In fact, we consider the random sampling step to quantify pd as a part of the stochastic properties of simulation-based metrics. For this reason, we performed a unique sampling for each BQL observation and obtained satisfactory type I errors as presented in Table 4. Extension to multiple imputation should be studied.

In this work, we did not include covariates into the model and the simulation study but it would be straightforward to perform another simulation study using this method to evaluate models with one or more covariates. This type of study has been conducted by Brendel *et al* to evaluate models with covariates on datasets containing no or only few censored data [19].

To test the distribution of npde under the null hypothesis, we used the KS test or the global test that is a combination of three sub-tests: the Wilcoxon test, the Fisher test and the Shapiro-Wilk test. The KS test is a general test that can be used to assess any distribution but it is conservative and may have lower power in comparison with other normality tests such as the Anderson - Darling, Cramer - von - Mises or Shapiro-Wilk tests [12]. In this simulation study, comparing results of the KS test and the global test, we also found that the KS test is less powerful than the global test. It is therefore worth performing the normalisation step to obtain normalised metrics in order to use the global test. In the global test, the Wilcoxon test is used to assess whether the mean of the npde is significantly different from 0 because in real life, we could never know if our npde (npd) follow a normal distributionn. However, the Wilcoxon does not test the mean but the ranks of our npde sample. In consequence, sometimes we observed a significant p-value of the Wilcoxon test while only the value of the variance was changed in the simulation (see Table 5, V_{var_1} assumption). In fact, for data received in a population-based clinical trial, the sample size is usually large enough to make the normal approximation and we can thus conduct a t-Student test to compare the mean of npde (npd) to 0. If the t-Student test can be used, it might be more powerful than the Wilcoxon test for detecting model deviation due to fixed effect parameters. A perspective of this work is to look for more appropriate tests to test the null hypothesis.

To evaluate a model, a large number of diagnostic tools have been developed, including qualitative (graphs) and quantitative tools (statistical tests). These tests were developed with a purpose to render interpretation more objective. For npde, statistical test results provide us another way interpreting model misspecifications instead of examining several graphs that may be not very easy to be visually investigated (for example, scatter plots of npde vs time or predictions when we have very large numbers of observations at each time). However, the use of statistical tests in model evaluation is still controversial and an agreed-upon solution may not exist. One should bear in mind that statistical tests can fail to reject a poor model due to lack of power or contrarily, a useful model can be refuted by high-power tests. For this reason, it is worth noting that statistical tests should never be used as a sole criterion for evaluating a model as they can be easily misinterpreted or misleading. It is, therefore, very important to understand the limitations of statistical tests and exercise caution in model evaluation.

Because of the complexity of NLMEM, evaluation should necessarily rely on several criteria and methods. As a consequence, it is important to develop new approaches to correctly handle BQL data for other evaluation methods (VPC, NPC, residuals or prediction errors and also individual - based metrics). In the meanwhile, the VPC plot in the presence of BQL data can be presented in several ways. One of the first approaches is to keep original simulations and to

plot the prediction intervals calculated from original simulation while for the observations, the percentiles lower than LOQ are not plotted [23]. Another approach is to impute BQL values in both observations and simulations at the LOQ in order to compute observed percentiles and the corresponding prediction intervals as proposed in MONOLIX 3.2. The final approach is to remove BQL values in both observations and simulations to compute VPC. These methods are usually used when there are BQL data in the validation dataset but they do not account for BQL data. One possible approach to take into account BQL data in the VPC plot is to use the proposed imputation method which is straightforward to be applied to compute metrics where a decorrelation step is not necessary (VPC). Another approach to impute BQL data in VPC plots was recently proposed by Lavielle and Mesa, which consisted in replacing BQL observations of an individual by values simulated with their conditional distribution (calculated using the model and data of that individual) [30]. VPC or residuals such as npde are then computed using these imputed BQL values. This method needs to be evaluated by simulation study.

In conclusion, the new pd and npde are useful diagnostic tools for evaluating NLMEM in presence of BQL data. Although they are not be able to detect model misspecification in some cases (modifications of variability, large fraction of BQL data), they offer better assessment of model adequacy than other classic metrics omitting all BQL data in the validation dataset or also in simulation data. The methods proposed in this paper to take BQL data into account will be implemented in the next version of the library npde for R.

Acknowledgements We would like to thank Professor Cecile Goujard (Hospital Bicêtre), principal investigator of the COPHAR 3 - ANRS 134 trial, to let us have access to the viral load data. We also want to thank Novartis Pharma for funding the PhD of T.H.T. Nguyen during which a part of this work was completed.

References

1. Sheiner LB, Rosenberg B, Melmon KL (1972) Modelling of individual pharmacokinetics for computer-aided drug dosage. Comput Biomed Res 5:441–459.
2. Sheiner LB, Rosenberg B, Marathe VV (1977) Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. J Pharmacokinet Biopharm 5:445–479.
3. Hughes JP (1999) Mixed effects models with censored data with application to HIV RNA levels. Biometrics 55:625–9.
4. Jacqmin-Gadda H, Thiebaut R, Chene G, Commenges D (2000) Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics 1:355–68.
5. Perelson AS, Ribeiro RM (2008) Estimating drug efficacy and viral dynamic parameters: HIV and HCV. Stat Med 27:4647–57.
6. Samson A, Lavielle M, Mentré F (2006) Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. Comput Stat Data An 51:1562–1574.
7. Ding AA, Wu H (1999) Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics. Math Biosci 160:63–82.
8. Ding AA, Wu H (2001) Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. Biostatistics 2:13–29.
9. Wu H, Wu L (2002) Identification of significant host factors for HIV dynamics modelled by non-linear mixed-effects models. Stat Med 21:753–771.
10. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582–6.
11. Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho DD (1997) Decay characteristics of HIV-1-infected compartments during combination therapy. Nature 387:188–91.
12. Mentré F, Escolano S (2006) Prediction discrepancies for the evaluation of nonlinear mixed-effects models. J Pharmacokinet Pharmacodyn 33:345–67.

13. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis Chapman & Hall/CRC, 2nd edn.
14. Yano Y, Beal SL, Sheiner LB (2001) Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn 28:171–192.
15. Holford NH (2005) The Visual Predictive Check: superiority to standard diagnostic (Rorschach) plots. PAGE 14 Abstr 738 [www.page-meeting.org/?abstract=738].
16. Brendel K, Comets E, Laffont C, Laveille C, Mentré F (2006) Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. Pharm Res 23:2036–49.
17. Comets E, Brendel K, Mentré F (2008) Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R. Comput Methods Programs Biomed 90:154–66.
18. Comets E, Brendel K, Mentré F (2010) Model evaluation in nonlinear mixed effect models with applications to pharmacokinetics. J Soc Fr Stat 151:106–128.
19. Brendel K, Comets E, Laffont C, Mentré F (2010) Evaluation of different tests based on observations for external model evaluation of population analyses. J Pharmacokinet Pharmacodyn 37:49–65 Comparative Study Evaluation Studies Journal Article United States.
20. Duval V, Karlsson MO (2002) Impact of omission or replacement of data below the limit of quantification on parameter estimates in a two-compartment model. Pharm Res 19:1835–1840.
21. Hing JP, Woolfrey SG, Greenslade D, Wright PMC (2001) Analysis of toxicokinetic data using nonmem: Impact of quantification limit and replacement strategies for censored data. J Pharmacokinet Pharmacodyn 28:465–479.
22. Ahn J, Karlsson M, Dunne A, Ludden T (2008) Likelihood based approaches to handling data below the quantification limit using nonmem vi. J Pharmacokinet Pharmacodyn 35:401–421.
23. Bergstrand M, Karlsson MO (2009) Handling data below the limit of quantification in mixed effect models. AAPS J 11:371–80.
24. Beal SL (2001) Ways to fit a PK model with some data below the quantification limit. J Pharmacokinet Pharmacodyn 28:481–504.
25. Yang S, Roger J (2010) Evaluations of bayesian and maximum likelihood methods in pk models with below-quantification-limit data. Pharm Stat 9:313–330.
26. Wu L (2004) Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. J Amer Statistical Assoc 99:700–709.
27. Goujard C, Barrail-Tran A, Duval X, Nembot G, Panhard X, Savic R, Descamps D, Vrijens B, Taburet A, Mentré F (2010) Virological response to atazanavir, ritonavir and tenofovir/emtricitabine: relation to individual pharmacokinetic parameters and adherence measured by medication events monitoring system (MEMS) in naïve HIV-infected patients (ANRS134 trial). International AIDS Society 2010 Abstr WEPE0094 [http://www.iasociety.org/Default.aspx?pageId=11&abstractId=200738161].
28. Laffont C, Concordet D (2011) A new exact test for the evaluation of population pharmacokinetic and/or pharmacodynamic models using random projections. Pharm Res 28:1948–62.
29. Bergstrand M, Hooker A, Wallin J, Karlsson M (2011) Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. AAPS J 13:143–51.
30. Lavielle M, Mesa H (2011) Improved diagnostic plots require improved statistical tools. Implementation in MONOLIX 4.0. PAGE 20 Abstr 2180 [www.page-meeting.org/?abstract=2180].

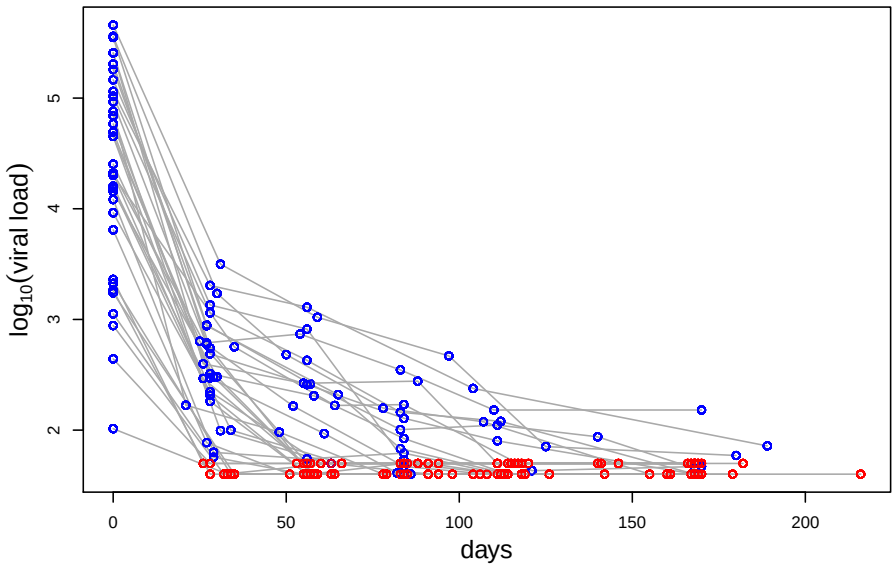


Fig. 1 Spaghetti plot of viral load in logarithmic scale versus time from COPHAR 3 - ANRS 134 trial. Data above LOQ are presented as blue circles, data below LOQ are imputed at LOQ in this graph and presented as red circles

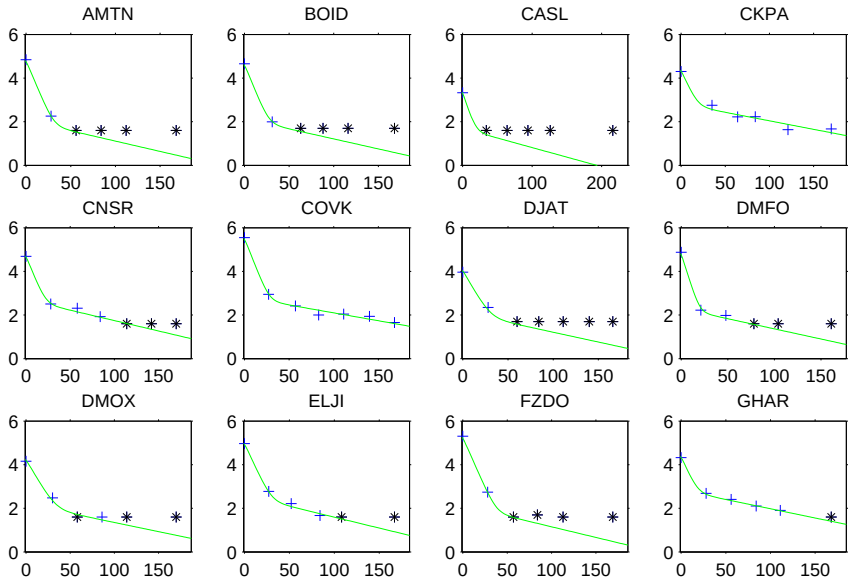


Fig. 2 Some examples of individual fits for the final model. Data above LOQ are presented as blue cross. In this plot, BQL data are imputed at LOQ and are presented as black star symbol

