HAL
open science

A New Exact Test for the Evaluation of Population Pharmacokinetic and/or Pharmacodynamic Models Using Random Projections

Céline M. Laffont, Didier Concordet

▶ **To cite this version:**

Céline M. Laffont, Didier Concordet. A New Exact Test for the Evaluation of Population Pharmacokinetic and/or Pharmacodynamic Models Using Random Projections. Pharmaceutical Research, 2011, 28 (8), pp.1948-1962. 10.1007/s11095-011-0422-9 . hal-01137013

HAL Id: hal-01137013

https://hal.science/hal-01137013v1

Submitted on 30 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pharm Res (2011) 28:1948–1962
DOI 10.1007/s11095-011-0422-9

RESEARCH PAPER

A New Exact Test for the Evaluation of Population Pharmacokinetic and/or Pharmacodynamic Models Using Random Projections

Celine Marielle Laffont • Didier Concordet

Received: 17 November 2010 / Accepted: 4 March 2011 / Published online: 14 April 2011
© Springer Science+Business Media, LLC 2011

ABSTRACT

Purpose Within-subject dependency of observations has a strong impact on the evaluation of population pharmacokinetic (PK) and/or pharmacodynamic (PD) models. To our knowledge, none of the current model evaluation tools correctly address this issue. We present a new method with a global test and easy diagnostic plot which relies on the use of a random projection technique that allows the analysis of dependent data.

Methods For each subject, the vector of standardised residuals is calculated and projected onto many random directions drawn uniformly from the unit sphere. Our test compares the empirical distribution of projections with their distribution under the model. Simulation studies assess the level of the test and compare its performance with common metrics including normalised prediction distribution errors and different types of weighted residuals. An application to real data is performed.

Results In contrast to other evaluated methods, our test shows adequate level for all models and designs investigated, which confirms its good theoretical properties. The weakness of other methods is demonstrated and discussed.

Conclusions This new test appears promising and could be used in combination with other tools to drive model evaluation in population PK/PD analyses.

KEY WORDS GUD · model evaluation or diagnostics · nonlinear mixed effects models · normalised prediction distribution errors NPDE · weighted residuals

NOTATIONS

KS	Kolmogorov-Smirnov
VPC	visual predictive check
GUD	global uniform distance

INTRODUCTION

Population pharmacokinetic (PK)/pharmacodynamic (PD) models are widely used by the pharmaceutical industry to model dose/concentration/effect relationships, to assess between-subject variability and to support decision making (1–3). Regulatory authorities (FDA, EMEA) greatly encourage the use of these models to identify differences in drug safety and efficacy among population subgroups (e.g. elderly, children, renally impaired) and to facilitate the drug development process (3–6). At the same time, FDA and EMEA guidance documents (5,6) stress the need for model evaluation depending on the purpose of the analysis: if population PK/PD models are used simply to describe the data, then basic evaluation methods can be applied. However, if models are used to predict drug exposure and effects in the target patient population with an impact on drug labelling as it is more and more the case (4,7), then more advanced and objective methods are needed.

The problem with population PK/PD models is that evaluation of them is not easy. Indeed, they are nonlinear

C. M. Laffont · D. Concordet
INRA, UMR 1331, Toxalim
F-31027 Toulouse, France

C. M. Laffont · D. Concordet
Université de Toulouse, INPT, ENVT, UPS, EIP
F-31076 Toulouse, France

C. M. Laffont (✉)
Laboratoire de Biostatistique
Ecole Nationale Vétérinaire de Toulouse
23 chemin des Capelles, BP 87614
31076 Toulouse cedex 03, France
e-mail: c.rousset@envt.fr

mixed effects models, and the classical evaluation methods developed for linear Gaussian models cannot be used in a straightforward manner. For standard linear Gaussian models, model evaluation relies mainly on the inspection of studentised or weighted (standardised) residuals, which should be centred, homoscedastic, independent and distributed according to a Student distribution. Actually, when the number of observations is large enough, this Student distribution is close to $N(0,1)$. In the case of nonlinear mixed effects models, weighted residuals do not have the same theoretical properties because individual vectors of observations are not Gaussian. The non-normality of observations is constitutive of nonlinear mixed effects models (observations are a nonlinear function of Gaussian random variables), and this is why the estimation is so difficult. The result is that weighted residuals cannot be normally distributed and independent under the model as commonly assumed (8,9). Of course, it can be verified that their mean is zero and their variance is 1, but as the shape of their actual distribution is unknown and as the effect of dependence cannot be easily anticipated, it is quite difficult to evaluate the overall adequacy of the model. Interpretation of residual plots is therefore very much dependent on the experience of the modeller, and, even though Monte Carlo simulations can be used to obtain reference patterns and improve decision making (10), they are not an optimal tool for the evaluation of population PK/PD models. On top of this, weighted residuals may not be properly calculated. In some cases (WRES, CWRES), they are computed using the expectation and variance-covariance matrix of observations derived from an approximated model (first-order linearisation around the null value of random effects or around their individual predictions), whatever the method of estimation. Simulation studies have shown that wrong decisions could be taken with these types of residuals, i.e., the correct model could be rejected while strong model misspecifications could be missed (9–11). A better way to compute weighted residuals is to use the expectation and variance-covariance matrix of observations estimated from Monte Carlo simulations, provided enough simulations are performed (11). This is a better method in the sense that there is no approximation of the model. PWRES (displayed by MONOLIX software (12) or EWRES (displayed by NONMEM version 7 (13)) are examples of such Monte Carlo-generated residuals.

In that context, more advanced evaluation tools have been proposed such as visual predictive checks (VPCs) and normalised prediction distribution errors (NPDE) (11). VPCs graphically compare observations with their predictive distribution under the model and are quite helpful in identifying misfits and suggesting model improvements (Karlsson and Holford. PAGE meeting 17 (2008) Abstr 1434). However, their interpretation remains subjective as no strict rule of decision exists. One usual approach consists in computing prediction intervals from model predictions at each time point and assessing whether the percentage of observations outside these prediction intervals matches the theoretical coverage. The problem is that no classical test can be applied due to the data dependency within subjects, e.g. a major increase in type I errors (around 13%) was reported for the exact binomial test applied to VPCs (14).

Data dependency within subjects is a real issue for the evaluation of nonlinear mixed effects models when there is more than one observation per subject. The reader may be more familiar with the term "data correlation," which is widely used by the PK/PD community, but clearly these two terms are not synonymous. Independence implies decorrelation, while the reverse is not true: decorrelation implies independence only when it is applied to Gaussian variables. The difference between the two can be shown graphically after plotting one time component *vs.* another (e.g. data at 1 h *vs.* data at 4 h). If no particular pattern is observed, this means that the time components are independent and thus that there is no data dependency within subjects. In this case, a linear regression through the data gives a horizontal line, indicating the absence of a correlation. Conversely, when there is a pattern in the data, there is data dependency, but still a horizontal linear regression line may be observed: in that particular case, data are uncorrelated but dependent. These different scenarios are further illustrated in the article for some examples. Note that the data dependency investigated here is a marginal data dependency that includes the serial correlation between observations, as has been previously investigated in ref. (15).

A recent method was developed to address the issue of data correlation within subjects and to propose an objective assessment of the model. This method is based on the computation of normalised prediction distribution errors (NPDE), which should follow an $N(0,I)$ distribution under the model, and a global test (11,16). If the test indicates a departure from the $N(0,1)$ distribution, the model should be rejected with a nominal risk of 5%; otherwise, it should not be rejected. The limitation of this method is that it does not cope with the issue of data dependency, only data correlation. This point has been rightly discussed by the authors (16), but the consequences on model evaluation have not been thoroughly investigated. Therefore, although NPDE represent a major improvement over weighted residuals, both in theory and in practice (11,14,16), they might not be an optimal tool for the objective assessment of population PK/PD models.

In the present paper, we thus introduce a new method for model evaluation that properly handles data dependency within subjects. The idea is to use a random projection technique (see (17) for general methodology and (18–21) for

some applications) that allows the analysis of dependent data. Based on this method, we present a global test, which is exact whatever the sample size, and an innovative diagnostic graph that does not require any subjective interpretation. Several simulation studies were performed under different models to assess the level of the test and to compare the performance of our method with those of NPDE and commonly used weighted residuals obtained by model linearisation (WRES, CWRES) or Monte Carlo simulation (PWRES).

Background

This section provides the notations used throughout the paper and gives details of the computation of NPDE and weighted residuals assessed in the simulation studies. All computations were done in Microsoft Visual C++ 6.0.

Notations

We consider the general mixed effects model:

$$Y_i = f_i(x_i, \theta, \eta_i) + \Gamma_i(x_i, \theta, \eta_i, \sigma)\varepsilon_i \quad (1)$$

where $Y_i = (Y_{ij})_{j=1\ldots n_i}$ is the n_i-vector of observations in subject i ($i = 1\ldots N$), x_i is the vector of independent variables for subject i, θ is an unknown p-vector of fixed parameters (or fixed effects), η_i is a q-vector of real random effects associated with subject i and accounting for interindividual variability, and $\varepsilon_i = (\varepsilon_{ij})_{j=1\ldots n_i}$ is a vector of random variables accounting for residual error (model misspecification, analytical error), σ being a real constant. In this model, Γ_i is a known positive diagonal $n_i \times n_i$-matrix depending on x_i, θ, η_i and σ, while f_i denotes a known real vector-valued function depending on x_i, θ and η_i. For nonlinear mixed effects models, f_i depends nonlinearly on η_i. We assume that η_i is independent and identically distributed (i.i.d.) $N_p(0,\Omega)$ and that ε_i is i.i.d. $N(0, I_{n_i})$. We also assume that η_i and ε_i are independent.

