Pharmaceutical Research, Vol. 23, No. 9, September 2006 (© 2006)
DOI: 10.1007/s11095-006-9067-5

Research Paper

Metrics for External Model Evaluation with an Application to the Population Pharmacokinetics of Gliclazide

Karl Brendel,[1,2,3,6] Emmanuelle Comets,[1] Céline Laffont,[4] Christian Laveille,[4,5] and France Mentré[1,2,3]

Received January 9, 2006; accepted May 18, 2006; published online August 12, 2006

Purpose. The aim of this study is to define and illustrate metrics for the external evaluation of a population model.

Materials and Methods. In this paper, several types of metrics are defined: based on observations (standardized prediction error with or without simulation and normalized prediction distribution error); based on hyperparameters (with or without simulation); based on the likelihood of the model. All the metrics described above are applied to evaluate a model built from two phase II studies of gliclazide. A real phase I dataset and two datasets simulated with the real dataset design are used as external validation datasets to show and compare how metrics are able to detect and explain potential adequacies or inadequacies of the model.

Results. Normalized prediction errors calculated without any approximation, and metrics based on hyperparameters or on objective function have good theoretical properties to be used for external model evaluation and showed satisfactory behaviour in the simulation study.

Conclusions. For external model evaluation, prediction distribution errors are recommended when the aim is to use the model to simulate data. Metrics through hyperparameters should be preferred when the aim is to compare two populations and metrics based on the objective function are useful during the model building process.

KEY WORDS: external validation; metrics; model evaluation; population pharmacokinetics; posterior predictive check.

INTRODUCTION

Population pharmacokinetic (PK) and/or pharmacodynamic (PD) analyses using nonlinear mixed-effects models are increasingly used during drug development (1–3) and for simulation of clinical trials (4 – 6). The use of population pharmacokinetic modelling in the drug development process is recommended in the FDA's guidance for industry to help identify differences in drug safety and efficacy among population subgroups (7). Model evaluation is also recommended in this guidance however there is no consensus today on an appropriate approach to assess a population model.

Two types of model evaluation can be performed. The first is internal evaluation and refers to the use of data splitting and resampling techniques (8); in the following, we only consider the second one, external evaluation, which refers to a comparison between a validation dataset and the predictions from the model built from the learning dataset using dosage regimen information and possibly covariates from the validation dataset. The validation dataset is not used for model building or for parameter estimation.

In this paper, we describe criterion which are often used for model evaluation as well as some new metrics or new approaches that can be used for external model evaluation in population PK or PD analyses. We then propose to compare the metrics for the evaluation of a population PK model with different tests and graphs.

Different approaches to model evaluation have been proposed (9), although none has yet proved universally preferable. We consider here metrics with or without Monte Carlo simulation. Metrics with Monte Carlo simulation are called posterior predictive check (PPC), and they evaluate the adequacy between the data and the model by comparing a given statistic, computed with the observed data, to its posterior predictive distribution according to the model. PPC was defined by Yano *et al.* (10). Several papers have been published that apply PPC to pharmacokinetic–pharmacodynamic models (11,12).

Evaluation through prediction errors on observations calculated by linearisation of the model are the most used model evaluation tool (13 – 17). The mean square prediction error (precision) and the mean prediction error (bias) are easily computed and used to compare predictions to observations. However prediction errors on observations are not independent within an individual (18). The standardized

[1] INSERM U738, Paris, France.
[2] University Paris 7, Paris, France.
[3] AP-HP, Bichat Hospital, Paris, France.
[4] Institut de Recherches Internationales Servier, Courbevoie, France.
[5] Exprimo NV, Lummen, Belgium.
[6] To whom correspondence should be addressed. (email: karl.brendel@bch.aphp.fr)

prediction errors, obtained using a first-order approximation, are also often used. In NONMEM, their computation takes into account correlation of the observations within an individual so that standardized prediction errors are decorrelated (9). A recent method consists of using PPC on observations, by computing for each observation prediction discrepancies as the percentile of the observation in the whole distribution of predictions (19). Computation of prediction discrepancies using Monte Carlo integration does not require model approximation but this metric is correlated within an individual. This method has been applied to detect the differences in the pharmacokinetics of S1, an oral anticancer agent, in Western and Japanese patients (20). As predictions discrepancies did not take into account correlation of the observations within an individual (19,20), we have decorrelated this metric in this paper.

A second approach to model evaluation is through the examination of population parameters or hyperparameters, comparing population estimates of the parameters between the learning and the validation datasets with appropriate tests based on the estimated standard error. Another method for hyperparameter comparison is to use PPC.

A third approach is to use the objective function for model evaluation. We describe in this paper two tests based on this metric.

These metrics were applied to the evaluation of a population pharmacokinetic model of gliclazide (an antidiabetic drug) which is a one compartment model with zero order absorption, built from two phase II studies. We show the results of the evaluation for three validation datasets: a real phase I dataset (V_{real}) and two datasets simulated with the design of V_{real}. The first (V_{true}) is simulated using the parameters of the model; the second (V_{false}) is simulated using the same model, but with a bioavailability multiplied by 2. All these metrics were applied as an illustration to these two simulated datasets to show how they are able to detect and explain potential adequacies and inadequacies of the model and to compare theoretical statistical properties of the metrics.

MATERIALS AND METHODS

Notations

Let B denote a learning dataset and V a separate external validation dataset. B is used to build a population pharmacokinetic model called M^{B}. External evaluation methods compare the predictions obtained by M^{B}, using the design of V, to the observations in V.

Let i denote the ith individual (i = 1,...,N) and j the jth measurement in an individual (j = 1,...,n_i, where n_i is the number of observations for subject i). Let Y_i be the $\boldsymbol{n}_i$-vector of observations observed in individual i. The function f is a nonlinear structural model, i.e., the pharmacokinetic model. The statistical model for the observation Y_{ij} in patient i at time t_{ij}, is given by:

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

where θ_i is the $\boldsymbol{p}$-vector of the pharmacokinetic 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:

$$Var\left(\varepsilon_{ij}\right) = \sigma^2_{\mathrm{inter}} + \sigma^2_{\mathrm{slope}} \times f\left(t_{ij,}\theta_i\right)^2 \tag{2}$$

where $\sigma^2_{\mathrm{inter}}$ and $\sigma^2_{\mathrm{slope}}$ are two parameters characterizing the variance. This combined variance model covers the case of an homoscedastic variance error model when $\sigma^2_{\mathrm{slope}} = 0$ and the case of a constant coefficient of variation error model when $\sigma^2_{\mathrm{inter}} = 0$. Let Σ be the parameters of the measurement error, $= \left(\sigma^2_{\mathrm{inter}}, \sigma^2_{\mathrm{slope}}\right)$.

We assume an exponential model for interindividual variability, so that:

$$\theta_i = \theta \times \exp\left(\eta_i\right) \tag{3}$$

where $\boldsymbol{\theta}$ is the population vector of the pharmacokinetic parameters, and $\boldsymbol{\eta}_i$ represents the vector of random effects of individual i. It is assumed that $\boldsymbol{\eta}_i \sim N(0, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega}$ defined as variance–covariance matrix so that each diagonal element ω^2_k represents the variance of the kth component of the random effects vector.

The vector of population parameters called hyperparameters, is denoted $\boldsymbol{\Psi}$ and has dimension Q. $\boldsymbol{\Psi}$ includes the vector of population means $\boldsymbol{\theta}$, the unknown elements in the variance–covariance matrix of the random effects $\boldsymbol{\Omega}$. Estimation of the hyperparameters is based on maximum likelihood (ML). $\boldsymbol{\Gamma}$ is the asymptotic variance–covariance matrix of estimation, i.e., the Fisher information matrix calculated in NONMEM using the inverse Hessian matrix for the hyperparameters $\boldsymbol{\Psi}$. SE$_q$, the standard errors of estimation for the qth hyperparameter $\boldsymbol{\Psi}_q$, is the square root of the qth diagonal element of $\boldsymbol{\Gamma}$.

Model M^{B} is defined by its structure and by the hyperparameters $\boldsymbol{\Psi}^{\mathrm{B}}$ estimated from the learning dataset B.

Illustrative Example

Phase II Studies (Learning Dataset)

Two phase II studies in a total of 209 Type II diabetic patients were pooled to create the dataset B, which was used to build M^{B}. These studies were performed during the clinical development of a modified release formulation of gliclazide (gliclazide MR) and were part of a larger dataset analyzed by Frey *et al*., who studied the relationship between the pharmacokinetics of gliclazide and its long-term pharmacodynamic effect in a large population of Type II diabetic patients (21).

The first study (N = 50 patients) was a phase II, dose-increase, monocentric, randomised, parallel double-blind placebo-controlled study. Patients were first treated with placebo for a 2-week wash-out period. At the end of this period, 50 patients were randomised to receive placebo (10 patients), 15 mg of gliclazide (20 patients) or 30 mg of gliclazide (20 patients) during 4 weeks (period 1). During the next 4 weeks (period 2), the ten patients treated by placebo continued to receive placebo. For the other 40 patients, two fasting plasma glucose (FPG) measurements were performed at the end of the first period. If the mean of the two FPG measurements was lower than 7.8 mmol/l, the patients received the same dose of gliclazide (15 or 30 mg) as in the first period. If the mean of the two FPG measurements was

7.8 mmol/l or more, the patients received a dose of gliclazide two times higher (30 or 60 mg) than in the first period.