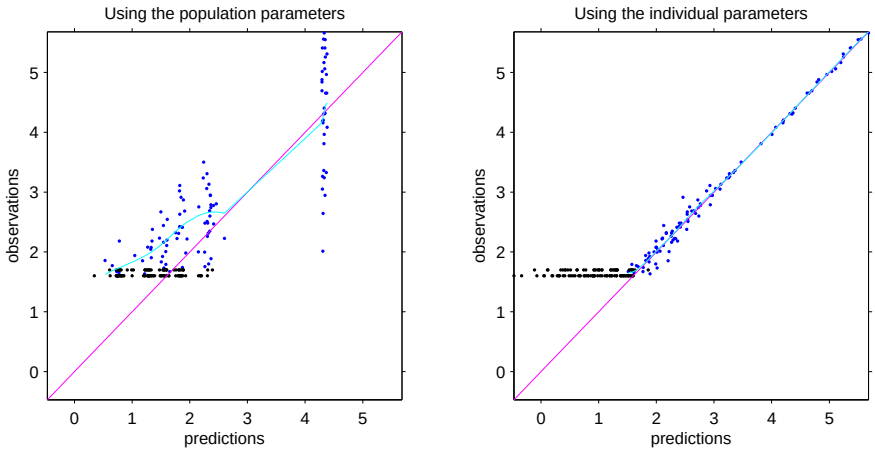


Fig. 3 Observations versus population predicted values and individual predicted values. Observations are plotted as blue closed circles. In these plots, BQL observations are imputed at LOQ values and are presented as black closed circles

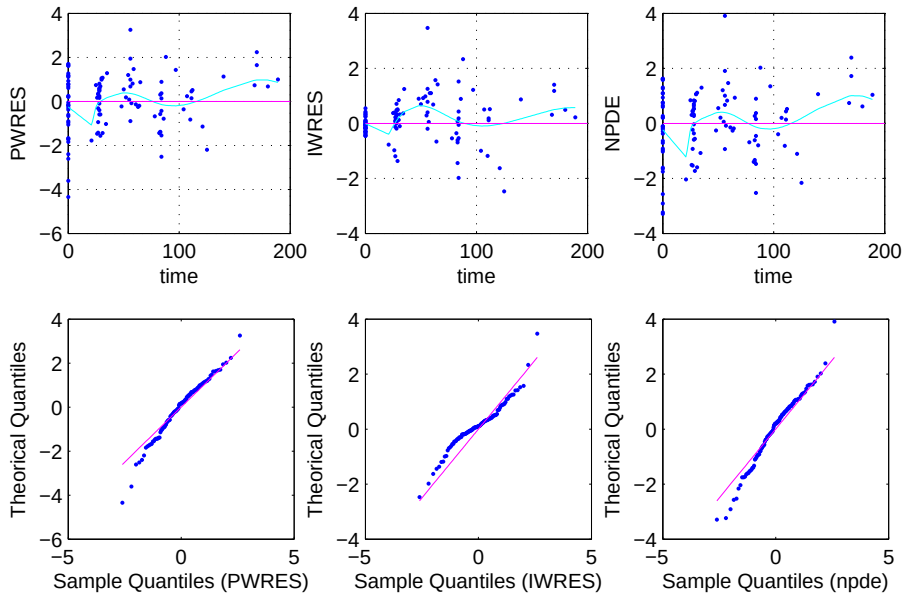
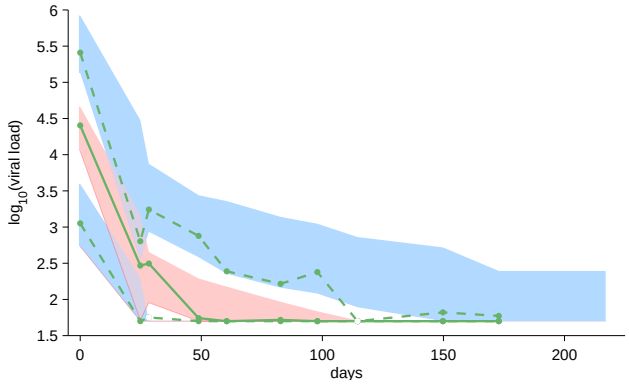
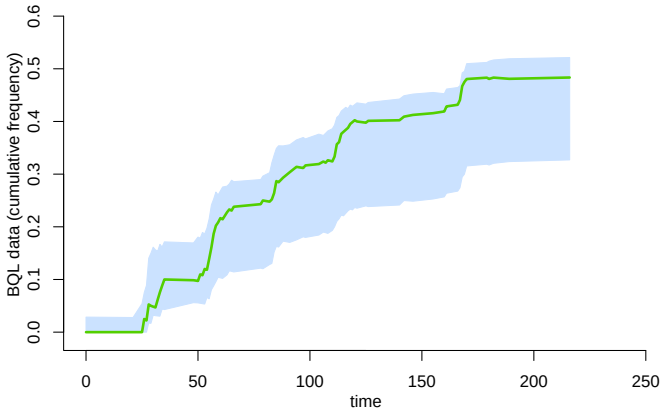


Fig. 4 Residuals (provided by MONOLIX) versus time plots and q-qplots versus the standard normal distribution $\mathcal{N}(0,1)$ of different types of residuals for the final model. In these plots, BQL data are omitted