All model evaluation tools described hereafter are usually calculated conditionally to model parameter estimates ($\widehat{\theta}$, $\widehat{\Omega}$ and $\widehat{\sigma}$), and all expectations/variances appearing in the text are conditional expectations/variances. To simplify the notations, these conditional expectations/variances are denoted $E()/Var()$. However, as the aim of the paper was not to discuss parameter estimations, all the calculations were done assuming that the parameters θ, Ω and σ were known.

Calculating Weighted Residuals

Three types of weighted residuals were selected for the comparison with our method, namely WRES, CWRES and PWRES. They are all calculated in different manners: either by linearisation of the model (WRES, CWRES) or through Monte Carlo simulations (PWRES), and are all commonly used for the evaluation of population PK/PD models. WRES is the default provided by all versions of NONMEM (22), which is the gold standard software for population PK/PD analyses. CWRES is provided by the latest (7th) version of NONMEM (13) or can be computed using verbatim code in NONMEM and a post processing step implemented in either R or MATLAB (9). PWRES is provided by Monolix software using default settings (12) and is also available from NONMEM version 7 under the name of EWRES (13) (with some possible difference in computation depending on the software and the number of simulations used).

Currently, all these residuals are obtained with the same general formula:

$$W_i = Var(Y_i)^{-1/2} \times (Y_i - E(Y_i)) \quad (2)$$

where W_i denotes the vector of weighted residuals in subject i, $E(Y_i)$ is the expectation vector of Y_i, and $Var(Y_i)$ is the full variance-covariance matrix of Y_i. In the present paper, $Var(Y_i)^{1/2}$ refers to the Cholesky decomposition of $Var(Y_i)$.

What differs between WRES, CWRES and PWRES is the way $E(Y_i)$ and $Var(Y_i)$ are obtained. For PWRES, $E(Y_i)$ and $Var(Y_i)$ are estimated empirically over K simulations of the study design without any approximation of the model (here, K = 2,000). In the case of WRES, $E(Y_i)$ and $Var(Y_i)$ are computed using the first-order (FO) approximation of the model, while for CWRES, $E(Y_i)$ and $Var(Y_i)$ are computed using the first-order conditional estimation (FOCE) approximation to the model, which is a less crude approximation than FO. In both cases, the model is linearised with respect to η_i, either around zero (FO) or around their individual predictions (FOCE). Further details on the computation of WRES and CWRES are available in ref. (9). By construction, when $E(Y_i)$ and $Var(Y_i)^{1/2}$ are properly calculated, the components of W_i are uncorrelated within subject i.

Calculating NPDE

NPDE, developed by Brendel *et al.* (11) and first presented under the name of NPDEYS, compare, for each subject i and each $j = 1\ldots n_i$, the observation Y_{ij} with its predictive distribution under the model. In order to remove part of the data dependency within subject i, weighted residuals W_i are computed from Y_i using Eq. 2 and Monte Carlo simulations so that W_i components are uncorrelated within subject i. In that case, W_i is equivalent to PWRES. Let F_{ij} denote the cumulative distribution function (cdf) of the predictive distribution of W_{ij} under the model for all $i = 1\ldots N$ and all $j = 1\ldots n_i$. By construction of the cdf, $F_{ij}(W_{ij})$ should follow a uniform distribution over [0,1] when the model is correct. Normalised prediction distribution errors,

$NPDE_{ij}$, are then obtained from $F_{ij}(W_{ij})$ by applying the inverse of the standard normal cdf Φ as follows:

$$NPDE_{ij} = \Phi^{-1}(F_{ij}(W_{ij})). \quad (3)$$

By construction, $NPDE_{ij}$ are thus expected to follow an $N(0,1)$ distribution under the model. This does not imply, however, that the overall joint distribution of NPDE is standard Gaussian. In order to have NPDE~$N(0,\mathrm{I})$, all W_i components must be independent for all $i = 1\ldots N$ and all subjects must be independent. While subject independence is a reasonable assumption, the independence of W_i components can only be achieved when Y_i is normally distributed because in that case, data decorrelation leads to data independency. As Y_i is not normally distributed for nonlinear mixed effects models, we do not expect the joint distribution of NPDE to be an $N(0,\mathrm{I})$.

As for $E(Y_i)$ and $Var(Y_i)$, F_{ij} is determined empirically over K simulations of the study design. Let $Y_{i,sim(k)}$ denote the vector of simulated "observations" in subject i for the kth simulation ($k = 1\ldots K$) with the tested model. The corresponding vector of weighted residuals $W_{i,sim(k)}$ is then calculated as $Var(Y_i)^{-1/2} \times (Y_{i,sim(k)} - E(Y_i))$ according to Eq. 2, and F_{ij} is obtained as the empirical cdf of $W_{ij,sim(k)}$ over the K simulations (see (16) for further details). Here, we chose $K = 2{,}000$.

MATERIALS AND METHODS

In the first part, we present our new method for the evaluation of population PK/PD models, which is based on the use of a random projection technique and on the calculation of a global uniform distance abbreviated GUD. Based on this distance, we then propose a global test (the GUD test) and a new diagnostic graph (the Q-Q ring plot). In the second part, we present the different simulation studies carried out to assess the level of our test and compare its performance with those of NPDE and different types of weighted residuals (PWRES, WRES, CWRES). Finally, we illustrate the use of our method by applying it to the evaluation of a previously developed model built on real clinical data.

All calculations were done in Microsoft Visual C++ 6.0 using a Dell Precision M6300 laptop (Intel Core2 Duo, T9500, 2.60 GHz, 3.5 Go RAM with Windows XP SP3) except for the application to real data where some Monte Carlo simulations were done with NONMEM software.

Random Projection Concept

Let us first explain the basic idea that makes the random projection technique work. The distribution of a random vector X of dimension n is fully characterised by its projection on all vectors $t \in R^n$. The projection of X on t is defined as the scalar product $\langle X; t\rangle = \sum_{i=1}^{n} X_i t_i$. When X has a variance-covariance matrix equal to identity, the distribution of X is then fully characterised by its projection on all vectors t of unit norm that define the unit sphere. In other words, projecting X on the unit sphere suffices to completely characterise its distribution. In practice, we cannot project X on an infinite number of directions but randomly draw a "sample" of t in the unit sphere according to the uniform distribution, and calculate the projection $\langle X, t\rangle$ for all t sampled. Here, the size of this sample is denoted B.

For the following test, we only consider the distribution of projections regardless of the sampled directions t. The result is that our method does not fully but "partially" characterises the distribution of X, i.e., it is expected to separate distributions that are sufficiently different.

Application of Random Projections to Model Evaluation

In the present case, X refers to the vector of weighted residuals W_i computed for each subject i according to Eq. 2. As for NPDE, $E(Y_i)$ and $Var(Y_i)$ used for the calculation of W_i are determined empirically over K (= 2,000) simulations of the study design performed with the tested model. W is thus equivalent to PWRES. Subjects are random, so W_i is random and we can "partially" characterise the distribution of W_i by projecting it onto B random directions drawn uniformly from the unit sphere of R^{n_i} since all components of W_i have variance 1. Practically, we generate, for each subject i, B independent vectors $a_{ib} \sim N(0, I_{n_i})$ of dimension n_i with $b = 1\ldots B$ (here we choose $B = 100$). Then, as the unit sphere is defined by a set of vectors with an L^2 norm equal to 1, we normalise a_{ib} in order to obtain B vectors of unit norm $e_{ib} = a_{ib} \Big/ \sqrt{\sum_{j=1}^{n_i} a_{ibj}{}^2}$. The projection of W_i on e_{ib} is denoted $P_{ib} = \langle W_i, e_{ib}\rangle = \sum_{j=1}^{n_i} W_{ij} e_{ibj}$. In the end, we obtain for each subject i a single object that is the empirical distribution of random projections $(P_{ib})_{b=1\ldots B}$ denoted $P_i(x) = \frac{1}{B}\sum_{j=1}^{B} \delta_{P_{ib}}(x)$, where $\delta_{P_{ib}}$ is a Dirac mass at P_{ib}. The sample is then defined by a mixture of individual random projection distributions with distribution $P(x) = \frac{1}{N}\sum_{i=1}^{N} P_i(x)$. We can compute the empirical cdf of this mixture, which is denoted C.