To obtain a better evaluation of patient compliance, MEMS (Medication Event Monitoring System), medication bottles in which the cap contains microelectronic components recording the dates and times of dose removals from the bottle, were used. Blood samples were drawn on the first day, at the end of the first 4 weeks and at the end of the study (8 weeks), according to one of two sampling schedules (S_1 or S_2) which were randomly assigned for a period to the 50 patients. For S_1, the sampling times were pre-dose, 2, 4, 6, 8 h after administration and before leaving the clinical research unit (13 h after administration). For S_2 the sampling times were pre-dose, 3, 5, 7, 9 h after administration and before leaving the clinical research unit (13 h after administration).

The second study ($N = 169$ patients) was a phase II, dose ranging, monocentric, randomised, parallel double-blind placebo-controlled study. After a 2-week wash-out period, patients were randomly divided into six parallel groups and given either placebo or one of the following doses of gliclazide: 15, 30, 60, 90 and 135 mg. Gliclazide was administered once a day during 8 weeks. Three blood samples were taken on the last day of treatment: just before the last administration, 2 h after dosing and between 5 and 6.5 h (half of the patients) or between 7.5 or 9 h after dosing (other half of the patients). The times of drug intake on the day of the visit and on the day before the visit were recorded.

Gliclazide plasma concentrations were measured using high-performance liquid chromatography with ultaviolet detection. The lowest concentration giving accuracy and precision within a limit of 20% was 50 ng ml^{-1}. This value was taken as the limit of quantification.

Population Pharmacokinetic Model from the Above Phase II Studies

Plasma concentration-time data were obtained from the 209 patients who received gliclazide. A one compartment model with zero-order absorption and first-order elimination was used to fit the concentration-time data of gliclazide. This model was parameterized with the apparent volume of distribution (V/F), the apparent clearance (CL/F) and the duration of absorption for the zero order absorption model (T_{abs}). An exponential random-effect model was chosen to describe inter-individual variability.

During model building, values below the quantification limit (BQL), with a quantification limit (QL) which was equal to 0.05 mg/l, were treated in one of the standard ways by imputing the first BQL measurement to QL/2 and omitting subsequent BQL measurements during the terminal phase (22). The symmetrical reverse procedure was applied to BQL measurements during the absorption phase. Only five samples were below the quantification limit.

The population analysis of the two phase II studies was performed using NONMEM software version V (University of San Francisco) with the FOCE method with interaction. SAS version 8.2 software was used to perform statistical analyses and to plot graphs (SAS Institute INC., 1990).

Model selection was based on comparison of the objective function given by NONMEM. For nested models, a likelihood ratio test (LRT) was performed with a p value of 0.05; i.e., the difference on objective function was compared to the limit of a chi-square distribution, with a number of degrees of freedom equal to the number of additional parameters in the full model. For non-nested models, models were compared using the Akaike criterion (AIC). Goodness-of-fit plots were performed during the building process (population or individual predictions *versus* observations, WRES *versus* population predictions (or time), individual WRES *versus* individual predictions (or time)). Decrease of the inter-individual variability of parameters estimation and decrease of the standard error of the fixed effects were also taken into account.

A proportional error model was found to best describe the residual error model. As study 1 used MEMS, the records of dates and times of drug administrations were more accurate than for study 2. Two different variances for the error, one for study 1 and another for study 2, were included, and this provided a significant improvement in the likelihood ($p < 0.0001$). Several models for the random effects were tested to determine the basic model that best fitted the data, and only random effects on CL/F and V/F were kept in the model. No inter-occasion variability was found in the model. As there were five different doses of gliclazide (15, 30, 60, 90 and 135 mg), the effect of dose was tested on the two pharmacokinetic parameters with inter-individual variability (CL/F and V/F). We therefore defined a categorical covariate, DF, which equals 1 for dose >60 mg and 0 for dose ≤60 mg and θ_D the fixed effect for the dose effect when dose >60 mg. Inclusion of this covariate in the population analysis on V/F provided a significant improvement ($p < 0.001$). The fixed effect equation for the volume was: $V\big/F \times \theta_D^{DF}$.

The estimates of the population parameters for this model are given in Table I. The CV of the error model were estimated to 0.06% (σ_1) for the study using MEMS and 0.11% (σ_2) for the study without MEMS.

Phase I Study (Validation Dataset)

The validation dataset V_{real} was obtained in a phase I open single dose study with a two-way-cross-over randomised and balanced design which aim was to evaluate the absolute bioavailability of gliclazide. Twelve healthy volunteers received gliclazide as a tablet of 30 mg and as a solution given in 1 h infusion, with a 7-day-wash-out between the two administrations (23). These volunteers were Caucasian males and were between 18 and 35 years old. We only considered here data obtained after oral administration of gliclazide.

Table I. Estimated Population Pharmacokinetic Parameters of Gliclazide (Estimate and Relative Standard Error of Estimation, RSE), Pooling the Data of Two Phase II Studies

Population parameters	Estimate	RSE (%)
CL/F (l/h)	1.0	4.2
θ_V (L)	21.6	5.8
T_{abs} (h)	6.6	3.3
θ_D	0.46	12.4
$\omega^2_{CL/F}$	0.35	16.2
$\omega^2_{V/F}$	0.17	24.9
σ_1^2	0.06	10.5
σ_2^2	0.11	12.7

For each volunteer, 16 blood samples were taken at 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 10, 16, 24, 36, 48, 60, 72 h after oral dosing.

Simulation

To illustrate and compare the methods presented in the following section, we simulated with NONMEM two validation datasets according to the design of the real phase I study (23). The first dataset (V_{true}) was simulated according to the model and to the hyperparameters values estimated from the learning dataset; the second (V_{false}) was simulated using the same model, but with the fixed effects for V/F and CL/F divided by two, corresponding to a bioavailability multiplied by two.

Patients were given the dose of gliclazide under controlled conditions, therefore we assumed a CV of 25% for the residual error, corresponding to the estimate obtained in the population where MEMS was used. Values below the quantification limit for V_{true}, V_{false} and V_{real} were treated in the same way as in the learning dataset. As all subjects received 30 mg of gliclazide the dose effect was not taken into account.

The various metrics proposed in the following were first applied to the two simulated datasets to illustrate the ability of each metric not to reject the "correct" dataset (V_{true}) and to reject the "false" dataset (V_{false}). We then applied these metrics to the real dataset (V_{real}).

Metrics for External Evaluation

The null hypothesis (H_0) is that data in the validation dataset V can be described by model M^{B}. In this section, we describe the metrics which can be proposed as tools for model evaluation, and we test their distribution under H_0. We consider metrics with or without Monte Carlo simulation. Metrics with simulation are called posterior predictive check (PPC), and evaluate the adequacy between the data and the model by comparing a given statistic, computed with the observed data, to its posterior predictive distribution according to the model (10). When performing Monte Carlo simulations with the model M^{B} applied to the design of V, we used the estimates of the parameters without taking into account the standard errors of estimation. This is reasonable for large enough datasets.

Since for some of the metrics multiple tests are involved, we used the Simes procedure, a modification of the Bonferroni procedure, to adjust for the increase of the type I error (24). This procedure, while preserving the family-wise error of the test, is less conservative than the Bonferroni correction but is still simple to apply. This method allows to test Q simultaneous assumptions (H_1,...,H_{Q}), and uses the following procedure.

The Q p values of each of the Q tests are sorted in ascending order, so that $p_1<p_2<...<p_{\text{Q}}$. We chose a family-wise error rate of 0.05. Starting with p_1, the smallest p value, each successive p_q for q = 2,...Q, is compared with the value $c_q = q \times 0.05/Q$. If $p_1 > c_1$, H_1,..., H_Q are not rejected, if not, H_1 is rejected and p_2 and c_2 are compared. The procedure is then iterated until p_q is found such that $p_q > c_q$. Then the hypotheses H_1... H_{q-1} are rejected while the remaining hypotheses H_q... H_Q are not rejected.

Metrics Based on Observations

Metrics based on observations are the most frequently used statistics to validate population models. Predictions are obtained using M^{B} and the design of V and compared to the observed values in V. Three metrics based on observations are tested.

*Standardized Prediction Error on Observations (**SPEY**).* Prediction errors are defined as the difference between the observations and the predictions obtained using M^{B}. The vector $\boldsymbol{PEY}_i$, of the ith subject is then:

$$PEY_i = Y_i - PRED_i \tag{4}$$

where $\boldsymbol{Y}_i$ is the vector of observations of the ith subject, and $\boldsymbol{PRED}_i$ the vector of population prediction (assuming $\eta = 0$) using M^B. It should be noted that, for nonlinear models, the prediction at $\eta = 0$ is not the mean of the predictive distribution. $\boldsymbol{PEY}_i$ is obtained with NONMEM under the name of RES$_i$.

Calculation of standardized prediction errors takes into account the variability. $\boldsymbol{SPEY}_i$, the vector of standardized prediction error for the ith subject is defined as (25):

$$SPEY_i = C_i^{-1/2} \times PEY_i \tag{5}$$

where the matrix $\mathbf{C}_i$ is the variance–covariance matrix of $\boldsymbol{Y}_i$ in the population derived using the first order approximation and $C_i^{-1/2}$ is obtained using the Cholesky decomposition of $\mathbf{C}_i$. $\boldsymbol{SPEY}_i$ is derived from the full variance matrix of predictions so are decorrelated assuming that the approximations made by linearization are negligible. The standardized predictions errors were derived, for each observation, from the mean value and its variance, computed using the first-order approximation around the mean of the model like in the FO linearization approach used in NONMEM. $\boldsymbol{SPEY}_i$ is obtained with NONMEM under the name weighted residual denoted WRES$_i$ using the first order approximation of the model.