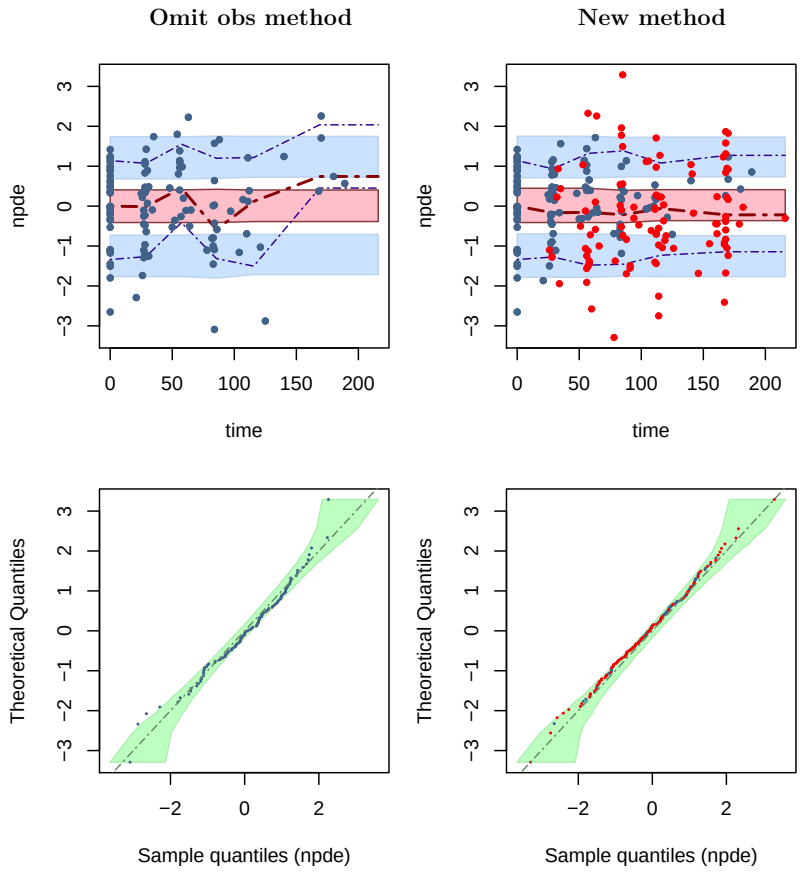


(a) VPC for observed data



(b) BQL data fraction over time

Fig. 5 Visual Predictive Check for final model. Figure 5(a) is the classical VPC graph for data above LOQ. BQL data are imputed at LOQ in the graph. The green (dashed and solid) lines represent the 10^{th}, 50^{th} and 90^{th} percentiles for observed data. The shaded blue and pink areas represent 90% prediction intervals for the corresponding percentiles calculated from simulated data. Figure 5(b) represents the observed cummulative fraction of BQL data versus time (the fraction of BQL data at a time point is calculated by the ratio between the number of BQL data and the total number of measurements obtained from the beginning to the study to this time) (green solid line) and its 90% prediction interval calculated under the final model (blue shaded area)



Wilcoxon test:	0.482	Wilcoxon test:	0.204
Fisher test:	0.344	Fisher test:	0.366
Shapiro-Wilk test:	0.131	Shapiro-Wilk test:	0.937
Global test:	0.393	Global test:	0.611
KS test:	0.461	KS test:	0.551

Fig. 6 Scatterplot of npde versus time and the q-qplot of npde for the final model. npde are calculated using the "omit obs" method (left) or the new method (right). npde for data above LOQ are presented by blue closed circles, npde for BQL data are presented by red closed circles. Dashed (blue and dark red) lines in the scatterplot represent 10^{th}, 50^{th} and 90^{th} percentiles of the npde corresponding to observed data. Light blue and pink shaded areas in the scatterplot are 95% prediction intervals of the selected percentiles calculated from K MC samples. Green shaded area in the q-qplot represents the 95% prediction intervals for the npde sample.