Global Test Based on the Calculation of the Global Uniform Distance (GUD)

Our test compares the empirical cdf curve C obtained for the sample to its distribution under the model, which is

assessed by Monte Carlo simulation. Practically, M replicates of the study design are simulated (here, M = 5,000). Let $Y_{i,sim(m)}$ denote the vector of "observations" simulated in subject i for the mth simulation ($m = 1 \ldots M$). We then proceed exactly as described above, i.e. for each simulation m, we compute for each subject $i = 1 \ldots N$ the vector of weighted residuals $W_{i,sim(m)} = Var(Y_i)^{-1/2} \times (Y_{i,sim(m)} - E(Y_i))$ and project this vector onto B (= 100) random vectors drawn uniformly from the unit sphere. We recall that $E(Y_i)$ and $Var(Y_i)$ are determined empirically over K (= 2,000) simulations of the study design. We then obtain, for each simulation m, an empirical cdf $C_{sim(m)}$ characterising the mixture of individual random projection distributions in the simulated sample. In the end, we get M cdf curves corresponding to the M simulations of the study design. The distribution of these curves is summarised by computing a uniform 95% prediction area A. For the calculation of A, we assess for each cdf curve its maximal absolute distance to the mean cdf curve C_{mean} with $C_{mean} = 1/M \times \sum_{m=1}^{M} C_{sim(m)}$. This distance corresponds to the sup-norm and is called $GUD_{sim(m)}$ (for Global Uniform Distance):

$$GUD_{sim(m)} = \left\| C_{sim(m)} - C_{mean} \right\|_{\infty} = \sup_{x} \left| C_{sim(m)}(x) - C_{mean}(x) \right| \quad (4)$$

Based on this distance, we then select 95% of the cdf curves that are the closest to C_{mean}. Practically, $GUD_{sim(m),m=1\ldots M}$ values are sorted and 95% of the cdf curves with the lowest GUD values are selected, enabling calculation of the 95% prediction area A as follows:

$$U(x) = \sup_{s=1\ldots \mathrm{int}(0.95\times 5000)} (C_{sim(s)})(x) \quad (5)$$

$$L(x) = \inf_{s=1\ldots \mathrm{int}(0.95\times 5000)} (C_{sim(s)})(x) \quad (6)$$

where int(a) is the integer part of a, $U(x)$ defines the upper limit of A, $L(x)$ defines the lower limit of A, and $C_{sim(s),s=1\ldots \mathrm{int}(0.95\times 5000)}$ correspond to the 95% selected cdf curves for M = 5,000. Here, x ranges from −5 to 5 with steps of 0.01. Our test evaluates whether or not the sample cdf curve C is fully contained in A. By construction, this should happen in exactly 95% of cases when the model is correct (null hypothesis or H_0). Under the correct model, comparing C to A is completely equivalent to comparing the GUD value obtained for the sample ($= \|C - C_{mean}\|_{\infty}$) to its "theoretical empirical" distribution. In that case, a p value can be calculated:

$$p = 1 - perc \quad (7)$$

where $perc$ corresponds to the number of $GUD_{sim(m),m=1\ldots M}$ below the GUD value obtained for the sample, divided by M. This test will be referred to as the GUD test in the rest of the article.

GUD-Based Diagnostic Plot (Q–Q Ring Plot)

Q–Q plots are commonly used to compare two probability distributions. Here, the objective is to compare the empirical distribution of random projections obtained for the sample (characterised by its cdf C) to its theoretical distribution under the model (characterised by the mean cdf C_{mean}). To do the Q–Q plot, quantiles are calculated for the sample (Q_{sample}) and from the lower/upper limits of A ($Q_{lower\ A}/Q_{upper\ A}$ respectively), using the inverse of the mean cdf denoted $C_{mean}{}^{-1}$:

$$Q_{sample}(x) = C_{mean}{}^{-1}(C(x)) \quad (8)$$

$$Q_{lower\ A}(x) = C_{mean}{}^{-1}(L(x)) \quad (9)$$

$$Q_{upper\ A}(x) = C_{mean}{}^{-1}(U(x)) \quad (10)$$

Figure 1 shows the resulting Q–Q plots for a correct and an incorrect model. As it is difficult to see what happens around zero, we decided to create a new plot, which we called the "Q–Q ring plot." This plot is also a quantile *vs.* quantile plot but uses polar coordinates to transform the 95% prediction area A into a ring delimited by two circles of radii 1 and 2. The inner circle is defined by Cartesian coordinates ($x_{lower} = \cos(2\pi \times (x+5)/10)$; $y_{lower} = \sin(2\pi \times (x+5)/10)$), while the external circle is defined by Cartesian coordinates ($x_{upper} = x_{lower} \times 2$; $y_{upper} = y_{lower} \times 2$). Recall that x ranges from −5 to 5 so that $2\pi \times (x+5)/10$ ranges from 0 to 2π. Regarding the sample curve, it is obtained with the following coordinates: ($x_{sample} = x_{lower} \times radius(x)$; $y_{sample} = y_{lower} \times radius(x)$) with $radius(x) = \max\left(0\ ;\ 1 + \left(Q_{sample}(x) - Q_{lower\ R}(x)\right) / \left(Q_{upper\ R}(x) - Q_{lower\ R}(x)\right)\right)$. An illustration of the Q–Q ring plot is given in Fig. 1 for a correct and an incorrect model.

Assessment of the Level of the Test and Comparison with Other Evaluation Tools (NPDE and Weighted Residuals) Through Simulation Studies

Three models were used for the simulation studies: one linear Gaussian model, one very simple population PK model and one standard population PD model. Observations simulated with these models are shown in Fig. 2 for a representative sample. Further details on the models and designs used in the simulation studies are provided below. Briefly, 5,000 replicates of the study design were simulated

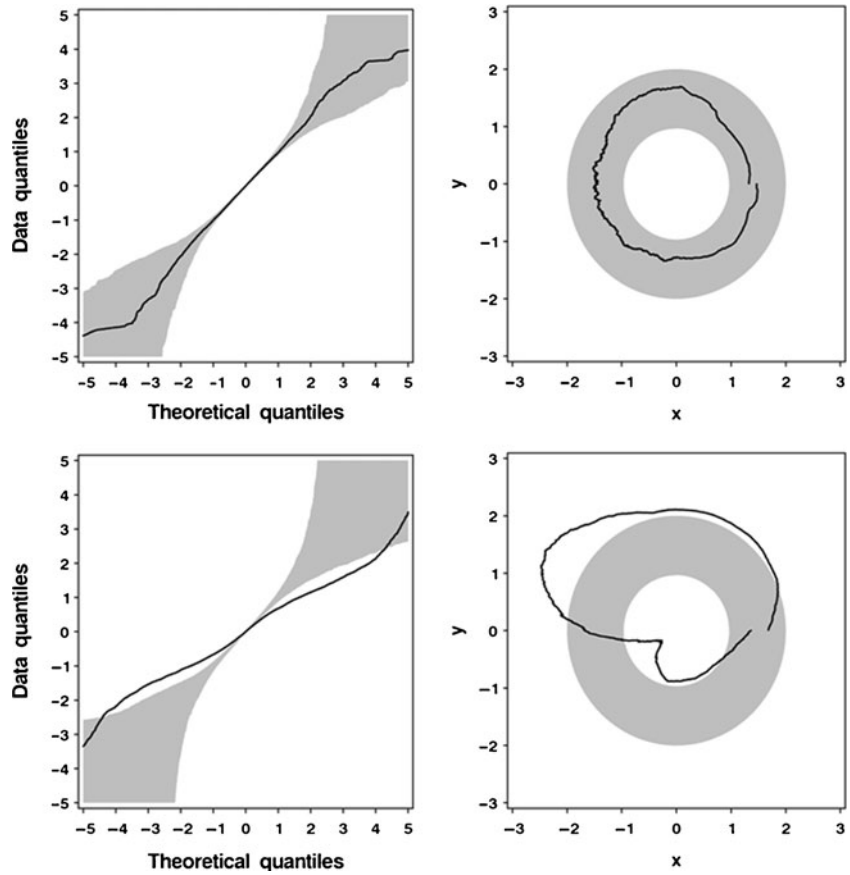

Fig. 1 Classical Q–Q plot on the *left* with the corresponding Q–Q ring plot on the *right* for the correct model (*top*) and an incorrect model (*bottom*). On both plots, the *black curve* refers to the observations, while the *grey area* delineates the 95% prediction area *A* under the model. When the *black curve* is fully contained in the *grey area*, the model is not rejected; otherwise, it is rejected with 5% risk of being wrong. Observations were generated for 100 i.i.d. subjects (four measurements per subject) using the population PK model (one-compartment model with intravenous bolus administration). The incorrect model was obtained by multiplying the geometric mean of clearance (θ_1) by two.

per model and study design. For each replicate, the GUD distance was calculated together with weighted residuals (WRES, CWRES, PWRES) and NPDE (details on the calculation of weighted residuals and NPDE are provided in the sections "Calculating Weighted Residuals" and "Calculating NPDE"). As the objective of the article was to compare the different methods for model evaluation apart from model estimation issues, no estimation of model parameters was carried out. Weighted residuals, NPDE and GUD distance were thus calculated using the true values of model parameters and the model was linearised around the true values of η_i for the computation of CWRES. The Kolmogorov–Smirnov (KS) test was then applied to NPDE and weighted residuals to test for a departure from the $N(0,1)$ distribution. The GUD test was performed in parallel. For each test, the type I error was computed as the percentage of simulated datasets for which the test was significant ($\alpha = 0.05$). For the GUD test and the KS test