Under H_0 and based on the first-order approximation, the prediction errors $SPEY_{ij}$ should have a normal distribution with mean 0. Testing the model adequacy using the assumed N(0, 1) distribution of the weighted residuals was first proposed by Vozeh (13).

Standardized Prediction Error on Observations with Simulation (SPEYS). Instead of using the model predictions at $\eta = 0$ to estimate $\boldsymbol{PRED}_i$ and a linearisation of the model to estimate $\mathbf{C}_i$, we can use Monte Carlo simulations to get better estimates of the mean and the variance of the predictive distribution of each $\boldsymbol{Y}_i$. Using the design of V and model M^B, we simulated K datasets V^{simk}. Let k denote the kth simulation (k = 1,..., K), and Y_i^{simk} the vector of simulated observation of the ith subject for this kth simulation. Let $\boldsymbol{E}(\boldsymbol{Y}_i)$ denote the vector of the mean of observations for the ith subject, estimated empirically over the k simulations as:

$$E(Y_i) = \frac{1}{K}\sum_k \left(Y_i^{\text{simk}}\right) \tag{6}$$

Let Var ($\boldsymbol{Y}_i$) be the full predicted variance of Y_i estimated empirically from the K simulations. We define the standardized prediction error on observations with simulations for the ith subject $SPEYS_i$ as:

$$SPEYS_i = Var(Y_i)^{-1/2} \times (Y_i - E(Y_i)) \qquad (7)$$

If K is large enough, under H_0 and based on the first-order approximation, the mean and variance of SPEYS_{ij} should be 0 and 1. By using NONMEM terminology SPEYS are a form of simulated WRES.

Normalized Prediction Distribution Errors on Observations with Simulation (NPDEYS). SPEY and SPEYS are standardized errors both defined by analogy to normal residuals, SPEYS using better estimates of mean and variance but their distribution is not normal for nonlinear models. As an alternative, we can consider the whole distribution to define prediction distribution errors (26). Let F_{ij} denote the cumulative distribution function (cdf) of the predictive distribution of Y_{ij} under M^B. We define the prediction distribution error $PDEYS_{ij}$ as the value of F_{ij} at Y_{ij}, F_{ij} (Y_{ij}). F_{ij} can be approached using Monte Carlo simulation of V^{simk} as described previously. $PDEYS_{ij}$ is then computed as the percentile of Y_{ij} in the empirical distribution of the Y_{ij}^{simk} :

$$PDEYS_{ij} = F_{ij} = \frac{1}{K}\sum_{k} \delta_{ijk} \qquad (8)$$

where δ_{ij} = 1 if $Y_{ij}^{\mathrm{simk}} \leq Y_{ij}$, and = 0 otherwise. These $PDEYS_{ij}$ are correlated within an individual i. To obtain decorrelated $PDEYS_{ij}$, we used $\boldsymbol{E}(\boldsymbol{Y}_i)$ and $\mathrm{Var}(\boldsymbol{Y}_i)$ estimated empirically from the K simulations and calculated $Y_i^{\mathrm{simk}*} = Var(Y_i)^{-1/2}\left(Y_i^{\mathrm{simk}} - E(Y_i)\right)$ and $Y_i{*} = Var(Y_i)^{-1/2}$ $(Y_i - E(Y_i))$. We then calculated F_{ij} based on these two new vectors $Y_i^{\mathrm{simk}*}$ and $\boldsymbol{Y}_i{*}$ instead of Y_i^{simk} and Y_i.

Under H_0, if K is large enough, the distribution of the PDEYS should follow a uniform distribution over [0,1] by construction of the cdf. Normalized prediction distribution errors (NPDEYS) can then be obtained using the inverse function of the normal cumulative density function implemented in most software. By construction NPDEYS_{ij} follow a N(0, 1) distribution without any approximation and are uncorrelated within an individual i.

Tests and Graphs. For each of the three metrics on concentrations, a Wilcoxon signed-rank test can be performed to test whether the mean is significantly different from 0, and a Fisher test can be performed to test whether the variance is significantly different from 1. Under the approximations mentioned previously, SPEY_{ij} and SPEYS_{ij} should follow a normal distribution if H_0 is true, while NPDEYS_{ij} should follow a normal distribution without any approximation. This can be tested using the Shapiro–Wilks test (SW), which tests the normality assumption with no constraints on mean and variance. We have to consider sequentially three tests (Wilcoxon, Fisher or Shapiro–Wilks tests) to decide whether to reject a validation dataset. Indeed, under the H_0, SPEY, SPEYS and NPDEYS should follow a normal distribution N(0, 1). The most important test is for the mean (Wilcoxon) than for the variance (Fisher), than for the distribution (Shapiro–Wilks) and we propose to do them in that order.

Graphically, the metrics can be represented by scatterplots *versus* time to look at the behaviour of the variances. We can also assess distributional assumptions using quantile–quantile plots (QQ-plots). Quantiles from the metrics distribution (SPEY, SPEYS and NPDEYS) can be plotted against quantiles of the theoretical distribution N(0, 1). Departures from the theoretical distribution can be visually assessed by plotting the unity line $y = x$. Histograms can be plotted instead of QQ-plots to represent the distribution of the metrics.

Metrics Based on Hyperparameters

Model evaluation can be performed on hyperparameters. The model developed on the learning dataset B can be used to estimate the hyperparameters in the validation dataset V. Let Ψ_q^V be the qth hyperparameter estimated with the model in V, which we compared to the qth hyperparameter estimated in B, Ψ_q^B .

Standardized Prediction Error on Hyperparameter (SPEH). We define SPEH_q for the qth hyperparameter as the following Wald statistic:

$$SPEH_q = \frac{\Psi_q^V - \Psi_q^B}{\sqrt{SE\left(\Psi_q^V\right)^2 + SE\left(\Psi_q^B\right)^2}} \qquad (9)$$

where $SE\left(\Psi_q^V\right)$ (respectively, $SE\left(\Psi_q^B\right)$) is the standard error of estimation for the qth hyperparameter in the analysis on V (respectively, on B). Asymptotically, maximum likelihood estimators follow a normal distribution. Therefore, under H_0, SPEH_q should follow N(0, 1).

Standardized Prediction Error on Hyperparameter with Simulation (SPEHS)

As previously, K datasets V^{simk} using M^B with design V are simulated. For each simulated dataset, the vector of hyperparameters $\boldsymbol{\Psi}^{\mathrm{simk}}$ is estimated. The qth hyperparameter estimated on V with M^B, Ψ_q^V is then compared to the empirical distribution of Ψ_q^{simk} .

Test and Graphs. The value of SPEH can be compared to the corresponding critical value of a N(0, 1). The hyperparameters can be compared one by one or with a global test (27). To compare the whole vector of the Q hyperparameters between the two analyses with a global approach, the null hypothesis: $\{\boldsymbol{\Psi}^B - \boldsymbol{\Psi}^V = 0\}$ can also be tested using the global Wald test, which statistic is given by:

$$T^2(\Psi) = \left(\Psi^B - \Psi^V\right)'\left(\Gamma^B + \Gamma^V\right)^{-1}\left(\Psi^B - \Psi^V\right) \qquad (10)$$

where $\boldsymbol{\Gamma}^V$ (respectively, $\boldsymbol{\Gamma}^B$) is the full variance matrix of estimation in V (respectively, in B). Asymptotically $T^2(\boldsymbol{\Psi})$ follows a chi-square with Q degrees of freedom under H_0.

For tests applied to SPEHS, the K values of Ψ_q^{simk} are sorted and the percentile of Ψ_q^{simk}, perc, is defined as the number of Ψ_q^{simk} below Ψ_q^V divided by K. Then the p value of the two sided test based on the empirical distribution are calculated as:

$$p = 2 \times \min\left(perc, (1 - perc)\right) \qquad (11)$$

p is compared with 0.05. A Simes procedure can be applied to the Q p values. To illustrate SPEHS, a histogram of the predictive distribution of simulated hyperparameters is plotted, on which the estimated value on V, Ψ_q^V is overlayed.

Metrics Based on the Objective Function

The objective function (OF) given in NONMEM corresponds to minus twice the log-likelihood plus some constant terms. OF can be determined on a dataset V with model M^B and hyperparameters Ψ^B without fitting ($\mathrm{OF}^V_{\mathrm{nofit}}$; all parameters fixed), or with hyperparameters Ψ^V after fitting the model on V ($\mathrm{OF}^V_{\mathrm{fit}}$; all parameters estimated). Several metrics can be defined from these objective functions, with and without Monte Carlo simulation.

Prediction Error on Objective Function (PEOF). We compute the difference $\Delta\mathrm{OF}^V$ between OF^V_{fit} and OF^V_{nofit} :

$$\mathrm{PEOF}^V = \mathrm{OF}^V = \mathrm{OF}^V_{\mathrm{nofit}} - \mathrm{OF}^V_{\mathrm{fit}} \tag{12}$$

Prediction Error on Objective Function with Simulation (PEOFS). OF^V_{nofit} can also be compared to the posterior predictive distribution of the objective function estimated from K simulated datasets with M^B, yielding to values $OF^{\mathrm{sim}k}_{\mathrm{nofit}}$. By using PEOFS, we make the assumption that the simulated dataset have the same number of observations.

Prediction Error on Gain in Objective Function with Simulation (PEGOFS). As the simulated datasets may have different number of values below the limit of quantification, they may have a different number of observations after treating the BQL. The empirical posterior distribution for PEOFS does not correct for the varying number of data involved in each simulated dataset and we think it is then preferable to compare the observed gain of objective function on the simulated dataset.