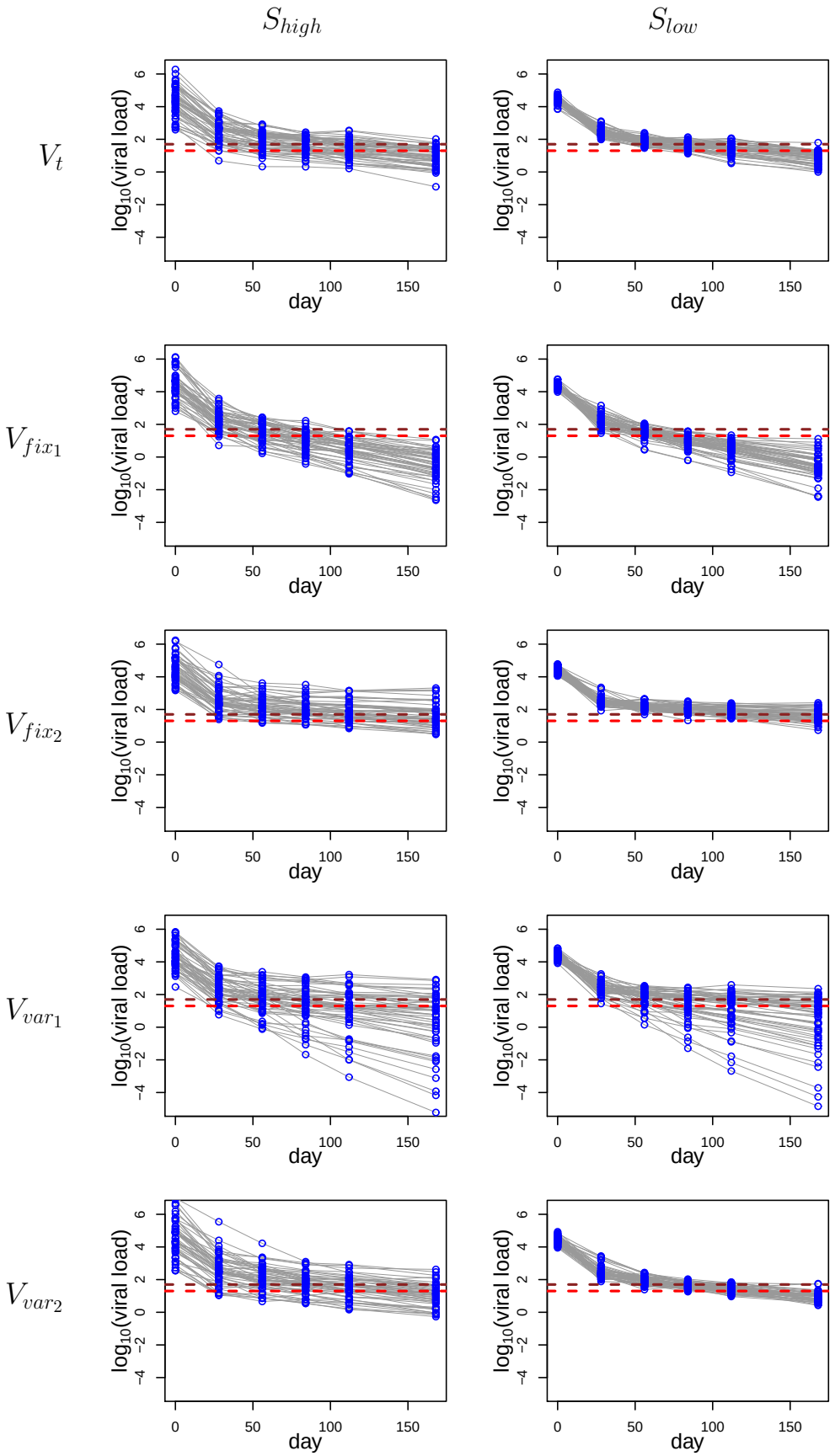


Fig. 7 Spaghetti plots of different validation datasets simulated using the rich design, under two variability settings S_{high}, S_{low}. Data are not censored and two LOQ levels (20 and 50 copies/mL) are co-plotted as brown and red dashed lines in the graphs

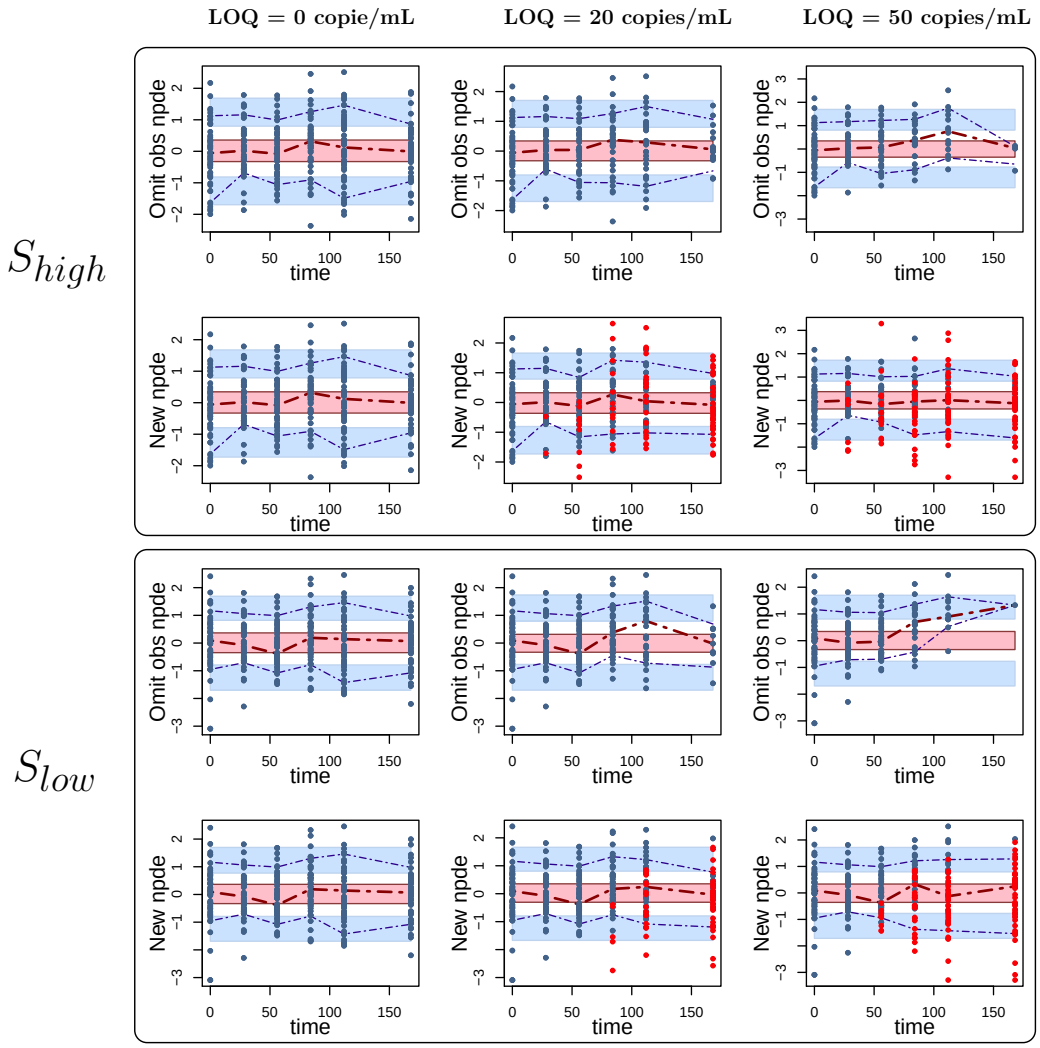


Fig. 8 Scatterplots of npde (calculated using "omit obs" and new methods) versus time for datasets V_t simulated under the true model. Three levels of censoring were applied on each datasets : 0 copie/mL (first column), 20 copies/mL (second column) and 50 copies/mL (third column).