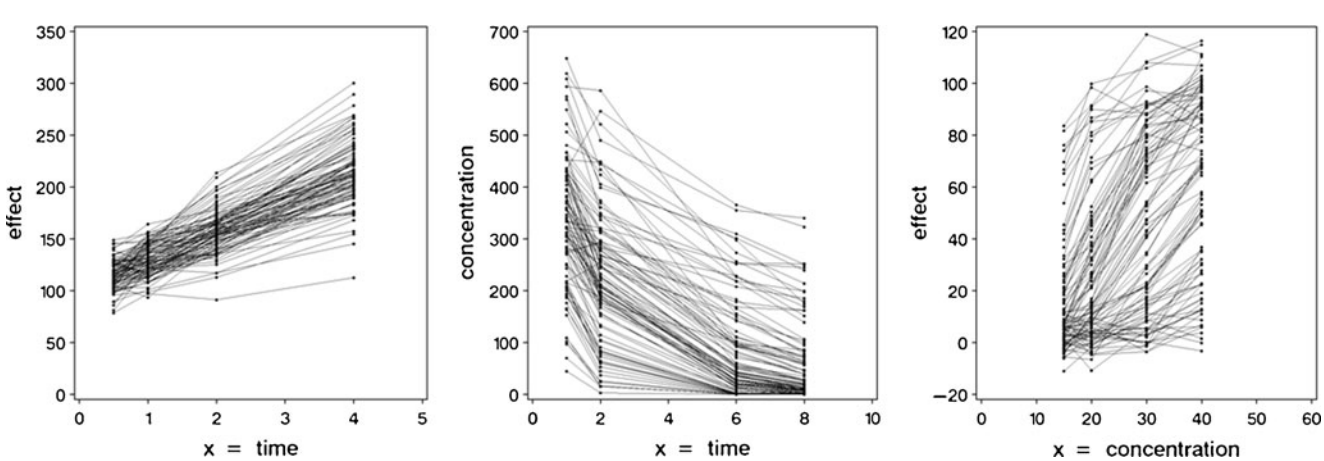

Fig. 2 Simulated data for one representative sample in the different simulation studies with i.i.d. subjects (100 subjects, four measurements per subject). *Left*: linear Gaussian model. *Middle*: population PK model (one-compartment model with intravenous bolus administration). *Right*: population PD model (sigmoidal E_{max} model).

applied to NPDE, probability–probability (P–P) plots were additionally done to see whether the 5,000 p values obtained for each test over the 5,000 replicates followed a uniform distribution over [0,1]. Indeed, if the level of the test is correct, 5% of the p values are expected to lie below 0.05, 10% below 0.10, 50% below 0.50, and so on. The resulting data points on the P–P plots (corresponding to the 5,000 p values obtained for the 5,000 replicates) should be distributed along the identity line. In order to account for the uncertainty related to the number of replicates, a 95% confidence interval was calculated and added to the P–P plots. This interval was derived from the Dvoretzky–Kiefer–Wolfowitz inequality, which measures how close an empirical distribution function is to its expectation (23). In the present case, this 95% prediction interval was equal to $[x-\varepsilon\,;x+\varepsilon]_{x\in[0;1]}$ where $\varepsilon=\sqrt{-\log(0.025)/2n}\approx 0.0192$, n being the number of replicates (= 5,000). As the number of replicates was very large, the 95% prediction interval was quite narrow.

Linear Gaussian Model

A drug effect was simulated for 100 subjects with four observations per subject using the following model:

$$f_i(x_i,\theta,\eta_i)=\psi_{1i}\times x_i+\psi_{2i} \text{ and } \Gamma_i(x_i,\theta,\eta_i,\sigma)=\mathrm{diag}(\sigma),$$

with $\psi_{1i}=\theta_1+\eta_{1i}$ and $\psi_{2i}=\theta_2+\eta_{2i}$, where x_i was the vector of measurement times in subject i. Measurement times were the same in all subjects with x_i=(0.5,1,2,4) hours. The model parameter values were (θ_1,θ_2) = (30,100), $\Omega=\begin{pmatrix}36 & 0\\ 0 & 100\end{pmatrix}$ and σ=10.

Population PK Model

Drug concentration data were simulated for 100 subjects with two to four observations per subject using a one-compartment model with a single intravenous bolus administration and first-order elimination. The model was parameterised in clearance (CL) and volume of distribution (V), with exponential interindividual variability set on CL and V and a proportional residual error:

$$f_i(x_i,\theta,\eta_i)=\frac{D}{V_i}\times\exp(-CL_i/V_i\times x_i) \text{ and } \Gamma_i(x_i,\theta,\eta_i,\sigma)$$
$$=\mathrm{diag}(\sigma\times f_i(x_i,\theta,\eta_i)),$$

D was the dose administered to each subject at time zero (D = 50,000) and x_i was the vector of measurement times in subject i. Two different cases were addressed: one case where subjects were i.i.d. and one case where they were not. In the i.i.d. case, measurement times were the same in all subjects with x_i = (1,2,6,8) hours, and no covariate was included in the model, i.e. $CL_i=\theta_1\times\exp(\eta_{1i})$ and $V_i=\theta_2\times\exp(\eta_{2i})$. The model parameter values were (θ_1,θ_2) = (40,100), $\Omega=\begin{pmatrix}0.49 & 0\\ 0 & 0.04\end{pmatrix}$ and σ = 0.10. In the non-i.i.d. case, measurement times could differ between subjects with x_i = (1,2,6,8) hours in 20% of subjects, x_i = (4,8) hours in 40% of subjects, and x_i = (3,6) hours in the rest of the subjects. In addition, one continuous covariate, creatinine clearance ($CrCL$), was included in the model as follows: $CL_i=(\theta_1+\theta_3\times(CrCL_i-75))\times\exp(\eta_{1i})$ and $V_i=\theta_2\times\exp(\eta_{2i})$ where 75 is the population median of $CrCL$. Individual $CrCL$ values were generated according to a uniform distribution over [30;120]. The model parameter values were $(\theta_1,\theta_2,\theta_3)$ = (38,100,0.4), $\Omega=\begin{pmatrix}0.25 & 0\\ 0 & 0.04\end{pmatrix}$ and σ = 0.10.

Population PD Model

A drug effect was simulated for 100 subjects with four observations per subject using a standard sigmoidal E_{max} model. Exponential interindividual variability was set on E_{max} and EC_{50} and an additive residual error was used.

$$f_i(x_i,\theta,\eta_i)=\frac{E_{\max,i}\times {x_i}^{\gamma}}{{EC_{50,i}}^{\gamma}+{x_i}^{\gamma}} \text{ and } \Gamma_i(x_i,\theta,\eta_i,\sigma)=\mathrm{diag}(\sigma),$$

with $EC_{50,i}=\theta_1\times\exp(\eta_{1i})$, $E_{\max,i}=\theta_2\times\exp(\eta_{2i})$, and $\gamma=\theta_3$, where x_i was the vector of drug concentrations in subject i. In this simple example, x_i was the same in all subjects with x_i=(15,20,30,40) mg/L. The model parameter values were $(\theta_1,\theta_2,\theta_3)$ = (30,100,4), $\Omega=\begin{pmatrix}0.49 & 0\\ 0 & 0.01\end{pmatrix}$ and σ=5.

Application of the GUD Test to Real Pharmacokinetic Data

To illustrate the use of our method, we applied it for the evaluation of a previously developed model (16) built on the well-known dataset for theophylline, an anti-asthmatic agent. Theophylline concentration-time data were obtained from a PK study reported by Boeckmann, Sheiner and Beal (22). In this experiment, the drug was given orally to twelve subjects, and drug serum concentrations were measured at ten time points per subject over the subsequent 25 h. The model under consideration was a one-compartment model with first-order absorption and elimination, parameterised in elimination rate constant (k), apparent distribution volume (V) and absorption rate constant (k_a) as follows:

$$f_i(x_i,\theta,\eta_i)=\frac{D_i\times k_{ai}}{V_i\times(k_{ai}-k_i)}\times(\exp(-k_i\times x_i)-\exp(-k_{ai}\times x_i)),$$

$\Gamma_i(x_i,\theta,\eta_i,\sigma_1,\sigma_2)=\mathrm{diag}(\sigma_1\times f_i(x_i,\theta,\eta_i)+\sigma_2)$ with σ_1 and σ_2 two real constants, $k_i=\theta_1\times\exp(\eta_{1i})$, $k_{a\,i}=\theta_2\times\exp(\eta_{2i})$ and $V_i=\theta_3\times\exp(\eta_{3i})$. x_i was the vector of

real measurement times in subject i and D_i was the dose (mg/kg) given to subject i on a per-weight basis. Details of the estimation method and final parameter estimates are available in (16). Our method was applied with the following settings: B = 100, K = 2,000 and M = 5,000. NPDE were calculated as well with K = 2,000. All computations were done in Microsoft Visual C++ 6.0, except for the K and M Monte Carlo simulations, which were carried out in NONMEM version 7. As for model estimation, the concentration measurements at time zero in the original dataset were removed for model evaluation.