A third approach compares therefore the $\Delta\mathrm{OF}^V$ with its posterior predictive distribution. For each simulated dataset k, we estimate parameters with M^B and calculate the gain in objective function $\left(\mathrm{OF}^{\mathrm{sim}k}_{\mathrm{nofit}} - \mathrm{OF}^{\mathrm{sim}k}_{\mathrm{fit}}\right)$ which is then compared to $\Delta\mathrm{OF}^V$.

Tests and Graphs. For the metric without Monte Carlo simulation, there is no test to compare $\mathrm{OF}^V_{\mathrm{nofit}}$ and $\mathrm{OF}^V_{\mathrm{fit}}$. If the model is true, the difference should be small. $\mathrm{OF}^V_{\mathrm{nofit}}$ can be compared to the empirical distribution of $\mathrm{OF}^{\mathrm{sim}k}_{\mathrm{nofit}}$ and a p value can be obtained as for SPEHS. To compare $\Delta\mathrm{OF}^V$ with the empirical distribution of $\Delta\mathrm{OF}^{\mathrm{sim}k}$, as $\mathrm{OF}^V_{\mathrm{nofit}}$ is necessarily higher than or equal to $\mathrm{OF}^V_{\mathrm{fit}}$, we calculate the p value of an unilateral test as:

$$p = (1 - perc) \tag{13}$$

The p value can be compared with 0.05.

To illustrate PEOFS, we plot histograms of the predictive distribution of $\mathrm{OF}^{\mathrm{sim}k}_{\mathrm{nofit}}$ or $\Delta\mathrm{OF}^{\mathrm{sim}k}$, and we show the estimated value on V, OF^V_q and $\Delta\mathrm{OF}^V$.

RESULTS

Metrics Illustration on the Two Simulated Datasets

Simulated concentrations *versus* time data for both V_{true} and V_{false} datasets are displayed in Fig. 1. The dashed lines represent the 80% prediction interval, obtained for each time-point as the 10th and 90th percentiles for 1,000 simulations under M^B. For V_{true}, 167 of 192 concentrations are inside the 80% prediction interval, *versus* only 74 of 190 for V_{false}. It is clear from this plot that V_{false} is not well described by M^B, with a large number of concentrations above the 80% prediction interval so that no further metric would be needed for a real example. In the following, we apply all the metrics described above to these two datasets to show how they are able to detect and explain potential adequacies and inadequacies of M^B and to compare theoretical statistical properties of the metrics.

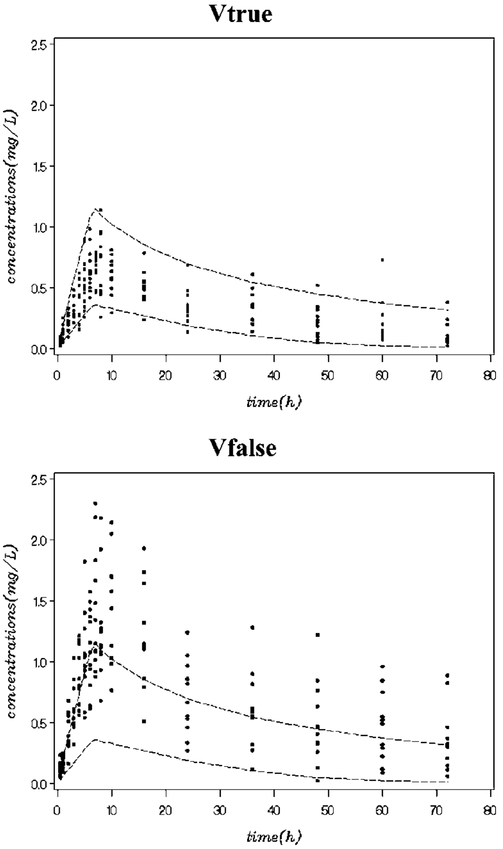

Fig. 1. Simulated concentrations *versus* time for V_{true} (*top*) and V_{false} (*bottom*). The *dashed lines* represent the 80% predicted interval, obtained for each time-point as the 10th and 90th percentiles of 1,000 simulations under M^{B}.

Metrics Based on Observations

The three standardized metrics based on observations are plotted *versus* time or *versus* predictions. The plots of these three metrics *versus* time are shown in Fig. 2. SPEY, SPEYS and NPDEYS have an homogeneous distribution for V_{true} with low variance (1.15, 1.08, 0.84, respectively) and are scattered around zero (means of 0.02, 0.04, 0.02, respectively). For V_{false}, these metrics have upper variance (4.50, 3.63, 1.68, respectively) and are mainly positive (mean are 0.73, 0.58 and 1.35, respectively). SPEYS and NPDEYS were calculated using K = 1,000 simulations.

The QQ-plot compares the distribution of each of these metrics with a normal $N(0, 1)$ distribution (Fig. 3). For V_{true}, points are close to the line $y = x$. On the contrary, for V_{false},

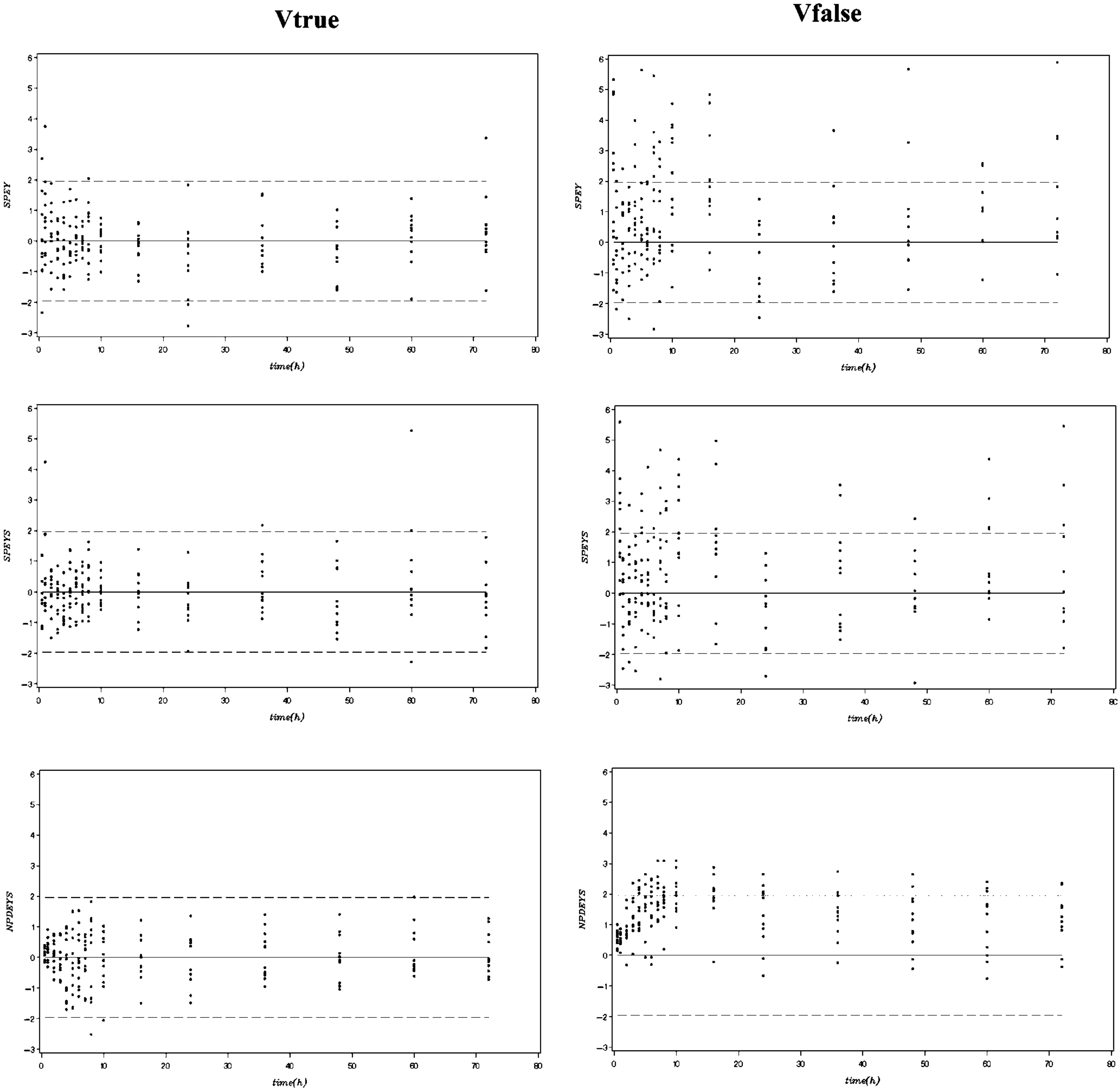

Fig. 2. Metrics based on observations plotted *versus* time on V_{true} (*left*) and on V_{false} (*right*). *Top*: SPEY; *middle*: SPEYS; *bottom*: NPDEYS. The *dashed lines* represent the 95% prediction interval for a normal distribution.

points are systematically biased away from the line $y = x$, which suggests that SPEY, SPEYS and NPDEYS do not follow a normal $N(0, 1)$ distribution. The NPDEYS seem more sensitive visually for the QQ-plot.

The results of the statistical tests performed on the two datasets for these metrics are given in Table II. The mean of the three metrics is not significantly different from 0 for V_{true}, and the variance does not differ from 1. For V_{true}, the distribution of both SPEY and SPEYS is found to differ significantly from a normal distribution with the SW test even though the data were simulated under the model, whereas NPDEYS do not deviate from a normal distribution. This illustrates that only NPDEYS have the good theoretical properties of following a $N(0, 1)$ under H_0 as discussed earlier, without any approximation.