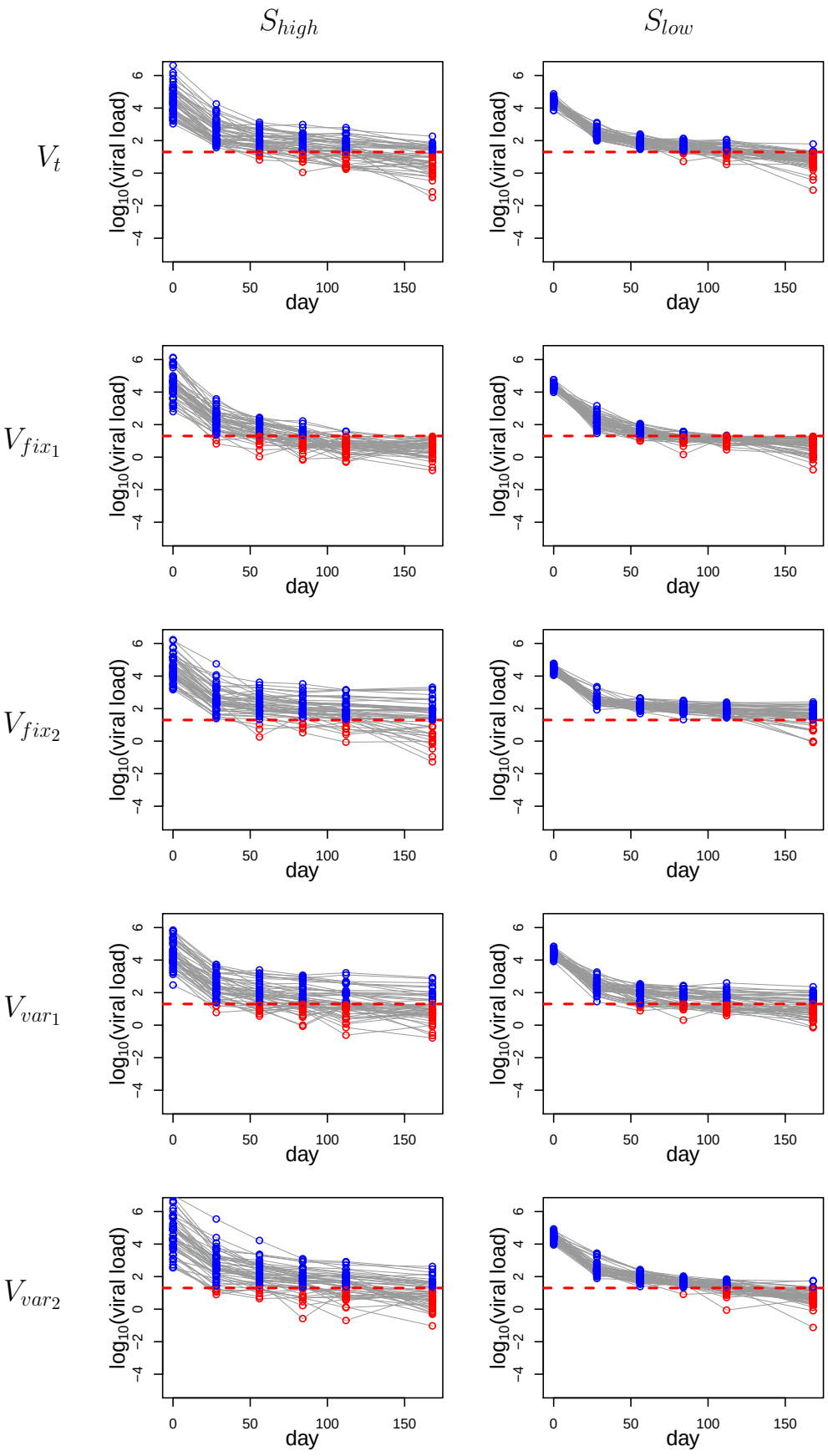


Fig. 9 Samples of spaghetti plots of validation datasets presented in the Fig. 7, at a censoring level of 20 copies/mL after the imputation step (see methods).