RESULTS

The results of the simulation studies are summarised in Table 1. This table presents the type I errors obtained for the GUD test and for the KS test applied to NPDE and weighted residuals and testing for an $N(0,1)$ distribution. For the linear Gaussian model, all the methods showed a good performance as type I errors were all within the 95% confidence interval of [4.4;5.6]% derived from the normal approximation to the binomial distribution $B(5{,}000;0.05)$ where 5,000 refers to the number of replicates and 0.05 is the nominal risk. These results support a standard Gaussian distribution of WRES, CWRES, PWRES and NPDE and suggest that all the methods can be used equivalently for the evaluation of these types of models.

For the population PK and PD models, however, the methods were not equivalent. Very high type I errors were obtained for the KS test applied to weighted residuals, irrespective of the model. The performances of WRES were particularly poor with type I errors around 100%. It thus appears that neither WRES, CWRES nor PWRES followed a standard Gaussian distribution under the tested models. By comparison, NPDE and GUD gave more reasonable results. Type I errors were very close to 5% for the GUD test. The results were less clear for NPDE with type I errors outside the 95% confidence interval of [4.4;5.6]%, either below for the PK model or above for the PD model. This prompted us to go further in the investigations and examine the distribution of p values over the 5,000 replicates of the study design. As previously mentioned, if the tests perform well under the model, the p values should follow a uniform distribution over [0,1]. This was assessed through P–P plots which are shown in Fig. 3 for NPDE and Fig. 4 for the GUD test. In the case of NPDE, the p values did not follow a uniform distribution over [0,1] whatever the model investigated: data points were not distributed along the identity line and some of the points were outside the 95% prediction interval. We recall that this interval, derived from the Dvoretzky–Kiefer–Wolfowitz inequality, is quite a rough (wide) interval because it is adapted to all possible distributions. Therefore, any data point out of this interval clearly indicates a strong departure from the expected distribution. In the case of the GUD test, all data points were distributed along the identity line and were within the 95% prediction interval. This supports a uniform distribution of p values over [0,1] for all models investigated and confirms the good theoretical properties of the GUD test under H_0. It is interesting to note that for the population PK model, type I errors were less for CWRES and PWRES in the non-i.i.d. case compared with the i.i.d. case, with a slightly better P–P plot for NPDE. However, given the nonlinearity of the model and the complexity of the data dependency phenomenon, it would be hazardous to convert this observation into a general result.

Residual plots were investigated for a sample of simulated datasets. Figure 5 illustrates the case where the GUD test and the KS test applied to NPDE do not reject the model while wrong visual patterns appear on the diagnostic plots. In the present example, WRES are clearly not centred on zero and are not homoscedastic. CWRES are not centred on zero as well. PWRES look much better, but an asymmetric distribution is observed at small values of the independent variable x. Finally, a clear trend appears on the plot of NPDE *vs.* model predictions. Thus, it turns out that only the plot of NPDE *vs.* x looks adequate. In this example, p values were 0.91 for the GUD test and 0.69 for the KS test on NPDE.

The pattern observed in the plot of NPDE *vs.* individual model predictions is explained by the dependency of NPDE data within subjects. This dependency is shown in Fig. 6 by plotting one component of the independent variable x against another (depending on the model, x refers to time after dose or to drug concentrations). If the data are truly independent, a rounded shape should be observed with no particular pattern. In contrast, if a particular pattern

Table 1 Type I Errors (%) Obtained for the GUD Test as well as the Kolmogorov–Smirnov (KS) Test Applied to NPDE or Weighted Residuals (WRES, CWRES, PWRES) in the Different Simulation Studies for a Nominal Level of 5%

		Linear Gaussian model	Population PK model		Population PD model
		i.i.d. case	i.i.d. case	non-i.i.d. case	i.i.d. case
GUD test	GUD	4.94	4.84	4.88	5.00
KS test[a]	NPDE	4.60	3.54	3.80	7.50
	WRES	4.88	100	99.8	100
	CWRES	4.88	50.1	19.2	13.5
	PWRES	4.50	69.3	22.3	62.0

[a] The reference distribution for the KS test is the $N(0,1)$ distribution.

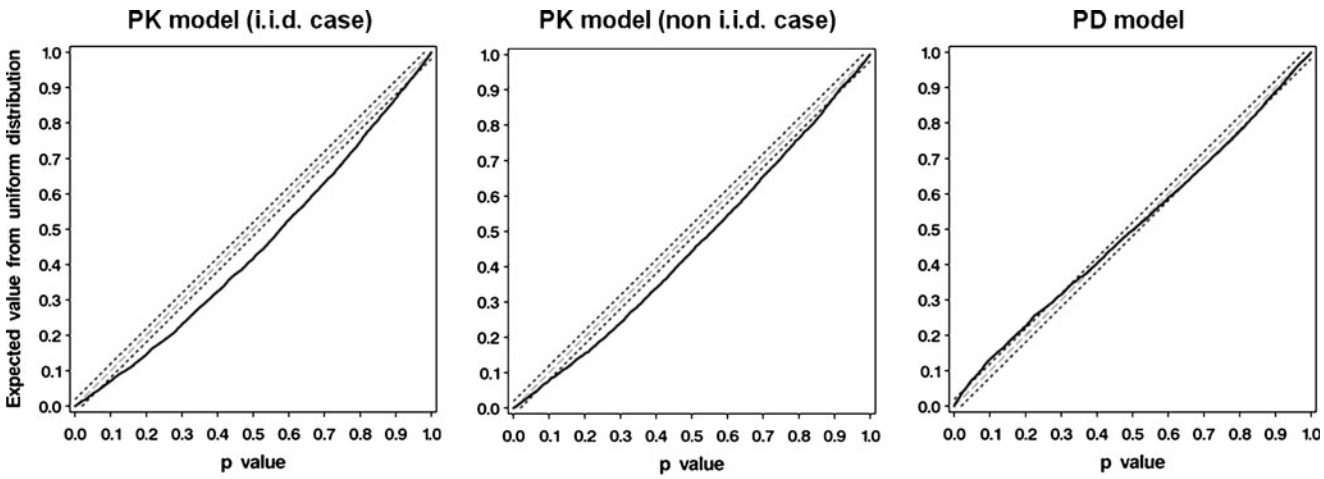

Fig. 3 P–P plot of the *p* values obtained for the KS test on NPDE to check for a uniform distribution over [0, 1]. *Left*: population PK model (i.i.d. case). *Middle*: population PK model (non i.i.d. case). *Right*: population PD model. The dashed grey line $y = x$ is shown to evaluate the adequacy between the theoretical and observed distribution. We can see that the *p*-values do not follow a uniform distribution as the data are not fully contained within the 95% prediction interval derived from Dvoretzky-Kiefer-Wolfowitz inequality (delineated by the *black dotted lines*).

appears in the data, this means that the data are not independent. For the PK and PD models tested, Fig. 6 shows that weighted residuals (PWRES) were uncorrelated but not independent. It results that NPDE were not independent, and Fig. 6 shows that some correlation could re-appear during the computation of NPDE from weighted residuals (i.e. the linear regression line through the data was clearly not horizontal). This correlation was globally negative for the PK model and globally positive for the PD model. With the linear Gaussian model, PWRES and NPDE were both uncorrelated and independent. This is in line with the theory that says that when observations are Gaussian, data decorrelation implies data independence.

Finally, the GUD test was applied to the evaluation of a previous population PK model developed on the theophylline data supplied with NONMEM. Fig. 7 shows theophylline concentration-time data *vs*. their predictive distribution under the model, which is summarised by a 90% prediction interval at each of the following time points which appeared to be the nominal times in the original study: 0.25, 0.50, 1, 2, 3.5, 5, 7, 9, 12 and 24 h. It appears from this plot that the model predicts the data reasonably well, even though interindividual variability is slightly overestimated. Fig. 7 also gives the Q–Q ring diagnostic plot obtained with our method. We can see here a very small departure of the black curve (representing the sample) from the ring (representing the 95% prediction region under the model), which indicates that the model should be rejected. In contrast, the KS test performed on NPDE was not significant ($p = 0.28$), suggesting that the model could not be rejected.