For V_{false}, the means of the three metrics are significantly different from 0 and the variances are significantly different from 1. The distribution of both SPEY and SPEYS is found to differ significantly from a normal distribution with the SW test. However NPDEYS do not differ from a normal distribution in this example but the mean and variance tests are significantly different from 0 and 1, respectively. Therefore, for V_{false}, the three metrics do not follow a $N(0,1)$ distribution, and H_0 was rejected.

The three metrics based on concentration can discriminate V_{true} and V_{false} by visual inspection but the trend in the

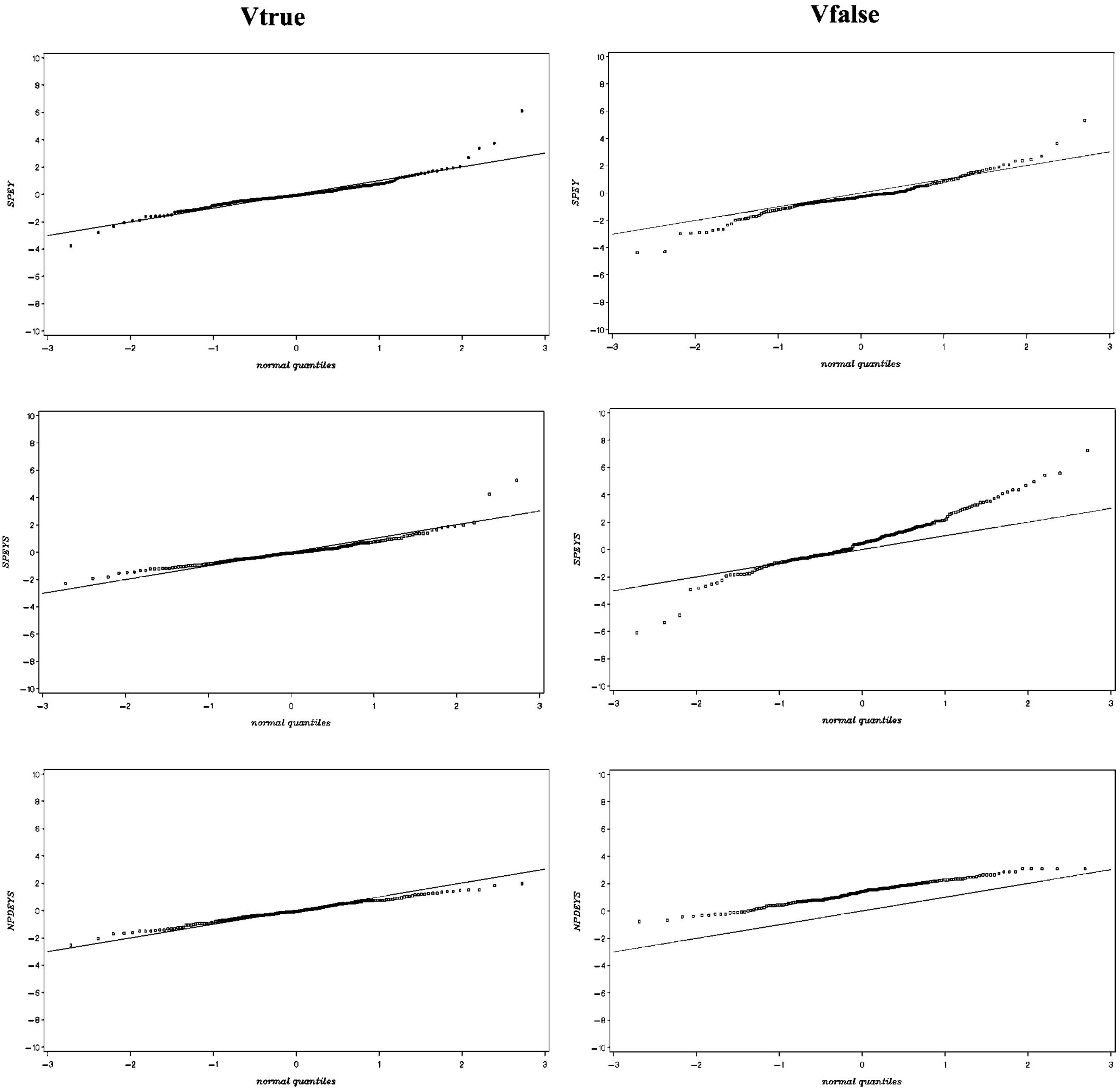

Fig. 3. QQ-plots of the metrics based on observations *versus* the theoretical $N(0,1)$ distribution for V_{true} (*left*) and V_{false} (*right*). The line $y = x$ is shown to evaluate the adequacy between the theoretical and the observed distribution. *Top*: SPEY; *middle*: SPEYS; *bottom*: NPDEYS.

plots are more apparent for NPDEYS than for the two other metrics.

Metrics Based on Hyperparameters

Table III shows the estimates of the hyperparameters and their standard errors on V_{true} and V_{false} after fitting, along with the estimates of the hyperparameters used for M^B. Using Wald tests on each hyperparameter, the estimates found with V_{true} are not significantly different from the previous estimates in B. As expected, there is a significant difference between B and V_{false} for CL/F and V/F. The global difference between the vector of estimates for B and V_{true} is non-significantly different from 0 with a global Wald test ($T^2_{V_{\text{true}}} = 0.79$, $p = 0.99$ for a chi-square with 6 degrees of freedom) but is significantly different from 0 between B and V_{false} ($T^2_{V_{\text{false}}} = 108$, $p < 0.0001$).

For hyperparameters with Monte Carlo simulation, the estimated values of the population parameters on V_{true} are within the simulated posterior predictive distribution of each parameter. This is illustrated on the histograms in Fig. 4 for CL/F and $\omega^2_{CL/F}$. There is no significant departure from the prediction distribution for CL/F ($p = 0.69$), V/F ($p = 0.72$), T_{abs} ($p = 0.75$), $\omega^2_{CL/F}$ ($p = 0.83$), $\omega^2_{V/F}$ ($p = 0.29$) and σ^2 ($p = 0.20$) using the Simes procedure.

For V_{false}, the test shows a significant departure, using the Simes procedure, for the predictive distribution for CL/F ($p < 0.0001$), V/F ($p < 0.0001$), but not for T_{abs} ($p = 0.35$),

Table II. P Values of the Tests Performed on the Three Standardised Metrics Based on Observations, for V_{true} and V_{false}: Mean, Variance and Shapiro–Wilks (SW) Normality Tests

Dataset	Metric	Mean test	Variance test	SW test
V_{true}	SPEY	0.21	0.15	<0.0001
	SPEYS	0.92	0.19	<0.0001
	NPDEYS	0.96	0.10	0.79
V_{false}	SPEY	<0.0001	<0.0001	<0.0001
	SPEYS	<0.0001	<0.0001	0.003
	NPDEYS	<0.0001	<0.0001	0.09

$\omega^2_{\mathrm{CL}/F}$ (p = 0.68), $\omega^2_{V/F}$ (p=0.086) and σ^2 (p = 0.26), as could be expected given that only CL and V were changed in the simulation for V_{false}.

Metrics Based on Objective Function

For V_{true}, the objective function given by NONMEM with M^B is −751 without fitting and −754 with fitting. The gain in objective function PEOF from fitting is 3. For the metrics based on objective function with Monte Carlo simulation, histograms of the predictive distribution of the objective function without fitting and of the gain in objective function are displayed in Fig. 5 to illustrate PEOFS and PEGOFS, respectively. The vertical line corresponds to the value of the objective function when M^B is applied to V_{true} without estimation (top graph in Fig. 5) or to the value of the gain in objective function from fitting (bottom graph in Fig. 5). Compared to the prediction distribution in the simulated datasets we do not reject V_{true} both for PEOFS (p = 0.092) and for PEGOFS (p = 0.90).

For V_{false}, the objective function with M^B is −421 without fitting and −474 with fitting, so PEOF is 53. V_{false} is rejected for PEOFS (p < 0.0001) based on its prediction distribution in the simulated dataset and is also rejected considering PEGOFS (p < 0.0001). Illustrations of these two metrics are shown in Fig. 5.

The three metrics perform similarly on the two datasets. For each dataset, 192 observations are simulated but after treating the BQLs, the two datasets have different number of observations (192 for V_{true} and 190 for V_{false}). So the method comparing the gain of objective function, PEGOFS is more adapted if we compare V_{true} and V_{false}.

Illustration with the Real Dataset

Finally these metrics were applied to the real phase I dataset (V_{real}), the design of which was used to simulate V_{true} and V_{false}. A plot of the concentration *versus* time data for V_{real} is displayed in Fig. 6. Here 144 concentrations (out of 179) are inside the 80% prediction interval but the variability of these concentrations seems to be smaller than for V_{true}.

The results of the tests performed for the metrics based on observations are given in Table IV. The mean of the SPEY is significantly different from 0, but the means of SPEYS and NPDEYS are not. The variance of the three metrics is significantly different from 1 and their distributions do not follow a normal distribution according to the SW test. The scatter plots *versus* time of the three metrics, SPEY, SPEYS and NPDEYS are displayed in Fig. 7. and visually rejected V_{real}.