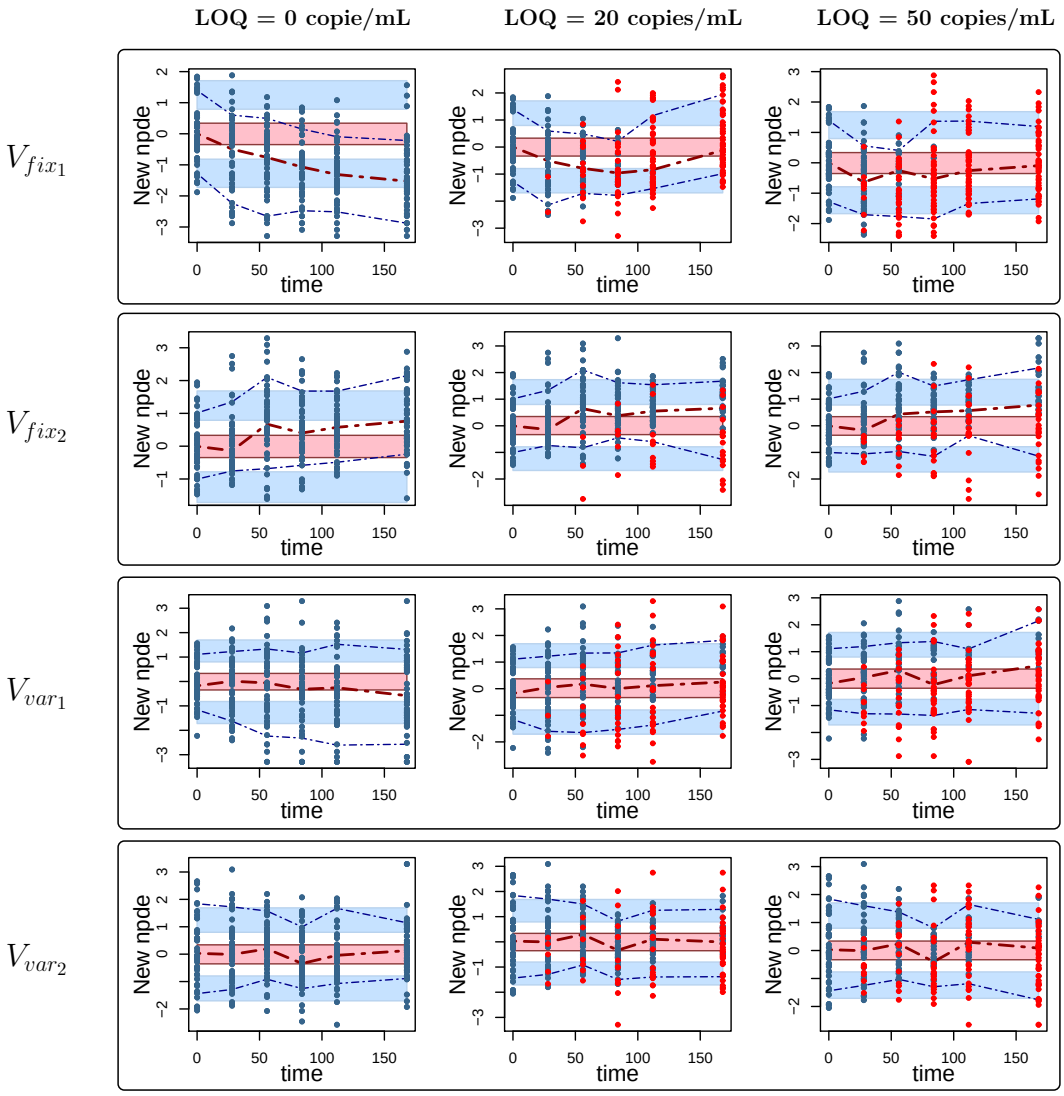


Fig. 10 Scatterplots of npde versus time for different validation datasets (V_{fix_1}, V_{fix_2}, V_{var_1}, V_{var_2}) simulated using the rich design and the high variability setting. npde are calculated by new method that accounts for BQL data. Data above LOQ are presented by blue closed circle, BQL data are presented by red closed circle. Dashed (blue and dark red) lines represent 10^{th}, 50^{th} and 90^{th} percentiles of the npde corresponding to observed data. Light blue and pink shaded areas are 95% prediction intervals of the selected percentiles corresponding to simulated data

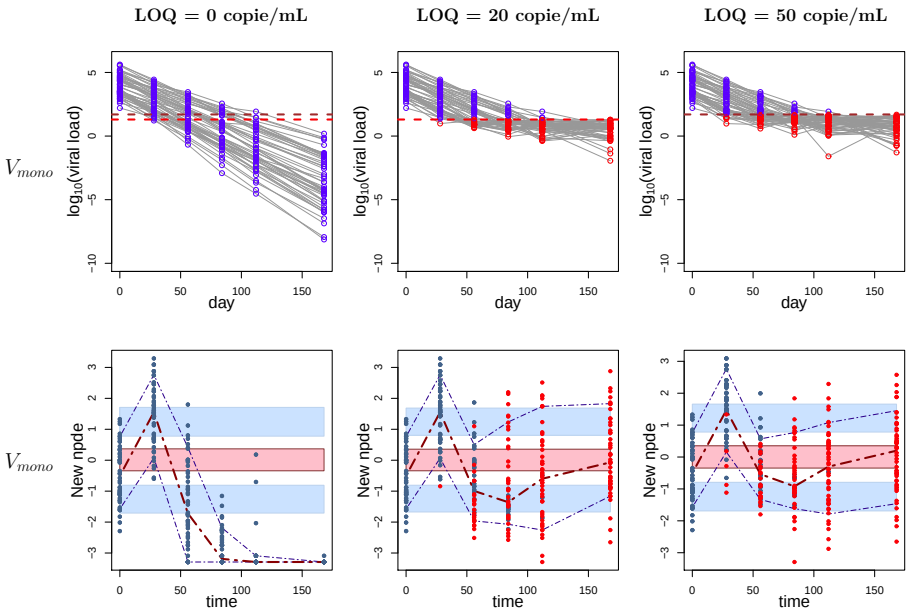


Fig. 11 Spaghetti plots (upper pattern) before (1^{st} column) and after imputation (last two columns) and scatterplots of new npde versus time for the dataset simulated under the false structure model (V_{mono}). BQL data and npde are presented by red circle, observed data and npde related to observed data are presented by blue circle. Dashed lines in spaghetti plots represent LOQ levels. Dashed (blue and dark red) lines in scatterplots of npde represent 10^{th}, 50^{th} and 90^{th} percentiles of the npde corresponding to observations. Light blue and pink shaded areas are 95% prediction intervals of the selected percentiles corresponding to simulated data

Table 1 Parameter estimates for the final model in the 35 patients of the COPHAR 3 - ANRS 134 trial

Parameters	Estimates	RSE (%)
P_1 (copie/mL)	22000	36
P_2 (copie/mL)	222	27
λ_1 (day^{-1})	0.222	11
λ_2 (day^{-1})	0.0198	8
ω_{P_1}	2.1	13
ω_{P_2}	1.31	14
ω_{λ_1}	0.232	48
ω_{λ_2}	0.218	37
$\rho_{(\eta_{P_1},\eta_{P_2})}$	0.758	13
σ_{inter}	0.15	5

Table 2 Parameters used for simulating V_{true}, V_{fix_1}, V_{fix_2}, V_{var_1}, V_{var_2}

	Parameters	True model	False models			
		V_t	V_{fix_1}	V_{fix_2}	V_{var_1}	V_{var_2}
	P_1 (copie/mL)	25000	25000	25000	25000	25000
	P_2 (copie/mL)	250	250	250	250	250
	λ_1 (day^{-1})	0.2	0.2	0.2	0.2	0.2
	λ_2 (day^{-1})	0.02	**0.04**	**0.01**	0.02	0.01
S_{high}	ω_{P_1}	2.1	2.1	2.1	2.1	2.1
	ω_{P_2}	1.4	1.4	1.4	1.4	1.4
S_{low}	ω_{P_1}	0.3	0.3	0.3	0.3	0.3
	ω_{P_2}	0.3	0.3	0.3	0.3	0.3
	ω_{λ_1}	0.3	0.3	0.3	0.3	0.3
	ω_{λ_2}	0.3	0.3	0.3	**0.9**	**0.1**
	$\rho_{(\eta_{P_1},\eta_{P_2})}$	0.8	0.8	0.8	0.8	0.8
	σ_{inter}	0.14	0.14	0.14	0.14	0.14