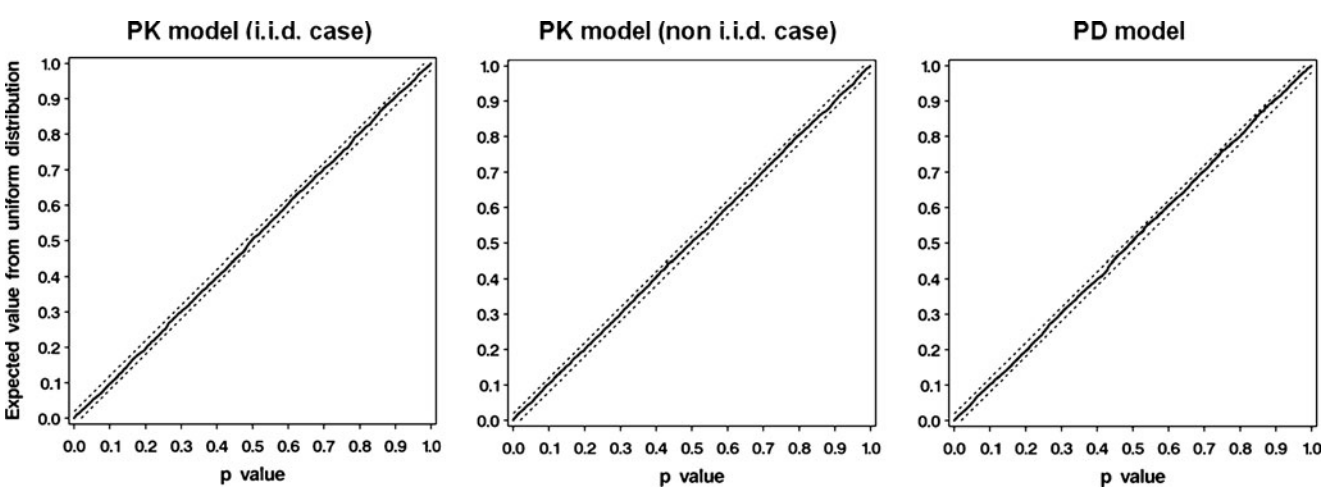

Fig. 4 P-P plot of the *p* values obtained for the GUD test to check for a uniform distribution over [0, 1]. *Left*: population PK model (i.i.d. case). *Middle*: population PK model (non i.i.d. case). *Right*: population PD model. The dashed grey line $y = x$ is shown to evaluate the adequacy between the theoretical and observed distribution. We can see that the *p*-values do follow a uniform distribution, irrespective of the model. Indeed, data points are all contained within the 95% prediction interval derived from Dvoretzky-Kiefer-Wolfowitz inequality (delineated by the *black dotted lines*).

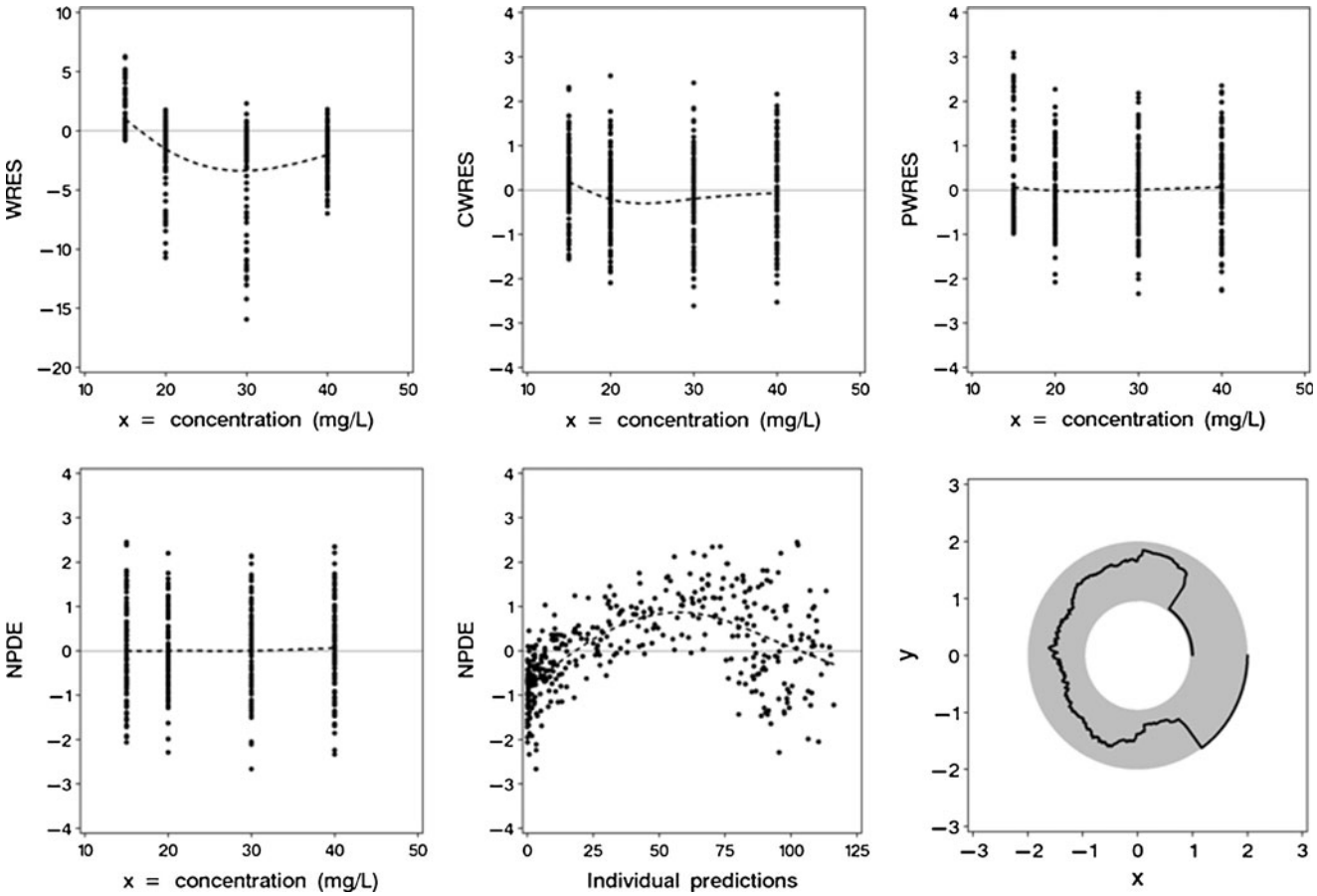

Fig. 5 Plots of weighted residuals *vs.* the independent variable *x* (*top left*: WRES, *top middle*: CWRES, *top right*: PWRES) and NPDE *vs*. *x* (*bottom left*) or individual model predictions (*bottom middle*) for a data set simulated with the population PD model. The last plot (*bottom right*) is the Q–Q ring plot obtained for the same simulated data set. As the black curve, representing observations, is fully contained in the ring (*grey area*), the model is not rejected with our method.

DISCUSSION

In this article, we propose a new method and global test (the GUD test) for the evaluation of population PK/PD models that correctly handle the issue of data dependency within subjects. Several simulation studies were conducted under the null hypothesis (H_0) that the tested model is correct to assess the level of our test and compared its performance with commonly used weighted residuals and NPDE.

A common belief is that weighted residuals (8,9) and NPDE (11,16) are $N(0,I)$ distributed under the true model. Our simulation studies show that this is only true for linear Gaussian models, not for nonlinear mixed effects models. These results directly corroborate the theory, i.e. for linear Gaussian models, individual vectors of observations are Gaussian; therefore, weighted residuals are Gaussian and are independent (decorrelation implies independence for Gaussian data). Their joint distribution under the model is thus an $N(0,I)$. As NPDE are computed from independent weighted residuals, they are also independent and their joint distribution is also an $N(0,I)$. It is noteworthy that in that case, the FO/FOCE linearisation does not result in any model approximation since the model is already linear in η_i. CWRES are therefore identical to WRES, and $E(Y_i)$ and $Var(Y_i)$ are accurately calculated for all types of residuals, either explicitly or by the means of simulations. For nonlinear mixed effects models, however, the distribution of observations is rarely Gaussian, which implies that the joint distribution of weighted residuals (PWRES) or NPDE cannot be an $N(0,I)$. The high type I errors observed for WRES and CWRES in our simulation studies were mainly due to model approximation. The GUD test was the only one of the evaluated tests to show adequate type I errors for all models and designs investigated.

It is noteworthy that a global test on weighted residuals is seldom performed in everyday practice and that modellers rely much more on the inspection of the residual plots where they search for trends. As illustrated by Fig. 5, we show that these diagnostic plots can be quite misleading. For WRES and CWRES, residuals may not be homoscedastic and centred on zero as a consequence of the