Concerning metrics based on hyperparameters without Monte Carlo simulation as shown in Table V, the differences between estimated CL/F, T_{abs}, $\omega^2_{\mathrm{CL}/F}$, and σ^2 for B and V_{real} are significantly different from 0 with a Wald test, while there was no significant difference for V/F and $\omega^2_{V/F}$. However, when the whole vector of the six hyperparameters between the dataset B and V_{real} are compared with a global Wald test, we do not reject the null hypothesis ($T^2_{V_{\mathrm{real}}} = 7.93$ and p=0.24 for a chi-square with 6 degrees of freedom). Using the predictive distribution for the hyperparameters obtained using Monte Carlo simulation, significant departures from the predictive distribution were found for CL/F (p < 0.0001), T_{abs} (p < 0.0001), $\omega^2_{\mathrm{CL}/F}$ (p < 0.001), $\omega^2_{V/F}$ (p = 0.03), and σ^2 (p < 0.0001) but not for V/F (p = 0.93) using the Simes procedure. Histograms of the predictive prediction of simulated hyperparameters (CL/F and $\omega^2_{\mathrm{CL}/F}$) are displayed in Fig. 4.

For V_{real}, the objective function with M^B is −600 without fitting and −661 with fitting. So PEOF is 61. For the metrics based on objective function with Monte Carlo simulation, histograms of the predictive distribution of the objective function without fitting and of the gain in objective function are displayed in Fig. 5, with the observed value as a vertical line. We therefore do not reject V_{real} using PEOFS (p = 0.062) but we reject it using PEGOFS (p < 0.0001). After treating the BQL measurements, V_{real} has finally 179 data (instead of 192 for V_{true}) which may explain the discrepancy.

In conclusion, model M^{B}, developed on a dataset of 209 phase II patients, did not adequately predict the data

Table III. Population Pharmacokinetic Parameters of Gliclazide (Estimate and Relative Standard Error of Estimation, RSE) Used for M^{B}

Hyperparameter	B		V_{true}			V_{false}		
	Estimate	RSE (%)	Estimate	RSE (%)	P	Estimate	RSE (%)	P
CL/F (l/h)	1.0	(4.0)	0.98	(16.3)	0.96	0.48	(16.7)	<0.0001
V/F (L)	40.0	(5.8)	42.0	(3.3)	0.59	20.0	(7.5)	<0.0001
T_{abs} (h)	6.6	(3.3)	6.5	(7.9)	0.78	7.0	(0.1)	0.08
$\omega^2_{\mathrm{CL}/F}$	0.35	(17.1)	0.27	(48.1)	0.56	0.34	(38.2)	0.99
$\omega^2_{V/F}$	0.11	(27.2)	0.09	(33.3)	0.55	0.06	(33.3)	0.16
σ^2	0.06	(10.0)	0.05	(10.0)	0.31	0.06	(10.0)	0.82

The second and third columns are the parameters of V_{true} and V_{false} estimated with independent population analyses. P is the p value of the Wald test for each population parameter of V_{true} and V_{false} compared to B.

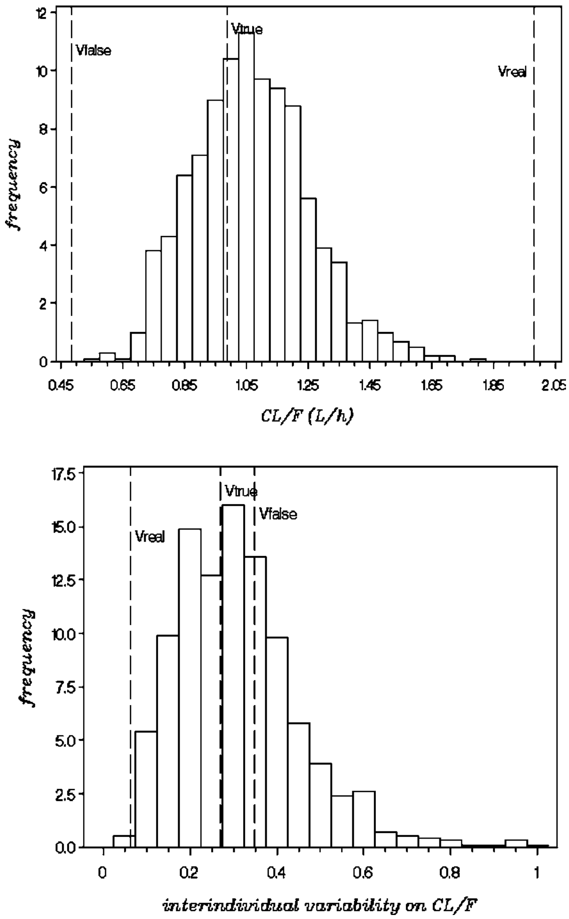

Fig. 4. Histogram of the predictive distribution of simulated hyperparameters estimated using M^B: for CL/*F* (*top*) and $\omega^2_{CL/F}$ (*bottom*). The values of the corresponding parameters found for V_{true}, V_{false} and V_{real} using an independent population analysis are shown as *vertical lines.*

observed in V_{real}, the dataset collected from 12 healthy volunteers. The main differences were lower number of subjects, higher clearance in the phase I subjects, as well as lower inter-individual variability and lower residual error. Metrics based on observations (concentrations here) were consistent in showing model misfit, while metrics based on hyperparameters highlighted the differences between the two datasets. Finally, PEGOFS was more powerful than metrics based on likelihood to detect the differences, because of the large number of BQL in V_{real}.

DISCUSSION

Model assessment consists in the evaluation of how well a model describes a dataset. In this paper, we consider external evaluation, a comparison between a validation dataset and the predictions from the population model built from the learning dataset using design information from the validation dataset. We illustrate known, as well as new or improved metrics to perform external evaluation using two simulated and one real validation datasets. These metrics are based on observations, hyperparameters or objective function. Some metrics are built with Monte Carlo simulations, which are performed using the estimated population model to be evaluated with the design of the validation dataset. In this example, the model M^B is a pharmacokinetic model, so the observations are concentrations but it is possible to apply these metrics to a pharmacodynamic model.

In this paper, we used real data obtained during the development of gliclazide, an antidiabetic drug. A population pharmacokinetic model of gliclazide was first built using data from two phase II studies. We found that the variance of the residual error was lower in the study where electronic pillboxes (MEMS) were used. Indeed, the observance was better taken into account using MEMS because the records of dates and times of drug administration were more accurate.

External evaluation of the model M^B was then performed using the dataset from a real phase I study (V_{real}). Two datasets were also simulated using the design of this Phase I study (V_{true} and V_{false}). V_{true} was simulated according to the model and to the hyperparameters values estimated in the phase II studies. V_{false} was simulated using the same model but with a bioavailability multiplied by two, that is, dividing by two the values obtained for CL/*F* and *V*/*F*. We simulated these two datasets to illustrate the ability of the metrics to validate V_{true} and reject V_{false}.

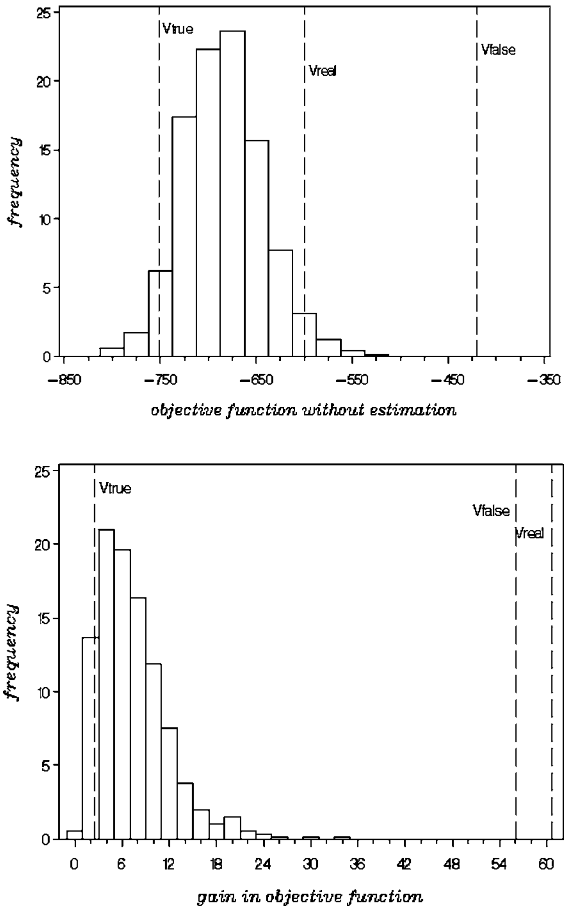

Fig. 5. Histogram of the predictive distribution of the objective function when model M^B is applied to the 1,000 datasets without estimation (*top*) and the gain in objective function (*bottom*). The values of the objective functions or of the gain found for V_{true}, V_{false} and V_{real} using M^B are shown as *dotted lines.*

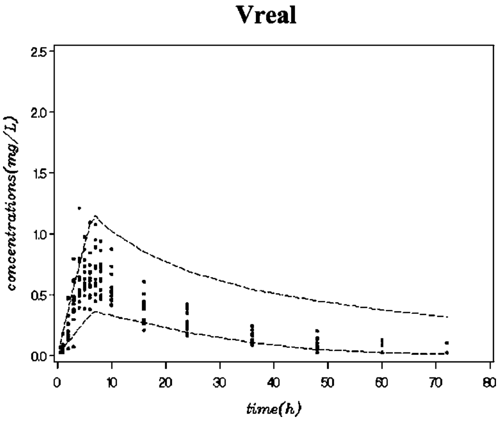

Fig. 6. Concentrations *versus* time for V_{real}. The *dashed lines* represent the 80% predicted interval, obtained for each time-point as the 10th and 90th percentiles of 1,000 simulations under M^{B}.