Table 3 Proportions (in %) of BQL data in different types of dataset (evaluated on 1000 validation datasets for each scenario)

	D_{sparse}, S_{high}		D_{rich}, S_{high}		D_{rich}, S_{low}	
	LOQ = 20	LOQ = 50	LOQ = 20	LOQ = 50	LOQ = 20	LOQ = 50
V_t	27.9	43.1	27.9	42.9	22.2	42.4
V_{fix_1}	50.5	60.5	50.6	60.5	51.3	62.8
V_{fix_2}	12.4	26.3	12.3	26.2	2.5	14.6
V_{var_1}	31.2	43.2	31.3	43.3	26.5	39.2
V_{var_2}	27.1	42.7	27.0	42.7	21.2	43.2
V_{mono}	-	-	53.5	59.0	-	-

Table 4 Type I errors under the null hypothesis (in %) of different statistical tests for npd (datasets simulated from D_{sparse}, S_{high}) and for npde (datasets obtained from D_{rich}, S_{high} and D_{rich}, S_{low}). Type I errors are evaluated on 1000 simulated datasets and at three LOQ levels. npde (npd) are calculated using different methods. Values significantly different from 5% are in bold

Methods	Tests	D_{sparse}, S_{high}			D_{rich}, S_{high}			D_{rich}, S_{low}		
		npd			npde			npde		
		LOQ (cp/mL)			LOQ (cp/mL)			LOQ (cp/mL)		
		0	20	50	0	20	50	0	20	50
"omit obs" method	Global test	5.5	**99.8**	**100**	5.4	**25.8**	**46.9**	5.6	**23.9**	**64.3**
"omit both" method	Global test	5.5	4.7	5.0	–	–	–	–	–	–
New method	Wilcoxon	5.5	4.3	5.5	5.1	6.2	5.6	5.3	4.8	6.2
	Fisher	4.9	5.1	4.8	5.7	**9.1**	**8.6**	5.8	6.3	6.5
	Shapiro-Wilk	5.3	5.4	6.2	4.3	4.0	3.9	6.1	5.0	4.9
	Global test	5.9	5.6	5.2	5.4	**6.7**	6.2	5.6	6.2	6.1
	KS	5.8	4.8	4.7	5.4	**6.5**	**7.0**	5.1	4.8	5.4

Table 5 Power under several alternative assumptions (in %) of different statistical tests for npd (D_{sparse}, S_{high}) and for npde D_{rich}, S_{high} and D_{rich}, S_{low}) evaluated on 1000 simulated datasets and at 3 LOQ levels. Values greater than 90% are in bold. npd and npde are calculated using the"omit both" or new method

Methods		Tests	D_{sparse}, S_{high}			D_{rich}, S_{high}			D_{rich}, S_{low}		
			npd			npde			npde		
			LOQ (cp/mL)			LOQ (cp/mL)			LOQ (cp/mL)		
			0	20	50	0	20	50	0	20	50
"omit both" method	V_{fix_1}	Wilcoxon	**100**	62.5	21.9	-	-	-	-	-	-
		Fisher	**99.9**	6.5	7.2	-	-	-	-	-	-
		Shapiro-Wilk	82.7	6.6	5.4	-	-	-	-	-	-
		Global test	**100**	47.1	14.0	-	-	-	-	-	-
		KS	**100**	55.6	17.7	-	-	-	-	-	-
New method	V_{fix_1}	Wilcoxon	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99.9**
		Fisher	**99.9**	34.2	57.4	**99.4**	**91.1**	35.8	**100**	**99.9**	28.2
		Shapiro-Wilk	82.7	46.4	21.1	70.5	47.4	23.0	**96.9**	18.7	5.5
		Global test	**100**	**100**	**100**	**100**	**100**	**98.8**	**100**	**100**	**99.9**
		KS	**100**	**100**	**100**	**100**	**100**	**99.7**	**100**	**100**	**99.6**
	V_{fix_2}	Wilcoxon	**100**	**100**	**100**	**100**	**100**	**99.8**	**100**	**100**	**100**
		Fisher	9.2	32.9	66.3	7.2	9.8	15.1	4.9	16.5	**98.3**
		Shapiro-Wilk	7.0	10.9	8.9	5.4	48.8	28.9	5.0	6.9	17.5
		Global test	**100**	**100**	**100**	**100**	**100**	**99.7**	**100**	**100**	**100**
		KS	**100**	**100**	**99.9**	**100**	**100**	**99.8**	**100**	**100**	**100**
	V_{var_1}	Wilcoxon	43.4	5.1	6.1	51.0	28.7	25.1	50.6	63.4	67.3
		Fisher	**100**	63.8	40.1	**100**	81.9	59.9	**100**	**100**	**99.9**
		Shapiro-Wilk	**99.5**	17.2	15.7	**95.2**	14.7	5.6	**98.5**	21.6	29.4
		Global test	**100**	53.6	31.6	**100**	78.7	53.6	**100**	**100**	**99.4**
		KS	**96.5**	19.1	9.5	**93.0**	56.6	39.6	**95.5**	**92.3**	82.0
	V_{var_2}	Wilcoxon	8.8	4.6	4.2	4.7	5.1	5.5	3.6	11.0	10.5
		Fisher	23.1	7.4	6.7	24.8	14.7	9.3	42.4	41.6	21.3
		Shapiro-Wilk	5.7	6.3	5.5	5.0	7.0	5.0	4.9	6.9	7.7
		Global test	17.1	6.1	5.9	14.2	11.2	8.2	30.5	30.7	17.8
		KS	8.5	4.5	4.2	5.2	6.8	6.7	6.0	14.7	11.1
	V_{mono}	Wilcoxon	-	-	-	**100**	67.5	22.9	-	-	-
		Fisher	-	-	-	**100**	**100**	**100**	-	-	-
		Shapiro-Wilk	-	-	-	**100**	**95.0**	78.4	-	-	-
		Global test	-	-	-	**100**	**100**	**100**	-	-	-
		KS	-	-	-	**100**	**100**	**99.0**	-	-	-