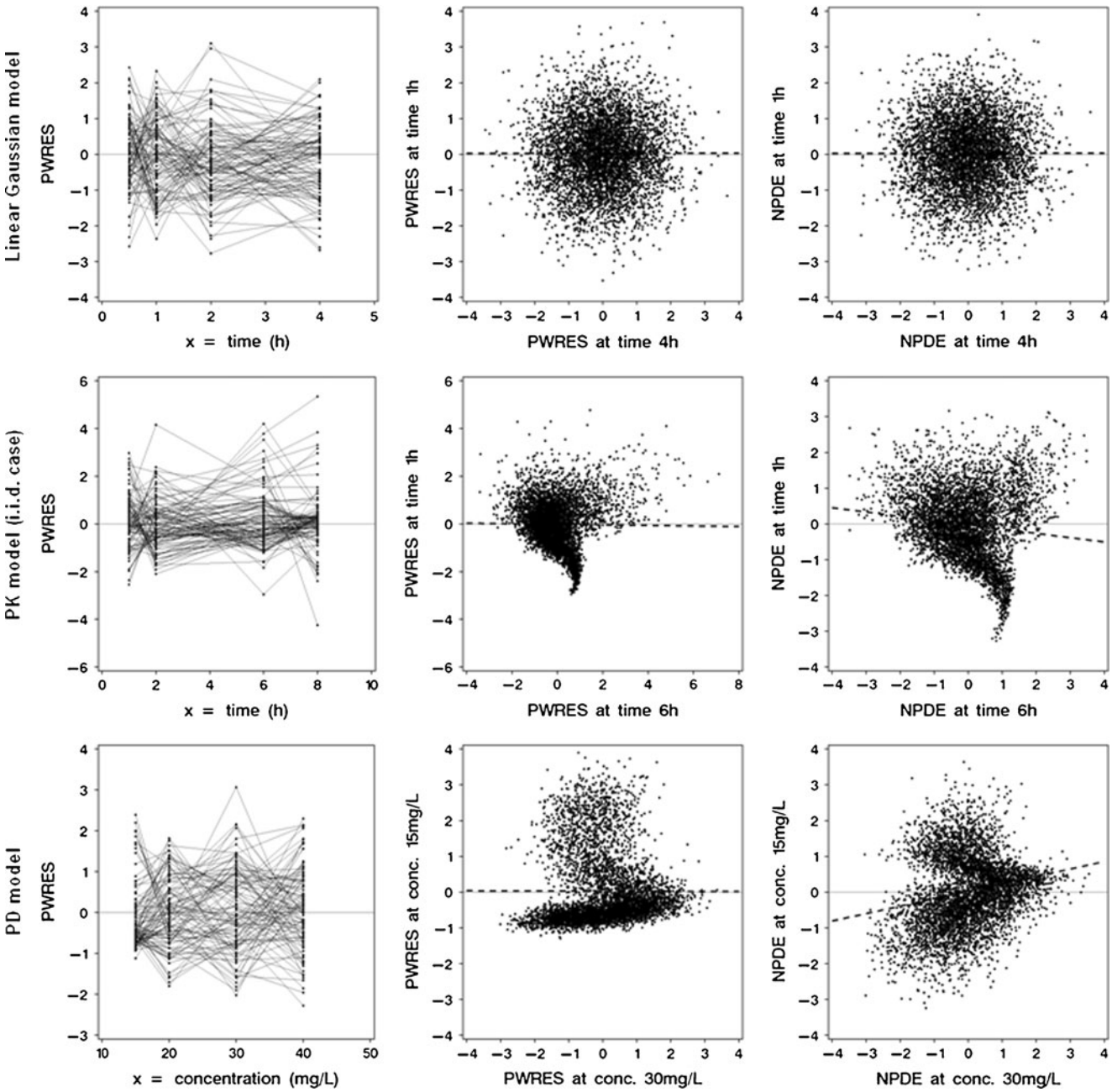

Fig. 6 Illustration of data dependency within subjects for the different models investigated: the linear Gaussian model (*top*), the population PK model (*middle*) and the population PD model (*bottom*). The graphs shows from *left* to *right* the plots of weighted residuals (PWRES) *vs.* the independent variable *x*, PWRES at one value of *x vs*. PWRES at another value of *x*, and NPDE at one value of *x vs*. NPDE at another value of *x* (the last two graphs were performed in 5,000 subjects instead of 100 to clearly illustrate data dependency). The dashed line refers to a linear regression line through the data. A horizontal dashed line indicates that the data are uncorrelated. A rounded shape for the data indicates that the data are independent.

approximation of the model. This approximation can be very crude for WRES, which makes WRES a very poor diagnostic tool, as has also been reported by previous studies (9,11,14). By construction, PWRES are expected to be homoscedastic and centred on zero; however, the non-normality of the data and their dependency within subjects might create artificial patterns (e.g. asymmetric distribution, presence of outliers). Many different kinds of residuals are provided by population PK/PD software, especially NONMEM version 7, which proposes an extensive list including NWRES, WRESI CWRES, CWRESI, EWRES, ECWRES (13). Actually, the distribution of these residuals under the model is *a priori* unknown (it depends on the model and study design), and unless simulations are done to give some indication of this distribution and improve decision making (10), it is quite difficult to interpret residual diagnostic plots in a straightforward manner. Note that weighted residuals are sometimes called "standardised prediction errors" in publications. The

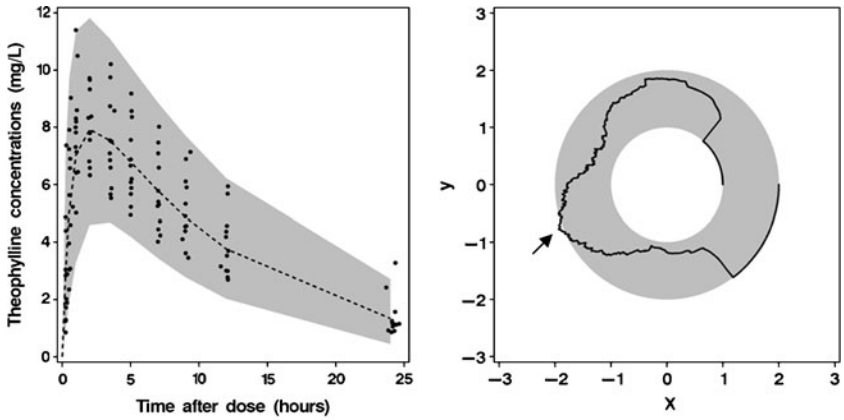

Fig. 7 GUD test applied to the evaluation of a previously developed model built on theophylline concentration-time data. The plot on the *left* represents observations (*dots*) superimposed with their predictive distribution under the model, which is summarised by 90% prediction intervals (*grey* area) and the median (*dashed line*) at each theoretical time point. The Q–Q ring plot is displayed on the *right*. As indicated by the *arrow*, the *black curve* representing observations is not fully contained within the ring (*grey area*), indicating that the model is rejected.

term "weighted residuals" is usually employed when residuals are computed with the dataset used for model estimation, while standardised prediction errors are computed using a different dataset in the context of external model evaluation (11).

Regarding NPDE, using two simple models revealed their limitations, and one may wonder what happens with more complex models. It is nonetheless true that NPDE should work in several situations, as indicated by previous studies (11,14,16). The problem is that detecting situations where they work is not so easy. From our experience, this depends on the dependency structure of NPDE. We have found in additional studies (not shown) that a negative correlation of the data leads to a decrease in type I errors for the KS test, while a positive correlation of the data leads to an increase in type I errors. We have also found that the extent of increase/decrease in type I errors depends on the magnitude of the correlation. This supports what we have observed in the present simulation studies, i.e. for the PK model, the decrease in type I errors for NPDE was associated with a negative correlation of NPDE, while for the PD model, the increase in type I errors was associated with a positive correlation of NPDE. As the sign and magnitude of NPDE correlation seem to depend both on the model and the study design, it may be very difficult to predict what will be the level of the test because it varies between models and datasets. The impact of data dependency on NPDE performance has been rightly discussed by the authors (16) but no real solution has been proposed. When there is some departure from the standard Gaussian distribution, it is advocated that the NPDE distribution be inspected graphically to see whether, in spite of a significant p value, the model may be considered as sufficient to describe the data. More precisely, it is recommended to plot NPDE *vs.* x (time) or *vs.* model predictions, assuming that no trend should be visible on the plots with the correct model (16). Although this assumption holds for the first plot (as subjects are independent, NPDE are expected to follow an $N(0,\mathrm{I})$ distribution at any value of x), this is not true for the plot of NPDE *vs.* model predictions as shown by Fig. 5. Indeed, due to NPDE dependency within subjects, we expect some pattern in the data provided the number of observations in the sample is large enough to detect such a pattern. It is noteworthy that by plotting NPDE *vs.* time or model predictions, we then come back to a subjective evaluation of the model and in that sense do not perform better than weighted residuals.

It is thus crucial to handle properly the issue of data dependency within subjects for the evaluation of population PK/PD models. The only way to get independent data is to reduce the dimension of each individual vector of observations to one. Dimensionality reduction has been a hot topic for the past few decades, and one of the most popular and contemporary techniques is the use of random projections. Random projections have been successfully applied to many different areas such as imaging, text processing, machine learning, optimisation, genomics and data mining in general (see 18–21 for some applications). Contrary to traditional dimensionality reduction methods, this technique is data independent. The idea is simply to project the data onto a chosen subspace of lower dimension: in the present case, we have chosen to project individual vectors of observations (standardised residuals to be exact) onto the unit sphere in many directions. We hence transform each individual vector of observations into a random variable whose distribution "partially" characterises the distribution of the observation vector provided enough projections are performed. We then obtain for the sample a mixture of individual distributions of random projections, and even in the case where subjects are non-identically distributed (e.g. with different measurements times and covariates), it is possible to determine the law of this mixture by computing

its empirical cdf C. We then compare the cdf obtained for the sample to its distribution under the model (assessed by simulation). For that purpose, we need to evaluate the distance between the sample cdf (C) and the mean cdf under the model (C_{mean}). Many metrics exist to express a distance between two distributions, but we have chosen the sup-norm (i.e. the maximal absolute distance) as it is easily and rapidly calculated. Furthermore, using the sup-norm, we are sure to only reject the model when it is really inadequate. We called this distance GUD (for Global Uniform Distance) as it allows the computation of a 95% prediction area under the model, which is uniform and constitutes the basis of our test. This test evaluates whether the sample cdf C belongs to this 95% prediction area. By construction, it is exact at a finite distance on condition that enough simulations are performed to determine the 95% prediction area with sufficient precision. The term "at a finite distance" means that no asymptotic condition is required to perform the test and that the level of the test is guaranteed irrespective of the study design and sample size. These good theoretical properties were confirmed by our simulation studies. In these studies, a p-value was calculated by comparing the GUD value obtained for the sample to its empirical distribution under the model, which is completely equivalent to comparing the sample cdf C to its empirical distribution under the model. For all the models investigated (linear or nonlinear) and all the designs (i.i.d. and non- i.i.d. subjects), the distribution of the p-values was uniform, which means that the level of the test was good whatever the nominal level.