The metrics most often used in model evaluation are prediction errors and standardized prediction errors (SPEY, called WRES in NONMEM) on observations (28). The term "WRES" is widely used by NONMEM users but "standardized prediction error" is a general term used in the FDA guidance and by other authors in the context of model validation. Indeed, these metrics are called "residuals" when they are applied to the same dataset (internal validation) but when applied to an external validation dataset, the denomination "error" is more appropriate than "residual" although they are computed similarly and are reported as WRES in NONMEM tables. SPEY is not an optimal metric for external model evaluation because pharmacokinetics models are generally nonlinear with respect to the parameters (although the pharmacokinetics of a drug is often assumed to be linear with respect to dose.) SPEY relies on a linear approximation of the mixed-effect model around the mean as in the FO estimation method even if the FOCE estimation methods is used for estimation (29). In the present work, this problem appears since SPEY did not follow a normal distribution even when the dataset has been simulated using the true model under H_0. A somehow more refined strategy consists in using simulations to recover the empirical mean and variance of the predictive distribution for each observation, thus computing what we called SPEYS. However these SPEYS suffer from some of the same theoretical flaws as SPEY. However SPEY and SPEYS present good behaviour to reject our V_{false} as judged by graphically inspection. Moreover, SPEY (or WRES) present the advantage to be automatically given by softwares such as NONMEM. Also noticing that WRES is a poor metric (because based on the FO approximation), Hooker *et al.* proposed computing another metric that they called conditional WRES (CWRES) in which the FOCE approximation is used for the computation of the mean and the variance of the model (30). We did not apply this metric here because we have computed SPEYS, which calculate the mean and of the variance of the model based on simulations as opposed to using the FOCE approximation. The main limitations of all these metrics (SPEY, SPEYS and CWRES) is that they come from the theory of linear models and that they implicitly assume that the observations are normally distributed around the mean which is not true for nonlinear models.

Table IV. Different Tests Proposed for Normalized Metrics Based on Concentration for V_{real}: Mean and Variance Tests and Shapiro–Wilks (SW) Normality Tests

Dataset	Metric	Mean test	Variance test	SW test
V_{real}	SPEY	0.03	<0.0001	<0.0001
	SPEYS	0.72	0.0044	<0.0001
	NPDEYS	0.85	0.0005	0.01

We use a new approach based on the calculation of prediction errors on observations, called PDEYS or NPDEYS in their normalised version (19). This metric does not require any assumption on the distribution of the observations, and,

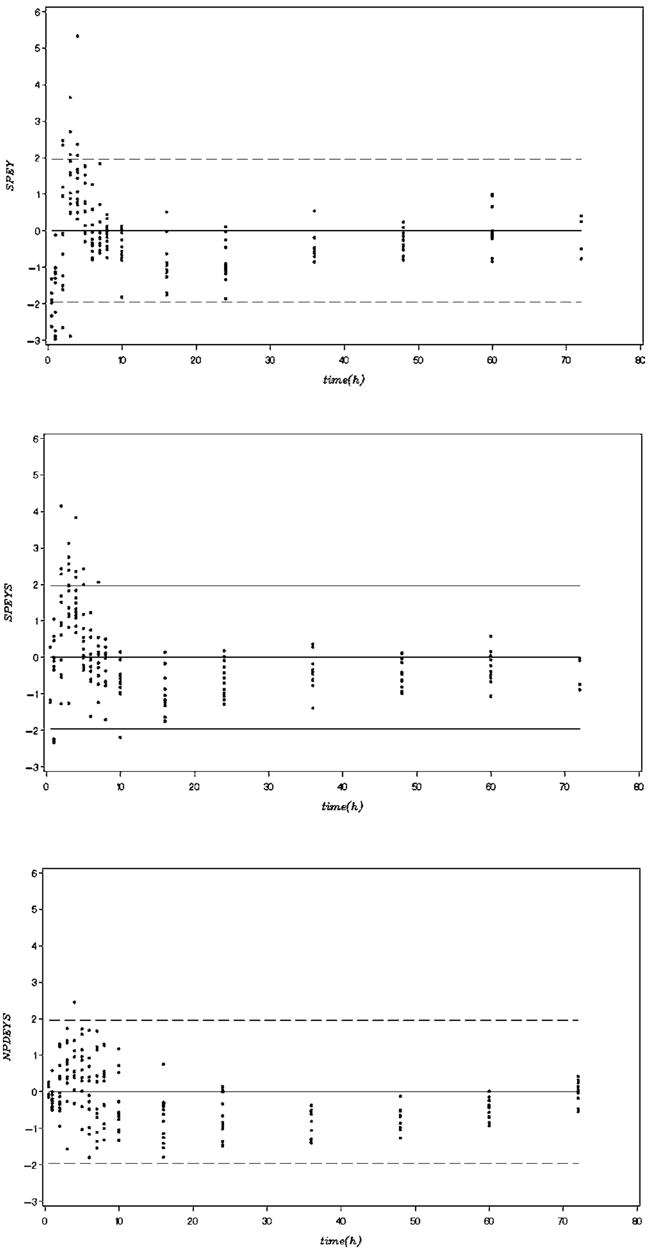

Fig. 7. Metrics based on observations plotted *versus* time on V_{real}. *Top*: SPEY; *middle*: SPEYS; *bottom*: NPDEYS. The *dashed lines* represent the 95% prediction interval for a normal distribution.

Table V. Population Pharmacokinetic Parameters of Gliclazide (Estimate and Relative Standard Error of Estimation, RSE) with Data from the Phase I Study (V_{real})

Population parameters	V_{real}		
	Estimate	RSE (%)	P
CL/F (l/h)	2.0	(7.5)	<0.0001
V/F (l)	40.6	(6.5)	0.82
T_{abs} (h)	5.7	(4.6)	0.009
$\omega^2_{\mathrm{CL}/F}$	0.06	(33.3)	<0.0001
$\omega^2_{V/F}$	0.05	(40.0)	0.054
σ^2	0.10	(10.0)	0.003

P is the p value of the Wald test for each hyperparameter compared to M^{B}.

when computed with a large number of simulations, has a known theoretical distribution which can be tested. This metric was applied in the present paper but was improved as compared to the previous applications (19,20) in that the within subject correlation between observations is now taken into account. Using the uncorrelated version of the NPDEYS, their variance was significantly different from 1 for V_{true} (0.65) as opposed to the correlated version of the metric. In the previous paper this metric was named the prediction discrepancy, here we use the name prediction errors to be more homogenous in the paper with the other metric (SPEY, SPEYS...) and also because now that they are normalized and decorrelated these metrics are more in the spirit of an error than of a measure of discrepancy.

Regarding the tests applied to the different metrics, the null hypothesis is that the model is correct, so that we can only invalidate a model when we reject H_0, never accept it. To test H_0 using metrics based on observations, we propose to use simultaneously a mean test, a variance test and a normality test. The Shapiro–Wilks has become the preferred test of normality because of its good power as compared to a wide range of alternative tests. The Kolmogorov–Smirnov test that was used in the previous paper (19) is very general to test any distribution and may have lower power compared to other normality tests like Anderson–Darling test, Cramer–von-Mises test or Shapiro–Wilks test. Because the Kolmogorov–Smirnov test is very conservative, there is a high likelihood of not rejecting normality. Mean tests are more efficient to detect problems with fixed effects, as here with V_{false}, and if significant the SW test is not needed.

The three metrics were able to reject V_{false} by visual inspection, however the trend is more visible for NPDEYS. Applying statistical tests on these metrics, SPEY and SPEYS showed a significant difference in the mean and variance tests for V_{false}, but the SW test was significant on both datasets V_{false} and V_{true}. On the other hand, the approach based on NPDEYS does not reject V_{true} and, based on the combination of the three tests, also rejected V_{false} corresponding to the expected theoretical properties. However, although the mean and variance tests were significant for V_{false}, the NPDEYS are not significantly different from a normal according to the SW test (p = 0.09) in this example. We do not have an explanation, and further investigations are needed.

Metrics based on hyperparameters without Monte Carlo simulation were also interesting for external evaluation. The Wald test assessed whether the population estimates were significantly different in the building and validation datasets, taking into account the precision in the estimation of the hyperparameter. A correction for multiple tests was applied with the Simes procedure, and we were able to pick up the differences in the fixed effects. The global Wald test was also used, but this test, when significant, did not allow us to detect which parameters were different between the two datasets. Note that the Wald test assumes the normality of the estimators. Metrics with Monte Carlo simulation did not make this assumption and provided the same results in this simple example. Simulation carries however a large computational burden, because all the simulated datasets have to be re-fitted.

Finally, we introduce new metrics based on objective function. The metrics performed similarly on the simulated validation datasets but we recommend the use of the metric based on the gain in objective function evaluated by simulation (PEGOFS) for the following reason. In this paper we deal with BQL data using a standard method in population pharmacokinetics which consists in replacing the first BQL measurement in a series with the value LOQ/2 and censoring the following BQL measurements. We applied this method to both the original and the simulated datasets. The distribution of the objective functions resulting from the fit of the simulated datasets therefore arises from datasets with potentially different number of observations, and objective functions are obtained as minus twice the log-likelihood up to a constant which depends on the number of observations. Therefore the empirical posterior distribution of the objective function (PEOFS) obtained using the simulated data may not be as accurate as expected. On the other hand, PEGOFS are defined as the difference in objective function between fitted and non-fitted models, and therefore the constant is eliminated from their expression, so that PEGOFS do not suffer from the same problem as PEOFS in the presence of a varying number of BQL measurements. As an alternative, we could also account for the presence of BQL by computing the contribution of BQL data to the likelihood, but this would require complex computations in NONMEM.

Yano, Beal and Sheiner defined posterior predictive check (PPC) (10) and proposed three approaches to compute the posterior distribution of the parameters estimated through the maximum likelihood estimation (31,32). Here the metrics NPDEYS, SPEHS, PEOFS and PEGOFS are all forms of PPC. They were built without considering the estimation error, the simplest of the three approaches implemented by Yano *et al.*, who have shown it to perform well in large enough datasets.

The metrics were finally applied to a real dataset, V_{real}. Model M^B was not found to be valid when applied to this dataset collected from 12 healthy volunteers according to most of the metrics proposed above. The metrics based on observations or on objective function demonstrated model misfit, while metrics based on hyperparameters highlighted the differences between the two datasets (learning and validation datasets). Using metrics based on hyperparameters, the main differences were a higher apparent clearance in the phase I subjects, as well as a lower interindividual variability for CL/F and V/F and a smaller residual error variance. These results can be explained by the differences between a phase I and a phase II study. In a phase I study, there are few subjects, participants are healthy volunteers (except for oncology studies), young, often male and have normal body weight and normal biological functions, and the pharmacokinetics in patients may show a number of modifications.

All the evaluation methods we presented aim at providing one or a small set of metrics to assess model adequacy. As such they are criteria combining information about various sources of model misspecification and it is not always easy to assert which part of the model should be improved. For instance for evaluation through predicted concentrations, a model with or without covariates should have correct standardized prediction errors if the estimation of inter-individual variability is adequate (28). Also misspecification of the error model may lead to errors in the model of the random effects, which are not always easy to find when exploring only the *post hoc* distribution (33). We therefore recommend to use several approaches or metrics to evaluate a model in order to provide a more informative overview.

CONCLUSION

In conclusion, the three groups of metrics discussed here can be used to evaluate a population model. The choice of the metrics depends on the objectives of the analysis. Model evaluation based on observations is crucial if the model is to be used for clinical trial simulation or for therapeutic drug monitoring. Amongst the first type of metrics based on concentration, SPEY (WRES) are easily computed and were able to pick out the problem in V_{false} according to the visual inspection of the metrics and two of the three statistical tests but their calculation are based on the first order method which is not always used in modelling. As simulations are often performed in population analyses to calculate prediction intervals or to perform visual predictive check, we can use the same simulations to compute SPEYS or NPDEYS and apply the statistical tests. We recommend in a final step to use NPDEYS over SPEY or SPEYS since they do not depend on an approximation of the model. If the aim is to compare two populations, metrics based on hyperparameters are very useful to highlight differences between the datasets. Model evaluation based on objective function is a good approach to evaluate a series of models during the building process. These last metrics based on objective function are an interesting new tool for external evaluation. Amongst the three metrics based on objective function proposed, we recommend using the metrics based on the empirical distribution of the difference in objective function between fitted and no-fitted models (PEGOFS), obtained through simulations. Metrics based on hyperparameter, SPEHS or on the delta of objective functions, PEGOFS, need a simulation and an estimation step. So these methods are time consuming and should be applied to the final model or in the building process only if the model is simple enough.

REFERENCES

1. L. B. Sheiner and J. L. Steimer. Pharmacokinetic/pharmacodynamic modeling in drug development. *Annu. Rev. Pharmacol. Toxicol.* **40**:67–95 (2000).
2. L. Aarons, M. O. Karlsson, F. Mentre, F. Rombout, J. L. Steimer, and A. van Peer. Role of modelling and simulation in Phase I drug development. *Eur. J. Pharm. Sci.* **13**:115–122 (2001).
3. R. Jochemsen, C. Laveille, and D. D. Breimer. Application of pharmacokinetic/pharmacodynamic modelling and population approaches to drug development. *Int. J. Pharm. Med.***13**:243–251 (1999).
4. N. H. Holford, H. C. Kimko, J. P. Monteleone, and C. C. Peck. Simulation of clinical trials. *Annu. Rev. Pharmacol. Toxicol.* **40**:209–234 (2000).
5. L. J. Lesko, M. Rowland, C. C. Peck, and T. F. Blaschke. Optimizing the science of drug development: opportunities for better candidate selection and accelerated evaluation in humans. *Pharm. Res.* **17**:1335–1344 (2000).
6. H. C. Kimko and S. B. Duffull. *Simulation for Designing Clinical Trials: A Pharmacokinetic–Pharmacodynamic Modeling Prospective*. Marcel Dekker, New York, 2003.
7. Food and Drug Administration. *Guidance for Industry:population pharmacokinetics* (available at http://www.fda.gov/cder/guidance/index.html,1999).
8. E. I. Ette. Stability and performance of a population pharmacokinetic model. *J. Clin. Pharmacol.* **37**:486–495 (1997).
9. P. J. Williams and E. I. Ette. *Determination of Model Appropriateness*. Marcel Dekker, New York, 2003.
10. Y. Yano, S. L. Beal, and L. B. Sheiner. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. *J. Pharmacokinet. Pharmacodyn.* **28**:171–192 (2001).
11. E. H. Cox, C. Veyrat-Follet, S. L. Beal, E. Fuseau, S. Kenkare, and L. B. Sheiner. A population pharmacokinetic–pharmacodynamic analysis of repeated measures time-to-event pharmacodynamic responses: the antiemetic effect of ondansetron. *J. Pharmacokinet. Biopharm.* **27**:625–644 (1999).
12. P. Girard, T. F. Blaschke, H. Kastrissios, and L. B. Sheiner. A Markov mixed effect regression model for drug compliance. *Stat Med.* **17**:2313–2333 (1998).
13. S. Vozeh, T. Uematsu, G. F. Hauf, and F. Follath. Performance of Bayesian feedback to forecast lidocaine serum concentration: evaluation of the prediction error and the prediction interval. *J. Pharmacokinet. Biopharm.* **13**:203 –212 (1985).
14. J. W. Mandema, R. F. Kaiko, B. Oshlack, R. F. Reder, and D. R. Stanski. Characterization and validation of a pharmacokinetic model for controlled-release oxycodone. *Br. J. Clin. Pharmacol.* **42**:747–756 (1996).
15. K. Fattinger, S. Vozeh, H. R. Ha, M. Borner, and F. Follath. Population pharmacokinetics of quinidine. *Br. J. Clin. Pharmacol.* **31**:279–286 (1991).
16. T. H. Grasela, J. B. Fiedler-Kelly, C. Salvadori, C. Marey, R. Jochemsen, and H. Loo. Predictive performance of population pharmacokinetic parameters of tianeptine as applied to plasma concentrations from a post-marketing study. *Eur. J. Clin. Pharmacol.* **45**:123–128 (1993).
17. L. Aarons, S. Vozeh, M. Wenk, P. Weiss, and F. Follath. Population pharmacokinetics of tobramycin. *Br. J. Clin. Pharmacol.* **28**:305–314 (1989).
18. L. B. Sheiner and S. L. Beal. Some suggestions for measuring predictive performance. *J Pharmacokinet. Biopharm.* **9**:503–512 (1981).
19. F. Mentré and S. Escolano. Prediction discrepancies for the evaluation of nonlinear mixed-effects Models. *J. Pharmacokinet. Pharmacodyn.* **33**:345–367 (2006).
20. E. Comets, K. Ikeda, P. Hoff, P. Fumoleau, J. Wanders, and Y. Tanigawara. Comparison of the pharmacokinetics of S-1, an oral anticancer agent, in Western and Japanese patients. *J. Pharmacokinet. Pharmacodyn.* **30**:257–283 (2003).
21. N. Frey, C. Laveille, M. Paraire, M. Francillard, N. H. Holford, and R. Jochemsen. Population PKPD modelling of the long-term hypoglycaemic effect of gliclazide given as a once-a-day modified release (MR) formulation. *Br. J. Clin. Pharmacol.* **55**:147–157 (2003).
22. S. L. Beal. Ways to fit a PK model with some data below the quantification limit. *J. Pharmacokinet. Pharmacodyn.* **28**:481–504 (2001).
23. P. Delrat, M. Paraire, and R. Jochemsen. Complete bioavailability and lack of food-effect on pharmacokinetics of gliclazide

30 mg modified release in healthy volunteers. *Biopharm. Drug Dispos*. **23**:151–157 (2002).

24. R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**:751–764 (1986).
25. S. L. Beal and L. B. Sheiner. NONMEM Users Guides (I–VIII). *Globomax LLC; Hanover, maryland*. (1989–1998).
26. F. Mesnil, F. Mentré, C. Dubruc, J. P. Thenot, and A. Mallet. Population pharmacokinetic analysis of mizolastine and validation from sparse data on patients using the nonparametric maximum likelihood method. *J. Pharmacokinet. Biopharm*. **26**:133–161 (1998).
27. 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).
28. F. Mentré and M. E. Ebelin. Validation of population pharmacokinetic/pharmacodynamic analyses: review of proposed approaches. *The Population Approach: Measuring and Managing Variability in Response Concentration and Dose*. Office for official publications of the European Communities, Brussels, 1997, pp. 141–158.
29. A. Gelman, J. B. Carlin, H. S. Stern, and R. D.B. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.
30. A. Hooker and M. O. Karlsson. Conditional weighted residuals. A diagnostic to improve population PK/PD model building and evaluation. *AAPS J*. **7**(S2), Abstract W5321 (2005).
31. M. J. Bayarri and P. Berger. *P* values for composite null models. *JASA* **95**:1143–1172 (2000).
32. J. M. Robins, A. van der Vaart, and V. Ventura. Assymptotic distribution of *P* values in composite null models. *JASA* **95**: 1143–1172 (2000).
33. G. Verbecke, and E. Lesaffre. A linear mixed-effects models with heterogeneity in random effects population. *JASA* **91**: 217–221 (1996).