It is noteworthy that dimensionality reduction is always accompanied by a loss of information but the important point is that random projections retain the approximate level of separation between the individual vectors of observations. In other words, individual distributions of random projections should represent the "fingerprint" of each subject on the unit sphere. In this respect, the choice of the distribution for the directions of projection is important. Here, it appeared natural to select the uniform distribution because it is isotropic (i.e. each direction has the same probability to be chosen), although other distributions could be chosen. With a uniform distribution, all components of W_i are treated equally. This results in the distribution of random projections for subject i not being affected by permuting W_i components, which may result in a loss of power. Another point is that the test is not performed on each individual distribution but on a mixture of individual distributions. Under these conditions, it is likely that the test will only reject truly inadequate models, and further work is necessary to assess the power of the test for a number of alternative hypotheses, using different models and study designs. We have also assumed that model parameter estimates are known with certainty. Although this approach is commonly used for evaluation of population PK/PD models and may be efficient in some cases (24), this is obviously not true. Note that it is possible to account for the uncertainty in parameter estimates $\widehat{\theta}$, $\widehat{\Omega}$ and $\widehat{\sigma}$ by computing their posterior distribution and drawing samples from this distribution when evaluating the 95% prediction area A. We will not go into details as it is not the scope of the paper, but the reader can refer to (24) regarding the different ways to compute the posterior distribution of $\widehat{\theta}$, $\widehat{\Omega}$ and $\widehat{\sigma}$.

As for all simulation-based methods, computational burden may be a limiting factor for the applicability of our method. Here, the objective was to evaluate precisely the level of the test, so we deliberately used a very high number of simulations: K was set to 2,000 based on (16), while B and M were set to 100 and 5,000, respectively. With these settings and given our numerical implementation, computation time was around 3.5 h for the population PK and PD models tested, with 100 subjects per sample and four observations per subject. For the application to real data (12 subjects with ten observations per subject), computation time was only 14 min. It is obvious that computation time will be less an issue in the future as the computational performance of computers increases. Furthermore, B and M can be optimised for large sample sizes, which will be investigated in further work. Note that part of the simulations required to do the test can be performed in population PK/PD software such as NONMEM, as has been done for the evaluation of the theophylline model. With complex observational designs (e.g. adaptative designs, drop-outs, censored or missing data), simulations *per se* may be a difficult task, which constitutes another limitation. In that case however, it seems that all simulation-based methods would be unhelpful (10).

Finally, the "Q–Q-ring" plot was created to detect a potential departure from the theoretical distribution under H_0. This plot is a quantile *vs.* quantile plot but uses polar coordinates to transform the 95% prediction area A into a ring. Thereby, each part of the plot is treated equally so there is more chance to detect a departure from the theoretical distribution than on a classical Q–Q plot. With the Q–Q ring plot, model diagnostics is straightforward: either the sample curve is somewhere outside the ring and the model is rejected with a 5% risk of being wrong, or the sample curve is fully contained in the ring and the model is not rejected. Furthermore, contrary to VPC (10), the Q–Q ring plot allows global model evaluation whatever the model (absence or presence of covariates) and the study design (i.e. doses and measurement times may vary among subjects). The counterpart is that it does not indicate where the misfit occurs when the model is wrong. If we use the terminology of Sheiner (25), the Q–Q ring plot can be viewed as an evaluation procedure for confirming, while the VPC is more an evaluation procedure for learning during the model building process.

It is obvious that all models are wrong and that all tests will reject the model when the sample size is large enough. What actually matters is to select models that predict data sufficiently well with regard to the objective of the analysis. In other words, the modeller may not come with the true model (model building involving many subjective steps) but this model can nevertheless be considered as useful. One of the challenges in coming years is thus the development of "bioequivalence"-like tests that accept useful models while still rejecting unacceptable ones. Clearly, this cannot be achieved with approximate methods whose performances vary among models or study designs. This implies the selection of methods with good statistical performances and the adaptation of testing hypotheses according to a bioequivalence approach. We believe that our method, based on random projections, could be a reasonable starting point for the elaboration of such tests as it is objective and as the nominal risk is controlled.

CONCLUSION

We show that data dependency within subjects has a strong impact on the evaluation of population PK/PD models and that even NPDE do not correctly address this issue. It is noteworthy that a global test on weighted residuals is seldom performed in everyday practice and that modellers rely much more on the inspection of the residual plots where they search for trends. However, these plots can be quite misleading and one single plot with trends might raise doubt about the model even though the latter is correct. The method we propose is the first to handle data dependency within subjects with the construction of an exact test and an easily interpretable diagnostic plot. The good theoretical properties of our test were confirmed through simulation studies using different models and study designs. Further investigations are now required to assess the power of this test under many different scenarios. In conclusion, our method allows a global, objective and easy evaluation of the model and could be used in combination with other tools to drive model evaluation in population PK/PD analyses.

ACKNOWLEDGMENTS

We kindly thank the two anonymous reviewers for their useful comments which have improved the quality of the paper.

REFERENCES

1. Jochemsen R, Laveille C, Breimer DD. Application of pharmacokinetic/pharmacodynamic modelling and population approaches to drug development. Int J Pharm Med. 1999;13:243–51.
2. Aarons L, Karlsson MO, Mentré F, Rombout F, Steimer JL, van Peer A. Role of modelling and simulation in Phase I drug development. Eur J Pharm Sci. 2001;13(2):115–22.
3. Zhang L, Sinha V, Forgue ST, Callies S, Ni L, Peck R, *et al.* Model-based drug development: the road to quantitative pharmacology. J Pharmacokinet Pharmacodyn. 2006;33 (3):369–93.
4. Williams PJ, Ette EI. The role of population pharmacokinetics in drug development in light of the Food and Drug Administration's 'Guidance for Industry: population pharmacokinetics'. Clin Pharmacokinet. 2000;39(6):385–95.
5. US Food and Drug Administration. Guidance for Industry: Population Pharmacokinetics; 1999 February. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072137.pdf.
6. Committee for Medicinal Products for Human use (CHMP). Guideline on reporting the results of population pharmacokinetic analyses; 2007 June 21. Available from: http://www.tga.gov.au/DOCS/pdf/euguide/ewp/18599006en.pdf.
7. Duan JZ. Applications of population pharmacokinetics in current drug labelling. J Clin Pharm Ther. 2007;32(1):57–79.
8. Vozeh S, Uematsu T, Hauf GF, Follath F. Performance of Bayesian feedback to forecast lidocaine serum concentration: evaluation of the prediction error and the prediction interval. J Pharmacokinet Biopharm. 1985;13(2):203–12.
9. Hooker AC, Staatz CE, Karlsson MO. Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. Pharm Res. 2007;24(12):2187–97.
10. Karlsson MO, Savic RM. Diagnosing model diagnostics. Clin Pharmacol Ther. 2007;82(1):17–20.
11. Brendel K, Comets E, Laffont C, Laveille C, Mentré F. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. Pharm Res. 2006;23 (9):2036–49.
12. Lavielle M. MONOLIX (MOdèles Non LInéaires à effets miXtes) Version 3.1 Users Guide; 2009 September. Available from: http://software.monolix.org/download-31/Monolix31_UsersGuide.pdf.
13. Bauer RJ. NONMEM Users Guide Introduction to NONMEM 7. ICON Development Solutions, Ellicott City, Maryland; 2009
14. Brendel K, Comets E, Laffont C, Mentré F. Evaluation of different tests based on observations for external model evaluation of population analyses. J Pharmacokinet Pharmacodyn. 2010;37 (1):49–65.
15. Silber HE, Kjellsson MC, Karlsson MO. The impact of misspecification of residual error or correlation structure on the type I error rate for covariate inclusion. J Pharmacokinet Pharmacodyn. 2009;36(1):81–99.
16. Comets E, Brendel K, Mentré F. Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R. Comput Methods Programs Biomed. 2008;90(2):154–66.
17. Vempala SS. The random projection method, dimacs series in discrete mathematics and theoretical computer science, volume 65. American Mathematical Society; 2004.
18. Majumdar A, Ward RK. Robust classifiers for data reduced via random projections. IEEE Trans Syst Man Cybern B Cybern. 2010;40(5):1359–71.
19. Avogadri R, Valentini G. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artif Intell Med. 2009;45(2–3):173–83.
20. Quera V. RAP: a computer program for exploring similarities in behavior sequences using random projections. Behav Res Methods. 2008;40(1):21–32.

21. Li W, Bebis G, Bourbakis NG. 3-D object recognition using 2-D views. IEEE Trans Image Process. 2008;17(11):2236–55.
22. Boeckmann AJ, Sheiner LB, Beal SL. NONMEM user's guide, part V, introductory guide. San Francisco: University of California; 1994.
23. Dvoretzky A, Kiefer J, Wolfowitz J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann Math Statist. 1956;27(3):642–69.
24. Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001;28(2):171–92.
25. Sheiner LB. Learning *versus* confirming in clinical drug development. Clin Pharmacol Ther. 1997;61(3):275–91.